Data-Driven Risk-Averse Two-Stage Stochastic Program with ζ-Structure Probability Metrics

Chaoyue Zhao and Yongpei Guan

Department of Industrial and Systems Engineering
University of Florida, Gainesville, FL 32611

April 7, 2015

Abstract

The traditional two-stage stochastic programming approach assumes the distribution of the random parameter in a problem is known. In most practices, however, the distribution is actually unknown. Instead, only a series of historic data are available. In this paper, we develop a data-driven stochastic optimization framework to provide a risk-averse decision making under uncertainty. In our approach, starting from a given set of historical data, we first construct a confidence set for the unknown probability distribution utilizing a family of ζ-structure probability metrics. Then, we describe the reference distributions and solution approaches to solving the developed two-stage risk-averse stochastic program, corresponding to the given set of historical data, for the cases in which the true probability distributions are discrete and continuous, respectively. More specifically, for the case in which the true probability distribution is discrete, we reformulate the risk-averse problem to a traditional two-stage robust optimization problem. For the case in which the true probability distribution is continuous, we develop a sampling approach to obtaining the upper and lower bounds of the risk-averse problem, and prove that these two bounds converge to the optimal objective value uniformly at the sample size increases. Furthermore, we prove that, for both cases, the risk-averse problem converges to the risk-neutral one as more data samples are observed, and derive the convergence rate, which indicates the value of data. Finally, the experiment results on newsvendor and facility location problems show how numerically the optimal objective value of the risk-averse stochastic program converges to the risk-neutral one, verifying the effectiveness of our proposed approach.

Keywords: stochastic program, ζ-structure probability metrics, risk-averse, value of data

1 Introduction

As an effective tool to solve optimization under uncertainty problems, stochastic programming has been widely applied to solve practical problems in energy, finance, production planning, and transportation scheduling, among others [5]. Among this, a specific class of stochastic programs, named two-stage stochastic programs, have been studied extensively. The traditional two-stage stochastic program (denoted as SP) can be described as follows:

$$\begin{aligned} & \min_x \quad c^T x + \mathbb{E}_{\mathbb{P}}[Q(x,\xi)] \\ \text{(SP)} \qquad & s.t. \quad x \in X, \end{aligned} \tag{1}$$

where x is the first-stage decision variable with its feasible region defined as a compact convex set X, $Q(x,\xi) = \min_{y(\xi)\in Y}\{H(y(\xi)) : G(x,y(\xi)) \leq 0\}$ (assumed a continuous function on ξ) represents the second-stage problem, Y represents the feasible region for the second-stage decision variable $y(\xi)$, and the random variable ξ is defined on a probability space $(\Omega, \sigma(\Omega), \mathbb{P})$, where Ω is the sample space for ξ, $\sigma(\Omega)$ is the σ-algebra of Ω, and $\mathbb{P}$ is the associate probability distribution. In this formulation, the probability distribution $\mathbb{P}$ is assumed known and significant research progress has been made in theoretical analysis and developing efficient algorithms such as sample average approximation approaches [35].

However, in practice, due to limited available information on the random parameters, it is generally difficult to obtain the true probability distribution. Instead, only a series of historical data taken from the true distribution can be observed. Although we can derive a distribution for the above (SP) that best fits the given set of historical data, so as to approximate the true distribution, this approach can lead to a final solution suboptimal and away from the true optimal solution, in particular for the case in which the amount of historical data are very limited. To address this issue, risk-averse stochastic optimization approaches allowing distribution ambiguity (also known as distributionally robust optimization) have been investigated recently [11]. In this approach, instead of deriving a unique true distribution for the above (SP), a confidence set $\mathcal{D}$ for the unknown true distribution is derived and accordingly the risk-averse two-stage stochastic program can be formulated as follows (denoted as RA-SP), with the objective of minimizing the

total expected cost under the worst-case distribution in $\mathcal{D}$:

$$\begin{aligned} & \min_x \quad c^T x + \max_{\mathbb{P}\in\mathcal{D}} \mathbb{E}_{\mathbb{P}}[Q(x,\xi)] \\ \text{(RA-SP)} \qquad & s.t. \quad x \in X. \end{aligned} \tag{2}$$

As compared to (SP), (RA-SP) allows the distribution ambiguity and the confidence set $\mathcal{D}$ ensures that the true description is within this set with certain confidence level (say, e.g., 95%). Therefore, (RA-SP) provides tolerance for the unknown probability distribution with the tradeoff of leading to a risk-averse solution which corresponds to a larger objective value.

It can be observed from (RA-SP) that constructing $\mathcal{D}$ is crucial for solving the distributionally robust stochastic program, because different shapes and sizes of $\mathcal{D}$ affect the computational complexity to solve the problem and the robustness of the final solution. One approach to constructing the uncertainty set is based on the moment information [34, 11]. In this approach, the mean values and covariance matrices of the random variables are estimated based on the historical data and the confidence set $\mathcal{D}$ is constructed by restricting the ambiguous distribution satisfying the constraints with the same mean value and covariant matrix. This uncertainty set can also be constructed by allowing the moment ambiguity [11]. That is, $\mathcal{D} = \{\mathbb{P} \in \mathcal{M}_+ : \ \mathbb{E}_{\mathbb{P}}[\xi] \in \mathrm{ES}(\mu_0),\ \mathbb{E}_{\mathbb{P}}[(\xi-\mu_0)(\xi-\mu_0)^\top] \preceq \gamma_2\Sigma_0\}$, where $\mu_0 \in \mathbb{R}^K$ and $\Sigma_0 \in \mathbb{R}^{K\times K}$ are given, $\mathrm{ES}(\mu_0)$ represents an ellipsoid centered at μ_0, and $\mathcal{M}_+$ represents the space of all probability distributions on $(\Omega, \sigma(\Omega))$. The related research works on moment-based confidence sets are described in [33], [41], [43], [44], and [2], among others. Along this direction, eventually the problem can be reformulated as a conic or semidefinite program and efficient algorithms can be developed accordingly. This approach can guarantee the system robustness by deriving the confidence level guarantee for the ambiguous distribution to be within the uncertainty set $\mathcal{D}$. However, the conservatism could not be vanished, because the mean value and covariance matrix can not uniquely define a probability distribution. Even for the case in which $\mathcal{D} = \{\mathbb{P} \in \mathcal{M}_+ : \ \mathbb{E}_{\mathbb{P}}[\xi] = \gamma_1 \text{ and } \mathbb{E}_{\mathbb{P}}[\xi^2] = \gamma_2\}$, where γ_1 and γ_2 are estimated first and second-moment values, there are still a significant amount of distributions in $\mathcal{D}$. The objective value of (RA-SP) could not converges to that of (SP).

The other approach to constructing the confidence set is based on the density (or distribution) information [8, 25]. With a reference distribution $\mathbb{P}_0$ determined by the historical data (e.g., the

empirical distribution function) and a predefined distance measure $d(\mathbb{P}_0, \mathbb{P})$ to measure the distance between $\mathbb{P}_0$ and the ambiguous distribution $\mathbb{P}$, the confidence set $\mathcal{D}$ can be represented as

$$\mathcal{D} = \{\mathbb{P} : d(\mathbb{P}, \mathbb{P}_0) \leq \theta\}, \tag{3}$$

where the tolerance θ is dependent on the size of historical data observed. Intuitively, $\mathcal{D}$ gets tighter around the true distribution $\mathbb{P}$ with more historical data observed. That is, θ becomes smaller and hopefully achieves zero eventually, and accordingly (RA-SP) becomes less conservative and achieves risk-neutral (i.e., (SP)) eventually. There has been significant research progress made by using the data-driven approach to constructing the confidence set $\mathcal{D}$. For instance, in [3], ϕ-divergences are studied to measure the distance between the empirical distribution and the true distribution $d(\mathbb{P}, \mathbb{P}_0)$. Under this setting, the distance is in the form $d_\phi(\mathbb{P}, \mathbb{P}_0) := \int_\Omega \phi(f(\xi)/f_0(\xi))f_0(\xi)d\xi$ with reference density function f_0 corresponding to $\mathbb{P}_0$ and true density function f corresponding to $\mathbb{P}$. By using this distance measure, the single-stage distributionally robust stochastic program can be successfully reformulated to be a tractable problem. This approach has also been extended to the two-stage case in [27] and the chance constrained case in [23]. Besides these works, it is also well-known that ϕ-divergences, especially KL-divergence $d_{\text{KL}}(\mathbb{P}, \mathbb{P}_0) := \int_\Omega \log(f(\xi)/f_0(\xi))f_0(\xi)d\xi$, has been widely applied in machine learning and information theory, e.g., [29] and [38].

Nevertheless, ϕ-divergences are in general not metrics because it can be observed from their definitions that most ϕ-divergences do not satisfy the triangle inequality and even the symmetric property $d_\phi(\mathbb{P}, \mathbb{P}_0) = d_\phi(\mathbb{P}_0, \mathbb{P})$. Meanwhile, although ϕ-divergences converge to χ^2 distributions for the discrete distribution cases [3], as indicated in [14] and the metrics' relationships in [18], there are no convergence results for general ϕ-divergences for the case in which the true density function f is continuous and an empirical probability distribution is selected as $\mathbb{P}_0$.

In this paper, we follow the direction of constructing the confidence set based on the density (or distribution) information. We ensure the distance measures to be metrics. More specifically, we study a family of metrics, named ζ-structure probability metrics, to help construct the confidence set $\mathcal{D}$. Given two probability distributions $\mathbb{P}$ and $\mathbb{Q}$, the ζ-structure probability metrics are defined as follows:

$$d_\zeta(\mathbb{P}, \mathbb{Q}) = \sup_{h \in \mathcal{H}} \left| \int_\Omega h d\mathbb{P} - \int_\Omega h d\mathbb{Q} \right|,$$

where $\mathcal{H}$ is a family of real-valued bounded measurable functions on Ω. The ζ-structure probability metrics family is first introduced in [42], and it has been applied in information theory [20], mathematical statistics [37], mass transportation problems [32], and several areas in computer science, including probabilistic concurrency, image retrieval, data mining, and bioinformatics [13]. The members in the ζ-structure probability metrics family are Kantorovich metric, Fortet-Mourier metric, Total Variation metric, Bounded Lipschitz metric, and Kolmogorov/Uniform metric. These metrics will be described in details in Section 2. The discrete distribution case for a family member of this family of probability metrics, Kantorovich metric, is utilized early to solve the portfolio selection [31] problem. In addition, Prohorov metric is studied in [17] to solve ambiguous chance constrained problems. In our approach, we utilize ζ-structure probability metrics to construct the confidence set for the probability distribution in (RA-SP) (see, e.g., (2)). Then, we develop algorithms to solve (RA-SP) for both discrete and continuous true distribution cases. Finally, we show that (RA-SP) converges to (SP) (see, e.g., (1)) as the size of historical data increases to infinity and further explore the "value of data" by deriving the convergence rate and conducting numerical experiments. Our contributions can be summarized as follows:

1. We study a new family of metrics, ζ-structure probability metrics, to construct the confidence set $\mathcal{D}$ for the ambiguous distributions and explore the relationships among the members in the ζ-structure probability metrics family.

2. We develop a solution framework to solve (RA-SP), corresponding to a given size of historical data, for both discrete and continuous distribution cases. For the case in which the true distribution $\mathbb{P}$ is discrete, we can reformulate (RA-SP) as a traditional two-stage robust optimization problem. For the case in which $\mathbb{P}$ is continuous, we propose a sampling approach, which uses discrete distributions to approximate the continuous distribution, to solving (RA-SP).

3. We perform the convergence analysis and derive the convergence rates for both outer and inner loops, where the outer loop shows the relationship between the conservatism of (RA-SP) and the size of historical data, and the inner loop describes the solution algorithms to solve each specific (RA-SP) once the size of the historical data set is given. More specifically, for the discrete case, we can prove that as the size of historical data increases to infinity, (RA-

SP) converges to (SP) exponentially fast (the outer loop). Meanwhile, for a given historical data set (the inner loop), (RA-SP) can be reformulated as a traditional two-stage robust optimization problem. For the continuous case, we can prove that the outer loop converges to the risk-neutral case uniformly. Meanwhile, in the inner loop for a given sample size, our solution method provides the lower and upper bounds of the optimal objective value of (RA-SP), and these bounds converge to the optimal objective value of (RA-SP) uniformly as the sample size goes to infinity.

The remaining part of this chapter is organized as follows: In Section 2, we introduce the ζ-structure probability metrics family and study the relationships among the members in this family. Since our developed algorithm to solve the continuous case requires the solution techniques to solve the discrete case, we first develop a solution approach including deriving convergence rates to solve (RA-SP) for the discrete case in Section 3. Then, in Section 4, we develop a solution framework to solve the continuous true distribution case, and accordingly explore the convergence rates. In Section 5, we perform numerical studies on data-driven risk-averse newsvendor and facility location problems. Finally, in Section 6, we conclude our research.

2 ζ-Structure Probability Metrics

In this section, we introduce the family of ζ-structure probability metrics. We first provide the definitions of each member of this family: Kantorovich metric, Fortet-Mourier metric, Total Variation metric, Bounded Lipschitz metric, and Kolmogorov/Uniform metric. Then, we investigate the relationships among these members. We show that if the supporting space Ω is bounded, the Total Variation metric is a dominating metric in the family, i.e., the convergence of the Total Variation metric can guarantee those of other metrics in the family.

