Numer. Math. Theor. Meth. Appl.
doi: 10.4208/nmtma.OA-2020-0125

Published online 5 January 2021

Discrete Approximation Scheme in Distributionally Robust Optimization

Yongchao Liu[1], Xiaoming Yuan[2] and Jin Zhang[3,*]

[1] *School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China*
[2] *Department of Mathematics, The University of Hong Kong, Hong Kong*
[3] *SUSTech International Center for Mathematics, Department of Mathematics, Southern University of Science and Technology, Shenzhen, China*

Received 15 August 2020; Accepted (in revised version) 9 November 2020

Abstract. Discrete approximation, which has been the prevailing scheme in stochastic programming in the past decade, has been extended to distributionally robust optimization (DRO) recently. In this paper, we conduct rigorous quantitative stability analysis of discrete approximation schemes for DRO, which measures the approximation error in terms of discretization sample size. For the ambiguity set defined through equality and inequality moment conditions, we quantify the discrepancy between the discretized ambiguity sets and the original set with respect to the Wasserstein metric. To establish the quantitative convergence, we develop a Hoffman error bound theory with Hoffman constant calculation criteria in a infinite dimensional space, which can be regarded as a byproduct of independent interest. For the ambiguity set defined by Wasserstein ball and moment conditions combined with Wasserstein ball, we present similar quantitative stability analysis by taking full advantage of the convex property inherently admitted by Wasserstein metric. Efficient numerical methods for specifically solving discrete approximation DRO problems with thousands of samples are also designed. In particular, we reformulate different types of discrete approximation problems into a class of saddle point problems with completely separable structures. The stochastic primal-dual hybrid gradient (PDHG) algorithm where in each iteration we update a random subset of the sampled variables is then amenable as a solution method for the reformulated saddle point problems. Some preliminary numerical tests are reported.

AMS subject classifications: 90C15, 90C46, 90C47

Key words: Quantitative stability analysis, Hoffman lemma, discrete approximation method, distributionally robust optimization, stochastic primal-dual hybrid gradient.

[*] Corresponding author. *Email addresses:* `lyc@dlut.edu.cn` (Y. Liu), `xmyuan@hku.hk` (X. Yuan), `zhangj9@sustech.edu.cn` (J. Zhang)

1. Introduction

Making an optimal decision under uncertain conditions is typically a challenging task in decision analysis. The quality of such decisions relies heavily on the information concerning the underlying uncertainties. Stochastic programming (SP) is a powerful tool if the uncertainty distributions are completely known. The history of SP can be traced back to the middle of the last century. So far, a variety of SP models have been proposed to handle the presence of random data in optimization problems. Prevailing examples include chance-constrained models, two-and multi-stage models, and models involving risk measures. For recent developments on SP, we refer the readers to a monograph [39] and references therein.

Unfortunately, in most real-life applications such as signal processing of mobile ad hoc networks, the number of samples collected is relatively small. Evaluating the exact probability of safe operation is somewhat challenging. One remedy for this difficulty is to adopt a distributionally robust approach. That is, constructing an ambiguity set of distributions with historical data, computer simulations or subjective judgements that contains the true distribution with certain confidence. Thus we may choose an optimal decision on the basis of the worst-case distribution over the ambiguity set. For example, some samples can be obtained, it may be more reliable to estimate the moment information than to evaluate the exact probability. This type of robust optimization framework can be traced back to the earlier work by Scarf [40]. It has been thoroughly investigated through research, see, e.g., Žáčková [51], Dupačová [14], and Shapiro and Ahmed [42]. Over the past few years, it has gained substantial popularity through further contributions by Bertsimas and Popescu [5], Betsimas *et al.* [4], Delage and Ye [11], Goldfarb and Iyengar [19], Mehrotra and Papp [27], Pflug *et al.* [31], Wiesemann *et al.* [46,47].

One important issue concerning distributionally robust optimization (DRO) is the design of numerical solvers. Unlike robust optimization problems, DRO problems usually contain functional variables. Designing implementable and efficient numerical schemes for solving DRO becomes extraordinarily challenging. Most studies of DRO have focused on a dual approach. In summary, the technical framework of such approach has three steps:

- Consider the Lagrange dual of the inner max problem.
- Reformulate the min-max problem as a min-min (combining the min-min by min) problem with semi-infinite constraints.
- Recast the semi-infinite constraints as a linear semi-definite constraint by S-Lemma or dual method again.

In particular, Wiesemann *et al.* [48] provides a unified framework of the semi-definite programming (SDP) reformulation for DRO problems where the ambiguity set is constructed through some probabilistic and moment constraints.

Apart from the mentioned conic reformulation, another important approach pioneered by Pflug and Wozabal [31] is to discretize the ambiguity set of DRO problems. The validity of this scheme depends on the fact that the discretized min-max optimization problem is readily tractable in the numerical optimization context. More recently, Mehrotra and Papp [27] popularize this discretization approach to a general class of DRO problems. In particular, they design a process that generates a cutting surface of the inner optimal value at each iteration. Xu *et al.* [49] also suggest discretizing DRO problems with moment ambiguity sets, while cutting-plane methods are used to solve the discretized subproblems. Liu *et al.* [26] introduce a primal-dual type method to solve the discrete approximations of DRO problems.

Unlike the dual (tractable reformulation) approach whose validity relies heavily on specific structures of involved functions and ambiguity set, the discrete approximation method usually works under less restrictive conditions. Thanks to this feature, the discrete approximation method allows decision makers to construct more informative ambiguity set and hence broaden the applicable horizon of DRO model to a wider setting. Typically, the discrete approximation method stays effective when the ambiguity set is constructed through moment condition with bounded support set, or characterized by combining the moment information and statistical-based distance. However, practical implementation of the discrete approximation method requires solving a sequence of discretized min-max subproblems. The efficiency of this approach then relies substantially on the load of computing subproblems. When the sample size is larger, solving subproblems becomes a more challenging task. In order to address this issue, it is crucial to establish quantitative proximity from the approximated solution to the optimal solution in terms of the sample size. In stochastic programming, this stability analysis technique which is pioneered by Römisch in the 1980's [36], has become widely-adopted. It has been intensively studied for a wide range of stochastic programming models, such as chance-constrained problems [38] and stochastic dominance [12]; see the survey paper [37] and references therein.

