Data-Driven Methods and Applications for Optimization under Uncertainty and Rare-Event Simulation

by

Zhiyuan Huang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in The University of Michigan
2020

Doctoral Committee:

Assistant Professor Ruiwei Jiang, Co-chair
Associate Professor Henry Lam, Co-chair
Associate Professor Eunshin Byon
Assistant Professor Gongjun Xu
Assistant Professor Ding Zhao

Zhiyuan Huang

zhyhuang@umich.edu

ORCID iD: 0000-0003-1284-2128

Dedicated to my parents
Weiping Liu and Xiaolei Huang

ACKNOWLEDGEMENTS

First of all, I am deeply thankful to my advisor Professor Henry Lam for his guidance and support throughout my Ph.D. experience. It has been one of the greatest honor of my life working with him as a graduate student. His wisdom, kindness, dedication, and passion have encouraged me to overcome difficulties and to complete my dissertation.

I also would like to thank my dissertation committee members. I want to thank the committee co-chair Professor Ruiwei Jiang for his support in my last two Ph.D. years. I am also grateful to Professor Ding Zhao for his consistent help and collaboration. My appreciation also goes to Professor Gongjun Xu for his valuable comments on my dissertation. I would like to thank Professor Eunshin Byon for being helpful both academically and personally.

Next, I want to thank other collaborators in my Ph.D. studies. I am very fortunate to encounter great mentor and collaborator Professor Jeff Hong, who made great advice on Chapter 2 of this dissertation. His perspective on research has inspired my future studies. Moreover, I also want to thank Qi Luo, Mansur Arief, and Yaohui Guo, who have been great collaborators and friends.

Finally, I want to thank my family and my friends for their love and supports in these five years. Their accompany and encouragement have given me endless courage and motivation towards my goal.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

For most of decisions or system designs in practice, there exist chances of severe hazards or system failures that can be catastrophic. The occurrence of such hazards is usually uncertain, and hence it is important to measure and analyze the associated risks. As a powerful tool for estimating risks, rare-event simulation techniques are used to improve the efficiency of the estimation when the risk occurs with an extremely small probability. Furthermore, one can utilize the risk measurements to achieve better decisions or designs. This can be achieved by modeling the task into a chance constrained optimization problem, which optimizes an objective with a controlled risk level. However, recent problems in practice have become more data-driven and hence brought new challenges to the existing literature in these two domains. In this dissertation, we will discuss challenges and remedies in data-driven problems for rare-event simulation and chance constrained problems. We propose a robust optimization based framework for approaching chance constrained optimization problems under a data-driven setting. We also analyze the impact of tail uncertainty in data-driven rare-event simulation tasks.

On the other hand, due to recent breakthroughs in machine learning techniques, the development of intelligent physical systems, e.g. autonomous vehicles, have been actively investigated. Since these systems can cause catastrophes to public safety, the evaluation of their machine learning components and system performance is crucial. This dissertation will cover problems arising in the evaluation of such systems.

We propose an importance sampling scheme for estimating rare events defined by machine learning predictors. Lastly, we discuss an application project in evaluating the safety of autonomous vehicle driving algorithms.

CHAPTER I

Introduction

For most of decisions or system designs in practice, there exist chances of severe hazards or system failures that can be catastrophic. The occurrence of such hazards is usually uncertain, and hence it is important to measure and analyze the associated risks. As a powerful tool for estimating risks, rare-event simulation techniques are used to improve the efficiency of the estimation when the risk occurs with an extremely small probability. Furthermore, one can utilize the risk measurements to achieve better decisions or designs. This can be achieved by modeling the task into a chance constrained optimization problem, which optimizes an objective with a controlled risk level. For example, in financial management, we can use the probability of large loss as a quantitative measure to assess the risk of the portfolio. In order to pursue better profit under risks, we can design a portfolio that maximizes the expected return while constrains the risk of large loss to be low.

Rare-event estimation and chance constrained programming have been extensively studied and developed in the last few decades. However, recent problems in practice have become more data-driven and hence brought new challenges to the existing literature in these two domains. In this dissertation, we will discuss challenges and remedies in data-driven problems for rare-event simulation and chance constrained problems. More specifically, in Chapter II we propose a robust optimization based framework for approaching chance constrained optimization problems under data-driven settings. In Chapter III we analyze the impact of tail uncertainty in data-driven rare-event simulation tasks. Partial results of these studies have been published in [86, 90].

On the other hand, due to recent breakthroughs in machine learning techniques, the development of intelligent physical systems, e.g. autonomous vehicles, have been actively investigated. Since these systems can cause catastrophes to public safety, the evaluation of their machine learning components and system performance is crucial. This dissertation will cover two problems arising in the evaluation of such systems. In Chapter IV, we propose an importance sampling scheme to estimate rare events defined by machine learning predictors. In Chapter V, we consider an application case in evaluating the safety of autonomous vehicle driving algorithms. The detailed overview of each chapter is given as follows. Partial results of these studies have been published in [92, 91].

In Chapter II, we propose the robust optimization based framework under a data-driven setting. Robust optimization is a common approach to tractably obtain safeguarding solutions for optimization problems with uncertain constraints. In this chapter, we study a statistical framework to integrate data into robust optimization (RO), based on learning a prediction set using (combinations of) geometric shapes that are compatible with established RO tools, and a simple data-splitting validation step that achieves finite-sample nonparametric statistical guarantees on feasibility. We demonstrate how our required sample size to achieve feasibility at a given confidence level is independent of the dimensions of both the decision space and the probability space governing the stochasticity, and discuss some approaches to improve the objective performances while maintaining these dimension-free statistical feasibility guarantees.

In Chapter III, we study the problem of designing good importance sampling (IS) schemes to simulate the probability that a sophisticated predictor, built for instance from an off-the-shelf machine learning toolbox, gives a prediction that exceeds a large threshold. This problem is motivated as a step towards building good learning algorithms that takes into account the extremal risks of the prediction. We provide a framework to design IS for two common machine learning models, namely random forest and a basic neural network. Our approach utilizes some available mathematical programming formulations to optimize over these models and a simple "cutting plane" idea to look for dominating points under Gaussian input distributions.

In Chapter IV, we analyze the impact of tail uncertainty in data-driven rare-event simulation. Rare-event probabilities and risk measures that quantify the likelihood of catastrophic or failure events can be sensitive to the accuracy of the underlying input models, especially regarding their tail behaviors. We investigate how the lack of tail information of the input can affect the output extremal measures, in relation to the level of data that are needed to inform the input tail. Using the basic setting of estimating the probability of the overshoot of an aggregation of i.i.d. input variables, we argue that heavy-tailed problems are much more vulnerable to input uncertainty than light-tailed problems. We explain this phenomenon via their large deviations behaviors, and substantiate with some numerical experiments.

In Chapter V, we discuss the safety evaluation of autonomous vehicle driving algorithms. Currently, the process to certify highly Automated Vehicles has not yet been defined by any country in the world. Companies are testing Automated Vehicles on public roads, which is time-consuming and inefficient. We proposed the Accelerated

Evaluation concept, which uses a modified statistics of the surrounding vehicles and the Importance Sampling theory to reduce the evaluation time by several orders of magnitude, while ensuring the evaluation results are statistically accurate. In this chapter, we further improve the accelerated evaluation concept by using Piecewise Mixture Distribution models, instead of Single Parametric Distribution models. We developed and applied this idea to forward collision control system reacting to vehicles making cut-in lane changes. The behavior of the cut-in vehicles was modeled based on more than 403,581 lane changes collected by the University of Michigan Safety Pilot Model Deployment Program. Simulation results confirm that the accuracy and efficiency of the Piecewise Mixture Distribution method outperformed single parametric distribution methods in accuracy and efficiency, and accelerated the evaluation process by almost four orders of magnitude.

CHAPTER II

Learning-based Robust Optimization

2.1 Introduction

Many optimization problems in industrial applications contain uncertain parameters in constraints where the enforcement of feasibility is of importance. This chapter aims to build procedures to find good-quality solutions for these problems that are tractable and statistically accurate for high-dimensional or limited data situations.

To locate our scope of study, we consider situations where the uncertainty in the constraints is “stochastic”, and a risk-averse modeler wants the solution to be feasible “most of the time” while not making the decision space overly conservative. One common framework to define feasibility in this context is via a chance-constrained program (CCP)

$$\text{minimize } f(x) \ \text{ subject to } \ P(g(x;\xi) \in \mathcal{A}) \geq 1 - \epsilon \tag{2.1}$$

where $f(x) \in \mathbb{R}$ is the objective function, $x \in \mathbb{R}^d$ is the decision vector, $\xi \in \mathbb{R}^m$ is a random vector (i.e. the uncertainty) under a probability measure P, and $g(x;\xi) : \mathbb{R}^d \times \mathbb{R}^m \to \Omega$ with $\mathcal{A} \subset \Omega$ for some space Ω. Using existing terminology, we sometimes call $g(x;\xi) \in \mathcal{A}$ the safety condition, and ϵ the tolerance level that controls the violation probability of the safety condition. In this chapter we will consider $g(x;\xi) \in \mathcal{A}$ as linear inequalities, which constitute the commonest class of CCPs.

We will focus on settings where ξ is observed via a finite amount of data, driven by the fact that in almost every application there is no exact knowledge about the uncertainty, and that data is increasingly ubiquitous. Our problem target is to find a solution feasible for (2.1) with a given statistical confidence (with respect to the data, in a frequentist sense) that has an objective value as small as possible.

First proposed by [44], [43], [123] and [141], the CCP framework (2.1) has been studied extensively in the stochastic programming literature (see [142] for a thorough introduction), with applications spanning across reservoir system design ([144, 143]), cash matching ([55]), wireless cooperative network ([156]), inventory ([107]) and production management ([126]). Though not always proper (notably when the uncertainty is deterministic or bounded; see e.g., [12] P.28–29), in many situations it is natural to view uncertainty as "stochastic", and (2.1) provides a rigorous definition of feasibility under these situations. Moreover, (2.1) sets a framework to assimilate data in a way that avoids over-conservativeness by focusing on the "majority" of the data, as we will exploit in this chapter.

Our main contribution is a framework to integrate data into robust optimization (RO) as a tool to obtain high-quality solutions feasible in the sense defined by (2.1). Instead of directly solving (2.1), which is known to be challenging in general, RO operates by representing the uncertainty via a (deterministic) set, often known as the uncertainty set or the ambiguity set, and enforces the safety condition to hold for any ξ within it. By suitably choosing the uncertainty set, RO is well-known to be a tractable approximation to (2.1). We will revisit these ideas by studying a procedural framework to construct an uncertainty set as a *prediction set* for the data. This consists of approximating a high probability region via combinations of tractable geometric shapes compatible with RO. As a key development, we propose a simple data-splitting scheme to determine the size of this region that ensures rigorous statistical performance. This framework is nonparametric and applies under minimal distributional requirements.

In terms of basic statistical property, our approach satisfies a finite-sample confidence guarantee on the feasibility of the solution in which the minimum required sample size in achieving a given confidence is provably *independent* of the dimensions of both the decision space and the underlying probability space. While finite-sample guarantees are also found in existing sampling-based methods, the dimension-free property of our approach makes it a suitable resort for certain high-dimensional and limited-data situations where previous methods break down.

The above property, which may appear very strong, needs nonetheless be complemented with good approaches to curb over-conservativeness and maintain tractability. In particular, to reduce conservativeness, a prediction set should accurately trace the shape of data. On the other hand, to retain tractability, the set should be expressible in terms of basic geometric shapes compatible with RO techniques. We will present some techniques to construct uncertainty sets that balance these two aspects, while simultaneously achieve the basic statistical property. Nonetheless, we caution that theses techniques tie conservativeness to the set volume, while often times the former is more intricate and depends on the optimization setting at hand (see, e.g., [100]). Along this line, we also discuss a method to iterate the construction of uncertainty sets that incorporate updated optimality beliefs to improve the objective performance.

Our approach is related to several existing methods for approximating (2.1). Scenario generation (SG), pioneered by [31, 33, 37, 38] and independently suggested in the context of Markov decision processes by [50], replaces the chance constraint in

(2.1) with a collection of sampled constraints. Related work include also the sample average approximation (SAA) studied in [116, 117, 115], which restricts the proportion of violated constraints and resembles the discarding approach in [38]. SG provides explicit statistical guarantees on the feasibility of the obtained solution in terms of the confidence level, the tolerance level and the sample size. It directly approximates the chance-constrained optimization without the need of a set-based representation of the uncertainty, and hence allows a high geometric flexibility in the resulting set of violation and leads to less conservative solutions. However, in general, the sample size needed to achieve a given confidence grows linearly with the dimension of the decision space, which can be demanding for large-scale problems (as pointed out by, e.g., [132], P.971). Recent work reduce dependence on the decision dimension (and its interplay with the tolerance parameter) by, for instance, regularization ([36]), tighter complexity results in terms of the support rank ([152]), solution-dependent number of support constraints ([39]), one-off calibration schemes ([40]), sequential validation ([34, 42, 32]), and hybrid approaches between RO and SG that translate scenario size requirements from decision to stochasticity space dimension ([121]). Among these, our proposed step to tune the set size is closest to the calibration approaches. However, instead of calibrating a solution obtained from a randomized program, we calibrate the coverage of an uncertainty set, and control conservativeness and tractability of the resulting RO through proper learning of its shape.

A classical approach to approximating (2.1) uses safe convex approximation (SCA), by replacing the intractable chance constraint with an inner approximating convex constraint (such that a solution feasible for the latter would also be feasible for the former) (e.g., [15, 131, 132]). This approach is intimately related to RO, as the approximating constraints are often equivalent to the robust counterparts (RC) of RO problems with properly chosen uncertainty sets (e.g., [12], Chapters 2 and 4). The statistical guarantees provided by these approximations come from probabilistic deviation bounds, which often rely on the stochastic assumptions and the constraint structure on a worst-case basis (e.g., [132], [12] Chapter 10, [13, 14, 63, 19, 20, 18, 46, 35]). Thus, although the approach carries several advantages (e.g., in handling extraordinarily small tolerance levels), the utilized bounds can be restrictive to use in some cases. Moreover, most of the results apply to a single chance constraint; when the safety condition involves several constraints that need to be jointly maintained (known as a joint chance constraint), one typically needs to reduce it to individual constraints via the Bonferroni correction, which can add pessimism (there are exceptions, however; e.g., [45]). On the other hand, these classical results in SCA and RO are capable of constructing uncertainty sets with well-chosen shapes, without directly using prediction set properties.

We mention two other lines of work in approximating (2.1) that can blend with data. Distributionally robust optimization (DRO), an approach dated back to [151] and of growing interest in recent years (e.g., [51, 170, 74, 11, 112]), considers using a worst-case probability distribution for ξ within an ambiguity set that represents partial distributional information. The two major classes of sets consist of distance-based constraints (statistical distance from a nominal distribution such as the empirical distribution; e.g., [11, 169]) and moment-and-support-type constraints (including moments, dispersion, covariance and/or support, e.g., [51, 170, 74, 81], and shape and unimodality, e.g., [140, 80, 162, 108, 104]). To provide statistical feasibility guarantee, these uncertainty sets need to be properly calibrated from data, either via direct estimation or using the statistical implications from Bayesian ([78])

or empirical likelihood ([105, 59, 23, 102]) methods. Another line of work takes a Monte Carlo viewpoint and uses sequential convex approximation ([87, 89]) that stochastically iterates the solution to a Karush-Kuhn-Tucker (KKT) point, which guarantees local optimality of the convergent solution. This approach can be applied to data-driven situations by viewing the data as Monte Carlo samples.

Finally, some recent RO-based approaches aim to utilize data more directly. For example, [75] calibrate uncertainty sets using linear regression under Gaussian assumptions. [17] study a tight value-at-risk bound on a single constraint and calibrate uncertainty sets via imposing a confidence region on the distributions that govern the bound. [160] study supervised prediction models to approximate uncertainty sets and suggest using sampling or relaxation to reduce to tractable problems. Our approach follows the general idea in these work in constructing uncertainty sets that cover the "truth" with high confidence.

The rest of this chapter is organized as follows. Section 2.2 presents our procedural framework and statistical implications. Section 2.3 discusses some approaches to construct tight and tractable prediction sets. Section 2.4 reports numerical results and comparisons with existing methods. Additional proofs, numerical results and useful existing theorems are presented in the rest sections.

2.2 Basic Framework and Implications

This section lays out our basic procedural framework and implications. First, consider an approximation of (2.1) via the RO:

$$\text{minimize } \ f(x) \ \text{ subject to } \ g(x;\xi) \in \mathcal{A} \ \ \forall\, \xi \in \mathcal{U} \tag{2.2}$$

where $\mathcal{U} \in \Omega$ is an uncertainty set. Obviously, for any x feasible for (2.2), $\xi \in \mathcal{U}$ implies $g(x;\xi) \in \mathcal{A}$. Therefore, by choosing $\mathcal{U}$ that covers a $1-\epsilon$ content of ξ (i.e.,

$\mathcal{U}$ satisfies $P(\xi \in \mathcal{U}) \geq 1 - \epsilon$), any x feasible for (2.2) must satisfy $P(g(x;\xi) \in \mathcal{A}) \geq P(\xi \in \mathcal{U}) \geq 1 - \epsilon$, implying that x is also feasible for (2.1). In other words,

Lemma II.1. *Any feasible solution of* (2.2) *using a* $(1-\epsilon)$*-content set* $\mathcal{U}$ *is feasible for* (2.1)*.*

Note that [12], P.33 discussion point B points out that it is not necessary for an uncertainty set to contain most values of the stochasticity to induce probabilistic guarantees. Nonetheless, Lemma II.1 provides a platform to utilize data structure easily and formulate concrete procedures, as we will describe.

2.2.1 Learning Uncertainty Sets

Assume a given i.i.d. data set $D = \{\xi_1, \ldots, \xi_n\}$, where $\xi_i \in \mathbb{R}^m$ are sampled under a continuous distribution P. In view of Lemma II.1, our basic strategy is to construct $\mathcal{U} = \mathcal{U}(D)$ that is a $(1-\epsilon)$-content prediction set for P with a prescribed confidence level $1 - \delta$. In other words,

$$\mathbb{P}_D\left(P(\xi \in \mathcal{U}(D)) \geq 1 - \epsilon\right) \geq 1 - \delta \tag{2.3}$$

where we use the notation $\mathbb{P}_D(\cdot)$ to denote the probability taken with respect to the data D. Using such a $\mathcal{U}$, any feasible solution of (2.2) is feasible for (2.1) with the same confidence level $1 - \delta$, i.e.,

Lemma II.2. *Any feasible solution of* (2.2) *using* $\mathcal{U}$ *that satisfies* (2.3) *is feasible for* (2.1) *with confidence* $1 - \delta$*.*

(2.3) only focuses on the feasibility guarantee for (2.1), but does not speak much about conservativeness. To alleviate the latter issue, we judiciously choose $\mathcal{U}$ according to two criteria:

1. We prefer $\mathcal{U}$ that has a smaller volume, which leads to a larger feasible region in (2.2) and hence a less conservative inner approximation to (2.1). Note that, with a fixed ϵ, a small $\mathcal{U}$ means a $\mathcal{U}$ that contains a high probability region (HPR) of ξ.

2. We prefer $\mathcal{U}$ such that $P(\xi \in \mathcal{U}(D))$ is close to, not just larger than, $1-\epsilon$ with confidence $1-\delta$. We also want the coverage probability $\mathbb{P}_D(P(\xi \in \mathcal{U}(D)) \geq 1-\epsilon)$ to be close to, not just larger than, $1-\delta$.

Moreover, $\mathcal{U}$ needs to be chosen to be compatible with tractable tools in RO. Though this tractability depends on the type of safety condition at hand and is problem-specific, the general principle is to construct $\mathcal{U}$ as an HPR that is expressed via a basic geometric set or a combination of them.

The above discussion motivates us to propose a two-phase strategy in constructing $\mathcal{U}$. We first split the data D into two groups, denoted D_1 and D_2, with sizes n_1 and n_2 respectively. Say $D_1 = \{\xi_1^1, \dots, \xi_{n_1}^1\}$ and $D_2 = \{\xi_1^2, \dots, \xi_{n_2}^2\}$. These two data groups are used as follows:

<u>Phase 1: Shape learning.</u> We use D_1 to approximate the shape of an HPR. Two common choices of tractable basic geometric shapes are:

1. *Ellipsoid:* Set the shape as $\mathcal{S} = \{(\xi - \mu)'\Sigma^{-1}(\xi - \mu) \leq \rho\}$ for some $\rho > 0$. The parameters can be chosen by, for instance, setting μ as the sample mean of D_1 and Σ as some covariance matrix, e.g., the sample covariance matrix, diagonalized covariance matrix, or identity matrix.

2. *Polytope:* Set the shape as $\mathcal{S} = \{\xi : a_i'\xi \leq b_i, i = 1, \dots, k\}$ where $a_i \in \mathbb{R}^m$ and $b_i \in \mathbb{R}$. For example, for low-dimensional data, this can be obtained from a convex hull (or an approximated version) of D_1, or alternately, of the data

that leaves out $\lfloor n_1\epsilon \rfloor$ of D_1 that are in the “periphery”, e.g., having the smallest Tukey depth (e.g., [154, 79]). It can also take the shape of the objective function when it is linear (a case of interest when using the self-improving strategy that we will describe later).

We can also combine any of the above two types of geometric sets, such as:

1. *Union of basic geometric sets:* Given a collection of polytopes or ellipsoids $\mathcal{S}_i$, take $\mathcal{S} = \bigcup_i \mathcal{S}_i$.
2. *Intersection of basic geometric sets:* Given a collection of polytopes or ellipsoids $\mathcal{S}_i$, take $\mathcal{S} = \bigcap_i \mathcal{S}_i$.

The choices of ellipsoids and polytopes are motivated from the tractability in the resulting RO, but they may not describe an HPR of ξ to sufficient accuracy. Unions or intersection of these basic geometric sets provide more flexibility in tracking the HPR of ξ. For example, in the case of multi-modal distribution, one can group the data into several clusters ([83]), then form a union of ellipsoids over the clusters as $\mathcal{S}$. For non-standard distributions, one can discretize the space into boxes and take the union of boxes that contain at least some data, inspired by the “histogram” method in the literature of minimum volume set learning ([153]). The intersection of basic sets is useful in handling segments of ξ where each segment appears in a separate constraint in a joint CCP.

Phase 2: Size calibration. We use D_2 to calibrate the size of the uncertainty set so that it satisfies (2.3) and moreover $P(\xi \in \mathcal{U}(D)) \approx 1-\epsilon$ with coverage $\approx 1-\delta$. The key idea is to use quantile estimation on a “dimension-collapsing” transformation of the data. More concretely, first express our geometric shape obtained in Phase 1 in the form $\{\xi : t(\xi) \leq s\}$, where $t(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is a transformation map from the space of ξ to $\mathbb{R}$, and $s \in \mathbb{R}$. For the two geometric shapes we have considered above,

1. *Ellipsoid:* We set $t(\xi) = (\xi - \mu)'\Sigma^{-1}(\xi - \mu)$. Then the $\mathcal{S}$ described in Phase 1 is equivalent to $\{\xi : t(\xi) \leq \rho\}$.
2. *Polytope:* Find a point, say μ, in $\mathcal{S}^\circ$, the interior of $\mathcal{S}$ (e.g., the Chebyshev center ([28]) of $\mathcal{S}$ or the sample mean of D_1 if it lies in $\mathcal{S}^\circ$). Let $t(\xi) = \max_{i=1,\ldots,k}(a_i'(\xi - \mu))/(b_i - a_i'\mu)$ which is well-defined since $\mu \in \mathcal{S}^\circ$. Then the $\mathcal{S}$ defined in Phase 1 is equivalent to $\{\xi : t(\xi) \leq 1\}$.

For the combinations of sets, we suppose each individual geometric shape $\mathcal{S}_i$ in Phase 1 possesses a transformation map $t_i(\cdot)$. Then,

1. *Union of the basic geometric sets:* We set $t(\xi) = \min_i t_i(\xi)$ as the transformation map for $\bigcup_i \mathcal{S}_i$. This is because $\bigcup_i \{\xi : t_i(\xi) \leq s\} = \{\xi : \min_i t_i(\xi) \leq s\}$.
2. *Intersection of the basic geometric sets:* We set $t(\xi) = \max_i t_i(\xi)$ as the transformation map for $\bigcap_i \mathcal{S}_i$. This is because $\bigcap_i \{\xi : t_i(\xi) \leq s\} = \{\xi : \max_i t_i(\xi) \leq s\}$

We overwrite the value of s in the representation $\{\xi : t(\xi) \leq s\}$ as $t(\xi^2_{(i^*)})$, where $t(\xi^2_{(1)}) < t(\xi^2_{(2)}) < \cdots < t(\xi^2_{(n_2)})$ are the ranked observations of $\{t(\xi^2_i)\}_{i=1,\ldots,n_2}$, and

$$i^* = \min\left\{r : \sum_{k=0}^{r-1} \binom{n_2}{k} (1-\epsilon)^k \epsilon^{n_2-k} \geq 1-\delta,\ 1 \leq r \leq n_2\right\} \tag{2.4}$$

This procedure is valid if such an i^* can be found, or equivalently $1-(1-\epsilon)^{n_2} \geq 1-\delta$.

2.2.2 Basic Statistical Guarantees

Phase 1 focuses on Criterion 1 in Section 2.2.1 by learning the shape of an HPR. Phase 2 addresses our basic requirement (2.3) and Criterion 2. The choice of s in Phase 2 can be explained by the elementary observation that, for any arbitrary i.i.d. data set of size n_2 drawn from a continuous distribution, the i^*-th ranked observation as defined by (2.4) is a valid $1-\delta$ confidence upper bound for the $1-\epsilon$ quantile of the distribution:

Lemma II.3. *Let $Y_1, \dots, Y_{n_2}$ be i.i.d. data in $\mathbb{R}$ drawn from a continuous distribution. Let $Y_{(1)} < Y_{(2)} < \dots < Y_{(n_2)}$ be the order statistics. A $1-\delta$ confidence upper bound for the $(1-\epsilon)$-quantile of the underlying distribution is $Y_{(i^*)}$, where*

$$i^* = \min\left\{r : \sum_{k=0}^{r-1} \binom{n_2}{k}(1-\epsilon)^k \epsilon^{n_2-k} \geq 1-\delta,\ 1 \leq r \leq n_2\right\}$$

If $\sum_{k=0}^{n_2-1} \binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k} < 1-\delta$ or equivalently $1-(1-\epsilon)^{n_2} < 1-\delta$, then none of the $Y_{(r)}$'s is a valid confidence upper bound.

Similarly, a $1-\delta$ confidence lower bound for the $(1-\epsilon)$-quantile of the underlying distribution is $Y_{(i_*)}$, where

$$i_* = \max\left\{r : \sum_{k=r}^{n_2} \binom{n_2}{k}(1-\epsilon)^k \epsilon^{n_2-k} \geq 1-\delta,\ 1 \leq r \leq n_2\right\}$$

If $\sum_{k=1}^{n_2} \binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k} < 1-\delta$ or equivalently $1-\epsilon^{n_2} < 1-\delta$, then none of the $Y_{(r)}$'s is a valid confidence lower bound.

Proof. Proof of Lemma II.3. Let $q_{1-\epsilon}$ be the $(1-\epsilon)$-quantile, and $F(\cdot)$ and $\bar{F}(\cdot)$ be the distribution function and tail distribution function of Y_i. Consider

$$\begin{aligned}
P(Y_{(r)} \geq q_{1-\epsilon}) &= P(\leq r-1 \text{ of the data } \{Y_1, \dots, Y_n\} \text{ are } < q_{1-\epsilon}) \\
&= \sum_{k=0}^{r-1} \binom{n_2}{k} F(q_{1-\epsilon})^k \bar{F}(q_{1-\epsilon})^{n_2-k} \\
&= \sum_{k=0}^{r-1} \binom{n_2}{k} (1-\epsilon)^k \epsilon^{n_2-k}
\end{aligned}$$

by the definition of $q_{1-\epsilon}$. Hence any r such that $\sum_{k=0}^{r-1} \binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k} \geq 1-\delta$ is a $1-\delta$ confidence upper bound for $q_{1-\epsilon}$, and we pick the smallest one. Note that if $\sum_{k=0}^{n_2-1} \binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k} < 1-\delta$, then none of the $Y_{(r)}$ is a valid confidence upper bound.

Similarly, we have

$$\begin{aligned} P(Y_{(r)} \leq q_{1-\epsilon}) &= P(\geq r \text{ of the data } \{Y_1, \ldots, Y_n\} \text{ are } \leq q_{1-\epsilon}) \\ &= \sum_{k=r}^{n_2} \binom{n_2}{k} F(q_{1-\epsilon})^k \bar{F}(q_{1-\epsilon})^{n_2-k} \\ &= \sum_{k=r}^{n_2} \binom{n_2}{k} (1-\epsilon)^k \epsilon^{n_2-k} \end{aligned}$$

by the definition of $q_{1-\epsilon}$. Hence any r such that $\sum_{k=r}^{n_2} \binom{n_2}{k} (1-\epsilon)^k \epsilon^{n_2-k} \geq 1-\delta$ will be a $1-\delta$ confidence lower bound for $q_{1-\epsilon}$, and we pick the largest one. Note that if $\sum_{k=1}^{n_2} \binom{n_2}{k} (1-\epsilon)^k \epsilon^{n_2-k} < 1-\delta$, then none of the $Y_{(r)}$ is a valid confidence lower bound.

□

Similar results in the above simple order statistics calculation can be found in, e.g., [155] Section 2.6.1. A key element of our procedure is that $t(\cdot)$ is constructed using only Phase 1 data D_1, which are independent of Phase 2. Lemma II.3 implies that, conditional on D_1, $P(t(\xi) \leq t(\xi^2_{(i^*)})) \geq 1-\epsilon$ with a (conditional) confidence $1-\delta$. From this, we can average over the realizations of D_1 to obtain a valid coverage for the resulting uncertainty set in the sense of satisfying (2.3). This is summarized formally as:

Theorem II.4 (Basic statistical guarantee)**.** *Suppose D is an i.i.d. data set drawn from a continuous distribution P on $\mathbb{R}^m$, and we partition D into two sets $D_1 = \{\xi^1_i\}_{i=1,\ldots,n_1}$ and $D_2 = \{\xi^2_i\}_{i=1,\ldots,n_2}$. Suppose $n_2 \geq \log\delta/\log(1-\epsilon)$. Consider the set $\mathcal{U} = \mathcal{U}(D) = \{\xi : t(\xi) \leq s\}$, where $t : \mathbb{R}^m \to \mathbb{R}$ is a map constructed from D_1 such that $t(\xi)$, with ξ distributed according to P, is a continuous random variable, and $s = t(\xi^2_{(i^*)})$ is calibrated from D_2 with i^* defined in* (2.4)*. Then $\mathcal{U}$ satisfies* (2.3)*. Consequently, an optimal solution obtained from* (2.2) *using this $\mathcal{U}$ is feasible for*

(2.1) *with confidence* $1-\delta$*.*

Proof. Proof of Theorem II.4. Since $t(\cdot)$ depends only on D_1 but not D_2, we have, conditional on any realization of D_1,

$$\mathbb{P}_{D_2}(P(\xi \in \mathcal{U}(D)) \geq 1-\epsilon|D_1) = \mathbb{P}_{D_2}(q_{1-\epsilon} \leq t(\xi^2_{(i^*)})|D_1) \geq 1-\delta \tag{2.5}$$

where $q_{1-\epsilon}$ is the $(1-\epsilon)$-quantile of $t(\xi)$ (which depends on D_1). The first equality in (3.3) follows from the representation of $\mathcal{U} = \{\xi : t(\xi) \leq t(\xi^2_{(i^*)})\}$, the second equality uses the definition of a quantile, and the last inequality follows from Lemma II.3 using the condition $1-(1-\epsilon)^{n_2} \geq 1-\delta$, or equivalently $n_2 \geq \log\delta/\log(1-\epsilon)$. Note that (3.3) holds given any realization of D_1. Thus, taking expectation with respect to D_1 on both sides in (3.3), we have

$$\mathbb{E}_{D_1}[\mathbb{P}_{D_2}(P(\xi \in \mathcal{U}(D)) \geq 1-\epsilon|D_1)] \geq 1-\delta$$

where $\mathbb{E}_{D_1}[\cdot]$ denotes the expectation with respect to D_1, which gives

$$\mathbb{P}_D(P(\xi \in \mathcal{U}(D)) \geq 1-\epsilon) \geq 1-\delta$$

We therefore arrive at (2.3). Finally, Lemma II.2 guarantees that an optimal solution obtained from (2.2) using the constructed $\mathcal{U}$ is feasible for (2.1) with confidence $1-\delta$. □

Theorem II.4 implies the validity of the approach in giving a feasible solution for CCP (2.1) with confidence $1-\delta$ for any finite sample size, as long as it is large enough such that $n_2 \geq \log\delta/\log(1-\epsilon)$. The reasoning of the latter restriction can be seen easily in the proof, or more apparently from the following argument: In order to get an upper confidence bound for the quantile by choosing one of the ranked statistics, we need the probability of at least one observation to upper bound the quantile to be at least $1-\delta$. In other words, we need $P\,(\text{at least one } t(\xi_i^2) \geq (1-\epsilon)\text{-quantile}) \geq 1-\delta$ or equivalently $1-(1-\epsilon)^{n_2} \geq 1-\delta$.

We also mention the convenient fact that, conditional on D_1,

$$P(\xi \in \mathcal{U}) = P(t(\xi) \leq t(\xi^2_{(i^*)})) = F(t(\xi^2_{(i^*)})) \overset{d}{=} U_{(i^*)} \tag{2.6}$$

where $F(\cdot)$ is the distribution function of $t(\xi)$ and $U_{(i^*)}$ is the i^*-th ranked variable among n_2 uniform variables on $[0,1]$, and "$\overset{d}{=}$" denotes equality in distribution. In other words, the theoretical tolerance level induced by our constructed uncertainty set, $P(\xi \in \mathcal{U})$, is distributed as the i^*-th order statistic of uniform random variables, or equivalently $Beta(i^*, n_2-i^*+1)$, a Beta variable with parameters i^* and n_2-i^*+1. Note that $P(Beta(i^*, n_2-i^*+1) \geq 1-\epsilon) = P(Bin(n_2, 1-\epsilon) \leq i^*-1)$ where $Bin(n_2, 1-\epsilon)$ denotes a binomial variable with number of trials n_2 and success probability $1-\epsilon$. This informs an equivalent expression of (2.4) as

$$\begin{aligned} & \min\{r : P(Beta(r, n_2-r+1) \geq 1-\epsilon) \geq 1-\delta,\ 1 \leq r \leq n_2\} \\ = \ & \min\{r : P(Bin(n_2, 1-\epsilon) \leq r-1) \geq 1-\delta,\ 1 \leq r \leq n_2\} \end{aligned}$$

To address Criterion 2 in Section 2.2.1, we use the following asymptotic behavior as $n_2 \to \infty$:

Theorem II.5 (Asymptotic tightness of tolerance and confidence levels)**.** *Under the same assumptions as in Theorem II.4, we have, conditional on D_1:*

1. *$P(\xi \in \mathcal{U}) \to 1-\epsilon$ in probability (with respect to D_2) as $n_2 \to \infty$.*
2. *$\mathbb{P}_{D_2}(P(\xi \in \mathcal{U}) \geq 1-\epsilon | D_1) \to 1-\delta$ as $n_2 \to \infty$.*

Theorem II.5 confirms that $\mathcal{U}$ is tightly chosen in the sense that the tolerance level and the confidence level are held asymptotically exact. This can be shown by using (2.6) together with an invocation of the Berry-Essen Theorem ([61]) applied on the normal approximation to binomial distribution. Section 2.5 shows the proof details, which use techniques similar to [110] and [155] Section 2.6. In fact, one could further obtain that our choice of i^* satisfies $\sqrt{n_2}\,(i^*/n_2 - (1-\epsilon)) \to \sqrt{(1-\epsilon)\epsilon}\Phi^{-1}(1-\delta)$ as $n_2 \to \infty$. As a result, the theoretical tolerance level $P(\xi \in \mathcal{U})$ given D_1 concentrates at $1-\epsilon$ by being approximately $(1-\epsilon) + Z/\sqrt{n_2}$ where $Z \sim N\left(\sqrt{\epsilon(1-\epsilon)}\Phi^{-1}(1-\delta), \epsilon(1-\epsilon)\right)$. For further details, see Section 2.5.

Note that, because of the discrete nature of our quantile estimate, the theoretical confidence level is not a monotone function of the sample size, and neither is there a guarantee on an exact confidence level at $1-\delta$ using a finite sample (see Section 2.6). On the other hand, Theorem II.5 Part 2 guarantees that asymptotically our construction can achieve an exact confidence level.