2.1 Definition

As described in Section 1, for any two probability distributions $\mathbb{P}$ and $\mathbb{Q}$, the ζ-structure probability metrics are defined by $d_\zeta(\mathbb{P},\mathbb{Q}) := \sup_{h\in\mathcal{H}} |\int_\Omega h d\mathbb{P} - \int_\Omega h d\mathbb{Q}|$, where $\mathcal{H}$ is a family of real-valued bounded measurable functions on Ω. In general, the ζ-structure metrics satisfy the properties of metrics, i.e., $d_\zeta(\mathbb{P},\mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$, $d_\zeta(\mathbb{P},\mathbb{Q}) = d_\zeta(\mathbb{Q},\mathbb{P})$ (symmetric property), and

$d_\zeta(\mathbb{P},\mathbb{Q}) \leq d_\zeta(\mathbb{P},\mathbb{O}) + d_\zeta(\mathbb{O},\mathbb{Q})$ for any probability distribution $\mathbb{O}$ (triangle inequality). In the following, we define $\rho(x,y)$ as the distance between two random variables x and y and n as the dimension of Ω. If a random variable x follows distribution $\mathbb{P}$, we denote it as $\mathbb{P} = \mathcal{L}(x)$. Then, we derive different types of metrics in this family based on variant definitions of $\mathcal{H}$.

- **Kantorovich metric**: For Kantorovich metric (denoted as $d_{\mathrm{K}}(\mathbb{P},\mathbb{Q})$), $\mathcal{H} = \{h : \|h\|_{\mathrm{L}} \leq 1\}$, where $\|h\|_{\mathrm{L}} := \sup\{(h(x) - h(y))/\rho(x,y) : x \neq y \text{ in } \Omega\}$. Many metrics known in statistics, measure theory, ergodic theory, and functional analysis, are special cases of the Kantorovich metric [39]. Kantorovich metric also has many applications in transportation theory [32] and in computer science (including probabilistic concurrency, image retrieval, data mining, and bioinformatics [13]).

- **Fortet-Mourier metric**: For Fortet-Mourier metric (denoted as $d_{\mathrm{FM}}(\mathbb{P},\mathbb{Q})$), $\mathcal{H} = \{h : \|h\|_{\mathrm{C}} \leq 1\}$, where $\|h\|_{\mathrm{C}} := \sup\{(h(x) - h(y))/c(x,y) : x \neq y \text{ in } \Omega\}$ and $c(x,y) = \rho(x,y)\max\{1, \rho(x,a)^{p-1}, \rho(y,a)^{p-1}\}$ for some $p \geq 1$ and $a \in \Omega$. Note here that when $p = 1$, Fortet-Mourier metric is the same as Kantorovich metric. Therefore, Fortet-mourier metric is usually utilized as a generalization of Kantorovich metric, with the applications on mass transportation problems [32].

- **Total Variation metric**: For Total Variation metric (denoted as $d_{\mathrm{TV}}(\mathbb{P},\mathbb{Q})$), $\mathcal{H} = \{h : \|h\|_{\infty} \leq 1\}$, where $\|h\|_{\infty} := \sup_{x\in\Omega} |h(x)|$. Another equivalent definition of the Total Variation metric is $d_{\mathrm{TV}}(\mathbb{P},\mathbb{Q}) := 2\sup_{B\in\sigma(\Omega)} |\mathbb{P}(B) - \mathbb{Q}(B)|$, for which we use often in the later analysis. The total variation metric has a coupling characterization (detailed proofs are shown in [26]):

  $$d_{\mathrm{TV}}(\mathbb{P},\mathbb{Q}) = 2\inf\{\mathbb{P}\mathrm{r}(X \neq Y) : \mathbb{P} = \mathcal{L}(X), \mathbb{Q} = \mathcal{L}(Y)\}.$$

  The total variation metric can be applied in information theory [9] and in studying the ergodicity of Markov Chains [28]. Later on, we will prove that the convergence with respect to the Total Variation metric implies the convergences with respect to other metrics in the general ζ-structure probability metrics family when Ω is bounded.

- **Bounded Lipschitz metric**: For Bounded Lipschitz metric (denoted as $d_{\mathrm{BL}}(\mathbb{P},\mathbb{Q})$), $\mathcal{H} = \{h : \|h\|_{\mathrm{BL}} \leq 1\}$, where $\|h\|_{\mathrm{BL}} := \|h\|_{\mathrm{L}} + \|h\|_{\infty}$. One important application of the Bounded

Lipschitz metric is to prove the convergence of probability measures in weak topology [15].

- **Uniform (Kolmogorov) metric**: For Uniform metric (also called Kolmogorov metric, denoted as $d_{\mathrm{U}}(\mathbb{P},\mathbb{Q})$), $\mathcal{H}=\left\{I_{(-\infty,t]},t\in R^n\right\}$. The Uniform metric is often used in proving the classical central limit theorem, and commonly utilized in the Kolmogorov-Smirnov statistic for hypothesis testing [36]. According to the definition, we have $d_{\mathrm{U}}(\mathbb{P},\mathbb{Q})=\sup_t|\mathbb{P}(x\leq t)-\mathbb{Q}(x\leq t)|$.

- **Wasserstein metric**: Wasserstein metric is defined as $d_{\mathrm{W}}(\mathbb{P},\mathbb{Q}):=\inf_\pi\{E_\pi[\rho(X,Y)]:\mathbb{P}=\mathcal{L}(X),\mathbb{Q}=\mathcal{L}(Y)\}$, where the infimum is taken over all joint distributions π with marginals $\mathbb{P}$ and $\mathbb{Q}$. Although Wasserstein metric is not a member in the general ζ-structure probability metrics family, we list it here because by the Kantorovich-Rubinstein theorem [24], the Kantorovich metric is equivalent to the Wasserstein metric. In particular, when $\Omega=R$,

$$d_{\mathrm{W}}(\mathbb{P},\mathbb{Q})=\int_{-\infty}^{+\infty}|F(x)-G(x)|dx,$$

where $F(x)$ and $G(x)$ are the distribution functions derived from $\mathbb{P}$ and $\mathbb{Q}$ respectively. This conclusion holds following the argument that $\inf_\pi\{E_\pi[\rho(X,Y)]:\mathbb{P}=\mathcal{L}(X),\mathbb{Q}=\mathcal{L}(Y)\}=\int_0^1|F^{-1}(t)-G^{-1}(t)|dt$, as stated in Theorem 6.0.2 in [1] and $\int_0^1|F^{-1}(t)-G^{-1}(t)|dt=\int_{-\infty}^{+\infty}|F(x)-G(x)|dx$. Wasserstein metric also has wide applications on transportation problems [32].

2.2 Relationships Among Metrics

In this subsection, we explore the relationships among the members in the ζ-structure probability metrics family. Based on the relationships, we demonstrate that the Total Variation metric is dominated by other members in the family if the supporting space Ω is bounded. That is, if we can prove the convergence for the Total Variation metric case in our later analysis for the data-driven risk-averse stochastic program, then the convergence results for other members are guaranteed with no smaller convergence rates.

First, we explore the relationships between the Total Variation metric and the Kantorovich (Wasserstein) metric. Meanwhile, we use Wasserstein metric and Kantorovich metric interchange-

ably since they are equivalent. We denote Ø as the diameter of Ω and have the following lemma.

Lemma 1. *The relationships between the Total Variation metric and the Kantorovich (Wasserstein) metric are as follows:*

- *If* Ω *is bounded,* $2d_K(\mathbb{P}, \mathbb{Q}) \leq \text{Ø} \cdot d_{TV}(\mathbb{P}, \mathbb{Q})$*;*
- *If* Ω *is a finite set,* $d_{min} \cdot d_{TV}(\mathbb{P}, \mathbb{Q}) \leq 2d_K(\mathbb{P}, \mathbb{Q})$*, where* $d_{min} = \min_{x \neq y} \rho(x, y)$*.*

Proof. Since Kantorovich and Wasserstein metrics are equivalent, we only need to show the relationship between the Wasserstein metric and the Total Variation metric, for which the detailed proofs are provided in Theorem 4 in [18]. □

Next, we study the relationships among the Bounded Lipschitz metric, the Kantorovich (Wasserstein) metric, and the Total Variation metric. Since $\|h\|_{\text{BL}} := \|h\|_{\text{L}} + \|h\|_{\infty}$, the feasible region of h for the Bounded Lipschitz metric is more restrictive than the one for the Kantorovich (Wasserstein) metric or the one for the Total Variation metric. Therefore we have the following lemma:

Lemma 2. *The relationships among the Bounded Lipschitz metric, the Kantorovich (Wasserstein) metric, and the Total Variation metric are as follows:*

- $d_{BL}(\mathbb{P}, \mathbb{Q}) \leq d_K(\mathbb{P}, \mathbb{Q})$*,*
- $d_{BL}(\mathbb{P}, \mathbb{Q}) \leq d_{TV}(\mathbb{P}, \mathbb{Q})$*.*

Moreover, the relationships between the Total Variation metric and the Uniform metric can be obtained by the definitions of each metric. The Total Variation metric is to find a set B among all the Borel sets in $\sigma(\Omega)$, to maximize the value of $2|\mathbb{P}(B) - \mathbb{Q}(B)|$. But the Uniform metric is to find a set B among all the sets in the form $(-\infty, x]$, to maximize the value of $|\mathbb{P}(B) - \mathbb{Q}(B)|$. Since set $(-\infty, x]$ is a borel set, we have the following conclusion holds.

Lemma 3. *The relationship between the Total Variation metric and the Uniform metric is:* $2d_U(\mathbb{P}, \mathbb{Q}) \leq d_{TV}(\mathbb{P}, \mathbb{Q})$*.*

Finally, we explore the relationships between the Kantorovich (Wasserstein) metric and the Fortet-Mourier metric. We have the following lemma:

Lemma 4. *The relationships between the Kantorovich (Wasserstein) metric and the Fortet-Mourier metric are as follows:*

- $d_K(\mathbb{P}, \mathbb{Q}) \leq d_{FM}(\mathbb{P}, \mathbb{Q})$,
- $d_{FM}(\mathbb{P}, \mathbb{Q}) \leq \Lambda \cdot d_K(\mathbb{P}, \mathbb{Q})$,

where $\Lambda = \max\{1, \mathcal{O}^{p-1}\}$.

Proof. The first statement is obvious following the definitions of the Kantorovich metric and the Fortet-Mourier metric. For the second statement, for the Fortet-Mourier metric, we have $|h(x) - h(y)| \leq c(x, y) = \rho(x, y) \max\{1, \rho(x, a)^{p-1}, \rho(y, a)^{p-1}\} \leq \rho(x, y)\Lambda$, where the equation holds following the definition of $c(x, y)$. Now we can observe

$$\begin{aligned} d_{\mathrm{FM}}(\mathbb{P}, \mathbb{Q}) &\leq \sup_{h:|h(x)-h(y)|\leq\Lambda\cdot\rho(x,y)} \left| \int_\Omega h d\mathbb{P} - \int_\Omega h d\mathbb{Q} \right| \\ &= \Lambda \cdot \sup_{h:|h(x)-h(y)|\leq\Lambda\cdot\rho(x,y)} \left| \int_\Omega h/\Lambda d\mathbb{P} - \int_\Omega h/\Lambda d\mathbb{Q} \right| \\ &= \Lambda \cdot \sup_{g:|g(x)-g(y)|\leq\rho(x,y)} \left| \int_\Omega g d\mathbb{P} - \int_\Omega g d\mathbb{Q} \right| \\ &= \Lambda \cdot d_{\mathrm{W}}(\mathbb{P}, \mathbb{Q}), \end{aligned}$$

where the last equality follows the definition of Kantorovich metric. Then the second statement of Lemma 4 holds. □

To summarize, the relationships among the members in the ζ-structure probability metrics family are shown in Figure 1.

Based on Lemmas 1 to 4, we conclude the following proposition without proof.

Proposition 1. *If the support* Ω *is bounded, the Kantorovich metric, the Fortet-Mourier metric, the Bounded Lipschitz metric, and the Kolmogorov metric are dominated by the Total Variation metric.*

With the above lemmas and proposition, we explore methodologies to solve (RA-SP) (i.e., Formulation (2)) and derive the corresponding convergence rates in the next two sections by considering two cases: (i) the true probability distribution is discrete, and (ii) the true probability distribution is continuous.