Stability analysis of DRO problems with respect to perturbation on the ambiguity set, however, is still in its infancy. It has been recently studied in two notable works [44, 52]. They present an asymptotic stability analysis of DRO problems under total variation metric. That is, as the difference between two ambiguity sets tends to zero under the total variation metric, the associated difference between the optimal solutions/optimal values vanishes. However, the total variation metric between a discrete probability and a continuous probability equal to 1. Apparently this distance characterization is far from suitably acceptable, so the total variation metric cannot be regarded as an appropriate measure for discrete approximation schemes on DRO problems. To overcome this issue, one work analyzes the quantitative stability for DRO problems under Wasserstein metric (see, e.g., [15, 25]). They derive a new variant of Hoffman's error bound lemma for the moment problem under the Slater condition [25], i.e., the distance between certain probability measure and the ambiguity set $\mathcal{P}$ under a generic metric with ζ-structure can be estimated through the residual of the moment system. Unfortunately, the prerequisite Slater condition excludes those cases where the ambiguity sets are defined through equality moment conditions. As commented in [25], "this is a significant limitation."

This present paper aims to complete the picture of stability analysis of the discrete approximation on DRO problems. The main contributions of our paper can be summarized as follows.

- Both recent articles [44, 52] are closely related to error bound theory, while the difference between these two papers is the spaces where the linear systems are embedded. Note that the probability space under total variation metric is indeed a Banach space. Based on this observation, in this present paper, we aim to illuminate and simplify some of the results in [44, 52]. In particular, inspired by the celebrated Hoffman lemma in Banach space [6, Theorem 2.200], we shall provide a new proof of error bound in Theorem 3.1 in order to streamline the arguments in two previous works [44, 52].

- By taking advantage of some established results in [30], together with the calculus criteria for Hoffman's error bound radius in Euclidean space [24], we shall establish a quantitative estimation for the distance between $\mathcal{P}$ and its discretized counterpart under Wasserstein metric in terms of the Hausdorff distance between the support sets of the two ambiguity sets of probability measures, see, e.g., Theorem 3.2. Our new results further streamline and illuminate the arguments in three previous works [25, 44, 52], while improving their results in two ways.

  1. Compared to [44,52], we present the quantitative stability of ambiguity sets through the more appropriate Wasserstein metric instead of total variation metric.
  2. Compared to [25], we focus on more general ambiguity sets which are defined through both equality and inequality moment conditions instead of only inequality constraints. Moreover, we encounter a surprise that the quantitative error bound estimation can be established in the absence of any regularity conditions. This error bound result significantly improves the previous works [25, 44, 52] where the Slater (or Slater type) conditions are required.

  We also consider the ambiguity set which is characterized through moment information together with probability distance. This new characterization helps us to exclude pathological distributions more efficiently. By taking advantage of the convexity inherently obsessed by Wasserstein metric, we present a quantitative connection between the distance between certain ambiguity set and its discretized counterpart under Wasserstein metric, see, e.g., Theorem 3.5. Moreover, with quantification of the difference between the ambiguity set and its discrete approximation, we present the quantitative stability analysis of discrete approximation schemes for DRO problems, which can be regarded as an extension of Römisch's [37] stability results on stochastic programming problem to DRO problem, see, e.g., Theorem 4.2.

- To efficiently reduce the computation cost associated with large sample size in an efficient manner, we reformulate the approximation problem as a saddle point problem with completely separable structures. A stochastic version primal-dual hybrid gradient (PDHG) algorithm [8] is therefore amenable as a solution method. We report the numerical results on a practical portfolio application that demonstrate the superiority of our approach.

An important byproduct of our analysis, worthy of independent interest, relates the error bound radius of a generalized linear system in infinite dimensional space. In fact, in Theorem 3.3, we improve the celebrated Hoffman's error bound in Banach space (see [6, Theorem 2.200]) in the sense that we shall provide an explicit calculus criterion of Hoffman constant.

Throughout this paper, we use the following notations. For vectors $a, b \in {\rm I\!R}^n$, a^Tb denotes the scalar product, $\|a\|$, $\|a\|_1$ and $\|a\|_\infty$ denote the Euclidean norm, 1-norm and supremum norm, respectively. $\langle \cdot, \cdot \rangle$ denotes a bilinear representation of the expected value. The dual norm of $\|\cdot\|_p$, $p = 1, 2, \infty$, is $\|a\|_{p^*} := \max\{a^Tb : b \in {\rm I\!R}^n, \|b\|_p = 1\}$. Let $d_p(x, D) := \inf_{x' \in D} \|x - x'\|_p$ denote the p-distance from a point x to a set D. For two compact sets $\mathcal{C}$ and $\mathcal{D}$, $\mathscr{D}(\mathcal{C}, \mathcal{D}) := \sup_{x \in \mathcal{C}} d_p(x, \mathcal{D})$ denotes the deviation of $\mathcal{C}$ from $\mathcal{D}$ and $\mathcal{H}(\mathcal{C}, \mathcal{D}) := \max(\mathscr{D}(\mathcal{C}, \mathcal{D}), \mathscr{D}(\mathcal{D}, \mathcal{C}))$ denotes the Hausdorff distance between $\mathcal{C}$ and $\mathcal{D}$. Moreover, $\mathcal{C} + \mathcal{D}$ denotes the Minkowski addition of the two sets, that is, $\{C + D : C \in \mathcal{C}, D \in \mathcal{D}\}$. For a sequence of subsets $\{\mathcal{C}_k\}$ in a metric space, we follow the standard notation [35] by using $\limsup_{k \to \infty} \mathcal{C}_k$ to denote its outer limit, that is, $\limsup_{k \to \infty} \mathcal{C}_k = \{x : \liminf_{k \to \infty} d(x, \mathcal{C}_k) = 0\}$.

2. Metrics of probability measures

In probability theory, various metrics have been introduced to quantify the difference between two probability measures; see [2, 18]. In this part, we specifically focus on the Wasserstein metric and the total variation metric, which have been widely used to study distributionally robust optimization problems, see, e.g., [15, 44].