The idea of using a dimension-collapsing transformation map $t(\cdot)$ resembles the notion of data depth in the literature of generalized quantile ([110, 154]). In particular, the data depth of an observation is a positive number that measures the position of the observation from the "center" of the data set. The larger the data depth, the closer the observation is to the center. For example, the half-space depth is the minimum number of observations on one side of any line passing through the chosen observation ([85, 159]), and the simplicial depth is the number of simplices formed by different combinations of observations surrounding an observation ([114]). Other common data depths include the ellipsoidally defined Mahalanobis depth ([119]) and projection-based depths ([58, 175]). Instead of measuring the position of the data relative to the center as in the data depth literature, our transformation map is constructed to create uncertainty sets with good geometric and tractability properties.

2.2.3 Dimension-free Sample Size Requirement

Theorem II.4 and the associated discussion above states that we need at least $n_2 \geq \log \delta / \log(1-\epsilon)$ observations in Phase 2 to construct an uncertainty set that guarantees a feasible solution for (2.1) with confidence $1-\delta$. From a *purely* feasibility viewpoint, this lower bound on n_2 is the minimum total sample size we need: Regardless of what shape we generate in Phase 1, as long as we can express it in terms of the $t(\cdot)$ and have $\log \delta / \log(1-\epsilon)$ Phase 2 observations, the basic feasibility guarantee (2.3) is attained. This number does not depend on the dimension of the decision space or the probability space. It does, however, depend roughly linearly on $1/\epsilon$ for small ϵ, a drawback that is also common among sampling-based approaches including both SG and SAA and gives more edge to using safe convex approximation when applicable.

We should caution, however, that if we take $n_1 = 0$ or choose an arbitrary shape in Phase 1, the resulting solution is likely extremely conservative in terms of objective performance. To combat this issue, it is thus recommended to set aside some data for Phase 1 with the help of established methods borrowed from statistical learning (Section 2.3 and Appendices 2.8 and 2.9 discuss these).

2.2.4 Enhancing Optimality Performance via Self-improving Reconstruction

We propose a mechanism, under the framework in Section 2.2.2, to improve the performance of an uncertainty set by incorporating updated optimality belief.

An Elementary Explanation

As indicated at the beginning of this section, the RO we construct is a conservative approximation to the CCP. A question is whether there is an "optimal" uncertainty set, in the sense that it is a $(1-\epsilon)$-level prediction set, and at the same time gives rise to the same solution between the RO and the CCP. As a first observation, the uncertainty set $\mathcal{U} = \{\xi : g(x^*;\xi) \in \mathcal{A}\}$, where x^* is an optimal solution to the CCP, satisfies both properties: By the definition of x^*, this set contains $(1-\epsilon)$-content of P. Moreover, when we use this $\mathcal{U}$ in (2.2), x^* is trivially a feasible solution. Since this RO is an inner approximation to CCP, x^* is optimal for both the RO and the CCP. The catch, of course, is that in reality we do not know what is x^*. Our suggestion is to replace x^* with some approximate solution $\hat{x}$, leading to a set $\{\xi : g(\hat{x},\xi) \in \mathcal{A}\}$.

Alternately, the conservativeness of the RO can be reasoned from the fact that $\xi \in \mathcal{U}$, independent of what the obtained solution $\hat{x}$ is in (2.2), implies that $g(\hat{x};\xi) \in \mathcal{A}$. Thus our target tolerance probability $P(g(\hat{x};\xi) \in \mathcal{A})$ satisfies $P(g(\hat{x};\xi) \in \mathcal{A}) \geq P(\xi \in \mathcal{U})$, and, in the presence of data, makes the actual confidence level (namely $\mathbb{P}_D(P(g(\hat{x};\xi) \in \mathcal{A}) \geq 1-\epsilon)$) potentially over-conservative. However, this inequality becomes an equality if $\mathcal{U}$ is exactly $\{\xi : g(\hat{x};\xi) \in \mathcal{A}\}$. This suggests again that, on a high level, an uncertainty set that resembles the form $g(\hat{x};\xi) \in \mathcal{A}$ is less conservative and preferable.

Using the above intuition, a proposed strategy is as follows. Consider finding a solution for (2.1). In Phase 1, find an approximate HPR of the data (using some suggestions in Section 2.3) with a reasonably chosen size (e.g., just enough to cover $(1-\epsilon)$ of the data points). Solve the RO problem using this HPR to obtain an initial solution $\hat{x}_0$. Then reshape the uncertainty set as $\{\xi : g(\hat{x}_0;\xi) \in \mathcal{A}\}$. Finally, conduct Phase 2 by tuning the size of this reshaped set, say we get $\{\xi : g(\hat{x}_0;\xi) \in \tilde{\mathcal{A}}\}$ where $\tilde{\mathcal{A}}$ is size-tuned. The final RO is:

$$\text{minimize } f(x) \text{ subject to } g(x,\xi) \in \mathcal{A} \;\; \forall\, \xi : g(\hat{x}_0;\xi) \in \tilde{\mathcal{A}} \tag{2.7}$$

Evidently, if the tuning step can be done properly, i.e., the set $\{\xi : g(\hat{x}_0;\xi) \in \mathcal{A}\}$ can be expressed in the form $\{\xi : t(\xi) \leq s\}$ and s is calibrated using the method in

Section 2.2.1, then the procedure retains the overall statistical confidence guarantees presented in Theorems II.4 and II.5. For convenience, we call the RO (2.7) created from $\hat{x}_0$ and the discussed procedure a "reconstructed" RO.

More explicitly, consider the safety condition $g(x;\xi) \in \mathcal{A}$ in the form of linear inequalities $Ax \leq b$ where $A \in \mathbb{R}^{l\times d}$ is stochastic and $b \in \mathbb{R}^l$ is constant. After we obtain an initial solution $\hat{x}_0$, we set the uncertainty set as $\mathcal{U} = \{A : A\hat{x}_0 \leq b + sk\}$ where $k = (k_i)_{i=1,\ldots,l} \in \mathbb{R}^l$ is some positive vector and $s \in \mathbb{R}$. The value of s is calibrated by letting $t(A) = \max_{i=1,\ldots,l}\{(a_i'\hat{x}_0 - b_i)/k_i\}$ where a_i' is the i-th row of A and b_i is the i-th entry of b, and s is chosen as $t(A^2_{(i^*)})$, the order statistic of Phase 2 data as defined in Section 2.2.1. Using the uncertainty set $\mathcal{U}$, the constraint $Ax \leq b \ \forall \ A \in \mathcal{U}$ becomes $\max_{a_i'\hat{x}_0 \leq b_i + sk_i} a_i'x \leq b_i, i = 1,\ldots,l$ via constraint-wise projection of the uncertainty set, which can be reformulated into linear constraints by using standard RO machinery (see, e.g., Theorem II.13).

Properties of Self-improving Reconstruction

We formalize the discussion in Section 2.2.4 by showing some properties of the optimization problem (2.7). We focus on the setting of inequalities-based safety conditions

$$\text{minimize } f(x) \ \text{ subject to } \ P(g(x;\xi) \leq b) \geq 1-\epsilon \tag{2.8}$$

where $g(x;\xi) = (g_j(x;\xi))_{j=1,\ldots,l} \in \mathbb{R}^l$ and $b = (b_j)_{j=1,\ldots,l} \in \mathbb{R}^l$. Suppose $\hat{x}_0$ is a given solution (not necessarily feasible). Suppose for now that there is a way to compute quantiles exactly for functions of ξ, and consider the reconstructed RO

$$\text{minimize } f(x) \ \text{ subject to } \ g(x,\xi) \leq b \ \ \forall \ \xi : g(\hat{x}_0;\xi) \leq b + \rho k \tag{2.9}$$

where $k = (k_j)_{j=1,\ldots,l} \in \mathbb{R}^l$ is a positive vector, and $\rho = \rho(\hat{x}_0)$ is the $(1-\epsilon)$-quantile of $\max_{j=1,\ldots,l}\{(g_j(\hat{x}_0;\xi) - b_j)/k_j\}$. A useful observation is:

Theorem II.6 (Feasibility guarantee for reconstruction). *Given any solution $\hat{x}_0$, if ρ is the $(1-\epsilon)$-quantile of $\max_{j=1,\ldots,l}\{(g_j(\hat{x}_0;\xi)-b_j)/k_j\}$, then any feasible solution of* (2.9) *is also feasible for* (2.8)*.*

Proof. Proof of Theorem II.6. Since $\{\xi : g(\hat{x}_0;\xi) \leq b+\rho k\}$ is by construction a $(1-\epsilon)$-content set for ξ under P, Lemma II.1 concludes the theorem immediately. □

Note that Theorem II.6 holds regardless of whether $\hat{x}_0$ is feasible for (2.8). That is, (2.9) is a way to output a feasible solution from the input of a possibly infeasible $\hat{x}_0$. What is more, in the case that $\hat{x}_0$ is feasible, (2.9) is guaranteed to give a solution at least as good:

Theorem II.7 (Monotonic objective improvement). *Under the same assumption as Theorem II.6, an optimal solution $\hat{x}$ of* (2.9) *is feasible for* (2.8)*. Moreover, if $\hat{x}_0$ is feasible for* (2.8)*, then $\hat{x}$ satisfies $f(\hat{x}) \leq f(\hat{x}_0)$.*

Proof. Proof of Theorem II.7. Note that if $\hat{x}_0$ is feasible for (2.8), we must have $\rho \leq 0$ (or else the chance constraint does not hold) and hence $\hat{x}_0$ must be feasible for (2.9). By the optimality of $\hat{x}$ for (2.9) we must have $f(\hat{x}) \leq f(\hat{x}_0)$. The theorem concludes by invoking Theorem II.6 that implies $\hat{x}$ is feasible for (2.8). □

Together, Theorems II.6 and II.7 give a mechanism to improve any input solution in terms of either feasibility or optimality for (2.8): If $\hat{x}_0$ is infeasible, then (2.9) corrects the infeasibility and gives a feasible solution; if $\hat{x}_0$ is feasible, then (2.9) gives a feasible solution that has an objective value at least as good.

Similar statements hold if the quantile ρ is only calibrated under a given statistical confidence. To link our discussion to the procedure in Section 2.2.1, suppose that a solution $\hat{x}_0$ is obtained from an RO formulation (or in fact, any other procedures) using only Phase 1 data. We have:

Corollary II.8 (Feasibility guarantee for reconstruction under statistical confidence)**.** *Given any solution $\hat{x}_0$ obtained using Phase 1 data, suppose ρ is the upper bound of the $(1-\epsilon)$-quantile of $\max_{j=1,\ldots,l}\{(g_j(\hat{x}_0;\xi)-b_j)/k_j\}$ with confidence level $1-\delta$ generated under Phase 2 data. Any feasible solution of* (2.9) *is also feasible for* (2.8) *with the same confidence.*

Corollary II.9 (Improvement from reconstruction under statistical confidence)**.** *Under the same assumptions as Corollary II.8, an optimal solution $\hat{x}$ of* (2.9) *is feasible for* (2.8) *with confidence $1-\delta$. Moreover, if $\rho \leq 0$, then $\hat{x}$ satisfies $f(\hat{x}) \leq f(\hat{x}_0)$.*

The proofs of Corollaries II.8 and II.9 are the same as those of Theorems II.6 and II.7, except that Lemma II.2 is invoked instead of Lemma II.1. Note that $\rho \leq 0$ in Corollary II.9 implies that $\hat{x}_0$ is feasible for (2.8) with confidence $1-\delta$. However, the case $\rho > 0$ in Corollary II.9 does not directly translate to a conclusion that $\hat{x}_0$ is infeasible under confidence $1-\delta$, since ρ is a confidence upper bound, instead of lower bound, for the quantile. This implies a possibility that $\hat{x}_0$ is feasible and close to the boundary of the feasible region. There is no guarantee of objective improvement under the reconstructed RO in this case, but there is still guarantee that the output $\hat{x}$ is feasible with confidence $1-\delta$.

Our numerical experiments in Section 2.4 show that, when applicable, such reconstructions frequently lead to notable improvements. Nonetheless, we caution that, depending on the constraint structure, the reconstruction step does not always lead to a significant or a strict improvement even if $\rho \leq 0$, and in these cases some transformation of the constraint is needed. For example, in the case of single linear chance constraint in the form (2.8) with $l = 1$ and a bilinear $g(x;\xi)$, the reconstructed uncertainty set consists of one linear constraint. Consequently, the dualization of the

RO (see Theorem II.13) consists of one dual variable, which optimally scales $\hat{x}_0$ by a scalar factor. When b in (2.8) (with $l = 1$) is also a stochastic source, no scaling adjustment is allowed because the "decision variable" associated with b (viewing b as a random coefficient in the linear constraint) is constrained to be 1. Thus, the proposed reconstruction will show no strict improvement. However, this behavior could be avoided by suitably re-expressing the constraint. When b is say positively distributed (or very likely so), one can divide both sides of the inequality by b to obtain an equivalent inequality with right hand side fixed to be 1. This equivalent constraint is now improvable by our reconstruction (and the new stochasticity now comprises the ratios of the original variables, which can still be observed from the data).

2.3 Constructing Uncertainty Sets

Our proposed strategy in Section 2.2 requires constructing an uncertainty set that is tractable for RO, and recommends to trace the shape of an HPR as much as possible. Regarding tractability, linear RO with the uncertainty set shapes mentioned in Section 2.2.1 can be reformulated into standard optimization formulations. For convenience we document some of these results in Section 2.7, along with some explanation on how to identify $t(\cdot)$ for the size calibration in our procedure.

Since taking unions or intersections of basic sets gives more capability to trace HPR, we highlight the following two immediate observations. First is that unions of basic sets preserve the tractability of the robust counterpart associated with each union component, with a linear growth of the number of constraints against the number of components.

Lemma II.10 (Reformulating unions of sets)**.** *The constraint*

$$g(x;\xi) \in \mathcal{A} \;\; \forall\, \xi \in \mathcal{U}$$

where $\mathcal{U} = \bigcup_{i=1}^{k} \mathcal{U}^i$ *is equivalent to the joint constraints*

$$g(x;\xi) \in \mathcal{A} \;\; \forall\, \xi \in \mathcal{U}^i, \quad i = 1, \dots, k$$

Second, in the special case of intersections of sets where each intersection component is on the portion of the stochasticity associated with each of multiple constraints, the projective separability property of uncertainty sets (e.g., [12]) gives the following:

Lemma II.11 (Reformulating intersections of sets)**.** *Let* $\xi \in \mathbb{R}^m$ *be a vector that can be represented as* $\xi = (\xi^i)_{i=1,\dots,k}$*, where* $\xi^i \in \mathbb{R}^{m^i}, i = 1, \dots, k$ *are vectors such that* $\sum_{i=1}^{k} m^i = m$*. Suppose that* $\mathcal{U} = \prod_{i=1}^{k} \mathcal{U}^i$ *where each* $\mathcal{U}^i$ *is a set on the domain of* ξ^i*. The set of constraints*

$$g(x;\xi^i) \in \mathcal{A}^i, i = 1, \dots, k \;\; \forall\, \xi \in \mathcal{U}$$

is equivalent to

$$g(x;\xi^i) \in \mathcal{A}^i \;\; \forall\, \xi^i \in \mathcal{U}^i, \quad i = 1, \dots, k$$

Note that in approximating a joint CCP, all the $\mathcal{U}^i$ in Lemma II.11 need to be jointly calibrated statistically to account for the simultaneous estimation error (which can be conducted by introducing a max operation for the intersection of sets). Intuitively, with weakly correlated data across the constraints, it fares better to use a separate $\mathcal{U}^i$ to represent the uncertainty of each constraint rather than using a single $\mathcal{U}$ and projecting it. Section 2.8 provides a formal statement to support this intuition, by arguing a lower level of conservativeness in using individual ellipsoids rather than a single aggregated block-diagonal ellipsoid.

In addition, we can borrow the following statistical tools to more tightly trace an HPR, i.e., a smaller-volume prediction set:

1. When data appears in multi-modal form, we can use clustering. Label the data into different clusters (using k-means, Gaussian mixture models, or any other techniques), form a simple set $\mathcal{U}_i$ like a ball or an ellipsoid for each cluster, and use the union $\bigcup_i \mathcal{U}_i$ as the final shape.

2. If the high-dimensional data set has an intrinsic low-dimensional representation, we can use dimension reduction tools like principal component analysis. Suppose $\tilde{\xi} = M\xi + N$, where $M \in \mathbb{R}^{r\times m}$ and $N \in \mathbb{R}^r$, is a low-dimensional representation of a raw random vector $\xi \in \mathbb{R}^m$. Then we can use uncertainty set in the form

$$\mathcal{U} = \{(M\xi - \mu)'\Sigma^{-1}(M\xi - \mu) \le s\}, \tag{2.10}$$

   where μ is the sample mean of $\tilde{\xi}$ and Σ is a covariance estimate of $\tilde{\xi}$. Tractability is preserved by a straightforward use of existing RO results (see Theorem II.15 in Section 2.7).

3. In situations of unstructured data where clustering or dimension reduction techniques do not apply, one approach is to view each data point as a “cluster” by taking the union of balls each surrounding one data point. Intriguingly, this scheme coincides with the one studied in [68] to approximate ambiguous CCP where the underlying distribution is within a neighborhood of some baseline measure.

We provide further illustrations of these tools in Section 2.9.

2.4 Numerical Examples

We present numerical examples to illustrate the performances of our RO approach. In all our examples,

1. We set $\epsilon = 0.05$ and $\delta = 0.05$.
2. For each setting, we repeat the experimental run $1,000$ times, each time generating a new independent data set.
3. We define $\hat{\epsilon}$ to be the estimated expected violation probability of the obtained solution. In other words, $\hat{\epsilon} = \hat{E}_D\left[P_{violation}\right]$, where $\hat{E}_D[\cdot]$ refers to the empirical expectation taken among the $1,000$ data sets, and $P_{violation}$ denotes the probability $P(g(\hat{x}(D);\xi) \notin \mathcal{A})$. For single linear CCPs with Gaussian distributed ξ, $P_{violation}$ can be computed analytically. In other cases, $P_{violation}$ is estimated using $10,000$ new independent realizations of ξ. For approaches that do not depend on data, e.g., SCA, we set $\hat{\epsilon} = P_{violation}$ directly.
4. We define $\hat{\delta} = \hat{P}_D(P_{violation} > \epsilon)$, where $\hat{P}_D(\cdot)$ refers to the empirical probability with respect to the $1,000$ data sets and $P_{violation}$ is similarly defined as for $\hat{\epsilon}$. For approaches that do not depend on data, the chance constraint is always satisfied and therefore we have $\hat{\delta} = 0$.
5. We denote "Obj. Val." as the average optimal objective value of the 1,000 solutions generated from the independent data sets.
6. When the reconstruction technique described in Section 2.2.4 is applied, the initial guessed solution is obtained from an uncertainty set with size calibrated to be just enough to cover $(1-\epsilon)$ of the Phase 1 data.

Recall that d is the decision space dimension, n is the total sample size, and n_1 and n_2 are the sample sizes for Phases 1 and 2. These numbers differ across the examples for illustration purpose.

Moreover, we compare our RO approaches with several methods:

1. Scenario approaches, including the classical SG ([37]) described in the introduction and its variant FAST ([40]). FAST was introduced to reduce the sample size requirement of the classical SG. It consists of two steps, each step using n_1 and n_2 samples respectively (the notations are unified with our method for easy comparisons). The first step of FAST is similar to SG, which solves a sampled program with n_1 constraints and obtains a tentative solution. The second step is a detuning step to adjust the tentative solution with the help of a "robust feasible solution", i.e., a solution feasible for any possible ξ. The adjusted solution is a convex combination of the tentative solution and the robust feasible solution so that the final solution satisfies the other n_2 sampled constraints. In our comparison, we use the minimum required sample sizes in the detuning step suggested in [40] so that the total required sample size is precisely the given overall size. We compare with FAST here since the latter elicits a small sample size requirement with the help of a validation-type scheme that is similar to our approaches applied to the RO setting.

2. DRO with first and second moment information, where the moments lie in an ellipsoidal joint confidence region. First, supposing we are given exact first and second moments, we can reformulate a distributionally robust linear chance constraint into a quadratic constraint suggested in [62]. On the other hand, using the delta method suggested in [120], we can construct ellipsoidal confidence regions for the vectorized mean and covariance matrix. Combining the quadratic constraint in [62] and the ellipsoidal set in [120], we can use Theorem 1 (II) and Example 4 in [120] to reformulate the DRO with ellipsoidal moment set

into a semidefinite program. We provide further details of this reformulation in Section 2.10.

3. DRO with uncertainty set defined by a neighborhood surrounding a reference distribution measured by a ϕ-divergence. We use the reformulation in [96] that transforms such a distributionally robust chance constraint into an ordinary chance constraint, under the reference distribution, with an adjusted tolerance level ϵ^*, which then allows us to resort to SG or SAA using Monte Carlo samples (as we will see momentarily, whichever method to resort to does not quite matter in our experiments). We use the Kullback-Leibler (KL) divergence, and construct the reference distribution using kernel density estimation (with Gaussian kernel). We set the size of the KL-divergence ball by estimating the divergence using the k-NN estimator, a provably consistent estimator proposed in [167, 139] (other related estimators and theoretical results are in [125, 113, 135, 138]). We use $k = 1$ in our experiments, as the experimental results indicate that the bias increases significantly as k increases. Moreover, to estimate the divergence properly, we split the data into two portions n_1 and n_2, first portion used to construct the reference kernel density, second portion used for the k-NN divergence estimation. The reason of this split is that, otherwise, the estimation of the reference distribution and the divergence would depend on and interfere with each others, leading to estimation accuracy so poor that the divergence estimate becomes negative all the time. We provide further implementation details in Section 2.11.3.

4. SCA. We will state the underlying a priori distributional assumptions in using the considered SCA, which differ case-by-case.

When applying moment-based DRO and SCA to joint CCPs, we use the Bonferroni correction (more details in the relevant examples). We also make two additional remarks. First, when comparing the objective values from different methods, since one can always translate or scale the problem by adding/multiplying constants to distort the apparent magnitudes, we mostly focus our comparisons on the direction (bigger or smaller), which is invariant under the above distortions. Second, even though we only report the point estimates of the mean objective values and ϵ, δ, our conclusions in comparing the objective values and constraint violation probabilities remain unchanged even if we consider the 95% confidence intervals of these estimates (from the $1,000$ experimental repetitions), and we do not report the confidence intervals for the sake of succinctness. Finally, our codes are available at https://github.com/zhyhuang/Learningbased-RO.

2.4.1 Test Case 1: Multivariate Gaussian on a Single Chance Constraint

We consider a single linear CCP

$$\text{minimize } c'x \ \text{ subject to } \ P(\xi'x \leq b) \geq 1-\epsilon \tag{2.11}$$

where $x \in \mathbb{R}^d$ is the decision vector, and $c \in \mathbb{R}^d$, $b \in \mathbb{R}$ are arbitrarily chosen constants. The random vector $\xi \in \mathbb{R}^d$ is drawn from a multivariate Gaussian distribution with an arbitrary mean (here we set it to $-c$) and an arbitrarily chosen positive definite covariance matrix. Since (2.11) is exactly solvable when the Gaussian distribution is known, we can verify that it has a bounded optimal solution.

We consider $d = 11$ and 100 as the dimension of the decision vector. Tables 2.1 and 2.2 show these two cases with a small sample size $n = 120$, whereas Tables 2.3 and 2.4 show these cases with a bigger sample size (336 and 2331 respectively) so that the classical SG provides provable feasibility guarantees. In each table, we show the results for our RO using ellipsoidal uncertainty set ("RO"), our reconstructed RO

Table 2.1: Optimality and feasibility performances on a single $d = 11$ dimensional linear CCP with Gaussian distribution for several methods, using sample size $n = 120$. The true optimal value is -1196.7.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	-
n_1	60	60	-	61	-	60	-
n_2	60	60	-	59	-	60	-
Obj. Val.	-1189.31	-1194.87	-1196.60	-1193.53	-1187.35	0	-1195.07
$\hat{\epsilon}$	1.34×10^{-5}	0.0164	0.090	0.0164	2.55×10^{-8}	0	0.0072
$\hat{\delta}$	0	0.048	0.957	0.043	0	0	0

Table 2.2: Optimality and feasibility performances on a single $d = 100$ dimensional linear CCP with Gaussian distribution for several methods, using sample size $n = 120$. The true optimal value is -1195.3. Results on moment-based DRO are based on 30 replications due to high computational demand.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	-
n_1	60	60	-	61	-	60	-
n_2	60	60	-	59	-	60	-
Obj. Val.	-832.12	-1112.11	unbounded	unbounded	-1193.21	0	-1193.0
$\hat{\epsilon}$	0	0.0158	-	-	0.195	0	0.0072
$\hat{\delta}$	0	0.041	-	-	1	0	0

("Recon"), SG ("SG"), FAST ("FAST"), DRO with ellipsoidal moment set ("DRO Mo"), DRO with KL-divergence set ("DRO KL") and SCA ("SCA"). The last approach does not need the data and instead assumes partial a priori distributional information.

For our RO approaches, we use ellipsoidal uncertainty sets with estimated covariance matrix for the case $d = 11$ (Tables 2.1 and 2.3), and diagonalized ellipsoidal sets (i.e., only using variance estimates) for $d = 100$ (Tables 2.2 and 2.4) to stabilize our estimates because n_1 is smaller than d in the latter case. The tables show that the solutions from our plain RO tend to be conservative, as $\hat{\delta} = 0$. Nonetheless, the reconstructed RO is less conservative across all settings, reflected by the better average optimal values and $\hat{\delta}$ close to the target confidence level 0.05. In all cases, both the plain RO and the reconstructed RO give valid (i.e., confidently feasible) solutions.

Table 2.3: Optimality and feasibility performances on a single $d = 11$ dimensional linear CCP with Gaussian distribution for several methods, using sample size $n = 336$. The true optimal value is -1196.7.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	336	336	336	336	336	336	-
n_1	212	212	-	318	-	168	-
n_2	124	124	-	18	-	168	-
Obj. Val.	-1190.33	-1195.82	-1195.67	-1195.14	-1188.48	0	-1195.07
$\hat{\epsilon}$	3.47×10^{-6}	0.0247	0.0331	0.0259	2.19×10^{-8}	0	0.0072
$\hat{\delta}$	0	0.04	0.056	0.043	0	0	0

Table 2.4: Optimality and feasibility performances on a single $d = 100$ dimensional linear CCP with Gaussian distribution for several methods, using sample size $n = 2331$. The true optimal value is -1195.3. Results on moment-based DRO are based on 30 replications due to high computational demand.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	2331	2331	2331	2331	2331	2331	-
n_1	1318	1318	-	2326	-	1166	-
n_2	1013	1013	-	5	-	1165	-
Obj. Val.	-1168.35	-1194.76	-1194.13	-1193.85	-1175.48	0	-1193.0
$\hat{\epsilon}$	0	0.0395	0.0428	0.0386	8.76×10^{-14}	0	0.0072
$\hat{\delta}$	0	0.051	0.039	0.052	0	0	0

We compare our ROs with scenario approaches. When the sample size is small (Tables 2.1 and 2.2), SG cannot obtain a valid solution. In the case $d = 11$, it gives $\hat{\delta}$ much greater than 0.05. Furthermore, in the case $d = 100$, SG gives unbounded solutions in all $1,000$ replications, as the number of sampled constraints is very close to the decision dimension. For FAST, since b is chosen to be positive, we can use the origin to be the robust feasible solution. Table 2.1 shows that, when $d = 11$, FAST gives confidently feasible solutions. The average optimal value from reconstructed RO (-1194.87) is (slightly) better than the value from FAST (-1193.53), while RO using ellipsoidal sets is more conservative (-1189.31). However, when $d = 100$ (Table 2.2), the first-step problem of FAST is unbounded in all $1,000$ replications.

When the sample size is adequate (Tables 2.3 and 2.4), the values of $\hat{\delta}$ from SG being less than or close to 0.05 confirms the validity of the solutions. Note that in these cases FAST gives more conservative solutions than SG (This is a general consequence from the construction of FAST that is designed to have a smaller feasible region than SG under the same dataset). RO with ellipsoidal sets obtains more conservative solutions than SG, as shown by the zero $\hat{\delta}$'s and worse average objective values. By using reconstruction, however, the $\hat{\delta}$'s become very close to the desired confidence level $\delta = 0.05$, and the average objective values are almost identical to (and slightly better than) those obtained from SG.

The above reveal that, when the sample size is large enough, SG can perform better than our RO using basic uncertainty sets. On the other hand, our RO can provide feasibility guarantees in small-sample situations where SG may fail. FAST is valid in small-sample situations, but is more likely to have unbounded solutions in high-dimensional problems than our RO. Thus, generally, our RO appears most useful for small sample sizes when compared with scenario approaches, a benefit postulated in the previous sections. It also appears that using reconstruction can boost our performance to a comparable level as SG (and hence also FAST) in situations where the latter is applicable in the shown examples. Note that our reconstruction by design can improve the objective performance compared to plain RO, whereas FAST is primarily used to reduce the sample size requirement and is necessarily more conservative than SG in terms of achieved objective value. Finally, we note that unbounded solutions in SG can potentially be avoided by adding artificial constraints. In this regard, we show in Section 2.11.1 the same example but with additional non-negativity constraints to illustrate the comparisons further.

Next, we compare with moment-based DRO. In low-dimensional cases with $d = 11$, moment-based DRO gives solutions more conservative than RO using ellipsoidal sets, as shown by the larger objective values, i.e. -1187.35 (DRO) versus -1189.31 (RO) in the small-sample case (Table 2.1) and -1184.48 (DRO) versus -1190.33 (RO) in the large-sample case (Table 2.3). The conservativeness of moment-based DRO is also revealed in the small $\hat{\epsilon}$ and $\hat{\delta} = 0$ in both cases. For high-dimensional problems with $d = 100$, we present the performance of moment-based DRO with only 30 replications (instead of 1000) due to the large program size and consequently the demanding computational effort when solving the reformulated semidefinite programs (although the replication size is smaller, conclusions can still be drawn rigorously, i.e., the confidence intervals of the estimated $\hat{\epsilon}$ and $\hat{\delta}$ turn out to either lie completely under or above 0.05). In the small-sample size case (Table 2.2), moment-based DRO fails to provide feasible solutions ($\hat{\delta} = 1$, i.e., obtained solutions violate the chance constraint in all 30 replications). This can be attributed to a poor estimation of the moment confidence region with small data and high dimension (Note that forming an ellipsoidal first-and-second-moment set for moment-based DRO requires estimating a covariance matrix of size $(3d+d^2)/2 \times (3d+d^2)/2$, as it uses the estimation variances of the first and second moments that involve even higher-order moments, in contrast to a size of $d \times d$ in our ellipsoidal RO). When the sample size is larger (Table 2.3), moment-based DRO provides valid feasible solutions ($\hat{\delta} = 0$). The average objective (-1175.48) is less conservative than our plain RO (-1168.35), but is more conservative than our reconstructed RO (-1194.76).

The above observations show that, when the moment information is well estimated (i.e., the sample size is sufficient relative to the dimension), moment-based DRO provides solutions with similar conservative level as our RO using ellipsoidal sets. However, when the sample size is too small to get reasonable estimates for the moments, moment-based DRO can fail to obtain feasible solutions. Reconstructed RO appears to outperform moment-based DRO generally. The benefits of our RO approaches in small sample and the boosted performance of reconstructed RO compared to moment-based DRO are in line with our comparisons with scenario approaches.

DRO with estimated KL-divergence set suffers from general setbacks in the experiments. In all cases we considered, the kernel density estimator cannot provide a good enough reference distribution f_0, so that the size of the divergence ball is too big and subsequently results in conservative solutions. The construction of f_0 is poor due to the curse of dimensionality in kernel density estimation whose accuracy deteriorates exponentially with the dimension, as we have a relatively high dimension compared with the data size. On the other hand, the performance of DRO, which relies on using the adjusted tolerance level ϵ^*, appears sensitive to the divergence ball size and demands a high accuracy in estimating f_0. Subsequently, the big divergence ball size leads to a zero ϵ^* in all replications, which in turn forces us to choose a solution x that satisfies the safety condition $\xi' x \leq b$ for all $\xi \in \mathbb{R}^d$. The origin is then output as the only such feasible solution, and the objective is 0, which are shown in Tables 2.1, 2.2, 2.3, and 2.4. This indicates that DRO with KL divergence, calibrated using density estimator and the divergence estimation technique suggested in the literature, gives overly conservative solutions for our considered problems.

Lastly, we compare with SCA. Consider a perturbation model for ξ given by $\xi = a_0 + \sum_{i=1}^{L} \zeta_i a_i$ where $a_i \in \mathbb{R}^d$ for all $i = 0, 1, \ldots, L$ and $\zeta_i \in \mathbb{R}$ are independent Gaussian variable with mean μ_i and variance s_i^2, such that $\mu_i \in [\mu_i^-, \mu_i^+]$ and $s_i^2 \leq \sigma_i^2$. A safe approximation of (2.11) is in [12]:

$$\min\ c'x \quad \text{s.t.} \quad (a_0' x - b) + \sum_{i=1}^{L} \max[a_i' x \mu_i^-, a_i' x \mu_i^+] + \sqrt{2 \log(1/\epsilon)} \sqrt{\sum_{i=1}^{L} \sigma_i^2 (a_i' x)^2} \leq 0.$$

To apply this SCA to (2.11), we set ζ_i to be independent $N(0,1)$ variables, $a_0 = \mu$ and a_i to be the i-th column of $\Sigma^{1/2}$, and $\mu_i^- = \mu_i^+ = 0$ and $\sigma_i^2 = 1$ for $i = 1, ..., d$. This in fact assumes knowledge on the mean and covariance of the Gaussian vector

ξ, thus giving an upper hand to SCA.

Tables 2.1, 2.3, 2.2 and 2.4 all show that the optimal objective values obtained from SCA (-1195.07 and -1193.0 respectively for $d = 11, 100$) are close to the true optimal values (-1196.7 and -1195.3) compared to other methods. Our ROs using ellipsoidal sets obtain more conservative solutions generally. The relative conservativeness also shows up in reconstructed RO with small sample sizes (Tables 2.1 and 2.2), but with more samples (Tables 2.3 and 2.4) our reconstructed RO outperforms the considered SCA.

Note that in this example the normality, and the mean and covariance information used in the SCA, makes the latter perform very well. Our RO using estimated ellipsoidal sets does not achieve this level of preciseness. However, the reconstructed RO can still outperform this SCA when the sample size is large enough. Note that the performance of SCA depends on the true distribution (as it is related to the tightness of the SCA constraint in approximating the chance constraint). In the next example, we consider an alternate underlying distribution where SCA does not perform as well.

2.4.2 Test Case 2: Beta Models on a Single Chance Constraint

We consider the single linear CCP in (2.11), where each component of ξ is now bounded. We use a perturbation model for ξ given by $\xi = a_0 + \sum_{i=1}^{L} \zeta_i a_i$ where $a_i \in \mathbb{R}^d$ for all $i = 0, 1, \ldots, L$ and $\zeta_i \in \mathbb{R}$ are independent random variables each with mean zero and bounded in $[-1, 1]$, where $d = 10$, $L = 10$ and $a_i \in \mathbb{R}^{10}$ being known arbitrarily chosen vectors. This allows the use of an SCA stated below. In particular, we set each ζ_i to be a Beta distribution with parameters $\alpha = 10$ and $\beta = 10$ that is multiplied by 2 and shifted by 1. Similar to Section 2.4.1, we set c to be the negative of the mean of ξ and $b \in \mathbb{R}$ is an arbitrarily chosen positive constant.

Regarding the comparison with SCA, this problem is supplementary to the Gaussian cases in Section 2.4.1 in that it presents performances of SCA when we use less information about ξ. Suppose that we have chosen a correct perturbation model in the SCA (i.e., knowledge of d, L, a_i and the boundedness on $[-1, 1]$). We use the Hoeffding inequality to replace the chance constraint with $\eta\sqrt{\sum_{i=1}^{L}(a_i'x)^2} \leq b - a_0'x$, where $\eta \geq \sqrt{2\log(1/\epsilon)}$. This SCA is equivalent to an RO imposing an uncertainty set $\mathcal{U} = \{\zeta : \|\zeta\|_2 \leq \eta\}$ where $\zeta = (\zeta_i)'_{i=1,\ldots,L}$ is the vector of perturbation random variables ([12] Section 2.3).