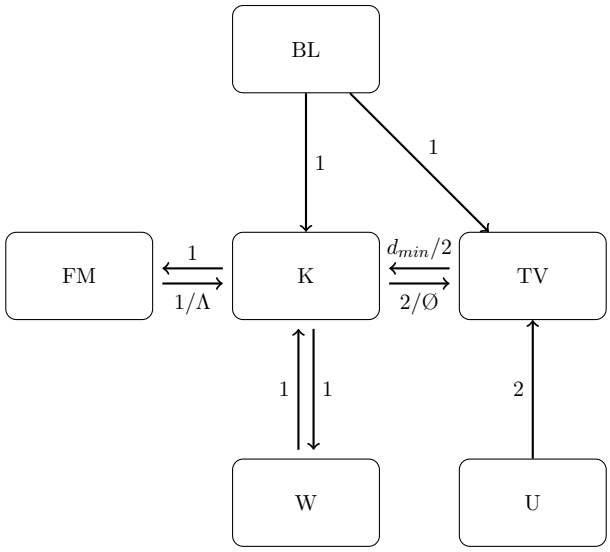

Figure 1: Relationships among the members in the ζ-structure probability metrics family

3 Discrete Case

In this section, we investigate the solution methodologies to solve the data-driven risk-averse two-stage stochastic program (i.e., RA-SP as described in (2)) with ζ-structure probability metrics, for the case in which the true distribution is discrete. For analysis convenience, we assume the supporting space Ω is bounded, which is common for a variety of practical problems. This section is organized by answering the following important questions:

(1) How to determine the reference distribution $\mathbb{P}_0$ as described in (3)?

(2) How to represent the value of θ depending on the amount of historical data, i.e., the convergence rate?

(3) How to solve the problem with respect to different ζ-structure probability metrics?

In our study, we first consider the case in which the supporting space Ω is finite (e.g., possible scenarios are denoted as $\xi^1, \xi^2, \cdots$, and ξ^N). For the case in which the supporting space Ω is infinite, we can apply the solution framework for the continuous probability distribution case, for which we will discuss later in Section 4.

3.1 Reference Distribution

Given M historical data $\xi_0^1, \xi_0^2, \cdots, \xi_0^M$, to estimate the reference distribution $\mathbb{P}_0$, we consider the empirical distribution of the historical data samples, i.e., the cumulative distribution function (cdf) that puts mass $1/M$ at each data point $\xi_0^i, i = 1, \cdots, M$. Formally, the empirical distribution is defined as

$$\mathbb{P}_0(x) = \frac{1}{M} \sum_{i=1}^{M} \delta_{\xi_0^i}(x),$$

where $\delta_{\xi_0^i}(x)$ is an indicator variable equal to 1 if $\xi_0^i \leq x$ and 0 otherwise. In this case, since the supporting space is discrete, the reference distribution $\mathbb{P}_0$ can be represented by its mass probability $p_0^1, p_0^2, \cdots, p_0^N$, where p_0^i is equal to the ratio between the number of historical data samples matching ξ^i and M.

3.2 Convergence Rate Analysis

After identifying the reference distribution $\mathbb{P}_0$, we discuss the value of θ, i.e., the convergence rate of the empirical distribution to the true distribution for the discrete case. We first study the convergence rate of the Total Variation metric. Then we explore the convergence rates of other metrics in the family.

For the Total Variation metric, if the true distribution is discrete, Pinsker's inequality [10] shows $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{P}_0) \leq \sqrt{d_{\mathrm{KL}}(\mathbb{P}, \mathbb{P}_0)}$, where $d_{\mathrm{KL}}(\mathbb{P}, \mathbb{P}_0)$ is the discrete case KL-divergence defined as $\sum_i \ln\left(p_i^0/p_i\right) p_i^0$. Since it is shown in [30] that $d_{\mathrm{KL}}(\mathbb{P}, \mathbb{P}_0)$ converges in probability to a χ^2 distributed random variable as the number of historical data samples M goes to infinity, we can claim that the convergence rate of $d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P})$ is bounded by that of a χ^2 distributed random variable as $M \to +\infty$. Meanwhile, note here that for the case in which the true distribution is continuous, the Total Variation metric does not converge as described in [14] if the empirical distribution is selected as the reference distribution $\mathbb{P}_0$, which is also the reason why the KL-divergence does not converge for the continuous case as indicated in Section 1.

For the Uniform (Kolmogorov) metric, the convergence property and the convergence rate can be obtained by utilizing the following Dvoretzky-Kiefer-Wolfowitz inequality.

Proposition 2 (Dvoretzky-Kiefer-Wolfowitz inequality [16])**.** *For a single dimension case (i.e.,*

$n = 1$), $\mathbb{P}(d_U(\mathbb{P}_0, \mathbb{P}) \leq \epsilon) \geq 1 - 2e^{-2M\epsilon^2}$. *For a high dimension case (i.e., $n > 1$), for any $\alpha > 0$, there exists a constant number C_α such that* $\mathbb{P}(d_U(\mathbb{P}_0, \mathbb{P}) \leq \epsilon) \geq 1 - C_\alpha e^{-(2-\alpha)M\epsilon^2}$.

Now we prove the convergence result for the Kantorovich (Wasserstein) metric. We can obtain the following conclusion.

Proposition 3. *For a general dimension case (i.e., $n \geq 1$), we have*

$$P(d_K(\mathbb{P}_0, \mathbb{P}) \leq \epsilon) \geq 1 - \exp\left(-\frac{\epsilon^2}{2\emptyset^2} M\right).$$

Proof. Let us define set

$$\mathcal{B} := \{\mu \in \mathcal{P}(\Omega) : d_{\mathrm{K}}(\mu, \mathbb{P}) \geq \epsilon\} \tag{4}$$

where $\mathcal{P}(\Omega)$ is the set of all probability measures defined on Ω. Define set $\mathcal{C}(\Omega)$ be the collection of all bounded continuous functions ϕ: $\Omega \to R$. Following the definitions, for each $\phi \in \mathcal{C}(\Omega)$, we have

$$\begin{aligned}
\mathbb{P}(d_{\mathrm{K}}(\mathbb{P}, \mathbb{P}_0) \geq \epsilon) &= Pr(\mathbb{P}_0 \in \mathcal{B}) && (5)\\
&\leq Pr\left(\int_\Omega \phi d\mathbb{P}_0 \geq \inf_{\mu\in\mathcal{B}} \int_\Omega \phi d\mu\right) && (6)\\
&\leq \exp\left(-M \inf_{\mu\in\mathcal{B}} \int_\Omega \phi d\mu\right) E\left(e^{M\int_\Omega \phi d\mathbb{P}_0}\right) && (7)\\
&= \exp\left(-M \inf_{\mu\in\mathcal{B}} \left\{\int_\Omega \phi d\mu - \frac{1}{M}\log E\left(e^{M\int_\Omega \phi d\mathbb{P}_0}\right)\right\}\right)\\
&= \exp\left(-M \inf_{\mu\in\mathcal{B}} \left\{\int_\Omega \phi d\mu - \frac{1}{M}\log E\left(e^{\sum_{i=1}^M \phi(\xi^i)}\right)\right\}\right) && (8)\\
&= \exp\left(-M \inf_{\mu\in\mathcal{B}} \left\{\int_\Omega \phi d\mu - \log \int_\Omega e^\phi d\mathbb{P}\right\}\right), && (9)
\end{aligned}$$

where inequality (5) follows the definition of $\mathcal{B}$, inequality (6) follows from the fact that $\mathbb{P}_0 \in \mathcal{B}$ (see (4)) and μ is the one in $\mathcal{B}$ that achieves the minimum of $\int_\Omega \phi d\mu$, inequality (7) follows from the Chebyshev's exponential inequality [21], equation (8) follows from the definition of $\mathbb{P}_0$, and equation (9) follows from the assumption that the historical data are independently drawn from the true distribution $\mathbb{P}$.

Now we define $\Delta(\mu) := \sup_{\phi\in\mathcal{C}(\Omega)} \int_\Omega \phi d\mu - \log \int_\Omega e^\phi d\mathbb{P}$. Following the continuity and boundedness from the definition of $\mathcal{C}(\Omega)$, there exists a series ϕ_n such that $\lim_{n\to\infty} \int_\Omega \phi_n d\mu - \log \int_\Omega e^{\phi_n} d\mathbb{P} = \Delta(\mu)$. Therefore, for any small positive number $\epsilon' > 0$, there exists a constant number n_0 such that

$\Delta(\mu) - (\int_\Omega \phi_n d\mu - \log \int_\Omega e^{\phi_n} d\mathbb{P}) \leq \epsilon'$ for any $n \geq n_0$. Therefore, according to (9) by substituting ϕ with ϕ_n, we have

$$\begin{aligned} Pr(\mathbb{P}_0 \in \mathcal{B}) &\leq \exp\left(-M \inf_{\mu \in \mathcal{B}} \left\{ \int_\Omega \phi_n d\mu - \log \int_\Omega e^{\phi_n} d\mathbb{P} \right\}\right) \\ &\leq \exp\left(-M \inf_{\mu \in \mathcal{B}} (\Delta(\mu) - \epsilon')\right). \end{aligned} \tag{10}$$

According to Lemma 6.2.13 in [12], we have

$$\Delta(\mu) = d_{\mathrm{KL}}(\mu, \mathbb{P}). \tag{11}$$

For the case $\mu \in \mathcal{B}$, following (4), we have $d_{\mathrm{K}}(\mu, \mathbb{P}) \geq \epsilon$. In addition, it is shown in "Particular case 5" in [7] that

$$d_{\mathrm{K}}(\mu, \mathbb{P}) \leq \text{Ø}\sqrt{2 d_{\mathrm{KL}}(\mu, \mathbb{P})} \tag{12}$$

holds for $\forall \mu \in \mathcal{P}(\Omega)$. Therefore, following (12) we have

$$d_{\mathrm{KL}}(\mu, \mathbb{P}) \geq \epsilon^2/(2\text{Ø}^2). \tag{13}$$

Therefore, combining (10), (11), and (13), we have

$$Pr(\mathbb{P}_0 \in \mathcal{B}) \leq \exp\left(-M\left(\frac{\epsilon^2}{2\text{Ø}^2} - \epsilon'\right)\right).$$

Let $\epsilon' = \delta/M$ for any arbitrary small positive δ. Then, we have

$$\mathbb{P}(d_{\mathrm{K}}(\mathbb{P}, \mathbb{P}_0) \geq \epsilon) = Pr(\mathbb{P}_0 \in \mathcal{B}) \leq \exp\left(-\frac{\epsilon^2}{2\text{Ø}^2} M + \delta\right).$$

Since δ can be arbitrarily small, we have $P(d_{\mathrm{K}}(\mathbb{P}_0, \mathbb{P}) \leq \epsilon) \geq 1 - \exp(-\epsilon^2 M/(2\text{Ø}^2))$. □

We then can immediately obtain the convergence rates as follows for the Fortet-Mourier and Bounded Lipschitz metrics, following the relationships among the ζ-structure probability metrics as described in Figure 1 in Subsection 2.2.

Corollary 1. *For a general $n \geq 1$, we have $P(d_{FM}(\mathbb{P}_0, \mathbb{P}) \leq \epsilon) \geq 1 - \exp(-\epsilon^2 M/(2\text{Ø}^2\Lambda^2))$ and*

$\mathbb{P}(d_{BL}(\mathbb{P}_0, \mathbb{P}) \leq \epsilon) \geq 1 - \exp(-\epsilon^2 M/(2\emptyset^2))$.

With the convergence rates, we can calculate the value of θ accordingly. For instance, let us take the convergence rate for the Kantorovich (Wasserstein) metric obtained in Proposition 3 as an example to illustrate the process. Assuming the confidence level is set to be η, i.e., $P(d_{\mathrm{K}}(\mathbb{P}_0, \mathbb{P}) \leq \theta) \geq 1 - \exp(-\theta^2 M/(2\emptyset^2)) = \eta$, we can obtain $\theta = \emptyset\sqrt{2\log(1/(1-\eta))/M}$. Similarly, we can calculate the value of θ for different metrics based on Corollary 1.

Given that the reference distribution converges to the true distribution exponentially fast, next, we explore the convergence property of the optimal solution and the optimal objective value of (RA-SP) to those of SP (i.e., Formulation (1)). We have the following conclusion.