Definition 2.1. Let $P, Q \in \mathscr{P}(\Xi)$. The Wasserstein metric between P and Q is defined as

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

where $\mathscr{P}(\Xi)$ denotes the set of all probability measures over the set Ξ and

$$\mathscr{G} = \big\{ g : \Xi \to \mathbb{R} : g \text{ is Lipschitz continuous and Lipschtiz modulus } L_g \leq 1 \big\}.$$

By the Kantorovich-Rubinstein theorem, the Wasserstein metric is equivalent to the Kantorovich metric. Then for any $P, Q \in \mathscr{P}(\Xi)$, we have

$$\mathsf{dl}_{\mathsf{W}}(P, Q) = \inf \left\{ \int_{\Xi^2} \|\xi_1 - \xi_2\| \pi(d\xi_1, d\xi_2) \right\},$$

where π is a joint distribution of ξ_1 and ξ_2 with marginal P and Q, respectively, and the 'inf' is taken over all joint distributions π.

Based on the Wasserstein metric, for subsets $\mathcal{Q}$ and $\mathcal{Q}'$ of $\mathscr{P}(\Xi)$, we may define the distance from a point Q to the set $\mathcal{Q}$ as

$$\mathsf{dl}_{\mathsf{W}}(Q,\mathcal{Q}) := \inf_{P\in\mathcal{Q}} \mathsf{dl}_{\mathsf{W}}(Q,P),$$

the deviation from the set $\mathcal{Q}'$ to the other set $\mathcal{Q}$ as

$$\mathbb{D}_{\mathsf{W}}(\mathcal{Q}',\mathcal{Q}) := \sup_{Q\in\mathcal{Q}'} \mathsf{dl}_{\mathsf{W}}(Q,\mathcal{Q}),$$

the Hausdorff distance between $\mathcal{Q}$ and $\mathcal{Q}'$ in the space of probability measures $\mathscr{P}(\Xi)$ as

$$\mathbb{H}_{\mathsf{W}}(\mathcal{Q}',\mathcal{Q}) := \max\left\{\mathbb{D}_{\mathsf{W}}(\mathcal{Q}',\mathcal{Q}),\ \mathbb{D}_{\mathsf{W}}(\mathcal{Q},\mathcal{Q}')\right\}.$$

Definition 2.2. Let $P, Q \in \mathscr{P}(\Xi)$. The total variation metric between P and Q is defined as

$$\mathsf{dl}_{\mathrm{T}}(P,Q) := \sup_{g\in\mathscr{G}} \left|\mathbb{E}_P[g(\xi)] - \mathbb{E}_Q[g(\xi)]\right|,$$

where

$$\mathscr{G} := \left\{g : \Xi \to \mathbb{R} : g \text{ is } \mathscr{B} \text{ measurable, } \sup_{\xi\in\Xi} |g(\xi)| \le 1\right\}.$$

The deviation from one set to the other and the Hausdorff distance between two sets in the space of probability measures $\mathscr{P}(\Xi)$ under total variation metric are defined respectively as

$$\begin{aligned}\mathbb{D}_{\mathrm{T}}(\mathcal{Q}',\mathcal{Q}) &:= \sup_{Q\in\mathcal{Q}'} \mathsf{dl}_{\mathrm{T}}(Q,\mathcal{Q}),\\ \mathbb{H}_{\mathrm{T}}(\mathcal{Q}',\mathcal{Q}) &:= \max\left\{\mathbb{D}_{\mathrm{T}}(\mathcal{Q}',\mathcal{Q}), \mathbb{D}_{\mathrm{T}}(\mathcal{Q},\mathcal{Q}')\right\}.\end{aligned}$$

3. Stability of the ambiguity sets

The key issue in stability analysis is to quantify stability of feasible set [12,37] with respect to some probability metrics. Under extra Lipscthiz and/or growth condition of the objective function, the stability of optimal value and optimal solutions can be further quantified. In DRO problems, the ambiguity set is regarded as the feasible set. In this section, we focus on the quantitative stability of ambiguity sets under certain perturbations.

3.1. Moment type ambiguity set

There are different approaches to construct the ambiguity set of distributions for the DRO problem. Among them, the moment type condition gains popularity. The moment condition is motivated by the fact that given some historical data, the estimation of the moments of random parameters is typically easier than the derivation of their true probability distributions. Over the past several years, DRO problems with moment constraints have been intensively studied, see, e.g., [11,44,48,55]. Our purpose in this subsection is to study the moment type ambiguity set

$$\mathcal{P} := \big\{P \in \mathscr{P}(\Xi) : \mathbb{E}_P[\psi_{\rm E}(\xi)] = \mu_{\rm E}, \mathbb{E}_P[\psi_{\rm I}(\xi)] \leq \mu_{\rm I}\big\}, \tag{3.1}$$

where $\mathscr{P}(\Xi)$ denotes the set of all probability measures over the set Ξ, $\psi_{\rm E} : \Xi \to {\rm I\!R}^p$ and $\psi_{\rm I} : \Xi \to {\rm I\!R}^{q-p}$ are random mappings and $(\mu_{\rm E}, \mu_{\rm I})$ is the prior moment information of $(\psi_{\rm E}, \psi_{\rm I})$.

The next theorem is first established by Sun and Xu in [44]. It states that the distance between a given probability Q and $\mathcal{P}$ under total variation metric is linearly bounded by the residual of the moment system. As promised, we significantly simplify the proof and illuminate the result in [44] from an error bound perspective.

Theorem 3.1. *Assume the Slater type conditions hold, i.e.,*

$$0_q \in {\rm int}\big\{\mathbb{E}_P[\psi(\xi)] : P \in \mathscr{P}(\Xi)\big\} - \mathcal{K},$$

where int *denotes the interior of a set,*

$$\psi(\xi) := \begin{bmatrix} \psi_{\rm I}(\cdot) \\ \psi_{\rm E}(\cdot) \end{bmatrix}$$

and $\mathcal{K} := 0_p \times {\rm I\!R}^{q-p}_+$. *Then there exists a positive constant* κ_1 *such that*

$$\mathsf{dl}_{\rm T}(Q, \mathcal{P}) \leq \kappa_0\big(\|\langle Q, \psi_{\rm E}(\xi)\rangle - \mu_{\rm E}\| + \|(\langle Q, \psi_{\rm I}(\xi)\rangle - \mu_{\rm I})_+\|\big), \quad \forall Q \in \mathscr{P}(\Xi). \tag{3.2}$$