Table 2.5: Optimality and feasibility performances on a single $d = 10$ dimensional linear CCP with the Beta-perturbation model for several methods, using sample size $n = 120$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	-
n_1	60	60	-	61	-	60	-
n_2	60	60	-	59	-	60	-
Obj. Val.	-988.78	-1087.85	-1114.57	-1071.77	-968.30	0	-815.06
$\hat{\epsilon}$	1.02×10^{-5}	0.0161	0.0643	0.0171	0	0	0
$\hat{\delta}$	0	0.037	0.723	0.063	0	0	0

Table 2.5 shows the results from different approaches with sample size $n = 120$. Our RO performs better than SCA in terms of achieved objective values (-988.78 against -815.06), the latter appearing more conservative than the example in Section 2.4.1 as shown by $\hat{\epsilon} = 0$. Also, as in the previous example, reconstruction boosts further our RO performance (from -988.78 to -1087.85). Our RO here performs better than SCA because the latter, derived on a worst-case basis, does not tightly apply to the "truth" in this example, i.e., the Hoeffding bound does not lead to tight performance guarantees on the scaled Beta distribution (putting aside the assumed knowledge of d, L, a_i and the boundedness on $[-1, 1]$ when applying the SCA). Note that, since SCA also has an RO interpretation, the above observations show the superiority of our geometry or size selection of the uncertainty set. Our fully

nonparametric approach shows full-fledged advantage than SCA in this example.

We also report the outcomes of SG, which breaks down as shown by $\hat{\delta}$ being much bigger than 0.05, as 120 observations is not enough to achieve the needed feasibility confidence. FAST obtains valid solutions, and outperforms our RO with ellipsoidal sets but underperforms our reconstructed RO in terms of achieved objective value. Moment-based DRO also obtains valid solutions, but is conservative as shown by $\hat{\delta} = 0$ and $\hat{\epsilon} = 0$. Its objective value underperforms our RO approaches. For divergence-based DRO, the poor construction of a reference distribution again leads to a large divergence ball size, which renders the adjusted tolerance level ϵ^* to be 0 in all but one out of 1000 replications (for the one replication where ϵ^* is non-zero, it is $\epsilon^* = 1.10 \times 10^{-11}$) and essentially outputs the origin as the solution all the time. In this example, our reconstructed RO performs the best among all considered approaches.

2.4.3 Test Case 3: Multivariate Gaussian on Joint Chance Constraints

We consider a joint CCP with $d = 11$ variables and $l = 15$ constraints in the form

$$\text{minimize } c'x \ \text{ subject to } \ P(Ax \le b) \ge 1 - \epsilon,\ x \ge 0 \tag{2.12}$$

where $c \in \mathbb{R}^{11}$ and $b \in \mathbb{R}^{15}$ are arbitrary constants, and b is positive in each element. The random vector $\xi = vec(A)$ is generated from a multivariate Gaussian distribution with mean $vec(\bar{A})$ and covariance matrix Σ, where $\bar{A} \in \mathbb{R}^{15\times 11}$ is arbitrary and $\Sigma \in \mathbb{R}^{165\times 165}$ is also an arbitrary positive definite matrix.

Tables 2.6 and 2.7 present the experimental results using two different sample sizes on the same problem. We use diagonalized ellipsoids in our RO, and conduct reconstruction with scaling parameters k_i described in Section 2.8.3. To use DRO and SCA, we apply the Bonferroni correction to decompose the joint CCP, by evenly

Table 2.6: Optimality and feasibility performances on a joint linear CCP with Gaussian distribution for several methods, using sample size $n = 120$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	-
n_1	60	60	-	61	-	60	-
n_2	60	60	-	59	-	60	-
Obj. Val.	-6956.49	-7920.12	-9283.35	-8925.74	-3996.87	0	-8927.71
$\hat{\epsilon}$	3.46×10^{-5}	0.0161	0.0581	0.0169	0	0	0.026
$\hat{\delta}$	0	0.044	0.607	0.045	0	0	0

Table 2.7: Optimality and feasibility performances on a joint linear CCP with Gaussian distribution for several methods, using sample size $n = 336$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	336	336	336	336	336	336	-
n_1	212	212	-	318	-	168	-
n_2	124	124	-	18	-	168	-
Obj. Val.	-7146.54	-8029.83	-9130.95	-9081.81	-4209.86	0	-8927.71
$\hat{\epsilon}$	7.32×10^{-5}	0.0235	0.0223	0.0185	0	0	0.026
$\hat{\delta}$	0	0.038	0.005	0.002	0	0	0

dividing the tolerance level into ϵ/m to create individual chance constraints. For each individual chance constraint, we construct DRO and SCA constraint following the scheme in Section 2.4.1.

Comparing with scenario approaches, we see that, much like the examples in Sections 2.4.1 and 2.4.2, SG fails with small sample size (confirmed by $\hat{\delta}$ much larger than 0.05 in Table 2.6), but obtains valid solutions as sample size grows (confirmed by $\hat{\delta} < 0.05$ in Table 2.7). While reconstruction improves the optimal values for RO in both cases, SG (and so is FAST) gives better optimal value (-9130.95) than reconstructed RO (-8029.83) under a big sample size. Moment-based DRO appears very conservative for both small and large sample cases, as the obtained average objective values (-3996.87 and -4209.86) are much greater than other approaches, including our ROs, and the associated $\hat{\epsilon}$ and $\hat{\delta}$ are 0. Like the previous experiments, divergence-based DRO outputs the origin as the solution and gives objective value 0 due to over-sized uncertainty sets. On the other hand, SCA obtains a better

Table 2.8: Optimality and feasibility performances on a joint linear CCP with beta distribution for several methods, using sample size $n = 120$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	-
n_1	60	60	-	61	-	60	-
n_2	60	60	-	59	-	60	-
Obj. Val.	-1241.05	-1796.74	-2105.77	-1732.73	-230.74	0	-361.079
$\hat{\epsilon}$	6.96×10^{-5}	0.0138	0.0577	0.0170	0	0	0
$\hat{\delta}$	0	0.022	0.576	0.045	0	0	0

Table 2.9: Optimality and feasibility performances on a joint linear CCP with beta distribution for several methods, using sample size $n = 336$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	336	336	336	336	336	336	-
n_1	212	212	-	318	-	168	-
n_2	124	124	-	18	-	168	-
Obj. Val.	-1304.89	-1911.36	-1881.69	-1828.98	-251.69	0	-361.079
$\hat{\epsilon}$	1.20×10^{-4}	0.0199	0.0229	0.0192	0	0	0
$\hat{\delta}$	0	0.023	0.004	0.003	0	0	0

solution than our ROs, thanks to the tightness of the approximation for Gaussian distributions.

2.4.4 Test Case 4: Beta Models on Joint Chance Constraints

We consider the joint CCP in (2.12) with a bounded random vector ξ. We use the perturbation model described in Section 2.4.2, where $d = 165$, $L = 165$ and $a_i \in \mathbb{R}^{165}, i = 1, ..., L$ are arbitrarily chosen vectors, and the same random variables for ζ_i's as in Section 2.4.2. Again, we apply the Bonferroni correction to invoke DRO and SCA as in Section 2.4.3, and the corresponding schemes for each individualized chance constraint as in Section 2.4.2.

Tables 2.8 and 2.9 show our experimental results. The major difference with Section 2.4.3 is that now our reconstructed RO outperforms all other methods including SG and SCA: It gives smaller objective values than FAST under both small and big sample sizes. It also gives smaller objective values than SG under big sample size, while SG does not give valid solutions under small sample size. SCA is very conservative in this case, and DROs (both moment- and divergence-based) continue to be very conservative, all of whom our RO significantly outperforms.

2.4.5 Test Case 5: t- and Log-Normal Distributions

We consider problems with two heavier-tailed distributions, namely t- and log-normal. We test both the single CCP (2.11) and the joint CCP (2.12) with different dimensions and sample sizes. Since the considered SCA does not apply to these distributions, we do not include it in our comparisons here.

Tables 2.10, 2.11 and 2.12 show the comparisons among different approaches for the single CCP, and Tables 2.13 and 2.14 show the counterparts for joint CCP, when ξ is generated from a multivariate t-distribution with degree of freedom 5 and an arbitrary positive definite dispersion matrix. The comparisons are largely consistent with the Gaussian and beta cases shown in the previous subsections. Compared with SG, our ROs output feasible solutions in the small-sample case ($n = 120$), whereas SG struggles to obtain feasible solutions ($\hat{\delta}$ much greater than 0.05 in Tables 2.10 and 2.13). In the large-sample case ($n = 336$), SG gains enough feasibility and outperforms our plain RO in average objective value (-1175.04 versus -1126.66 in the single CCP case in Table 2.11, and -7387.98 versus -5778.44 in the joint CCP case in Table 2.14), but underperforms our reconstructed RO (-1175.64 and -7562.60 for single and joint CCPs respectively). FAST remedies the infeasibility issue of SG in the small-sample cases and outperforms our plain RO. On the other hand, our reconstructed RO performs competitively against FAST. Among all four cases where $d = 11$, reconstructed RO outperforms FAST in three cases but underperforms in the case of small-sample joint CCP (average objective values -1166.52, -1175.64 and -7562.60 versus -1158.27, -1170.35 and -7173.97 in Tables 2.10, 2.11 and 2.14 respectively, and -6499.93 versus -7220.37 in Table 2.13). Note that, when the dimension

Table 2.10: Optimality and feasibility performances on a single $d = 11$ dimensional linear CCP with t-distribution for several methods, using sample size $n = 120$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	120	120	120	120	120	120
n_1	60	60	-	61	-	60
n_2	60	60	-	59	-	60
Obj. Val.	-1112.75	-1166.52	-1182.20	-1158.27	-1134.38	0
$\hat{\epsilon}$	0.000252	0.0161	0.0910	0.0172	0.000461	0
$\hat{\delta}$	0	0.046	0.961	0.064	0	0

Table 2.11: Optimality and feasibility performances on a single $d = 11$ dimensional linear CCP with t-distribution for several methods, using sample size $n = 336$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	336	336	336	336	336	336
n_1	212	212	-	318	-	168
n_2	124	124	-	18	-	168
Obj. Val.	-1126.66	-1175.64	-1175.04	-1170.35	-1137.19	0
$\hat{\epsilon}$	0.00023	0.024	0.0334	0.0259	0.000407	0
$\hat{\delta}$	0	0.055	0.069	0.04	0	0

is large ($d = 100$ in Table 2.12), SG and FAST output unbounded solutions in all 1000 experimental replications, whereas plain and reconstructed RO output feasible bounded solutions.

Like in the previous subsections, our reconstructed RO outperforms moment-based DRO in all cases. When the dimension is large ($d = 100$ in Table 2.12), moment-based DRO fails to obtain feasible solutions in all 30 replications, attributed to the difficulty in estimating valid moment confidence regions. Compared to our plain RO, moment-based DRO outperforms in single CCP (-1134.38 and -1137.19 versus -1112.75 and -1126.66 in Tables 2.10 and 2.11 respectively), but underperforms in joint CCP (-3888.63 and -3891.83 versus -4229.6 and -5778.44 in Tables 2.13 and 2.14 respectively). Lastly, divergence-based DRO is once again very conservative, resulting in zero objective values all the time.

Next we consider ξ generated from log-normal distributions with arbitrarily chosen means and covariance matrices. Tables 2.15, 2.16 and 2.17 show the results for

Table 2.12: Optimality and feasibility performances on a single $d = 100$ dimensional linear CCP with t-distribution for several methods, using sample size $n = 120$. Results on moment-based DRO are based on 30 replications due to high computational demand.

	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	120	120	120	120	120	120
n_1	60	60	-	61	-	60
n_2	60	60	-	59	-	60
Obj. Val.	-1077.56	-1184.45	unbounded	unbounded	-1190.70	0
$\hat{\epsilon}$	6.00×10^{-7}	0.0156	-	-	0.22	0
$\hat{\delta}$	0	0.045	-	-	1	0

Table 2.13: Optimality and feasibility performances on a joint $d = 11$ dimensional linear CCP with t-distribution for several methods, using sample size $n = 120$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	120	120	120	120	120	120
n_1	60	60	-	61	-	60
n_2	60	60	-	59	-	60
Obj. Val.	-4229.6	-6499.93	-8313	-7220.37	-3888.63	0
$\hat{\epsilon}$	0.00108	0.00847	0.0404	0.0152	4.17×10^{-4}	0
$\hat{\delta}$	0	0.002	0.284	0.048	0	0

the single CCP, while Tables 2.18 and 2.19 show those for the joint CCP. The comparisons are quite similar to the t-distribution cases. SG in small sample outputs invalid solutions ($\hat{\delta}$ much greater than 0.05), and in large sample outputs solutions with average objective values (e.g. -683.60 in Table 2.16) better than our plain RO (-354.10) but worse than our reconstructed RO (-685.01). FAST remedies the infeasibility issue of SG in the small-sample cases, but underperforms our reconstructed RO in all cases. Moment-based DRO outperforms our plain RO but underperforms our reconstructed RO in all cases, and it continues to struggle in obtaining feasible solutions for high-dimensional problems ($\hat{\delta} = 1$ in Table 2.17). Lastly, divergence-based DRO continues to be conservative and outputs zero objective values. In all considered settings, reconstructed RO appears the best among all compared methods in terms of feasibility and optimality.

Table 2.14: Optimality and feasibility performances on a joint $d = 11$ dimensional linear CCP with t-distribution for several methods, using sample size $n = 336$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	336	336	336	336	336	336
n_1	212	212	-	318	-	168
n_2	124	124	-	18	-	168
Obj. Val.	-5778.44	-7562.60	-7387.98	-7173.97	-3891.83	0
$\hat{\epsilon}$	0.00248	0.0133	0.0144	0.0126	3.97×10^{-4}	0
$\hat{\delta}$	0	0	0	0	0	0

Table 2.15: Optimality and feasibility performances on a single $d = 11$ dimensional linear CCP with log-normal distribution for several methods, using sample size $n = 120$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	120	120	120	120	120	120
n_1	60	60	-	61	-	60
n_2	60	60	-	59	-	60
Obj. Val.	-294.00	-588.58	-784.27	-510.38	-418.30	0
$\hat{\epsilon}$	1.45×10^{-4}	0.0164	0.0902	0.0159	5.11×10^{-4}	0
$\hat{\delta}$	0	0.041	0.961	0.048	0	0

2.4.6 Summary on the Experiment Results

From the results in this section (and additional ones in Section 2.11), we highlight the following situations where our method is the most recommended.

The competitiveness of our method compared with scenario approaches is most seen in small-sample situations. Classical SG needs a much larger sample size than ours to achieve feasibility. FAST is capable of obtaining feasible solutions in small-sample cases, but appears more susceptible than RO in generating unbounded solutions. With reconstruction, our approach tends to work as well as SG and FAST for large sample (when they are all applicable). Moreover, our reconstruction has the capability to improve the optimality over plain RO, whereas FAST is by design always more conservative than SG in terms of optimality. Nonetheless, we should mention that some constraint removal approaches like sampling-and-discarding ([38]) can improve SG performances in large-sample situations.

Compared to our ROs, moment-based DRO can generate infeasible solutions when

Table 2.16: Optimality and feasibility performances on a single $d = 11$ dimensional linear CCP with log-normal distribution for several methods, using sample size $n = 336$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	336	336	336	336	336	336
n_1	212	212	-	318	-	168
n_2	124	124	-	18	-	168
Obj. Val.	-354.10	-685.01	-683.60	-646.83	-429.75	0
$\hat{\epsilon}$	8.07×10^{-5}	0.0243	0.0333	0.0261	3.33×10^{-4}	0
$\hat{\delta}$	0	0.057	0.052	0.033	0	0

Table 2.17: Optimality and feasibility performances on a single $d = 100$ dimensional linear CCP with log-normal distribution for several methods, using sample size $n = 120$. Results on moment-based DRO are based on 30 replications due to high computational demand.

	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	120	120	120	120	120	120
n_1	60	60	-	61	-	60
n_2	60	60	-	59	-	60
Obj. Val.	-309.93	-784.24	unbounded	unbounded	-1030.52	0
$\hat{\epsilon}$	6.00×10^{-6}	0.0174	-	-	0.2772	0
$\hat{\delta}$	0	0.063	-	-	1	0

the problem dimension is high compared to data size (e.g., $d = 100$ and $n = 120$), attributed to the difficulty in constructing valid moment confidence regions. In cases where moment-based DRO generates valid solutions, the solution performances seem to be sometimes better, sometimes worse than our plain RO, but in all considered instances they perform worse than our reconstructed RO. KL-divergence-based DRO appears to perform poorly in the experiments due to the challenge in obtaining a small enough divergence ball size (To get a further sense of this behavior, we investigate a very low-dimensional problem ($d = 3$) with sufficient sample size in Section 2.11.3, where divergence-based DRO provides nontrivial but still conservative solutions).

Lastly, compared with SCA, our performance is best seen when the data is non-normal. In this case the approximate constraint in SCA may not tightly approximate the original chance constraint and tends to be significantly more conservative than our approach. Moreover, SCA generally requires at least some partial distributional

Table 2.18: Optimality and feasibility performances on a joint $d = 11$ dimensional linear CCP with log-normal distribution for several methods, using sample size $n = 120$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	120	120	120	120	120	120
n_1	60	60	-	61	-	60
n_2	60	60	-	59	-	60
Obj. Val.	-0.1284	-1.1166	-4.5359	-1.0369	-0.8360	0
$\hat{\epsilon}$	0.00228	0.0157	0.0598	0.0165	0.0131	0
$\hat{\delta}$	0	0.043	0.646	0.044	0.006	0

Table 2.19: Optimality and feasibility performances on a joint $d = 11$ dimensional linear CCP with log-normal distribution for several methods, using sample size $n = 336$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	336	336	336	336	336	336
n_1	212	212	-	318	-	168
n_2	124	124	-	18	-	168
Obj. Val.	-0.0844	-1.9373	-1.7135	-1.4058	-1.2021	0
$\hat{\epsilon}$	0.0074	0.0239	0.0238	0.0197	0.0131	0
$\hat{\delta}$	0	0.05	0.011	0.007	0.026	0

knowledge (e.g., moments, support) in deriving the needed relaxing constraint, in contrast to our approach that is fully data-driven and nonparametric.

2.5 Missing Proofs in Section 2.2

Proof. Proof of Theorem II.5. Proof of 1. Let $Bin(n, p)$ be a binomial variable with number of trials n and success probability p. Then (2.4) can be written as

$$i^* = \min\{r : P(Bin(n_2, 1-\epsilon) \le r-1) \ge 1-\delta,\ 1 \le r \le n_2\} \tag{2.13}$$

Note that by the Berry-Essen Theorem,

$$\begin{aligned} & P(Bin(n_2, 1-\epsilon) \le r-1) - \Phi\left(\frac{r-1-n_2(1-\epsilon)}{\sqrt{n_2(1-\epsilon)\epsilon}}\right) \\ = \ & P\left(\frac{Bin(n_2,1-\epsilon) - n_2(1-\epsilon)}{\sqrt{n_2(1-\epsilon)\epsilon}} \le \frac{r-1-n_2(1-\epsilon)}{\sqrt{n_2(1-\epsilon)\epsilon}}\right) \\ & -\Phi\left(\frac{r-1-n_2(1-\epsilon)}{\sqrt{n_2(1-\epsilon)\epsilon}}\right) \\ = \ & O\left(\frac{1}{\sqrt{n_2}}\right) \end{aligned} \tag{2.14}$$

uniformly over $r \in \mathbb{N}^+$, where Φ is the distribution function of standard normal. Since i^* in (2.13) is chosen such that $P(Bin(n_2, 1-\epsilon) \leq i^* - 1) \geq 1 - \delta$ (where we define $i^* = n_2 + 1$ if no choice of r is valid), we have, for any $\gamma > 0$, i^* satisfies

$$\Phi\left(\frac{i^* - 1 - n_2(1-\epsilon)}{\sqrt{n_2(1-\epsilon)\epsilon}}\right) + \gamma \geq 1 - \delta$$

for large enough n_2, which gives

$$i^* \geq 1 + n_2(1-\epsilon) + \sqrt{n_2(1-\epsilon)\epsilon}\,\Phi^{-1}(1 - \delta - \gamma) \tag{2.15}$$

for large enough n_2.

On the other hand, we claim that i^* also satisfies, for any $\gamma > 0$,

$$\Phi\left(\frac{i^* - 1 - n_2(1-\epsilon)}{\sqrt{n_2(1-\epsilon)\epsilon}}\right) \leq 1 - \delta + \gamma \tag{2.16}$$

for large enough n_2. If not, then there exists an $\gamma > 0$ such that

$$\Phi\left(\frac{i^* - 1 - n_2(1-\epsilon)}{\sqrt{n_2(1-\epsilon)\epsilon}}\right) > 1 - \delta + \gamma$$

infinitely often, which implies

$$P(Bin(n_2, 1-\epsilon) \leq i^* - 1) + O\left(\frac{1}{\sqrt{n_2}}\right) > 1 - \delta + \gamma$$

or

$$P(Bin(n_2, 1-\epsilon) \leq i^* - 1) > 1 - \delta + \tilde{\gamma}$$

infinitely often for some $0 < \tilde{\gamma} < \gamma$. By the choice of i^*, we conclude that there is no r that satisfies

$$1 - \delta \leq P(Bin(n_2, 1-\epsilon) \leq r - 1) \leq 1 - \delta + \tilde{\gamma}$$

infinitely often, which is impossible. Therefore, (2.16) holds for large enough n_2, and we have

$$i^* \leq 1 + n_2(1-\epsilon) + \sqrt{n_2(1-\epsilon)\epsilon}\,\Phi^{-1}(1 - \delta + \gamma) \tag{2.17}$$

Combining (2.15) and (2.17), and noting that γ is arbitrary, we have

$$\sqrt{n_2}\left(\frac{i^*}{n_2}-(1-\epsilon)\right)\to\sqrt{(1-\epsilon)\epsilon}\Phi^{-1}(1-\delta) \tag{2.18}$$

almost surely. The same argument also shows that i^* is well-defined for large enough n_2 almost surely.

It suffices to show that

$$\mathbb{P}_{D_2}(1-\epsilon-\gamma\leq P(\xi\in\mathcal{U})\leq 1-\epsilon+\gamma|D_1)\to 1 \tag{2.19}$$

for any small $\gamma>0$. Note that, conditional on D_1, we have $P(\xi\in\mathcal{U})=P(t(\xi)\leq t(\xi^2_{(i^*)}))=F(t(\xi^2_{(i^*)}))$ where $F(\cdot)$ is the distribution function of $t(\xi)$. Since $F(t(\xi))\sim U[0,1]$ by the continuity of $t(\xi)$, we have,

$$\begin{aligned}
(2.20)\quad & \mathbb{P}_{D_2}(1-\epsilon-\gamma\leq P(\xi\in\mathcal{U})\leq 1-\epsilon+\gamma|D_1)\\
= \; & P(\#\{U_i<1-\epsilon-\gamma\}\leq i^*-1,\ \#\{U_i>1-\epsilon+\gamma\}\leq n_2-i^*)\\
& \quad \text{where } \{U_i\} \text{ denotes } n_2 \text{ realizations of i.i.d. } U[0,1] \text{ variables,}\\
& \quad \#\{U_i<1-\epsilon-\gamma\} \text{ and } \#\{U_i>1-\epsilon+\gamma\} \text{ count}\\
& \quad \text{the numbers of } U_i\text{'s that are } <1-\epsilon-\gamma \text{ and}\\
& \quad >1-\epsilon+\gamma \text{ respectively}\\
(2.21)\geq \; & 1-P(\#\{U_i<1-\epsilon-\gamma\}>i^*-1)-P(\#\{U_i>1-\epsilon+\gamma\}>n_2-i^*)
\end{aligned}$$

Consider the second term in (2.21). We have

$$\begin{aligned}
& P(\#\{U_i < 1-\epsilon-\gamma\} > i^* - 1) \\
= \;& P(Bin(n_2, 1-\epsilon-\gamma) > i^* - 1) \\
= \;& \bar{\Phi}\left(\frac{i^* - 1 - n_2(1-\epsilon-\gamma)}{\sqrt{n_2(1-\epsilon-\gamma)(\epsilon+\gamma)}}\right) + O\left(\frac{1}{\sqrt{n_2}}\right) \\
& \text{by the Berry-Essen Theorem, where } \bar{\Phi} \text{ is the tail distribution function} \\
& \text{of standard normal} \\
= \;& \bar{\Phi}\left(\frac{i^* - 1 - n_2(1-\epsilon)}{\sqrt{n_2(1-\epsilon)\epsilon}}\sqrt{\frac{1-\epsilon}{1-\epsilon-\gamma}\frac{\epsilon}{\epsilon+\gamma}} + \frac{\sqrt{n_2}\gamma}{\sqrt{(1-\epsilon-\gamma)(\epsilon+\gamma)}}\right) + O\left(\frac{1}{\sqrt{n_2}}\right) \\
\to \;& 0 \;\text{ by (2.18)}
\end{aligned}$$

Similarly, for the third term in (2.21), we have

$$\begin{aligned}
& P(\#\{U_i > 1-\epsilon+\gamma\} > n_2 - i^*) \\
= \;& P(Bin(n_2, \epsilon-\gamma) > n_2 - i^*) \\
= \;& \bar{\Phi}\left(\frac{n_2 - i^* - n_2(\epsilon-\gamma)}{\sqrt{n_2(\epsilon-\gamma)(1-\epsilon+\gamma)}}\right) + O\left(\frac{1}{\sqrt{n_2}}\right) \\
& \text{by the Berry-Essen Theorem} \\
= \;& \bar{\Phi}\left(-\frac{i^* - n_2(1-\epsilon)}{\sqrt{n_2(1-\epsilon)\epsilon}}\sqrt{\frac{\epsilon}{\epsilon-\gamma}\frac{1-\epsilon}{1-\epsilon+\gamma}} + \frac{\sqrt{n_2}\gamma}{\sqrt{(\epsilon-\gamma)(1-\epsilon+\gamma)}}\right) + O\left(\frac{1}{\sqrt{n_2}}\right) \\
\to \;& 0 \;\text{ by (2.18)}
\end{aligned}$$

Hence (2.21) converges to 1.

Proof of 2. Using again the fact that, conditional on D_1, $F(t(\xi)) \sim U[0,1]$ and

$P(\xi \in \mathcal{U}) = F(t(\xi^2_{(i^*)}))$, we have

$$\begin{aligned} & \mathbb{P}_{D_2}(P(\xi \in \mathcal{U}) \geq 1-\epsilon | D_1) \\ = \ & P(\#\{U_i < 1-\epsilon\} \leq i^* - 1) \\ = \ & P(Bin(n_2, 1-\epsilon) \leq i^* - 1) \\ = \ & \Phi\left(\frac{i^* - 1 - n_2(1-\epsilon)}{\sqrt{n_2(1-\epsilon)\epsilon}}\right) + O\left(\frac{1}{\sqrt{n_2}}\right) \quad \text{by using (2.14)} \\ \to \ & 1-\delta \quad \text{by (2.18)} \end{aligned}$$

which concludes Part 2 of the theorem.

□

Note that (2.18) is mentioned in [155] Section 2.6.1, and implies that, given D_1,

(2.22)

$$\sqrt{n_2}(P(\xi \in \mathcal{U}) - (1-\epsilon)) = \sqrt{n_2}(F(t(\xi^2_{(i^*)})) - (1-\epsilon)) \Rightarrow N\left(\sqrt{\epsilon(1-\epsilon)}\Phi^{-1}(1-\delta), \epsilon(1-\epsilon)\right)$$

by using [155] Corollary 2.5.2, which can be used to prove Part 1 of the theorem as well (as in [155] Section 2.6.3). From (2.22), we see that $P(\xi \in \mathcal{U})$ concentrates at $1-\epsilon$, as it is approximately $(1-\epsilon) + Z/\sqrt{n_2}$ where $Z \sim N\left(\sqrt{\epsilon(1-\epsilon)}\Phi^{-1}(1-\delta), \epsilon(1-\epsilon)\right)$.

2.6 Illustration of Attained Theoretical Confidence Levels

The argument in Lemma II.3 and the discussion after Theorem II.4 implies that the theoretical confidence level for a given Phase 2 sample size n_2 is

$$1 - \delta_{theoretical} = \mathbb{P}_D(P(\xi \in \mathcal{U}) \geq 1-\epsilon) = \sum_{k=0}^{i^*-1} \binom{n_2}{k} (1-\epsilon)^k \epsilon^{n_2-k}$$

This quantity is in general not a monotone function of the sample size, but it does converge to $1-\delta$ as n_2 increases, as shown in Theorem II.5 Part 2. Figures 2.1 and 2.2 illustrate how $\delta_{theoretical}$ changes with n_2 for two pairs of ϵ and δ. The changes follow a zig-zag pattern, with a general increasing trend. In the case $\delta = 0.05$ and $\epsilon = 0.05$ for example, local maxima of $\delta_{theoretical}$ occur at $n_2 = 59, 93, 124, 153, 181, \ldots$

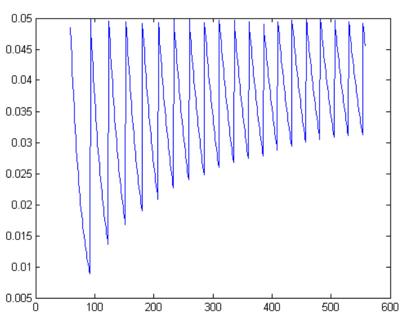


Figure 2.1: $\delta_{theoretical}$ against n_2 when $\delta = 0.05$ and $\epsilon = 0.05$

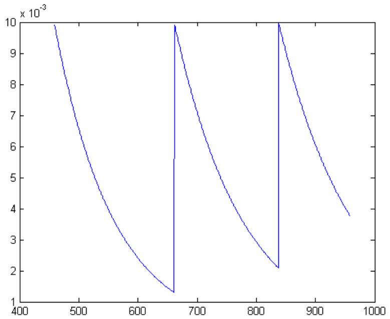


Figure 2.2: $\delta_{theoretical}$ against n_2 when $\delta = 0.01$ and $\epsilon = 0.01$

2.7 Using RO Reformulations

Results from the following discussion are adapted from [16]. Further details can be found therein and in, e.g., [12]. Along with reviewing these results, we also describe how to cast them in our procedure in Section 2.2.1.

We focus on linear safety conditions in (2.1), i.e., $g(x;\xi) \in \mathcal{A}$ is in the form $Ax \leq b$, where $A \in \mathbb{R}^{l \times d}$ is uncertain and $b \in \mathbb{R}^l$ is constant. Here A is identified with the random vector ξ. The following discussion also holds if x is further constrained to lie in some deterministic set, say $\mathcal{B}$. For convenience, we denote each row of A as a_i' and each entry in b as b_i, so that the safety condition can also be written as $a_i'x \leq b_i, i = 1, \ldots, l$.

It is well-known that in solving the robust counterpart (RC), it suffices to consider uncertainty sets in the form $\mathcal{U} = \prod_{i=1}^{l} \mathcal{U}_i$ where $\mathcal{U}_i$ is the uncertainty set projected onto the portion associated with the parameters in each constraint, and so typically we consider the RC of each constraint separately.

We first consider ellipsoidal uncertainty:

Theorem II.12 (c.f. [14])**.** *The constraint*

$$a_i'x \leq b_i \ \forall a_i \in \mathcal{U}_i$$

where $\mathcal{U}_i = \{a_i = a_i^0 + \Delta_i u : \|u\|_2 \leq \rho_i\}$ *for some fixed* $a_i^0 \in \mathbb{R}^d$*,* $\Delta_i \in \mathbb{R}^{d\times r}$*,* $\rho_i \in \mathbb{R}$*, for* $u \in \mathbb{R}^r$*, is equivalent to*

$$a_i^{0'}x + \rho_i\|\Delta_i'x\|_2 \leq b_i$$

Note that $\mathcal{U}_i$ in Theorem II.12 is equivalent to $\{a_i : \|\Delta_i^{-1}(a_i - a_i^0)\|_2 \leq \rho_i\}$ if Δ_i is invertible. Thus, given an ellipsoidal set (for the uncertainty in constraint row i) calibrated from data in the form $\{a_i : (a_i - \mu)'\Sigma^{-1}(a_i - \mu) \leq s\}$ where Σ is positive definite and $s > 0$, we can take $a_i^0 = \mu$, Δ_i as the square-root matrix in the Cholesky decomposition of Σ, and $\rho_i = \sqrt{s}$ in using the depicted RC.

Next we have the following result on polyhedral uncertainty:

Theorem II.13 (c.f. [14] and [16])**.** *The constraint*

$$a_i'x \leq b_i \ \forall a_i \in \mathcal{U}_i$$

where $\mathcal{U}_i = \{a_i : D_i a_i \leq e_i\}$ *for fixed* $D_i \in \mathbb{R}^{r\times d}$*,* $e_i \in \mathbb{R}^r$ *is equivalent to*

$$\begin{aligned} &p_i'e_i \leq b_i \\ &p_i'D_i = x' \\ &p_i \geq 0 \end{aligned}$$

where $p_i \in \mathbb{R}^r$ *are newly introduced decision variables.*

The following result applies to the collection of constraints $Ax \leq b$ with the uncertainty on $A \in \mathbb{R}^{l\times d}$ represented via a general norm on its vectorization.

Theorem II.14 (c.f. [18])**.** *The constraint*

$$Ax \leq b \ \ \forall A \in \mathcal{U}$$

where

$$\mathcal{U} = \{A : \|Q(vec(A) - vec(\bar{A}))\| \leq \rho\}, \tag{2.23}$$

for fixed $\bar{A} \in \mathbb{R}^{l\times d}$*,* $Q \in \mathbb{R}^{ld\times ld}$ *invertible,* $\rho \in \mathbb{R}$*,* $vec(A)$ *as the concatenation of all the rows of* A*,* $\|\cdot\|$ *any norm, is equivalent to*

$$\bar{a}_i'x + \rho\|(Q')^{-1}x_i\|^* \leq b_i, i = 1, ..., l$$

where $\bar{a}_i' \in \mathbb{R}^d$ *is the* i*-th row of* $\bar{A}$*,* $x_i \in \mathbb{R}^{(ld)\times 1}$ *contains* $x \in \mathbb{R}^d$ *in entries* $(i-1)d+1$ *through* $i\,d$ *and 0 elsewhere, and* $\|\cdot\|^*$ *is the dual norm of* $\|\cdot\|$*.*

When $\|\cdot\|$ denotes the L_2-norm, Theorem II.14 can be applied in much the same way as Theorem II.12, with $vec(\bar{A})$ denoting the center, Q taken as the square root of the Cholesky decomposition of Σ^{-1} where Σ is the covariance matrix, and $\rho = \sqrt{s}$ where s is the squared radius in an ellipsoidal set constructed for the data of $vec(A)$.

Next we have the following theorem to handle (2.10), which can be proved similarly as for Theorem II.12 or by standard conic duality.

Theorem II.15. *The constraint*

$$\xi'x \leq b \ \forall \xi \in \mathcal{U}$$

where $\mathcal{U}$ *is defined in* (2.10)*, and* Σ *has full rank, is equivalent to*

$$\mu'\Sigma^{-1/2}u + \sqrt{s}\lambda \leq b$$
$$M'\Sigma^{-1/2}u = x$$
$$\|u\|_2 \leq \lambda,$$

where $\lambda \in \mathbb{R}$*,* $u \in \mathbb{R}^r$ *are additional decision variables.*

2.8 Further Discussion on Choices of Uncertainty Sets

This section extends the discussions in Section 2.3 on choosing suitable uncertainty sets.

2.8.1 Comparing Individualized Ellipsoids and a Single Ellipsoids for Joint Chance Constraints

We state the following result that compares, in the case of joint chance constraints, between the use of individual ellipsoids for the stochasticities on different constraints and a single ellipsoid for all.

Proposition 1. *Let $\xi \in \mathbb{R}^m$ be a vector that can be represented as $\xi = (\xi^i)_{i=1,\dots,k}$ with $\xi^i \in \mathbb{R}^{r^i}$ and $\sum_{i=1}^k r^i = m$. Let $\mathcal{U}_{joint} = \{\xi : \|M(\xi-\mu)\|_2^2 \leq \rho_{joint}\}$ where M is a block diagonal matrix*

$$M = \begin{pmatrix} M^1 & & & \\ & M^2 & & \\ & & \cdots & \\ & & & M^k \end{pmatrix}, \tag{2.24}$$

and each $M^i \in \mathbb{R}^{r^i \times r^i}$. Let $\mathcal{U}_{individual} = \prod_{i=1}^k \mathcal{U}^i$ where $\mathcal{U}^i = \{\xi^i : \|M^i(\xi^i - \mu^i)\|_2^2 \leq \rho_{individual}\}$ and $(\mu^i)_{i=1,\dots,k}$ is defined such that $\mu = (\mu^i)_{i=1,\dots,k}$ analogously as in $(\xi^i)_{i=1,\dots,k}$ for ξ. Suppose that $\mathcal{U}_{joint}$ and $\mathcal{U}_{individual}$ are calibrated using the same Phase 2 data, with the transformation maps defined as $t_{joint}(\xi) = \|M(\xi-\mu)\|_2^2$ and $t_{individual}(\xi) = \max_{i=1,\dots,k} \|M^i(\xi^i-\mu^i)\|_2^2$ respectively.

