Journal of the Operations Research Society of China (2021) 9:525–542
https://doi.org/10.1007/s40305-020-00313-w

Data-driven Stochastic Programming with Distributionally Robust Constraints Under Wasserstein Distance: Asymptotic Properties

Yu Mei[1,2] · **Zhi-Ping Chen**[1,2] · **Bing-Bing Ji**[1,2] · **Zhu-Jia Xu**[1,2] · **Jia Liu**[1,2]

Received: 17 November 2019 / Revised: 12 June 2020 / Accepted: 23 June 2020 /
Published online: 5 August 2020

Abstract

Distributionally robust optimization is a dominant paradigm for decision-making problems where the distribution of random variables is unknown. We investigate a distributionally robust optimization problem with ambiguities in the objective function and countably infinite constraints. The ambiguity set is defined as a Wasserstein ball centered at the empirical distribution. Based on the concentration inequality of Wasserstein distance, we establish the asymptotic convergence property of the data-driven distributionally robust optimization problem when the sample size goes to infinity. We show that with probability 1, the optimal value and the optimal solution set of the data-driven distributionally robust problem converge to those of the stochastic optimization problem with true distribution. Finally, we provide numerical evidences for the established theoretical results.

This work was supported by the National Natural Science Foundation of China (Nos. 11991023, 11901449, 11735011).

✉ Zhi-Ping Chen
zchen@mail.xjtu.edu.cn

Yu Mei
meiyu414@stu.xjtu.edu.cn

Bing-Bing Ji
bingji0225@stu.xjtu.edu.cn

Zhu-Jia Xu
xzj19930326@163.com

Jia Liu
jialiu@xjtu.edu.cn

[1] Department of Computing Science, School of Mathematics and Statistics, Xi’an Jiaotong University, Xi’an 710049, China

[2] Center for Optimization Technique and Quantitative Finance, Xi’an International Academy for Mathematics and Mathematical Technology, Xi’an 710049, China

Keywords Distributionally robust optimization · Wasserstein distance · Ambiguity set · Asymptotic analysis · Empirical distribution

Mathematics Subject Classification 90C59 · 90C34

1 Introduction

Stochastic programming is a useful decision-making paradigm for dealing with optimization problems under parameter uncertainty [1]. For the modeling or analyzing of stochastic programming problems, it may be essential to take into account multiple criteria [2,3]. To deal with this issue, the decision model may involve infinitely many constraints. For example, the stochastic optimization model with stochastic dominance constraints proposed in [4,5] is a semi-infinite constrained stochastic programming problem. One way to handle the semi-infinite constraints is to view them as a robust risk measure constraint and use tractable risk measures to approximate it [6]. [7] utilized the sample average approximation method to solve a stochastic programming problem with second-order stochastic dominance constraints.

Classical stochastic programming is sometimes questioned in reality, because the true distribution of random parameters cannot be precisely known. An alternative modeling scheme is the *distributionally robust optimization*, where one considers the worst-case expectation instead of the expectation under the true distribution. The worst-case expectation is taken over an ambiguity set, which is a collection of all possible probability distributions characterized by specific known properties. Here, the true distribution is supposed to be in the ambiguity set (at least with a high confidence level). Distributionally robust optimization was first introduced in the seminal works [8,9] and has developed rapidly in the last decade [10,11].

Different ambiguity sets have been proposed in the literature. The following two types of ambiguity sets have been widely adopted. Moment-based ambiguity sets contain distributions characterized by certain moment information. [12] considered an ambiguity set with known variance or covariance matrix and known bounds on the mean value. [11] studied an ambiguity set based on support and moment information obtained from samples. [13] developed a unified framework where the ambiguity set is based on known mean and nested cones. [14] studied a robust two-stage stochastic linear programming model with mean-CVaR recourse under the moment ambiguity set. [15] investigated the approximation scheme for distributionally robust stochastic dominance constrained problems under a moment-based ambiguity set, which has infinitely many constraints. An alternative method for specifying the ambiguity set is to contain all the distributions close to a nominal distribution under a prescribed probability metric. [16] considered an ambiguous chance-constrained problem with an ambiguity set determined by the Prohorov metric. Wasserstein distance was adopted in [17,18] to construct ambiguity sets. The phi-divergence family, such as the Kullback–Leibler divergence, total variation, and χ^2-divergence, has also been used to define ambiguity sets [10,19,20].

The Wasserstein distance-based ambiguity set has attracted much attention among all these ambiguity sets [18,21,22]. It has the following three advantages: firstly, Wasserstein distance intuitively describes the minimum cost to move from one mass distribution to another [23]; secondly, there are some probabilistic guarantees on the a priori estimation such that the true distribution belongs to the Wasserstein ambiguity set [24]; thirdly, [22] established the out-of-sample performance guarantee for stochastic optimization problems under the Wasserstein ambiguity set. Therefore, in this paper, we consider the Wasserstein distance-based ambiguity set centered at the empirical distribution, which is constituted by N historical i.i.d. samples. Different from most of the current models, we consider distributionally robust counterparts in both the objective and the countably infinite constraints. We discuss the asymptotic convergence property of the data-driven distributionally robust semi-infinite optimization problem when the sample size goes to infinity.

The main differences between our work and that in [22] lie in three aspects. Firstly, we consider the Wasserstein distance with any order $p \geqslant 1$, while [22] only investigated the case $p = 1$. Secondly, the distributionally robust optimization problem we consider involves infinite constraints, which includes a broad class of problems such as the stochastic dominance constrained problem. Infinite constraints naturally increase the difficulties in analyzing asymptotic convergence properties. Finally, the convergence of the optimal solution set is also examined in this paper, which is not considered in [22].

We use the following notations. The m-dimensional random vector ξ is governed by a probability distribution P. Let $\Xi \subset \mathbb{R}^m$ be the support of ξ. The N-fold Cartesian product of distribution P on Ξ is denoted by P^N, which is supported on the Cartesian product space Ξ^N. $\mathcal{P}_p(\Xi)$ denotes the collection of probability distributions Q supported on Ξ with $\int_\Xi \|\xi\|^p Q(d\xi) < \infty$. For a fixed distribution $Q \in \mathcal{P}_p(\Xi)$, $\mathcal{L}^1(Q)$ denotes the space of all Q-integrable functions. The distance between two sets A and B is defined as $D(A, B) := \sup_{x\in A} \mathrm{dist}(x, B) = \sup_{x\in A} \inf_{y\in B} \|x - y\|$.