Proof. Obviously, we have

$$\begin{aligned}\mathcal{P} &= \big\{P \in \mathscr{P}(\Xi) : \mathbb{E}_P[\psi_{\rm E}(\xi)] = \mu_{\rm E}, \mathbb{E}_P[\psi_{\rm I}(\xi)] \leq \mu_{\rm I}\big\} \\ &= \big\{P \in \mathscr{M}_+(\Xi) : \langle P, 1\rangle = 1, \langle P, \psi_{\rm E}(\xi)\rangle = \mu_{\rm E}, \langle P, \psi_{\rm I}(\xi)\rangle \leq \mu_{\rm I}\big\},\end{aligned}$$

where $\mathscr{M}_+(\Xi)$ denotes the set of all measures on Ξ. Note that total variation is a norm defined on the space of measures with bounded variation and thus $\mathscr{M}_+(\Xi)$ is a Banach space equipped with total variation norm. By [6, Theorem 2.200], straightforwardly there exists a positive number κ_0 such that

$$\begin{aligned}\mathsf{dl}_{\rm T}(Q, \mathcal{P}) \leq \kappa_0\big(&\|\langle Q, 1\rangle - 1\| + \|\langle Q, \psi_{\rm E}(\xi)\rangle - \mu_{\rm E}\| \\ &+ \|(\langle Q + \psi_{\rm I}(\xi)\rangle - \mu_{\rm I})_+\|\big), \quad \forall Q \in \mathscr{M}_+(\Xi).\end{aligned}$$

Subsequently, by restricting $Q \in \mathscr{P}(\Xi)$, we prove (3.2). □

Following [44], Zhang *et al.* [52] extend the error bound result to a general cone constrained moment system. The proofs in both [44,52] depend on the reformulation of the distance as a min-max linear programming problem through the Lagrangian duality. We provide a more transparent proof in terms of celebrated error bound theory which streamlines the arguments in [44] and [52].

Moreover, based on Theorem 3.1, Sun and Xu [44] consider the canonically perturbed system

$$\tilde{\mathcal{P}} := \left\{P \in \mathscr{P}(\Xi) : \mathbb{E}_P[\psi_{\mathrm{E}}(\xi)] = \mu_{\mathrm{E}}^{\mathrm{N}}, \mathbb{E}_P[\psi_{\mathrm{I}}(\xi)] \le \mu_{\mathrm{I}}^{\mathrm{N}}\right\},$$

and quantify the proximity from $\tilde{\mathcal{P}}$ to $\mathcal{P}$ under the total variation metric

$$\mathbb{H}_{\mathrm{T}}(\tilde{\mathcal{P}}, \mathcal{P}) \le \kappa_0\left(\|\mu_{\mathrm{I}}^{\mathrm{N}} - \mu_{\mathrm{I}}\| + \|\mu_{\mathrm{E}}^{\mathrm{N}} - \mu_{\mathrm{E}}\|\right). \tag{3.3}$$

Note that (3.3) has been used to study the stability of one-stage DRO problems when the prior information $(\mu_{\mathrm{E}}, \mu_{\mathrm{I}})$ is perturbed. However, Theorem 3.1 fails to offer an appropriate measure for a discrete approximation of the ambiguity set since the total variation between a discrete probability and a continuous probability is identically 1.

In order to characterize the convergence of discrete distributions of the ambiguity set, Liu *et al.* [25] study the stability under Wasserstein metric instead of total variation metric. Moreover, an explicit expression of the error bound constant κ_0 has been provided. In particular, [25] considers the case where the ambiguity set is defined through moment inequality constraints

$$\mathcal{C} = \left\{P \in \mathscr{P}(\Xi) : \mathbb{E}_P[\psi_{\mathrm{I}}(\xi)] \le \mu_{\mathrm{I}}\right\}$$

as well as its discrete approximation

$$\mathcal{C}_{\mathrm{N}} = \left\{P \in \mathscr{P}(\Xi^{\mathrm{N}}) : \mathbb{E}_P[\psi_{\mathrm{I}}(\xi)] \le \mu_{\mathrm{I}}\right\},$$

where $\mathscr{P}(\Xi^{\mathrm{N}})$ denotes the set of all probability measures over the support Ξ^{N} and $\Xi^{\mathrm{N}} := \{\xi^1, \cdots, \xi^{\mathrm{N}}\}$ is a set of points in Ξ. As Ξ^{N} is a discrete set, $\mathcal{C}_{\mathrm{N}}$ is set of discrete distributions. Under the Slater condition, i.e., there exists $\bar{P} \in \mathscr{P}(\Xi)$ and a positive number $\delta > 0$ such that $\mathbb{E}_{\bar{P}}[\psi_{\mathrm{I}}(\xi)] + \delta \le \mu_{\mathrm{I}}$, [25, Theorem 12] provides a quantitative description for the difference between two ambiguity sets

$$\mathbb{H}_{\mathrm{W}}(\mathcal{C}_{\mathrm{N}}, \mathcal{C}) \le \kappa_2 \beta_{\mathrm{N}}, \tag{3.4}$$

where κ_2 is a positive number depending on the diameter of Ξ and the Lipschitz modulus of ψ_{I}, and

$$\beta_{\mathrm{N}} := \max_{\xi \in \Xi} \min_{1 \le i \le N} d(\xi, \xi_i). \tag{3.5}$$

Compared with the results in [44] (see also (3.3)), (3.4) sheds some light on numerical implementation of the discrete approximation scheme for DRO problems; see, e.g., [31,49]. However, the Slater condition restricts the application of (3.4) to moment constraint system with equality constraints. This fact limits the scientific contributions of (3.4) as a large number of interesting moment based ambiguity sets in the literature involving equality constraints.

The limitation inspires us to revisit discrete approximation techniques for DRO problems. We shall establish error bound results for ambiguity sets defined through equality and inequality moment constraints, which can be regarded as an extension of (3.4). Compared with existing results, we mainly focus on the proximity of $\mathcal{P}_{\rm N}$ in (3.6) to $\mathcal{P}$ in (3.1) and hence present a quantitative stability analysis without any Slater type conditions. To streamline the idea of discretization, we let $\Xi^{\rm N} := \{\xi^1, \cdots, \xi^{\rm N}\} \subset \Xi$ be a set of points in Ξ. These points may be samples of ξ or selected in deterministic manner. We then consider the discrete approximation of $\mathcal{P}$ defined in (3.1) as

$$\mathcal{P}_{\rm N} := \big\{P \in \mathscr{P}(\Xi^{\rm N}) : \mathbb{E}_P[\psi_{\rm E}(\xi)] = \mu_{\rm E}, \mathbb{E}_P[\psi_{\rm I}(\xi)] \leq \mu_{\rm I}\big\}, \tag{3.6}$$

where $\mathscr{P}(\Xi^{\rm N})$ denotes the set of all probability measures over the support $\Xi^{\rm N}$. Obviously, $\mathcal{P}_{\rm N} \subset \mathcal{P}$ as $\Xi^{\rm N} \subseteq \Xi$. However, the difference between $\mathcal{P}_{\rm N}$ and $\mathcal{P}$ is unclear.