Consider the RO

$$\text{minimize } f(x) \quad \text{subject to} \quad g_i(x;\xi^i) \in \mathcal{A}_i, i = 1,\dots,l, \ \forall \xi \in \mathcal{U} \tag{2.25}$$

Let x_{joint} be an optimal solution obtained by setting $\mathcal{U} = \mathcal{U}_{joint}$, and $x_{individual}$

be an optimal solution obtained by setting $\mathcal{U} = \mathcal{U}_{individual}$. We have $f(x_{joint}) \geq f(x_{individual})$. In other words, using $\mathcal{U}_{joint}$ is more conservative than using $\mathcal{U}_{individual}$.

Proof. Proof of Proposition 1. The ρ_{joint} calibrated using Phase 2 data is set as $t_{joint}(\xi^2_{(i^*_{joint})}) = \|M(\xi^2_{(i^*_{joint})} - \mu)\|_2^2$ where i^*_{joint} is defined similarly as (2.4). On the other hand, the $\rho_{individual}$ in the set $\mathcal{U}_{individual}$ (equal among all $\mathcal{U}^i$), is set as $t_{individual}(\xi^2_{(i^*_{individual})}) = \max_{i=1,\ldots,k} \|M^i(\xi^{i,2}_{(i^*_{individual})} - \mu^i)\|_2^2$ where $(\xi^{i,2}_{(i^*_{individual})})_{i=1,\ldots,k}$ is the corresponding partition of $\xi^2_{(i^*_{individual})}$. Using $\|M(\xi-\mu)\|_2^2 = \sum_{i=1}^k \|M^i(\xi^i - \mu^i)\|_2^2$ and the fact that $\sum_{i=1}^k y_i \geq \max_{i=1,\ldots,k} y_i$ for any $y_i \geq 0$, we must have $\|M(\xi-\mu)\|_2^2 \geq \max_{i=1,\ldots,k} \|M^i(\xi^i - \mu^i)\|_2^2$, and so $\rho_{joint} \geq \rho_{individual}$. Note that, when projecting to each constraint, the considered RO is written as

$$\text{minimize } f(x) \ \text{ subject to } \ g_i(x;\xi^i) \in \mathcal{A}_i, \forall \xi^i \in \mathcal{U}^i, \ i = 1,\ldots,l$$

where $\mathcal{U}^i = \{\xi : \|M^i(\xi^i - \mu^i)\|_2^2 \leq \rho_{joint}\}$ and $\{\xi : \|M^i(\xi^i - \mu^i)\|_2^2 \leq \rho_{individual}\}$ for the two cases respectively. Since $\rho_{joint} \geq \rho_{individual}$, we conclude that $f(x_{joint}) \geq f(x_{individual})$. □

Proposition 1 is evident in that the relation $t_{joint}(\xi) \geq t_{individual}(\xi)$ leads to a larger $\mathcal{U}_{joint}$ and hence a smaller resulting feasible region for (2.25) compared with $\mathcal{U}_{individual}$. It hints that, if the data across the constraints are uncorrelated, it is always better to use constraint-wise individual ellipsoids that are calibrated jointly. The same holds if we choose to use diagonalized ellipsoids in our representation, as these satisfy the block-diagonal structural assumption in the proposition. On the other hand, if the data across individual constraints are dependent and we want to capture their correlations in our ellipsoidal construction, the comparison between the two approaches is less clear.

2.8.2 Complexity of Uncertainty Sets

Another consideration in choosing uncertainty set in our framework is the set complexity. For example, we can use an ellipsoidal set with a full covariance matrix, a diagonalized matrix and an identity matrix, the latest leading to a ball. The numbers of parameters in these sets are in decreasing order, making the sets less and less “complex”. Generally, more data supports the use of higher complexity representation, because they are less susceptible to over-fitting. In terms of the average optimal value obtained by the resulting RO, we observe the following general phenomena:

1. Ellipsoidal sets with full covariance matrices are generally better than diagonalized elliposids and balls when the Phase 1 data size is larger than the dimension of the stochasticity. However, if the data size is close to or less than the dimension, the estimated full covariance matrix may become singular, causing numerical instabilities.
2. In the case where ellipsoidal sets are problematic (due to the issue above), diagonalized ellipsoids are preferable to balls unless the data size is much smaller than the stochasticity dimension.

Note that the above observations are consistent with theoretical results in covariance matrix estimation. In particular, it is known that the data size required to accurately estimate the covariance matrix of an m-dimensional random vector is of order (arbitrarily higher than) m if the vector is sub-Gaussian (Theorem 4.7.1 in [165]) and $m\log m$ for more and very general vectors (Theorem 5.6.1 in [165]). This suggests that using fully estimated covariance matrix is desirable over diagonalized matrix when data size is slightly above the dimension.

Table 2.20: Comparing the optimality and feasibility performances between single diagonalized ellipsoid and individually constructed diagonalized ellipsoids, under sample size $n = 120$, and we use $n_1 = 60$ and $n_2 = 60$.

	RO(Single Diagonalized Ellipsoid)	RO(Individual Diagonalized Ellipsoids)
Obj. Val.	-4529.51	-6957.26
$\hat{\epsilon}$	0	3.55×10^{-5}
$\hat{\delta}$	0	0

Table 2.21: Comparing the optimality and feasibility performances between two scaling strategies for reconstructing the uncertainty set.

	Reconstructed RO (Scale 1)	Reconstructed RO (Scale 2)
Obj. Val.	-7880.06	-7541.29
$\hat{\epsilon}$	0.0127	0.0017
$\hat{\delta}$	0.029	0

2.8.3 Missing Details for Section 2.4.3

The example in Section 2.4.3 utilizes the observations discussed in Appendices 2.8.1 and 2.8.2, which we detail below. Since the sample size is less than the stochasticity dimension, we use diagonalized ellipsoids in our constructions. Next, we compare using individualized ellipsoids each for the stochasticity in each constraint versus a single ellipsoid, as depicted in Proposition 1. Table 2.20 column 2 shows the results using a single ellipsoid over vectorized A, and column 3 shows the counterparts for individually constructed ellipsoids. We observe that the latter has a smaller average optimal value (-6957.26 versus -4529.51), which is consistent with the implication from Proposition 1.

We further investigate the use of reconstruction for joint CCP. We use

$$\max_{j=1,\ldots,l}\{(a_j'\hat{x}_0 - b_j)/k_j\}$$

to determine the quantile for calibrating the size of the uncertainty set, where k_j is a scale parameter assigned to constraint j. Table 2.21 compares two natural choices of k_j for the same problem as above but with a different Σ. Column 2 uses $k_j = b_j - \mu_j'\hat{x}_0$, where μ_j' is the sample mean of the Phase 1 data of a_j'. Column

Table 2.22: Optimality and feasibility performances on a single linear CCP with mixture Gaussian distributions for several methods, under sample size 300, and we use $n_1 = 240$ and $n_2 = 60$.

	RO(Unclustered)	RO(Clustered)	Reconstructed RO(Unclustered)	Reconstructed RO(Clustered)
Obj. Val.	-940.502	-961.434	-1074.63	-1087.66
$\hat{\epsilon}$	2.18×10^{-7}	3.01×10^{-6}	0.0162	0.0163
$\hat{\delta}$	0	0	0.05	0.049

3 uses $k_j = std(a_j'\hat{x}_0)$, the standard deviation of the Phase 1 data of $a_j'\hat{x}_0$. While the performances using these two scale parameters can be problem dependent, we observe that the former works better in this example (with a better average optimal value) and hence adopt it for our experiment.

2.9 Integrating with Machine Learning Tools

We provide some numerical results to support the use of the machine learning tools described in Section 2.3. Throughout this section we use the single CCP (2.11) as an example.

2.9.1 Cluster Analysis

To illustrate the use of clustering, suppose ξ follows a mixture of $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ with probabilities $\pi_1 = \pi_2 = 0.5$. Table 2.22 column 2 shows the performance of our RO using a single ellipsoidal set. Column 3 shows the result when we first apply 2-mean clustering to Phase 1 data and construct a union of ellipsoids. The average objective value (-961.434) is demonstrably improved compared to using a single ellipsoid (-940.502). Similarly, the reconstructed RO from using clustering performs better than RO using a single ellipsoid, and both are better than the non-reconstructed counterparts.

2.9.2 Dimension Reduction

To illustrate the use of dimension reduction, we specify ξ as follows. We first generate $\tilde{\xi} \in \mathbb{R}^{11}$ under $N(\mu, \Sigma)$, where μ and Σ are arbitrary vector and positive

Table 2.23: Optimality and feasibility performances on a $d = 1100$ dimensional single linear CCP using PCA, under sample size $n = 120$, and we use $n_1 = 60$ and $n_2 = 60$.

	RO(Diagonalized Ellipsoid)	RO(PCA with 11 Components)
Obj. Val.	-1039	-1189.32
$\hat{\epsilon}$	4.54×10^{-16}	1.43×10^{-5}
$\hat{\delta}$	0	0

definite matrix. We create a higher dimensional $\xi \in \mathbb{R}^{1100}$ by $\xi = P\tilde{\xi} + \omega$, where ω is a "perturbation" vector with each element distributed uniformly on [-0.0005,0.0005] and $P \in \mathbb{R}^{1100\times 11}$.

Table 2.23 column 2 shows the results using RO with a diagonalized ellipsoid on the data of ξ. Diagonalized ellipsoid is used here because the dimension $d = 1100$, which is much larger than the Phase 1 data size $n_1 = 60$, causes singularity issue when constructing a full ellipsoid. Column 3 shows the results when we apply principal component analysis (PCA) to reduce the data to the 11 components having the largest variances and use the linearly transformed ellipsoid (2.10). The number of components 11 is chosen from the cutoff of leaving out 0.01% of the total variance, which we declare as negligible. The PCA approach outperforms the use of a basic diagonalized ellipsoid in terms of average optimal value (-1189.32 versus -1039).

As can be seen in this example, the dimension reduction brought by PCA allows to use a full ellipsoid that captures the shape of the data better on the relevant directions than using the original data, whose high dimension forces one to adopt a simpler geometric set such as diagonalized ellipsoid. Our recommendation in selecting the number of components in PCA is to be conservative, in the sense of choosing one as large as possible so long as it is small enough to support the use of a full ellipsoid (roughly speaking, this means it is smaller than the Phase 1 data size).

Table 2.24: Optimality and feasibility performances on a single linear CCP for basis learning and other methods, under sample size $n = 80$, and we use $n_1 = 21$ and $n_2 = 59$.

	RO(Ellipsoid)	RO(Diagonalized Ellipsoid)	RO(Basis)
Obj. Val.	-1186.86	-946.33	-1016.95
$\hat{\epsilon}$	0.0002	3.03×10^{-4}	1.22×10^{-8}
$\hat{\delta}$	0	0	0

2.9.3 "Basis" Learning

We consider the last approach described in Section 2.3 that surrounds each observation with a ball. For convenience, we call this approach "basis" learning (as we view each of these created balls as a "basis"). We set $\xi \sim \mathcal{N}(\mu, \Sigma)$ for some arbitrarily chosen μ and Σ and $d = 11$. Table 2.24 shows that the basis learning approach (column 4) outperforms the use of a diagonalized ellipsoid (column 3), but underperforms the use of a full ellipsoid (column 2), in terms of average optimal value (-1016.95, -946.33 and -1186.86 respectively). All three approaches are conservative however ($\hat{\delta} \approx 0$). This roughly indicates that basis learning is capable of capturing some covariance information.

Next we generate ξ from a mixture of Gaussian distribution with 5 components and $d = 11$. Table 2.25 shows that basis learning (column 4) outperforms ellipsoid (column 2) in terms of average optimal value (-1033.84 versus -845.973). However, it does not perform as well compared to using the union of 5 ellipsoids from clustering (column 3, with an average optimal value -1090.57). This supports the guidance that, when applying to convoluted data, basis learning is better than using over-simplified shape, but may not work as well compared to other established machine learning tools.

Table 2.25: Optimality and feasibility performances on a single linear CCP for basis learning and other methods, using sample size $n = 300$. For learning-based RO, we use $n_1 = 240$ and $n_2 = 60$.

	SG	RO(Ellipsoid)	RO(Clustered)	RO(Basis)
Obj. Val.	-1191.82	-845.973	-1090.57	-1033.84
$\hat{\epsilon}$	0.037	2.20×10^{-5}	8.73×10^{-12}	0
$\hat{\delta}$	0.125	0	0	0

2.10 Tractable Reformulation of DRO under Ellipsoidal Moment-Based Uncertainty Set

We review the tractable reformulation of moment-based DRO. In particular, we focus on the extension of the DRO reformulation under first and second moment information in [62] using the ellipsoidal uncertainty set suggested in [120].

For single linear CCP with constraint $P(\xi' x \leq b) \geq 1 - \epsilon$, [62] shows that the worst-case constraint, among all distributions generating ξ that have exactly known mean μ and covariance matrix Σ, can be reformulated as

$$\sqrt{\frac{1-\epsilon}{\epsilon}} \|\Sigma^{\frac{1}{2}} x\|_2 - \mu' x - b \leq 0. \tag{2.26}$$

In the situation where μ and Σ are unknown but i.i.d. data are available, we can construct an ellipsoidal moment set $\mathcal{V}$ such that $P\left((\mu, \Sigma) \in \mathcal{V}\right) \geq 1 - \delta$, using the delta method in Section 5 of [120]. We then consider the worst-case chance constraint over distributions with mean and covariance matrix inside $\mathcal{V}$, i.e.,

$$\inf_{Q:(E_Q[\xi], E_Q[(\xi - E_Q[\xi])(\xi - E_Q[\xi])'])\in\mathcal{V}} Q(\xi' x \leq b) \geq 1 - \epsilon \tag{2.27}$$

where Q is a distribution generating $\xi \in \mathbb{R}^d$, $E_Q[\xi]$ is the mean and $E_Q[(\xi - E_Q[\xi])(\xi - E_Q[\xi])']$) the covariance matrix under Q. Given (2.26), the following theorem that extends the result in [120] can be used to provide a tractable reformulation for this worst-case chance constraint.

Theorem II.16. *Let $u \in \mathbb{R}$, $\hat{\Gamma} \in \mathbb{R}^{d \times d}$, $\hat{w} \in \mathbb{R}^d$, $B \in \mathbb{R}^{\frac{d^2+3d}{2} \times \frac{d^2+3d}{2}}$, $\rho \in \mathbb{R}$ be given.*

We set $svec(\Gamma) = [\Gamma_{11}, \sqrt{2}\Gamma_{12}, ..., \sqrt{2}\Gamma_{1n}, \Gamma_{22}, ..., \sqrt{2}\Gamma_{(n-1)n}, \Gamma_{nn}]'$. *The constraint*

$$\sqrt{x'\Gamma x} + w'x + u \leq 0, \forall \begin{pmatrix} w \\ svec(\Gamma) \end{pmatrix} \in \mathcal{U} \tag{2.28}$$

with decision variable $x \in \mathbb{R}^d$, *where* $\mathcal{U} = \mathcal{U}_1 \cap \mathcal{U}_2$ *and*

$$\mathcal{U}_1 = \left\{ \begin{pmatrix} w \\ svec(\Gamma) \end{pmatrix} = B\nu + \begin{pmatrix} \hat{w} \\ svec(\hat{\Gamma}) \end{pmatrix} : \|\nu\|_2 \leq \rho, \nu \in \mathbb{R}^{\frac{d^2+3d}{2}} \right\},$$

$$\mathcal{U}_2 = \left\{ \begin{pmatrix} w \\ svec(\Gamma) \end{pmatrix} : w \in \mathbb{R}^d,\ \Gamma \in S_d^+ \right\},$$

is equivalent to

$$\hat{w}'x + trace(\hat{\Gamma}W) + \rho \left\| B' \begin{pmatrix} x \\ svec(W) \end{pmatrix} \right\|_2 + u + \frac{\eta}{4} \leq 0, \begin{bmatrix} W & x \\ x' & \eta \end{bmatrix} \succeq 0_{(d+1)\times(d+1)} \tag{2.29}$$

where $W \in \mathbb{R}^{d\times d}$ *and* $\eta \in \mathbb{R}$ *are additional (dummy) variables, and* $0_{(d+1)\times(d+1)}$ *is a zero matrix of size* $(d+1) \times (d+1)$.

Theorem II.16 is an application of Theorem 1 (II) in [120] on ellipsoidal uncertainty sets in the form of $\mathcal{U}$. Note that $\mathcal{U}$ consists of two intersecting sets, the ellipsoidal set $\mathcal{U}_1$ constructed from the delta method discussed in [120] that is designed to contain the true moments of ξ with confidence $1 - \delta$, and the set $\mathcal{U}_2$ that constrains the covariance matrix to be positive semidefinite. We reformulate the worst-case chance constraint (2.27) into a semidefinite constraint by rewriting the former in the form (2.28) using (2.26) and applying Theorem II.16.

When ξ has dimension d, the total number of the first and second moments is $(3d + d^2)/2$. To form an ellipsoidal set for all these moments using the delta method, one would need to use the estimated covariance matrix for all these moments, which requires estimating higher-order moments and has size $(3d + d^2)/2 \times (3d + d^2)/2$ (for more details, see Section 5 of [120]). The resulting optimization problem is a semidefinite program with $(5d + 3d^2)/2 + 1$ decision variables.

2.11 Additional Numerical Results

This section shows three additional sets of numerical results. The first is the same example as Section 2.4.1 but with additional non-negativity constraints. These constraints are added to make sure that SG and FAST do not generate unbounded solutions. The second set of results contain a random right hand side quantity in a linear chance constraint. It illustrates how one can use our reconstruction to enhance performance by transforming the safety condition, in the case that a direct use seems un-usable at first. Lastly, we present some further numerical investigation of divergence-based DRO.

2.11.1 Multivariate Gaussian on a Single Chance Constraint with Non-negativity Conditions

We consider a modification of the example in Section 2.4.1

$$\text{minimize } c'x \;\; \text{subject to} \;\; P(\xi' x \leq b) \geq 1 - \epsilon, \; x \geq 0 \tag{2.30}$$

where we add a non-negativity constraint and keep all other parts unchanged. We again consider $d = 11$ and 100. The main purpose of the modification is to eliminate the unbounded solutions that occurred in the $d = 100$ case of (2.11) when we apply SG and FAST. The comparisons among different approaches on this problem, shown in Tables 2.26, 2.27, 2.28 and 2.29, are largely similar to those in Section 2.4.1, but also bear some notable differences that we highlight here.

In the $d = 100$ case, when sample size is small ($n = 120$), SG and FAST can now obtain bounded solutions. However, SG fails to obtain feasible solutions as shown

Table 2.26: Optimality and feasibility performances on a single $d = 11$ dimensional linear CCP with non-negativity constraints for several methods, using sample size $n = 120$. The true optimal value is -1106.23.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	-
n_1	60	60	-	61	-	60	-
n_2	60	60	-	59	-	60	-
Obj. Val.	-924.05	-1070.75	-1068.17	-1060.04	-893.84	0	-1065.59
$\hat{\epsilon}$	4.99×10^{-7}	0.0158	0.0155	0.0119	8.46×10^{-10}	0	0.0072
$\hat{\delta}$	0	0.032	0.019	0.008	0	0	0

Table 2.27: Optimality and feasibility performances on a single $d = 11$ dimensional linear CCP with non-negativity constraints for several methods, using sample size $n = 336$. The true optimal value is -1106.23.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	336	336	336	336	336	336	-
n_1	212	212	-	318	-	168	-
n_2	124	124	-	18	-	168	-
Obj. Val.	-956.63	-1086.28	-1050.52	-1049.82	-921.232	0	-1065.59
$\hat{\epsilon}$	1.34×10^{-6}	0.0244	0.00534	0.00523	6.15×10^{-9}	0	0.0072
$\hat{\delta}$	0	0.045	0	0	0	0	0

Table 2.28: Optimality and feasibility performances on a single $d = 100$ dimensional linear CCP with non-negativity constraints for several methods, using sample size $n = 120$. The true optimal value is -1195.3. Results on moment-based DRO are based on 30 replications due to high computational demand.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	-
n_1	60	60	-	61	-	60	-
n_2	60	60	-	59	-	60	-
Obj. Val.	-832.142	-1111.04	-1195.26	-980.64	-1120.37	0	-1152.35
$\hat{\epsilon}$	0	0.0159	0.458	0.0170	0.0095	0	0.0072
$\hat{\delta}$	0	0.046	1	0.064	0	0	0

Table 2.29: Optimality and feasibility performances on a single $d = 100$ dimensional linear CCP with non-negativity constraints for several methods, using sample size $n = 2331$. The true optimal value is -1195.3. Results on moment-based DRO are based on 30 replications due to high computational demand.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	2331	2331	2331	2331	2331	2331	-
n_1	1318	1318	-	2326	-	1166	-
n_2	1013	1013	-	5	-	1165	-
Obj. Val.	-1005.62	-1164.47	-1156.76	-1155.51	-1033.58	0	-1152.35
$\hat{\epsilon}$	0	0.0397	0.0293	0.0272	5.18×10^{-11}	0	0.0072
$\hat{\delta}$	0	0.058	0	0	0	0	0

by $\hat{\delta} = 1$ in Table 2.28, because the sample size is far smaller than the minimum requirement (2331). FAST obtains confidently feasible solutions that perform better in objective value than our plain RO (-980.64 versus -832.142), but worse than our reconstructed RO (-1111.04), the latter plausibly attributed to the initial solutions of FAST that are not in good quality.

In the $d = 11$ case, SG now achieves feasibility with $n = 120$ samples, and when the minimum required sample size $n = 336$ is used, the solution appears more conservative compared to the counterpart in Section 2.4.1, as shown by $\hat{\delta} = 0$ in Table 2.27 versus $\hat{\delta} = 0.056$ in Table 2.3. This can be explained by the obtained solutions in the current problem being non-fully-supported (i.e., the number of support constraints is less than d, which gives the problem a lower “intrinsic” dimension). Note that when the sample size increases from 120 to 336, the solutions of SG necessarily become more conservative (regardless of the dimension in consideration), which is a consequence of the nature of constraint addition in SG. On the other hand, the solutions in our RO improve as sample size increases, plausibly attributed to a better estimation of HPR. Reconstructed RO provides better solutions than SG and FAST in all four sets of experiments. Nonetheless, we should mention that some constraint removal approaches like sampling-and-discarding in [38] are available to enhance the performances of SG. Finally, since the performances of DROs and SCA follow similarly as in Section 2.4.1, we do not restate the comparisons with them here.

2.11.2 Multivariate Gaussian on a Single Chance Constraint with Random Right Hand Side

We continue to consider the single linear CCP in (2.11), but with the right hand side quantity b being random. Specifically, we set b to be generated from a Gaussian distribution with mean 1200 and variance 100 (in this case, b is almost positive for sure). The rest of the problem follows from Section 2.4.1. Note that, by the discussion at the end of Section 2.2.4, a direct use of reconstruction would not improve the solution in this example. However, we can divide b on both sides of the inequality in the safety condition, which now gives a right hand side value 1 and transformed stochastiticities as the ratios of ξ and b.

Tables 2.30 and 2.31 present the experiments on a $d = 11$ dimensional problem with $n = 120$ and $n = 336$ sample sizes respectively. The performances of the presented approaches are consistent with the experiments in Sections 2.4.1 and 2.4.2. Specifically, when the sample size is small ($n = 120$), our RO is preferable to SG, as it obtains feasible solutions while SG fails. Reconstruction applied on the described transformed problem continues to work and perform competitively against FAST and SCA. In particular, when $n = 120$, it outperforms FAST in terms of achieved objective value, but slightly falls short of SCA. When $n = 336$, reconstructed RO, SG, FAST and SCA all perform very similarly. Note that SCA have assumed moment information and hence are given an upper hand in this example.

DROs contine to be conservative in this experiment. Moment-based DRO is outperformed by both plain and reconstructed ROs in both the $n = 120$ and $n = 336$ cases. Similar to the example in Section 2.4.1, KL-divergence-based DRO obtains an adjusted tolerance level $\epsilon^* = 0$, which forces the decision x to satisfy the safety condition $\xi' x \leq b$ for all $\xi \in \mathbb{R}^d, b \in \mathbb{R}$, and in this case leads to an infeasible problem.

2.11.3 Additional Numerical Investigation on DRO with KL Divergence

We provide more details on constructing KL-divergence balls in DRO, which has been used in our numerical comparisons. In the case of continuous distributions for generating ξ, constructing KL balls requires estimating a reference distribution f_0

Table 2.30: Optimality and feasibility performances on a single $d = 11$ dimensional linear CCP with random right hand side for several methods, using sample size $n = 120$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	-
n_1	60	60	-	61	-	60	-
n_2	60	60	-	59	-	60	-
Obj. Val.	-1143.45	-1173.62	-1182.90	-1167.61	-1138.49	infeasible	-1175.05
$\hat{\epsilon}$	7.60×10^{-6}	-	0.0170	0.0910	1.00×10^{-7}	-	0.0074
$\hat{\delta}$	0	0.045	0.958	0.053	0	-	0

Table 2.31: Optimality and feasibility performances on a single $d = 11$ dimensional linear CCP with random right hand side for several methods, using sample size $n = 336$.

	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	336	336	336	336	-	336	-
n_1	212	212	-	318	-	168	-
n_2	124	124	-	18	-	168	-
Obj. Val.	-1149.31	-1175.70	-1178.00	-1178.01	-1143.70	infeasible	-1175.05
$\hat{\epsilon}$	1.60×10^{-6}	0.0253	0.051	0.0238	1.00×10^{-7}	-	0.0074
$\hat{\delta}$	0	0.035	0.051	0.052	0	-	0

(center of the ball) using kernel density estimation, and then a k-NN or other similar methods to estimate the set size. This selection of the reference distribution aims to approximate the true distribution as much as possible, and the set size is chosen such that the divergence ball contains the true distribution with high confidence. Below we detail these procedures, followed by a very low-dimensional example where these procedures work in calibrating DRO and allow illustrative comparisons with other approaches.

Bandwidth Selection for Kernel Density Estimation

Following [96], we use kernel density estimation to estimate the reference distribution f_0. This estimation procedure requires the proper selection of a bandwidth parameter, whose theoretical optimal choice is of order $N^{-\frac{1}{m+4}}$, where N is the sample size and m is the dimension of the randomness. In the following, we consider bandwidth in the form of $BN^{-\frac{1}{m+4}}$ for some $B \in \mathbb{R}$.

We investigate how the divergence between the reference and the true distributions

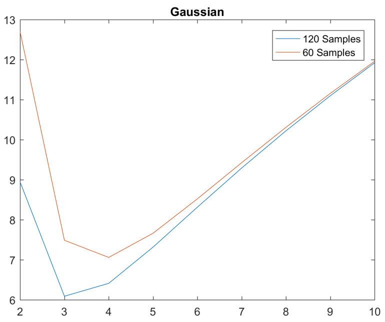


Figure 2.3: Divergence with different bandwidth parameter B and sample size $N = 120, 60$. The randomness is Gaussian distributed with dimension $m = 11$.

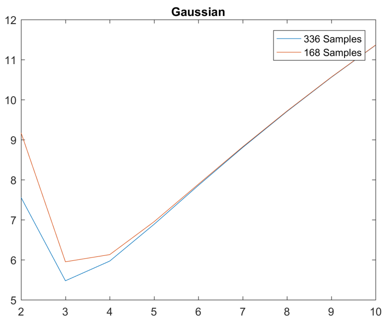


Figure 2.4: Divergence with different bandwidth parameter B and sample size $N = 336, 168$. The randomness is Gaussian distributed with dimension $m = 11$.

varies with the bandwidth parameter used to estimate the reference. We consider a Gaussian distribution with dimension $m = 11$, and sample sizes $N = 120$ and $N = 336$ (which are considered in Section 2.4.1). Figures 2.3 and 2.4 show the KL divergence (estimated from 100,000 Monte Carlo samples drawn from the true distribution) against the bandwidth choice. In the figures we also show results with half of the samples sizes to give a sense of the sensitivity (and also motivated from the necessity of data splitting to be discussed momentarily). Among all the choices, $B = 3$ appears the best as it gives the smallest divergence in three out of four different sample sizes. Figures 2.5 and 2.6 further show the divergences between reference and true distributions when the truth follows other distributions, namely a Gaussian distribution with dimension $m = 100$ and a log-normal distribution with dimension $m = 11$ respectively. We see that the graphs behave very differently from each other and the optimal bandwidth choices now deviate from 3, thus showing that the optimal bandwidth can depend heavily on the underlying distribution.

We note that the constructed f_0's using kernel density estimation seem to be quite far from the true distribution. For example, in the problem considered in Section 2.4.1, the KL divergence needs to be smaller than 1.25 in order to achieve a non-

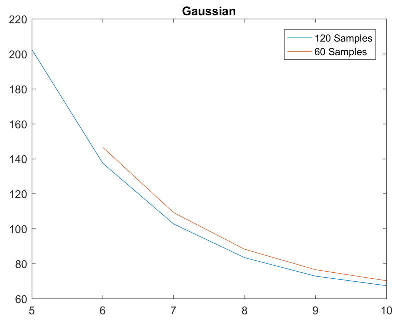


Figure 2.5: Divergence with different bandwidth parameter B. The randomness is Gaussian distributed with dimension $m = 100$.

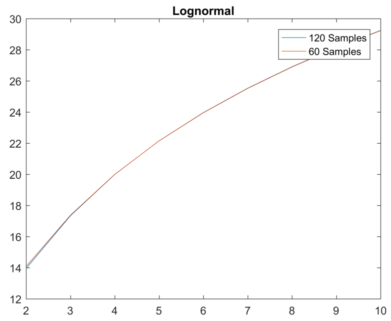


Figure 2.6: Divergence with different bandwidth parameter B. The randomness is log-normal distributed with dimension $m = 11$.

trivial solution. This is substantially smaller than 5.5, the lowest observed divergence value among all of Figures 2.3–2.6. In other words, kernel density estimation is not efficient enough to obtain a good enough reference distribution for implementing DRO in this case.

Construction of Divergence-Based Uncertainty Sets

Once we obtain a reference distribution, the next task is to calibrate the size of the uncertainty set. More precisely, we need to determine γ for the set $\{f : \mathcal{D}(f\|f_0) \leq \gamma\}$, where $\mathcal{D}$ denotes the KL divergence, to cover the true distribution (with high confidence). This calls for the literature of divergence estimation. Here, we discuss the k-NN estimator studied by [167, 139]. But before we proceed, we note that since f_0 itself is estimated from data, we need to be careful in controlling the statistical error in simultaneously estimating f_0 and γ. We consider two approaches. One is to use all the data to construct f_0 and reuse the same data to estimate γ. Another is to split the data into two groups, one for estimating f_0 and another for γ. In our experiments, the first approach turns out to consistently give a negative γ, indicating a poor estimation error (which is expected as the combined statistical error from f_0 and γ is hard to control). Therefore, we adopt the second approach that splits the data.

We investigate the quality of k-NN estimation with different choices of k, using an example of Gaussian distribution with $m = 11$ and sample size 336. Here we split the data into two equal halves, and use the first half to estimate f_0 with bandwidth $B = 3$ and the second half to estimate the divergence to calibrate γ. Figure 2.7 shows the average point estimate of the divergence using k-NN among 1000 experimental replications, against k. We see that $k = 1$ gives the closest estimate to the true divergence (5.5, using $B = 3$ in Figure 2.4). This observation is consistent with the known result in the literature that $k = 1$ gives the smallest bias. However, even in this case the bias is still substantial, likely due to insufficient sample size. The performance is worse as k increases.

Figure 2.8 further shows the histogram of divergence estimates from 1000 experimental replications with $k = 1$. The distribution of the estimates appears very spread out. Moreover, the biggest realized estimate (less than 3.5) is still far away from the true divergence (5.5 in Figure 2.4). As noted in [167], estimating divergence for high-dimensional distributions with small sample typically incurs large variances and is challenging, in line with our observations here. For problems with even higher dimension (e.g., the setting in Figure 2.5), we expect it to be even more difficult to obtain a reasonable divergence estimate.

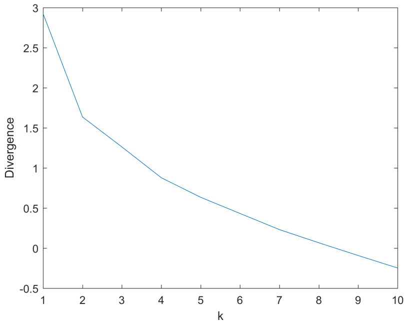


Figure 2.7: Estimated divergence with different k-NN parameter k using 336 samples.

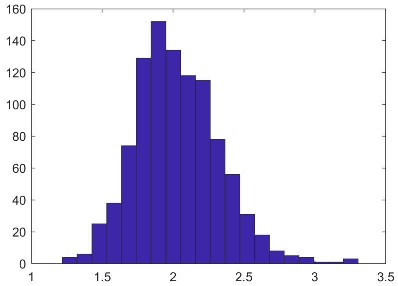


Figure 2.8: Histogram of the divergence estimates for γ with $k = 1$ and 336 samples.

CHAPTER III

Tail Uncertainty in Rare-event Simulation

3.1 Introduction

Assessing rare-event probabilities and extremal measures for the likelihood of catastrophic events is ubiquitous in risk analysis and management. Examples include the prediction of large asset losses in finance, imbalance of cash flows and ruin in insurance, and system overloads in service operations. In many cases, these extremal quantities are outputs that rely on underlying, granular stochastic components. For example, a financial portfolio may consist of a weighted combination of assets each having its own (correlated) return pattern, and an insurance portfolio consists of the cash flows of many different policyholders. Estimating these extremal quantities hinges on the provision of accurate probabilistic descriptions of these input components, with any deviations away from the reality leading to potential errors or even meaningless estimates.

The latter issue has been studied and has gathered growing literature in recent years, generally known as the problem of model uncertainty or input uncertainty. Its main focus is to develop methodologies that can quantify the impact of model misspecifications or errors that propagate to output estimation or decision-making. See, e.g., [8, 84, 47, 9, 157, 101] in the stochastic simulation literature, and [137, 82, 73, 111] in finance, economics, control and operations management applications. In the extremal estimation setting, this problem is intimately related to extreme value theory, in which one attempts to extrapolate the tail beyond the scope of data in a statistically justified fashion, along with uncertainty quantification [65, 64]. Recently, the framework of so-called distributionally robust optimization [51, 170, 11, 62] has been studied to construct bounds on extremal measures with additional robustness properties beyond statistical asymptotics. This approach utilizes postulations such as the acknowledgement of the true distribution within a neighborhood of a baseline model measured by a suitable statistical distance [6, 26], marginal information and extremal coefficients [66, 146, 166, 56, 145, 67, 171], moments and shape assumptions on the tail such as monotonicity or convexity [104, 163, 109]. In simulation-based rare-event analysis, [128, 129] studied methods to efficiently compute sensitivities of rare-event probabilities with respect to model parameters, and [130] proposed an averaging of distributions to fit input models to enhance tail performances.

In contrast to most past literature that focused on the technique in quantifying model uncertainty impacts, here we address several validity questions that arise when, given input data, a modeler chooses to use "standard" approaches to obtain estimates and quantify uncertainty, namely:

1. By simply using the empirical distribution as my input model fit, would the rare-event estimate be reasonably close to the truth? (assuming computational or Monte Carlo noise is negligible)
2. Following the point estimate in Question 1, would it work if one runs a bootstrap to obtain a confidence interval that accounts for the input data noise?
3. If the bootstrap does not work, would incorporating extreme value theory in fitting the input tail helps with more reliable uncertainty quantification?

Our viewpoint is that the main source of uncertainty in determining the accuracy of rare-event estimation comes from the lack of knowledge of the tail of the input models. The main body (i.e., non-tail) part of the input distribution can be fit by both parametric and nonparametric techniques, where there are typically adequate data to perform such fit (and in Question 1 above, we simply use the empirical distribution as the fit). However, it is the portion beyond the scope of data that determines the distributional tail and in turn the rare-event behaviors. Thus, before we go to the above questions, we first focus on:

0. How does truncating the tail of the input model affect the rare-event estimate?

Our main contention is that heavy-tailed problems could be much more challenging than light-tailed counterparts regarding estimation and uncertainty quantification using the standard approaches in Questions 1-3. This challenge roots from Question 0 in that truncating the input tail in a heavy-tailed system exerts a huge effect on the rare-event estimate, when the truncation level represents the typical level of knowledge that the data informs (e.g., the top 1% or 0.1% of the data). As a consequence, using empirical distribution, or bootstrap on the empirical distribution, which significantly ignores the tail content, would fail to estimate the rare-event quantity and vastly under-estimate the uncertainty. Using extreme value theory in Question 3 to extrapolate tail (such as the peak-over-threshold method, e.g. [106]) helps to an extent, but could introduce extra bias, at least using our fitting methods (though we should point out that better techniques are available). On the other hand, the effect of missing tails on light-tailed estimation is relatively milder.

