sustainability

MDPI

Article

Convergence Analysis on Data-Driven Fortet-Mourier Metrics with Applications in Stochastic Optimization

Zhiping Chen , He Hu and Jie Jiang *

School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China; zchen@mail.xjtu.edu.cn (Z.C.); sxytjxyhh@stu.xjtu.edu.cn (H.H.)
* Correspondence: jiangjiecq@163.com

Abstract: Fortet-Mourier (FM) probability metrics are important probability metrics, which have been widely adopted in the quantitative stability analysis of stochastic programming problems. In this study, we contribute to different types of convergence assertions between a probability distribution and its empirical distribution when the deviation is measured by FM metrics and consider their applications in stochastic optimization. We first establish the quantitative relation between FM metrics and Wasserstein metrics. After that, we derive the non-asymptotic moment estimate, asymptotic convergence, and non-asymptotic concentration estimate for FM metrics, which supplement the existing results. Finally, we apply the derived results to four kinds of stochastic optimization problems, which either extend the present results to more general cases or provide alternative avenues. All these discussions demonstrate the motivation as well as the significance of our study.

Keywords: Fortet-Mourier metric; discrete approximation; stochastic optimization; stochastic dominance; distributionally robust

MSC: 90C15; 91B70

Citation: Chen, Z.; Hu, H.; Jiang, J. Convergence Analysis on Data-Driven Fortet-Mourier Metrics with Applications in Stochastic Optimization. *Sustainability* **2022**, *14*, 4501. https://doi.org/10.3390/su14084501

Academic Editor: David Barilla

Received: 22 March 2022
Accepted: 6 April 2022
Published: 10 April 2022

1. Introduction

The estimation of the distance between a distribution and its empirical approximation obtained from some independent and identically distributed (iid) samples is an important subject in probability theory, mathematical statistics, and information theory. It has vast applications in many fields, such as quantization, optimal matching, density estimation, clustering, and so on (see [1] and the references therein for more details). To quantify the distance between two probability distributions, some rules have been adopted to generate probability metrics, such as the commonly used ζ-structure metric. By selecting different generators of the ζ-structure metric, we obtain a number of well-known probability metrics, such as the Wasserstein metric (in this study, when we refer to Wasserstein metric, it means the 1-Wasserstein metric, which is also called the Kantorovich–Rubinstein metric or Kantorovich metric), FM metric, and total variation metric.

Among probability metrics with ζ-structure, the Wasserstein metric is the most popular one which has been widely applied in statistics, probability, and machine learning [2]. It originates from the optimal transportation problem and thus can be interpreted as an optimal mass transportation plan. Except for its practical meaning in transportation, the Wasserstein metric has some good properties. For example, convergence in the Wasserstein metric is equivalent to weak convergence plus the convergence of the first order absolute moment [2].

There is some literature concentrated on the convergence analysis under Wasserstein metrics between a distribution and its empirical approximation. From now on, we refer to this as the data-driven Wasserstein metric for simplicity, and other probability metrics do the same. These convergence analyses can be mainly divided into two parts: moment estimates which aim at providing the rate of convergence for the expectation of the Wasserstein distance between a distribution and its empirical approximation and concentration

estimates which focus on the violation probability under a given tolerance. As for moment estimates, some earlier results can be found in [3,4], which provide a relatively loose convergence rate. More recently, Weed and Bach [5] focused on the compact support set case and obtained a sharp convergence rate. Dereich et al. [6] conducted the almost optimal convergence analysis. However, they put some restrictions on the range of parameters. An interesting result was given in [1], which extends some results in [6] from a limited range of parameters to the general case. As for concentration estimates, only a few results are available. The corresponding results can be found in [7,8] under some strong assumptions. Moreover, it requires that the violation parameter is large enough. In [9], Zhao and Guan investigated the case with a discrete and bounded support set. Particularly, an elaborate result on the rate of convergence of data-driven Wasserstein distance was presented in [1].

As pointed out in [2] (p. 110), the Wasserstein metric is a rather strong probability metric. Intuitively, it needs harsh conditions to establish the strong Wasserstein type upper bound estimate. Actually, we know from the definition of the Wasserstein metric that its generator is the set of Lipschitz continuous functions with Lipschitz modulus being one.

Compared with the Wasserstein metric, FM metrics are more general; their generator is a class of locally Lipschitz continuous functions. Therefore, it is more friendly to obtain some upper bounds by adopting FM metrics. In view of this, FM metrics have been widely used in the quantitative stability analysis of stochastic programming problems when the underlying probability distribution is perturbed and approximated, see for example [10–13]. Moreover, FM metrics have a close relationship with Wasserstein metrics through dual representation (Kantorovich–Rubinstein theorem). Generally, the pth order FM metric degrades into the Wasserstein metric when $p = 1$. From this point of view, the FM metric can be viewed as an extension of the Wasserstein metric. Nevertheless, there are few results concerning the convergence analysis for data-driven FM metrics. To the best of our knowledge, only Strugarek [14] examined the asymptotic convergence analysis under the FM distance.

In view of the above situations, in this article we study the data-driven FM metric. The main contributions of this study can be summarized as follows:

- We establish the quantitative connection between the Wasserstein metric and the FM metric. Based on this connection, we investigate the non-asymptotic moment estimate, asymptotic convergence, and non-asymptotic concentration estimate for data-driven FM metrics.
- We provide an alternative avenue for the convergence analysis of discrete approximations for two-stage stochastic programming problems. Different from the convergence or exponential rate of convergence analysis in [15,16], where some complex conditions are required, our approach is straightforward and brief.
- We reestablish the quantitative stability results for stochastic optimization problems with stochastic dominance constraints through FM metrics. Compared with that in [17], our conditions are weaker and different probability metrics are adopted. More importantly, we can apply the convergence conclusion to examine the discrete approximation method which is crucial for numerical solution.
- We consider data-driven distributionally robust optimization (DRO) problems with FM ball, which extends the results in [18] from the ambiguity set constructed by Wasserstein ball to the FM ball case. We prove the finite sample guarantee and asymptotic consistency, which lay the theoretical foundation for the data-driven approach for the DRO model.
- We analyze the discrete approximation of the DRO problem whose ambiguity set is constructed with the general moment information. Compared with the existing work [19] under the bounded support set, we weaken their conditions and extend their results to the case with an unbounded support set.

The remainder of this study is organized as follows. In Section 2, we give some prerequisites for further discussion. In Section 3, we discuss different kinds of convergence results for data-driven FM metrics. We consider four applications to verify our convergence

results and to further demonstrate the motivation and significance of this study in Section 4. Finally, we have some concluding remarks in Section 5.

2. Prerequisites

Let $\xi : \Omega \to \Xi \subseteq \mathbb{R}^s$ be a random vector defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, its induced probability distribution (sometimes it is called probability measure) on Ξ is $P := \mathbb{P} \circ \xi^{-1}$. We use $\mathcal{P}(\Xi)$ to denote all the probability distributions on Ξ. The set of probability distributions having finite pth order absolute moments is denoted by $\mathcal{P}_p(\Xi) := \{P \in \mathcal{P}(\Xi) : \int_\Xi \|\xi\|^p P(d\xi) < +\infty\}$.

Probability metrics measure the distance between two probability distributions. Generally, they do not satisfy the three axioms of usual distance in metric space. A commonly used class of probability metrics is the probability metric with ζ-structure, whose definition is as follows.

Definition 1. *Let $\mathcal{G}$ be a set of measurable functions from Ξ to $\mathbb{R}$. Then, for any $P, Q \in \mathcal{P}(\Xi)$,*

$$\mathbb{D}_\mathcal{G}(P, Q) := \sup_{g \in \mathcal{G}} \left| \mathbb{E}_P[g(\xi)] - \mathbb{E}_Q[g(\xi)] \right|$$

is called the ζ-structure probability metric induced by $\mathcal{G}$.

The $\mathcal{G}$ in Definition 1 totally determines the resulting ζ-structure probability metric, so it is called the generator of the ζ-structure probability metric. FM metrics and Wasserstein metrics can be deduced from the ζ-structure probability metric by choosing specific generators. Particularly, we have the following definitions.

Definition 2. *Let $P, Q \in \mathcal{P}_p(\Xi)$ for some $p \geq 1$ and $\mathcal{G}_{\mathrm{FM}_p}$ denote a set of locally Lipschitz continuous functions given by*

$$\left\{ g : \Xi \to \mathbb{R} : |g(\xi_1) - g(\xi_2)| \leq \max\{1, \|\xi_1\|, \|\xi_2\|\}^{p-1} \|\xi_1 - \xi_2\|, \forall \xi_1, \xi_2 \in \Xi \right\}.$$

Then, the pth order FM metric between P and Q is

$$\zeta_p(P, Q) := \sup_{g \in \mathcal{G}_{\mathrm{FM}_p}} \left| \mathbb{E}_P[g(\xi)] - \mathbb{E}_Q[g(\xi)] \right|.$$

Definition 3. *Let $P, Q \in \mathcal{P}_1(\Xi)$ and*

$$\mathcal{G}_W := \{ g : \Xi \to \mathbb{R} : |g(\xi_1) - g(\xi_2)| \leq \|\xi_1 - \xi_2\|, \forall \xi_1, \xi_2 \in \Xi \}.$$

Then, the Wasserstein metric between P and Q is

$$\mathbb{D}_W(P, Q) := \sup_{g \in \mathcal{G}_W} \left| \mathbb{E}_P[g(\xi)] - \mathbb{E}_Q[g(\xi)] \right|.$$

It is easy to see from the above definitions that $\zeta_1(P, Q) = \mathbb{D}_W(P, Q)$ for any $P, Q \in \mathcal{P}(\Xi)$. Moreover, we have that: if $g \in \mathcal{G}_{\mathrm{FM}_p}$, then, $-g \in \mathcal{G}_{\mathrm{FM}_p}$, so does $\mathcal{G}_W$. Therefore, we can ignore the absolute value operator in Definitions 2 and 3 when we take supremum. Moreover, both FM metrics and Wasserstein metrics have a close relationship with weak convergence. One can refer to [10] (p. 490) and [2] (Theorem 6.9) for more details.

The Wasserstein metric has an alternative definition which corresponds to the coupling marginal distributions. Specifically, the Wasserstein metric between P and Q is defined as (see [2], Definition 6.1):

$$\mathbb{D}_W(P, Q) = \inf \left\{ \int_{\Xi \times \Xi} \|\xi_1 - \xi_2\| \pi(d\xi_1, d\xi_2) : \pi \in \Pi \right\}, \tag{1}$$

where Π is the collection of all joint distributions of ξ_1 and ξ_2 with marginal distributions P and Q, respectively. It is known from Kantorovich–Rubinstein theorem [20] that Definition 3 is the dual representation of (1).

We have the following extension theorem for Lipschitz functions in Hilbert space (see [21], Theorems 4 and 5).

Lemma 1. *Let X and Y be Hilbert spaces and $g : B \subseteq X \to Y$ be a Lipschitz function with Lipschitz modulus L_g. Then, there exists a Lipschitz function $\hat{g} : X \to Y$ such that $\hat{g}(x) = g(x)$ for any $x \in B$ and L_g is also the Lipschitz modulus of $\hat{g}$.*

Lemma 1 is important for the following discussion. In [1], the authors assumed that the support set Ξ is the whole space $\mathbb{R}^s$. They obtained the non-asymptotic moment estimate [1] (Theorem 1) and concentration estimate [1] (Theorem 2) for the Wasserstein metric. For any $P, Q \in \mathcal{P}(\Xi)$, we can view them as probability distributions $\tilde{P}, \tilde{Q} \in \mathcal{P}(\mathbb{R}^s)$ through the following correspondence:

$$\tilde{P}(A) := P(A \cap \Xi) \text{ and } \tilde{Q}(A) := P(Q \cap \Xi)$$

for all $A \subseteq \mathbb{R}^s$. That is, we set the probability of the area $\mathbb{R}^s \backslash \Xi$ to be zero. Generally, we have $\mathbb{D}_W(P, Q) = \mathbb{D}_W(\tilde{P}, \tilde{Q})$. The details are as follows:

$$\begin{aligned}\mathbb{D}_W(P, Q) &= \sup_{g \in \mathcal{G}_W(\Xi)} \left| \int_\Xi g(\xi)(P - Q)(d\xi) \right| \\ &= \sup_{\hat{g} \in \hat{\mathcal{G}}_W(\mathbb{R}^s)} \left| \int_{\mathbb{R}^s} \hat{g}(\xi)(\tilde{P} - \tilde{Q})(d\xi) \right|,\end{aligned}$$

where $\mathcal{G}_W(\Xi)$ denotes the collection of all the Lipschitz continuous functions with Lipschitz modulus 1 on Ξ, and $\hat{\mathcal{G}}_W(\mathbb{R}^s)$ is the extension of $\mathcal{G}_W(\Xi)$ according to Lemma 1. Obviously, $\hat{\mathcal{G}}_W(\mathbb{R}^s) \subseteq \mathcal{G}_W(\mathbb{R}^s)$ which is the set of Lipschitz continuous functions with Lipschitz modulus 1 over $\mathbb{R}^s$. Thus, we have the estimation

$$\begin{aligned}\sup_{\hat{g} \in \hat{\mathcal{G}}_W(\mathbb{R}^s)} \left| \int_{\mathbb{R}^s} \hat{g}(\xi)(\tilde{P} - \tilde{Q})(d\xi) \right| &\leq \sup_{\bar{g} \in \mathcal{G}_W(\mathbb{R}^s)} \left| \int_{\mathbb{R}^s} \bar{g}(\xi)(\tilde{P} - \tilde{Q})(d\xi) \right| \\ &= \mathbb{D}_W(\tilde{P}, \tilde{Q}).\end{aligned}$$

That is, $\mathbb{D}_W(P, Q) \leq \mathbb{D}_W(\tilde{P}, \tilde{Q})$.