Before presenting the quantitative convergence of $\mathcal{P}_{\rm N}$ to $\mathcal{P}$, we recall some necessary preliminaries. For given $\Xi^{\rm N} = \{\xi^1, \cdots, \xi^{\rm N}\}$, let $\{\Xi_1, \cdots, \Xi_{\rm N}\}$ be a Voronoi tessellation of Ξ (see [30]), i.e.,

$$\Xi_{\rm i} \subseteq \Big\{y \in \Xi : \|y - \xi^i\| = \min_{1\leq k\leq N} \|y - \xi^k\|\Big\} \quad \text{for} \quad i = 1, \ldots, N$$

are pairwise disjoint subsets forming a partition of Ξ. For a fixed $P \in \mathscr{P}(\Xi)$, let $p_i = P\{\xi \in \Xi_{\rm i}\}$ for $i = 1, \ldots, N$ and define

$$P^r_{\rm N}(\cdot) := \sum_{i=1}^N p_i \, \delta_{\xi_i}(\cdot).$$

We call $P^r_{\rm N}(\cdot)$ the Voronoi projection of the probability measure P on space $\mathscr{P}(\Xi^{\rm N})$. By definition of Wasserstein metric $\mathsf{dl}_{\rm W}(\cdot)$ and (3.5), the following estimation holds [30]

$$\mathsf{dl}_{\rm W}(P, P^r_{\rm N}) = \int \min_{1\leq i\leq N} d(\xi, \xi_i) dP = \sum_{i=1}^N \int_{\Xi_i} d(\xi, \xi_i) dP \leq \beta_{\rm N}. \tag{3.7}$$

Another result we need is the error bound condition of the linear system in Euclidean space. Since the first paper by Hoffman [20] on 1950's, many efforts have been devoted to the calculus of error bound radius. We recall the calculus criteria introduced in [24]. Before we can do so, we first define a linear system

$$\mathcal{F} := \big\{x \in {\rm I\!R}^n : Ax \leq b, Cx = d\big\}.$$

Then, according to [24], the distance from any given point x to $\mathcal{F}$ can be estimated,

$$d_1(x, \mathcal{F}) \leq \kappa\big(\|(Ax - b)_+\|_1 + \|Cx - d\|_1\big), \tag{3.8}$$

where

$$\kappa := \sup\left\{ \|u, v\|_\infty, \begin{array}{c} \|A^Tu + C^Tv\|_\infty = 1, \text{ the rows of} \\ \begin{pmatrix} A \\ C \end{pmatrix} \text{corresponding to nonzero} \\ \text{elements of} \begin{pmatrix} u \\ v \end{pmatrix} \text{are linearly independent} \end{array} \right\}. \tag{3.9}$$

We are now ready to present the main theorem of this section, i.e., a quantitative approximation of $\mathcal{P}_{\rm N}$ to $\mathcal{P}$ under Wasserstein metric. Together with the Voronoi projection and Hoffman error bound radius calculation in finite dimensional spaces, we take full advantage of some intrinsic properties of Wasserstein metric.

Theorem 3.2. *Suppose, (a) Ξ is compact, and (b) $\psi_{\rm E}(\cdot)$ and $\psi_{\rm I}(\cdot)$ are Lipschitz continuous on Ξ with bounded modulus L_ψ. Then, for any N,*

$$\mathbb{H}_{\rm W}(\mathcal{P}_{\rm N},\mathcal{P}) \le \kappa_1\beta_{\rm N}, \tag{3.10}$$

where $\beta_{\rm N}$ is defined in (3.5) *and*

$$\kappa_1 := (1+\kappa_{\rm N} q L_\psi \mathrm{Diam}_\Xi) \tag{3.11}$$

with $\kappa_{\rm N}$ estimated by (3.9)*, q denoting the dimension of $(\psi_{\rm E},\psi_{\rm I})$ and* Diam$_\Xi$ *denoting the diameter of Ξ.*

Proof. By the definitions of $\mathcal{P}$ and $\mathcal{P}_{\rm N}$, $\mathcal{P}_{\rm N}\subset\mathcal{P}$ in that $\Xi^{\rm N}\subset\Xi$. It is sufficient to show that (3.10) holds for the deviation $\mathbb{D}_{\rm W}(\mathcal{P},\mathcal{P}_{\rm N})$. Recall that for any fixed $P\in\mathcal{P}$, $P_{\rm N}^r$ denotes Voronoi projection of P. If $P_{\rm N}^r\in\mathcal{P}_{\rm N}$, then

$$\mathsf{dl}_{\rm W}(P,\mathcal{P}_{\rm N}) \le \mathsf{dl}_{\rm W}(P,P_{\rm N}^r) \le \beta_{\rm N}, \tag{3.12}$$

where the second inequality follows from (3.7). Thus, we are left with the case with $P_{\rm N}^r\notin\mathcal{P}_{\rm N}$. As Ξ is a compact set, $\mathcal{P}_{\rm N}$ is weakly compact with respect to topology of weak convergence [30]. Then we may denote $Q_{\rm N}$ as the projection of $P_{\rm N}^r$ on $\mathcal{P}_{\rm N}$, that is, $Q_{\rm N}\in\mathcal{P}_{\rm N}$ and $\mathsf{dl}_{\rm W}(P_{\rm N}^r,Q_{\rm N})=\mathsf{dl}_{\rm W}(P_{\rm N}^r,\mathcal{P}_{\rm N})$. By the triangle inequality of the Wasserstein metric, we have

$$\mathsf{dl}_{\rm W}(P,\mathcal{P}_{\rm N}) \le \mathsf{dl}_{\rm W}(P,Q_{\rm N}) \le \mathsf{dl}_{\rm W}(P,P_{\rm N}^r)+\mathsf{dl}_{\rm W}(P_{\rm N}^r,Q_{\rm N}). \tag{3.13}$$

In what follows, we present an upper bound of $\mathsf{dl}_{\rm W}(P_{\rm N}^r,Q_{\rm N})$ with $\beta_{\rm N}$.