2 Data-driven Distributionally Robust Optimization Under Wasserstein Distance

We consider the following infinitely constrained stochastic optimization model

$$\text{(SP)} \qquad \begin{aligned} \min_{z\in Z_0} \quad & f(z) := \mathrm{E}_{\mathrm{P}}[f(z,\xi)] \\ \text{s.t.} \quad & \mathrm{E}_{\mathrm{P}}[h(\eta,z,\xi)] \leqslant 0, \quad \forall \eta \in \Gamma, \end{aligned}$$

where $f(z,\xi) : \mathbb{R}^n \times \Xi \to \mathbb{R}$ is continuous with respect to z for every ξ, $h(\eta, z, \xi) : \Gamma \times \mathbb{R}^n \times \xi \to \bar{\mathbb{R}}$ is continuous with respect to z for every (η, ξ) and is continuous with respect to η for every (z, ξ), Γ is a set with infinitely many elements, and $Z_0 \subset \mathbb{R}^n$ is a compact set. Denote the optimal value and the optimal solution set of problem (SP) by J^* and S^*, respectively. We assume that P is unknown, but can be estimated from i.i.d. samples $\{\widetilde{\xi}_i\}_{i=1}^N$. The sample set $\widetilde{\Xi}_N := \{\widetilde{\xi}_i\}_{i=1}^N (\subset \Xi)$ can be considered as a random collection of samples governed by the distribution P^N. We always use superscript '$\sim$' to emphasize that a variable is treated as random. We first recall the definition of Wasserstein distance.

Definition 1 Let $p \geqslant 1$. The *Wasserstein distance* $W_p(Q_1, Q_2)$ between $Q_1, Q_2 \in \mathcal{P}_p(\Xi)$ is defined via

$$W_p(Q_1, Q_2) := \left(\inf \left\{ \int_{\Xi^2} \|\xi_1 - \xi_2\|^p \Pi(d\xi_1, d\xi_2) : \begin{array}{l} \Pi \text{ is a joint distribution of } \xi_1 \text{ and } \xi_2 \\ \text{with marginals } Q_1 \text{ and } Q_2, \text{ respectively} \end{array} \right\} \right)^{\frac{1}{p}}. \tag{1}$$

Wasserstein distance corresponds to the minimum cost of moving from one mass distribution Q_1 to another Q_2. When Q_1 and Q_2 are both discrete distributions, the optimization problem in Wasserstein distance can be viewed as Monge's mass transportation problem by treating Π as the transportation plan [23]. Wasserstein distance has the following dual representation [17, eq. (7)]

$$W_p^p(Q_1, Q_2) = \sup_{u \in \mathcal{L}^1(Q_1), v \in \mathcal{L}^1(Q_2)} \left\{ \begin{array}{l} \int_\Xi u(\xi_1) Q_1(d\xi_1) + \int_\Xi v(\xi_2) Q_2(d\xi_2) : \\ u(\xi_1) + v(\xi_2) \leqslant \|\xi_1 - \xi_2\|^p, \forall \xi_1, \xi_2 \in \Xi \end{array} \right\}. \tag{2}$$

Due to the representation (2), we immediately have the following observation.

Lemma 1 [17] *Let* $\Psi : \Xi \to \mathbb{R}$. *Suppose that* Ψ *satisfies* $|\Psi(\xi_1) - \Psi(\xi_2)| \leqslant L_0 \|\xi_1 - \xi_2\|^p + M_0$ *for all* $\xi_1, \xi_2 \in \Xi$ *and some* $L_0, M_0 \geqslant 0$. *Then,*

$$|E_{Q_1}[\Psi(\xi)] - E_{Q_2}[\Psi(\xi)]| \leqslant L_0 W_p^p(Q_1, Q_2) + M_0.$$

We now define

$$\widetilde{\mathcal{Q}}_N = \{Q \in \mathcal{P}_p(\Xi) : W_p(Q, \widetilde{\mathrm{P}}_N) \leqslant \epsilon_N\}, \tag{3}$$

where $\widetilde{\mathrm{P}}_N := \frac{1}{N} \sum_{i=1}^N \delta_{\widetilde{\xi}_i}$ is the empirical distribution, and ϵ_N is a given radius. We consider the following data-driven distributionally robust counterpart of problem (SP)

$$\begin{array}{lll} \text{(RP)} & \min\limits_{z \in Z_0} & \widetilde{f}_N(z) := \sup\limits_{Q \in \widetilde{\mathcal{Q}}_N} \mathrm{E}_Q[f(z, \xi)] \\ & \text{s.t.} & \sup\limits_{Q \in \widetilde{\mathcal{Q}}_N} \mathrm{E}_Q[h(\eta, z, \xi)] \leqslant 0, \quad \forall \eta \in \Gamma. \end{array}$$

Denote the optimal value and the optimal solution set of problem (RP) by $\widetilde{J}_N^*$ and $\widetilde{S}_N^*$, respectively.

It is worth noting that problem (RP) differs from another kind of distributionally robust model

$$\min_{z \in Z_0} \sup_{Q \in \widetilde{\mathcal{Q}}_N} \{\mathrm{E}_Q[f(z, \xi)] : \mathrm{E}_Q[h(\eta, z, \xi)] \leqslant 0, \ \forall \eta \in \Gamma\}. \tag{4}$$

Problem (RP) aims to define the distributionally robust counterparts in the objective function and in constraints, separately. While, problem (4) tries to define the distributionally robust counterpart in terms of the optimal value function. In problem (4), all the expectations in the objective function and the constraints are taken under the same worst-case distribution. While, in problem (RP), the worst-case distribution for the objective function $Q_f^* \in \operatorname{argmax}_{Q \in \widetilde{\mathcal{Q}}_N} \mathrm{E}_Q[f(z,\xi)]$ is probably different from the worst-case probability distributions for the constraint functions $Q_\eta^* \in \operatorname{argmax}_{Q \in \widetilde{\mathcal{Q}}_N} \mathrm{E}_Q[h(\eta, z, \xi)],\ \eta \in \Gamma$. Problem (RP) derives a more robust solution, which keeps the constraints feasible for all possible distributions in the ambiguity set. While, the optimal solution of problem (4) is only feasible for those constraints under the worst-case distribution. In this paper, we focus on the model (RP), and we do not require that $Q_f^* = Q_\eta^*,\ \eta \in \Gamma$.

3 Asymptotic Convergence Property

In this section, we will show that with probability 1, the optimal value and the optimal solution set of problem (RP) tend to those of problem (SP) when $N \to \infty$. For this purpose, we assume that the tail of the distribution P decays at a fast speed. Concretely, we introduce the following assumption.

Assumption 1 There exist $\alpha > p, \gamma > 0$ such that

$$A := \int_\Xi \exp\{\gamma \|\xi\|^\alpha\} \mathrm{P}(d\xi) < \infty.$$

This assumption is mild and has been widely adopted in related researches, such as [18,22,25]. If Ξ is compact, Assumption 1 holds trivially.