Theorem 1. *For the case in which the true distribution is discrete, the optimal solution and the optimal objective value of (RA-SP) (i.e., Formulation (2)), converge to those of SP (i.e., Formulation (1)), respectively.*

Proof. According to Proposition 3 and Corollary 1, we know as the historical data size M goes to infinity, the true distribution $\mathbb{P}^*$ converges to $\mathbb{P}_0$ in probability. Since as M goes to infinity, θ goes to zero, so the worst-case probability distribution, denoted as $\mathbb{P}'_{\mathrm{M}}$, converges to $\mathbb{P}_0$. Therefore, the worst-case distribution $\mathbb{P}'_{\mathrm{M}}$ converges to the true distribution $\mathbb{P}^*$ in probability. Due to the assumption that the function $Q(x, \xi)$ is bounded and continuous with ξ, according to the Helly-Bray theorem [4], we can claim that for any given $x \in X$,

$$\lim_{M\to\infty} \sup_{\mathbb{P}: d(\mathbb{P}, \mathbb{P}_0) \leq \theta_M} \mathbb{E}_{\mathbb{P}}[Q(x, \xi)] = \lim_{M\to\infty} \mathbb{E}_{\mathbb{P}'_{\mathrm{M}}}[Q(x, \xi)] = \mathbb{E}_{\mathbb{P}^*}[Q(x, \xi)]. \tag{14}$$

With equation (14), we first explore the convergence property of the optimal objective values. We represent $\hat{v}$ as the optimal objective value and $\hat{x}$ as the optimal solution of the following problem:

$$\begin{aligned} \min_x \quad & c^T x + \lim_{M\to\infty} \sup_{\mathbb{P}: d(\mathbb{P}, \mathbb{P}_0) \leq \theta_M} \mathbb{E}_{\mathbb{P}}[Q(x, \xi)] \\ s.t. \quad & x \in X. \end{aligned} \tag{15}$$

Besides, we represent $\bar{v}$ as the optimal objective value and $\bar{x}$ as the optimal solution of the risk-neutral problem SP (i.e., the problem as described in (1)). Then we need to prove that $\hat{v} = \bar{v}$ by

a contradiction method. Due to the fact that $\hat{v} \geq \bar{v}$, if the equation $\hat{v} = \bar{v}$ does not hold, we have $\hat{v} > \bar{v}$. According to equation (14), we can observe that

$$c^T \bar{x} + \lim_{M \to \infty} \sup_{\mathbb{P}: d(\mathbb{P}, \mathbb{P}_0) \leq \theta_M} \mathbb{E}_{\mathbb{P}}[Q(\bar{x}, \xi)] = c^T \bar{x} + \mathbb{E}_{\mathbb{P}^*}[Q(\bar{x}, \xi)].$$

Therefore,

$$\begin{aligned} & c^T \hat{x} + \lim_{M \to \infty} \sup_{\mathbb{P}: d(\mathbb{P}, \mathbb{P}_0) \leq \theta_M} \mathbb{E}_{\mathbb{P}}[Q(\hat{x}, \xi)] \\ & = \hat{v} > \bar{v} = c^T \bar{x} + \mathbb{E}_{\mathbb{P}^*}[Q(\bar{x}, \xi)] \\ & = c^T \bar{x} + \lim_{M \to \infty} \sup_{\mathbb{P}: d(\mathbb{P}, \mathbb{P}_0) \leq \theta_M} \mathbb{E}_{\mathbb{P}}[Q(\bar{x}, \xi)], \end{aligned}$$

which violates that $\hat{x}$ is the optimal solution to problem (15). Consequently, we have $\hat{v} = \bar{v}$, which indicates the convergence property of the optimal objective values. Besides, since X is compact and accordingly the optimal solution of formulation (2) converges to $\hat{x}$, and $\hat{x}$ is also an optimal solution of formulation (1) due to $\hat{v} = \bar{v}$, we can claim that the optimal solution of formulation (2) converges to that of formulation (1). □

3.3 Solution Approach

In this section, we explore the methodology to solve the problem. Assuming $\Omega = \{\xi^1, \xi^2, \cdots, \xi^N\}$, (RA-SP) (i.e., Formulation (2)) can be reformulated as:

$$\begin{aligned} \min_x \quad & c^T x + \max_{p_i} \sum_{i=1}^{N} p_i Q(x, \xi^i) \\ s.t. \quad & \sum_i p_i = 1, \\ & \max_{h_i} \quad \sum_{i=1}^{N} h_i p_i^0 - \sum_{i=1}^{N} h_i p_i \leq \theta, \quad \forall h_i : \|h\|_\zeta \leq 1. \end{aligned} \tag{16}$$

For $\|h\|_\zeta \leq 1$, according to different members in the ζ family, we have

- Kantorovich: $|h_i - h_j| \leq \rho(\xi^i, \xi^j), \quad \forall i, j,$
- Fortet-Mourier: $|h_i - h_j| \leq \rho(\xi^i, \xi^j) \max\{1, \rho(\xi^i, a)^{p-1}, \rho(\xi^j, a)^{p-1}\}, \quad \forall i, j,$

- Bounded-Lipschitz: $|h_i - h_j| \leq \rho(\xi^i, \xi^j), |h_i| \leq 1, \quad \forall i, j,$
- Total Variation: $|h_i| \leq 1, \quad \forall i.$

For the above four metrics, the constraints can be summarized as $\sum_i a_{ij} h_i \leq b_j, j = 1, \cdots, J$. First, we develop the reformulation of constraint (16). Considering the problem

$$\begin{aligned} \max_{h_i} \quad & \sum_{i=1}^{N} h_i p_i^0 - \sum_{i=1}^{N} h_i p_i \\ s.t. \quad & \sum_{i}^{N} a_{ij} h_i \leq b_j, \quad j = 1, \cdots, J, \end{aligned}$$

we can get its dual formulation as follows:

$$\begin{aligned} \min \quad & \sum_{j=1}^{J} b_j u_j \\ s.t. \quad & \sum_{j=1}^{J} a_{ij} u_j \geq p_i^0 - p_i, \forall i = 1, \cdots, N, \end{aligned}$$

where u is the dual variable. Therefore, for the discrete distribution case, (RA-SP) can be reformulated as follows:

$$\text{(FR-M)} \quad \begin{aligned} \min_x \quad & c^T x + \max_{p_i} \sum_{i=1}^{N} p_i G(x, \xi^i) \\ s.t. \quad & \sum_{i=1}^{N} p_i = 1, \quad \sum_{j=1}^{J} b_j u_j \leq \theta, \\ & \sum_{j=1}^{J} a_{ij} u_j \geq p_i^0 - p_i, \forall i = 1, \cdots, N. \end{aligned}$$

For the uniform metric, we can obtain the reformulation directly from the definition of Uniform metric:

$$\begin{aligned} & \min \quad c^T x + \max_{p_i} \sum_{i=1}^{N} p_i G(x, \xi^i) \\ \text{(FR-U)} \quad & s.t. \quad \sum_i p_i = 1, \\ & \left| \sum_{i=1}^{j} (p_i^0 - p_i) \right| \leq \theta, \forall j = 1, \cdots, J. \end{aligned}$$

Next, we summarize the algorithm to solve the case in which the true distribution is discrete.

Algorithm 1: Algorithm for the discrete case

Input: Historical data $\xi^1, \xi^2, \cdots, \xi^N$ i.i.d. drawn from the true distribution. The confidence level of set $\mathcal{D}$ is set to be η.

Output: The objective value of the risk-averse problem (RA-SP).

1 Obtain the reference distribution $\mathbb{P}_0(x) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi^i}(x)$ and the value of θ based on the historical data.

2 Use the reformulation (FR-M) or (FR-U) to solve the problem.

3 Output the solution.

4 Continuous Case

In this section, we discuss the case in which the true distribution is continuous. We accordingly develop the methodologies to solve the continuous case by answering the three questions (1), (2), and (3) in Section 3, respectively.

4.1 Reference Distribution

For the continuous case, we consider using Kernel Density Estimation to construct the reference probability density function (pdf) from the given set of historical data. Assuming there are M historical data samples, e.g., ξ^1,ξ^2,$\cdots$,ξ^M, i.i.d. drawn from the true distribution, the kernel density function is defined as

$$f_{\text{M}}(x) = \frac{1}{M h_{\text{M}}^n} \sum_{i=1}^{M} K\left(\frac{x - \xi^i}{h_{\text{M}}} \right), \tag{17}$$

where n is the dimension of x, h_{M} is the bandwidth, and $K(x) = \prod_{k=1}^{n} k(x_k)$, in which k is a Borel measurable function (kernel) with dimension 1, satisfying $k \geq 0$ and $\int_R k(x)dx = 1$. Note here that $K(x)$ also satisfies $K \geq 0$ and $\int_{R^n} K(x)dx = 1$. By using the kernel density estimation, we can first observe that the reference probability distribution $\mathbb{P}_0$ derived from $f_{\mathrm{M}}(x)$ is continuous.

4.2 Convergence Analysis

By using kernel density estimation, we analyze the convergence properties including the convergence rates of the reference distribution to the true distribution, corresponding to different metrics in the ζ-structure probability metrics family. To analyze the convergence rate of reference density $f_{\mathrm{M}}(x)$ to the true density $f(x)$, we first obtain the convergence rates of $f_{\mathrm{M}}(x)$ to $E[f_{\mathrm{M}}(x)]$, and $E[f_{\mathrm{M}}(x)]$ to $f(x)$. Then following the triangle inequality property of a metric (i.e., $d(f_{\mathrm{M}}(x), f(x)) \leq d(f_{\mathrm{M}}(x), E[f_{\mathrm{M}}(x)]) + d(E[f_{\mathrm{M}}(x)], f(x))$), we can obtain the convergence rate of $f_{\mathrm{M}}(x)$ to $f(x)$. Before describing the main results, we introduce the following commonly used hypothesis on kernel function $K(u)$, true density function $f(x)$, and bandwidth h_{M}.

(H1): The density function $f(x)$ is bounded and uniformly continuous.

(H2): $\nabla^2 f(x)$ (the second differentiation of $f(x)$) is piecewise continuous and square integrable.

(H3): The kernel function K is bounded.

(H4): The kernel function K is symmetric, i.e., $\int_\Omega K(u)udu = 0$.

(H5): The kernel function K is a square integrable function in the linear span of functions $k \geq 0$ such that the subgraph of k, $(s, u) : k(s) \geq u$, can be represented as a finite number of Boolean operations among sets of the form $(s, u) : p(s, u) \geq \phi(u)$, where p is a polynomial on $R^n \times R$ and $\phi(u)$ is an arbitrary real function [19].

(H6): The bandwidth h_{M} satisfies $h_{\mathrm{M}} \to 0$ and $Mh_{\mathrm{M}}^n \to \infty$.

Lemma 5. *[19] Under hypothesis (H1), (H3), (H5) and (H6), we have*

$$\lim_{M\to\infty} \sqrt{\frac{Mh_M^n}{2\log(h_M^{-n})}} \|f_M - E[f_M]\|_\infty = \|K\|_2 \|f\|_\infty^{1/2}, \quad a.s.$$

Lemma 6. *Under hypothesis (H1), (H3), (H5) and (H6), we have*

$$\lim_{M\to\infty}\sqrt{\frac{Mh_M^n}{2\log(h_M^{-n})}}\int_\Omega |f_M(x)-E[f_M(x)]|dx \le V(\Omega)\|K\|_2\|f\|_\infty^{1/2},\quad a.s.,$$

where $V(\Omega)$ is the volume of Ω.

Proof. Since $\int_\Omega |f_{\mathrm{M}}(x)-E[f_{\mathrm{M}}(x)]|dx \le \|f_{\mathrm{M}}-E[f_{\mathrm{M}}]\|_\infty \int_\Omega dx$, Lemma 6 follows from Lemma 5. □

Lemma 7. *Under hypothesis (H2), (H3), and (H4), we have*

$$\int_\Omega |f(x)-E[f_M(x)]|dx \le \frac{1}{2}\sqrt{V(\Omega)}\nu_2(K)h_M^2\|tr\{\nabla^2 f(x)\}\|_2,$$

where $tr\{A\}$ denotes the trace of A and $\nu_2(K)I=\int_\Omega uu^T K(u)du$ (I is an $n\times n$ Identity Matrix).

Proof. As indicated in [40], if hypothesis (H2), (H3), and (H4) are satisfied, then

$$E[f_{\mathrm{M}}(x)]-f(x)=\frac{1}{2}\nu_2(K)h_{\mathrm{M}}^2 tr\{\nabla^2 f(x)\}+o(h_{\mathrm{M}}^2). \tag{18}$$

Note here $o(h_{\mathrm{M}}^2)$ can be omitted for analysis convergence. According to the Hölder's inequality [22], we have

$$\begin{aligned}\int_\Omega |E[f_{\mathrm{M}}(x)]-f(x)|\,dx &\le \sqrt{V(\Omega)\int_\Omega (E[f_{\mathrm{M}}(x)]-f(x))^2dx}\\ &\le \frac{1}{2}\sqrt{V(\Omega)}\nu_2(K)h_{\mathrm{M}}^2\|tr\{\nabla^2 f(x)\}\|_2,\end{aligned} \tag{19}$$

where inequality (19) holds because of (18). □

Now we define

$$\begin{aligned}C_1 &= V(\Omega)\|K\|_2\|f\|_\infty^{1/2},\\ C_2 &= \frac{1}{2}\sqrt{V(\Omega)}\nu_2(K)\|tr\{\nabla^2 f(x)\}\|_2.\end{aligned}$$

According to Lemmas 6 and 7, and the triangle inequality of metrics, we can derive that

$$\begin{aligned}\int_\Omega |f(x) - f_{\mathrm{M}}(x)|dx &\leq \int_\Omega |E[f_{\mathrm{M}}(x)] - f_{\mathrm{M}}(x)|dx + \int_\Omega |f(x) - E[f_{\mathrm{M}}(x)]|dx \\ &\leq C_1\sqrt{\frac{2\log(h_{\mathrm{M}}^{-n})}{Mh_{\mathrm{M}}^n}} + C_2 h_{\mathrm{M}}^2. \end{aligned} \tag{20}$$

Note here that different settings of h_{M} lead to different convergence rates. For instance, if we set $h_{\mathrm{M}} = M^{-\frac{1}{2n}}$, then

$$\int_\Omega |f(x) - f_{\mathrm{M}}(x)|dx \leq C_1\sqrt{\log M}M^{-\frac{1}{4}} + C_2 M^{-\frac{1}{n}}.$$

With the convergence rate of $\int_\Omega |f(x) - f_{\mathrm{M}}(x)|dx$ as described in (20), we can analyze the convergence rates of reference distribution $\mathbb{P}_0$ to the true distribution $\mathbb{P}$ for the members in the ζ-structure probability metrics family.