By the Kantorovich-Rubinstein theorem, the Wasserstein metric is equivalent to the Kantorovich metric. Then by the definition of Kantorovich metric, we have

$$\mathsf{dl}_{\rm W}(P_{\rm N}^r,Q_{\rm N}) \le \mathrm{Diam}_\Xi\|P_{\rm N}^r-Q_{\rm N}\|_1, \tag{3.14}$$

where Diam$_\Xi$ denotes the diameter of Ξ. Note that $P_{\rm N}^r$ and $Q_{\rm N}$ are in Euclidean space, and Hoffman error (3.8) implies

$$\begin{aligned}\|P_{\rm N}^r-Q_{\rm N}\|_1 &= d_1(P_{\rm N}^r,\mathcal{P}_{\rm N})\\ &\le \kappa_{\rm N}\big(\|\langle P_{\rm N}^r,\psi_{\rm E}(\xi)\rangle-\mu_{\rm E}\|_1+\|(\langle P_{\rm N}^r,\psi_{\rm I}(\xi)\rangle-\mu_{\rm I})_+\|_1\big)\\ &\le \kappa_{\rm N}\big(\|\langle P_{\rm N}^r,\psi_{\rm E}(\xi)\rangle-\langle P,\psi_{\rm E}(\xi)\rangle\|_1+\|\langle P_{\rm N}^r,\psi_{\rm I}(\xi)\rangle-\langle P,\psi_{\rm I}(\xi)\rangle\|_1\big)\\ &\le \kappa_{\rm N} q L_\psi \mathsf{dl}_{\rm W}(P,P_{\rm N}^r)\\ &\le \kappa_{\rm N} q L_\psi \beta_{\rm N},\end{aligned} \tag{3.15}$$

where the first inequality follows from the Hoffman error bound (3.8), the second inequality follows from the feasibility of P which means

$$\langle P, \psi_{\rm E}(\xi)\rangle = 0, \quad \langle P, \psi_{\rm I}(\xi)\rangle \le 0,$$

the third inequality follows from the definition of $\mathsf{dl}_{\rm W}(\cdot)$ and the Lipschitz continuity of $\psi_{\rm E}(\cdot)$ and $\psi_{\rm E}(\cdot)$.

Combining (3.12)-(3.15), we obtain that desired inequality that

$$\mathsf{dl}_{\rm W}(P, \mathcal{P}_{\rm N}) \le (1 + \kappa_{\rm N} q L_{\psi} \text{Diam}_{\Xi})\beta_{\rm N}.$$

This completes the proof. □

As we emphasized, Theorem 3.2 illuminates the arguments in [44, 52] from an error bound perspective, while simplifying and improving their results. Moreover, in the absence of the Slater condition, Theorem 3.2 also extends [25] to the case where the ambiguity set is defined through equality and inequality constraints.

Remark 3.1. It is known that the propagation of the discrepancy (approximation error) for a DRO problem quantifies the difference between the discretized ambiguity set and the original one under some appropriate metrics. Based on Theorem 3.2, it is not difficult to quantify the optimal values and optimal solutions of one-stage DRO problem [25,44,52] or two-stage DRO problem [32] where the ambiguity set is defined by generalized prior equality and inequality moment conditions.

The discrete set $\Xi^{\rm N}$ can be generated in different ways. If $\Xi^{\rm N}$ is constructed in a deterministic way, we can explicitly characterize $\kappa_{\rm N}$ through (3.9). Moreover, for the discrete sets $\Xi^{\rm N'} \subset \Xi^{\rm N''}$, the Hoffman constant $\kappa_{\rm N''} \le \kappa_{\rm N'}$. Therefore, κ_1 in Theorem 3.2 is monotonically decreasing with respect to increasing N. This observation inspires an approachable scheme to estimate $\kappa_{\rm N}$; see the following simple Example 3.1 for illustration.

Example 3.1. Consider the following ambiguity set

$$\mathcal{P} = \left\{ P \in \mathscr{P}([a,b]) : \begin{array}{l} \mathbb{E}_P[\xi] = \mu_0, \\ \mathbb{E}_P\big[(\xi - \mu_0)(\xi - \mu_0)\big] \le \sigma_0 \end{array} \right\},$$

where μ_0 and σ_0 denote the estimation of mean and variance of the random variable ξ respectively, $0 < a < b < \infty$. The first equality constraint means that we have complete information on mean value. Let $\Xi^{\rm N} := \{\xi_1, \cdots, \xi_N\} \subseteq [a,b]$. We may consider the discrete approximation

$$\mathcal{P}_{\rm N} = \left\{ P \in \mathscr{P}(\Xi^{\rm N}) : \begin{array}{l} \mathbb{E}_P[\xi] = \mu_0, \\ \mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)] \le \sigma_0 \end{array} \right\}.$$

The estimation of the Hoffman constants can be obtained by solving a convex programming problem in the form of (3.9). Obviously, the feasible set of optimization problem in the form of (3.9) turns out to be

$$\big\{(u,v); v \ge 0,\ |u\xi_i + v(\xi_i - \mu_0)^2| \le 1, i = 1, \ldots, N\big\}.$$

When N increases, the above feasible set is shrinking. Thus $\Xi^{\mathrm{N'}} \subseteq \Xi^{\mathrm{N''}}$ if $N'' > N'$, which implies that $\kappa_{\mathrm{N''}} \leq \kappa_{\mathrm{N'}}$. Subsequently, we may first calculate κ_{M} with $\Xi^{\mathrm{M}} := \{\xi_1, \cdots, \xi_M\}$ and then set $\kappa_{\mathrm{N}} := \kappa_{\mathrm{M}}$ for $N \geq M$ if the sample is chosen in a proper way. Note that in this setting κ_{N} can be determined regardless of the value of N.

Example 3.1 reveals the underlying monotonicity of the Hoffman constant of discrete ambiguity set $\mathcal{P}_{\mathrm{N}}$ with respect to the change of N. This observation motivates an approachable scheme to calculate a predeterminable error bound constant κ_1 in (3.10). Note that this scheme is implementable without the Slater condition. The results in [25] have been improved by Theorem 3.2.