Based on [24, Theorem 2], Esfahani and Kuhn stated a measure concentration property in [22, Theorem 3.4] for $p = 1$. We generalize this result to any integer order $p \geqslant 1$.

Lemma 2 (*Concentration Inequality*) *Given Assumption 1, there exist positive constants c_1 and c_2 depending only on α, γ, A and m such that*

$$P^N\{W_p(P, \widetilde{P}_N) \geqslant \epsilon_N\} \leqslant \begin{cases} c_1 \exp\left(-c_2 N \epsilon_N^{\max\{m,2p\}}\right), & \text{if } \epsilon_N \leqslant 1, \\ c_1 \exp\left(-c_2 N \epsilon_N^{\alpha}\right), & \text{if } \epsilon_N > 1, \end{cases} \tag{5}$$

for all $N \geqslant 1,\ m \neq 2p$.

This lemma can be easily proved by using [24, Theorem 2]. When $m = 2p$, a similar inequality also holds. The detailed proof is thus omitted here.

Lemma 2 provides a probabilistic estimation that the true distribution P lies outside the Wasserstein ball $\mathcal{B}(\widetilde{\mathrm{P}}_N, \epsilon_N)$. This probability can be set as some prescribed disaster level β_N. Solving ϵ_N in the equation

$$\beta_N = \begin{cases} c_1 \exp\left(-c_2 N \epsilon_N^{\max\{m,2p\}}\right), & \text{if } \epsilon_N \leqslant 1, \\ c_1 \exp\left(-c_2 N \epsilon_N^{\alpha}\right), & \text{if } \epsilon_N > 1, \end{cases} \tag{6}$$

we obtain the smallest radius of the Wasserstein ball containing P with confidence $1 - \beta_N$. Namely

$$\epsilon_N(\beta_N) := \begin{cases} \left(\frac{\log(\frac{c_1}{\beta_N})}{c_2 N}\right)^{\frac{1}{\max\{m,2p\}}}, & \text{if } N \geqslant \frac{\log(\frac{c_1}{\beta_N})}{c_2}, \\ \left(\frac{\log(\frac{c_1}{\beta_N})}{c_2 N}\right)^{\frac{1}{\alpha}}, & \text{if } N < \frac{\log(\frac{c_1}{\beta_N})}{c_2}, \end{cases} \tag{7}$$

such that

$$\mathrm{P}^N\{W_p(\mathrm{P}, \widetilde{\mathrm{P}}_N) \leqslant \epsilon_N(\beta_N)\} \geqslant 1 - \beta_N. \tag{8}$$

Note that for a fixed level $\beta_N \equiv \beta > 0$, the radius $\epsilon_N(\beta_N)$ goes to zero as N tends to infinity. We now want to show that, when N tends to infinity, $\widetilde{f}_N$ converges to f with probability 1. To this end, we introduce the following assumptions.

Assumption 2 There exists an $L \geqslant 0$ such that $|f(z, \xi_1) - f(z, \xi_2)| \leqslant L\|\xi_1 - \xi_2\|^p$ for all $z \in Z_0$.

Assumption 3 $\beta_N \in (0, 1)$ satisfies $\sum_{N=1}^{\infty} \beta_N < \infty$ and $\lim_{N\to\infty} \frac{\log \beta_N}{N} = 0$.

The following theorem establishes the pointwise convergence result.

Theorem 1 (*Convergence*) *Given Assumptions 1, 2 and 3, P^∞-almost surely we have that $\widetilde{f}_N$ converges pointwise to f.*

Proof We have from (8) that

$$\mathrm{P}^N\{\mathrm{P} \in \widetilde{\mathcal{Q}}_N\} \geqslant 1 - \beta_N,$$

which further yields

$$\mathrm{P}^N\{\widetilde{f}_N(z) \geqslant f(z),\ \forall z \in Z_0\} \geqslant 1 - \beta_N.$$

Applying Borel–Cantelli Lemma (see, e.g., [26, Theorem 2.18]), we obtain

$$\mathrm{P}^\infty\{\widetilde{f}_N(z) \geqslant f(z),\ \forall z \in Z_0,\ \text{for all sufficiently large } N\} = 1.$$

Hence, it holds that

$$\mathrm{P}^\infty\left\{\liminf_{N\to\infty} \widetilde{f}_N(z) \geqslant f(z),\ \forall z \in Z_0\right\} = 1. \tag{9}$$

On the other hand, from the definition of supremum and the fact that $\widetilde{\mathcal{Q}}_N$ is connected, for any $\epsilon > 0$, there exists a $\widetilde{Q}_N \in \widetilde{\mathcal{Q}}_N$ such that

$$\mathrm{E}_{\widetilde{Q}_N}[f(z,\xi)] \geqslant \widetilde{f}_N(z) - \epsilon/2.$$

Then, we have

$$\begin{aligned}\widetilde{f}_N(z) &\leqslant \mathrm{E}_{\widetilde{Q}_N}[f(z,\xi)] + \epsilon/2\\ &= \int_{\varXi} f(z,\xi)\mathrm{P}(d\xi) + \int_{\varXi} f(z,\xi)\widetilde{Q}_N(d\xi) - \int_{\varXi} f(z,\xi)\mathrm{P}(d\xi) + \epsilon/2\\ &\leqslant \mathrm{E}_{\mathrm{P}}[f(z,\xi)] + LW_p^p(\mathrm{P},\widetilde{Q}_N) + \epsilon/2\\ &\leqslant \mathrm{E}_{\mathrm{P}}[f(z,\xi)] + L[W_p(\mathrm{P},\widetilde{\mathrm{P}}_N) + W_p(\widetilde{\mathrm{P}}_N,\widetilde{Q}_N)]^p + \epsilon/2.\end{aligned}$$

The last but one inequality follows directly from Lemma 1 and Assumption 2. Thus by (8), we obtain

$$\mathrm{P}^N\left\{\widetilde{f}_N(z) \leqslant \mathrm{E}_{\mathrm{P}}[f(z,\xi)]+L(2\epsilon_N(\beta_N))^p+\epsilon/2,\ \forall z\in Z_0\right\} \geqslant \mathrm{P}^N\left\{\mathrm{P}\in\widetilde{\mathcal{Q}}_N\right\} \geqslant 1-\beta_N.$$

Since $\lim_{N\to\infty}\epsilon_N(\beta_N) = 0$, there exists an N_1 such that for all $N \geqslant N_1$, we have $L(2\epsilon_N(\beta_N))^p < \epsilon/2$. This further implies that

$$\mathrm{P}^N\left\{\widetilde{f}_N(z) \leqslant \mathrm{E}_{\mathrm{P}}[f(z,\xi)] + \epsilon,\ \forall z\in Z_0\right\} \geqslant 1-\beta_N.$$

Again by Borel-Cantelli Lemma, we have

$$\mathrm{P}^\infty\left\{\widetilde{f}_N(z) \leqslant f(z)+\epsilon,\ \forall z\in Z_0,\ \text{for all sufficiently large } N\right\} = 1.$$

Therefore, it holds that

$$\mathrm{P}^\infty\left\{\limsup_{N\to\infty}\widetilde{f}_N(z) \leqslant f(z)+\epsilon,\ \forall z\in Z_0\right\} = 1.$$

Since ϵ can be chosen arbitrarily, we obtain

$$\mathrm{P}^\infty\left\{\limsup_{N\to\infty}\widetilde{f}_N(z) \leqslant f(z),\ \forall z\in Z_0\right\} = 1. \tag{10}$$

The proof follows immediately from (9) and (10).