On the other hand, for any $\bar{g} \in \mathcal{G}_W(\mathbb{R}^s)$, its restriction on Ξ is Lipschitz continuous with Lipschitz modulus 1. Thus,

$$\begin{aligned}\sup_{\bar{g} \in \mathcal{G}_W(\mathbb{R}^s)} \left| \int_{\mathbb{R}^s} \bar{g}(\xi)(\tilde{P} - \tilde{Q})(d\xi) \right| &= \sup_{\bar{g} \in \mathcal{G}_W(\mathbb{R}^s)} \left| \int_\Xi \bar{g}(\xi)(\tilde{P} - \tilde{Q})(d\xi) + \int_{\Xi^c} \bar{g}(\xi)(\tilde{P} - \tilde{Q})(d\xi) \right| \\ &= \sup_{\bar{g} \in \mathcal{G}_W(\mathbb{R}^s)} \left| \int_\Xi \bar{g}(\xi)(\tilde{P} - \tilde{Q})(d\xi) \right| \\ &\leq \mathbb{D}_W(P, Q).\end{aligned}$$

Finally, we have $\mathbb{D}_W(P, Q) = \mathbb{D}_W(\tilde{P}, \tilde{Q})$. Therefore, although all the convergence results in [1] were derived under $\mathbb{R}^s$, we can extend them to any support set $\Xi \subseteq \mathbb{R}^s$.

Lemma 2 ([1], Theorem 1)**.** *Let $P \in \mathcal{P}_p(\Xi)$ for some $p > 1$. Then, there exists a constant C depending only on s (the dimension of Ξ) and p such that, for all $N \geq 1$,*

$$\mathbb{E}[\mathbb{D}_W(P_N,P)] \le C\big(\mathbb{E}_P[\|\xi\|^p]\big)^{1/p} \times \begin{cases} N^{-1/2}+N^{-(p-1)/p} & \text{if } s=1 \text{ and } p\neq 2, \\ N^{-1/2}\log(1+N)+N^{-(p-1)/p} & \text{if } s=2 \text{ and } p\neq 2, \\ N^{-1/s}+N^{-(p-1)/p} & \text{if } s>2 \text{ and } p\neq s/(s-1), \end{cases}$$

where log *is the natural logarithm.*

Lemma 2 cannot cover all the pairs (s,p), for example, $(s,p)=(1,2)$ or $(s,p)=(2,2)$. However, we can always reset p such that Lemma 2 holds by the following procedures. If $s=1$ or 2 and $p=2$, P must belong to $\mathcal{P}_q(\Xi)$ for any $q\in(1,2)$. If $s>2$ and $p=2$, we can select $q\in(1,2)$ such that $q\neq s/(s-1)$. If $s>2$ and $p=s/(s-1)$, we can choose any $q\in(1,s/(s-1))$. Then, we let $p=q$ and have that Lemma 2 holds with $s=1$ or 2 and $p\in(1,2)$ or $s>2$ and $p\neq s/(s-1)$. Therefore, Lemma 2 is applicable for any $s\in\mathbb{N}$ through carefully prepared p. In the following discussion, without loss of generality, we always assume that Lemma 2 holds for any pair $(s,p)\in\mathbb{N}\times[1,+\infty)$. Further, we can, according to Lemma 2, obtain the following uniform upper bound:

$$\mathbb{E}[\mathbb{D}_W(P_N,P)] \le C\big(\mathbb{E}_P[\|\xi\|^p]\big)^{1/p}\Big(N^{-1/\max\{2,s\}}\log(1+N)+N^{-(p-1)/p}\Big)$$

for any $s\in\mathbb{N}$ and $N\ge 1$.

Assumption 1. *Let $P\in\mathcal{P}(\Xi)$ satisfy*

$$A := \mathbb{E}_P\Big[\exp\Big(\|\xi\|^b\Big)\Big] = \int_\Xi \exp\Big(\|\xi\|^b\Big)P(\mathrm{d}\xi) < \infty \tag{2}$$

for some constant b.

Lemma 3. *Suppose that Assumption 1 holds for some $b>1$. Then, we have for $\epsilon\in(0,1]$ that*

$$\mathbb{P}\{\mathbb{D}_W(P,P_N)\ge\epsilon\} \le \alpha\times\begin{cases} \exp(-\beta N\epsilon^2) & \text{if } s=1, \\ \exp(-\beta N(\epsilon/\log(2+1/\epsilon))^2) & \text{if } s=2, \\ \exp(-\beta N\epsilon^s) & \text{if } s>2 \end{cases}$$

for all $N\ge 1$, where α and β are two positive constants depending only on P, b, and s.

Proof. Based on Assumption 1, we know that Condition (1) in [1] holds. Then, due to $\epsilon\in(0,1]$, Lemma 3 directly follows from [1] (Theorem 2). □

For a more comprehensive version of Lemma 3, one can refer to [1] (Theorem 2). Here, we focus on the case $\epsilon\in(0,1]$ because it is more interesting for us to investigate a smaller violation rather than a bigger one. A simplified version can also be found in [18] (Theorem 3.4) where the assumption $s\neq 2$ is imposed.

To simplify the following discussion, we derive a uniform upper bound for the right-hand side in Lemma 3. Note the fact that $1+\delta\le e^\delta$ for any $\delta\in\mathbb{R}$. We have

$$\begin{aligned}\log\Big(2+\frac{1}{\epsilon}\Big) &= 1+\log\Big(2+\frac{1}{\epsilon}\Big)-\log(e) \\ &= 1+\log\Big(\frac{2}{e}+\frac{1}{e\epsilon}\Big) \le \frac{2}{e}+\frac{1}{e\epsilon}.\end{aligned}$$

Letting $\epsilon\in(0,1/2]$ gives us that

$$\Big(\frac{\epsilon}{\log(2+1/\epsilon)}\Big)^2 \ge \frac{e^2\epsilon^4}{4}.$$

When $s = 2$, we have

$$\mathbb{P}\{\mathbb{D}_{\mathrm{W}}(P, P_N) \geq \epsilon\} \leq \alpha \exp\Big(-\beta N(\epsilon/\log(2+1/\epsilon))^2\Big) \leq \alpha \exp\Big(-\frac{e^2\beta N\epsilon^4}{4}\Big).$$

Moreover, for $\epsilon \in (0, 1/2]$,

$$\exp\Big(-\beta N\epsilon^4\Big) \geq \exp\Big(-\frac{e^2\beta N\epsilon^4}{4}\Big) \geq \exp\Big(-\beta N\epsilon^2\Big).$$

Therefore, we can obtain a loose but uniform upper bound estimation

$$\mathbb{P}\{\mathbb{D}_{\mathrm{W}}(P, P_N) \geq \epsilon\} \leq \alpha \exp\Big(-\beta N\epsilon^{\max\{4,s\}}\Big) \tag{3}$$

for any $\epsilon \in (0, 1/2]$ and $s \in \mathbb{N}$.

3. Convergence Analyses of Data-Driven FM Metrics

In this section, we will investigate different kinds of convergence for data-driven FM metrics. To this end, let $\xi^1, \xi^2, \cdots, \xi^N$ be N iid samples generated according to P. These samples are viewed here as the random sample $\xi^i : \Omega \to \Xi$, $1 \leq i \leq N$, on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, we obtain the empirical distribution P_N defined as

$$P_N = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\xi^i},$$

where $\mathbf{1}_{\xi^i}(\cdot)$ is the indicator function, that is, $\mathbf{1}_{\xi^i}(\xi) = 1$ for $\xi = \xi^i$ and $\mathbf{1}_{\xi^i}(\xi) = 0$ otherwise.

We first give the following vital lemma.

Lemma 4. *Let $P, Q \in \mathcal{P}_p(\Xi)$ for some $p \geq 1$. Then,*

$$\zeta_p(P, Q) \leq R^{p-1}\mathbb{D}_W(P, Q) + 4\int_{\{\xi\in\Xi:\|\xi\|>R\}} \|\xi\|^p (P + Q)(d\xi)$$

for any R satisfying $R \geq 1$ and $\mathbb{B}(\mathbf{0}, R) \cap \Xi \neq \emptyset$. Here $\mathbf{0}$ is the original point in $\mathbb{R}^s$ and $\mathbb{B}(\mathbf{0}, R)$ is the closed ball centered at $\mathbf{0}$ with radius R.

The proof of Lemma 4 can be found in Appendix A.

If we define

$$\phi_P(P, Q) := \inf_R \Big\{ R^{p-1}\mathbb{D}_W(P, Q) + 4\int_{M_R^c} \|\xi\|^p (P + Q)(d\xi) : R \geq 1, \mathbb{B}(\mathbf{0}, R) \cap \Xi \neq \emptyset \Big\},$$

then, we can obtain a tighter upper bound estimation of $\zeta_p(P, Q)$, that is

$$\zeta_p(P, Q) \leq \phi_P(P, Q).$$

The first convergence result is about the non-asymptotic moment estimate. It provides an upper bound for the expectation of the FM distance between P and its empirical approximation distribution.

Theorem 1 (Non-asymptotic moment estimates for FM metrics)**.** *Suppose that $P \in \mathcal{P}_p(\Xi)$ for some $p > 1$. Then, for sufficiently large N, we have*

$$\mathbb{E}[\zeta_p(P, P_N)] \leq \beta_N,$$

where $\{\beta_N\}$ is a sequence of positive numbers satisfying $\beta_N \to 0$ as $N \to \infty$.

The proof of Theorem 1 can be found in Appendix A.

Theorem 1 establishes the convergence in the sense of expectation. However, it fails to tell us the sample-wise convergence. The following theorem states the asymptotic convergence under FM metrics for almost every sample.

Theorem 2 (Asymptotic convergence of FM metrics)**.** *Suppose that $P \in \mathcal{P}_p(\Xi)$. Then,*

$$\zeta_p(P, P_N) \to 0$$

with probability 1*, as $N \to \infty$. Here P_N is defined at the beginning of this section.*

The proof of Theorem 2 can be found in Appendix A.

Theorems 1 and 2 claim the convergence. As we know, the rate of convergence is quite important for guiding the solution process in practice. The following theorem gives the estimate of the convergence rate under certain assumptions.

Theorem 3 (Non-asymptotic concentration estimates for FM metrics)**.** *Suppose that $P \in \mathcal{P}_p(\Xi)$ $(p \geq 1)$ and Assumption 1 holds with $b > p$. Then, for any $\epsilon \in (0, 1/2]$, we have*

$$\mathbb{P}(\zeta_p(P, P_N) \geq \epsilon) \leq \hat{\alpha} \exp(-\hat{\beta} N)$$

for some constants $\hat{\alpha} > 0$ depending on P, b, and s, and $\hat{\beta} > 0$ depending on P, b, s, and ϵ.

The proof of Theorem 3 can be found in Appendix A.

Remark 1. *Here we assume that $\epsilon \in (0, 1/2]$. The main reason is that we want to give a relatively simple proof. Fortunately, it is more interesting for us to consider a small violation rather than a large one.*

Under certain assumptions, we can obtain an estimation for $I\left(\frac{\epsilon}{16}\right)$. For example, if $M(t) \leq \exp\left(\sigma^2 t^2/2\right)$ for $t \in \mathbb{R}$, here σ is a positive constant, we have $\log M(t) \leq \sigma^2 t^2/2$. Then, according to the properties of convex quadratic functions, the rate function has the lower bound

$$I\left(\frac{\epsilon}{16}\right) \geq \frac{\epsilon^2}{512\sigma^2}.$$

Thus, we can further obtain a concrete estimate for $\hat{\beta}$. For more details in this aspect, one can refer to [16].

4. Applications

In this section, we consider four applications of convergence conclusions about FM metrics obtained in Section 3. Specifically, we study the discrete approximation of two-stage stochastic programming problems, stochastic optimization problems with dominance constraints, data-driven distributionally robust optimization problems with FM ball, and the discrete approximation for distributionally robust optimization problems with general moment ambiguity set. They will not only further illustrate the motivations of this study but also provide alternative avenues or extensions for the current results.

4.1. Two-Stage Stochastic Linear Programming Problems

Discrete approximation is an important issue in stochastic optimization, which is crucial for its numerical solution. In this subsection, by employing the convergence results in Section 3, we give an alternative avenue for analyzing the discrete approximation of two-stage stochastic programming problems.

Consider the two-stage stochastic programming problem:

$$\min_{x \in X} c^\top x + \mathbb{E}_P[\Phi(x, \xi)], \tag{4}$$

where $c \in \mathbb{R}^n$; $X \subseteq \mathbb{R}^n$ is a polyhedron; the probability measure P is supported on $\Xi \subseteq \mathbb{R}^s$, which is a polyhedron; and

$$\Phi(x,\xi) := \inf\Big\{q(\xi)^\top y(\xi) : Wy(\xi) + T(\xi)x = h(\xi), y(\xi) \geq 0\Big\}. \tag{5}$$

Here $W \in \mathbb{R}^{r\times m}$, $T(\xi) \in \mathbb{R}^{r\times n}$, $q(\xi) \in \mathbb{R}^m$, $h(\xi) \in \mathbb{R}^r$. $q(\xi), T(\xi)$ and $h(\xi)$ depend affine linearly on ξ.

Denote $f(x,\xi) = c^\top x + \Phi(x,\xi)$, and let $v(P)$ and $S(P)$ denote the optimal value and optimal solution set of Problem (4). Moreover, we use Pos$W$ to denote the set $\{W\hat{y} : \hat{y} \in \mathbb{R}^m_+\}$. Denote $D = \{u \in \mathbb{R}^m : \{z \in \mathbb{R}^r : W^\top z \leq u\} \neq \emptyset\}$.