The larger effect from truncating the input tail on heavy-tailed problems can be explained from their large deviations behaviors that pertain to the one or several "big jumps" [65, 54, 147], i.e., to invoke a rare event, one or several input components exhibit huge values. On the other hand, light-tailed systems invoke rare events by having each component contributing a small shift and adding these contributions. Thus, to accurately estimate a heavy-tail rare event, one needs to accurately estimate the far tail of each input component, whereas this is not necessary in light-tailed systems. In fact, ignoring the tail of heavy-tail inputs would lead to estimates as if the system is light-tail, and the ultimate effect could be that the estimation error is as large as the rare-event probability of interest, deeming the estimation meaningless.

We point out that, regarding Question 0, our study is related to [60, 22, 103] and especially [134] and [95]. [60] investigated the sensitivities on the large deviations rate when the input model deviates within a Rényi divergence ball. [22] showed in a similar context that imposing a single ball over all inputs, thus allowing the distortion of dependency structure among the inputs, can lead to a substantially heavier tail than the original model when the Kullback-Leibler divergence is used. [103] studied robust rare-event simulation when the input tail is unknown but subject to geometric assumptions. [134] studied the impacts on the waiting times when the tail of service times is misspecified or truncated. Relating to [95], they investigated the truncation threshold needed to retain the heavy-tail characteristic of a system. They also contrasted it with the light-tail case and observed that the required threshold is higher for heavy tail. Our observation in this regard is thus similar to [134], but with a different setting (aggregation of i.i.d. variables) and focus on the statistical implications asked in Questions 1-3. Moreover, we investigate extensively on the numerical evidence and identify situations where the theoretical findings hold or deviate.

In the remainder of this chapter, we will focus on a basic setup on the overshoot of an aggregation of i.i.d. variables. Section 3.2 describes the estimation target and explains the impacts of tail truncation in light- versus heavy-tailed cases. Section

3.3 shows the numerical results and comparisons in input tail truncation, and the use of empirical distributions and bootstrapping. We leave the full derivations and generalizations to the journal version of this work.

3.2 Setting and Theory

We consider estimating the overshoot of an aggregation of n i.i.d. variables, i.e., consider $p = P(S_n > \gamma)$ where $S_n = X_1 + \cdots + X_n$ and $X_i \in \mathbb{R}$ are i.i.d. variables drawn from the distribution F. We denote X as a generic copy of X_i for convenience. We assume the density of X exists and denote as f. Correspondingly, we let $\bar{F}(x) = 1 - F(x)$ be the tail distribution function. We let $\mu = E[X] < \infty$. Suppose $\gamma = \gamma(n)$ is a high level that grows to ∞ as $n \to \infty$. Throughout this chapter, for any sequences $a_n, b_n \in \mathbb{R}$ we write $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$, $a_n = \omega(b_n)$ if $a_n/b_n \to \infty$ as $n \to \infty$, and $a_n = \Theta(b_n)$ if there exists an integer n_0 such that $\underline{M} \leq |a_n/b_n| \leq \overline{M}$ for $n \geq n_0$ and $0 < \underline{M} \leq \overline{M} < \infty$.

It is well known that, if $\gamma = bn$ for some constant $b > \mu$ and X possesses exponential moments, then under mild additional assumptions p decays exponentially in n [52]. On the other hand, if X is Pareto-tailed, then, as $\gamma = \omega(\sqrt{n})$, p approximately equals $P(\max_i X_i > \gamma - n\mu)$ or $n\bar{F}(\gamma - n\mu)$, which corresponds to the one-big-jump behavior.

Our investigation pertaining to Question 0 is the following. Suppose we truncate the distribution $F(x)$ at the point u so that the density becomes 0 for $x > u$, i.e., consider the truncated distribution function given by

$$\tilde{F}_u(x) = \begin{cases} F(x)/F(u) & \text{for } \ x \leq u \\ 1 & \text{for } \ x > u \end{cases}$$

and correspondingly the truncated density $\tilde{f}_u(x) = (f(x)/F(u))I(x \leq u)$, where

$I(\cdot)$ denotes the indicator function. For convenience, denote $p(G)$ as the probability $P_G(S_n > \gamma)$ where X_i's are governed by an arbitrary distribution G, and we simply denote $P(S_n > \gamma)$ if X_i's are governed by F. We consider the approximation error $p(\tilde{F}_u) - p(F)$.

Note that, roughly speaking, this situation captures the case where we use the empirical distribution to plug into our input model, so that the probability mass is zero for regions outside the scope of data or close to zero at the very tail of the data. The proportional constant $F(u)$ is introduced to ensure a proper truncated distribution and has little effect on the mass below u when u is reasonably big.

By definition, the approximation error is

$$p(\tilde{F}_u) - p(F) = \frac{P(S_n > \gamma, X_i \leq u \ \forall i = 1, \ldots, n)}{F(u)^n} - P(S_n > \gamma). \tag{3.1}$$

3.2.1 Heavy-Tail Case

We first consider the Pareto-tail case. Suppose that $\bar{F}$ has a regularly varying tail in the form

$$\bar{F}(x) = L(x)x^{-\alpha}(1 + o(1)) \tag{3.2}$$

for some slowly varying function $L(\cdot)$ and $\alpha > 2$, and $E|X_i|^{2+\delta} < \infty$.

Suppose $n \to \infty$ and $\gamma = \Theta(n)$ (or more generally $\gamma = \omega(\sqrt{n \log n})$). In this case, it is known that $P(S_n > \gamma)$ is approximately $P(\text{at least one } X_i > \gamma - n\mu)$, or probabilistically, that the rare event $S_n > \gamma$ happens most likely due to a big jump from one of the X_i's (e.g. [65]). Thus, if the truncation level u is too small compared to $\gamma - n\mu$, then the big jump that contributes to the dominating mass of the rare event is barred, making $P(S_n > \gamma, X_i \leq u \ \forall i = 1, \ldots, n)$ substantially smaller than $P(S_n > \gamma)$. In this situation, $p(\tilde{F}_u)$ becomes negligible compared to $P(S_n > \gamma)$, and the approximation error (3.1) is effectively $-P(S_n > \gamma)$. In other words, using a truncated input distribution leads to a substantial under-estimate with a bias almost equal to the magnitude of the rare-event probability itself.

Alternately, we can write the approximation error (3.1) as

$$\frac{-P(S_n > \gamma,\ \text{at least one } X_i > u) + P(S_n > \gamma)(1 - F(u)^n)}{F(u)^n}. \tag{3.3}$$

Again, when u is relatively small compared to $\gamma - n\mu$, then the event $\{\text{at least one } X_i > u\}$ inside the probability $P(S_n > \gamma,\ \text{at least one } X_i > u)$ is redundant, making this probability asymptotically equivalent to $P(S_n > \gamma)$ and that (3.3) is asymptotically equivalent to $-P(S_n > \gamma)$.

We summarize the above as:

Theorem III.1. *Suppose X_i's are i.i.d. random variables with regularly varying tail distribution $\bar{F}$ in the form* (3.2) *with $\alpha > 2$ and $E|X|^{2+\delta} < \infty$. Let $n \to \infty$ and $\gamma = n\mu + \omega(\sqrt{n \log n})$. Assume $u \leq (\gamma - n\mu)/\sqrt{\log n}$. The discrepancy between using a truncated distribution $\tilde{F}_u$ and the original distribution F in evaluating the probability $p(F) = P(S_n > \gamma)$ as $n \to \infty$ is given by*

$$p(\tilde{F}_u) - p(F) = -p(F)(1 + o(1)).$$

Proof. Using equation (1.45) in [127], when $(\gamma - n\mu)/\sqrt{n \log n} \to \infty$ and $u \leq (\gamma - n\mu)/\sqrt{\log n}$, we have $P(S_n > \gamma, X_i \leq u\ \forall i = 1, \ldots, n) = o(n\bar{F}(\gamma - n\mu))$. Moreover, by Theorem 1.9 in [127] (or equation (1.25b) therein), we have $P(S_n > \gamma) = n\bar{F}(\gamma - n\mu)(1 + o(1))$ under the given conditions. Thus $P(S_n > \gamma, X_i \leq u\ \forall i = 1, \ldots, n) = o(P(S_n > \gamma))$.

Moreover, note that if $u = (\gamma - n\mu)/\sqrt{\log n}$, we have $F(u)^n \to 1$. Thus

$$\frac{P(S_n > \gamma, X_i \leq u \ \forall i = 1, \ldots, n)}{F(u)^n} = o(P(S_n > \gamma))$$

if $u = (\gamma - n\mu)/\sqrt{\log n}$. However, since the truncated distribution $\tilde{F}_u$ stochastically dominates $\tilde{F}_{u'}$, i.e., $\bar{\tilde{F}}_u(\cdot) \geq \bar{\tilde{F}}_{u'}(\cdot)$, for any u, u' such that $u > u'$, we must have, for given γ, $P(S_n > \gamma, X_i \leq u \ \forall i = 1, \ldots, n)/F(u)^n$ non-decreasing in u. Therefore we have

$$\frac{P(S_n > \gamma, X_i \leq u \ \forall i = 1, \ldots, n)}{F(u)^n} = o(P(S_n > \gamma))$$

for any $u \leq (\gamma - n\mu)/\sqrt{\log n}$. This concludes the theorem. □

Consider the case $\gamma = bn$ for some $b > \mu$. Theorem III.1 states that when u is below $(b - \mu)n/\sqrt{\log n}$, the rare-event estimation is essentially void, at least asymptotically. Note that this threshold is approximately linear in n. When the number of input components n is large, it could be difficult to sustain an accuracy level given a finite set of input data.

3.2.2 Light-Tail Case

We now consider X that possesses finite exponential moment, i.e., the logarithmic moment generating function $\psi(\theta) = \log E[e^{\theta X}] < \infty$ for θ in a neighborhood of 0. Consider $\gamma = bn$ for some constant $b > \mu$. Suppose that there exists a unique solution θ^* to the equation $b = \psi'(\theta)$. Then $p(F) = P(S_n > \gamma)$ exhibits exponential decay as $n \to \infty$, i.e., $-(1/n) \log p(F) \to I$ where I is the rate function given by the Legendre transform or the convex conjugate of $\psi(\theta)$

$$I = \sup_{\theta} \left\{ b\theta - \psi(\theta) \right\}.$$

In fact, if X is further assumed non-lattice, we have the following more accurate asymptotic [30]

$$P(S_n > \gamma) = \frac{1}{\theta^* \sqrt{2\pi\psi''(\theta^*)n}} e^{-nI}(1 + o(1)).$$

We have:

Theorem III.2. *Consider $\gamma = bn$ for some constant $b > \mu$. Suppose that F is non-lattice, satisfies $\psi(\theta) < \infty$ for θ in a neighborhood of 0, and there exists a unique solution θ^* to the equation $b = \psi'(\theta)$. Then, as long as the truncation level u is chosen such that $ne^{\theta^* u}\bar{F}(u) \to 0$, the discrepancy between using a truncated distribution $\tilde{F}_u$ and the original distribution F in evaluating the probability $p(F) = P(S_n > \gamma)$ is asymptotically negligible, i.e.,*

$$p(\tilde{F}_u) - p(F) = -o(p(F)).$$

Sketch of Proof. We consider the rate function corresponding to $p(\tilde{F}_u)$, given by

$$I_u = \sup_\theta \{b\theta - \psi_u(\theta)\}$$

where $\psi_u(\theta)$ denotes the logarithmic moment generating function of $\tilde{F}_u$, namely $\log(E[e^{\theta X}; X \leq u]/F(u))$. Now, consider a change of variable $r = F(u)$, and abuse notation slightly to write $\psi_r(\theta) = \log(E[e^{\theta X}; X \leq F^{-1}(r)]/r)$ and the corresponding rate function as I_r. By Taylor series expansion we have, as $u \to \infty$ or $r \to 1$,

$$I_u \approx I - \frac{d}{dr}\psi_r(\theta^*)(F(u) - 1) \tag{3.4}$$

where $-\frac{d}{dr}\psi_r(\theta^*)$ is the derivative of I_r by the generalized Danskin's Theorem [48].

Note that

$$\begin{aligned}\frac{d}{dr}\psi_r(\theta^*) &= \frac{e^{\theta^* F^{-1}(r)} f(F^{-1}(r))}{f(F^{-1}(r)) E[e^{\theta^* X}; X \leq F^{-1}(r)]} - \frac{1}{r} \\ &= \frac{e^{\theta^* u}}{E[e^{\theta^* X}; X \leq u]} - \frac{1}{F(u)}.\end{aligned}$$

Hence from (3.4) we have

$$I_u \approx I + \left(\frac{e^{\theta^* u}}{E[e^{\theta^* X}; X \leq u]} - \frac{1}{F(u)} \right) \bar{F}(u).$$

Now, one can show that, as $u \to \infty$, we have $\theta^* + \delta$ with $\delta \to 0$. Thus the approximation error $p(\tilde{F}_u) - p(F)$ is given by

$$\begin{aligned} & \frac{P(S_n > \gamma, X_i \leq u \ \forall i = 1, \ldots, n)}{F(u)^n} - P(S_n > \gamma) \\ \approx \ & \frac{1}{(\theta^* + \delta)\sqrt{2\pi \psi''(\theta^* + \delta, u) n}} e^{-nI - n\left(\frac{e^{\theta^* u}}{E[e^{\theta^* X}; X \leq u]} - \frac{1}{F(u)}\right)\bar{F}(u)} \qquad (3.5) \\ & - \frac{1}{\theta^* \sqrt{2\pi \psi''(\theta^*) n}} e^{-nI}. \qquad (3.6)\end{aligned}$$

Thus, when

$$n \left(\frac{e^{\theta^* u}}{E[e^{\theta^* X}; X \leq u]} - \frac{1}{F(u)} \right) \bar{F}(u) \to 0 \tag{3.7}$$

we have (3.6) being asymptotically negligible compared to $(1/(\theta^* \sqrt{2\pi\psi''(\theta^*) n}))e^{-nI}$, which would conclude our claim. Finally, we only need to observe that (3.7) is equivalent to $ne^{\theta^* u}\bar{F}(u) \to 0$. □

Theorem III.2 postulates that as long as the truncation level u is chosen high enough relative to n such that $ne^{\theta^* u}\bar{F}(u) \to 0$, then the model error in using the truncated input is negligible. In contrast to the heavy-tail case, this condition on u dictates typically a logarithmic requirement on n. For instance, if F is an exponential distribution, say with rate λ, then we have $ne^{-(\lambda - \theta^*)u} \to 0$ which holds as long as $n = \omega(\log n)$. If F is a Gaussian distribution, say with mean μ and variance σ^2, then we have $ne^{\theta^* u - (u-\mu)^2/(2\sigma^2)}/\sqrt{2\pi u} \to 0$ which holds as long as $u = \omega(\sqrt{\log n})$.

3.3 Numerical Experiments

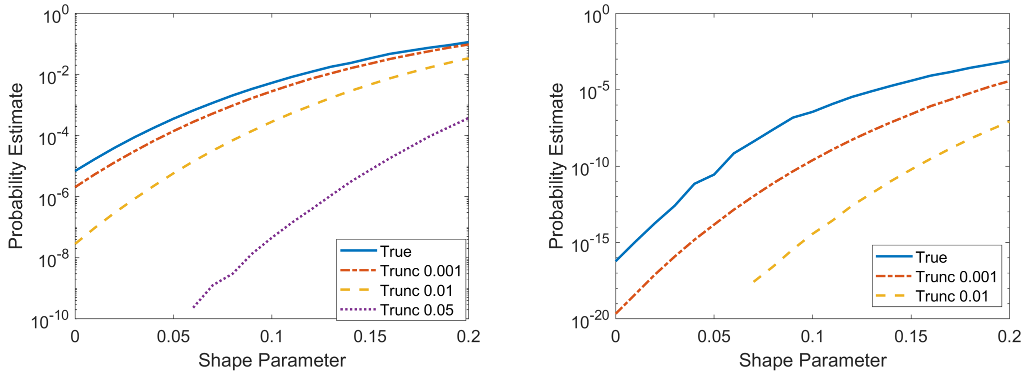
We consider estimating the probability $p = P(\sum_{i=1}^{n} X_i > \gamma)$, with different number of variables n, rarity levels γ and distributions of X_i. In each of our experiments, we implement variance reduction techniques (including importance sampling and conditional Monte Carlo; for further details, see [5]) to achieve better computation efficiency and use sufficient samples to ensure negligible simulation noise. We investigate the effect of input tail truncation (Question 0) in Section 3.3.1, using empirical distribution (Question 1) in Section 3.3.2, bootstrap (Question 2) in Section 3.3.3, and using peak-over-threshold (Question 3) in Section 3.3.4.

3.3.1 Truncating Input Tail

We test with truncation points on the input distribution corresponding to 0.05, 0.01 and 0.001 tail probability masses respectively, i.e., t such that $P(X > t) = \alpha$ where $\alpha = 0.05,\ 0.01,\ 0.001$ (We shall refer to as the α tail quantile). When we truncate the distribution at t, we only use $f(x|x < t)$ to generate X_i's. The estimate with untruncated distribution is set as a baseline for comparison.

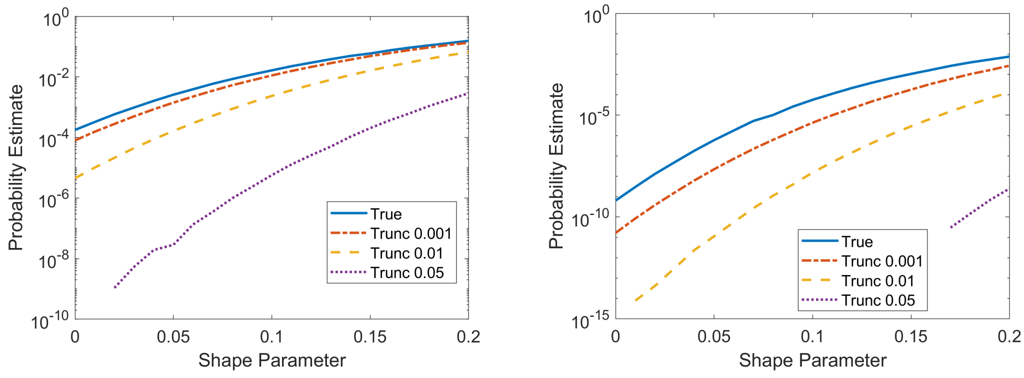
We generate X_i's using the generalized Pareto distribution, varying the shape parameter ξ and the scale parameter σ and fixing the threshold parameter to be 0. Note that when $\xi = 0$, the distribution is light-tailed (equivalent to exponential distribution); when $\xi > 0$ the distribution is heavy-tailed and larger ξ gives heavier tail. We vary ξ from 0 to 0.2 to observe what would change if the tail part grows heavier. When ξ varies, we keep the mean of the distribution to be 1 by letting $\sigma = 1 - \xi$. Our experiments also include different settings of γ and n. Figures 3.1 and 3.2 show the experiment results.

Before comparing light and heavy tails, we note that the tail part of the dis-



(a) (b)

Figure 3.1: The probability estimation with untruncated and truncated distributions. “Trunc 0.001” denotes the probability estimate using distribution turncated at 0.001 tail quantile. (a) $\gamma = 60,\ n = 30$. (b) $\gamma = 100,\ n = 30$.



(a) (b)

Figure 3.2: The probability estimation with untruncated and truncated distributions. “Trunc 0.001” denotes the probability estimate using distribution turncated at 0.001 tail quantile. (a) $\gamma = 40,\ n = 20$. (b) $\gamma = 60,\ n = 20$.

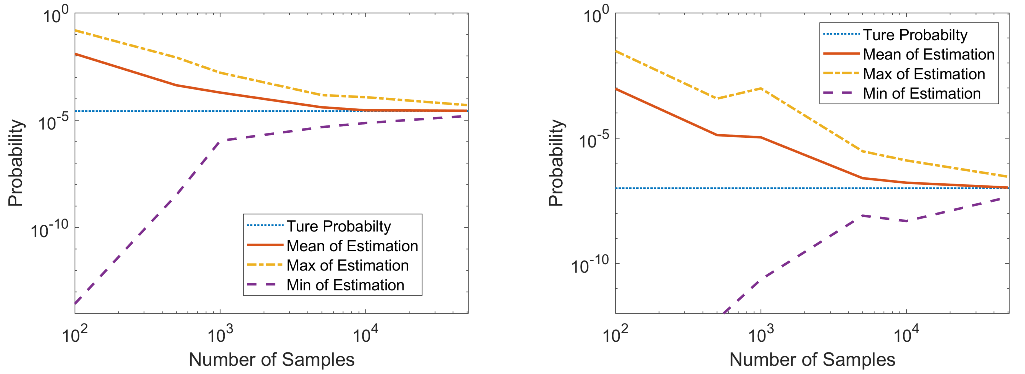
tribution is generally quite important to the probability estimation. This claim is supported by the gaps between the probability estimates using true distribution and truncated distributions in Figures 3.1 and 3.2. For instance, suppose we truncate at the 0.01 tail quantile. The gap between the estimate with the truth (between the blue solid line and the yellow dash) is roughly greater than one order of magnitude in almost all cases, which means we are not able to estimate a correct scale of the probability without the 0.01 tail information. Moreover, for fixed γ and n, we see a smooth trend of the estimates (in Figures 3.1 and 3.2) when the shape parameter increases from 0.

Next we compare the impacts between light and heavy tails. Although the trends in the figures seem to suggest smaller gaps as ξ increases (heavier tail), a more proper comparison should fix the target probability level and the truncation level. In this case, the impact of the heavier tail appears larger. In particular, we compare the gaps between the estimate with true distribution (blue solid line) and truncated distribution (orange dash-dot line) at the 0.001 tail quantile in Figure 3.1a at shape parameter value around 0 and in Figure 3.1b at value around 0.2. In these two cases, the objective probabilities have similar values and the truncated tail quantile are the same, and in the heavy-tail case, a larger gap can be seen. More specifically, the gap is smaller than one order of magnitude in light tail (Figure 3.1a at 0) and larger than one order of magnitude in heavy tail (Figure 3.1b at 0.2).

3.3.2 Data-driven Rare-Event Simulation

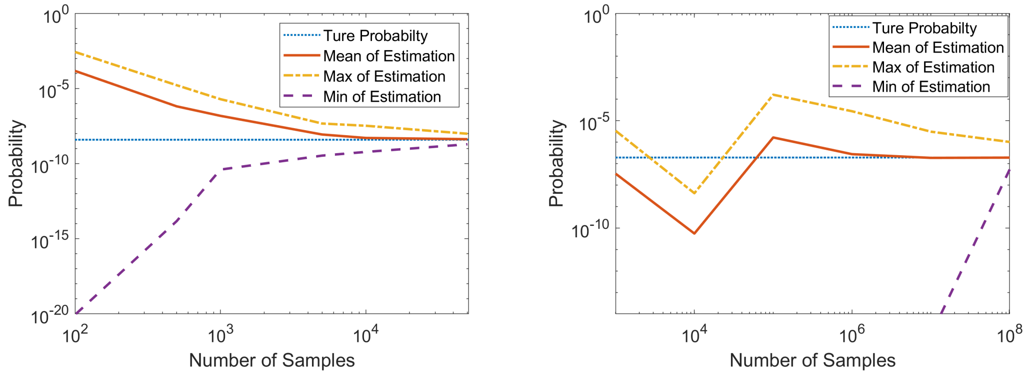
We consider the data-driven situation (i.e. when distribution of X is unknown but data is available) and use empirical distributions to drive the simulation of p. Here we set X_i as Gaussian distribution and generalized Pareto distribution with $\xi = 0.2$. We independently generate data sets of X_i's for 100 replications to construct empirical distributions to drive the simulation. Figures 3.3 and 3.4 show the true probability, the averaged estimates from all replications, and also the maximum and minimum estimates among the replications to provide a measure of variability.

In the light-tail cases, Figures 3.3 and 3.4a show similar trends in that the estimation variability reduces as the number of samples increases (the maximum and minimum values approach to the true probability as sample increases). The variability is higher when the rarity level grows, which can be observed from the slower convergence of the maximum and minimum estimates (Figure 3.4a versus Figure



(a) (b)

Figure 3.3: The estimation performance with different number of samples based on 100 replications. (a) Gaussian distribution, $\gamma = 70$. (b) Gaussian distribution, $\gamma = 90$.



(a) (b)

Figure 3.4: The estimation performance with different number of samples based on 100 replications. (a) Gaussian distribution, $\gamma = 100$. (b) Generalized Pareto distribution, $\gamma = 100$.

3.3a).

On the other hand, the performance in the heavy-tail cases (Figure 3.4b) is more "abnormal". When the sample size is small, e.g. 10^4, the maximum estimate from the 100 replications is smaller than the true probability. Even with more samples (10^5 and 10^6), in most (92 and 75 out of the 100 respectively) replications, we obtain overly small estimates compared to the true probability (a difference of more than 5 orders of magnitude). These suggest a severe underestimation in the heavy-tail problem with limited data.

Table 3.1: The coverage and the width of plain bootstrap confidence interval. The results are computed from 30 replications. The rare event problem is defined by the sum of standard Gaussian variables.

	$\gamma = 70$, $p = 2.6561 \times 10^{-5}$		$\gamma = 100$, $p = 3.882 \times 10^{-9}$		$\gamma = 115$, $p = 1.5734 \times 10^{-11}$	
Samples	Coverage	CI Width	Coverage	CI Width	Coverage	CI Width
100	0.9	0.4019	0.9	0.07700	0.9	0.024
500	0.93	0.0203	0.93	2.75×10^{-4}	0.93	1.20×10^{-5}
1000	0.97	0.0051	0.97	1.89×10^{-5}	0.97	4.25×10^{-7}
5000	0.97	2.33×10^{-4}	0.97	1.155×10^{-7}	0.97	9.36×10^{-10}

3.3.3 Using Nonparametric Bootstrap

Next we investigate the use of bootstrapping to assess input uncertainty. Such a technique has been studied in the simulation literature (e.g., [10]). In our experiment, we construct bootstrapped empirical distributions by repeatedly resampling with replacement and with full size from the data and using them to drive enough simulation runs per resample. The bootstrap size is $B = 100$. We use the empirical quantiles of the bootstrap estimates to construct a confidence interval for the estimate. We repeat our experiments 100 times. Here we again consider X_i with Gaussian distribution and generalized Pareto distribution with $\xi = 0.2$. We examine whether the confidence intervals constructed from the bootstrap scheme provide the target coverage (95% in our experiment).

Tables 3.1 and 3.2 suggest that the bootstrap works well in light-tailed problems, but fails in heavy-tail problems. This ties to our explanation in Section 3.2 that the impact from tail uncertainty is more profound in the heavy-tail case, which cannot be captured through the standard bootstrap. Table 3.1 shows that, for light-tail problems., the coverages of the confidence intervals are above 90% in all the considered cases (different numbers of samples and rarity levels). Note that when the number of samples is small (e.g. 100) the confidence interval width is relatively big compared to the estimated probability (0.4 to 0.024 when $p = 2.66 \times 10^{-5}$ to 1.57×10^{-11}). Though this could be pessimistic, these wide intervals successfully

Table 3.2: The coverage and the width of bootstrap confidence interval. The results are computed from 100 replications. The problem is defined by the sum of generalized Pareto variables. The true probability is 1.9195×10^{-7}. "# of 0 Width" presents the number of replications with 0 confidence interval width.

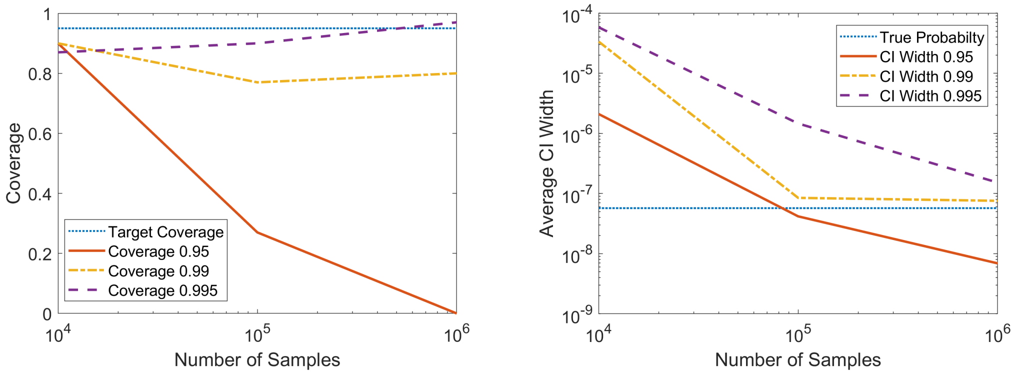
Sample Size	Coverage	CI Width (exclude 0)	# of 0 Width
10^4	0.02	1.15×10^{-5}	97
10^6	0.02	2.40×10^{-6}	80
10^7	0.04	2.00×10^{-7}	3

detect the unreliable probability estimate (see experiments in Section 3.3.2).

On the other hand, Table 3.2 shows that the coverage from the standard bootstrap are close to 0 in all considered cases, including the case of using 10^7 samples to simulate a rare-event probability of order 10^{-7}. Also, the last column shows that when the sample size is smaller than 10^6, most of the constructed confidence intervals have a 0 width. This suggests that in heavy-tail problems, the lack of tail information not only causes problems in estimating the probability itself, but can also deem the assessment of input uncertainty very challenging.

3.3.4 Bootstrap using Generalized Paerto Distribution

Lastly, we attempt to overcome the challenges in Section 3.3.3 by fitting a generalized Pareto distribution to the tail of the data. We then run the bootstrap similar to before taking into the fitted tail from each resample. We again consider different truncation points with 0.05, 0.01 and 0.005 tail quantile in our experiment. Here we use X_i with a t-distribution with degree of freedom $\nu = 4$. The considered numbers of samples (from 10^4 to 10^6) are not enough for standard bootstrap to work (see Table 3.2). Similar to the experiments in Section 3.3.3, each confidence interval is calculated from 100 bootstrap size, and the reported results are based on 30 experiments. To fit the generalized Pareto, we implement the maximum likelihood estimation (MLE), the method-of-moments (MOM) and probability-weighted moments (PWM) [41, 88].



(a)

(b)

Figure 3.5: The confidence interval coverage and width of the generalized Pareto tail bootstrap scheme (fitted using MLE) on the problem with $p = 5.7095 \times 10^{-8}$. (a) Coverage. (b) Average CI Width.

Table 3.3: The coverage and the width of confidence interval from bootstrap using generalized Pareto distribution. "# Spl" represents the number of samples and "Tail Qtl" represents the tail quantile of the truncation points. The problem has a rare event probability $p = 5.7095 \times 10^{-8}$.

	# Spl	10^4		10^5		10^6	
Tail Qtl	Method	Coverage	CI Width	Coverage	CI Width	Coverage	CI Width
0.05	MLE	0.9	2.10×10^{-6}	0.27	4.17×10^{-8}	0	6.89×10^{-9}
	MOM	0.67	1.23×10^{-6}	0.53	4.89×10^{-7}	0.30	3.19×10^{-7}
	PWM	0.87	1.81×10^{-6}	0.30	4.57×10^{-8}	0	7.04×10^{-9}
0.01	MLE	0.90	3.36×10^{-5}	0.77	8.46×10^{-7}	0.80	7.56×10^{-8}
	MOM	0.60	1.26×10^{-6}	0.70	8.10×10^{-7}	0.77	5.61×10^{-7}
	PWM	0.90	1.09×10^{-5}	0.77	8.13×10^{-7}	0.77	8.84×10^{-8}
0.005	MLE	0.87	5.81×10^{-5}	0.90	1.46×10^{-6}	0.97	1.42×10^{-7}
	MOM	0.63	1.17×10^{-6}	0.67	8.43×10^{-7}	0.87	6.10×10^{-7}
	PWM	0.87	1.25×10^{-5}	0.83	1.58×10^{-6}	0.97	1.70×10^{-7}

The experiment results show that although the overall performance of this Pareto tail bootstrap scheme is better than the standard bootstrap, the obtained confidence interval can still be misleading. The latter is caused by the model biasedness from the generalized Pareto in finite sample that the bootstrap cannot overcome. As shown in Figure 3.5, the coverage could drop to 0 as we increase the number of samples. This is because when the interval width shrinks as the number of samples increases, the model biasedness starts to surface.

Among the three approaches for fitting generalized Pareto distribution, MLE and PWM turn out to be more reliable than MOM (see Table 3.3, where the coverage of MOM is less than the other two approaches in most cases). The performance matches the documented fact that MOM is unreliable when the shape parameter $\xi > 0.2$. When the sample size is smaller (with 10^4 samples), PWM gives a smaller average confidence interval width than MLE, while providing similar coverage (e.g. with 0.01 tail quantile the widths are 3.36×10^{-5} for MLE and 1.09×10^{-5} for PWM). It therefore suggests that PWM is more suitable for smaller samples. When the sample size is large (10^6), MLE has an upper hand in terms of confidence interval width (e.g. with 0.005 tail quantile the widths are 1.42×10^{-7} for MLE and 1.70×10^{-7} for PWM).

CHAPTER IV

Importance Samplers for Rare Events in Prediction Models

4.1 Introduction

We consider the problem of estimating $P(g(X) > \gamma)$ via simulation, where $X \in \mathbb{R}^d$ is a random input and $\gamma \in \mathbb{R}$ is a high threshold, so that the probability is small and leads to a rare-event simulation problem. We are interested particularly in $g(\cdot)$ that is the response of a sophisticated predictor, built for instance from an off-the-shelf machine learning toolbox. This problem is motivated as a step towards building good learning methods that take into account the extremal risks of machine learning prediction.

To be more concrete, suppose we are interested in training the parameter θ of a predictor, say $Y_\theta(X)$, by minimizing a risk criterion $E[L(Y_\theta(X), Y(X))]$, where $Y(\cdot)$ is the true response function. Suppose that this loss function L exerts a big value in some "hidden region" where the occurrence of X is rare. The data in this case may not reveal satisfactorily this hidden risk. An approach to learn a good risk-conscious predictor in this situation is to fit the data using some probability distribution, and use it to train the predictor. A judicious choice of the distribution allows one to generate Monte Carlo samples, and to use importance sampling (IS) to populate more samples in the rare-event set. Of course, this approach requires various aspects of considerations, both statistically and practically. However, one step towards handling such type of problems involves how to speed up simulation for a machine learning model, which is precisely what we address in this chapter.

As a motivating example, autonomous vehicle (AV) design often uses predictions of the movements of surrounding cars to decide its own behavior (e.g. [94]). Predictions that seriously deviate from the reality can post substantial danger and catastrophic failure. However, unusual surrounding driving movements that lead to serious deviations of the predictions are rare. In this case, one can represent the driving environment as a stochastic model (the X), which provides inputs for the prediction model (the $Y_\theta(X)$) (e.g. [168, 91]), and a risk-sensitive predictive training could involve rare-event simulation on functionals associated with $Y_\theta(X)$.

4.2 Problem Setting

We state more precisely our problem setting. Suppose we are given a prediction model $g(\cdot)$, with the input $X \in \mathbb{R}^d$ and the output $g(X) \in \mathbb{R}$. Suppose that the input follows a standard Gaussian distribution, i.e, $X \sim N(0, I_d)$, where I_d is the $d \times d$ identity matrix. We want to estimate the probability $p = P(g(X) \geq \gamma)$, where $\gamma \in \mathbb{R}$ is a threshold that triggers a rare event. We note that the Gaussian assumption can be relaxed without much difficulty in our framework to, for instance, mixtures of Gaussians, which can expand our scope of applicability. For simplicity, however, we will concentrate on the standard Gaussian distribution in this chapter.

When p is small, estimation using crude Monte Carlo is challenging due to large variance. A common approach to speed up simulation in such contexts is to use IS (see, e.g. the surveys [30, 5, 149, 72, 97, 24], 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

$$Z = I(g(X) \geq \gamma)\frac{f(X)}{\tilde{f}(X)}, \tag{4.1}$$

where X is now sampled from $\tilde{f}$. This output is unbiased if f is absolutely continuous with respect to $\tilde{f}$ over the rare-event set $\{x : g(x) \geq \gamma\}$. Moreover, by choosing $\tilde{f}$ appropriately, one can substantially reduce the simulation variance.

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. Starting from a standard Gaussian distribution, a well-known IS scheme is as follows (see, e.g., [150]; [57]; [25]; [2]):

1. If there exists a dominating point a, i.e., a point a such that $a = \arg\min_x\{\|x\|^2 : g(x) \geq \gamma\}$ and $\{x : g(x) \geq \gamma\} \subseteq \{x : a'(x-a) \geq 0\}$, then we use a Gaussian distribution with mean at a as the IS distribution $\tilde{f}$.
2. If we can split $\{x : g(x) \geq \gamma\}$ into $\mathcal{R}_1, ..., \mathcal{R}_r$, and for each $\mathcal{R}_i, i = 1, ..., r$ there exists a dominating point a_i such that $a_i = \arg\min_x\{\|x\|^2 : x \in \mathcal{R}_i\}$ and $\mathcal{R}_i \subseteq \{x : a_i'(x-a_i) \geq 0\}$, then we use a Gaussian mixture distribution with r components as the IS distribution $\tilde{f}$, where the ith component has mean a_i.