We notice that if moreover, $f(z,\xi)$ is Lipschitz continuous with respect to z, then $\widetilde{f}_N$ converges uniformly to f.

Assumption 4 There exists a $\kappa(\xi)$ such that $|f(z_1,\xi) - f(z_2,\xi)| \leqslant \kappa(\xi)\|z_1 - z_2\|$ for all $z_1, z_2 \in Z_0$ and $K := \sup_{Q\in\mathcal{P}_p(\varXi)}\mathrm{E}_Q[\kappa(\xi)] < \infty$.

Theorem 2 (*Uniform convergence*) *Suppose that Assumptions* 1, 2, 3 *and* 4 *hold, then* P^∞*-almost surely* $\widetilde{f}_N$ *converges uniformly to* f *on* Z_0.

Proof According to the Arzelà-Ascoli theorem, a sequence in a compact Hausdorff space converges uniformly if and only if it is equicontinuous and converges pointwise. Therefore, what remains to prove is that $\{\widetilde{f}_N(z)\}$ is equicontinuous. By

$$\begin{aligned}
&|\widetilde{f}_N(z_1) - \widetilde{f}_N(z_2)| = |\sup_{Q\in\tilde{\mathcal{Q}}_N} \mathrm{E}_Q[f(z_1,\xi)] - \sup_{Q\in\tilde{\mathcal{Q}}_N} \mathrm{E}_Q[f(z_2,\xi)]| \\
&\leqslant \sup_{Q\in\tilde{\mathcal{Q}}_N} |\mathrm{E}_Q[f(z_1,\xi)] - \mathrm{E}_Q[f(z_2,\xi)]| \leqslant \sup_{Q\in\tilde{\mathcal{Q}}_N} \mathrm{E}_Q|f(z_1,\xi) - f(z_2,\xi)| \\
&\leqslant \sup_{Q\in\tilde{\mathcal{Q}}_N} \mathrm{E}_Q[\kappa(\xi)]\|z_1 - z_2\| \leqslant K\|z_1 - z_2\|,
\end{aligned}$$

the equicontinuity follows directly.

Equipped with the convergence of the objective function, we can establish the convergence of the optimal value and the optimal solution set. To make a clear statement, we need to consider the following intermediate problem:

$$\begin{aligned}
\text{(RCP)} \quad &\min_{z\in Z_0} \quad f(z) := \mathrm{E}_{\mathrm{P}}[f(z,\xi)] \\
&\text{s.t.} \quad \sup_{Q\in\tilde{\mathcal{Q}}_N} \mathrm{E}_Q[h(\eta,z,\xi)] \leqslant 0, \quad \forall \eta \in \varGamma.
\end{aligned}$$

Denote the optimal value and the optimal solution set of problem (RCP) by $\widetilde{J}_N$ and $\widetilde{S}_N$, respectively.

Firstly, we establish the finite sample guarantee and the asymptotic convergence property between the intermediate problem (RCP) and the true problem (SP).

Theorem 3 (*Finite sample guarantee*) *Given Assumption* 1,

$$P^N\left\{\widetilde{\varXi}_N : J^* \leqslant \widetilde{J}_N\right\} \geqslant 1 - \beta_N.$$

Proof Let $\widetilde{z}_N \in \widetilde{S}_N$. (8) implies that

$$\mathrm{P}^N\left\{\mathrm{E}_{\mathrm{P}}\left[h(\eta,\widetilde{z}_N,\xi)\right] \leqslant \sup_{Q\in\widetilde{\mathcal{Q}}_N} \mathrm{E}_{\mathrm{P}}\left[h(\eta,\widetilde{z}_N,\xi)\right],\ \forall \eta \in \varGamma\right\} \geqslant 1 - \beta_N.$$

Note that problems (RCP) and (SP) have the same objective function. Hence, if $\widetilde{z}_N$ satisfies all the constraints in problem (SP), then it would hold that $J^* \leqslant f(\widetilde{z}_N)$. Therefore, we have

$$\begin{aligned}
&\mathrm{P}^N\{J^* \leqslant \widetilde{J}_N\} = \mathrm{P}^N\left\{J^* \leqslant f(\widetilde{z}_N)\right\} \geqslant \mathrm{P}^N\left\{\mathrm{E}_{\mathrm{P}}[h(\eta,\widetilde{z}_N,\xi)] \leqslant 0,\ \forall \eta \in \varGamma\right\} \\
&\quad \geqslant \mathrm{P}^N\left\{\mathrm{E}_{\mathrm{P}}\left[h(\eta,\widetilde{z}_N,\xi)\right] \leqslant \sup_{Q\in\widetilde{\mathcal{Q}}_N} \mathrm{E}_{\mathrm{P}}\left[h(\eta,\widetilde{z}_N,\xi)\right],\ \forall \eta \in \varGamma\right\} \geqslant 1 - \beta_N.
\end{aligned}$$

To establish the asymptotic convergence property, we need the following technical assumptions.

Assumption 5 There exists an $\mathcal{L}(\eta)$ such that $|h(\eta, z, \xi_1) - h(\eta, z, \xi_2)| \leqslant \mathcal{L}(\eta)\|\xi_1 - \xi_2\|^p$ for all $z \in Z_0$.

Assumption 6 Γ is a countable and compact set.

For simplicity of exposition, let $\vartheta(\eta, z) := \mathrm{E}_{\mathrm{P}}[h(\eta, z, \xi)]$, $\widetilde{\vartheta}_N(\eta, z) := \sup_{Q\in\widetilde{\mathcal{Q}}_N} \mathrm{E}_Q[h(\eta, z, \xi)]$, $v(z) := \sup_{\eta\in\Gamma} \vartheta(\eta, z)$ and $\widetilde{v}_N(z) := \sup_{\eta\in\Gamma} \widetilde{\vartheta}_N(\eta, z)$. Since h is continuous with respect to z and Γ is compact, ϑ, $\widetilde{\vartheta}_N$, v, and $\widetilde{v}_N$ are all continuous with respect to z. Problems (SP) and (RCP) can be rewritten in the following compact forms, respectively

$$\min_{z\in Z_0} f(z) \quad \text{s.t. } v(z) \leqslant 0,$$

and

$$\min_{z\in Z_0} f(z) \quad \text{s.t. } \widetilde{v}_N(z) \leqslant 0.$$

Next, we will prove the pointwise convergence result of $\widetilde{v}_N(z)$.