Proposition 4. *Under hypothesis (H1) to (H6), we have*

$$\begin{aligned} d_{TV}(\mathbb{P}, \mathbb{P}_0) &\leq C_1\sqrt{\frac{2\log(h_M^{-n})}{Mh_M^n}} + C_2 h_M^2, \\ d_K(\mathbb{P}, \mathbb{P}_0) &\leq \frac{C_1\emptyset}{2}\sqrt{\frac{2\log(h_M^{-n})}{Mh_M^n}} + \frac{C_2\emptyset}{2}h_M^2, \\ d_{BL}(\mathbb{P}, \mathbb{P}_0) &\leq C_1\sqrt{\frac{2\log(h_M^{-n})}{Mh_M^n}} + C_2 h_M^2, \\ d_{FM}(\mathbb{P}, \mathbb{P}_0) &\leq \frac{C_1\emptyset\Lambda}{2}\sqrt{\frac{2\log(h_M^{-n})}{Mh_M^n}} + \frac{C_2\emptyset\Lambda}{2}h_M^2, \\ d_U(\mathbb{P}, \mathbb{P}_0) &\leq \frac{C_1}{2}\sqrt{\frac{2\log(h_M^{-n})}{Mh_M^n}} + \frac{C_2}{2}h_M^2. \end{aligned}$$

Proof. We first prove that $\int_\Omega |f(x) - f_{\mathrm{M}}(x)|dx = d_{\mathrm{TV}}(\mathbb{P}, \mathbb{P}_0)$, where $f(x)$ and $f_{\mathrm{M}}(x)$ are the density

functions of $\mathbb{P}$ and $\mathbb{P}_0$ respectively. First, letting $A := \{x \in \Omega : f(x) > f_{\mathrm{M}}(x)\}$, we have

$$\begin{aligned}\int_\Omega |f(x) - f_{\mathrm{M}}(x)|dx &= \int_A |f(x) - f_{\mathrm{M}}(x)|dx + \int_{A^c} |f(x) - f_{\mathrm{M}}(x)|dx \\ &= \int_A (f(x) - f_{\mathrm{M}}(x))dx + \int_{A^c} (f_{\mathrm{M}}(x) - f(x))dx \\ &= 2\int_A (f(x) - f_{\mathrm{M}}(x))dx = 2(\mathbb{P}(A) - \mathbb{P}_0(A)),\end{aligned}$$

where the third equation is due to the fact that $\int_\Omega (f(x) - f_{\mathrm{M}}(x))dx = 0$. That is, we find a set A such that $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{P}_0) = \int_\Omega |f(x) - f_{\mathrm{M}}(x)|dx$.

Second, for any $B \in \sigma(\Omega)$, we have

$$\begin{aligned}|\mathbb{P}(B) - \mathbb{P}_0(B)| &= \left|\int_B f(x)dx - \int_B f_{\mathrm{M}}(x)dx\right| \\ &= \left|\int_{B\cap A} (f(x) - f_{\mathrm{M}}(x))dx + \int_{B\cap A^c} (f(x) - f_{\mathrm{M}}(x))dx\right| \\ &\le \max\left\{\int_{B\cap A} (f(x) - f_{\mathrm{M}}(x))dx, \int_{B\cap A^c} (f_{\mathrm{M}}(x) - f(x))dx\right\} \\ &\le \max\left\{\int_A (f(x) - f_{\mathrm{M}}(x))dx, \int_{A^c} (f_{\mathrm{M}}(x) - f(x))dx\right\} \\ &= \frac{1}{2}\int |f(x) - f_{\mathrm{M}}(x)|dx.\end{aligned}$$

Therefore, we have $\int_\Omega |f(x) - f_{\mathrm{M}}(x)|dx = d_{\mathrm{TV}}(\mathbb{P}, \mathbb{P}_0)$. Thus the convergence rate of $\int_\Omega |f(x) - f_{\mathrm{M}}(x)|dx$ as shown in (20) can be directly applied to $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{P}_0)$. In addition, based on the metrics' relationships provided in Lemmas 1 to 4, we can derive the convergence rates of other metrics in the ζ-structure probability metrics family. □

For the convergence property of the optimal solution and the optimal objective value of (RA-SP) to those of SP (i.e., Formulation (1)), we can prove the following theorem in the same way as that for Theorem 1. The proof is omitted here.

Theorem 2. *For the case in which the true distribution is continuous, the optimal solution and the optimal objective value of (RA-SP) (i.e., Formulation (2)), converge to those of SP (i.e., Formulation (1)), respectively.*

4.3 Solution Approach

For the case in which the true probability distribution is continuous, we derive a sampling approximation approach to solving (RA-SP) (i.e., Formulation (2)) for a fixed value θ (θ can be calculated following Theorem 4) for any member in the ζ-structure probability metrics family. We denote

$$f(x) = c^T x + \sup_{\mathbb{P}: d_\zeta(\mathbb{P}, \mathbb{P}_0) \leq \theta} \mathbb{E}_{\mathbb{P}}[Q(x, \xi)], \tag{21}$$

where $\mathbb{P}_0$ is the reference probability distribution derived from f_{M} as indicated in (17). To make the solution framework for (21) more general, we assume that $\mathbb{P}$ can be any general distribution (discrete, continuous, or mixed). In general, this provides a risk-averse solution for the case in which $\mathbb{P}$ is limited to be continuous. Meanwhile, as $\theta \to 0$, the problem (21) converges to the risk-neutral one (i.e., Formulation (1)).

Before describing the main results, we first construct a discrete distribution counterpart for a general distribution $\mathbb{W}$. Since Ω is bounded, we can find a n-dimensional hypercube H that contains Ω with equal edge length ℓ. We can partition H into S^n equal-volume hypercubes $B_1, B_2, \cdots, B_{S^n}$ with edge length of each hypercube equal to ℓ/S. Let $v = \int_{B_i} dx$ be the volume of each hypercube, then we have $v = (\ell/S)^n$, and the volume of H: $V(H) = vS^n$. In addition, we can define $\rho(x, y) = \|x - y\|_2$. Then for any B_i and for any $x, y \in B_i$, following triangular inequality, we have

$$\rho(x, y) \leq \sqrt{n}\ell/S. \tag{22}$$

Now we construct a discrete distribution $\mathbb{W}_S$ corresponding to the continuous distribution $\mathbb{W}$ in the following way. For each hypercube B_i, select a sample $\xi_i \in B_i$ (for instance, the center of B_i), and let $\mathbb{W}_S$ be the discrete distribution on $\{\xi_1, \xi_2, \cdots, \xi_{S^n}\}$ with the probability p_i on ξ_i equal to $\int_{B_i} d\mathbb{W}$, for $\forall i = 1, 2, \cdots, S^n$.

We now first analyze the convergence rate of $\mathbb{W}_S$ to $\mathbb{W}$ under Kantorovich metric.

Lemma 8. *For a general dimension $n \geq 1$, we have*

$$d_K(\mathbb{W}_S, \mathbb{W}) \leq \sqrt{n}\ell/S.$$

Proof. According to the definition of Kantorovich metric, we have $d_{\mathrm{K}}(\mathbb{W}_S, \mathbb{W}) = \sup_{h\in\mathcal{H}} |\int_H h d\mathbb{W}_S - \int_H h d\mathbb{W}|$, with $|h(x) - h(y)|/\rho(x, y) \leq 1, \forall x, y$. Then for each $h \in \mathcal{H}$, we have

$$\begin{aligned}
\left|\int_H h d\mathbb{W}_S - \int_H h d\mathbb{W}\right| &= \left|\sum_{i=1}^{S^d} h(\xi_i)p_i - \sum_{i=1}^{S^d}\int_{B_i} h(\xi) d\mathbb{W}\right| \\
&= \left|\sum_{i=1}^{S^d}\int_{B_i} (h(\xi_i) - h(\xi)) d\mathbb{W}\right| \qquad (23)\\
&\leq \sum_{i=1}^{S^d}\int_{B_i} |h(\xi_i) - h(\xi)| d\mathbb{W} \\
&\leq \sum_{i=1}^{S^d}\int_{B_i} \rho(\xi_i, \xi) d\mathbb{W} \qquad (24)\\
&\leq \sup_{x,y\in B_i} \rho(x, y) \sum_{i=1}^{S^d}\int_{B_i} d\mathbb{W} \\
&= \sup_{x,y\in B_i} \rho(x, y) \\
&\leq \sqrt{n}\ell/S, \qquad (25)
\end{aligned}$$

where (23) follows the definition of p_i (i.e., $p_i = \int_{B_i} d\mathbb{W}$), (24) follows the property $|h(x) - h(y)|/\rho(x, y) \leq 1, \forall x, y$ for the Kantorovich metric, and (25) follows inequality (22). Therefore, we have $d_{\mathrm{K}}(\mathbb{W}_S, \mathbb{W}) = \sup_{h\in\mathcal{H}} |\int_H h d\mathbb{W}_S - \int_H h d\mathbb{W}| \leq \sqrt{n}\ell/S$. □

For the Bounded Lipschetz and Fortet-Mourier metrics in the ζ-structure probability metrics family, the convergence rates, which can be derived from Lemmas 2 and 4, are shown in the following corollary.

Corollary 2. *For a general dimension $n \geq 1$, we have*

$$\begin{aligned}
d_{BL}(\mathbb{W}_S, \mathbb{W}) &\leq \sqrt{n}\ell/S, \\
d_{FM}(\mathbb{W}_S, \mathbb{W}) &\leq \Lambda\sqrt{n}\ell/S.
\end{aligned}$$

For the Total Variation and Uniform metrics, the supporting space Ω of distribution $\mathbb{W}$ (which is assumed to be continuous in this section) is not a finite set. Accordingly, Lemmas 1 and 3 cannot be applied to derive the corresponding convergence rates. In the following, we study the convergence properties separately.

For the Total Variation metric, since $\mathbb{W}_S$ has a finite support, we define the corresponding sample space $\Omega_S = \{\xi_1, \xi_2, \cdots, \xi_{S^n}\}$. Then, based on the definition of Total Variation metric, we have

$$\begin{aligned} d_{\text{TV}}(\mathbb{W}_S, \mathbb{W}) &= 2 \sup_{B \in \sigma(\Omega)} |\mathbb{W}_S(B) - \mathbb{W}(B)| \\ &\geq 2|\mathbb{W}_S(\Omega_S) - \mathbb{W}(\Omega_S)| \\ &= 2, \end{aligned}$$

where the last inequality holds since $\mathbb{W}_S(\Omega_S) = 1$ and $\mathbb{W}(\Omega_S) = 0$ if $\mathbb{W}$ is a continuous distribution. Thus, we do not have the convergence property for the Total Variation metric.

For the Uniform metric, we have the following convergence results based on the assumption that the density function f of distribution $\mathbb{W}$ is bounded.

Lemma 9. *For a general dimension $n \geq 1$, we have*

$$d_{U}(\mathbb{W}_S, \mathbb{W}) \leq \Gamma(\ell/S)^n,$$

where Γ is the bound of density function f, i.e., $\Gamma = \max_{t \in \Omega} f(t)$.

Proof. For analysis convenience, we assume for each small hypercube, the smallest value (bottom left) of this hypercube is selected as our sample. Then, for any $t \in \Omega$, we can find the corresponding hypercube (defined as $B(t)$), the sample $\xi(t)$, and the largest value (top right) $\bar{\xi}(t)$ in this hypercube $B(t)$. Based on the approach described above on selecting the samples, we have $\bar{\xi}(t) > t \geq \xi(t)$ (note here that if $t = \bar{\xi}(t)$, then t belongs to the hypercube in which the sample is $\bar{\xi}(t)$). According to the definition of Uniform metric, we have

$$\begin{aligned} d_{\text{U}}(\mathbb{W}_S, \mathbb{W}) &= \sup_t |\mathbb{W}_S(x \leq t) - \mathbb{W}(x \leq t)| = \sup_t \left| \int_{[-\infty, t]} d\mathbb{W}_S - \int_{[-\infty, t]} d\mathbb{W} \right| \\ &= \sup_t \left| \left(\sum_{i:\xi_i < \xi(t)} \int_{B_i} d\mathbb{W}_S + \int_{[\xi(t), t]} d\mathbb{W}_S \right) - \left(\sum_{i:\xi_i < \xi(t)} \int_{B_i} d\mathbb{W} + \int_{[\xi(t), t]} d\mathbb{W} \right) \right| \\ &= \sup_t \left| \left(\sum_{i:\xi_i < \xi(t)} p_i + \int_{[\xi(t), t]} d\mathbb{W}_S \right) - \left(\sum_{i:\xi_i < \xi(t)} p_i + \int_{[\xi(t), t]} d\mathbb{W} \right) \right| \end{aligned}$$

$$
\begin{aligned}
&= \sup_t \left| \int_{[\xi(t),t]} d\mathbb{W}_S - \int_{[\xi(t),t]} d\mathbb{W} \right| = \sup_t \left| p_{B(t)} - \int_{[\xi(t),t]} d\mathbb{W} \right| \\
&= \sup_t \left| \int_{[\xi(t),\bar{\xi}(t))} d\mathbb{W} - \int_{[\xi(t),t]} d\mathbb{W} \right| = \sup_t \int_{(t,\bar{\xi}(t))} d\mathbb{W} \\
&= \sup_t \int_{t<x<\bar{\xi}(t)} f(x)dx \leq \Gamma \int_{t<x<\bar{\xi}(t)} dx = \Gamma(\ell/S)^n.
\end{aligned}
$$

Therefore, the conclusion holds. □

Now we provide the approximation tool to address (RA-SP) (i.e., Formulation (2)) by deriving its upper and lower bounds. For analysis convenience, we write down the following two problems:

$$
H_S^+(x,\epsilon) = c^T x + \sup_{F_S : d_\zeta(F_S, F_S^0) \leq \theta + 2\epsilon} \mathbb{E}_{F_S}[Q(x,\xi)] + \Delta_S, \tag{26}
$$

$$
H_S^-(x,\epsilon) = c^T x + \sup_{F_S : d_\zeta(F_S, F_S^0) \leq \theta - \epsilon} \mathbb{E}_{F_S}[Q(x,\xi)], \tag{27}
$$

where

$$
\Delta_S = \max_{\xi \in \Omega} \|\nabla Q(x,\xi)\| \sqrt{n}\ell/S, \tag{28}
$$

F_S^0 is the discrete distribution counterpart of $\mathbb{P}_0$ following the above construction process, and F_S is any discrete distribution of $\xi_1, \xi_2, \cdots, \xi_{S^d}$. Note here ϵ values in (27) can be decided based on different metrics. For instance, for the Kantorovich metric case, we have

$$
\epsilon = \sqrt{n}\ell/S. \tag{29}
$$

First, we consider deriving upper and lower bounds for (21) when the first-stage solution x is fixed (say $x = \bar{x}$ in the above (26) and (27)). We have the following proposition.