Note that the quantity $\arg\min_x\{\|x\|^2 : g(x) \geq \gamma\}$ is equivalent to $\arg\max_x\{\phi(x) : g(x) \geq \gamma\}$, where ϕ is the standard Gaussian density. The condition $2a'(x-a) \geq 0$ is the first order condition of optimality for the optimization $\min_x \|x\|^2$ over a convex set for x, and so the first situation above occurs if $\{x : g(x) \geq \gamma\}$ is a convex set (though it can be more general). The proposals above guarantee the so-called asymptotic efficiency or logarithmic efficiency (e.g., [5]) of the IS, meaning that the relative error, measured by the ratio of standard deviation of a single IS output over the probability of interest, grows at most polynomially in the location of the rare-event set.

Given the proposal above, one can follow Algorithm 1 to obtain the dominating points $a_1, ..., a_r$ to build an efficient IS distribution in our prediction model context. 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 the existing dominating points. The set A in the procedure serves to store the dominating points we have found throughout the procedure. At the end of the procedure, one obtain a set A that contains all the dominating points $a_1, ..., a_r$.

Algorithm 1: Procedure to find all dominating points for the set $\{x : g(x) \geq \gamma\}$.

```
Input: Prediction model $g(x)$, threshold $\gamma$.
Output: Dominating-points set $A$.
Start with $A = \emptyset$;
While $\{x : g(x) \geq \gamma, a_i'(x - a_i) < 0,\ \forall a_i \in A\} \neq \emptyset$ do
    Find a dominating point $a$ by solving the optimization problem
    (4.2)    $a = \arg\min_x \ \|x\|^2$
             $s.t.\ \ g(x) \geq \gamma$
             $a_i'(x - a_i) < 0,\ \text{for } \forall a_i \in A$
    and update $A \leftarrow A \cup \{a\}$;
End
```

To apply Algorithm 1 to find all dominating points, the key is to be able to solve the optimization problems (4.2). We will investigate this for two popular prediction models, random forest and neural network (NN). Section 4.3 studies the tractable formulations using some recent work on optimization over these models, while Section 4.4 applies them to design IS schemes and demonstrate some numerical results.

4.3 Tractable Formulation For Prediction Models

We discuss how to formulate the optimization problems in Algorithm 1 as a mixed integer program (MIP) with quadratic objective function and linear constraints. Sections 4.3.1 and 4.3.2 focus on a basic NN and random forest respectively. Section 4.3.3 then further discusses solving larger-scale problems using a simple bisection method that leverages existing techniques in Benders decompositions in the random forest case.

4.3.1 Tractable Formulation for Neural Network

We consider a NN with L layers and for each layer, the number of neurons is $n_1, ..., n_L$. (As we focus on output $g(x) \in \mathbb{R}$, we have $n_L = 1$.) We use rectified linear unit (ReLU) for each neuron, i.e., the activation function for each neuron is $\max\{0, x\}$. The input of the jth neuron in layer i is weighted from the output of the previous layer by a vector $w_i^j \in \mathbb{R}^{n_{i-1}}$ and is added by a bias $b_i^j \in \mathbb{R}$. We use $s_i \in \mathbb{R}^{n_i}, i = 1, ..., L$ to represent the output of the ith layer. At the ith layer, given the output from the $(i-1)$th layer s_{i-1}, we have $s_i \in \mathbb{R}^{n_i}$, where the jth element of s_i is given by $s_i^j = \max\{{w_i^j}^T s_{i-1} + b_i^j, 0\}$. For further details on such type of NNs, see, e.g., [76].

Here we reformulate (4.2) by replacing $g(x) \geq \gamma$ with constraints that represent the structure of the NN:

$$\begin{aligned}
& s_L \geq \gamma \\
& s_i^j = \max\{{w_i^j}^T s_{i-1} + b_i^j, 0\}, \ i = 1, ..., L, \ j = 1, ..., n_i \\
& s_0 = x.
\end{aligned}$$

[158] discussed one approach to express ReLU as a mixed integer model, by introducing auxiliary binary variables z. For $y = max(x, 0)$, if we have $l \leq x \leq u$, where $l \leq 0$ and $u \geq 0$ are the smallest and largest possible values of x, then we set $z = I(x \geq 0)$. We can reformulate $y = max(x, 0)$ as a set of linear constraints: $y \leq x - l(1 - z); \ y \geq x; \ y \leq uz; \ y \geq 0; \ z \in \{0, 1\}$.

Using this approach, we rewrite problem (4.2) with $A = \emptyset$ as

$$\begin{aligned}
(4.3)\qquad \min_{x, s_0, ..., s_L, z_1 ..., z_L} \quad & \|x\|^2 \\
s.t. \quad & s_L \geq \gamma \\
& s_i \leq {W_i}^T s_{i-1} + b_i - l(1 - z_i),\ i = 1, ..., L \\
& s_i \geq {W_i}^T s_{i-1} + b_i,\ i = 1, ..., L \\
& s_i \leq u z_i,\ i = 1, ..., L \\
& s_i \geq 0,\ i = 1, ..., L \\
& z_i \in \{0, 1\}^{n_i},\ i = 1, ..., L \\
& s_0 = x,
\end{aligned}$$

where $W_i = [w_i^1, ..., w_i^{n_i}] \in \mathbb{R}^{n_{i-1} \times n_i}$, $b_i = [b_i^1, ..., b_i^{n_i}]^T \in \mathbb{R}^{n_i}$, and $s_i = [s_i^1, ..., s_i^{n_i}]^T \in \mathbb{R}^{n_i}$.

Note that this optimization has a quadratic objective and linear constraints. Similarly, we can formulate (4.2) by adding linear constraints $a_i'(x - a_i) < 0$, $\forall a_i \in A$ to (4.3), which arrives at the same optimization class. Medium-size instances of these problems can be handled by standard solvers.

4.3.2 Tractable Formulation for Random Forest

To look for dominating points in a random forest or tree ensemble, we follow the route in [124] that studies optimization over these models. We consider a random forest as follows. The input x has d dimensions. Suppose the model consists of T trees $f_1, ..., f_T$. In each tree f_t, we use $a_{i,j}$ to denote the jth unique split point for the ith dimension of the input x, such that $a_{i,1} < a_{i,2} < ... < a_{i,K_i}$, where K_i is the number of unique split points for the ith dimension of x.

Following the notations of [124], let **leaves**(t) be the set of leaves (terminal nodes) of tree t and $\mathbf{splits}(t)$ be the set of splits (non-terminal nodes) of tree t. In each split s, we let $\mathbf{left}(t)$ be the set of leaves that are accessible from the left branch (the query at s is true), and $\mathbf{right}(t)$ be the set of leaves that are accessible from the right branch (the query at s is false). For each node s, we use $\mathbf{V}(s) \in \{1, ..., d\}$ to denote the dimension that participate in the node and $\mathbf{C}(s) \in \{1, ..., K_{\mathbf{V}(s)}\}$ to denote the set of values of dimension i that participate in the split query of s ($\mathbf{C}(s) = \{j\}$ and $\mathbf{V}(s) = \{i\}$ indicates the query $x_i \leq a_{i,j}$).

We use λ_t to denote the weight of tree t ($\sum_{t=1}^{T} \lambda_t = 1$). For each $l \in \mathbf{leaves}(t)$, $p_{t,l}$ denotes the output for the lth leaf in tree t.

To formulate the random forest optimization as an MIP, we introduce binary decision variables $z_{i,j}$ and $y_{t,l}$. Firstly, we have

$$z_{i,j} = I(x_i \leq a_{i,j}),\ i = 1, ..., d,\ j = 1, ..., K_i. \tag{4.4}$$

We then 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. We use $\mathbf{z}, \mathbf{y}$ to represent the vectors of $z_{i,j}$ and $y_{t,l}$ respectively. For the input x, we assume that $x \in [-B, B]^d$ and $|a_{i,j}| \leq B$. Then (4.4) is represented by the following constraints

$$\begin{aligned} x_i &\leq a_{i,j} + 2(1 - z_{i,j})B \\ x_i &> a_{i,j} - 2z_{i,j}B. \end{aligned}$$

Now we formulate (4.2) with $A = \emptyset$ as the following MIP

$$
\begin{aligned}
(4.5) \qquad & \min_{x,\mathbf{y},\mathbf{z}} \ \|x\|^2 \\
& s.t. \ \sum_{t=1}^{T} \sum_{l \in \mathbf{leaves}(t)} \lambda_t p_{t,l} y_{t,l} \geq \gamma \\
& \sum_{l \in \mathbf{leaves}(t)} y_{t,l} = 1 \\
& \sum_{l \in \mathbf{left}(t)} y_{t,l} \leq \sum_{j \in \mathbf{C}(s)} z_{\mathbf{V}(s),j}, \ \forall t \in \{1, ..., T\}, \ s \in \mathbf{splits}(t) \\
& \sum_{l \in \mathbf{right}(t)} y_{t,l} \leq 1 - \sum_{j \in \mathbf{C}(s)} z_{\mathbf{V}(s),j}, \ \forall t \in \{1, ..., T\}, \ s \in \mathbf{splits}(t) \\
& z_{i,j} \leq z_{i,j+1}, \ \forall i \in \{1, ..., d\}, \ j \in \{1, ..., K_i - 1\} \\
& z_{i,j} \in \{0, 1\}, \ \forall i \in \{1, ..., d\}, \ j \in \{1, ..., K_i\} \\
& y_{t,l} \geq 0, \ \forall t \in \{1, ..., T\}, \ l \in \mathbf{leaves}(t) \\
& x_i \leq a_{i,j} + 2(1 - z_{i,j})B \\
& x_i > a_{i,j} - 2z_{i,j}B.
\end{aligned}
$$

This formulation again has a quadratic objective function and linear constraints. Similarly, we can formulate (4.2) with $A \neq \emptyset$ by adding linear constraints $a_i'(x - a_i) < 0, \ \forall a_i \in A$ to (4.5).

4.3.3 Bisection Algorithm and the Benders Decomposition for Solving Larger-scale Problems

The MIPs (4.5) are already tractable for small- and medium-size problems. Nonetheless, because of the special structure of (4.5), we can obtain, for larger-scale problems, an efficient algorithm based on a bisection on a "dual" form of the problem. The latter is a linear MIP and can be solved by the Benders decomposition considered in [124].

We consider the “dual” form problem of (4.2) with $A = \emptyset$:

$$h(\eta) = \max_{x} \;\; g(x) \qquad (4.6)$$
$$s.t. \;\; \|x\|^2 \leq \eta.$$

Note that (4.6) is non-decreasing in η. Moreover, we know that if η is above the optimal objective value for (4.2) with $A = \emptyset$, then (4.6) will output an objective value at least γ. On the other hand, if η is below that, then (4.6) will output a value less than γ. This motivates us to propose a bisection method that does a line search on the smallest value of η that gives $g(x) \geq \gamma$. This is summarized in Algorithm 2.

Algorithm 2: Bisection algorithm for solving (4.2) **with** $A = \emptyset$**.**

```
Input: Prediction model $g(x)$, threshold $\gamma$, tolerance parameters $\epsilon$.
Output: Optimal solution $a$.
Set $\eta_1 = 0$ and $\eta_2 = M$, where $M$ is large enough (e.g. $M = dB^2$);
Solve
  (4.7)            $\gamma_1 = h(\eta_1)$
  and
  (4.8)            $\gamma_2 = h(\eta_2)$;
While $\eta_2 - \eta_1 \geq \epsilon$ do
    Update $\eta_m = \frac{\eta_1+\eta_2}{2}$;
    Solve
  (4.9)            $\gamma_m = h(\eta_m)$;
    If $\gamma_m \geq \gamma$ do
        Update $\eta_2 \leftarrow \eta_m$;
    Else do
        Update $\eta_1 \leftarrow \eta_m$;
    End;
End;
Update $a$ using the optimal solution $x$ from (4.8) ;
```

Note that Algorithm 2 converges to the solution of (4.5) as ϵ goes to 0. Also, the number of iterations we need is given by $N_{iter} < \log \frac{M}{\epsilon} / \log 2$. Therefore, as long as we can solve (4.6) efficiently, the same is true for the dominating point problem. But, because the quadratic constraint in (4.6) only results in eliminating some of the dummy binary variables in the MIP formulation, we recover the formulation suggested in [124], but with less variables, to represent (4.6). This means that (4.6) is equivalent to

$$\begin{aligned}
(4.10)\qquad & \max_{x,\mathbf{y},\mathbf{z}} \sum_{t=1}^{T} \sum_{l\in\mathbf{leaves}(t)} \lambda_t p_{t,l} y_{t,l} \\
& s.t. \quad \sum_{l\in\mathbf{leaves}(t)} y_{t,l} = 1 \\
& \quad \sum_{l\in\mathbf{left}(t)} y_{t,l} \le \sum_{j\in\mathbf{C}(s)} z_{\mathbf{V}(s),j},\ \forall t\in\{1,...,T\},\ s\in\mathbf{splits}(t) \\
& \quad \sum_{l\in\mathbf{right}(t)} y_{t,l} \le 1 - \sum_{j\in\mathbf{C}(s)} z_{\mathbf{V}(s),j},\ \forall t\in\{1,...,T\},\ s\in\mathbf{splits}(t) \\
(4.11)\qquad & \quad \|x\|^2 \le \eta \\
(4.12)\qquad & \quad z_{i,j} \le z_{i,j+1},\ \forall i\in\{1,...,d\},\ j\in\{1,...,K_i-1\} \\
(4.13)\qquad & \quad z_{i,j} \in \{0,1\},\ \forall i\in\{1,...,d\},\ j\in\{1,...,K_i\} \\
(4.14)\qquad & \quad x_i \le a_{i,j} + 2(1-z_{i,j})B \\
(4.15)\qquad & \quad x_i > a_{i,j} - 2z_{i,j}B \\
& \quad y_{t,l} \ge 0,\ \forall t\in\{1,...,T\},\ l\in\mathbf{leaves}(t).
\end{aligned}$$

In (4.10), let us consider $x, \mathbf{z}$ as one set of variables and $\mathbf{y}$ as the other set of variables, where $\mathbf{y}$ can be further partitioned as $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_T)$ and $\mathbf{y}_T$ consists of $y_{t,l}$'s. We observe that for any two trees $t \neq t'$, $\mathbf{y}_t$ and $\mathbf{y}_{t'}$ does not appear together in any constraints and are only linked through $\mathbf{z}$. This observation allows us to use the Benders decomposition as [124] suggests.

We rewrite problem (4.10) as follows:

$$\begin{aligned}
(4.16)\qquad & \max_{x,\mathbf{z}} \sum_{t=1}^{T} \lambda_t G_t(\mathbf{z}) \\
& s.t.\ \text{ constraints (4.11)-(4.15)},
\end{aligned}$$

where $G_t(\mathbf{z})$ is the optimal value of the following subproblem:

(4.17)

$$\begin{aligned} G_t(\mathbf{z}) = \max_{\mathbf{y},\mathbf{z}} \quad & \sum_{l\in\mathbf{leaves}(t)} \lambda_t p_{t,l} y_{t,l} \\ s.t. \quad & \sum_{l\in\mathbf{leaves}(t)} y_{t,l} = 1 \\ & \sum_{l\in\mathbf{left}(t)} y_{t,l} \leq \sum_{j\in\mathbf{C}(s)} z_{\mathbf{V}(s),j}, \ \forall t\in\{1,...,T\}, \ s\in\mathbf{splits}(t) \\ & \sum_{l\in\mathbf{right}(t)} y_{t,l} \leq 1 - \sum_{j\in\mathbf{C}(s)} z_{\mathbf{V}(s),j}, \ \forall t\in\{1,...,T\}, \ s\in\mathbf{splits}(t) \\ & y_{t,l} \geq 0, \ \forall t\in\{1,...,T\}, \ l\in\mathbf{leaves}(t). \end{aligned}$$

The dual of the linear problem (4.17) is

$$\begin{aligned} \min_{\alpha_t,\beta_t,\zeta_t} \quad & \sum_{s\in\mathbf{splits}} \alpha_{t,s}\left[\sum_{j\in\mathbf{C}(s)} z_{\mathbf{V}(s),j}\right] + \sum_{s\in\mathbf{splits}} \beta_{t,s}\left[1-\sum_{j\in\mathbf{C}(s)} z_{\mathbf{V}(s),j}\right] + \zeta_t \\ s.t. \quad & \sum_{s:l\in\mathbf{left}(t)} \alpha_{t,s} + \sum_{s:l\in\mathbf{right}(t)} \beta_{t,s} + \zeta_t \geq p_{t,l}, \ \forall l\in\mathbf{leaves}(t) \\ & \alpha_{t,s}, \beta_{t,s} \geq 0, \ s\in\mathbf{splits}. \end{aligned} \tag{4.18}$$

We use $\mathcal{D}_t$ to denote the set of feasible $(\alpha_t, \beta_t, \zeta_t)$ for subproblem (4.18) of tree t. We write (4.16) in terms of the dual variables $(\alpha_t, \beta_t, \zeta_t)$:

$$\max_{x,\mathbf{z},\theta} \quad \sum_{t=1}^{T} \lambda_t\theta_t \tag{4.19}$$

$$\begin{aligned} s.t. \quad & \sum_{s\in\mathbf{splits}} \alpha_{t,s}\left[\sum_{j\in\mathbf{C}(s)} z_{\mathbf{V}(s),j}\right] + \sum_{s\in\mathbf{splits}} \beta_{t,s}\left[1-\sum_{j\in\mathbf{C}(s)} z_{\mathbf{V}(s),j}\right] + \zeta_t \\ & \geq \theta_t, \ \forall(\alpha_t,\beta_t,\zeta_t)\in\mathcal{D}_t, t\in\{1,...,T\} \end{aligned} \tag{4.20}$$

constraints (4.11)-(4.15).

With problem (4.19), we can use a constraint generating scheme. We start with solving problem (4.19) using a subset of $\bar{\mathcal{D}}_t \subseteq \mathcal{D}_t$ in the constraint (4.20). We then solve (4.18) for each tree t to check if there exists a solution $(\alpha_t, \beta_t, \zeta_t)$ for which the constraint (4.20) is violated. If so, we add the constraint to (4.19) and solve it again; otherwise, we obtain the optimal solution. For a quick method to solve problem (4.18), please refer to Proposition 5 in [124].

The decomposition of (4.2) with non-empty A follows the same route as above and the resulting formulation is very similar. The dual of the subproblem is exactly the same as (4.18), while the main problem is (4.19) with a set of additional constraints $a_i'(x - a_i) < 0, \ \forall a_i \in A$.

4.4 Applications to Importance Sampling and Numerical Experiments

In this section, we use several sets of simple experiments to illustrate the IS scheme using dominating points generated from Algorithm 1, for the described NN and random forest. In the first set of problems, we consider an example where there is only one dominating point. In the second set of problems, we solve a problem with multiple dominating points.

To illustrate the efficiency of the proposed IS scheme, we compare with a naive use of a uniform IS estimator as follows. 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 is given by:

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

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 grow, but the efficiency also depends significantly on the size of the bounded set, i.e., l, u, d.

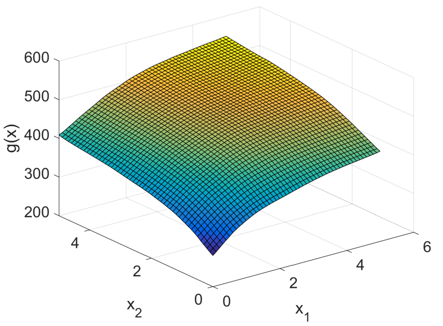


Figure 4.1: Response surface of the neural network.

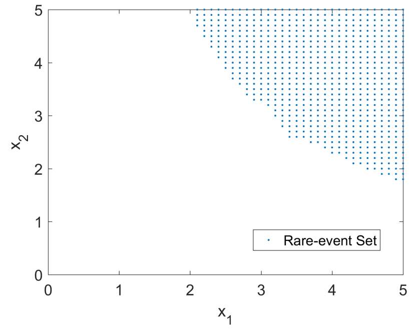


Figure 4.2: Rare-event set $\{x : g(x) \geq \gamma\}$ of the neural network.

4.4.1 Neural Network Example: IS with A Single Dominating Point

We recall the problem setting that the input X follows a standard Gaussian distribution, and we are interested in estimating the probability $P(g(X) \geq \gamma)$. Here $g(x)$ represents the output of a NN prediction at x and γ is a real-valued threshold. The NN has 3 layers with 100 neurons in each of the 2 hidden layers, and all neurons are ReLU. We consider only X in the region $[0, 5]^2$, so that $g(x)$ can be thought of as being set to 0 outside this box.

The NN is trained as follows. We generate 2,601 samples using a uniform grid over the space $[0, 5]^2$ with a mesh of 0.1 on each coordinate. For the input $x = [x_1, x_2]$, we use the function

$$y(x) = (x_1 - 5)^3 + (x_2 - 4.5)^3 + (x_1 - 1)^2 + x_2^2 + 500 \tag{4.21}$$

to generate output value as the “truth” . We obtain the dataset $D = \{(X_n, Y_n)\}$ and use it to train the NN with gradient descent. We present the response surface of $g(x)$ in Figure 4.1. We use $\gamma = 500$ in this example and the shape of the rare-event set $\{x : g(x) \geq \gamma\}$ in this case is presented in Figure 4.2. We observe that the set is roughly convex and should have a single dominating point. By solving (4.3), we obtain the dominating point for the set at $(3.3676, 2.6051)$.

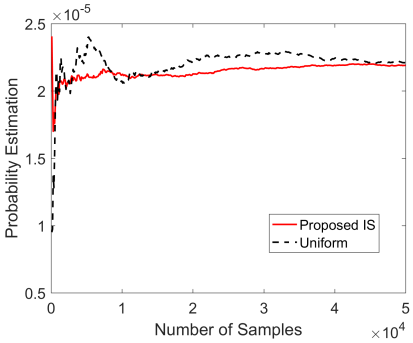


Figure 4.3: Probability estimation with different numbers of samples.

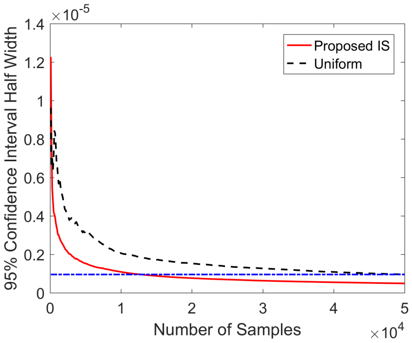


Figure 4.4: 95% confidence interval half-width with different numbers of samples.

We use the obtained dominating point to construct the IS estimator described in (4.1). In Figures 4.3 and 4.4, we compare the performance of the proposed IS scheme with that of the uniform IS estimator. Figure 4.3 shows the estimated probabilities of the two estimators as the number of samples increases in a single sample path. We observe that the uniform IS (black dash line) has more fluctuations than the proposed IS (red solid line), which indicates that the proposed IS gives more stable estimates. This observation is confirmed in Figure 4.4 that shows the half-width of the 95% confidence intervals of the two estimators as the number of samples varies. Our IS appears to have shorter confidence intervals and it only takes about 12,000 samples for our IS to reach the accuracy that the uniform IS reaches with 50,000 samples.

4.4.2 Neural Network Example: IS with Multiple Dominating Points

We now consider true output values generated according to the function

$$y(x) = 10 \times e^{-\left(\frac{x_1-5}{3}\right)^2-\left(\frac{x_2-5}{4}\right)^2} + 10 \times e^{-{x_1}^2-(x_2-4.5)^2}. \tag{4.22}$$

We use a uniform grid over $[0,5]^2$ with a mesh of 0.1 on each coordinate to train a 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.

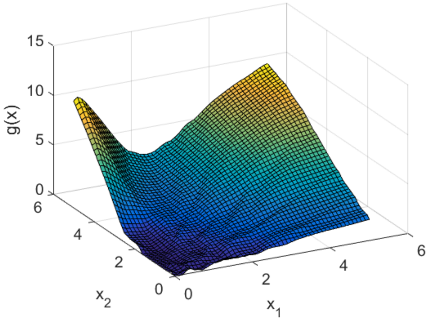


Figure 4.5: Response surface of the neural network.

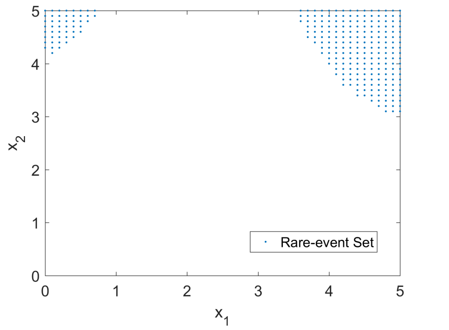


Figure 4.6: Rare-event set $\{x : g(x) \geq \gamma\}$ of the neural network.

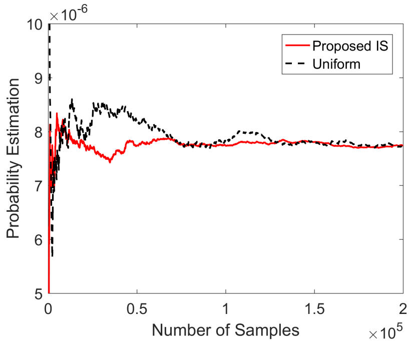


Figure 4.7: Probability estimation with different numbers of samples.

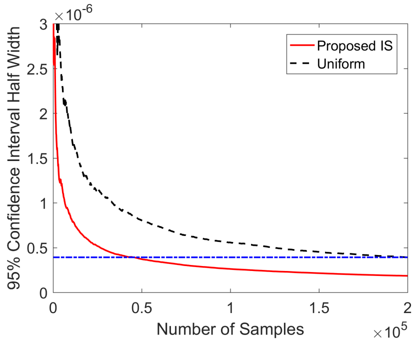


Figure 4.8: 95% confidence interval half-width with different numbers of samples.

The response surface of the trained model $g(x)$ is shown in Figure 4.5. We set $\gamma = 8$. The shape of the rare-event set is shown in Figure 4.6. We observe that the set now consists of two separate regions and therefore we expect to obtain multiple dominating points. Using Algorithm 1 with the formulation in Section 4.3.1, we obtain two dominating points, $(0.113, 4.162)$ and $(4.187, 3.587)$. We use these dominating points to construct a mixture distribution, as discussed in Section 4.2, as the IS distribution.

The comparison between the proposed IS scheme and the uniform IS is shown in Figures 4.7 and 4.8. In Figure 4.7, the probability estimates of the proposed IS (red solid line) appear more stable than that of the uniform IS (black dash line). Figure 4.8 further shows that the confidence intervals of the proposed IS are shorter. The

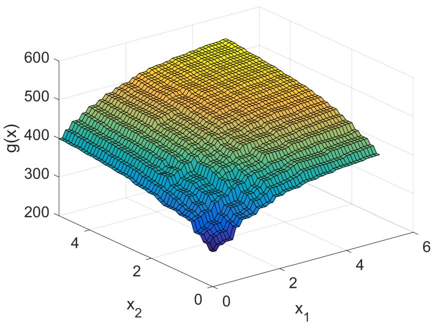


Figure 4.9: Response surface of the random forest.

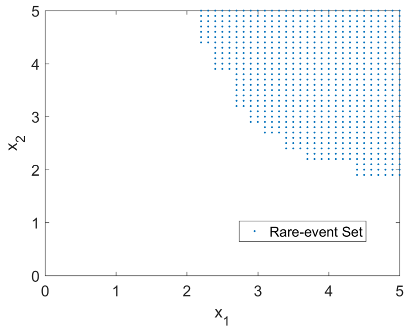


Figure 4.10: Rare-event set $\{x : g(x) \geq \gamma\}$ of the random forest.

efficiency of the proposed IS is roughly 4 times better than the uniform IS, considering that the confidence interval half-width of the proposed IS at 50,000 samples is similar to the uniform IS at 200,000 samples.

Thus, comparing with uniform IS, the estimation accuracy of the proposed IS is better in both experiments, demonstrating the efficiency of the IS scheme obtained in the formulation in Section 4.3.1.

4.4.3 Random Forest Example: IS with A Single Dominating Point

Consider now that $g(x)$ represents a random forest. The random forest is trained from samples generated uniformly over the space $[0, 5]^2$ with a mesh 0.25 on each coordinate, with output values from (4.21). The trained random forest model consists of 3 trees with 563, 535, 565 nodes respectively. The response surface of the model $g(x)$ is shown in Figure 4.9. Here we use $\gamma = 500$ and the shape of the rare-event set $\{x : g(x) \geq \gamma\}$ in presented in Figure 4.10. We use Algorithm 2 to obtain the dominating point of the rare-event set. In the algorithm, we choose the tolerance parameters to be $\epsilon = 0.01$. This gives us the dominating point for the set at $(3.00001, 2.62501)$.

We use the proposed IS and the uniform IS to estimate the probability $P(g(x) \geq \gamma)$. In Figure 4.11, we observe that the estimates from the uniform IS (black dash

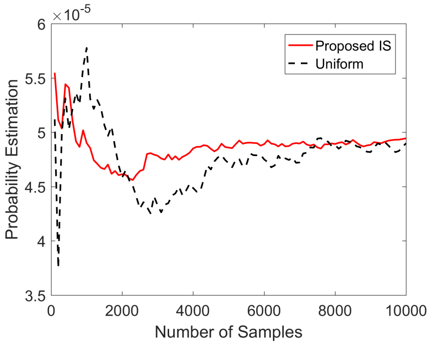


Figure 4.11: Probability estimation with different numbers of samples.

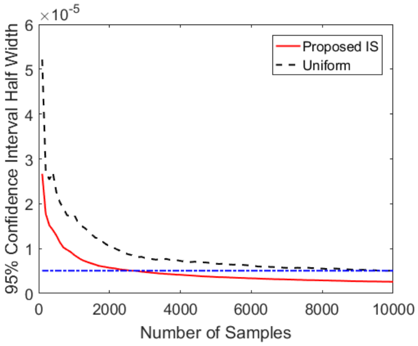


Figure 4.12: 95% confidence interval half-width with different numbers of samples.

line) are relatively unstable, compared to those from the proposed IS (red solid line). In Figure 4.12, the confidence intervals of the uniform IS (black dash line) appear wider than the proposed IS. These comparisons are similar to the neural network case.

4.4.4 Random Forest Example: IS with Multiple Dominating Points

We now generate input samples uniformly over $[0, 5]^2$ with a mesh 0.2 on each coordinate, with the output values drawn from (4.22). The random forest has 2 trees in this example. The trained model using the generated dataset has 865 nodes in the first tree and 835 nodes in the second tree. Figure 4.13 shows the response surface of the random forest. We consider $\gamma = 8$ and the shape of the rare-event set is shown in Figure 4.14, which shows that the rare event set consists of two separate regions.

We find dominating points by implementing Algorithm 1 combined with Algorithm 2. Again, we use $\epsilon = 0.01$ and $\delta = 0.01$ for the tolerance level. Note that the tolerance level we choose could generate an error on the dominating point we obtain. For $\epsilon = 0.01$, the error is given by $\|a\|^2 - \|\hat{a}\|^2 \leq 0.02$, where a is the true dominating point and $\hat{a}$ is the dominating point we obtain from the bisection. This error might bring an issue to Algorithm 1. That is, the constraint $\hat{a}(x - \hat{a}) < 0$ we add may not

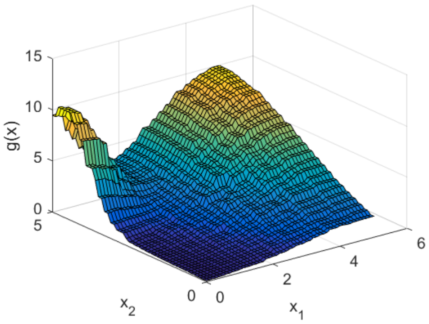


Figure 4.13: Response surface of the random forest.

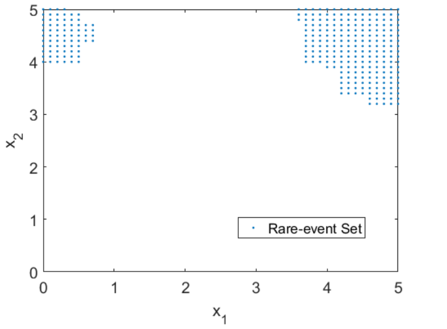


Figure 4.14: Rare-event set $\{x : g(x) \geq \gamma\}$ of the random forest.

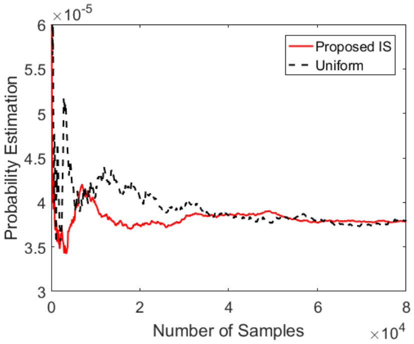


Figure 4.15: Probability estimation with different numbers of samples.

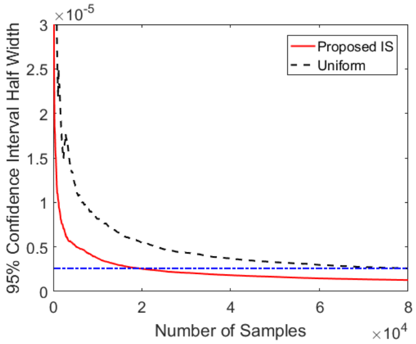


Figure 4.16: 95% confidence interval half-width with different numbers of samples.

cut off the corresponding half-space. This can possibly lead to an infinite loop for Algorithm 1. Here we resolve this issue by modifying the additional constraint as $\hat{a}(x - \hat{a}) < -\delta$, where $\delta > 0$ is approximately in the scale of $\sqrt{\epsilon}$. In this case, we use $\delta = 0.1$. We obtain two dominating points, $(0.1, 3.9)$ and $(4.1, 3.3)$.

The obtained dominating points allow us to implement the proposed IS. Compared to the uniform IS, Figures 4.15 and 4.16 show that the proposed approach is more efficient as in the previous experiments. The estimates of the proposed approach are more stable (red solid line in Figure 4.15) and the confidence intervals are shorter (red solid line in Figure 4.16). The uniform IS (black dash line in Figure 4.16) requires roughly 3 times more samples to achieve the same level of accuracy as the proposed IS.

CHAPTER V

Accelerating Autonomous Vehicle Tests using Importance Sampling and Piecewise Mixture Models

5.1 Introduction

It is critical to thoroughly and rigorously test and evaluate an Automated Vehicle (AV) before its release. Recent crashes involving a Google self-driving car [77] and a Tesla Autopilot vehicle [1] attracted the public's attention to AV testing and evaluation. While these AVs are generally considered as industrial leaders, because they use public roads for testing, statistically they have not yet accumulated enough miles. The Tesla Autopilot, in particular, was criticized for being released too early in the hands of the general public [69].

Currently, there are no standards or protocols to test AVs at automation level 2 or higher. Many companies adopt the Naturalistic Field Operational Tests (N-FOT) approach [71]. However, this method is inefficient because safety critical scenarios rarely happen in daily driving. The Google Self-driving cars accumulated 1.9 million driving. This distance, although sounds a lot, provides limited exposure to critical events, given that U.S. drivers encounter a police reported crash every five hundred thousand miles on average and fatal crash every one hundred million miles [133]. In the meantime, both Google and Tesla update their software throughout the process, which may have improved safety, but the newest version of the AV has not accumulated that many miles as they have claimed. In summary, today's best practice adopted by the industry is time-consuming and inefficient. A better approach is needed.

Besides the N-FOT, the test matrix approach [136, 7] and the worst-case scenarios approach [98, 161, 118] are two alternative methods for vehicle evaluation. These alternative methods also face some challenges for AV evaluation. The test matrix approach uses fixed and predefined test scenarios, which allows AVs to be tuned to perform well in these tests [136]. Moreover, it is not clear how to correlate the test results with real-world conditions [7]. The worst-case evaluation can identify the weakness of a vehicle control system, but it does not provide sufficient information about the risk of the vehicle system.