Theorem 4 (*Convergence of constraints*) *Given Assumptions 1, 3, 5, for every $\eta \in \Gamma$, P^∞-almost surely we have that $\widetilde{\vartheta}_N(\eta, z)$ converges to $\vartheta(\eta, z)$ pointwise. Moreover, if Assumption 6 also holds, then $\widetilde{v}_N(z)$ converges pointwise to $v(z)$.*

Proof Similar to the proof of Theorem 1, the first conclusion can be established immediately.

We know that the intersection set of countable sets with probability 1 also has probability 1. If Γ is countable, then P^∞-almost surely it holds that

$$\lim_{N\to\infty} \widetilde{\vartheta}_N(\eta, z) - \vartheta(\eta, z) = 0, \forall \eta \in \Gamma,\ \forall z \in Z_0. \tag{11}$$

Since $h(\eta, z, \xi)$ is continuous with respect to η and Γ is compact, we can easily show that for any $z \in Z_0$, there exists an η^* such that

$$|\widetilde{v}_N(z) - v(z)| \leqslant \sup_{\eta\in\Gamma} \left|\widetilde{\vartheta}_N(\eta, z) - \vartheta(\eta, z)\right| = \left|\widetilde{\vartheta}_N(\eta^*, z) - \vartheta(\eta^*, z)\right|. \tag{12}$$

(11) and (12) together ensure that P^∞-almost surely, we have

$$\lim_{N\to\infty} \widetilde{v}_N(z) = v(z)$$

for all $z \in Z_0$. This completes the proof.

To establish the asymptotic convergence property of problem (RCP), we need the following assumption like that in [27].

Assumption 7 Assume that there exists an optimal solution $\bar{z}$ of the true problem (SP) such that for any $\epsilon > 0$, there is a $z \in Z_0$ with $\|z - \bar{z}\| \leqslant \epsilon$ and $v(z) < 0$.

Theorem 5 (*Asymptotic convergence property*) *Given Assumptions 1, 3, 5, 6 and 7, P^∞-almost surely $\widetilde{J}_N \to J^*$ and $D(\widetilde{S}_N, S^*) \to 0$ as $N \to \infty$.*

Proof By Theorem 3 and Borel-Cantelli Lemma, we have

$$\mathrm{P}^\infty \left\{ J^* \leqslant \liminf_{N\to\infty} \widetilde{J}_N \right\} = 1.$$

On the other hand, for any $\epsilon > 0$, there exists a $z_\epsilon \in Z_0$ with $\|z_\epsilon - \bar{z}\| \leqslant \epsilon$ and $v(z_\epsilon) < 0$. Assume that $z_\epsilon \to \bar{z}$ when $\epsilon \to 0$, by passing to a subsequence if necessary. It is known from Theorem 4 that such z_ϵ satisfies

$$\mathrm{P}^\infty \{\widetilde{v}_N(z_\epsilon) < 0 \text{ for all sufficiently large } N\} = 1,$$

and consequently,

$$\mathrm{P}^\infty \left\{ f(z_\epsilon) \geqslant \widetilde{J}_N \text{ for all sufficiently large } N \right\} = 1.$$

We immediately get

$$\mathrm{P}^\infty \left\{ f(z_\epsilon) \geqslant \limsup_{N\to\infty} \widetilde{J}_N \right\} = 1.$$

Since f is continuous, we have that

$$\mathrm{P}^\infty \left\{ J^* = f(\bar{z}) = f(\lim_{\epsilon\to 0} z_\epsilon) = \lim_{\epsilon\to 0} f(z_\epsilon) \geqslant \limsup_{N\to\infty} \widetilde{J}_N \right\} = 1.$$

For the second claim, the following discussions are all understood in the P^∞-almost surely sense. Assume that $D(\widetilde{S}_N, S^*) \nrightarrow 0$. Then, there must exist an $\epsilon_0 > 0$ and $\widetilde{z}_N \in \widetilde{S}_N$ such that $\mathrm{dist}(\widetilde{z}_N, S^*) \geqslant \epsilon_0$ for all sufficiently large N. Since Z_0 is compact, we assume by passing to a subsequence if necessary that $\widetilde{z}_N \to z^*$. Therefore, $z^* \notin S^*$. Noticing $\widetilde{z}_N \in \widetilde{S}_N$, from the facts that $\widetilde{v}_N$ converges to v pointwise and $\widetilde{v}_N$ is continuous, we know that $v(z^*) = \lim_{N\to\infty} \widetilde{v}_N(\widetilde{z}_N) \leqslant 0$ and thus z^* is a feasible solution of problem (SP). Hence $f(z^*) > J^*$. By the continuity of f, we have

$$\lim_{N\to\infty} \widetilde{J}_N = \lim_{N\to\infty} f(\widetilde{z}_N) = f(z^*) > J^*,$$

which contradicts $\widetilde{J}_N \to J^*$.

Next, let us investigate problem (RP). We will discuss how $\widetilde{J}_N^*$ approximates J^* and how $\widetilde{S}_N^*$ approximates S^* when $N \to \infty$.

Theorem 6 *Given Assumptions 1-7, P^∞-almost surely $\widetilde{J}_N^* \to J^*$ and $D(\widetilde{S}_N^*, S^*) \to 0$ as $N \to \infty$.*

Proof The following discussions are all understood in the P^∞-almost surely sense.