It is worth mentioning that (3.10) in Theorem 3.2 is not a standard error bound admitted by the moment system (3.1). Indeed, the residual function in (3.10) is β_{N} which cannot match the canonical perturbation to the moment system (3.1). We may next present the typical Hoffman's error bound of system (3.1) under Wasserstein metric, where the insight in Example 3.1 is still indispensable. Before we can do so, for simplicity, we rewrite system (3.1) as

$$\mathcal{P} := \big\{P \in \mathscr{P}(\Xi) : \langle P, \psi_{\mathrm{E}}(\xi)\rangle = \mu_{\mathrm{E}}, \langle P, \psi_{\mathrm{I}}(\xi)\rangle \leq \mu_{\mathrm{I}}\big\},$$

where μ_{E} and μ_{I} are scalars.

Theorem 3.3. *Suppose, (a) Ξ is compact, and (b) $\psi_{\mathrm{E}}(\cdot)$ and $\psi_{\mathrm{I}}(\cdot)$ are Lipschitz continuous on Ξ with bounded modulus L_ψ. Then,*

$$\mathsf{dl}_{\mathrm{W}}(P, \mathcal{P}) \leq \kappa_{\mathrm{W}}\big(|\langle P, \psi_{\mathrm{E}}(\cdot)\rangle - \mu_{\mathrm{E}}| + (\langle P, \psi_{\mathrm{I}}(\cdot)\rangle - \mu_{\mathrm{I}})_+\big), \quad \forall P \in \mathscr{P}(\Xi), \tag{3.16}$$

where $\kappa_{\mathrm{W}} = \mathrm{Diam}_\Xi \kappa_{\mathrm{M}}$, with Diam_Ξ denoting the diameter of Ξ and κ_{M}[†] *is the Hoffman constant of the discrete approximation $\mathcal{P}_{\mathrm{M}}$ which can be calculated by* (3.8).

Proof. In order to prove (3.16), we shall show that, there exists a constant κ_{W} such that, for any $\epsilon > 0$,

$$\mathsf{dl}_{\mathrm{W}}(P, \mathcal{P}) \leq \epsilon + \kappa_{\mathrm{W}}\big(|\langle P, \psi_{\mathrm{E}}(\cdot)\rangle - \mu_{\mathrm{E}}| + (\langle P, \psi_{\mathrm{I}}(\cdot)\rangle - \mu_{\mathrm{I}})_+\big), \quad \forall P \in \mathscr{P}(\Xi). \tag{3.17}$$

Let $\Xi^{\mathrm{N}} := \{\xi_1, \cdots, \xi_N\} \subseteq \Xi$. Throughout this proof, we set that $\Xi^{\mathrm{N'}} \subseteq \Xi^{\mathrm{N''}}$ for any $\mathrm{N}' \leq \mathrm{N}''$. Denote the Hoffman constant of linear system

$$\mathcal{P}_{\mathrm{N}} := \big\{P \in \mathscr{P}(\Xi^{\mathrm{N}}) : \langle P, \psi_{\mathrm{E}}(\xi)\rangle = \mu_{\mathrm{E}}, \langle P, \psi_{\mathrm{I}}(\xi)\rangle \leq \mu_{\mathrm{I}}\big\}$$

under 1-norm as κ_{N}, that is

$$d_1(P_{\mathrm{N}}, \mathcal{P}_{\mathrm{N}}) \leq \kappa_{\mathrm{N}}\big(\|\langle P_{\mathrm{N}}, \psi_{\mathrm{E}}(\xi)\rangle - \mu_{\mathrm{E}}\|_1 + \|(\langle P_{\mathrm{N}}, \psi_{\mathrm{I}}(\xi)\rangle - \mu_{\mathrm{I}})_+\|_1\big), \quad \forall P \in \mathscr{P}(\Xi^{\mathrm{N}}).$$

Then, by the analysis of Example 3.1, $\kappa_{\mathrm{N''}} \leq \kappa_{\mathrm{N'}}$ for any $\mathrm{N}' \leq \mathrm{N}''$.

[†] κ_{M} is a constant which is monotonically decreasing in M, that is, $\kappa_{\mathrm{M''}} \leq \kappa_{\mathrm{M'}}$ if $\Xi^{\mathrm{M'}} \subseteq \Xi^{\mathrm{M''}}$, see Example 3.1 for details.

For any given $\epsilon > 0$, there exists a sufficiently large N and the corresponding set $\Xi^{\rm N} := \{\xi_1, \cdots, \xi_N\} \subseteq \Xi$ such that $\beta_{\rm N} \le \epsilon/\kappa^*$, where $\beta_{\rm N}$ is defined as in (3.5),

$$\kappa^* = 2\mathrm{Diam}_\Xi \kappa_{\rm M} L_\psi + 1$$

and $\kappa_{\rm M}$ $(M \le N)$ is the Hoffman constant of linear system under 1-norm which is estimated in terms of (3.9). Recall that for any fixed $P \in \mathscr{P}(\Xi)$, $P^r_{\rm N}$ denotes Voronoi projection of P onto $\mathscr{P}(\Xi^{\rm N})$. Hence,

$$\begin{aligned}
&\mathsf{dl}_{\rm W}(P, \mathcal{P})\\
&\le \mathsf{dl}_{\rm W}(P, P^r_{\rm N}) + \mathsf{dl}_{\rm W}(P^r_{\rm N}, \mathcal{P})\\
&\le \mathsf{dl}_{\rm W}(P, P^r_{\rm N}) + \mathsf{dl}_{\rm W}(P^r_{\rm N}, \mathcal{P}_{\rm N})\\
&\le \mathsf{dl}_{\rm W}(P, P^r_{\rm N}) + \mathrm{Diam}_\Xi \kappa_{\rm M}\big(|\langle P^r_{\rm N}, \psi_{\rm E}(\cdot)\rangle - \mu_{\rm E}| + (\langle P^r_{\rm N}, \psi_{\rm I}(\cdot)\rangle - \mu_{\rm I})_+\big)\\
&\le \mathsf{dl}_{\rm W}(P, P^r_{\rm N}) + \mathrm{Diam}_\Xi \kappa_{\rm M}\big(|\langle P^r_{\rm N} - P, \psi_{\rm E}(\cdot)\rangle| + |\langle P, \psi_{\rm E}(\cdot)\rangle - \mu_{\rm E}|\\
&\qquad\qquad + |\langle P^r_{\rm N} - P, \psi_{\rm I}(\cdot)\rangle| + (\langle P, \psi_{\rm I}(\cdot)\rangle - \mu_{\rm I})_+\big)\\
&\le \frac{\epsilon}{\kappa^*} + \frac{2}{\kappa^*}\mathrm{Diam}_\Xi L_\psi \kappa_{\rm M}\epsilon + \mathrm{Diam}_\Xi \kappa_{\rm M}\big(|\langle P, \psi_{\rm E}(\cdot)\rangle - \mu_{\rm E}| + (\langle P, \psi_{\rm I}(\cdot)\rangle - \mu_{\rm I})_+\big)\\
&\le \epsilon + \mathrm{Diam}_\Xi \kappa_{\rm M}\big(|\langle P, \psi_{\rm E}(\cdot)\rangle - \mu_{\rm E}| + (\langle P, \psi_{\rm I}(\cdot)\rangle - \mu_{\rm I})_+\big),
\end{aligned}$$