Proposition 5. *For any given first-stage decision $\bar{x}$, $H_S^-(\bar{x},\epsilon)$ as defined in (27) is a lower bound of $f(\bar{x})$ and $H_S^+(\bar{x},\epsilon)$ as defined in (26) is an upper bound of $f(\bar{x})$ as defined in (21).*

Proof. For the lower bound part, we first assume that F_S^* is the worst-case distribution of Problem (27). Then F_S^* should satisfy the constraint $d_\zeta(F_S^*, F_S^0) \leq \theta - \epsilon$. Also, based on Lemmas 1 to 4 and Lemma 8, we have $d_\zeta(F_S^0, \mathbb{P}_0) \leq \epsilon$ following (29) and (25). Since d_ζ is a metric, it satisfies the triangle inequality, i.e., $d_\zeta(F_S^*, \mathbb{P}_0) \leq d_\zeta(F_S^*, F_S^0) + d_\zeta(F_S^0, \mathbb{P}_0)$. Thus we have $d_\zeta(F_S^*, \mathbb{P}_0) \leq \theta$.

That is, F_S^* belongs to the set $\{\mathbb{P} : d_\zeta(\mathbb{P}, \mathbb{P}_0) \leq \theta\}$ as described in (21), which immediately leads to $\mathbb{E}_{F_S^*}[Q(\bar{x}, \xi)] \leq \sup_{\mathbb{P}: d_\zeta(\mathbb{P}, \mathbb{P}_0) \leq \theta} \mathbb{E}_{\mathbb{P}}[Q(\bar{x}, \xi)]$. Therefore,

$$\begin{aligned} H_S^-(\bar{x}, \epsilon) - f(\bar{x}) &= (c^T \bar{x} + \mathbb{E}_{F_S^*}[Q(\bar{x}, \xi)]) - (c^T \bar{x} + \sup_{\mathbb{P}: d_\zeta(\mathbb{P}, \mathbb{P}_0) \leq \theta} \mathbb{E}_{\mathbb{P}}[Q(\bar{x}, \xi)]) \\ &= \mathbb{E}_{F_S^*}[Q(\bar{x}, \xi)] - \sup_{\mathbb{P}: d_\zeta(\mathbb{P}, \mathbb{P}_0) \leq \theta} \mathbb{E}_{\mathbb{P}}[Q(\bar{x}, \xi)] \leq 0. \end{aligned}$$

For the upper bound part, we first prove $\hat{H}_S^+(\bar{x}, \epsilon)$ defined below is an upper bound of $f(\bar{x})$ for any given $\bar{x}$:

$$\hat{H}_S^+(\bar{x}, \epsilon) = c^T \bar{x} + \sup_{\mathbb{Q}: d_\zeta(\mathbb{Q}, F_S^0) \leq \theta + \epsilon} \mathbb{E}_{\mathbb{Q}}[Q(\bar{x}, \xi)], \tag{30}$$

where $\mathbb{Q}$ is a general probability distribution. Now assume $\mathbb{Q}^*$ is the worst-case distribution of Problem (21). Then $\mathbb{Q}^*$ should satisfy the constraint $d_\zeta(\mathbb{Q}^*, \mathbb{P}_0) \leq \theta$. Also, from Lemma 8, we have $d_\zeta(F_S^0, \mathbb{P}_0) \leq \epsilon$. According to the triangle inequality, we have $d_\zeta(\mathbb{Q}^*, F_S^0) \leq d_\zeta(\mathbb{Q}^*, \mathbb{P}_0) + d_\zeta(\mathbb{P}_0, F_S^0) \leq \theta + \epsilon$. That is, $\mathbb{Q}^*$ belongs to the set $\{\mathbb{Q} : d_\zeta(\mathbb{Q}, F_S^0) \leq \theta + \epsilon\}$, which immediately yields $E_{\mathbb{Q}^*}[Q(\bar{x}, \xi)] \leq \sup_{\mathbb{Q}: d_\zeta(\mathbb{Q}, F_S^0) \leq \theta + \epsilon} E_{\mathbb{Q}}[Q(\bar{x}, \xi)]$. Therefore,

$$\begin{aligned} f(\bar{x}) - \hat{H}_S^+(\bar{x}, \epsilon) &= (c^T \bar{x} + E_{\mathbb{Q}^*}[Q(\bar{x}, \xi)]) - (c^T \bar{x} + \sup_{\mathbb{Q}: d_\zeta(\mathbb{Q}, F_S^0) \leq \theta + \epsilon} E_{\mathbb{Q}}[Q(\bar{x}, \xi)]) \\ &= E_{\mathbb{Q}^*}[Q(\bar{x}, \xi)] - \sup_{\mathbb{Q}: d_\zeta(\mathbb{Q}, F_S^0) \leq \theta + \epsilon} E_{\mathbb{Q}}[Q(\bar{x}, \xi)] \leq 0. \end{aligned} \tag{31}$$

Now, we prove that $H_S^+(\bar{x}, \epsilon)$ as defined in (26) is an upper bound of $\hat{H}_S^+(\bar{x}, \epsilon)$ as defined in (30) for any given $\bar{x}$. Let $\mathbb{Q}^*$ be the worst-case distribution of $\hat{H}_S^+(\bar{x}, \epsilon)$. Then

$$\hat{H}_S^+(\bar{x}, \epsilon) = c^T \bar{x} + \mathbb{E}_{\mathbb{Q}^*}[Q(\bar{x}, \xi)], \tag{32}$$

and

$$d_\zeta(\mathbb{Q}^*, F_S^0) \leq \theta + \epsilon. \tag{33}$$

Next, let F_S^* be the discrete distribution counterpart of $\mathbb{Q}^*$ following our construction. Then

according to Lemma 8, we have

$$d_\zeta(\mathbb{Q}^*, F_S^*) \leq \epsilon. \tag{34}$$

Therefore, with inequalities (33) and (34), we have

$$d_\zeta(F_S^*, F_S^0) \leq d_\zeta(F_S^*, \mathbb{Q}^*) + d_\zeta(\mathbb{Q}^*, F_S^0) \leq \theta + 2\epsilon.$$

Hence, F_S^* belongs to the set $\{F_S : d_\zeta(F_S, F_S^0) \leq \theta + 2\epsilon\}$, which leads to

$$\mathbb{E}_{F_S^*}[Q(\bar{x}, \xi)] \leq \sup_{F_S : d_\zeta(F_S, F_S^0) \leq \theta + 2\epsilon} E_{F_S}[Q(\bar{x}, \xi)] \tag{35}$$

for any given $\bar{x}$. Now we are ready to prove that $H_S^+(\bar{x}, \epsilon)$ is an upper bound of $\hat{H}_S^+(\bar{x}, \epsilon)$ as follows:

$$\begin{aligned} H_S^+(\bar{x}, \epsilon) - \hat{H}_S^+(\bar{x}, \epsilon) &= \sup_{F_S : d_\zeta(F_S, F_S^0) \leq \theta + 2\epsilon} \mathbb{E}_{F_S}[Q(\bar{x}, \xi)] + \Delta_S - \mathbb{E}_{\mathbb{Q}^*}[Q(\bar{x}, \xi)] \\ &\geq \mathbb{E}_{F_S^*}[Q(\bar{x}, \xi)] - \mathbb{E}_{\mathbb{Q}^*}[Q(\bar{x}, \xi)] + \Delta_S \qquad (36) \\ &= \sum_{i=1}^{S^n} \int_{B_i} [Q(\bar{x}, \xi_i) - Q(\bar{x}, \xi)] d\mathbb{Q}^* + \Delta_S, \qquad (37) \end{aligned}$$

where (36) holds because of (35). According to the Mean Value Theorem, for every hypercube B_i, there exists a $\bar{\xi}_i \in [\xi, \xi_i]$, such that $Q(\bar{x}, \xi_i) - Q(\bar{x}, \xi) = \nabla Q(\bar{x}, \bar{\xi}_i)(\xi_i - \xi)$, because $Q(\bar{x}, \xi)$ is continuous in ξ. Hence,

$$\begin{aligned} \left| \textstyle\sum_{i=1}^{S^n} \int_{B_i} [Q(\bar{x}, \xi_i) - Q(\bar{x}, \xi)] d\mathbb{Q}^* \right| &\leq \sum_{i=1}^{S^n} \int_{B_i} |Q(\bar{x}, \xi_i) - Q(\bar{x}, \xi)| d\mathbb{Q}^* \\ &\leq \sum_{i=1}^{S^n} \|\nabla Q(\bar{x}, \bar{\xi}_i)\| \int_{B_i} \|\xi_i - \xi\| d\mathbb{Q}^* \\ &\leq \max_{\xi \in \Omega} \|\nabla Q(\bar{x}, \xi)\| \max_{\xi \in B_i} \rho(\xi, \xi_i) \sum_{i=1}^{S^n} \int_{B_i} d\mathbb{Q}^* \\ &\leq \max_{\xi \in \Omega} \|\nabla Q(\bar{x}, \xi)\| \sqrt{n} \ell / S, \end{aligned}$$

where the last inequality follows from (22). Then

$$\sum_{i=1}^{S^n} \int_{B_i} [Q(\bar{x}, \xi_i) - Q(\bar{x}, \xi)] d\mathbb{Q}^* \geq - \max_{\xi \in \Omega} \|\nabla Q(\bar{x}, \xi)\| \sqrt{n} \ell / S$$

and accordingly,

$$(37) \quad \geq \Delta_S - \max_{\xi \in \Omega} \|\nabla Q(\bar{x}, \xi)\| \sqrt{n} \ell / S \geq 0,$$

where the last inequality follows (28) and this implies $H_S^+(\bar{x}, \epsilon) - \hat{H}_S^+(\bar{x}, \epsilon) \geq 0$. Since $\hat{H}_S^+(\bar{x}, \epsilon)$ is an upper bound of $f(\bar{x})$ following (31), the conclusion holds. □

Now we are ready to analyze the upper and lower bounds of problem (RA-SP) and their convergence properties. Accordingly, we let

$$v_S^+ = \min_{x \in X} H_S^+(x, \epsilon), \tag{38}$$

$$v_S^- = \min_{x \in X} H_S^-(x, \epsilon), \tag{39}$$

and v^* represents the optimal objective value of (RA-SP) (i.e., Formulation (2)). The following conclusion holds.

Theorem 3. *The optimal objective value of (RA-SP) (i.e., v^*) is bounded above and below by v_S^+ and v_S^- as described in (38) and (39). That is,*

$$v_S^- \leq v^* \leq v_S^+. \tag{40}$$

Proof. We first verify the lower bound part. Denote x^* and x_S^- be the optimal solutions to problem (RA-SP) and problem (39), respectively. Then

$$\begin{aligned} v_S^- - v^* &= H_S^-(x_S^-, \epsilon) - f(x^*) \\ &\leq H_S^-(x^*, \epsilon) - f(x^*) \qquad (41) \\ &\leq 0, \qquad (42) \end{aligned}$$

where inequality (41) holds since x_S^- is an optimal solution to problem (39) and x^* is a feasible solution to problem (39), and inequality (42) follows Proposition 5. For the upper bound part, denote x^* and x_S^+ be the optimal solutions to problem (RA-SP) and problem (38), respectively. Then

$$\begin{aligned} v_S^+ - v^* &= H_S^+(x_S^+, \epsilon) - f(x^*) \\ &\geq H_S^+(x_S^+, \epsilon) - f(x_S^+) \qquad (43) \\ &\geq 0, \qquad (44) \end{aligned}$$

where inequality (43) holds since x^* is the optimal solution to problem (RA-SP) and x_S^+ is a feasible solution to problem (RA-SP), and inequality (44) follows Proposition 5. □

Now we analyze the convergence property of the upper and lower bounds as the number of hypercubes S^n goes to infinity. We have the following conclusion.