To quantify the upper semicontinuity or the deviation distance of the optimal solution set, we define the growth function $\psi_P : \mathbb{R}_+ \to \mathbb{R}_+$ as

$$\psi_P(\tau) := \min\{\mathbb{E}_P[f(x,\xi)] - v(P) : d(x, S(P)) \geq \tau, x \in X\}.$$

Its inverse function ψ_P^{-1} is given by

$$\psi_P^{-1}(t) := \sup\{\tau \in \mathbb{R}_+ : \psi_P(\tau) \leq t\}.$$

Thus, we can define the associated conditioning function $\Psi_P : \mathbb{R}_+ \to \mathbb{R}_+$ as

$$\Psi_P(\eta) = \eta + \psi_P^{-1}(2\eta).$$

It is easy to verify that ψ_P is nondecreasing and Ψ_P is increasing. Both ψ_P and Ψ_P are lower semicontinuous on $\mathbb{R}_+$ and vanish at 0. One can refer to [10] for more details.

Moreover, we have $\psi_P^{-1}(t) \to 0_+$ as $t \to 0_+$. We illustrate this fact by contradiction. Suppose that there exists a sequence $\{t_n\}$ satisfying $t_n \to 0_+$ as $n \to \infty$, such that $\psi_P^{-1}(t_n) \nrightarrow 0_+$. Denote $\tau_n = \psi_P^{-1}(t_n)$. The lower semicontinuity of ψ_P means that $\{\tau \in \mathbb{R}_+ : \psi_P(\tau) \leq t_n\}$ is closed. Thus, $\psi_P(\tau_n) \leq t_n$. Due to the nondecreasing property of ψ_P and $t_n \to 0_+$ as $n \to \infty$, $\{\tau_n\}$ must be bounded. Without loss of generality, we assume that $\tau_n \to \tau^*$ as $n \to \infty$, where τ^* is a positive constant. According to the lower semicontinuity of ψ_P, we have

$$0 = \liminf_{n\to\infty} t_n \geq \liminf_{n\to\infty} \psi_P(\tau_n) \geq \psi_P(\tau^*) > 0,$$

which leads to a contradiction.

According to the definition of Ψ_P, we can immediately deduce that $\Psi_P(\eta) \to 0_+$ as $\eta \to 0_+$.

To introduce the following discussion, we make some standard assumptions (see [11]).

Assumption 2. *Let the following assertions hold:*

(1) For each pair $(x,\xi) \in X \times \Xi$, $h(\xi) - T(\xi)x \in$ Pos$W$ and $q(\xi) \in D$;
(2) $P \in \mathcal{P}_2(\Xi)$.

Under the above assumptions, we have the following quantitative stability results about the optimal value and optimal solution set of Problem (4).

Lemma 5 ([11], Theorem 3.3)**.** *Suppose that Assumption 2 holds and $S(P)$ is nonempty and bounded. Then, there exist constants $L > 0$ and $\delta > 0$ such that*

$$\begin{aligned} |v(P) - v(Q)| &\leq L\zeta_2(P,Q), \\ \emptyset \neq S(Q) &\subseteq S(P) + \Psi_P(L\zeta_2(P,Q))\mathbb{B} \end{aligned}$$

when $Q \in \mathcal{P}_2(\Xi)$ and $\zeta_2(P,Q) < \delta$, where $\mathbb{B}$ is the closed unit ball in $\mathbb{R}^n$.

Based on Lemma 5 and the convergence results in Section 3, we have the following convergence conclusions between the two-stage stochastic programming problem (4) and its empirical approximation.

Theorem 4. *Suppose that: (i) Assumption 2 holds; (ii) $S(P)$ is nonempty and bounded. Then,*

$$|v(P) - v(P_N)| \to 0,$$
$$d(S(P_N), S(P)) \to 0$$

with probability 1, *as* $N \to \infty$.

Proof. For the first assertion, we have from Theorem 2 that $\zeta_2(P, P_N) \to 0$ with probability 1. This means that: for the δ defined in Lemma 5, there exists a positive number $N_0 = N_0(\delta, \omega)$ such that for any $N \geq N_0$, $\zeta_2(P, P_N) \leq \delta$ for almost every $\omega \in \Omega$. Then, by Lemma 5, we have that

$$|v(P) - v(P_N)| \leq L\zeta_2(P, P_N),$$
$$d(S(P_N), S(P)) \leq \Psi_P(L\zeta_2(P, P_N))$$

hold almost surely as $N \geq N_0$, here L is defined in Lemma 5. According to Theorem 2 and the property of Ψ_P, we have

$$\zeta_2(P, P_N) \to 0$$

and thus

$$\Psi_P(L\zeta_2(P, P_N)) \to 0$$

with probability 1, as $N \to \infty$. These facts imply that

$$|v(P) - v(P_N)| \to 0,$$
$$d(S(P_N), S(P)) \to 0$$

with probability 1, as $N \to \infty$. □

Theorem 5. *Suppose that: (i) Assumption 1 holds with $b > 2$; (ii) Assumption 2 holds; (iii) $S(P)$ is nonempty and bounded. Then, for any $\epsilon \in (0, 1/2]$, there exist $\bar{\alpha} > 0$ depending on P and s, and $\bar{\beta} > 0$ depending on P, s and ϵ, such that*

$$\mathbb{P}(|v(P) - v(P_N)| \geq L\epsilon) \leq \bar{\alpha}\exp(-\bar{\beta}N),$$
$$\mathbb{P}(d(S(P_N), S(P)) \geq \Psi_P(L\epsilon)) \leq \bar{\alpha}\exp(-\bar{\beta}N).$$

Proof. If $|v(P) - v(P_N)| \leq L\zeta_2(P, P_N)$ for L defined in Lemma 5, we have from Theorem 3 that

$$\mathbb{P}(|v(P) - v(P_N)| \geq L\epsilon) \leq \mathbb{P}(\zeta_2(P, P_N) \geq \epsilon) \leq \bar{\alpha}\exp(-\bar{\beta}(\epsilon)N)$$

for any $\epsilon \in (0, 1/2]$, where $\bar{\alpha} > 0$ depends on P and s, and $\bar{\beta} > 0$ depends on P, s and ϵ. Here we use $\bar{\beta}(\epsilon)$ to stress its dependence on ϵ.

As shown in Theorem 4, a sufficient condition for

$$|v(P) - v(P_N)| \leq L\zeta_2(P, P_N)$$

is $\zeta_2(P, P_N) < \delta$, where δ is defined in Lemma 5. Without loss of generality, we assume that $\delta \in (0, 1/2]$. Analogously, we have that

$$\mathbb{P}(\zeta_2(P, P_N) < \delta) \geq 1 - \bar{\alpha}\exp(-\bar{\beta}(\delta)N).$$

Then, we obtain

$$\begin{aligned}\mathbb{P}(|v(P)-v(P_N)|\geq L\epsilon) &\leq \bar{\alpha}\exp(-\bar{\beta}(\epsilon)N)\cdot\left(1-\bar{\alpha}\exp(-\bar{\beta}(\delta)N)\right)\\ &\leq \bar{\alpha}\exp(-\bar{\beta}(\epsilon)N).\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbb{P}(d(S(P_N),S(P))\geq \Psi_P(L\epsilon)) &\leq \mathbb{P}(\Psi_P(L\zeta_2(P,P_N))\geq \Psi_P(L\epsilon))\\ &= \mathbb{P}(\zeta_2(P,P_N)\geq\epsilon),\end{aligned}$$

where the equality follows from the strictly increasing property of $\Psi_P(\cdot)$. By the same procedure, we can derive the second assertion. □

Remark 2. *The convergence analysis about two-stage stochastic programming problems can also be found in [11] (Section 4), where the covering and bracketing numbers are introduced. However, it seems difficult to verify the growth rate of the covering or bracketing number in the general case (see [11], Proposition 4.2). Our convergence results are more straightforward. Compared with [11] (Proposition 4.2), instead of the growth rate of the covering or bracketing number, we use the light-tailed distribution assumption. This assumption is commonly used in the literature, see for example [1,18].*

4.2. Stochastic Optimization Problems with Stochastic Dominance Constraints

In this part, we consider stochastic optimization problems with stochastic dominance constraints. Stochastic dominance is an important ingredient in economics, decision theory, statistics, and nowadays in modern optimization. It has been widely studied in the last two decades, see for example [17,22–26] and their references therein. Different from classical stochastic optimization models which cope with random variables by taking expectation, stochastic dominance can better reflect the relationship between two random variables. It is known that expected utility theory can also provide the comparison of two random variables. However, it is hardly possible for us to explicitly express the utility functions of decision makers [27]. From this point of view, stochastic dominance is more friendly in practice. Actually, stochastic dominance has a close relationship with expected utility theory. Generally, a random variable $\mathscr{X}$ dominates another random variable $\mathscr{Y}$ in the kth $(k\geq 1)$ order, denoted by $\mathscr{X}\succeq_{(k)}\mathscr{Y}$, if $\mathbb{E}[u(\mathscr{X})]\geq\mathbb{E}[u(\mathscr{Y})]$ for every nondecreasing function $u(\cdot)$ from a certain set of utility functions [17]. Specially, $\mathscr{X}\succeq_{(1)}\mathscr{Y}$ if and only if $\mathbb{E}[u(\mathscr{X})]\geq\mathbb{E}[u(\mathscr{Y})]$ for every nondecreasing utility function $u(\cdot)$. $\mathscr{X}\succeq_{(2)}\mathscr{Y}$ if and only if $\mathbb{E}[u(\mathscr{X})]\geq\mathbb{E}[u(\mathscr{Y})]$ for every nondecreasing and concave utility function $u(\cdot)$ [27].

The convex stochastic optimization model with the kth order stochastic dominance constraint can be described as (see [22,27]):

$$\min\{f(x): x\in D, G(x,\xi)\succeq_{(k)} Y\}, \tag{6}$$

where D is a nonempty closed and convex subset of $\mathbb{R}^n$; $f:\mathbb{R}^n\to\mathbb{R}$ is a convex function; Y is a random variable supported on $\mathcal{Y}\subseteq\mathbb{R}$, which can be treated as the random benchmark; and $G:\mathbb{R}^n\times\Xi\to\mathbb{R}$. Moreover, we assume that G is locally Lipschitz continuous with respect to ξ in the following sense:

$$\left|G(x,\xi')-G(x,\xi)\right|\leq L_G\max\left\{1,\|\xi'\|,\|\xi\|\right\}^{p-1}\|\xi'-\xi\| \tag{7}$$

for any $\xi',\xi\in\Xi$, where $p\geq 1$ and $L_G>0$. G satisfies the linear growth condition:

$$|G(x,\xi)|\leq C_G(B)\max\{1,\|\xi\|\} \tag{8}$$

for every $x\in B$ and $\xi\in\Xi$, where B is any bounded subset of $\mathbb{R}^n$, and $C_G(B)>0$ depends on B.

Actually, we can impose a more general growth condition on G, for example,

$$|G(x,\xi)| \le C_G(B)\max\{1,\|\xi\|\}^q$$

for $q \ge 1$ and the following discussion still holds. Here a linear growth condition simplifies the demonstration. The above requirements for $G(x,\xi)$ can be met easily. For instance, the objective function of the two-stage stochastic programming problem with fixed recourse satisfies the above conditions (see [11], Proposition 3.2).

Due to its attractive modeling technique, the quantitative stability analysis of stochastic optimization models with dominance constraints has been recently investigated in several works. Dencheva et al. first studied in [22] stochastic optimization problems with first order stochastic dominance constraints, which was extended by Dencheva and Römisch in [17] to the problem with general kth ($k \ge 2$) order stochastic dominance constraints. In [24], Chen and Jiang weaken the assumptions of the quantitative stability analysis in [17] by considering the case that $G(x,\xi)$ is generated by the two-stage fully random stochastic programming problem.

To establish the convergence results, we first investigate the quantitative stability of model (6). By convention, we consider its relaxed problem (see also [17,24]):

$$\min\Big\{f(x) : x \in D,\ \mathbb{E}[(\eta - G(x,\xi))_+^{k-1}] \le \mathbb{E}[(\eta - Y)_+^{k-1}] \text{ for } \forall \eta \in I\Big\}, \tag{9}$$

where $I \subseteq \mathbb{R}$ is a compact interval: $(\cdot)_+ := \max\{0,\cdot\}$.

In view of our focus in this study, we reestablish the quantitative stability conclusions of Problem (9) in what follows. We use P_ξ and P_Y to denote the probability distributions of ξ and Y, respectively. We denote the feasible solution set of Problem (9) by

$$\mathcal{X}(P_\xi, P_Y) = \Big\{x \in D : \mathbb{E}_{P_\xi}[(\eta - G(x,\xi))_+^{k-1}] \le \mathbb{E}_{P_Y}[(\eta - Y)_+^{k-1}] \text{ for } \forall \eta \in I\Big\}$$

and its perturbed feasible solution set under (Q_ξ, Q_Y) by

$$\mathcal{X}(Q_\xi, Q_Y) = \Big\{x \in D : \mathbb{E}_{Q_\xi}[(\eta - G(x,\xi))_+^{k-1}] \le \mathbb{E}_{Q_Y}[(\eta - Y)_+^{k-1}] \text{ for } \forall \eta \in I\Big\}.$$

First we examine the quantitative stability of the feasible solution set.