Our approach follows the Accelerated Evaluation concept we proposed [173] to provide a brand-new alternative that can handle these challenges. The basic concept is that as high-level AVs just began to penetrate the market, they mainly interact with human-controlled vehicles (HVs). Therefore we focus on modeling the interaction between the AV and the HV around it. The evaluation procedure involves four steps:

- Model the behaviors of the "primary other vehicles" (POVs) represented by $f(x)$ (original distribution) as the major disturbance to the AV using large-scale naturalistic driving data.
- Skew the disturbance statistics from $f(x)$ to modified statistics $f^*(x)$ (accelerated distribution) to generate more frequent and intense interactions between AVs and POVs.
- Conduct "accelerated tests" with $f^*(x)$.
- Use the Importance Sampling (IS) theory to "skew back" the results to under-

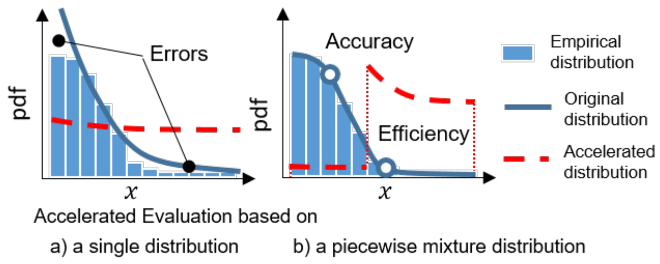


Figure 5.1: Acceleration evaluation based on single parametric distribution and Piecewise Mixture Distribution.

stand real-world behavior and safety benefits.

This approach has been successfully applied to evaluate AVs in the frontal crash with a cut-in vehicle [173] and also frontal crash with a lead vehicle [174, 172]. This approach was confirmed to significantly reduce the evaluation time while accurately preserving the statistical behavior of the AV-HV interaction. In the previous studies, the evaluation time was reduced by two to five orders of magnitudes - the accelerated rate depends on the test scenarios, where rarer events achieve higher accelerated rate. The non-accelerated models and the accelerated models were built based on signal component distributions. While this method does benefit from its simple mathematical form, it has a few drawbacks as illustrated in Fig. 5.1 a). i) Lack of accuracy, i.e. the fitting of the rare events (usually the tail part of the statistical distributions) would be dominated by the fitting of the normal driving behaviors (the majority part of the distributions), which may induce large errors. ii) Lack of efficiency, i.e. the full potential in higher accelerated rate is not achieved due to the lack of flexibility of the modified accelerated models.

In this chapter, we proposed a more general framework for the Accelerated Evaluation method to overcome the aforementioned limitations based on Piecewise Mixture Distribution Models as illustrated in Fig. 5.1 b). The piecewise model is a more flexible structure that can better captures the tail part of the data (more accurate) and provides better efficiency for accelerating the evaluation. In this chapter, we implemented the Accelerated Evaluation method on the lane change scenario to illustrate the benefits of using the proposed framework. In this chapter, we thoroughly discuss the model fitting and Cross Entropy method with proposed framework and present practical tips to overcome numerical issues and reduce computational efforts. We demonstrate this method by evaluating the longitudinal control system reacting to vehicles making cut-in lane changes. Some preliminary work are present in a conference version [93].

Section 5.2 will introduce the lane change model based on single parametric distributions. In Section 5.3, we present the new lane change model with Piecewise Mixture Distributions. We establish the Accelerated Evaluation in Section 5.4 and discuss the Cross Entropy method with Piecewise Mixture Distribution models in Section 5.5. Simulation results are discussed in Section 5.6. Section 5.7 concludes this chapter.

5.2 Accelerated Evaluation with Single Parametric Distributions

The lane change events were extracted from the Safety Pilot Model Deployment (SPMD) database [21]. With over 2 million miles of vehicle driving data collected from 98 cars over 3 years, we identify 403,581 lane change events. As shown in Fig. 5.2, the lane change events are detected by the SPMD vehicles and parameters in the lane changes are collected. Previously [173], we used 173,692 events with a negative range rate to build a statistical model focusing on three key variables that captured the effects of gap acceptance of the lane changing vehicle: velocity of the lead vehicle (v), range to the lead vehicle (R) and time to collision (TTC). TTC was defined as:

$$TTC = -\frac{R}{\dot{R}}, \tag{5.1}$$

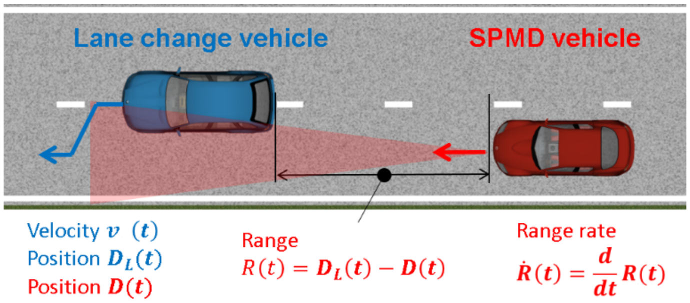


Figure 5.2: Lane change data collected by SPMD vehicle.

where $\dot{R}$ is the relative speed.

The modeling of these three variables was hard to handle because of dependency, so we simplified it based on a crucial observation. Although TTC is dependent on v generally, we split the data into 3 segments: v at 5 to 15 m/s, 15 to 25 m/s and 25 to 35 m/s. Within each segment, R is independent with v and TTC. This allowed us to model TTC and R independently with regard to the value of v. By comparing among 17 types of commonly used distribution templates [173], we selected the Pareto distribution to model R^{-1} and used the exponential distribution for TTC^{-1} segments.

Using the empirical distribution of v and parametric distributions of R and TTC, we drew values from these distributions as inputs to simulate the AV-HV interaction. We used an AV model designed from existing vehicle system [173] in the simulation. The simulation outputs whether a type of critical event (for example, crash or injury) happens. We use an event indicator function $I_\varepsilon(x)$ that returns $\{1, 0\}$ to represent the simulation procedure with input x, where ε stands for the set of the critical event of interest. Given the stochastic distribution of the variables and the event indicator function, we obtained the optimal exponential distribution for Importance Sampling by implementing the Cross Entropy method [148]. As we have shown in Fig. 5.1 a), we used only single parametric distributions. In the next section, we introduce our new approach using Piecewise Mixture Distributions.

5.3 Lane Change Model with Piecewise Mixture Distributions

Although many commonly used parametric distributions have concise and elegant forms, they do not always describe the data distribution well. Instead, a better fitting can be achieved by dividing the dataset into several subsets. We estimate the model parameters using the Maximum Likelihood Estimation (MLE) [3] in each subset. The general process of MLE is as follow.

Assume we have a family of distribution with Cumulative Distribution Function (CDF) $F(x|\theta)$, where θ is the parameter vector of F. The corresponding Probability Density Function (PDF) of F is $f(x|\theta)$. Assuming that data $D = \{X_1, X_2, ..., X_N\}$ is independently and identically distributed and the distribution is in the family of $F(x|\theta)$, we want to find the most "likely" parameter $\hat{\theta}$.

We define the likelihood function [49] as

$$L(\theta|D) = P(D|\theta) = \Pi_{n=1}^{N} f(X_n|\theta). \tag{5.2}$$

We call the estimation of $\hat{\theta}$ that maximizes the likelihood function the mostly likely estimation MLE.

For computation convenience, we introduce the log-likelihood function

$$\mathcal{L}(\theta|D) = \ln L(\theta|D) = \sum_{n=1}^{N} \ln f(X_n|\theta). \tag{5.3}$$

Since the logarithm is monotone, the log-likelihood function preserves the optimizer of the original function. [27] The optimizer of log-likelihood function, $\hat{\theta}$, is the MLE of distribution family F. We have the MLE as

$$\hat{\theta} = \arg\max_{\theta} \ \mathcal{L}(\theta|D). \tag{5.4}$$

In the following, we describe the Piecewise Mixture Distribution fitting concept based on MLE and we present the bounded distribution fitting results. All optimization problems presented in this section are tractable and can be solved by **fminunc** in MATLAB.

5.3.1 General Framework of the Piecewise Mixture Distribution Lane Change Model

We define Piecewise Mixture Distribution to be distribution with PDF in the form of

$$f(x) = \sum_{i=1}^{k} \pi_i f_{\theta_i}(x|\gamma_{i-1} \leq x < \gamma_i). \tag{5.5}$$

where k is the number of truncation, $\sum_{i=1}^{k} \pi_i = 1$, and $f_i(x|\gamma_{i-1} \leq x < \gamma_i)$ is the conditional density distribution function, meaning that $f_i(x|\gamma_{i-1} \leq x < \gamma_i) = 0$ for any $x \notin \{x|\gamma_{i-1} \leq x < \gamma_i\}$. θ_i denotes the parameter(s) for f_i. We can consider that $\pi_i = P(\gamma_{i-1} \leq x < \gamma_i)$ and when $x \geq 0$, we have $\gamma_0 = 0$ and $\gamma_k = \infty$.

In our case, $\theta = \{\pi_1, ..., \pi_k, \theta_1, ..., \theta_k\}$. Splitting D into pieces regarding the truncation points $\{\gamma_1, ..., \gamma_{k-1}\}$, gives data index sets $S_i = \{j|\gamma_{i-1} \leq X_j < \gamma_i\}$ for $i = 1, ..., k$. We can write the log-likelihood function as

$$\mathcal{L}(\theta|D) = \textstyle\sum_{i=1}^{k} \sum_{n \in S_i} \ln \pi_i + \sum_{i=1}^{k} \sum_{n \in S_i} \ln f_{\theta_i}(X_n|\gamma_{i-1} \leq x < \gamma_i). \tag{5.6}$$

We obtain the MLE of θ can be obtained by maximizing $\mathcal{L}(\theta|D)$ over θ. Since $\mathcal{L}$ is concave over π_i, we take

$$\frac{\partial \mathcal{L}}{\partial \pi_i} = 0 \tag{5.7}$$

and get

$$\hat{\pi}_i = |S_i|/N. \tag{5.8}$$

Note that for parameters θ_i in F_i, it is known (5.6) to be the same as computing the MLE of θ_i with corresponding dataset $D_i = \{X|\gamma_{i-1} \leq X < \gamma_i \; \textit{and} \; X \in D\}$.

Since we use bounded distribution for each F_i, below we explain the estimation of parameters for the three distributions we applied in later sections.

To sample from a Piecewise Mixture Distribution, we could use the inverse function approach.

5.3.2 Bounded Distribution

We develop three bounded distributions and use them in the lane change model. One can use criterion for goodness of fitting, e.g. Bayesian information criterion (BIC) [70], to select the distribution for fitting.

MLE for bounded exponential distribution

The bounded exponential distribution with rate θ has the form

$$f(x|\gamma_1 \leq x < \gamma_2) = \frac{\theta e^{-\theta x}}{e^{-\theta\gamma_1} - e^{-\theta\gamma_2}} \tag{5.9}$$

for $\gamma_1 \leq x < \gamma_2$.

For dataset $D = \{X_1, ..., X_N\}$, the log-likelihood function is

$$\mathcal{L}(D|\theta) = \sum_{n=1}^{N} \ln\theta - \theta X_n - \ln(e^{-\theta\gamma_1} - e^{-\theta\gamma_2}), \tag{5.10}$$

where $\mathcal{L}$ is concave over θ. Although we cannot solve the maximization analytically, it is solvable through numerical methods.

Therefore, the MLE of θ is given by the optimization

$$\max_{\theta} \ N\ln\theta - N\ln(e^{-\theta\gamma_1} - e^{-\theta\gamma_2}) - \sum_{n=1}^{N} \theta X_n. \tag{5.11}$$

MLE for bounded normal distribution

Consider a bounded normal distribution with mean 0 and variance θ^2 conditional on $0 \leq \gamma_1 \leq x < \gamma_2$. The PDF is

$$f(x|\gamma_1 \leq x < \gamma_2) = \frac{\frac{1}{\theta}\phi(\frac{x}{\theta})}{\Phi(\frac{\gamma_2}{\theta}) - \Phi(\frac{\gamma_1}{\theta})}. \tag{5.12}$$

The MLE of the bounded normal distribution is given by

$$\max_{\theta} -\frac{\sum_{n=1}^{N} X_n^2}{2\theta^2} - N\ln\theta - N\ln(\Phi(\frac{\gamma_2}{\theta}) - \Phi(\frac{\gamma_1}{\theta})). \tag{5.13}$$

Fitting algorithm for bounded mixture distribution

Compared to single parametric distributions, mixture distribution combines several classes of distribution and thus is more flexible. We consider the fitting problem of mixture bounded normal distribution.

The PDF of mixture of m bounded normal distribution can be written as

$$f(x|\gamma_1 \le x < \gamma_2) = \sum_{j=1}^{m} p_j f_j(x|\gamma_1 \le x < \gamma_2) \tag{5.14}$$

where f_j is bounded Gaussian distribution with mean 0 and variance σ_j^2. The parameters here are $\theta = \{p_1, ..., p_m, \sigma_1^2, ..., \sigma_m^2\}$. We want to find MLE of p_j and σ_j^2 for $j = 1, ..., m$.

The log-likelihood function for data $D = \{X_n\}_{n=1}^N$ is

$$\mathcal{L}(\theta|D) = \sum_{n=1}^{N} \ln \sum_{j=1}^{m} p_j f_j(X_n|\gamma_1 \le x < \gamma_2). \tag{5.15}$$

We note that this is hard to solve directly, because there is a sum within the log function. Therefore, we apply the Expectation-Maximization (EM) [53] algorithm to find the optimizer, i.e. MLE, for the parameters.

We define Z_n^j to denote whether or not the random number X_n comes from mixture distribution j, $j = 1, ..., m$, and $Z_n^j = \{0, 1\}$. We also introduce the expectation

$$E[Z_n^j|X_n] := \tau_n^j. \tag{5.16}$$

The EM algorithm starts with initial parameters $\{p_j, \sigma_j\}$, $j = 1, ..., m$. For data $D = \{X_n\}_{n=1}^N$, we set complete data as $D_c = \{X_n, Z_n\}_{n=1}^N$. The EM algorithm optimizes $E[\mathcal{L}(\theta|D_c)|D]$ in every step. The E step updates $E[\mathcal{L}(\theta|D_c)|D]$, and the M step optimizes this function. The algorithm iterates E step and M step until reaching the convergence criterion.

In our case,

$$E[\mathcal{L}(\theta|D_c)|D] = \textstyle\sum_{n=1}^{N} \sum_{j=1}^{m} \tau_n^j \left(\ln p_j + \ln f_j(X_n)\right). \tag{5.17}$$

Since objective $E[l_c(\theta|D_c)|D]$ in the M step is concave over p_j and σ_j, we could maximize the objective function through an analytic approach for p_j:

$$p_j = \frac{\sum_{n=1}^{N} \tau_n^j}{N}. \tag{5.18}$$

For σ_j, we can solve the following maximization problem through numerical approach.

$$\sigma_j = \arg\min_{\sigma_j} -\tau_n^j \ln \sigma_j + \tau_n^j \ln \phi\left(\frac{X_n}{\sigma_j}\right) - \tau_n^j \ln\left(\Phi(\frac{\gamma_2}{\sigma_j}) - \Phi(\frac{\gamma_1}{\sigma_j})\right). \tag{5.19}$$

5.3.3 Selection of Truncation Points

The framework we show in this section is based on the truncation points $\gamma_0, ..., \gamma_k$ are given. Here we discuss the selection of the truncation number k and the value these points.

The motivation of using Piecewise Mixture Distribution is to improve the fitting on the tail of the variables, because the tail fitting is crucial to the probability estimation of the event of interest. A basic criterion is that the tail truncation should not exceed the value that is "likely" to lead to an event of interest. Such value of each variable is roughly known in the AV testing scenarios. This allows us to assign the value of the tail truncation point. In the cases where such information is not available, one can use the mean excess plot to determine the tail truncation point and the tail distribution [122].

The body part of the variable is not as important, so we select the truncation points from data observation. Note that if we use the same distribution family for each piece of the distribution, adding a truncation point would always leads to a better fitting in the sense of likelihood. Here, we suggest to use as less truncation as possible to avoid over-fitting problem. One can use criterion for goodness of fitting to determine the number of selection.

5.4 Accelerated Evaluation with Importance Sampling

Importance Sampling (IS) is thus used to accelerate the evaluation process, because crude Monte Carlo simulations for rare events can be time-consuming. Here we describe the IS theory, which guarantees the unbiasedness of the probability estimation after the skewing-and-skewing-back procedure in the accelerated evaluation and provides the baseline for searching an efficient accelerated distribution.

5.4.1 Important Sampling and Optimal IS distribution

Let x be a random variable generated from distribution F, and $\varepsilon \subset \Omega$ where ε is the rare event of interest and Ω is the sample space. Our objective is to estimate

$$P(X \in \varepsilon) = E[I_\varepsilon(X)] = \int I_\varepsilon(x) dF \tag{5.20}$$

where

$$I_\varepsilon(x) = \begin{cases} 1 & x \in \varepsilon, \\ 0 & otherwise. \end{cases} \tag{5.21}$$

We can write the evaluation of rare events as the sample mean of $I_\varepsilon(x)$

$$\hat{P}(X \in \varepsilon) = \frac{1}{N} \sum_{n=1}^{N} I_\varepsilon(X_n), \tag{5.22}$$

where X_i's are drawn from distribution F.

Since we have

$$E[I_\varepsilon(X)] = \int I_\varepsilon(x)dF = \int I_\varepsilon(x)\frac{dF}{dF^*}dF^*, \tag{5.23}$$

we can compute the sample mean of $I_\varepsilon(X)\frac{dF}{dF^*}$ over the distribution F^*, which has the same support with F, to obtain an unbiased estimation of $P(X \in \varepsilon)$. By appropriately selecting F^*, the evaluation procedure obtains an estimation with smaller variance. This is known as Importance Sampling [29] and F^* is the IS distribution.

For estimating $P(X \in \varepsilon)$,we note that an optimal IS distribution

$$F^{**}(x) = F(x|\varepsilon) = \frac{P(X \leq x,\ \varepsilon)}{P(x \in \varepsilon)} \tag{5.24}$$

could reduce the variance of IS estimation to 0, but the optimal requires the knowledge of $P(X \in \varepsilon)$. However, it guides the selection of the IS distribution.

5.4.2 Exponential Change of Measure

Exponential change of measure is commonly used to construct F^*. Although the exponential change of measure cannot guarantee convergence to optimal distribution, it is easy to implement and the new distribution generally stays within the same class of distribution.

Exponential change of measure distribution takes the form of

$$f_\theta(x) = \exp(\theta x - \kappa(\theta))f(x), \tag{5.25}$$

where θ is the change of measure parameter and $\kappa(\theta)$ is the log-moment generating function of original distribution f. When $\theta = 0$, we have $f_\theta(x) = f(x)$.

For a bounded exponential distribution, the exponential change of measure distribution is

$$f_\theta(x|\gamma_1 \leq x < \gamma_2) = \frac{(\lambda - \theta)e^{-(\lambda-\theta)x}}{e^{-(\lambda-\theta)\gamma_1} - e^{-(\lambda-\theta)\gamma_2}}, \tag{5.26}$$

where λ is the parameter for exponential distribution. We note that f_θ is still a bounded exponential distribution with parameter $\lambda - \theta$.

For a bounded normal distribution, the exponential change of measure distribution is

$$f_\theta(x|\gamma_1 \leq x < \gamma_2) = \frac{\frac{1}{\sigma}\phi(\frac{x-\sigma^2\theta}{\sigma})}{\Phi(\frac{\gamma_2-\theta\sigma^2}{\sigma}) - \Phi(\frac{\gamma_1-\theta\sigma^2}{\sigma})}, \tag{5.27}$$

where the original distribution truncated from a normal distribution with parameters mean 0 and variance σ^2. We note that the change of measure distribution is still a bounded normal distribution with mean $\theta\sigma^2$ and variance σ^2.

5.5 Cross Entropy Method and Implementation

Section 5.4 discussed optimal IS distribution F^{**} providing 0 variance estimation to the value of interest, whereas this section describes the Cross Entropy method used to estimate the "optimal" parameters θ, which minimizes the "distance" between a parametric distribution F_θ and F^{**} without knowing F^{**}. The description below is based on the Piecewise Mixture Distribution structure.

5.5.1 Introduction

The Cross Entropy, which is also known as Kullback-Leibler distance [164], measures the similarity between distributions. We define the Cross Entropy between function g and h as

$$\mathcal{D}(g,h) = E_g[ln\frac{g(X)}{f(X)}] = \int g(x)\ln g(x)dx - \int g(x)\ln f(x)dx. \tag{5.28}$$

From (5.24), we know that the PDF of the optimal IS distribution F^{**} is

$$f^{**}(x) = \frac{I_\varepsilon(x)f(x)}{P(x \in \varepsilon)}. \tag{5.29}$$

Since $P(x \in \varepsilon)$ is generally unavailable, we use a parametric distribution F_θ to approach the optimal IS distribution. We want to find the parameter θ^* that minimizes the Cross Entropy [99] between f^{**} and f_θ. We denote θ^* as the optimal parameter for the parametric distribution. Then the minimization problem

$$\min_\theta \mathcal{D}(f_\theta, f^{**}) \tag{5.30}$$

is equivalent to

$$\max_\theta \; E_{\theta_s}[I_\varepsilon(X)\frac{f(X)}{f_{\theta_s}(X)} \ln f_\theta(X)], \tag{5.31}$$

where f_{θ_s} denotes the sampling distribution with parameters θ_s. We note that this is a generalized setting, since we can use any sampling distribution f_{θ_s} as long as it has the same support with f. This is the baseline for iterations in the Cross Entropy method. We use the same form as f_θ because in the following sections, we use a sampling distribution which is in the same family as the parametric distribution.

We estimate θ^* by solving the stochastic counterpart of (5.31)

$$\max_\theta \; \frac{1}{N}\sum_{n=1}^{N} I_\varepsilon(X_n)\frac{f(X_n)}{f_{\theta_s}(X_i)} \ln f_\theta(X_n), \tag{5.32}$$

where samples $\{X_1, ..., X_N\}$ are drawn from the sampling distribution f_{θ_s}.

We note that if $I_\varepsilon(X_n) = 0$ for all $n = 1, .., N$ in (5.32), the objective equals to 0 constantly. To avoid this situation, we select a sampling distribution which emphasizes the rarer events.

Fig. 5.3 shows the iteration procedure of the Cross Entropy method. The core part of the Cross Entropy method is to use the optimizer of the objective function (5.32) in the ith iteration, θ_i^*, as the parameters for the sampling distribution in the next iteration. The underlying idea is that the IS distribution in distribution family f_θ should better approach the optimal IS distribution. Therefore, as we iterate, we obtain more "critical" rare events and have a better estimation of the optimizer which leads to even more "critical" rare events in the next iteration. We define the stopping criterion regarding the parameter or the objective value. In practice, we want to start with an appropriate sampling distribution to get a good solution with less iteration. See section 5.5.3 for a discussion of initializing a sampling distribution.

We note that if we have two independent variables where $f(x, y) = f(x)f(y)$, we can take a parametric distribution for each variable and have $f_\Theta(x, y) = f_{\theta_1}(x)f_{\theta_2}(y)$, where $\Theta = \{\theta_1, \theta_2\}$. The objective function corresponding to (5.32) is

$$\max_\theta \frac{1}{N}\sum_{n=1}^{N} I_\varepsilon(X_n, Y_n)\frac{f(X_n, Y_n)}{f_{\Theta_s}(X_n, Y_n)}(\ln f_{\theta_1}(X_n) + \ln f_{\theta_2}(Y_n)), \tag{5.33}$$

which can be decoupled into two optimization problem over θ_1 and θ_2 respectively and $I_\varepsilon(X_n, Y_n)\frac{f(X_n,Y_n)}{f_{\Theta_s}(X_n,Y_n)}$ is a known constant given $\{X_n, Y_n\}$.

We implement the Cross Entropy on the Piecewise Mixture Distribution with one variable. We note that we can apply the results to the lane change model, since the Cross Entropy objective function of independent variables can be implemented in (5.33).

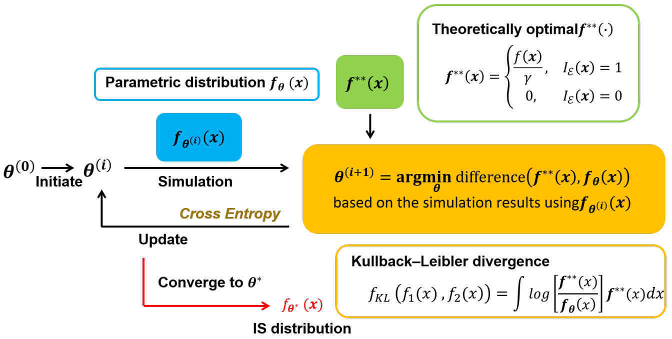

Theoretically optimal $f^{**}(\cdot)$
$f^{**}(x) = \begin{cases} \frac{f(x)}{\gamma}, & I_\varepsilon(x) = 1 \\ 0, & I_\varepsilon(x) = 0 \end{cases}$
Parametric distribution $f_\theta(x)$
$f^{**}(x)$
$f_{\theta^{(i)}}(x)$
$\theta^{(0)}$
Initiate
$\theta^{(i)}$
Simulation
$\theta^{(i+1)} = \operatorname{argmin}_\theta$ difference$(f^{**}(x), f_\theta(x))$
based on the simulation results using $f_{\theta^{(i)}}(x)$
Cross Entropy
Update
Converge to θ^*
$f_{\theta^*}(x)$
IS distribution
Kullback–Leibler divergence
$f_{KL}(f_1(x), f_2(x)) = \int log\left[\frac{f^{**}(x)}{f_\theta(x)}\right]f^{**}(x)dx$

Figure 5.3: Iterations of Cross Entropy.

5.5.2 Optimization Function for Piecewise Mixture Distributions

We propose a parametric family of IS distribution for Piecewise Mixture Distribution

$$f_\theta(x) = \sum_{i=1}^{k} \tilde{\pi}_i \exp(\theta_i x - \kappa(\theta_i)) f_i(x|\gamma_{i-1} \le x < \gamma_i), \tag{5.34}$$

where we use exponential change of measure for each piece of distribution and adjust the proportion parameter to $\tilde{\pi}_i$. The parameter is $\theta = \{\theta_1, ..., \theta_k, \tilde{\pi}_1, ..., \tilde{\pi}_k\}$.

In (5.32), $c_n = I_\varepsilon(X_n)\frac{f(X_n)}{f_{\theta_s}(X_n)}$ is a known constant given the data, so we simplify the function as

$$\max_\theta \; \frac{1}{N} \sum_{n=1}^{N} c_n \ln f_\theta(X_n). \tag{5.35}$$

We split the samples into index sets $S_i = \{j|\gamma_{i-1} \le X_j < \gamma_i\}$ for $i = 1, ..., k$ for each bounded segment. Since $f_i(X_n|\gamma_{i-1} \le x < \gamma_i) \ne 0$ only if $n \in S_i$, for each θ_i and $\tilde{\pi}_i$, the optimization function is equivalent to

$$\max_{\theta_i, \tilde{\pi}_i} \frac{1}{N} \sum_{n \in S_i} c_n \ln(\tilde{\pi}_i \exp(\theta_i X_n - \kappa(\theta_i)) f_i(X_n|x < \gamma_{i-1} \le x < \gamma_i)). \tag{5.36}$$

We can further rewrite the optimization function regarding θ_i and $\tilde{\pi}_i$ respectively. For $\tilde{\pi}_i$, we have

$$\max_{\tilde{\pi}_i} \frac{1}{N} \sum_{n \in S_i} c_n \ln \tilde{\pi}_i, \tag{5.37}$$

which obtains an analytical form for the optimizer

$$\tilde{\pi}_i = \frac{\sum_{n \in S_i} c_n \mathbf{1}\{n \in S_i\}}{\sum_{n \in S_i} c_n}. \tag{5.38}$$

For θ_i, we have

$$\max_{\theta_i} \frac{1}{N} \sum_{n \in S_i} c_n \ln \exp(\theta_i X_n - \kappa(\theta_i)) f_i(X_n|\gamma_{i-1} \le x < \gamma_i), \tag{5.39}$$

which is an exponential change of measure with D_i only. We note that we can simplify this optimization function by rewriting the log term as

$$\max_{\theta_i} \frac{1}{N} \sum_{n\in S_i} c_n(\ln \exp(\theta_i X_n - \kappa(\theta_i)) + \ln f_i(X_n|\gamma_{i-1} \leq x < \gamma_i)), \tag{5.40}$$

which is equivalent to

$$\max_{\theta_i} \frac{1}{N} \sum_{n\in S_i} c_n(\theta_i X_n - \kappa(\theta_i)), \tag{5.41}$$

since the latter term does not depend on θ_i.

For a bounded exponential distribution with parameter λ, the Cross Entropy iteration solves

$$\max_{\theta_i} \frac{1}{N} \sum_{n\in S_i} c_n(\theta_i X_n - \ln \frac{e^{-(\lambda-\theta_i)\gamma_{i-1}} - e^{-(\lambda-\theta_i)\gamma_i}}{\lambda - \theta_i}). \tag{5.42}$$

For a bounded normal distribution with parameters $\mu = 0$ and σ, the optimization function for the Cross Entropy iteration is

$$\max_{\theta_i} \sum_{n\in S_i} c_n X_n \theta_i - (\sum_{n\in S_i} c_n)(\frac{\sigma^2\theta_i^2}{2} + \ln \frac{\Phi(\frac{\gamma_i-\theta_i\sigma^2}{\sigma}) - \Phi(\frac{\gamma_{i-1}-\theta_i\sigma^2}{\sigma})}{\Phi(\frac{\gamma_i}{\sigma}) - \Phi(\frac{\gamma_{i-1}}{\sigma})}). \tag{5.43}$$

5.5.3 Discussion on Numerical Implementation

We have presented the optimization functions for Cross Entropy iterations, but we cannot reliably apply these equations in practice without considering some of the problematical numerical details. In this section, we discuss methods to overcome these numerical issues.

Initializing Cross Entropy Iterations for Rare Events

Since rare events occur with small probability, using the original distribution as sampling distribution to start the Cross Entropy iterations it becomes computationally burdensome to sample a single rare event. One possible approach is to initialize with guess of sampling distribution. When we have some rough knowledge about the optimal IS distribution, we can use the knowledge to construct a proper sampling distribution.

For cases where we have little knowledge about the optimal IS distribution, we construct adaptive events that gradually reduce the rarity. For rare events denoted by ε, we define the sequence of events to be $\varepsilon_1 \supset \varepsilon_2 \supset ... \supset \varepsilon_n \supset \varepsilon$, where ε_1 is not rare for our initializing sampling density. For each iteration t, we gradually reduce the rare event set ε_t and use ε_t to replace ε in the objective function. Since ε_t is a subset of ε_{t-1}, the IS distribution for ε_{t-1} also provides more chances for samples from ε_t. We use the optimal solution in $(t-1)$th iteration θ^*_{t-1} as the sampling parameter θ_t for the next iteration and choose ε_t to have a relatively larger probability to occur under f_{θ_t}. Since ε_t gradually approaches ε as we iterate, eventually we obtain the optimal parameters for ε.

Adjusting sample size N

The choice of sample size N should not only depend on the total number of rare events obtained in each iteration. For each parameter of interest, we need sufficient non-zero c_n's to guarantee the qualification of the estimation. We note that the parameters estimation depend only on the rare event in the corresponding piece, so we adjust sample size N to ensure that each piece with large portion $\tilde{\pi}_i$ contains enough rare event samples.

Setting a lower bound for $\tilde{\pi}_i$

When we update $\tilde{\pi}_i$ in (5.38), if $c_n = 0$ for all $n \in S_i$, meaning that there is no rare event sample in the piece, we have $\tilde{\pi}_i = 0$. When we have $\tilde{\pi}_i = 0$, the support of the IS distribution will differ from the original distribution. We note that it might cause bias in our simulation analysis. On the other hand, once $\tilde{\pi}_i$ hits 0, it will be 0 in the following iterations. Therefore, we need to keep $\tilde{\pi}_i > 0$. Setting a low bound for $\tilde{\pi}_i$, for example, 0.01, when there is no rare event for piece i, gives an efficient IS distribution while avoiding the problems.

Updating parameter θ_i

The absence of rare event samples also leads to failures in updating θ_i. In this case, we use either the value of θ_i in the last iteration, or we set it to 0, i.e. reset the distribution as the real distribution. We note that we can tolerant some inaccurate estimation if $\tilde{\pi}_i$ is small, since a small $\tilde{\pi}_i$ indicates that this piece might not be important to the rare events.

Changing truncation γ_i

The truncations of the Piecewise Mixture Distribution are fixed throughout the Cross Entropy method. Thus, if there is a bad selection of truncation in our original distribution model, the Cross Entropy cannot give an efficient IS distribution. The changing of truncation points is hard to implement by optimization, so we use a heuristic approach for adjusting the truncation points to emphasize the tail part of the Piecewise IS distribution.

In any iteration, if the number of rare events is not enough to properly update the parameters, we check $\tilde{\pi}_i$ of the current sampling distribution. If the $\tilde{\pi}_k$ of the tail piece is the largest possible value, we increase the value of the all truncation points except γ_0 with a certain value. Shifting the truncation gives more weight to the tail part. Then by sampling from the adjusted distribution, we check if the number of events of interest is sufficient. We repeat these actions until we obtain enough rare events in the iteration.

We propose this heuristic approach, since the flexibility of the Piecewise Mixture Distribution is not fully exploited if we cannot change the truncation points. We note that finding a more systematic procedure to locate the knots remains an open question.

5.6 Simulation Analysis

5.6.1 Automated Vehicle Model

First, we present our Piecewise Mixture Models for R^{-1} and TTC^{-1} and then compare the results with the single parametric distribution model used in [173]. For both approaches, we divide the data of TTC^{-1} into three segments regarding the range of v. Since the three segments are similar in distribution, we only show the results of the segment for v in the range of 5 to 15 m/s. We use BIC as the criterion for the goodness of fitting.

Piecewise mixture models for R^{-1} and TTC^{-1}

In Fig. 5.4, we truncated the data into two parts. For the tail part, we use the exponential distribution. For the body part, the mixture of two normal distributions gives a better fit (BIC is -2.6931×10^5, BIC for exponential and normal is -2.6905×10^5 and -2.6865×10^5 respectively). The Piecewise Mixture Models enable us to use different distributions for the body part and the tail part.

Comparison with single parametric distribution models

Fig.5.5 compare the new model and the previous model for R^{-1}. The Piecewise Mixture Distribution with two truncation points (BIC is -1.0428×10^6) provides a better fitting than the single parametric distribution (BIC is -1.0426×10^6). We have the same observation for the fitting of TTC^{-1} in Fig.5.6. The piecewise model provides BIC with -2.6931×10^5, where the single parametric distribution gives BIC with -2.6686×10^5. The result indicates that Piecewise Mixture Models provide more flexibility in data fitting.

5.6.2 Cross Entropy Results

Here, we use the lane change model to exemplify the Cross Entropy method. For the three variables R, TTC, v, the distribution is $f(R, TTC, v) = f(v)f(R)f(TTC|v)$ where $f(v)$ is the empirical distribution. Since we have three conditional distributions of TTC regarding the value of v, we find the IS distributions independently for each case. We present the results for v from 5 to 15 m/s.

We assume that we have less information about the relation between the distribution of variables and the rare events. Our objective is to construct adaptive rare events to help us approach the IS distribution. We recall that our original lane change model determines whether a crash happens by checking to see if the value of R, the range between two vehicles, reaches 0. Meanwhile, the TTC also goes to 0 when a crash happens. To construct events less rare than a crash as mentioned in Section 5.5.3, we relax the criterion for crash to be either R hits $t_R > 0$ or TTC hits $t_{TTC} > 0$. By changing these two thresholds, t_R and t_{TTC} as shown in Fig. 5.7, we construct the adaptive rare events sequence for the Cross Entropy iterations. The value of threshold is picked by taking the smaller number between the 0.95 quantile

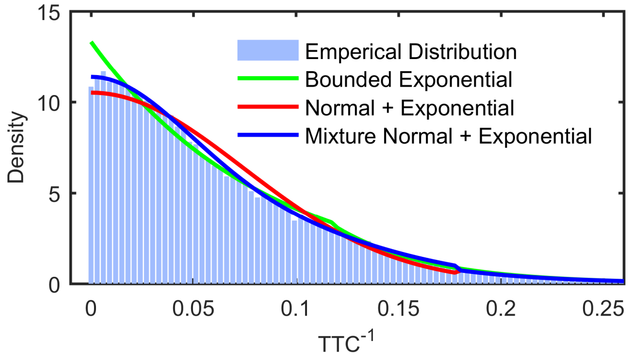


Figure 5.4: Piecewise Mixture Distribution fitting for TTC^{-1} given v between 5 and 15 m/s.

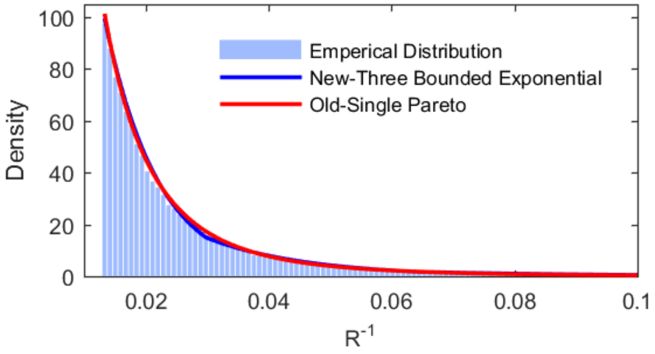


Figure 5.5: Comparison of fitting for R^{-1}.