Theorem 4. *The upper bound v_S^+ and lower bound v_S^- defined in (38) and (39) of (RA-SP) converge, i.e., $\lim_{S\to\infty} v_S^- = \lim_{S\to\infty} v_S^+ = v^*$, and so do the corresponding optimal solutions.*

Proof. For the first statement, since we have $v_S^- \leq v^* \leq v_S^+$ according to Theorem 3, we only need to prove $\lim_{S\to\infty} v_S^- = \lim_{S\to\infty} v_S^+$. To prove this, first, based on (26) to (28), we have $\epsilon \to 0$ and $\Delta_S \to 0$ as $S \to \infty$. Thus, for any given first-stage solution $\bar{x}$, we have

$$\lim_{S\to\infty} H_S^-(\bar{x}, \epsilon) = c^T\bar{x} + \sup_{F: d_\zeta(F, F_\infty^0) \leq \theta} \mathbb{E}_F[Q(x, \xi)] = \lim_{S\to\infty} H_S^+(x, \epsilon). \qquad (45)$$

Next, we prove the claim $\lim_{S\to\infty} v_S^- = \lim_{S\to\infty} v_S^+$. Since $\lim_{S\to\infty} v_S^- \leq \lim_{S\to\infty} v_S^+$ following (40), if the equation $\lim_{S\to\infty} v_S^- = \lim_{S\to\infty} v_S^+$ does not hold, we have

$$\lim_{S\to\infty} v_S^- < \lim_{S\to\infty} v_S^+. \qquad (46)$$

Now we prove the claim by a contradiction method. Let x^+ and x^- be the corresponding optimal solutions to $\lim_{S\to\infty} v_S^+$ and $\lim_{S\to\infty} v_S^-$, respectively. Then, accordingly, we have

$$\lim_{S\to\infty} H_S^+(x^+, \epsilon) = \lim_{S\to\infty} v_S^+ > \lim_{S\to\infty} v_S^- = \lim_{S\to\infty} H_S^-(x^-, \epsilon) = \lim_{S\to\infty} H_S^+(x^-, \epsilon), \qquad (47)$$

where the first equation follows (38), the inequality follows (46), the second equation follows (39), and the last equation follows (45). Note here (47) indicates that x^- is a better solution for $\lim_{S\to\infty} v_S^+$, as compared to x^+, which violates that x^+ is the corresponding optimal solution to $\lim_{S\to\infty} v_S^+$. Thus, the original conclusion holds, i.e., $\lim_{S\to\infty} v_S^- = \lim_{S\to\infty} v_S^+$. For the second statement, since X is compact and accordingly the optimal solutions of $\lim_{S\to\infty} v_S^+$ and $\lim_{S\to\infty} v_S^-$ converge. Because $\lim_{S\to\infty} v_S^- = \lim_{S\to\infty} v_S^+ = v^*$, these solutions must converge to some solution x^*, which is also an optimal solution of (RA-SP) (i.e., Formulation (2)). □

Now we can derive an approximation formulation of problem (RA-SP) as follows:

$$v_S = \min_{x\in X} \left\{ H(x) := c^T x + \sup_{F_S : d_\zeta(F_S, F_S^0) \le \theta} \mathbb{E}_{F_S}[Q(x,\xi)] \right\}. \tag{48}$$

Since $H_S^+(x,\epsilon) \ge H(x) \ge H_S^-(x,\epsilon)$, we have $v_S^+ \ge v_S \ge v_S^-$. Therefore, v_S also converges to the objective of (RA-SP) (i.e., Formulation (2)).

Next, we conclude the algorithm to solve the case in which the true distribution is continuous in Algorithm 2.

Algorithm 2: Algorithm for the continuous case

Input: Historical data $\xi^1, \xi^2, \cdots, \xi^M$ i.i.d. drawn from the true distribution.
Output: The objective value of the risk-averse problem (RA-SP).

1 Obtain the kernel density function $f_{\mathrm{M}}(x)$ as shown in (17) and θ based on the historical data.
2 Set $Gap = 1000$, $S = 0$, $\Delta S = 100$.
3 While $Gap > \sigma$, do
4 1) $S = S + \Delta S$.
5 2) Construct the discrete distribution counterpart (F_S^0) of $\mathbb{P}_0$ based on $f_{\mathrm{M}}(x)$.
6 3) Obtain ϵ based on Lemma 8.
7 4) Obtain the upper bound by solving (38).
8 5) Obtain the lower bound by solving (39).
9 6) Obtain the optimality gap Gap between the lower and upper bounds.
10 Output the solution.

Remark 1. *Note here that for the case in which the true distribution is discrete but the supporting space Ω is infinite, we can employ the same algorithm here as the continuous case.*

5 Numerical Experiments

In this section, we conduct computational experiments to show the effectiveness of the proposed ζ-structure probability metrics. We test the performance for both discrete and continuous distribution cases through two instances: Newsvendor problem and facility location problem.

5.1 Newsvendor Problem

For the newsvendor problem, we consider the case in which a news vendor places an order y of newspapers one day ahead at a unit cost of c_1, and on the second day, she observes the real demand d and places a supplemental order x at a unit cost of c_2. The unit sale price is p, and there is no salvage value for unsold newspapers. Without loss of generality, we can assume $c_1 < c_2 < p$. Otherwise, if $c_1 \geq c_2$, then the news vendor can put all purchase orders on the second day after demand is realized, and if $c_2 \geq p$, then the news vendor will not put any order on the second day to avoid loss of money. Accordingly, the data-driven risk-averse stochastic newsvendor problem can be formulated as follows:

$$\max_{y\geq 0} -c_1 y + \min_{\mathbb{P}: d_\zeta(\mathbb{P},\mathbb{P}_0)\leq\theta} \mathbb{E}_{\mathbb{P}} \max_{x\geq 0}[p\min\{x+y, d(\xi)\} - c_2 x], \tag{49}$$

where ζ can be any metric in the ζ-structure probability metrics family, and θ can be calculated based on the convergence rates derived in Subsection 3.2 if the true distribution is discrete and the convergence rates derived in Subsection 4.2 if the true distribution is continuous. To illustrate the performance of our proposed approach, we set $c_1 = 3, c_2 = 4$, and $p = 5$. In this setting, we assume that the demand follows a discrete distribution with two scenarios: 10 and 20, each with probabilities 0.4 and 0.6, respectively, and use this distribution to generate our historical data sets for testing. We study the effects of the number of observed historical data, by setting the confidence level to be 99% and varying the number of historical data from 10 to $10,000$.

The results are reported in Figure 2. From the figure, we can observe that, no matter what kind of metrics we use, as the size of historical data increases, the objective value of the risk-averse stochastic newsvendor problem tends to increase. This result conforms to the intuition, because the value of θ decreases as the number of historical data samples increases. Therefore, accordingly,

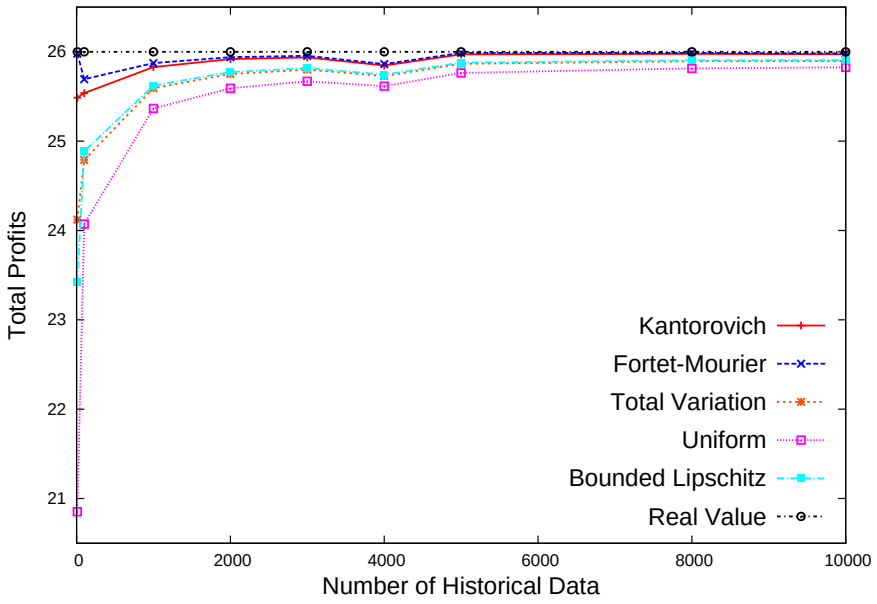

Figure 2: Effects of historical data

the risk-averse stochastic newsvendor problem becomes less conservative. We can also observe that, when the size of historical data samples exceeds $2,000$, the gaps between the risk-averse problem (RA-SP) and the risk-neutral problem (SP) are very small (less than 0.2) already under the Wasserstein and Fortet-Mourier metrics. Furthermore, when the size of historical data samples exceeds $5,000$, the gaps between the risk-averse and risk-neutral ones are small under all metrics.

In addition, we explore the effects of confidence level on the objective value of the risk-averse stochastic newsvendor problem. We set the number of historical data samples to be $5,000$, and test five different confidence levels: 0.7, 0.8, 0.9, 0.95, and 0.99. The results are reported in Figure 3.

From the figure, we can observe that as the confidence level increases, the gaps between the risk-averse and risk-neutral problems increase. This is due to the fact that, as the confidence level increases, the value of θ increases. Thus, the problem becomes more conservative and the true probability distribution is more likely to be in the confidence set $\mathcal{D}$.

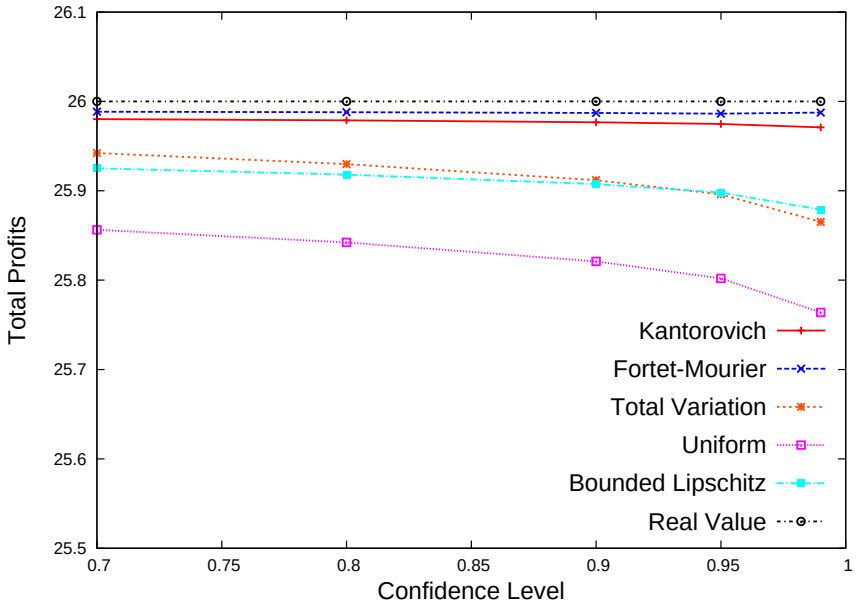

Figure 3: Effects of confidence level

5.2 Facility Location Problem

For the facility location problem, we consider the case in which the demand is assumed continuous, while the actual distribution is unknown. For the problem, the decisions include deciding what facilities, among candidate locations $i = 1, \cdots, M$, to open, to satisfy a set of demand sites $j = 1, \cdots, N$, whose demands are independently distributed. Each facility i associates with a fixed cost F_i and a capacity C_i if it is open. Meanwhile, there is a deterministic unit transportation cost T_{ij} for shipping products from each facility i to each demand site j. In our problem setting, in the first stage, we decide which facilities to open (e.g., using the binary variable "y_i" to indicate if facility i is open). In the second stage, after realizing the demand d_j at site j, we decide the amount of products to be shipped from facility i to demand site j, denoted as x_{ij}. Accordingly, the

data-driven risk-averse two-stage stochastic facility location problem can be formulated as follows:

$$\begin{aligned} \min_y \quad & \sum_{i=1}^{M} F_i y_i + \max_{\mathbb{P}\in\mathcal{D}} \mathbb{E}_{\mathbb{P}} \left[\sum_{i=1}^{M} \sum_{j=1}^{N} T_{ij} x_{ij}(\xi) \right] \\ s.t. \quad & \sum_{j=1}^{N} x_{ij} \le C_i y_i, i = 1, \cdots, M, \\ & \sum_{i=1}^{M} x_{ij} = d_j(\xi), j = 1, \cdots, N, \\ & y_i \in \{0,1\}, x_{ij} \ge 0, i = 1, \cdots, M, j = 1, \cdots, N. \end{aligned} \tag{50}$$

In our experiment setting, we assume there are 10 facility locations and 10 demand sites. For each location i, the capacity is $15 + i$, and the fixed cost is $100 + i$. Meanwhile, the unit shipping cost from location i to demand site j is $5 + 0.008i$. We test the cases in which the true distributions are Uniform distribution, Normal distribution, Gamma distribution, and Weibull distribution, respectively. That is, we generate historical data sets sampled from these distributions.