Let $\widetilde{z}_N \in \widetilde{S}_N$ and $\widetilde{z}_N^* \in \widetilde{S}_N^*$. From Theorem 2, for any $\epsilon > 0$, there exists an $N_1 = N_1(\epsilon)$ such that for all $N \geqslant N_1$, it holds that

$$\left|\widetilde{f}_N(z) - f(z)\right| \leqslant \epsilon/2,\ \forall z \in Z_0.$$

Thus we have

$$\widetilde{f}_N(\widetilde{z}_N) - f(\widetilde{z}_N) \leqslant \epsilon/2 \tag{13}$$

and

$$f(\widetilde{z}_N^*) - \widetilde{f}_N(\widetilde{z}_N^*) \leqslant \epsilon/2. \tag{14}$$

Notice that the constraints in problems (RP) and (RCP) are the same and it is obvious that

$$\widetilde{f}_N(\widetilde{z}_N) \geqslant \widetilde{J}_N^* := \widetilde{f}_N(\widetilde{z}_N^*) \tag{15}$$

and

$$f(\widetilde{z}_N^*) \geqslant \widetilde{J}_N := f(\widetilde{z}_N). \tag{16}$$

Therefore, (13) and (15) mean that

$$\widetilde{J}_N^* - \widetilde{J}_N \leqslant \epsilon/2, \tag{17}$$

while (14) and (16) lead to

$$\widetilde{J}_N - \widetilde{J}_N^* \leqslant \epsilon/2. \tag{18}$$

By Theorem 5, for the above ϵ, there must exist an $N_2 = N_2(\epsilon)$ such that for all $N \geqslant N_2$, it holds that $|\widetilde{J}_N - J^*| \leqslant \epsilon/2$. Let $N_0 = \max\{N_1, N_2\}$. We obtain

$$|\widetilde{J}_N^* - J^*| \leqslant |\widetilde{J}_N^* - \widetilde{J}_N| + |\widetilde{J}_N - J^*| \leqslant \epsilon, \quad \forall N \geqslant N_0.$$

Hence it holds that $\widetilde{J}_N^* \to J^*$.

Assume that $D(\widetilde{S}_N^*, S^*) \nrightarrow 0$. Then, there must exist an $\epsilon_0 > 0$ and $\widetilde{z}_N^* \in \widetilde{S}_N^*$ such that $\mathrm{dist}(\widetilde{z}_N^*, S^*) \geqslant \epsilon_0$ for all sufficiently large N. Since Z_0 is compact, we assume by passing to a subsequence if necessary that $\widetilde{z}_N^* \to z^*$. Therefore, $z^* \notin S^*$. From the facts that $\widetilde{v}_N$ converges to v pointwise and $\widetilde{v}_N$ is continuous, we know that

$v(z^*) = \lim_{N\to\infty} \widetilde{v}_N(\widetilde{z}_N^*) \leqslant 0$ and thus z^* is a feasible solution of problem (SP). Hence $f(z^*) > J^*$. By the uniform convergence of $\widetilde{f}_N$, we have

$$\lim_{N\to\infty} \widetilde{J}_N^* = \lim_{N\to\infty} \widetilde{f}_N(\widetilde{z}_N^*) = f(z^*) > J^*,$$

which contradicts $\widetilde{J}_N^* \to J^*$.

Theorem 6 guarantees that problem (RP) is a "good" approximation to problem (SP) in the sense of the optimal value and the optimal solution set. Thus, it is reasonable to consider problem (RP) instead of problem (SP) in practical applications.

4 Numerical Experiments

To examine the asymptotic convergence results in Theorem 6, we consider a data-driven distributionally robust portfolio selection problem with second-order stochastic dominance constraints.

4.1 Portfolio Optimization Models

We recall the portfolio optimization model with second-order stochastic dominance constraints proposed in [4]

$$\begin{aligned}
\min_{z\in\mathbb{R}^n} \quad & \mathrm{E}_{\mathrm{P}}[-z^{\mathrm{T}}\xi] \\
\text{s.t.} \quad & \mathrm{E}_{\mathrm{P}}[(\eta_k - z^{\mathrm{T}}\xi)_+] \leqslant \mathrm{E}_{\mathrm{P}}[(\eta_k - Y(\xi))_+],\ k = 1, \cdots, J, \\
& \sum_{j=1}^{n} z_j = 1, \\
& z_j \geqslant 0,\ j = 1, \cdots, n.
\end{aligned} \tag{19}$$

Here, we assume that there are n risky assets, z denotes the portfolio vector, and ξ denotes the random return rate vector of the risky assets. We assume that the support set Ξ of the random return rate vectors is finite [17, Corollary 4]. Y represents the benchmark which is a prespecified random variable with finite realizations $\eta_k = Y(\xi_k),\ k = 1, \cdots, J$. $(\cdot)_+$ denotes the positive part function, i.e., $(\cdot)_+ = \max(0, \cdot)$.

We further consider the data-driven distributionally robust counterpart of model (19)

$$
\begin{aligned}
\min_{z\in\mathbb{R}^n}\ & \sup_{Q\in\widetilde{\mathcal{Q}}_N}\ \mathrm{E}_Q[-z^{\mathrm{T}}\xi] \\
\text{s.t.}\ & \sup_{Q\in\widetilde{\mathcal{Q}}_N}\ \mathrm{E}_Q[(\eta_k - z^{\mathrm{T}}\xi)_+ - (\eta_k - Y(\xi))_+] \leqslant 0,\ k = 1,\cdots,J, \\
& \sum_{j=1}^{n} z_j = 1, \\
& z_j \geqslant 0,\ j = 1,\cdots,n,
\end{aligned}
\tag{20}
$$

where $\widetilde{\mathcal{Q}}_N$ is the ambiguity set defined in (3).

From the strong duality result in [17, Corollary 2], problem (20) can be equivalently written as

$$
\begin{aligned}
\min_{z\in\mathbb{R}^n_+,\lambda_0\geqslant 0,\lambda\in\mathbb{R}^J_+}\ & \lambda_0\epsilon_N^p + \frac{1}{N}\sum_{i=1}^{N}\sup_{\xi\in\Xi}[-z^{\mathrm{T}}\xi - \lambda_0\|\xi - \widetilde{\xi}_i\|^p] \\
\text{s.t.}\ & \lambda_k\epsilon_N^p + \frac{1}{N}\sum_{i=1}^{N}\sup_{\xi\in\Xi}[(\eta_k - z^{\mathrm{T}}\xi)_+ - (\eta_k - Y(\xi))_+ - \lambda_k\|\xi - \widetilde{\xi}_i\|^p] \leqslant 0, \\
& k = 1,\cdots,J, \\
& \sum_{j=1}^{n} z_j = 1.
\end{aligned}
\tag{21}
$$

By introducing auxiliary variables, problem (21) can be reformulated as

$$
\begin{aligned}
\min_{z,\lambda_0,\lambda,\alpha,\beta,s}\ & \lambda_0\epsilon_N^p + \frac{1}{N}\sum_{i=1}^{N}\alpha_i \\
\text{s.t.}\ & \lambda_k\epsilon_N^p + \frac{1}{N}\sum_{i=1}^{N}\beta_{ik} \leqslant 0, k = 1,\cdots,J, \\
& \alpha_i \geqslant -z^{\mathrm{T}}\xi_j - \lambda_0\|\xi_j - \widetilde{\xi}_i\|^p,\ i = 1,\cdots,N,\ j = 1,\cdots,J, \\
& \beta_{ik} \geqslant s_{jk} - (\eta_k - Y(\xi_j))_+ - \lambda_k\|\xi_j - \widetilde{\xi}_i\|^p, \\
& i = 1,\cdots,N,\ j = 1,\cdots,J,\ k = 1,\cdots,J, \\
& s_{jk} \geqslant \eta_k - z^T\xi_j,\ j = 1,\cdots,J,\ k = 1,\cdots,J, \\
& \sum_{j=1}^{n} z_j = 1, \\
& z\in\mathbb{R}^n_+, \lambda_0 \geqslant 0, \lambda\in\mathbb{R}^J_+, \alpha\in\mathbb{R}^N, \beta\in\mathbb{R}^{N\times J}, s\in\mathbb{R}^{J\times J}_+.
\end{aligned}
\tag{22}
$$