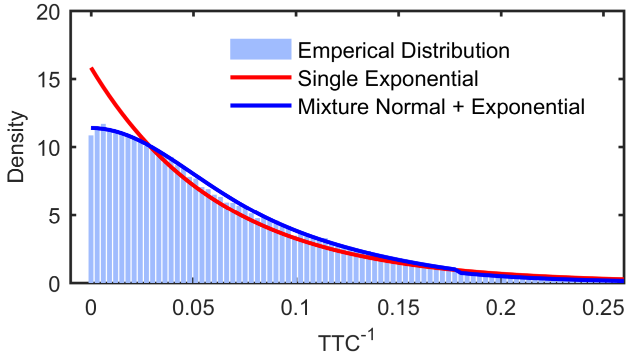


Figure 5.6: Comparison of fitting for TTC^{-1} given v between 5 and 15 m/s.

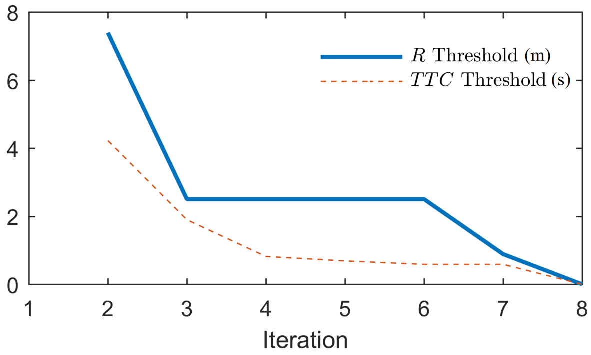


Figure 5.7: Cross Entropy iterations with sequence of events with thresholds for crash. We leave iteration 1 blank to keep the x-axis consistent with Fig. 5.8 and 5.9.

of the generated data and the current threshold. We set the thresholds to be 0 for both variable, when the value of the current thresholds are close to zero (less than 0.5 in this case). We use sample size $N = 1000$ for each iteration.

Fig. 5.8 and 5.9 show the parameters present in each of the iterations. We observe that the parameters stabilize gradually. Fig. 5.10 shows how the distribution changes gradually from the original distribution to the IS distribution. We note that the density moves toward the tail part as we iterate. This observation shows that the algorithm gradually learns the "importance" of the tail part.

5.6.3 Simulation Results

In our simulation experiments, we set the convergence criterion as the relative half-width of $100(1-\alpha)\%$ confidence interval drops below β. In this case, we use $\alpha = 0.2$ and $\beta = 0.2$ to study the number of samples needed for convergence. Our goal is to compare the efficiency of the Piecewise Mixture Distribution and single exponential distribution models in estimating the probability of crashes in the lane change scenario for the testing AV system.

Fig. 5.11 shows that both models give a similar estimation as the number of experiments grows large, and that the Piecewise Mixture Distribution model converges

Table 5.1: Number of samples (N) needed to converge.

	Piecewise	Single	Crude
N	7840	12320	5.5×10^7
Ratio to Piecewise	1	1.57	7×10^3

slightly faster than the single parametric model. The circles show that the relative half-width of the Piecewise Mixture Distribution model reaches the target confidence value after 7800 samples, whereas the single parametric model needs about 13800 samples. Using the Piecewise Mixture Distribution model reduced the sample size by 44%.

To reduce stochastic uncertainty, we repeat the tests 10 times and calculate the average. It takes 7840 samples on average to obtain a converged estimation using the Piecewise Mixture Distribution model, whereas it takes 12320 samples on average using the single accelerated distribution model to converge. Table 5.1 compares the two models with the crude Monte Carlo method [4]. We estimate the number needed for convergence of crude Monte Carlo by using the fact that the number of events of interest occurring is Binomial distributed. We compute the standard deviation of the crude Monte Carlo estimation $\hat{P}(x \in \varepsilon)$ by

$$std(\hat{P}(x \in \varepsilon)) = \sqrt{\frac{\hat{P}(x \in \varepsilon)(1 - \hat{P}(x \in \varepsilon))}{n}}, \tag{5.44}$$

which allows us to estimate

$$\hat{N} = \frac{z^2_{\alpha/2}(1 - \hat{P}(x \in \varepsilon))}{\beta^2 \hat{P}(x \in \varepsilon)}, \tag{5.45}$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of normal distribution. We calculate the required sample size N of crude Monte Carlo in Table 5.1 from an estimation $\hat{P}(x \in \varepsilon) = 7.4 \times 10^{-7}$ with 80% confidence interval $(7.0 \times 10^{-7}, 7.8 \times 10^{-7})$.

Finally, we apply the heuristic approach in Section 5.5.3 to the data segment with v from 5 to 15 m/s. We run simulations with this segment and compare the results with the standard approach for the Piecewise Mixture Distribution and single parametric distribution models. Fig. 5.12 shows the convergence of confidence half-width. We determine convergence as the relative confidence half-width smaller than β (the dash line). We note that the relative half-width of the heuristic, which is smaller than the standard approach for the Piecewise Mixture Distribution model, indicates that the latter model's performance can be further improved.

5.7 Conclusions

This chapter proposed a new model for accelerated evaluation of AVs. The Piecewise Mixture Distribution Models provide more accurate fitting to the surrounding human-controlled vehicle behaviors than the single parametric model used in the literature. The proposed model was more efficient and reduced the evaluation time by almost half than single parametric model. The Cross Entropy procedure described in this chapter effectively worked in this scenario analysis. We provided practical solutions to deal with the numerical issues which occurred while calculating the optimal parameters. The heuristic approach exploited the flexibility of the Piecewise Mixture Distribution structure. Testing the proposed model on a large dataset of cut-in crashes caused by improper lane changes, the Piecewise Mixture Distribution model reduced the simulation cases by about 33% compared with the single parametric model under the same convergence requirement. Moreover, the proposed model was 7000 times faster than the Crude Monte Carlo method.

Table 5.2 summarizes the comparison of the computation efforts between the models. We note that using the Piecewise Mixture Distribution model increases the number of parameters estimated, where the estimation of parameters is almost instant. In the Cross Entropy stage, the number of simulations required for the

Table 5.2: Comparison of the computation time between single parametric model and piecewise model.

Stages	Crude	Single	Piecewise
Fitting	-	4 parameters to estimate	18 parameters to estimate
Cross Entropy	-	30,000 simulations 4 parameters	24,000 simulations 18 parameters
Simulation	5.5×10^7 simulations	12,320 simulations	7840 simulations

Piecewise model is not significantly less than the single parametric model, because we assume no knowledge about the optimal IS distribution for the Piecewise model. Overall, the Piecewise model needs fewer simulations to reach the same confidence level compared to single parametric models.

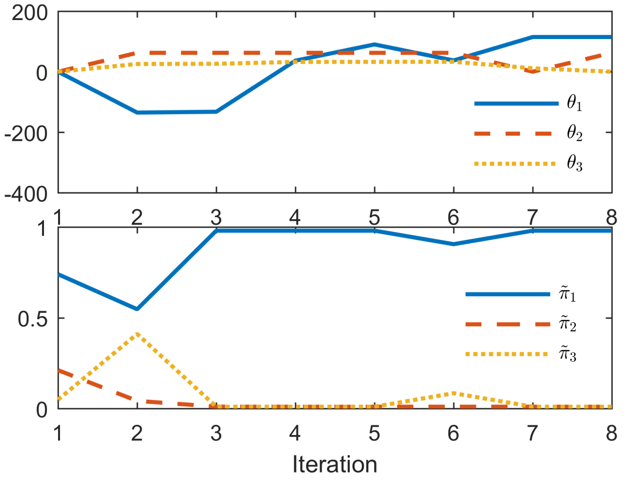


Figure 5.8: Cross Entropy iterations with sequence of events of R^{-1} for v from 5 to 15 m/s.

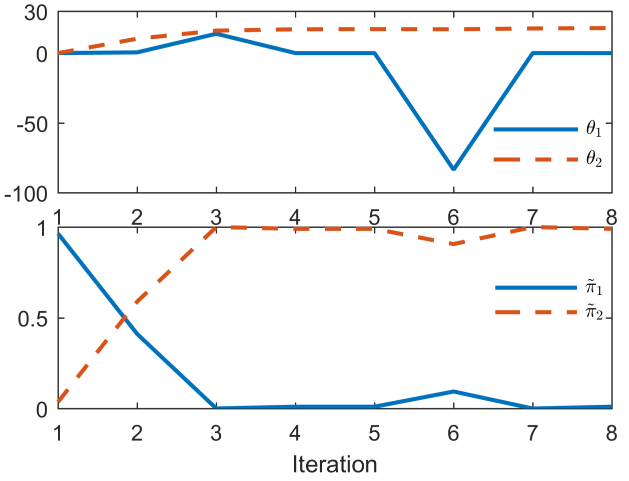


Figure 5.9: Cross Entropy iterations with sequence of events of TTC^{-1} for v from 5 to 15 m/s.

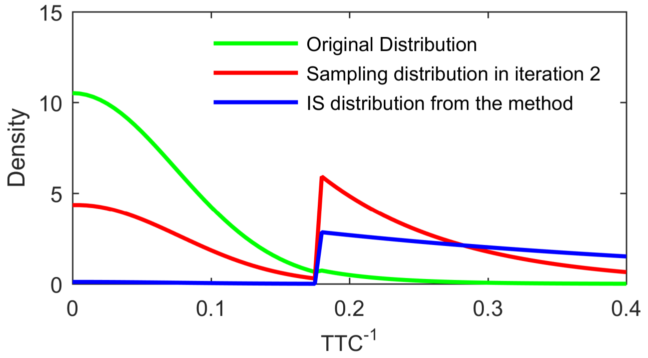


Figure 5.10: Distribution change through Cross Entropy iterations with sequence of events of TTC^{-1} for v from 5 to 15 m/s.

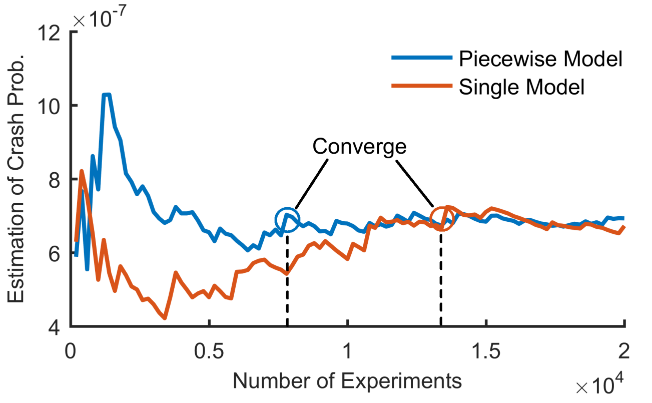


Figure 5.11: Estimation of crash probability for one lane change using piecewise and single accelerated distributions. The x-axis is truncated for illustrating the major change of the distributions.

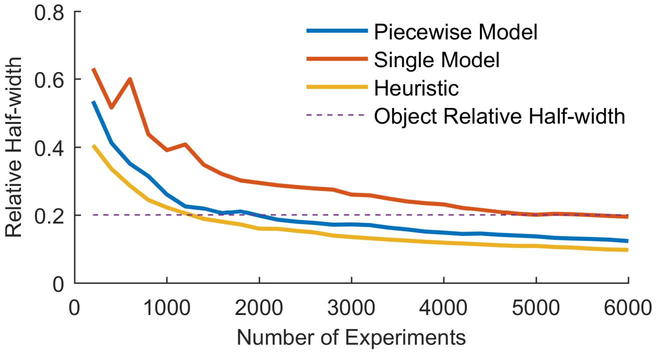


Figure 5.12: Relative half-width of crash probability estimation for one lane change with leading vehicle's speed in range of 5 to 15 m/s, comparing single, piecewise and heuristic accelerated distributions.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Preliminary Report, ?Highway HWY16FH018.

[2] Robert J Adler, Jose H Blanchet, Jingchen Liu, et al. Efficient monte carlo for high excursions of gaussian random fields. *The Annals of Applied Probability*, 22(3):1167–1214, 2012.

[3] J Aldrich. RA Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science*, 1997.

[4] S Asmussen and PW Glynn. *Stochastic Simulation: Algorithms and Analysis.* Springer, 2007.

[5] Søren Asmussen and Peter W Glynn. *Stochastic Simulation: Algorithms and Analysis*, volume 57. Springer Science & Business Media, 2007.

[6] Rami Atar, Kenny Chowdhary, and Paul Dupuis. Robust bounds on risk-sensitive functionals via rényi divergence. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):18–33, 2015.

[7] M Aust. Evaluation Process for Active Safety Functions: Addressing Key Challenges in Functional, Formative Evaluation of Advanced Driver Assistance Systems. 2012.

[8] RR Barton, SE Chick, RC Cheng, SG Henderson, AM Law, BW Schmeiser, LM Leemis, LW Schruben, and JR Wilson. Panel discussion on current issues in input modeling. In E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, editors, *Proceedings of the 2002 Winter Simulation Conference*, pages 353–369, Piscataway, New Jersey, 2002. Institute of Electrical and Electronics Engineers, Inc.

[9] Russell R Barton. Input uncertainty in outout analysis. In C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, editors, *Proceedings of the 2012 Winter Simulation Conference*, pages 67–78, Piscataway, New Jersey, 2012. Institute of Electrical and Electronics Engineers, Inc.

[10] Russell R Barton and Lee W Schruben. Resampling methods for input modeling. In B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, editors, *Proceedings of the 2001 Winter Simulation Conference*, pages 372–378, Piscataway, New Jersey, 2001. Institute of Electrical and Electronics Engineers, Inc.

[11] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[12] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization.* Princeton University Press, 2009.

[13] Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

[14] Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, 1999.

[15] Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, 88(3):411–424, 2000.

[16] Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

[17] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

[18] Dimitris Bertsimas, Dessislava Pachamanova, and Melvyn Sim. Robust linear optimization under general norms. *Operations Research Letters*, 32(6):510–516, 2004.

[19] Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[20] Dimitris Bertsimas and Melvyn Sim. Tractable approximations to robust conic optimization problems. *Mathematical Programming*, 107(1-2):5–36, 2006.

[21] Debby Bezzina and James R Sayer. Safety Pilot: Model Deployment Test Conductor Team Report. Technical Report June, NHTSA, 2014.

[22] Jose Blanchet, Christopher Dolan, and HK Lam. Robust rare-event performance analysis with natural non-convex constraints. In *Proceedings of the 2014 Winter Simulation Conference (WSC)*, pages 595–603. IEEE, 2014.

[23] Jose Blanchet and Yang Kang. Sample out-of-sample inference based on Wasserstein distance. *arXiv preprint arXiv:1605.01340*, 2016.

[24] Jose Blanchet and Henry Lam. State-dependent importance sampling for rare-event simulation: An overview and recent advances. *Surveys in Operations Research and Management Science*, 17(1):38–59, 2012.

[25] Jose Blanchet and Chenxin Li. Efficient simulation for the maximum of infinite horizon discrete-time gaussian processes. *Journal of Applied Probability*, 48(2):467–489, 2011.

[26] Jose Blanchet and Karthyek RA Murthy. On distributionally robust extreme value analysis. *arXiv preprint arXiv:1601.06858*, 2016.

[27] S Boyd and L Vandenberghe. *Convex Optimization.* 2004.

[28] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge university press, 2004.

[29] J Bucklew. *Introduction to Rare Event Simulation.* Springer Science & Business Media, 2004.

[30] James Bucklew. *Introduction to Rare Event Simulation.* Springer Science & Business Media, 2013.

[31] Giuseppe Calafiore and Marco C Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102(1):25–46, 2005.

[32] Giuseppe C Calafiore. Repetitive scenario design. *IEEE Transactions on Automatic Control*, 62(3):1125–1137, 2017.

[33] Giuseppe C Calafiore and Marco C Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753, 2006.

[34] Giuseppe C Calafiore, Fabrizio Dabbene, and Roberto Tempo. Research on probabilistic methods for control system design. *Automatica*, 47(7):1279–1293, 2011.

[35] Giuseppe Carlo Calafiore and Laurent El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(1):1–22, 2006.

[36] Marco C Campi and Algo Carè. Random convex programs with L_1-regularization: sparsity and generalization. *SIAM Journal on Control and Optimization*, 51(5):3532–3557, 2013.

[37] Marco C Campi and Simone Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.

[38] Marco C Campi and Simone Garatti. A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. *Journal of Optimization Theory and Applications*, 148(2):257–280, 2011.

[39] Marco C Campi and Simone Garatti. Wait-and-judge scenario optimization. *Mathematical Programming*, 167(1):155–189, 2018.

[40] Algo Carè, Simone Garatti, and Marco C Campi. FAST: Fast algorithm for the scenario technique. *Operations Research*, 62(3):662–671, 2014.

[41] Enrique Castillo and Ali S Hadi. Fitting the generalized pareto distribution to data. *Journal of the American Statistical Association*, 92(440):1609–1620, 1997.

[42] Mohammadreza Chamanbaz, Fabrizio Dabbene, Roberto Tempo, Venkatakrishnan Venkataramanan, and Qing-Guo Wang. Sequential randomized algorithms for convex optimization in the presence of uncertainty. *IEEE Transactions on Automatic Control*, 61(9):2565–2571, 2016.

[43] Abraham Charnes and William W Cooper. Chance-constrained programming. *Management Science*, 6(1):73–79, 1959.

[44] Abraham Charnes, William W Cooper, and Gifford H Symonds. Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil. *Management Science*, 4(3):235–263, 1958.

[45] Wenqing Chen, Melvyn Sim, Jie Sun, and Chung-Piaw Teo. From CVaR to uncertainty set: Implications in joint chance-constrained optimization. *Operations Research*, 58(2):470–485, 2010.

[46] Xin Chen, Melvyn Sim, and Peng Sun. A robust optimization perspective on stochastic programming. *Operations Research*, 55(6):1058–1071, 2007.

[47] Stephen E Chick. Bayesian ideas and discrete event simulation: Why, what and how. In L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, editors, *Proceedings of the 2006 Winter Simulation Conference*, pages 96–105, Piscataway, New Jersey, 2006. Institute of Electrical and Electronics Engineers, Inc.

[48] Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

[49] DR Cox and DV Hinkley. *Theoretical Statistics*. 1979.

[50] Daniela Pucci De Farias and Benjamin Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, 29(3):462–478, 2004.

[51] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[52] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications.* Springer Science & Business Media, 2nd edition, 1998.

[53] AP Dempster, NM Laird, and DB Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the royal statistical society*, 1977.

[54] Denis Denisov, Antonius Bernardus Dieker, Vsevolod Shneer, et al. Large deviations for random walks under subexponentiality: the big-jump domain. *The Annals of Probability*, 36(5):1946–1991, 2008.

[55] Darinka Dentcheva, Bogumila Lai, and Andrzej Ruszczyński. Dual methods for probabilistic optimization problems. *Mathematical Methods of Operations Research*, 60(2):331–346, 2004.

[56] Anulekha Dhara, Bikramjit Das, and Karthik Natarajan. Worst-case expected shortfall with univariate and bivariate marginals. *arXiv preprint arXiv:1701.04167*, 2017.

[57] Antonius Bernardus Dieker and Michel Mandjes. Fast simulation of overflow probabilities in a queue with gaussian input. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 16(2):119–151, 2006.

[58] David L Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, pages 1803–1827, 1992.

[59] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

[60] Paul Dupuis, Markos A Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. Sensitivity analysis for rare events based on rényi divergence. *arXiv preprint arXiv:1805.06917*, 2018.

[61] Rick Durrett. *Probability: Theory and Examples.* Cambridge university press, 2010.

[62] Laurent El Ghaoui, Maksim Oks, and Francois Oustry. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, 51(4):543–556, 2003.

[63] Laurent El Ghaoui, Francois Oustry, and Hervé Lebret. Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization*, 9(1):33–52, 1998.

[64] Paul Embrechts, Rdiger Frey, and Alexander McNeil. Quantitative risk management. *Princeton Series in Finance, Princeton*, 10, 2005.

[65] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events: For Insurance and Finance*, volume 33. Springer Science & Business Media, 2013.

[66] Paul Embrechts and Giovanni Puccetti. Bounds for functions of multivariate risks. *Journal of Multivariate Analysis*, 97(2):526–547, 2006.

[67] Paul Embrechts, Giovanni Puccetti, and Ludger Rüschendorf. Model uncertainty and var aggregation. *Journal of Banking & Finance*, 37(8):2750–2764, 2013.

[68] E Erdoğan and Garud Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1-2):37–61, 2006.

[69] A. Evan. Fatal Tesla Self-Driving Car Crash Reminds Us That Robots Aren't Perfect. *IEEE Spectrum*, 2016.

[70] Julian J Faraway. *Linear models with R.* CRC press, 2014.

[71] FESTA-Consortium. FESTA Handbook Version 2 Deliverable T6.4 of the Field opErational teSt supporT Action. Technical report, FESTA, 2008.

[72] Paul Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53. Springer Science & Business Media, New York, 2013.

[73] Paul Glasserman and Xingbo Xu. Robust risk measurement and model risk. *Quantitative Finance*, 14(1):29–58, 2014.

[74] Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4-part-1):902–917, 2010.

[75] Donald Goldfarb and Garud Iyengar. Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38, 2003.

[76] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep Learning*, volume 1. MIT press Cambridge, Massachusetts, 2016.

[77] Google Auto LLC. Monthly reports ? Google Self-Driving Car Project.

[78] Vishal Gupta. Near-optimal ambiguity sets for distributionally robust optimization. *Management Science*, Articles in advance, 2019.

[79] Marc Hallin, Davy Paindaveine, Miroslav Šiman, Ying Wei, Robert Serfling, Yijun Zuo, Linglong Kong, and Ivan Mizera. Multivariate quantiles and multiple-output regression quantiles: From L_1 optimization to halfspace depth. *The Annals of Statistics*, pages 635–703, 2010.

[80] Grani A Hanasusanto, Vladimir Roitch, Daniel Kuhn, and Wolfram Wiesemann. A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming*, 151(1):35–62, 2015.

[81] Grani A Hanasusanto, Vladimir Roitch, Daniel Kuhn, and Wolfram Wiesemann. Ambiguous joint chance constraints under mean and dispersion information. *Operations Research*, 65(3):751–767, 2017.

[82] Lars Peter Hansen and Thomas J Sargent. *Robustness*. Princeton university press, 2008.

[83] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. In *The Elements of Statistical Learning*, pages 485–585. Springer, 2009.

[84] Shane G Henderson. Input model uncertainty: Why do we care and what should we do about it? In S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, editors, *Proceedings of the 2003 Winter Simulation Conference*, volume 1, pages 90–100, Piscataway, New Jersey, 2003. Institute of Electrical and Electronics Engineers, Inc.

[85] Joseph L Hodges. A bivariate sign test. *The Annals of Mathematical Statistics*, 26(3):523–527, 1955.

[86] L Jeff Hong, Zhiyuan Huang, and Henry Lam. Learning-based robust optimization: Procedures and statistical guarantees. *arXiv preprint arXiv:1704.04342*, 2017.

[87] L Jeff Hong, Yi Yang, and Liwei Zhang. Sequential convex approximations to joint chance constrained programs: A Monte Carlo approach. *Operations Research*, 59(3):617–630, 2011.

[88] Jonathan RM Hosking and James R Wallis. Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3):339–349, 1987.

[89] Zhaolin Hu, L Jeff Hong, and Liwei Zhang. A smooth Monte Carlo approach to joint chance-constrained programs. *IIE Transactions*, 45(7):716–735, 2013.

[90] Zhiyuan Huang and Henry Lam. On the impacts of tail model uncertainty in rare-event estimation. In *2019 Winter Simulation Conference (WSC)*, pages 950–961. IEEE, 2019.

[91] Zhiyuan Huang, Henry Lam, David J LeBlanc, and Ding Zhao. Accelerated evaluation of automated vehicles using piecewise mixture models. *IEEE Transactions on Intelligent Transportation Systems*, 19(9):2845–2855, 2017.

[92] Zhiyuan Huang, Henry Lam, and Ding Zhao. Designing importance samplers to simulate machine learning predictors via optimization. In *2018 Winter Simulation Conference (WSC)*, pages 1730–1741. IEEE, 2018.

[93] Zhiyuan Huang, Ding Zhao, Henry Lam, David J. LeBlanc, and Huei Peng. Evaluation of Automated Vehicles in the Frontal Cut-in Scenario - an Enhanced Approach using Piecewise Mixture Models. 10 2016.

[94] Ashesh Jain, Hema S Koppula, Bharad Raghavan, Shane Soh, and Ashutosh Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3182–3190, 2015.

[95] Predrag R Jelenkovic. Network multiplexer with truncated heavy-tailed arrival streams. In *Proceedings of the Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 2, pages 625–632. IEEE, 1999.

[96] Ruiwei Jiang and Yongpei Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, 158(1-2):291–327, 2016.

[97] Sandeep Juneja and Perwez Shahabuddin. Rare-event simulation techniques: an introduction and recent advances. *Handbooks in Operations Research and Management Science*, 13:291–350, 2006.

[98] Y Kou. Development and Evaluation of Integrated Chassis Control Systems. 2010.

[99] Dirk P Kroese, Reuven Y Rubinstein, and Peter W Glynn. *The Cross-Entropy Method for Estimation*, volume 31. Elsevier B.V., 2013.

[100] Constantino M Lagoa and B Ross Barmish. Distributionally robust Monte Carlo simulation: A tutorial survey. In *Proceedings of the IFAC World Congress*, pages 1–12. IFAC New York, NY, 2002.

[101] Henry Lam. Advanced tutorial: Input uncertainty and robust analysis in stochastic simulation. In T. M. K. Roeder, P. I. Frazier, R. Szechtman, and E. Zhou, editors, *Proceedings of the 2016 Winter Simulation Conference*, pages 178–192, Piscataway, New Jersey, 2016. Institute of Electrical and Electronics Engineers, Inc.

[102] Henry Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

[103] Henry Lam and Clementine Mottet. Simulating tail events with unspecified tail models. In *Proceedings of the 2015 Winter Simulation Conference*, pages 392–402. IEEE Press, 2015.

[104] Henry Lam and Clementine Mottet. Tail analysis without parametric models: A worst-case perspective. *Operations Research*, 65(6):1696–1711, 2017.

[105] Henry Lam and Enlu Zhou. The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters*, 45(4):301–307, 2017.

[106] Malcolm R Leadbetter. On a basis for 'peaks over threshold' modeling. *Statistics & Probability Letters*, 12(4):357–362, 1991.

[107] Miguel A Lejeune and Andrzej Ruszczynski. An efficient trajectory method for probabilistic production-inventory-distribution problems. *Operations Research*, 55(2):378–394, 2007.

[108] Bowen Li, Ruiwei Jiang, and Johanna L Mathieu. Ambiguous risk constraints with moment and unimodality information. *Mathematical Programming*, 173(1-2):151–192, 2019.

[109] Bowen Li, Ruiwei Jiang, and Johanna L Mathieu. Ambiguous risk constraints with moment and unimodality information. *Mathematical Programming*, 173(1-2):151–192, 2019.

[110] Jun Li and Regina Y Liu. Multivariate spacings based on data depth: I. Construction of nonparametric multivariate tolerance regions. *The Annals of Statistics*, 36(3):1299–1323, 2008.

[111] Andrew EB Lim and J George Shanthikumar. Relative entropy, exponential utility, and robust dynamic pricing. *Operations Research*, 55(2):198–214, 2007.

[112] Andrew EB Lim, J George Shanthikumar, and ZJ Max Shen. Model uncertainty, robust optimization, and learning. *Tutorials in Operations Research: Models, Methods, and Applications for Innovative Decision Making*, pages 66–94, 2006.

[113] Han Liu, Larry Wasserman, and John D Lafferty. Exponential concentration for mutual information estimation with application to forests. In *Advances in Neural Information Processing Systems*, pages 2537–2545, 2012.

[114] Regina Y Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414, 1990.

[115] James Luedtke. A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support. *Mathematical Programming*, 146(1-2):219–244, 2014.

[116] James Luedtke and Shabbir Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699, 2008.

[117] James Luedtke, Shabbir Ahmed, and George L Nemhauser. An integer programming approach for linear programs with probabilistic constraints. *Mathematical Programming*, 122(2):247–272, 2010.

[118] WH Ma and H Peng. A Worst-case Evaluation Method for Dynamic Systems. *Journal of dynamic systems,*, 1999.

[119] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

[120] Ahmadreza Marandi, Aharon Ben-Tal, Dick den Hertog, and Bertrand Melenberg. Extending the scope of robust quadratic optimization. *Available on Optimization Online*, 2017.

[121] Kostas Margellos, Paul Goulart, and John Lygeros. On the road between robust optimization and the scenario approach for chance constrained optimization problems. *IEEE Transactions on Automatic Control*, 59(8):2258–2263, 2014.

[122] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: Concepts, techniques and tools*. Princeton university press, 2015.

[123] Bruce L Miller and Harvey M Wagner. Chance constrained programming with joint constraints. *Operations Research*, 13(6):930–945, 1965.

[124] Velibor V Mišic. Optimization of tree ensembles. *Working Paper: arXiv preprint arXiv:1705.10883*, 2017.

[125] Kevin Moon and Alfred Hero. Multivariate f-divergence estimation with confidence. In *Advances in Neural Information Processing Systems*, pages 2420–2428, 2014.

[126] Michael R Murr and András Prékopa. Solution of a product substitution problem using stochastic programming. In *Probabilistic Constrained Optimization*, pages 252–271. Springer, 2000.

[127] Sergey V Nagaev et al. Large deviations of sums of independent random variables. *The Annals of Probability*, 7(5):745–789, 1979.

[128] Marvin K Nakayama. Asymptotics of likelihood ratio derivative estimators in simulations of highly reliable markovian systems. *Management Science*, 41(3):524–554, 1995.

[129] Marvin K Nakayama. On derivative estimation of the mean time to failure in simulations of highly reliable markovian systems. *Operations Research*, 46(2):285–290, 1998.

[130] B. L. Nelson, A. T. K. Wan, S. Fan, and X. Zhang. Reducing simulation input-model risk via input model averaging. *working paper*, 2019.

[131] Arkadi Nemirovski. On tractable approximations of randomly perturbed convex constraints. In *Proceedings of 42nd IEEE Conference on Decision and Control*, volume 3, pages 2419–2422. IEEE, 2003.

[132] Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2006.

[133] NHTSA. Traffic Safety Facts 2014. Technical report, DOT, 2014.

[134] Mariana Olvera-Cravioto. The single-server queue with heavy tails. *PhD dissertation*, 2006.

[135] Dávid Pál, Barnabás Póczos, and Csaba Szepesvári. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Advances in Neural Information Processing Systems*, pages 1849–1857, 2010.

[136] H Peng and D Leblanc. Evaluation of the Performance and Safety of Automated Vehicles. *White Pap. NSF Transp. CPS Work*, 2012.

[137] Ian R Petersen, Matthew R James, and Paul Dupuis. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control*, 45(3):398–412, 2000.

[138] Barnabás Póczos and Jeff G Schneider. Nonparametric estimation of conditional information and divergences. In *AISTATS*, pages 914–923, 2012.

[139] Barnabás Póczos, Liang Xiong, and Jeff Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. *arXiv preprint arXiv:1202.3758*, 2012.

[140] Ioana Popescu. A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operations Research*, 30(3):632–657, 2005.

[141] Andras Prékopa. On probabilistic constrained programming. In *Proceedings of the Princeton Symposium on Mathematical Programming*, pages 113–138. Princeton University Press Princeton, NJ, 1970.

[142] András Prékopa. Probabilistic programming. *Handbooks in Operations Research and Management Science*, 10:267–351, 2003.

[143] András Prékopa, Tamás Rapcsák, and István Zsuffa. Serially linked reservoir system design using stochastic programing. *Water Resources Research*, 14(4):672–678, 1978.

[144] András Prékopa and Tamás Szántai. Flood control reservoir system design using stochastic programming. In *Mathematical Programming in Use*, pages 138–151. Springer, 1978.

[145] Giovanni Puccetti and Ludger Rüschendorf. Computation of sharp bounds on the distribution of a function of dependent risks. *Journal of Computational and Applied Mathematics*, 236(7):1833–1840, 2012.

[146] Giovanni Puccetti and Ludger Rüschendorf. Sharp bounds for sums of dependent risks. *Journal of Applied Probability*, 50(01):42–53, 2013.

[147] Chang-Han Rhee, Jose Blanchet, and Bert Zwart. Sample path large deviations for lèvy processes and random walks with regularly varying increments. *arXiv preprint arXiv:1606.02795*, 2016.

[148] R Y Rubinstein. Rare Event Simulation via Cross-entropy and Importance Sampling. In *Second International Workshop on Rare Event Simulation*, pages 1–17, 1999.

[149] Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo Method*, volume 10. John Wiley & Sons, New Jersey, 2016.

[150] John S Sadowsky and James A Bucklew. On large deviations theory and asymptotically efficient monte carlo estimation. *IEEE transactions on Information Theory*, 36(3):579–588, 1990.

[151] Herbert Scarf, KJ Arrow, and S Karlin. A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, 10:201–209, 1958.

[152] Georg Schildbach, Lorenzo Fagiano, and Manfred Morari. Randomized solutions to convex programs with multiple chance constraints. *SIAM Journal on Optimization*, 23(4):2479–2501, 2013.

[153] Clayton D Scott and Robert D Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7(Apr):665–704, 2006.

[154] Robert Serfling. Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, 56(2):214–232, 2002.

[155] Robert J Serfling. *Approximation Theorems of Mathematical Statistics*, volume 162. John Wiley & Sons, 2009.

[156] Yuanming Shi, Jun Zhang, and Khaled B Letaief. Optimal stochastic coordinated beamforming for wireless cooperative networks with CSI uncertainty. *IEEE Transactions on Signal Processing*, 63(4):960–973, 2015.

[157] Eunhye Song, Barry L Nelson, and C Dennis Pegden. Advanced tutorial: Input uncertainty quantification. In A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, editors, *Proceedings of the 2014 Winter Simulation Conference*, pages 162–176, Piscataway, New Jersey, 2014. Institute of Electrical and Electronics Engineers, Inc.

[158] Vincent Tjeng and Russ Tedrake. Verifying neural networks with mixed integer programming. *Working Paper: arXiv preprint arXiv:1711.07356*, 2017.

[159] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531, 1975.

[160] Theja Tulabandhula and Cynthia Rudin. Robust optimization using machine learning for uncertainty sets. *arXiv preprint arXiv:1407.1097*, 2014.

[161] AY Ungoren and H Peng. An Adaptive Lateral Preview Driver Model. *Vehicle System Dynamics*, 43(4):245–259, 4 2005.

[162] Bart PG Van Parys, Paul J Goulart, and Daniel Kuhn. Generalized Gauss inequalities via semidefinite programming. *Mathematical Programming*, 156(1-2):271–302, 2016.

[163] Bart PG Van Parys, Paul J Goulart, and Manfred Morari. Distributionally robust expectation inequalities for structured distributions. *Mathematical Programming*, 173(1-2):251–280, 2019.

[164] VN Vapnik. *Statistical Learning Theory*. 1998.

[165] Roman Vershynin. *High-dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

[166] Bin Wang and Ruodu Wang. The complete mixability and convex minimization problems with monotone marginal densities. *Journal of Multivariate Analysis*, 102(10):1344–1360, 2011.

[167] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

[168] Wenshuo Wang, Ding Zhao, Junqiang Xi, David J LeBlanc, and J Karl Hedrick. Development and evaluation of two learning-based personalized driver models for car-following behaviors. In *Proceedings of American Control Conference (ACC), 2017*, pages 1133–1138. IEEE, 2017.

[169] Zizhuo Wang, Peter W Glynn, and Yinyu Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261, 2016.

[170] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[171] Robert Yuen, Stilian Stoev, and Dan Cooley. Distributionally robust inference for extreme value-at-risk. *working paper*, 2019.

[172] Ding Zhao, Xianan Huang, Huei Peng, Henry Lam, and David J Leblanc. Accelerated Evaluation of Automated Vehicles in Car-Following Maneuvers. *ArXiv*, page 12, 2016.

[173] Ding Zhao, Henry Lam, Huei Peng, Shan Bao, David J. LeBlanc, Kazutoshi Nobukawa, and Christopher S. Pan. Accelerated Evaluation of Automated Vehicles Safety in Lane-Change Scenarios Based on Importance Sampling Techniques. *IEEE Transactions on Intelligent Transportation Systems*, 2016.

[174] Ding Zhao, Huei Peng, Shan Bao, Kazutoshi Nobukawa, David J. LeBlanc, and Christopher S. Pan. Accelerated evaluation of automated vehicles using extracted naturalistic driving data. In *Proceeding for 24th International Symposium of Vehicles on Road and Tracks*, 2015.

[175] Yijun Zuo. Projection-based depth functions and associated medians. *Annals of Statistics*, pages 1460–1490, 2003.