Proposition 1. *Let D be compact, $P_\xi \in \mathcal{P}_{k+p-2}(\Xi)$, $P_Y \in \mathcal{P}_{k-1}(\mathcal{Y})$, and G satisfy the locally Lipschitz continuity condition (7) and the linear growth condition (8). Then, there exist constants $L > 0$ and $\delta > 0$ such that*

$$d_H(\mathcal{X}(P_\xi, P_Y), \mathcal{X}(Q_\xi, Q_Y)) \le L(\zeta_{k+p-2}(P_\xi, Q_\xi) + \zeta_{k-1}(P_Y, Q_Y))$$

whenever $Q_\xi \in \mathcal{P}_{k+p-2}(\Xi)$, $Q_Y \in \mathcal{P}_{k-1}(\mathcal{Y})$ and the pair (Q_ξ, Q_Y) satisfies $\zeta_{k+p-2}(P_\xi, Q_\xi) + \zeta_{k-1}(P_Y, Q_Y) < \delta$, where $d_H(\cdot,\cdot)$ denotes the Pompeiu–Hausdorff distance.

Proof. We know from the proof of [17] (Proposition 3.2) that

$$\begin{aligned} & d_H(\mathcal{X}(P_\xi, P_Y), \mathcal{X}(Q_\xi, Q_Y)) \\ \le\ & \frac{1}{(k-1)!}\max_{\eta\in I}\Big|\mathbb{E}_{P_\xi}\Big[(\eta - G(x,\xi))_+^{k-1}\Big] - \mathbb{E}_{Q_\xi}\Big[(\eta - G(x,\xi))_+^{k-1}\Big]\Big| \\ & + \frac{1}{(k-1)!}\max_{\eta\in I}\Big|\mathbb{E}_{P_Y}\Big[(\eta - Y)_+^{k-1}\Big] - \mathbb{E}_{Q_Y}\Big[(\eta - Y)_+^{k-1}\Big]\Big| \end{aligned}$$

whenever the right-hand side is less than or equal to some positive scalar $\bar{\delta}$.

In view of this, we estimate

$$\max_{\eta\in I}\Big|\mathbb{E}_{P_\xi}\Big[(\eta - G(x,\xi))_+^{k-1}\Big] - \mathbb{E}_{Q_\xi}\Big[(\eta - G(x,\xi))_+^{k-1}\Big]\Big|$$

and

$$\max_{\eta\in I}\left|\mathbb{E}_{P_Y}\left[(\eta-Y)_+^{k-1}\right]-\mathbb{E}_{Q_Y}\left[(\eta-Y)_+^{k-1}\right]\right|,$$

respectively.

Note the fact that (see [17], (3.9))

$$\left|(\eta-t)_+^{k-1}-(\eta-\hat{t})_+^{k-1}\right|\le K_I(k-1)\max\{1,|t|,|\hat{t}|\}^{k-2}|t-\hat{t}|$$

for some positive constant K_I and any $\eta\in I$. Then, we have

$$\begin{aligned}
&\left|(\eta-G(x,\xi))_+^{k-1}-(\eta-G(x,\hat{\xi}))_+^{k-1}\right|\\
\le\ & K_I(k-1)\max\{1,|G(x,\xi)|,|G(x,\hat{\xi})|\}^{k-2}|G(x,\xi)-G(x,\hat{\xi})|\\
\le\ & K_I(k-1)\max\{1,C_G(D)\max\{1,\|\xi\|\},C_G(D)\max\{1,\|\hat{\xi}\|\}\}^{k-2}\\
&\cdot L_G\max\{1,\|\xi\|,\|\hat{\xi}\|\}^{p-1}\|\xi-\hat{\xi}\|\\
\le\ & K_I(k-1)(C_G(D)+1)L_G\max\{1,\|\xi\|,\|\hat{\xi}\|\}^{k+p-3}\|\xi-\hat{\xi}\|.
\end{aligned}$$

This means that

$$\max_{\eta\in I}\left|\mathbb{E}_{P_\xi}\left[(\eta-G(x,\xi))_+^{k-1}\right]-\mathbb{E}_{Q_\xi}\left[(\eta-G(x,\xi))_+^{k-1}\right]\right|\le L_1\zeta_{k+p-2}(P_\xi,Q_\xi),$$

where $L_1:=K_I(k-1)(C_G(D)+1)L_G$.

Similarly, we have

$$\left|(\eta-Y)_+^{k-1}-(\eta-\hat{Y})_+^{k-1}\right|\le K_I(k-1)\max\{1,|Y|,|\hat{Y}|\}^{k-2}|Y-\hat{Y}|,$$

which means that

$$\max_{\eta\in I}\left|\mathbb{E}_{P_Y}\left[(\eta-Y)_+^{k-1}\right]-\mathbb{E}_{Q_Y}\left[(\eta-Y)_+^{k-1}\right]\right|\le K_I(k-1)\zeta_{k-1}(P_Y,Q_Y).$$

Taking $L:=\frac{1}{(k-1)!}\max\{L_1,K_I(k-1)\}$, we have

$$d_H(\mathcal{X}(P_\xi,P_Y),\mathcal{X}(Q_\xi,Q_Y))\le L(\zeta_{k+p-2}(P_\xi,Q_\xi)+\zeta_{k-1}(P_Y,Q_Y)),$$

whenever $\zeta_{k+p-2}(P_\xi,Q_\xi)+\zeta_{k-1}(P_Y,Q_Y)\le\delta:=\frac{\bar{\delta}}{L}$. □

The quantitative stability result in Proposition 1 differs in two perspectives from the corresponding results in [17]. One is the locally Lipschitz continuity of G; the other is the probability metric we choose. In [17], the authors assumed that G is Lipschitz continuous, and adopted Rachev metrics and the $(k-1)$th order Wasserstein metric. As far as we know, there does not exist data-driven results under Rachev metrics.

Let $v(P_\xi,P_Y)$ and $S(P_\xi,P_Y)$ denote the optimal value and optimal solution set of Problem (9), respectively. Similar to that in Section 4.1, we can define the growth function of Problem (9) as

$$\psi_{P_\xi,P_Y}(\tau)=\inf\{f(x)-v(P_\xi,P_Y):d(x,S(P_\xi,P_Y))\ge\tau,x\in\mathcal{X}(P_\xi,P_Y)\}.$$

Then, its inverse function and the associated conditioning function are

$$\psi^{-1}_{P_\xi,P_Y}(t):=\sup\{\tau\in\mathbb{R}_+:\psi_{P_\xi,P_Y}(\tau)\le t\}$$

and

$$\Psi_{P_\xi,P_Y}(\eta):=\eta+\psi^{-1}_{P_\xi,P_Y}(2\eta).$$

Proposition 2. *Under the conditions of Proposition 1, there exist constants $\hat{L} > 0$ and $\hat{\delta} > 0$ such that*

$$\begin{aligned}&\left|v(P_\xi, P_Y) - v(Q_\xi, Q_Y)\right| \le \hat{L}(\zeta_{k+p-2}(P_\xi, Q_\xi) + \zeta_{k-1}(P_Y, Q_Y)),\\ &d(S(Q_\xi, Q_Y), S(P_\xi, P_Y)) \le \Psi_{P_\xi, P_Y}(\hat{L}(\zeta_{k+p-2}(P_\xi, Q_\xi) + \zeta_{k-1}(P_Y, Q_Y)))\end{aligned}$$

whenever $\zeta_{k+p-2}(P_\xi, Q_\xi) + \zeta_{k-1}(P_Y, Q_Y) < \hat{\delta}$.

Proof. Since f is convex, f is locally Lipschitz continuous. Since D is compact, f is in fact Lipschitz continuous over D. Then, the assertions follow from a similar proof as that for [17] (Theorem 3.3). □

Now we consider the iid samples of ξ and Y. For convenience, we assume that the samples drawn from ξ and Y have the same sample size N. The N iid samples of ξ are $\xi^1, \xi^2, \cdots, \xi^N$ and the N iid samples of Y are $Y^1, Y^2, \cdots, Y^N$. Then, we have the following empirical distributions:

$$P_{\xi,N} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\xi^i},$$

and

$$P_{Y,N} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{Y^i}.$$

With these preparations, we can establish the following convergence results.

Theorem 6. *Let D be compact, $P_\xi \in \mathcal{P}_{k+p-2}(\Xi)$, $P_Y \in \mathcal{P}_{k-1}(\mathcal{Y})$, and G satisfy the locally Lipschitz continuity condition (7) and the linear growth condition (8).*

(i) We have

$$d_H(\mathcal{X}(P_\xi, P_Y), \mathcal{X}(P_{\xi,N}, Q_{Y,N})) \to 0$$

and

$$\begin{aligned}\left|v(P_\xi, P_Y) - v(P_{\xi,N}, Q_{Y,N})\right| &\to 0,\\ d(S(P_{\xi,N}, Q_{Y,N}), S(P_\xi, P_Y)) &\to 0\end{aligned}$$

with probability 1, as $N \to \infty$.

(ii) If, moreover,

$$\mathbb{E}_{P_\xi}\left[\exp\left(\|\xi\|^b\right)\right] = \int_\Xi \exp\left(\|\xi\|^b\right) P_\xi(\mathrm{d}\xi) < +\infty$$

and

$$\mathbb{E}_{P_Y}\left[\exp\left(|Y|^c\right)\right] = \int_{\mathcal{Y}} \exp\left(|y|^c\right) P_Y(\mathrm{d}y) < +\infty$$

for some $b > k + p - 2$ and $c > k - 1$, then, for $\epsilon \in (0, 1/2]$, there exist positive scalers α_1 depending on P_ξ, b and s; β_1 depending on P_ξ, b, s and ϵ; α_2 depending on P_Y and c, and β_2 depending on P_Y, c and ϵ, such that

$$\mathbb{P}\left(d_H(\mathcal{X}(P_\xi, P_Y), \mathcal{X}(P_{\xi,N}, Q_{Y,N})) \ge L\epsilon\right) \le \alpha_1 \exp(-\beta_1 N) + \alpha_2 \exp(-\beta_2 N)$$

and

$$\begin{aligned}\mathbb{P}\left(\left|v(P_\xi, P_Y) - v(P_{\xi,N}, Q_{Y,N})\right| \ge \hat{L}\epsilon\right) &\le \alpha_1 \exp(-\beta_1 N) + \alpha_2 \exp(-\beta_2 N),\\ \mathbb{P}\left(d(S(P_{\xi,N}, Q_{Y,N}), S(P_\xi, P_Y)) \ge \Psi_{P_\xi, P_Y}(\hat{L}\epsilon)\right) &\le \alpha_1 \exp(-\beta_1 N) + \alpha_2 \exp(-\beta_2 N),\end{aligned}$$

where L and $\hat{L}$ are defined in Propositions 1 and 2, respectively.

Proof. Part (i) can be similarly proved as that in Theorem 4 by utilizing Theorem 2 and Proposition 1.

For Part (ii), we have

$$\begin{aligned}
&\mathbb{P}\big(d_H(\mathcal{X}(P_\xi, P_Y), \mathcal{X}(P_{\xi,N}, P_{Y,N})) \geq L\epsilon\big) \\
\leq\ &\mathbb{P}\Big(\zeta_{k+p-2}(P_\xi, P_{\xi,N}) + \zeta_{k-1}(P_Y, P_{Y,N}) \geq \epsilon\Big) \\
\leq\ &\mathbb{P}\Big(\zeta_{k+p-2}(P_\xi, P_{\xi,N}) \geq \frac{\epsilon}{2}\Big) + \mathbb{P}\Big(\zeta_{k-1}(P_Y, P_{Y,N}) \geq \frac{\epsilon}{2}\Big) \\
\leq\ &\alpha_1 \exp(-\beta_1 N) + \alpha_2 \exp(-\beta_2 N),
\end{aligned}$$

where the last inequality follows from Theorem 3; α_1 depends on P_ξ, b, and s; β_1 depends on P_ξ, b, s, and ϵ; α_2 depends on P_Y and c; and β_2 depends on P_Y, c, and ϵ.

The second and third probability inequalities can be analogously verified and thus we omit the proof here. □

4.3. Data-Driven DRO Problems with FM Ball

A general stochastic optimization model can be formulated as

$$\min_{x \in X} \mathbb{E}_P[g(x, \xi)], \tag{10}$$

where $g : X \times \Xi \to \mathbb{R}$, $X \subseteq \mathbb{R}^n$, and $\Xi \subseteq \mathbb{R}^s$ is the support set of ξ. The sample average approximation (SAA) is usually used to solve Problem (10) numerically. The SAA method acquiescently assumes that we can generate any number of samples based on P. To better approximate Problem (10), a large sample size is needed [28]. However, in practice, the true probability distribution P cannot be known exactly, and thus we cannot generate a sufficiently large number of samples to make the SAA method well-defined, due to the expensive cost for more samples. However, it is possible for us to obtain a limited number of samples or scenarios, such as historical data. Under these settings, the data-driven DRO model is proposed [18,29,30]. The natural idea is to use the partial information to construct an ambiguity set such that the true probability distribution is included in the ambiguity set. As pointed out in [18], under certain conditions, it offers powerful out-of-sample performance guarantees.

For further discussion, we denote the limited finite samples by $\xi^1, \xi^2, \cdots, \xi^N$ and the corresponding empirical distribution by P_N. Since the number of samples N is limited, we cannot adopt the classical SAA method, which requires that the sample size tends to infinity. However, we can use the limited information to construct a set of probability measures which contains the true one, that is, the ambiguity set. In this subsection, we consider the following FM ball-based ambiguity set:

$$\mathbb{B}_r(P_N) := \{Q \in \mathcal{P}(\Xi) : \zeta_p(Q, P_N) \leq r\},$$

where $p \geq 1$, the positive constant r stands for the confidence parameter determined by the decision maker. Then, we have the data-driven DRO problem with the FM ball-based ambiguity set of Problem (10) as follows:

$$\min_{x \in X} \sup_{Q \in \mathbb{B}_r(P_N)} \mathbb{E}_Q[g(x, \xi)]. \tag{11}$$

It is common for us to see that the Wasserstein ball is used to build the ambiguity set, for example, [18]. To further explain the reasonability and motivations for us to adopt the FM metric, we have the following comments.