Therefore, problem (20) can be solved through the linear programming reformulation (22), which can be efficiently solved by many optimization software. We solve it by Mosek solver in CVX package in MATLAB R2016a on a Dell G7 laptop with Windows 10 operating system, Intel Core i7-8750H processor, and 16 GB RAM.

4.2 Data

We select eight risky assets to constitute the stock pool, which are U.S. three-month treasury bills, U.S. long-term government bonds, S& P 500, Willshire 5000, NASDAQ, Lehmann Brothers corporate bond index, EAFE foreign stock index, and gold. We use the same historical annual return rate data as that in [4], whose statistics and more details can be found in Table 8.1 therein. The benchmark is the return rate of the equally weighted portfolio.

4.3 Numerical Evidences

Firstly, we examine the conservativeness of the data-driven distributionally robust model (20). We fix the sample set to be the support set (i.e., $\widetilde{\xi}_i = \xi_i,\ i = 1, \cdots, N$ with $N = J$) and solve problem (22) for $\epsilon_N = 0.2$, $p = 1, 2$. We also solve problem (19) with the empirical distribution $\widetilde{\mathrm{P}}_N$ as a comparison by using the solution method in [4]. The optimal values and the optimal solutions of the three models are shown in Table 1. We can see that problem (22) always gives a more conservative solution than problem (19) since $\widetilde{\mathrm{P}}_N$ is contained in the ambiguity set $\widetilde{\mathcal{Q}}_N$. The optimal values of the distributionally robust model are larger than that of the stochastic programming model under the true distribution, which can be viewed as the price of robustness.

Next, we investigate the trend of the optimal value when the sample size increases. We carry out 5 groups of tests with the sample size being $N = 5, 10, 20, 50, 100$, respectively. For each group of tests, we randomly generate N independent samples and solve the tested problems. Due to the randomness of sampling, for each group, we repeatedly generate the samples and test the model for 20 times, which provide 20 optimal solutions as well as 20 optimal values. Here, we set $\epsilon_N = 5/N$ to satisfy Assumption 3. We summarize in Table 2 the descriptive statistics of the optimal values for each group, which include maximum (max.), minimum (min.), median, mean, and standard deviation (std.). We can see from Table 2 that as the sample size increases, the maximum value, the minimum value, the median value, and the mean value of the optimal values all increase. Then, it is reasonable to infer that the optimal value increases with a high probability when the sample size increases. The standard devi-

Table 1 Comparison of data-driven distributionally robust model and the empirical model

Problem	Optimal value/%	Optimal solution
(22) with $p = 1$	-10.9371	(0, 0, 0.0299, 0.2329, 0, 0.3907, 0.2236, 0.1230)
(22) with $p = 2$	-11.0077	(0, 0, 0.0677, 0.1883, 0, 0.3914, 0.2309, 0.1217)
(19)	-11.0082	(0, 0, 0.0680, 0.1880, 0, 0.3914, 0.2309, 0.1217)

Table 2 Descriptive statistics of the optimal values under different sample sizes

Statistics	$N = 5$	$N = 10$	$N = 20$	$N = 50$	$N = 100$
Max/%	−7.9369	−8.9760	−8.7072	−7.9071	−9.3030
Min/%	−17.6376	−17.4964	−16.8976	−14.0841	−12.9365
Median/%	−14.0631	−13.0935	−11.9106	−11.5660	−11.2986
Mean/%	−13.2652	−12.9046	−12.2369	−11.5416	−11.2008
Std.	3.1050	2.0927	2.4285	1.4256	1.0013

ation of the optimal values decreases, which means that model (20) becomes more robust with the increase in the sample size.

Then, we adopt a box-plot to characterize the optimal values between mean±std., shown in Fig. 1. From Fig. 1, we can see that the box gets smaller as the sample size increases. This means that the optimal values fluctuate less and thus the model (20) becomes more robust with the increase in the sample size. We also observe that the mean value and the median value of the optimal values increase when the sample size increases, but their increase rates are decreasing. These observations verify the asymptotic convergence results in Theorem 6.

Finally, we briefly show the influence of the order p in the data-driven distributionally robust model (20). We carry out 4 groups of tests with $(p, N) = (1, 20), (2, 20), (1, 50), (2, 50)$, respectively. For each group, we repeat the tests for 20 times. Let $\epsilon = 5/N$. The box-plot showing the max., min., mean, mean±std, and median of the optimal values for the four groups is exhibited in Fig. 2. We can see that for fixed N, the model (20) with $p = 2$ generates a larger optimal value than that with