Similar to the newsvendor problem described in Subsection 5.1, we first study the effects of historical data. We report the relationships between the objective values of the risk-averse stochastic facility location problems and the numbers of historical data ($\times 10^6$), and the results are shown in Figure 4. Corresponding to each risk-averse stochastic facility location problem (50) for a given set of historical data, we compute the objective value of the risk-averse stochastic facility location problem (50) following the approximation formulation (48). Meanwhile, to obtain F_S^0, we set the number of samples taken from the reference continuous distribution based on (17) to be 100.

From Figure 4 we can observe that, as the size of historical data increases, the objective values tend to decrease. This is because as the number of historical data increases, the value of θ decreases and the problem becomes less conservative.

Finally, we set the size of historical data to be 5000 to evaluate the performance of the algorithm for the inner loop problem. For this experiment, we take Bounded Lipschitz as an example. We test the effects of the number of samples for the proposed sampling approximation approach. We obtain the upper and lower bounds, and the estimated values corresponding to different sample sizes in solving the inner loop problem, respectively. The computational results are shown in Figure 5.

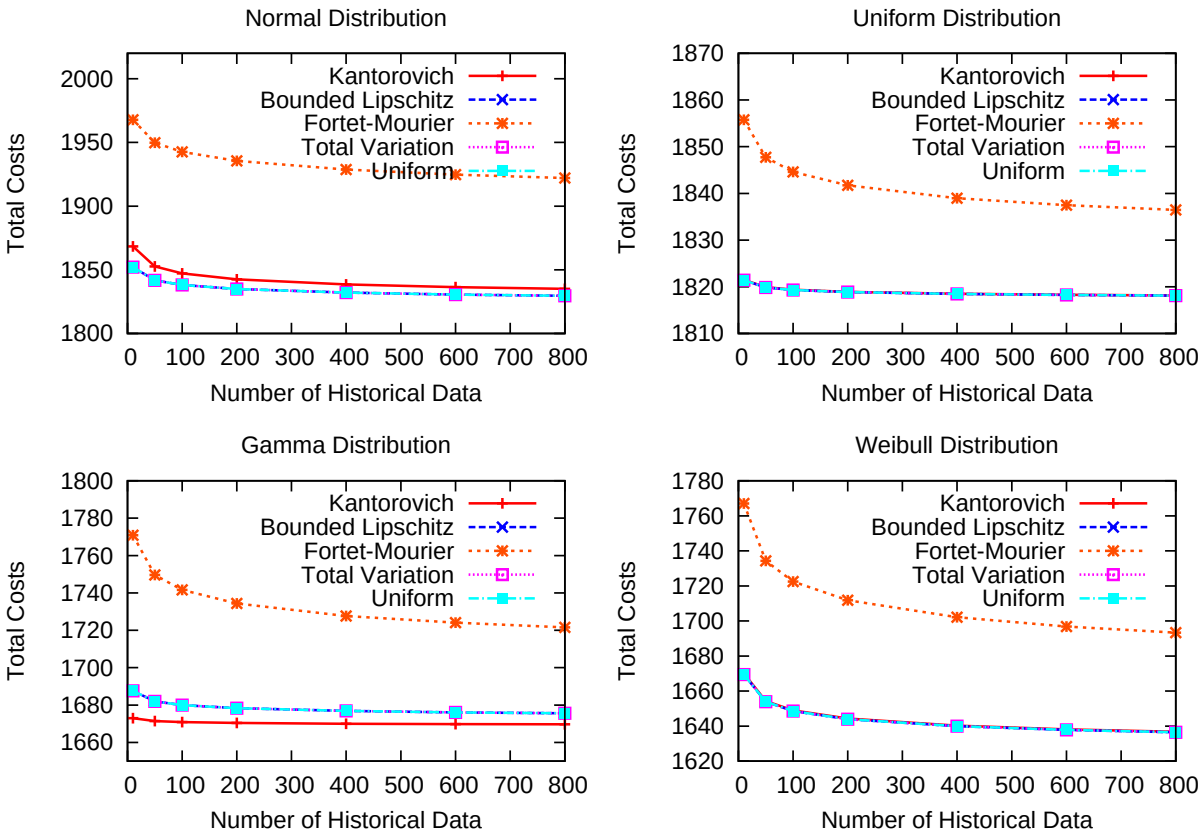

Figure 4: Effects of historical data

From Figure 5 we can observe that, no matter what the true distribution is, as the number of samples increases, the gap between the upper bound and the lower bound tends to decrease.

6 Conclusion

In this paper, we studied a new family of probability metrics, ζ-structure probability metrics, to construct the confidence set of ambiguous distributions, by learning from the historical data. To the best of our knowledge, this research provided one of the the first studies on exploring this whole family of probability metrics to solve risk-averse stochastic programs considering general distributions (both discrete and continuous cases). Based on the constructed confidence set, we

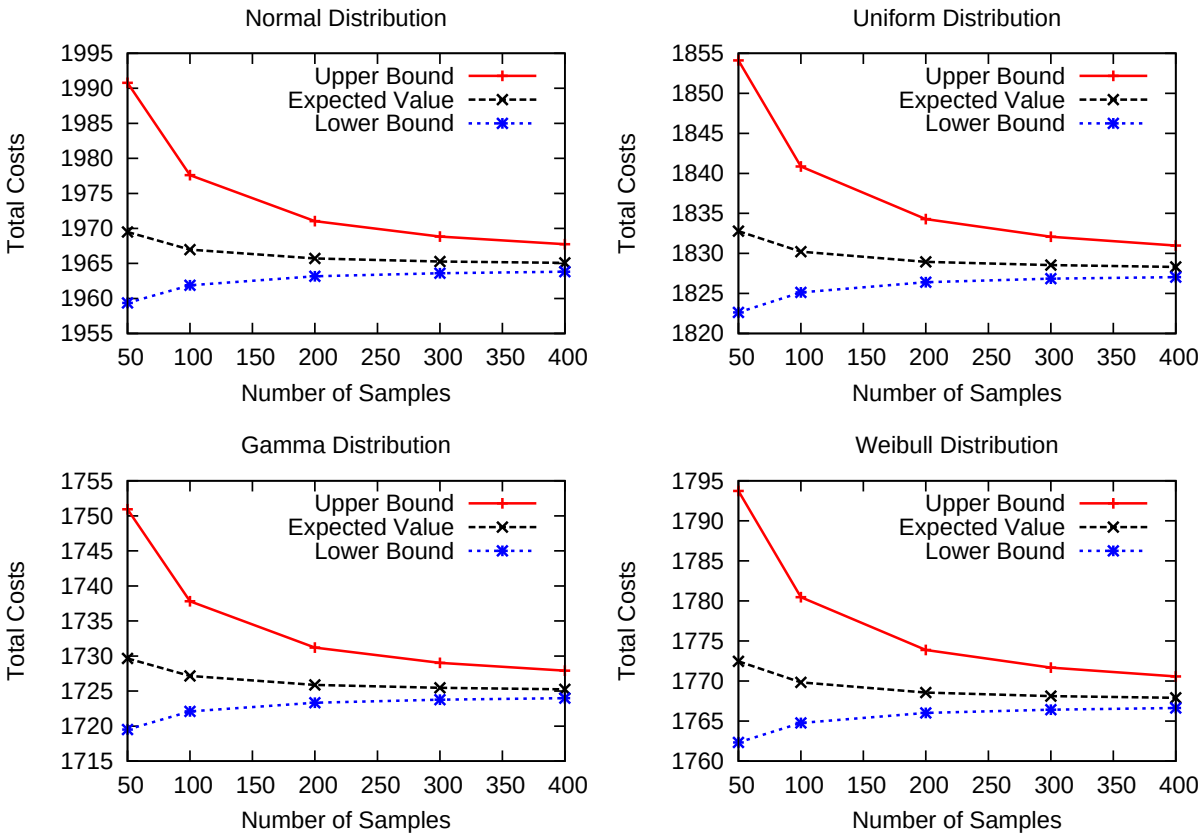

Figure 5: Effects of inner loop samples

further developed a framework to solve the risk-averse two-stage stochastic program. In particular, we reformulated the risk-averse problem as a traditional robust optimization problem for the discrete case. For the continuous case, we proposed a sampling approach to deriving the upper and lower bounds for the optimal objective value of the risk-averse problem corresponding to a given set of historical data. In addition, these bounds are proved to converge to the optimal objective value as the sample size increases to infinity. We also proved that under ζ-structure probability metrics, the risk-averse problem converges to the risk-neutral one uniformly as the size of historical data increases to infinity. The final experimental results of the newsvendor and facility location problems verified the effectiveness of the proposed approach and numerically showed the value of data.

Acknowledgements

The authors would like to thank Professor Anton Kleywegt from the Georgia Institute of Technology for the suggestions on improving the quality of this paper.

References

[1] L. Ambrosio, N. Gigli, and G. Savar. *Gradient Flows in Metric Spaces and in the Spaces of Probability Measures*. Springer, 2000.

[2] J. Ang, F. Meng, and J. Sun. Two-stage stochastic linear programs with incomplete information on uncertainty. *European Journal of Operational Research*, 233(1):16–22, 2014.

[3] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[4] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013.

[5] J. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, 1997.

[6] F. Bolley, A. Guillin, and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3-4):541–593, 2007.

[7] F. Bolley and C. Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des Sciences de Toulouse*, volume 14, pages 331–352, 2005.

[8] G. C. Calafiore. Ambiguous risk measures and optimal robust portfolios. *SIAM Journal on Optimization*, 18(3):853–877, 2007.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[10] A. B. Cybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

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

[12] A. Dembo and Z. Ofer. *Large Deviations Techniques and Applications*, volume 38. Springer, 2010.

[13] Y. Deng and W. Du. The Kantorovich metric in computer science: A brief survey. *Electronic Notes in Theoretical Computer Science*, 253(3):73–82, 2009.

[14] L. Devroye and L. Gyorfi. No empirical probability measure can converge in the total variation sense for all distributions. *The Annals of Statistics*, 18(3):1496–1499, 1990.

[15] R. M. Dudley. *Real Analysis and Probability*, volume 74. Cambridge University Press, 2002.

[16] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

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

[18] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[19] E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

[20] R. M. Gray, D. L. Neuhoff, and P. C. Shields. A generalization of Ornstein's $\bar{d}$ distance with applications to information theory. *The Annals of Probability*, 3(2):315–328, 1975.

[21] J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods*, volume 1. Methuen London, 1964.

[22] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 1952.

[23] R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. *Technical Report, Available at Optimization-Online*, 2013.

[24] L. V. Kantorovich and G. S. Rubinshtein. On a space of totally additive functions. *Vestn Lening. Univ.*, 13(7):52–59, 1958.

[25] D. Klabjan, D. Simchi-Levi, and M. Song. Robust stochastic lot-sizing by means of histograms. *Production and Operations Management*, 22(3):691–710, 2013.

[26] T. Lindvall. *Lectures on the Coupling Method*. Courier Dover Publications, 2002.

[27] D. Love and G. Bayraksan. Phi-divergence constrained ambiguous stochastic programs. Technical report, University of Arizona, 2012.

[28] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2009.

[29] P. J. Moreno, P. P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *Advances in Neural Information Processing Systems*, 2003.

[30] L. Pardo. *Statistical Inference Based on Divergence Measures*. CRC Press, 2006.

[31] G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

[32] S. T. Rachev. *Mass Transportation Problems*, volume 2. Springer, 1998.

[33] H. Scarf. A min-max solution of an inventory problem. In K. Arrow, S. Karlin, and H. Scarf, editors, *Studies in the Mathematical Theory of Inventory and Production*, pages 201–209. Stanford University Press, 1958.

[34] A. Shapiro and S. Ahmed. On a class of minimax stochastic programs. *SIAM Journal on Optimization*, 14(4):1237–1249, 2004.

[35] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*, volume 9. SIAM, 2009.

[36] G. R. Shorack. *Probability for Statisticians*. Springer New York, 2000.

[37] G. R. Shorack and J. A. Wellner. *Empirical Processes with Applications to Statistics*, volume 59. SIAM, 2009.

[38] F. Thollard, P. Dupont, and C. D. L. Higuera. Probabilistic DFA inference using Kullback-Leibler divergence and minimality. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 975–982. Morgan Kaufmann Publishers Inc., 2000.

[39] A. M. Vershik. Kantorovich metric: Initial history and little-known applications. *Journal of Mathematical Sciences*, 133(4):1410–1417, 2006.

[40] M. P Wand and M. C. Jones. *Kernel Smoothing*. CRC Press, 1994.

[41] J. Žáčková. On minimax solutions of stochastic linear programming problems. *Časopis pro pěstování matematiky*, 91(4):423–430, 1966.

[42] V. M. Zolotarev and M. Vladimir. Probability metrics. *Teoriya Veroyatnostei i ee Primeneniya*, 28(2):264–287, 1983.

[43] S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198, 2013.

[44] S. Zymler, D. Kuhn, and B. Rustem. Worst-case value at risk of nonlinear portfolios. *Management Science*, 59(1):172–188, 2013.