Remark 3. *As we know, a key issue for DRO problems is how to build the ambiguity set. Different kinds of ambiguity sets have been proposed, such as moment information [31], ζ-ball [32], and so on. Of course, the FM metric, as a specific case of the ζ-structure probability metric, can be employed to construct the ambiguity set.*

More importantly, the decision maker can utilize the limited empirical distribution P_N to obtain an approximate optimal value, say $v(P_N)$. By prior experience, the decision maker usually has some confidence, measured by the derivation constant $\hat{r} > 0$, that the true optimal value, denoted by $v(P)$, locates in the interval $[v(P_N) - \hat{r}, v(P_N) + \hat{r}]$. Frequently, $g(x, \xi)$ is locally Lipschitz continuous in the following sense:

$$\left|g(x,\xi) - g(x,\xi')\right| \leq L \max\{1, \|\xi\|, \|\xi'\|\}^{p-1} \|\xi' - \xi\|$$

for some positive constant L. A typical example is that $g(x, \xi)$ is the objective function of the two-stage stochastic programming problem, here $p = 2$ (see [11], Proposition 3.2). Then, we have the quantitative relationship:

$$|v(P) - v(P_N)| \leq L\zeta_p(P, P_N).$$

Therefore, it is reasonable for the decision maker to consider the ambiguity set

$$\{Q \in \mathcal{P}(\Xi) : \zeta_p(Q, P_N) \leq r := \hat{r}/L\}.$$

Moreover, since $\zeta_p(P, P_N) \to 0$ with probability 1, as $N \to \infty$ (see Theorem 2), P must be included in $\{Q \in \mathcal{P}(\Xi) : \zeta_p(Q, P_N) \leq r\}$ for suitable N and r.

Finally, we have $\zeta_p(P, Q) \geq \mathbb{D}_W(P, Q)$. The equality holds if $p = 1$. Thus,

$$\{Q \in \mathcal{P}(\Xi) : \zeta_p(Q, P_N) \leq r\} \subseteq \{Q \in \mathcal{P}(\Xi) : \mathbb{D}_W(Q, P_N) \leq r\}.$$

This tells us that the ambiguity set constructed by the FM ball is tighter than that constructed with the Wasserstein ball.

All these arguments motivate us to consider the data-driven DRO problem with the FM ball-based ambiguity set.

We use v^* and v_N to denote the optimal values of Problems (10) and (11), respectively.

To quantify the out-of-sample performance of the data-driven DRO problem (11), we examine the following probability

$$\mathbb{P}(\mathbb{E}_P[g(x_N, \xi)] \leq v_N),$$

where x_N is any optimal solution of Problem (11). Of course, we hope that, for sufficiently small $\epsilon > 0$, there exists a finite positive integer N_0 such that

$$\mathbb{P}(\mathbb{E}_P[g(x_N, \xi)] \leq v_N) \geq 1 - \epsilon \tag{12}$$

for any $N \geq N_0$.

If P satisfies $\zeta_p(P, P_N) < r$, we have $P \in \mathbb{B}_r(P_N)$. Thus,

$$\mathbb{E}_P[g(x,\xi)] \leq \sup_{Q \in \mathbb{B}_r(P_N)} \mathbb{E}_Q[g(x,\xi)]$$

for any $x \in X$, which of course implies that

$$\mathbb{E}_P[g(x_N,\xi)] \leq \sup_{Q \in \mathbb{B}_r(P_N)} \mathbb{E}_Q[g(x_N,\xi)] = v_N.$$

From Theorem 3, we have for any $r \in (0, 1/2]$ that

$$\mathbb{P}(\zeta_p(P, P_N) \geq r) \leq \hat{\alpha} \exp(-\hat{\beta}(r)N),$$

here we use the notation $\hat{\beta}(r) > 0$ to stress the dependence of $\hat{\beta}$ on r. Consequently,

$$\begin{aligned}\mathbb{P}(\zeta_p(P, P_N) \le r) &\ge \mathbb{P}(\zeta_p(P, P_N) < r)\\ &\ge 1 - \hat{\alpha}\exp(-\hat{\beta}(r)N).\end{aligned}$$

A sufficient condition to satisfy (12) is

$$1 - \hat{\alpha}\exp(-\hat{\beta}(r)N) \ge 1 - \epsilon,$$

which is equivalent to

$$N \ge \frac{1}{\hat{\beta}(r)}\log\left(\frac{\hat{\alpha}}{\epsilon}\right).$$

Denote

$$N_0 = \left\lceil \frac{1}{\hat{\beta}(r)}\log\left(\frac{\hat{\alpha}}{\epsilon}\right)\right\rceil, \tag{13}$$

where $\lceil\cdot\rceil$ stands for rounding up to an integer. Sometimes, we use the notation $N_0(\epsilon, r)$ to stress the dependence of N_0 on ϵ and r.

Summarizing the above discussions, we obtain the following so-called finite sample guarantee property (see also [18,30]).

Proposition 3 (Finite sample guarantee)**.** *Let $P \in \mathcal{P}_p(\Xi)$ $(p \ge 1)$ and Assumption 1 hold for some $b > p$. For any $r \in (0, 1/2]$, $\epsilon \in (0, 1)$ and N_0 defined in (13), we have that (12) holds for every $N \ge N_0$.*

Proposition 3 tells us, for the fixed confidence parameter r, at least how large the sample size should be to ensure the significance level ϵ. Now we slightly modify model (11) and consider the following data-driven DRO problem:

$$\min_{x\in X}\sup_{Q\in\mathbb{B}_{r_N}(P_N)}\mathbb{E}_Q[g(x,\xi)], \tag{14}$$

where $r_N > 0$ and $r_N \to 0$ as $N \to \infty$. It reflects the natural fact that the decision maker becomes more confident with more information. Meanwhile, the model (11) emphasizes the fixed limited information. We use $\hat{v}_N$ to denote the optimal value of Problem (14). In what follows, we investigate the asymptotic consistency whenever N tends to infinity. To this end, we need the following lemma.

Lemma 6. *Let $\{A_N\}$ and $\{B_N\}$ be two sequences of random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $\{B_N\}$ converges almost surely and*

$$\mathbb{P}\{A_N \le B_N\} \ge 1 - \kappa_N \tag{15}$$

for $N \in \mathbb{N}$, where $\kappa_N \in (0, 1)$ with $\sum_{N=1}^{\infty}\kappa_N < +\infty$, we have

$$\mathbb{P}\left\{\liminf_{N\to\infty} A_N \le \lim_{N\to\infty} B_N\right\} = 1.$$

Proof. We prove by contradiction. That is, we assume that

$$\mathbb{P}\left\{\liminf_{N\to\infty} A_N \le \lim_{N\to\infty} B_N\right\} < 1.$$

This implies that there exists a subset $\hat{\Omega} \subseteq \Omega$ with $\mathbb{P}(\hat{\Omega}) > 0$ such that

$$\liminf_{N\to\infty} A_N(\hat{\omega}) > \lim_{N\to\infty} B_N(\hat{\omega})$$

for every $\hat{\omega} \in \hat{\Omega}$. Define the sequence $\{\Omega_N\}$ as

$$\Omega_N = \{\omega \in \Omega : A_N(\omega) > B_N(\omega)\}$$

for $N \in \mathbb{N}$. Obviously, according to (15), we have $\mathbb{P}(\Omega_N) \leq \kappa_N$, which implies that

$$\sum_{N=1}^{\infty} \mathbb{P}(\Omega_N) \leq \sum_{N=1}^{\infty} \kappa_N < +\infty.$$

Then, we can always choose a sufficiently large $\hat{N} \in \mathbb{N}$, such that

$$\sum_{N=\hat{N}}^{\infty} \mathbb{P}(\Omega_N) < \mathbb{P}(\hat{\Omega}).$$

Choose

$$\omega \in \hat{\Omega} \backslash \left(\cup_{N=\hat{N}}^{\infty} \Omega_N \right),$$

and we have

$$A_N(\omega) \leq B_N(\omega)$$

for all $N \geq \hat{N}$, which implies that

$$\liminf_{N \to \infty} A_N(\omega) \leq \lim_{N \to \infty} B_N(\omega).$$

This contradicts the definition of $\hat{\Omega}$. We complete the proof. □

The following proposition states that the optimal value and optimal solution set of the data-driven DRO problem (14) converge to those of the original Problem (10), which verifies the reasonability of our data-driven DRO model (14).

Proposition 4 (Asymptotic consistency)**.** *Suppose that $g(x, \xi)$ is locally Lipschitz continuous in the following sense:*

$$\left| g(x, \xi) - g(x, \xi') \right| \leq L \max\{1, \|\xi\|, \|\xi'\|\}^{p-1} \|\xi' - \xi\|$$

for every $x \in X$, where $L > 0$ and $p \geq 1$. Let x_N be any optimal solution of Problem (14). Then, the following assertions hold:

(i) *$\hat{v}_N \to v^*$ with probability 1, as $N \to \infty$;*

(ii) *If, moreover, X is closed, $g(\cdot, \xi)$ is lower semicontinuous for every $\xi \in \Xi$ and $g(x, \xi)$ dominates some P-integrable function uniformly with respect to $x \in X$, then, any accumulation point of $\{x_N\}$ is an optimizer of Problem (10) almost surely.*

Proof. Part (i): Notice that

$$\begin{aligned} |\hat{v}_N - v^*| &= \left| \min_{x \in X} \sup_{Q \in \mathbb{B}_{r_N}(P_N)} \mathbb{E}_Q[g(x, \xi)] - \min_{x \in X} \mathbb{E}_P[g(x, \xi)] \right| \\ &\leq \max \Bigg\{ \sup_{Q \in \mathbb{B}_{r_N}(P_N)} \mathbb{E}_Q[g(x^*, \xi)] - \mathbb{E}_P[g(x^*, \xi)], \\ &\qquad \mathbb{E}_P[g(x_N, \xi)] - \sup_{Q \in \mathbb{B}_{r_N}(P_N)} \mathbb{E}_Q[g(x_N, \xi)] \Bigg\}, \end{aligned}$$

where x^* and x_N are any optimal solutions of Problems (10) and (14), respectively. For the first term on the right-hand side, we have

$$\begin{aligned}\sup_{Q\in\mathbb{B}_{r_N}(P_N)} \mathbb{E}_Q[g(x^*,\xi)] - \mathbb{E}_P[g(x^*,\xi)] &\le \mathbb{E}_{Q_N}[g(x^*,\xi)] - \mathbb{E}_P[g(x^*,\xi)] + \delta_N \\ &\le L\zeta_p(Q_N,P) + \delta_N \\ &\le Lr_N + \delta_N \to 0,\end{aligned}$$

almost surely, as $N \to \infty$, where the first inequality is due to the definition of supremum for some $\delta_N > 0$ with $\delta_N \to 0$ almost surely and $Q_N \in \mathbb{B}_{r_N}(P_N)$; the second inequality follows from the definition of FM metric. Similarly, we can derive

$$\mathbb{E}_P[g(x_N,\xi)] - \sup_{Q\in\mathbb{B}_{r_N}(P_N)} \mathbb{E}_Q[g(x_N,\xi)] \to 0$$

almost surely, as $N \to \infty$. Thus, we obtain that

$$|\hat{v}_N - v^*| \to 0 \text{ almost surely, as } N \to \infty.$$

Part (ii): Without loss of generality, in the following discussion, we assume $x_N \to \hat{x}$ with probability 1 as $N \to \infty$. Moreover, we select a sequence $\{\epsilon_k\}$ with $\epsilon_k \in (0,1)$ and $\sum_{k=1}^{\infty} \epsilon_k < \infty$. According to (12), for each pair (ϵ_k, r_k) with r_k defined in (14), we can select an $N_k \ge N_0(\epsilon_k, r_k)$ ($N_0(\epsilon_k, r_k)$ is defined in (13)) such that

$$\mathbb{P}\big(\mathbb{E}_P[g(x_{N_k},\xi)] \le \hat{v}_{N_k}\big) \ge 1 - \epsilon_k.$$

We know from Lemma 6 and assertion (i) that

$$\mathbb{P}\left(\liminf_{k\to\infty} \mathbb{E}_P[g(x_{N_k},\xi)] \le \lim_{k\to\infty} \hat{v}_{N_k} = v^*\right) = 1. \tag{16}$$

Then, the following inequalities hold almost surely:

$$v^* \overset{(a)}{\le} \mathbb{E}_P[g(\hat{x},\xi)] \overset{(b)}{\le} \mathbb{E}_P\left[\liminf_{k\to\infty} g(x_{N_k},\xi)\right] \overset{(c)}{\le} \liminf_{k\to\infty} \mathbb{E}_P\left[g(x_{N_k},\xi)\right] \overset{(d)}{\le} v^*,$$

where (a) follows from $\hat{x} \in X$ due to the closedness of X; (b) follows from the lower semi-continuity of $g(\cdot,\xi)$ for every $\xi \in \Xi$; (c) is due to Fatou's lemma; (d) follows from (16). □

Remark 4. *Propositions 3 and 4 establish the finite sample guarantee and the asymptotic consistency, which are two desirable properties of the data-driven DRO problem [18,30]. Different from the existing results in [18] where the Wasserstein ball is used to construct the ambiguity set, we adopt the FM ball. Due to the feature of Wasserstein metric, to ensure the existence of the significance parameter ϵ, they explicitly derived the radius $r(\epsilon, N)$ depending on ϵ and N and the finite sample size $N_0(\epsilon)$ depending only on ϵ. In Proposition 3, we view both r and ϵ as parameters because $\hat{\beta}$ couples with ϵ implicitly in Theorem 3. Moreover, the assumptions for the asymptotic consistency (Proposition 4) are different from those in [18] (Theorem 3.6), where the upper semicontinuity and linear growth were employed. Here we use the locally Lipschitz continuity but a weaker assumption of the lower bound. Specially, Ref. [18] (Theorem 3.6) employs Borel–Cantelli lemma to obtain*

$$\mathbb{P}\{\mathbb{E}_P[g(x_N,\xi)] \le \hat{v}_N \text{ for sufficiently large } N\} = 1.$$

This is not applicable for our case, so we need Lemma 6.