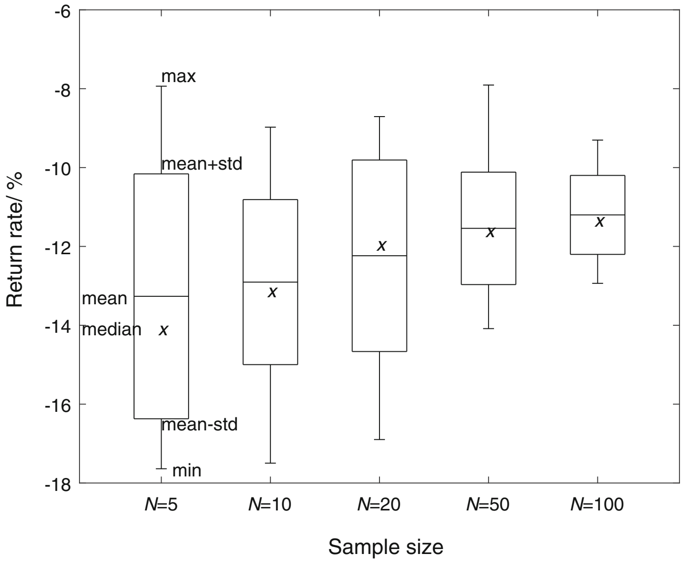

Fig. 1 Variation of the optimal value with respect to the sample size

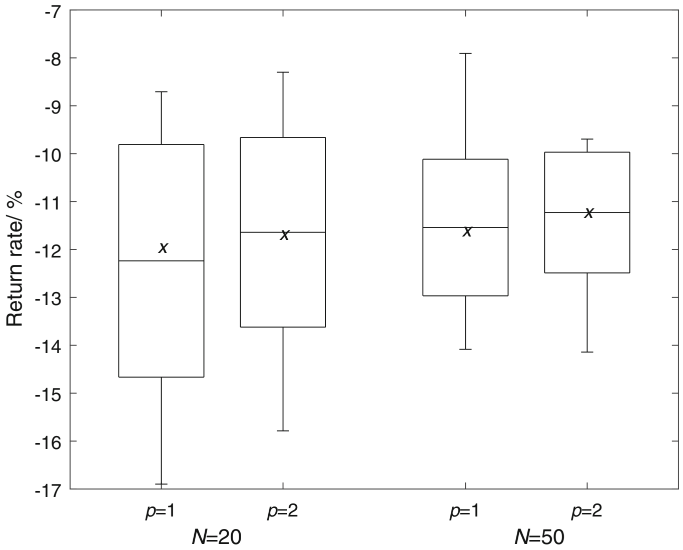

Fig. 2 Variation of the optimal value with respect to p and N

$p = 1$. Additionally, Fig. 2 verifies the asymptotic convergence results for $p = 2$ as well.

5 Conclusion

We studied a data-driven distributionally robust stochastic optimization problem with countably infinite constraints. We considered an ambiguity set which contains all probability distributions close to the empirical distribution measured under the Wasserstein distance.

We established the asymptotic convergence property of the distributionally robust optimization problem when the sample size goes to infinity. We proved that with probability 1, the optimal value and the optimal solution set of the data-driven distributionally robust optimization problem tend to those of the stochastic programming problem under the true distribution.

The asymptotic convergence properties lay a foundation for the practical solution and application of distributionally robust optimization problems with infinite constraints. Finally, we solved a data-driven distributionally robust portfolio optimization problem with second-order stochastic dominance constraints to numerically verify the theoretical results.

One of the future research topics would be the relaxation of assumptions in order to generalize the asymptotic convergence properties to non-smooth distributionally robust optimization problems.

Acknowledgements We are grateful to two anonymous referees for their insightful comments, which help us to improve the paper significantly. We also thank Dr. Mohajerin Esfahani for helpful discussions on conditions for the uniform convergence.

References

[1] Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming: Modeling and Theory. Society for Industrial and Applied Mathematics and Mathematical Programming Society, Philadelphia (2009)

[2] Noyan, N., Rudolf, G.: Optimization with stochastic preferences based on a general class of scalarization functions. Oper. Res. **66**(2), 463–486 (2018)

[3] Ji, Y., Qu, S., Wu, Z., Liu, Z.: A fuzzy-robust weighted approach for multicriteria bilevel games. IEEE Trans. Ind. Inf. (2020). https://doi.org/10.1109/TII.2020.2969456

[4] Dentcheva, D., Ruszczyński, A.: Optimization with stochastic dominance constraints. SIAM J. Optim. **14**(2), 548–566 (2003)

[5] Liu, Y., Xu, H.: Stability analysis of stochastic programs with second order dominance constraints. Math. Prog. **142**, 435–460 (2013)

[6] Liu, Y., Sun, H., Xu, H.: An approximation scheme for stochastic programs with second order dominance constraints. Numer. Algeb. Control Optim. **6**(4), 473–490 (2016)

[7] Sun, H., Xu, H., Wang, Y.: A smoothing penalized sample average approximation method for stochastic programs with second-order stochastic dominance constraints. Asia-Pacific J. Oper. Res. **30**(3), 548–554 (2013)

[8] Arrow, K.J., Karlin, S., Scarf, H.: Studies in the mathematical theory of inventory and production. Rev. Econ. Stat. **14**(69), 64–108 (1958)

[9] Žáčková, J.: On minimax solutions of stochastic linear programming problems. Čas. Pěst. Mat. **91**(4), 423–430 (1966)

[10] Ben-Tal, A., den Hertog, D., De Waegenaere, A., Melenberg, B., Rennen, G.: Robust solutions of optimization problems affected by uncertain probabilities. Manag. Sci. **59**(2), 341–357 (2013)

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

[12] Goh, J., Sim, M.: Distributionally robust optimization and its tractable approximations. Oper. Res. **58**(4), 902–917 (2010)

[13] Wiesemann, W., Kuhn, D., Sim, M.: Distributionally robust convex optimization. Oper. Res. **62**(6), 1358–1376 (2014)

[14] Ling, A., Sun, J., Xiu, N., Yang, X.: Robust two-stage stochastic linear optimization with risk aversion. Eur. J. Oper. Res. **256**(1), 215–229 (2017)

[15] Guo, S., Xu, H., Zhang, L.: Probability approximation schemes for stochastic programs with distributionally robust second-order dominance constraints. Optim. Methods Software **32**(4), 770–789 (2016)

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

[17] Gao, R., J.Kleywegt, A.: Distributionally robust stochastic optimization with Wasserstein distance, Working Paper. (2016). arXiv:1604.02199

[18] Zhao, C., Guan, Y.: Data-driven risk-averse stochastic optimization with Wasserstein metric. Oper. Res. Lett. **46**, 262–267 (2018)

[19] Bayraksan, G., Love, D.K.: Data-driven stochastic programming using phi-divergences. Turorials Oper. Res. (2015). https://doi.org/10.1287/educ.2015.0134

[20] Huang, R., Qu, S., Yang, X., Liu, Z.: Multi-stage distributionally robust optimization with risk aversion. J. Ind. Manag. Optim. **13**, 1–27 (2017)

[21] Pflug, G.C., Pichler, A., Wozabal, D.: The $1/n$ investment strategy is optimal under high model ambiguity. J. Bank. Finance **36**, 410–417 (2012)

[22] Mohajerin Esfahani, P., Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Math. Program. **171**(1–2), 115–166 (2018)

[23] Rachev, S.T., Rüschendorf, L.: Mass Transportation Problems. Springer, New York (1998)

[24] Fournier, N., Guillin, A.: On the rate of convergence in Wasserstein distance of the empirical measure. Probab. Theory Relat. Fields **162**(3), 707–738 (2015)

[25] Guo, S., Xu, H.: Distributionally robust shortfall risk optimization model and its approximation. Math. Program. **174**(1–2), 473–498 (2019)

[26] Kallenberg, O.: Foundations of Modern Probability. Springer, New York (1997)

[27] Pagnoncelli, B.K., Ahmed, S., Shapiro, A.: Sample average approximation method for chance constrained programming: theory and applications. J. Optim. Theory Appl. **142**(2), 399–416 (2009)