4.4. Discrete Approximation for DRO Problems with General Moment Information

We consider the following general DRO problem:

$$\min_{x\in X}\sup_{P\in\mathcal{P}}\mathbb{E}_P[h(x,\xi)], \tag{17}$$

where $X\subseteq\mathbb{R}^n$ is a compact set, $h: X\times\Xi\to\mathbb{R}$, $\mathcal{P}:=\{P\in\mathcal{P}(\Xi):\mathbb{E}_P[\Gamma(\xi)]\in\mathcal{K}\}$, $\mathcal{K}$ is a closed and convex set in the Cartesian product of some finite dimensional vector and/or matrix spaces, and Γ is a general mapping on Ξ. We implicitly assume that, for each $x\in X$, $\mathbb{E}_P[h(x,\xi)]<+\infty$ for all $P\in\mathcal{P}$.

The above ambiguity set $\mathcal{P}$ is very general, and it covers almost all the available ambiguity sets with moment information (see, e.g., [19], Examples 3–5). Zhang et al. discussed in [33] the quantitative stability of the DRO problem with a general moment information ambiguity set. There are usually two ways to numerically solve Problem (17): One is to use some kind of duality argument to reformulate Problem (17) as a solvable Problem [18,31]; the other is to discretize the ambiguity set, which leads to a saddle point problem in the finite dimensional space [19]. For instance, the discrete approximation in [19] is conducted under a bounded support set. In this part, by employing our results in Section 3, we consider the discrete approximation for problem (17) under weaker conditions.

Denote by $\hat{\mathcal{P}}_N$ the collection of all discrete distributions which have at most N supporting elements, that is,

$$\hat{\mathcal{P}}_N=\left\{\sum_{i=1}^N p_i\mathbf{1}_{\xi^i}:\sum_{i=1}^N p_i=1,p_j\geq 0,\xi^j\in\Xi,j=1,2,\cdots,N\right\}.$$

We define the discrete approximation of $\mathcal{P}$ as

$$\mathcal{P}_N=\{Q\in\hat{\mathcal{P}}_N:\mathbb{E}_Q[\Gamma(\xi)]\in\mathcal{K}\}.$$

Obviously, $\mathcal{P}_N\subseteq\mathcal{P}$. Then, the discrete approximation of Problem (17) can be written as

$$\min_{x\in X}\sup_{P\in\mathcal{P}_N}\mathbb{E}_P[h(x,\xi)]. \tag{18}$$

We use $v(\mathcal{P})$ and $S(\mathcal{P})$ to denote the optimal value and optimal solution set of Problem (17). $v(\mathcal{P}_N)$ and $S(\mathcal{P}_N)$ are the optimal value and optimal solution set of Problem (18). To make sense of the discrete approximation, we hope that Problem (18) can approximately solve Problem (17) when N is sufficiently large.

To continue the following discussion, we define the growth function of Problem (17) $\psi_{\mathcal{P}}:\mathbb{R}_+\to\mathbb{R}_+$ as

$$\psi_{\mathcal{P}}(\tau)=\inf\left\{\sup_{P\in\mathcal{P}}\mathbb{E}_P[h(x,\xi)]-v(\mathcal{P}):d(x,S(\mathcal{P}))\geq\tau,x\in X\right\}$$

and its inverse function is

$$\psi_{\mathcal{P}}^{-1}(t):=\sup\{\tau\in\mathbb{R}_+:\psi_{\mathcal{P}}(\tau)\leq t\}.$$

Thus, the associated conditioning function $\Psi_{\mathcal{P}}:\mathbb{R}_+\to\mathbb{R}_+$ is defined as

$$\Psi_{\mathcal{P}}(\eta):=\eta+\psi_{\mathcal{P}}^{-1}(2\eta).$$

Immediately, we have the following quantitative stability results:

Proposition 5. *Suppose that: (i) $\mathcal{P}\subseteq\mathcal{P}_p(\Xi)$; (ii) $h(x,\xi)\geq g(\xi)$ for each $\xi\in\Xi$ and a measurable function $g:\Xi\to\mathbb{R}$ with $\mathbb{E}_P[g(\xi)]>-\infty$ for any $P\in\mathcal{P}$; (iii) $h(\cdot,\xi)$ is lower semicontinuous for each $\xi\in\Xi$; (iv)*

$$\left|h(x,\xi)-h(x,\xi')\right| \le L_h \max\left\{1,\|\xi\|,\|\xi'\|\right\}^{p-1}\|\xi'-\xi\|$$

for each $x \in X$. *Then,* $S(\mathcal{P}) \neq \emptyset$ *and*

$$|v(\mathcal{P})-v(\mathcal{P}_N)| \le L_h\zeta_p(\mathcal{P},\mathcal{P}_N),$$
$$\emptyset \neq S(\mathcal{P}_N) \subseteq S(\mathcal{P}) + \Psi_{\mathcal{P}}(L_h\zeta_p(\mathcal{P},\mathcal{P}_N)).$$

Proof. Since $h(x,\xi) \ge g(\xi)$ and $h(\cdot,\xi)$ is lower semicontinuous for each $\xi \in \Xi$, we have from Fatou's lemma that

$$\liminf_{k\to\infty} \mathbb{E}_P[h(x_k,\xi)] \ge \mathbb{E}_P\left[\liminf_{k\to\infty} h(x_k,\xi)\right] \ge \mathbb{E}_P[h(\bar{x},\xi)]$$

holds for any $\{x_k\} \subseteq X$ such that $\lim_{k\to\infty} x_k = \bar{x}$. This implies that $\mathbb{E}_P[h(\cdot,\xi)]$ is lower semicontinuous. According to [34] (Lemma 4.1), $\sup_{P\in\mathcal{P}} \mathbb{E}_P[h(\cdot,\xi)]$ is lower semicontinuous too. This, together with the compactness of X, ensures that $S(\mathcal{P}) \neq \emptyset$. Similarly, we can prove that $S(\mathcal{P}_N) \neq \emptyset$.

Note that

$$\begin{aligned}
|v(\mathcal{P})-v(\mathcal{P}_N)| &\overset{(a)}{=} \min_{x\in X}\sup_{P\in\mathcal{P}} \mathbb{E}_P[h(x,\xi)] - \min_{x\in X}\sup_{Q\in\mathcal{P}_N} \mathbb{E}_Q[h(x,\xi)] \\
&\le \max_{x\in X}\left(\sup_{P\in\mathcal{P}} \mathbb{E}_P[h(x,\xi)] - \sup_{Q\in\mathcal{P}_N} \mathbb{E}_Q[h(x,\xi)]\right) \\
&= \max_{x\in X}\left(\sup_{P\in\mathcal{P}}\inf_{Q\in\mathcal{P}_N} \left(\mathbb{E}_P[h(x,\xi)] - \mathbb{E}_Q[h(x,\xi)]\right)\right) \\
&\overset{(b)}{\le} \max_{x\in X}\left(\sup_{P\in\mathcal{P}}\inf_{Q\in\mathcal{P}_N} L_h\zeta_p(P,Q)\right) \\
&= L_h\zeta_p(\mathcal{P},\mathcal{P}_N),
\end{aligned}$$

where (a) follows from the fact $\mathcal{P}_N \subseteq \mathcal{P}$; (b) is due to the definition of the pth order FM metric.

Finally, based on the first assertion, the inclusion for the optimal solution sets can be analogously derived as that in [11]. □

For simplicity as well as to show the linear relationship more clearly, we write $\mathbb{E}_P[\Gamma(\xi)]$ as $\langle P,\Gamma(\xi)\rangle$ in what follows. We need the following technical assumption to proceed.

Assumption 3 (see [19]). *The system* $\mathcal{P} := \{P : \mathbb{E}_P[\Gamma(\xi)] \in \mathcal{K}\}$ *satisfies the following Slater condition:*

$$\left\langle \tilde{P},\Gamma(\xi)\right\rangle + \delta\mathbb{B} \subseteq \mathcal{K}$$

for some $\tilde{P} \in \mathcal{P}(\Xi)$ *and* $\delta > 0$.

Proposition 6. *Suppose that Assumption 3 holds and* $\mathcal{P} \subseteq \mathcal{P}_p(\Xi)$. *Then, there exists an* $\Omega_0 \subseteq \Omega$ *with* $\mathbb{P}(\Omega_0) = 0$, *such that for any* $\hat{\delta} < \delta$ *and* $\omega \in \Omega\backslash\Omega_0$, *we have*

$$\zeta_p(Q,\mathcal{P}_{2N}) \le \frac{\zeta_p(Q,\tilde{P})+1}{\hat{\delta}} \inf_{K\in\mathcal{K}}\|K-\langle Q,\Gamma(\xi)\rangle\|$$

for any $Q \in \hat{\mathcal{P}}_N$ *and* $N \ge \hat{N}(\hat{\delta},\omega)$, *where* $\hat{N}(\hat{\delta},\omega)$ *is a positive integer depending on* $\hat{\delta}$ *and* ω.

Proof. Let the empirical approximation of $\tilde{P}$ be $\tilde{P}_N$. Then, we have from the law of large numbers that

$$\left\langle \tilde{P}_N,\Gamma(\xi)\right\rangle \to \left\langle \tilde{P},\Gamma(\xi)\right\rangle$$

with probability 1, as $N \to \infty$. Equivalently, there exists an $\Omega_1 \subseteq \Omega$ with $\mathbb{P}(\Omega_1) = 0$, such that for any $\hat{\delta} < \delta$ and $\omega \in \Omega \backslash \Omega_1$, we have

$$\left\| \left\langle \tilde{P}_N(\omega), \Gamma(\xi) \right\rangle - \left\langle \tilde{P}, \Gamma(\xi) \right\rangle \right\| \leq \delta - \hat{\delta}$$

for $N \geq N_1(\hat{\delta}, \omega)$. This implies that

$$\left\langle \tilde{P}_N(\omega), \Gamma(\xi) \right\rangle \in \left\langle \tilde{P}, \Gamma(\xi) \right\rangle + (\delta - \hat{\delta})\mathbb{B},$$

or equivalently,

$$\left\langle \tilde{P}_N(\omega), \Gamma(\xi) \right\rangle + \hat{\delta}\mathbb{B} \subseteq \left\langle \tilde{P}, \Gamma(\xi) \right\rangle + \delta\mathbb{B} \subseteq \mathcal{K}$$

for $N \geq N_1(\hat{\delta}, \omega)$, where $\mathbb{B}$ is the unit closed ball in the space of $\mathcal{K}$.

Notice that $\tilde{P}_N \in \mathcal{P}_N$, and hence, for $N \geq N_1(\hat{\delta}, \omega)$, the Slater condition holds with respect to $\hat{\delta}$ for the system

$$\mathcal{P}_N := \{P \in \hat{\mathcal{P}}_N : \mathbb{E}_P[\Gamma(\xi)] \in \mathcal{K}\}.$$

Now we define, for any $Q \in \hat{\mathcal{P}}_N$, $\rho_Q = \inf_{K \in \mathcal{K}} \|K - \langle Q, \Gamma(\xi) \rangle\|$ and

$$\bar{Q} = \left(1 - \frac{\rho_Q}{\rho_Q + \hat{\delta}}\right) Q + \frac{\rho_Q}{\rho_Q + \hat{\delta}} \tilde{P}_N.$$

Obviously, we have $\bar{Q} \in \hat{\mathcal{P}}_{2N}$. Similar to that proof of [19] (Theorem 2), we again obtain $\bar{Q} \in \mathcal{P}_{2N}$. Then, we have

$$\begin{aligned}
\zeta_p(Q, \mathcal{P}_{2N}) \leq \zeta_p(Q, \bar{Q}) &= \sup_{g \in \mathcal{G}_{\mathrm{FM}_p}} \left| \langle Q, g \rangle - \left\langle \bar{Q}, g \right\rangle \right| \\
&= \frac{\rho_Q}{\rho_Q + \hat{\delta}} \sup_{g \in \mathcal{G}_{\mathrm{FM}_p}} \left| \langle Q, g \rangle - \left\langle \tilde{P}_N(\omega), g \right\rangle \right| \\
&\leq \frac{\zeta_p(Q, \tilde{P}_N(\omega))}{\hat{\delta}} \inf_{K \in \mathcal{K}} \|K - \langle Q, \Gamma(\xi) \rangle\| \\
&\leq \frac{\zeta_p(Q, \tilde{P}) + \zeta_p(\tilde{P}, \tilde{P}_N(\omega))}{\hat{\delta}} \inf_{K \in \mathcal{K}} \|K - \langle Q, \Gamma(\xi) \rangle\| \\
&\leq \frac{\zeta_p(Q, \tilde{P}) + 1}{\hat{\delta}} \inf_{K \in \mathcal{K}} \|K - \langle Q, \Gamma(\xi) \rangle\|
\end{aligned}$$

for $N \geq N_2(\omega)$ and $\omega \in \Omega \backslash \Omega_2$ with $\mathbb{P}(\Omega_2) = 0$, where the last inequality follows from Theorem 2.

Finally, letting $\Omega_0 := \Omega_1 \cup \Omega_2$ and $\hat{N}(\hat{\delta}, \omega) := \max\{N_1(\hat{\delta}, \omega), N_2(\omega)\}$ completes the proof. □

The following theorem states that the discrete approximation ambiguity set $\mathcal{P}_N$ converges to $\mathcal{P}$ as $N \to \infty$ in the sense of FM metrics.

Theorem 7. *Suppose that: (i) Assumption 3 holds; (ii)* $\mathcal{P} \subseteq \mathcal{P}_p(\Xi)$*; (iii)*

$$\sup_{P \in \mathcal{P}} \|\mathbb{E}_P[\Gamma(\xi)]\| < +\infty \text{ and } C_{\mathcal{P}} := \sup_{P, Q \in \mathcal{P}} \zeta_p(P, Q) < +\infty.$$

Then,

$$\lim_{N \to \infty} \zeta_p(\mathcal{P}, \mathcal{P}_N) = 0$$

with probability 1.

Proof. For any $P \in \mathcal{P}$, by the triangle inequality, we have

$$\zeta_p(P, \mathcal{P}_N) \leq \zeta_p(P, P_N) + \zeta_p(P_N, \mathcal{P}_N)$$

where P_N is the empirical distribution of P with N samples. Since $P_N \in \hat{\mathcal{P}}_N \subseteq \mathcal{P}$, we know from Proposition 6 that

$$\begin{aligned}\zeta_p(P_N, \mathcal{P}_N) &\leq \frac{\zeta_p(P_N, \tilde{P}) + 1}{\hat{\delta}} \inf_{K \in \mathcal{K}} \|K - \langle P_N, \Gamma(\xi)\rangle\| \\ &\leq \frac{C_{\mathcal{P}} + 1}{\hat{\delta}} \|\langle P, \Gamma(\xi)\rangle - \langle P_N, \Gamma(\xi)\rangle\|\end{aligned}$$

for $N \geq \hat{N}(\hat{\delta}, \omega)$ and almost every $\omega \in \Omega$. Thus, we have

$$\zeta_p(P, \mathcal{P}_N) \leq \zeta_p(P, P_N) + \frac{C_{\mathcal{P}} + 1}{\hat{\delta}} \|\langle P, \Gamma(\xi)\rangle - \langle P_N, \Gamma(\xi)\rangle\|.$$

Subsequently,

$$\begin{aligned}\zeta_p(\mathcal{P}, \mathcal{P}_N) &= \sup_{P \in \mathcal{P}} \zeta_p(P, \mathcal{P}_N) \\ &\leq \sup_{P \in \mathcal{P}} \zeta_p(P, P_N) + \frac{C_{\mathcal{P}} + 1}{\hat{\delta}} \sup_{P \in \mathcal{P}} \|\langle P, \Gamma(\xi)\rangle - \langle P_N, \Gamma(\xi)\rangle\|.\end{aligned}$$

For the first term on the right-hand side, the definition of supremum, the boundedness of $\mathcal{P}$, and Theorem 2 give rise to

$$\lim_{N \to \infty} \sup_{P \in \mathcal{P}} \zeta_p(P, P_N) \leq \lim_{N \to \infty} \zeta_p(P^k, P_N^k) + \epsilon_k = \epsilon_k$$

with probability 1, where $\{P^k\}$ is a sequence included in $\mathcal{P}$ such that

$$\sup_{P \in \mathcal{P}} \zeta_p(P, P_N) \leq \zeta_p(P^k, P_N^k) + \epsilon_k$$

and $\{\epsilon_k\}$ is a positive sequence with $\epsilon_k \to 0$ as $N \to \infty$. Thus, we obtain

$$\lim_{N \to \infty} \sup_{P \in \mathcal{P}} \zeta_p(P, P_N) = 0$$

with probability 1.

Analogously, by the law of large numbers, we can derive that

$$\lim_{N \to \infty} \sup_{P \in \mathcal{P}} \|\langle P, \Gamma(\xi)\rangle - \langle P_N, \Gamma(\xi)\rangle\| = 0$$

with probability 1.

Then, we complete the proof. □

The following corollary shows the reasonability for the approximation of Problem (18) to Problem (17).

Corollary 1. *Under the conditions of Proposition 5 and Theorem 7, we have*

$$\begin{aligned}&|v(\mathcal{P}) - v(\mathcal{P}_N)| \to 0, \\ &d(S(\mathcal{P}_N), S(\mathcal{P})) \to 0.\end{aligned}$$

with probability 1, *as* $N \to \infty$.

Remark 5. *In this subsection, we investigated the discrete approximation of the DRO problem with the general moment information ambiguity set. Compared with the existing work [19], we have*

further weakened the necessary assumptions and extended them to a more general case. Firstly, the Lipschitz continuity of the objective function is required in [19] (Theorem 14) due to the adoption of the Wasserstein metric, so that the upper bound between the discrete approximation of the DRO problem and the original DRO problem can be derived [19] (Proposition 7). We only call for the locally Lipschitz continuity. More importantly, they restricted their discussion to the bounded support set case because the upper bound in [19] (Proposition 7) would be infinity when the support set is unbounded, which is not well defined in this case. However, our support set can be unbounded by employing our convergence results in Section 3.

5. Concluding Remarks

In this study, we investigated different kinds of convergence assertions about data-driven FM metrics and their possible applications. In view of the rich results about Wasserstein metrics (Lemmas 2 and 3), we first established the relationship between the FM metric and the Wasserstein metric (Lemma 4). Based on these results, the non-asymptotic moment estimate (Theorem 1), asymptotic convergence estimate (Theorem 2), and non-asymptotic concentration estimate (Theorem 3) for FM metrics were presented. These convergence assertions for FM metrics were applied to the asymptotic analyses of the empirical approximations of four kinds of stochastic optimization problems. The results sufficiently show the motivations of this study and its importance.

There are still some topics to settle in the future. For example, we leave the numerical tractability for the results in Sections 4.3 and 4.4 for future work.

Author Contributions: Supervision, Z.C.; Writing—original draft, J.J.; Writing—review & editing, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by China Postdoctoral Science Foundation (Grant Number 2020M673117), the National Natural Science Foundation of China (Grant Numbers 11991023 and 11735011).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Proof of Lemma 4. According to the definition of FM metric, we have

$$\zeta_p(P,Q) = \sup_{g\in\mathcal{G}_p} \int_{\Xi} g(\xi)(P-Q)(d\xi).$$

Moreover, it is easy to verify that $g(\xi)$ adding any constant will not change the value of the integral. For simplification of the following discussion, without loss of generality, we hereafter set $g(\xi_0) = 0$ for any fixed $\xi_0 \in \arg\min_{\xi\in\Xi}\|\xi\|$. We denote $M_R = \{\xi \in \Xi : \|\xi\| \le R\}$ and M_R^c is the complementary set of M_R. Since $\mathbb{B}(\mathbf{0}, R) \cap \Xi \neq \emptyset$ and thus $\xi_0 \in M_R$, we have an upper bound estimation of $g(\xi)$ as follows:

$$\begin{aligned} |g(\xi)| = |g(\xi) - g(\xi_0)| &\le \max\{1, \|\xi\|, \|\xi_0\|\}^{p-1}\|\xi - \xi_0\| \\ &\le \max\{1, \|\xi\|, \|\xi_0\|\}^{p-1}(\|\xi\| + \|\xi_0\|). \end{aligned}$$

If $\xi \in M_R^c$, then, $\|\xi\| > R \ge \|\xi_0\|$, and we have the following upper bound of $|g(\xi)|$:

$$|g(\xi)| \le \|\xi\|^{p-1}(\|\xi\| + \|\xi_0\|) \le 2\|\xi\|^p.$$

Then, we continue

$$\begin{aligned}\zeta_p(P,Q) &= \sup_{g\in\mathcal{G}_p}\left(\int_{M_R} g(\xi)(P-Q)(d\xi)+\int_{M_R^c} g(\xi)(P-Q)(d\xi)\right)\\ &\le \sup_{g\in\mathcal{G}_p}\left(R^{p-1}\int_{M_R} h(\xi)(P-Q)(d\xi)+2\int_{M_R^c}\|\xi\|^p(P+Q)(d\xi)\right),\end{aligned}$$

where $h: M_R \to \mathbb{R}$ is defined by $h(\cdot) := g(\cdot)/R^{p-1}$. It is easy to see that h is Lipschitz continuous on M_R with Lipschitz modulus 1. Based on Lemma 1, we can extend h to $\mathbb{R}^s$, and its restriction on Ξ is denoted by $\tilde{h}(\cdot)$. Then, $\tilde{h}(\cdot)$ is Lipschitz continuous on Ξ with Lipschitz modulus 1. Thus, we can continue

$$\begin{aligned}\zeta_p(P,Q) \le \sup_{g\in\mathcal{G}_p}\Bigg(&R^{p-1}\int_{\Xi}\tilde{h}(\xi)(P-Q)(d\xi)-R^{p-1}\int_{M_R^c}\tilde{h}(\xi)(P-Q)(d\xi)\\ &+2\int_{M_R^c}\|\xi\|^p(P+Q)(d\xi)\Bigg).\end{aligned}$$

So, for any $\xi\in\Xi$, we have

$$\left\|\tilde{h}(\xi)-\tilde{h}(\xi_0)\right\| \le \|\xi-\xi_0\|.$$

Note that $\tilde{h}(\xi_0) = g(\xi_0)/R^{p-1} = 0$, so

$$\begin{aligned}\left\|\tilde{h}(\xi)\right\| &= \left\|\tilde{h}(\xi)-\tilde{h}(\xi_0)\right\|\\ &\le \|\xi-\xi_0\| \le \|\xi\|+\|\xi_0\|.\end{aligned}$$

Similarly, this means that $\left\|\tilde{h}(\xi)\right\| \le 2\|\xi\|$ for any $\xi\in M_R^c$. Then, we continue

$$\begin{aligned}\zeta_p(P,Q) &\le R^{p-1}\mathbb{D}_W(P,Q)+2R^{p-1}\int_{M_R^c}\|\xi\|(P+Q)(d\xi)+2\int_{M_R^c}\|\xi\|^p(P+Q)(d\xi)\\ &\le R^{p-1}\mathbb{D}_W(P,Q)+4\int_{M_R^c}\|\xi\|^p(P+Q)(d\xi).\end{aligned}$$

The proof is complete. □

Proof of Theorem 1. According to Lemma 4 with $Q = P_N$, we have

$$\begin{aligned}\mathbb{E}[\zeta_p(P,P_N)] &\le \mathbb{E}\left[R^{p-1}\mathbb{D}_W(P,P_N)+4\int_{M_R^c}\|\xi\|^p(P+P_N)(d\xi)\right]\\ &= R^{p-1}\mathbb{E}[\mathbb{D}_W(P,P_N)]+4\int_{M_R^c}\|\xi\|^p P(d\xi)+4\mathbb{E}\left[\int_{M_R^c}\|\xi\|^p P_N(d\xi)\right].\end{aligned}$$

Moreover, since

$$\begin{aligned}\mathbb{E}\left[\int_{M_R^c}\|\xi\|^p P_N(d\xi)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{M_R^c}(\xi^i)\cdot\left\|\xi^i\right\|^p\right]\\ &= \mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\cdot\|\xi\|^p\right],\end{aligned}$$

we obtain

$$\mathbb{E}[\zeta_p(P,P_N)] \le R^{p-1}\mathbb{E}[\mathbb{D}_W(P,P_N)]+8\mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\cdot\|\xi\|^p\right].$$

Meanwhile, we know from Lemma 2 that

$$\mathbb{E}[\mathbb{D}_W(P_N,P)] \le C\left(\mathbb{E}_P[\|\xi\|^p]\right)^{1/p}\alpha(N),$$

where $\alpha(N) = \left(N^{-1/\max\{2,s\}}\log(1+N) + N^{-(p-1)/p}\right) \to 0$ as $N \to \infty$. Then, we take $R = \alpha(N)^{-1/p}$. Since $R \to +\infty$ as $N \to \infty$, we have $R \geq 1$ and $\mathbb{B}(\mathbf{0}, R) \cap \Xi \neq \emptyset$ for sufficiently large N. Therefore, we have

$$\begin{aligned}\mathbb{E}[\zeta_p(P, P_N)] &\leq R^{p-1}\mathbb{E}[\mathbb{D}_W(P, P_N)] + 8\mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\cdot\|\xi\|^p\right] \\ &\leq C\left(\mathbb{E}_P[\|\xi\|^p]\right)^{1/p}\alpha(N)^{1/p} + 8\mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\cdot\|\xi\|^p\right]\end{aligned}$$

and

$$\mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\cdot\|\xi\|^p\right] \to 0, \text{ as } N \to \infty.$$

Thus, letting

$$\beta_N = C\left(\mathbb{E}_P[\|\xi\|^p]\right)^{1/p}\alpha(N)^{1/p} + 8\mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\cdot\|\xi\|^p\right]$$

completes the proof. □

Proof of Theorem 2. To prove this assertion, we need to verify that: for any $\epsilon > 0$, there exists a positive number $N(\epsilon, \omega)$ such that

$$\zeta_p(P, P_N) \leq \epsilon \tag{A1}$$

as $N \geq N(x, \omega)$ for almost every $\omega \in \Omega$. Notice from Lemma 4 that

$$\zeta_p(P, P_N) \leq R^{p-1}\mathbb{D}_W(P, P_N) + 4\int_{M_R^c}\|\xi\|^p(P + P_N)(d\xi)$$

for sufficiently large R.

We can deduce from $P \in \mathcal{P}_p(\Xi)$ and Lemma 2 that

$$\int_{M_R^c}\|\xi\|^p P(d\xi) \to 0 \text{ as } R \to +\infty$$

and

$$\int_{M_R^c}\|\xi\|^p P_N(d\xi) \to \int_{M_R^c}\|\xi\|^p P(d\xi) \text{ as } N \to \infty$$

with probability 1. Thus, there always exists a sufficiently large positive number $R(\epsilon)$ such that

$$\int_{M_R^c}\|\xi\|^p P(d\xi) \leq \frac{\epsilon}{32} \tag{A2}$$

as $R \geq R(\epsilon)$. Moreover, there exists a positive number $N_1 := N_1(\epsilon, \omega)$ such that

$$\left|\int_{M_R^c}\|\xi\|^p P_N(d\xi) - \int_{M_R^c}\|\xi\|^p P(d\xi)\right| \leq \frac{\epsilon}{16}$$

as $N \geq N_1$ with probability 1, which implies from the triangle inequality that

$$\int_{M_R^c}\|\xi\|^p P_N(d\xi) \leq \frac{3\epsilon}{32} \tag{A3}$$

with probability 1. Combining (A2) with (A3), we have

$$4\int_{M_R^c}\|\xi\|^p(P + P_N)(d\xi) \leq 4\left(\frac{\epsilon}{32} + \frac{3\epsilon}{32}\right) = \frac{\epsilon}{2} \tag{A4}$$

as $R \geq R(\epsilon)$ and $N \geq N_1$, with probability 1.

On the other hand, we know from the Glivenko-Cantelli theorem [35] that

$$\mathbb{D}_W(P, P_N) \to 0 \text{ as } N \to \infty \text{ with probability 1,}$$

which implies that there exists a positive number $N_2 := N_2(\epsilon, \omega)$ such that

$$R^{p-1}\mathbb{D}_W(P, P_N) \leq \frac{\epsilon}{2} \tag{A5}$$

when $N \geq N_2$.

(A5) and (A4) mean (A1) by letting $N = \max\{N_1, N_2\}$. This completes the proof. □

Proof of Theorem 3. We know from Lemma 4 that

$$\zeta_p(P, P_N) \leq R^{p-1}\mathbb{D}_W(P, P_N) + 4\int_{M_R^c} \|\xi\|^p (P + P_N)(d\xi).$$

Then, we have

$$\begin{aligned}\mathbb{P}(\zeta_p(P, P_N) \geq \epsilon) &\leq \mathbb{P}\left(R^{p-1}\mathbb{D}_W(P, P_N) + 4\int_{M_R^c} \|\xi\|^p (P + P_N)(d\xi) \geq \epsilon\right) \\ &\leq \mathbb{P}\left(R^{p-1}\mathbb{D}_W(P, P_N) \geq \epsilon/2\right) + \mathbb{P}\left(4\int_{M_R^c} \|\xi\|^p (P + P_N)(d\xi) \geq \epsilon/2\right).\end{aligned}$$

For the first term, we know from (3) that

$$\begin{aligned}\mathbb{P}\left(R^{p-1}\mathbb{D}_W(P, P_N) \geq \epsilon/2\right) &= \mathbb{P}\left(\mathbb{D}_W(P, P_N) \geq \epsilon/(2R^{p-1})\right) \\ &\leq \alpha \exp\left(-\beta N\left(\frac{\epsilon}{2R^{p-1}}\right)^{\max\{4,s\}}\right).\end{aligned} \tag{A6}$$

We, in what follows, consider the estimation of the second term:

$$\mathbb{P}\left(4\int_{M_R^c} \|\xi\|^p (P + P_N)(d\xi) \geq \epsilon/2\right).$$

Since $P \in \mathcal{P}_p(\Xi)$, we can choose a sufficiently large $R = R(\epsilon)$ such that

$$\int_{M_R^c} \|\xi\|^p P(d\xi) \leq \frac{\epsilon}{32}.$$

Then, we have

$$\begin{aligned}&\mathbb{P}\left(4\int_{M_R^c} \|\xi\|^p (P + P_N)(d\xi) \geq \epsilon/2\right) \\ &\leq \mathbb{P}\left(\int_{M_R^c} \|\xi\|^p P_N(d\xi) \geq \frac{3\epsilon}{32}\right) \\ &= \mathbb{P}\left(\int_{\Xi} \mathbf{1}_{M_R^c}(\xi)\|\xi\|^p P_N(d\xi) \geq \frac{3\epsilon}{32}\right) \\ &= \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{M_R^c}(\xi^i)\left\|\xi^i\right\|^p \geq \frac{3\epsilon}{32}\right) \\ &= \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{M_R^c}(\xi^i)\left\|\xi^i\right\|^p - \mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\|\xi\|^p\right] \geq \frac{3\epsilon}{32} - \mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\|\xi\|^p\right]\right) \\ &\leq \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{M_R^c}(\xi^i)\left\|\xi^i\right\|^p - \mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\|\xi\|^p\right] \geq \frac{\epsilon}{16}\right).\end{aligned}$$

Furthermore, according to Cramér's large deviation theorem, we have

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{M_R^c}(\xi^i)\left\|\xi^i\right\|^p - \mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\|\xi\|^p\right] \geq \frac{\epsilon}{16}\right) \leq \exp\left\{-NI\left(\frac{\epsilon}{16}\right)\right\},$$

where $I(\cdot)$ is the so-called (large deviations) rate function defined as

$$I\left(\frac{\epsilon}{16}\right) := \sup_{t\in\mathbb{R}}\left\{\frac{\epsilon}{16}t - \log M(t)\right\}$$

and

$$\begin{aligned} M(t) &:= \mathbb{E}\left[\exp\left\{t\left(\mathbf{1}_{M_R^c}(\xi)\|\xi\|^p - \mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\|\xi\|^p\right]\right)\right\}\right] \\ &= \frac{\mathbb{E}\left[\exp\left\{t\mathbf{1}_{M_R^c}(\xi)\|\xi\|^p\right\}\right]}{\exp\left\{t\mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\|\xi\|^p\right]\right\}} \\ &\leq \frac{\mathbb{E}\left[\exp\{t\|\xi\|^p\}\right]}{\exp\left\{t\mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\|\xi\|^p\right]\right\}} \\ &\leq \frac{\mathbb{E}\left[\exp\{\|\xi\|^p\}\right]}{\exp\{-\mathbb{E}\left[\|\xi\|^p\right]\}} < +\infty \end{aligned}$$

for $t \in [-1, 1]$, where the last inequality follows from Assumption 1 with $b > p$.

We know from [28] (Section 7.2.9) that $M(t)$ is positive, convex, and infinitely differentiable at the interior of its domain. This means that $\log M(t)$ is also convex and infinitely differentiable at the interior of its domain, which is consistent with the domain of $M(t)$. Since $M(t)$ is finite on $[-1, 1]$, $M(t)$ is differentiable on $(-1, 1)$. Note that

$$M'(0) = \mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\|\xi\|^p - \mathbb{E}\left[\mathbf{1}_{M_R^c}(\xi)\|\xi\|^p\right]\right] = 0.$$

Then, the derivative of

$$\frac{\epsilon}{16}t - \log M(t),$$

which is

$$\frac{\epsilon}{16} - \frac{M'(t)}{M(t)}, \tag{A7}$$

is larger than 0 at $t = 0$. Due to its differentiability, which implies the continuity, there exists a sufficiently small $0 < \bar{t} \leq 1$ such that (A7) is larger than 0 for any $t \in [0, \bar{t}]$. Then, for any $t \in (0, \bar{t}]$, we have

$$\frac{\epsilon}{16}t - \log M(t) > \frac{\epsilon}{16}\cdot 0 - \log M(0) = 0.$$

Therefore, we obtain that $I\left(\frac{\epsilon}{16}\right)$ is positive.

Finally, we obtain

$$\begin{aligned} \mathbb{P}(\zeta_p(P, P_N) \geq \epsilon) &\leq \alpha\exp\left(-\beta N\left(\frac{\epsilon}{2R^{p-1}}\right)^{\max\{4,s\}}\right) + \exp\left\{-NI\left(\frac{\epsilon}{16}\right)\right\} \\ &\leq (1+\alpha)\exp\left(-\min\left\{\beta\left(\frac{\epsilon}{2R^{p-1}}\right)^{\max\{4,s\}}, I\left(\frac{\epsilon}{16}\right)\right\}N\right). \end{aligned}$$

Letting $\hat{\alpha} := 1 + \alpha$ and

$$\hat{\beta} := \min\left\{\beta\left(\frac{\epsilon}{2R^{p-1}}\right)^{\max\{4,s\}}, I\left(\frac{\epsilon}{16}\right)\right\}$$

completes the proof. □

References

1. Fournier, N.; Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Relat. Fields* **2015**, *162*, 707–738. [CrossRef]
2. Villani, C. *Optimal Transport: Old and New*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008; Volume 338.
3. Rachev, S.T.; Rüschendorf, L. *Mass Transportation Problems: Volume I: Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1998; Volume 1.
4. Horowitz, J.; Karandikar, R.L. Mean rates of convergence of empirical measures in the Wasserstein metric. *J. Comput. Appl. Math.* **1994**, *55*, 261–273. [CrossRef]
5. Weed, J.; Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv* **2017**, arXiv:1707.00087.
6. Dereich, S.; Scheutzow, M.; Schottstedt, R. Constructive quantization: Approximation by empirical measures. *Ann. L'Ihp Probab. Stat.* **2013**, *49*, 1183–1203. [CrossRef]
7. Bolley, F.; Guillin, A.; Villani, C. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probab. Theory Relat. Fields* **2007**, *137*, 541–593. [CrossRef]
8. Boissard, E. Simple bounds for the convergence of empirical and occupation measures in 1-Wasserstein distance. *Electron. J. Probab.* **2011**, *16*, 2296–2333. [CrossRef]
9. Zhao, C.; Guan, Y. Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics. *Optim. Online* **2015**, *2*, 1–40.
10. Römisch, W. Stability of Stochastic Programming Problems. *Handb. Oper. Res. Manag. Sci.* **2003**, *10*, 483–554.
11. Rachev, S.T.; Römisch, W. Quantitative stability in stochastic programming: The method of probability metrics. *Math. Oper. Res.* **2002**, *27*, 792–818. [CrossRef]
12. Römisch, W.; Vigerske, S. Quantitative stability of fully random mixed-integer two-stage stochastic programs. *Optim. Lett.* **2008**, *2*, 377–388. [CrossRef]
13. Han, Y.; Chen, Z. Quantitative stability of full random two-stage stochastic programs with recourse. *Optim. Lett.* **2015**, *9*, 1075–1090. [CrossRef]
14. Strugarek, C. *On the Fortet-Mourier Metric for The Stability of Stochastic Optimization Problems, An Example*; Humboldt-Universität zu Berlin: Berlin, Germany, 2004.
15. Shapiro, A. Monte Carlo sampling methods. *Handb. Oper. Res. Manag. Sci.* **2003**, *10*, 353–425.
16. Shapiro, A.; Xu, H. Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation. *Optimization* **2008**, *57*, 395–418. [CrossRef]
17. Dentcheva, D.; Römisch, W. Stability and sensitivity of stochastic dominance constrained optimization models. *SIAM J. Optim.* **2013**, *23*, 1672–1688. [CrossRef]
18. Esfahani, P.M.; Kuhn, D. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Program.* **2018**, *171*, 115–166. [CrossRef]
19. Liu, Y.; Pichler, A.; Xu, H. Discrete approximation and quantification in distributionally robust optimization. *Math. Oper. Res.* **2018**, *44*, 19–37. [CrossRef]
20. Kantorovich, L.V.; Rubinstein, G.S. On a space of completely additive functions. *Vestn. Leningrad. Univ.* **1958**, *13*, 52–59.
21. Valentine, F.A. A Lipschitz condition preserving extension for a vector function. *Am. J. Math.* **1945**, *67*, 83–93. [CrossRef]
22. Dentcheva, D.; Henrion, R.; Ruszczyński, A. Stability and sensitivity of optimization problems with first order stochastic dominance constraints. *SIAM J. Optim.* **2007**, *18*, 322–337. [CrossRef]
23. Dentcheva, D.; Ruszczyński, A. Robust stochastic dominance and its application to risk-averse optimization. *Math. Program.* **2010**, *123*, 85–100. [CrossRef]
24. Chen, Z.; Jiang, J. Stability analysis of optimization problems with kth order stochastic and distributionally robust dominance constraints induced by full random recourse. *SIAM J. Optim.* **2018**, *28*, 1396–1419. [CrossRef]
25. Sun, H.; Xu, H. Convergence analysis of stationary points in sample average approximation of stochastic programs with second order stochastic dominance constraints. *Math. Program.* **2014**, *143*, 31–59. [CrossRef]
26. Liu, Y.; Xu, H. Stability analysis of stochastic programs with second order dominance constraints. *Math. Program.* **2013**, *142*, 435–460. [CrossRef]
27. Dentcheva, D.; Ruszczyński, A. Optimization with stochastic dominance constraints. *SIAM J. Optim.* **2003**, *14*, 548–566. [CrossRef]
28. Shapiro, A.; Dentcheva, D.; Ruszczyński, A. *Lectures on Stochastic Programming: Modeling and Theory*; SIAM: Philadelphia, PA, USA, 2014.
29. Bertsimas, D.; Gupta, V.; Kallus, N. Data-driven robust optimization. *Math. Program.* **2018**, *167*, 235–292. [CrossRef]
30. Bertsimas, D.; Gupta, V.; Kallus, N. Robust sample average approximation. *Math. Program.* **2018**, *171*, 217–282. [CrossRef]
31. Delage, E.; Ye, Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* **2010**, *58*, 595–612. [CrossRef]
32. Pichler, A.; Xu, H. Quantitative stability analysis for minimax distributionally robust risk optimization. *Math. Program.* **2022**, *191*, 47–77. [CrossRef]

33. Zhang, J.; Xu, H.; Zhang, L. Quantitative stability analysis for distributionally robust optimization with moment constraints. *SIAM J. Optim.* **2016**, *26*, 1855–1882. [CrossRef]
34. Jiang, J.; Chen, Z. Quantitative stability analysis of two-stage stochastic linear programs with full random recourse. *Numer. Funct. Anal. Optim.* **2019**, *40*, 1847–1876. [CrossRef]
35. Varadarajan, V.S. On the convergence of sample probability distributions. *Sankhyā Indian J. Stat.* **1958**, *19*, 23–26.