where the first inequality follows from the triangle inequality of Wasserstein metric, the second inequality follows from the fact that $\mathcal{P}_{\rm N} \subseteq \mathcal{P}$, and the third inequality follows from the (3.14), (3.15) and the fact that $\kappa_{\rm N} \le \kappa_{\rm M}$ for any $\rm M \le N$, the last inequality follows from the definition of κ^*. The proof is then complete. □

Theorem 3.3 acts as the Hoffman's lemma in infinite dimension space. The modulus $\kappa_{\rm M}$ can be calculated first for any given sample $\Xi^{\rm M} := \{\xi_1, \cdots, \xi_M\} \subseteq \Xi$. In fact, the error bound modulus calculation can be further enhanced while larger M induces tighter estimation. As we commented at the beginning of this section, [6, Theorem 2.200] presents a similar result for a linear system in Banach Space. Consider the linear system

$$\Psi(y, b) := \big\{x \in X, Ax = y, \langle x^*, x\rangle \le b\big\},$$

where X and Y are Banach spaces and X^* is the dual space of X. [6, Theorem 2.200] shows that there exists a constant κ such that

$$dist\big(x, \Psi(y, b)\big) \le \kappa\big(\|Ax - y\| + [\langle x^*, x\rangle - b]_+\big).$$

Theorem 3.3 differs from [6, Theorem 2.200] on several aspects. First, under the Wasserstein metric, the probability space is actually a Polish space instead of a Banach space. Second, [6, Theorem 2.200], which is based on the open mapping Theorem, only states the existence of the Hoffman constant. However, Theorem 3.3 which takes full advantage of the monotonicity of Hoffman constant in finite dimensional space, presents an explicit estimation of the Hoffman constant. Thus Theorem 3.3 offers more appropriate a tool for quantitative stability analysis; see, e.g., Example 3.2. Together with Theorem 3.1, Theorem 3.3 significantly improves the results in [44, 52]. On the other hand, [25, Theorem 2] also presents a Hoffman's lemma with explicit modulus estimation equipped with a more general probability metric, i.e., ζ-structure metric. Unfortunately, their contribution is limited by strict assumption, i.e., the Slater constraint qualification.

Remark 3.2. As far as we know, Theorem 3.3 turns out to be the first Hoffman's lemma in infinite dimensional non-Banach space with explicit error bound modulus estimation. The simple technique in the proof of Theorem 3.3 may shed some light on studying Hoffman error bound admitted by distributionally robust ambiguity set under different probability metrics such as bounded Lipschitz metric, Fortet-Mourier metric.

We close this part with an illustrative example which emphasizes our improvement to [44, 52].

Example 3.2. Sun and Xu [44] consider the canonically perturbed ambiguity set (3.1) to $\mathcal{P}$ as

$$\mathcal{P}' := \left\{P \in \mathscr{P}(\Xi) : \mathbb{E}_P[\psi_{\text{E}}(\xi)] = \mu'_{\text{E}}, \mathbb{E}_P[\psi_{\text{I}}(\xi)] \leq \mu'_{\text{I}}\right\}.$$

In particular, [44] shows that there exists a constant κ such that

$$\mathbb{H}_{\text{T}}(\mathcal{P}', \mathcal{P}) \leq \kappa\big(\|\mu'_{\text{E}} - \mu_{\text{E}}\| + \|\mu'_{\text{I}} - \mu_{\text{I}}\|\big).$$

By Theorem 3.3, we arrive at the error bound

$$\mathbb{H}_{\text{W}}(\mathcal{P}', \mathcal{P}) \leq \kappa_{\text{W}}\big(\|\mu'_{\text{E}} - \mu_{\text{E}}\| + \|\mu'_{\text{I}} - \mu_{\text{I}}\|\big), \tag{3.18}$$

where κ_{W} admits an explicit estimation while [44] only concerns the existence of κ. Moreover, we may consider the following perturbed system:

$$\mathcal{P}'_{\text{N}} := \left\{P \in \mathscr{P}(\Xi^{\text{N}}) : \mathbb{E}_P[\psi_{\text{E}}(\xi)] = \mu'_{\text{E}}, \mathbb{E}_P[\psi_{\text{I}}(\xi)] \leq \mu'_{\text{I}}\right\}.$$

By Theorems 3.2 and 3.3, it is easy to see that

$$\mathbb{H}_{\text{W}}(\mathcal{P}'_{\text{N}}, \mathcal{P}) \leq \max\{\kappa_{\text{W}}, \kappa_1\}\big(\|\mu'_{\text{E}} - \mu_{\text{E}}\| + \|\mu'_{\text{I}} - \mu_{\text{I}}\| + \beta_{\text{N}}\big). \tag{3.19}$$

Unfortunately, [44] fails to characterize the proximity of $\mathcal{P}'_{\text{N}}$ to $\mathcal{P}$ as the total variation metric between a discrete probability distribution and a continuous probability distribution is identically 1. Leveraging (3.18) and (3.19) opens a new door to quantifying the stability of DRO problem; details are discussed in the next section.

3.2. Distance type ambiguity set

Another popular characterization of ambiguity set is through certain distance defined in probability space, such as Kullback-Leibler divergence [21, 22], Wasserstein metric [15, 30, 31, 53]. In particular, the ambiguity set contains a collection of distributions that are sufficiently close to a given nominal distribution with respect to given metric. Moreover, the distance type ambiguity sets offer powerful out-of-sample performance guarantees and enjoy convergence properties [15].

In this part, we consider the case that the ambiguity set is defined as a Wasserstein ball

$$\mathcal{P}_{\mathrm{W}} := \big\{Q \in \mathscr{P}(\Xi) : \mathsf{dl}_{\mathrm{W}}(Q, P_0) \leq c\big\}, \tag{3.20}$$

where P_0 is a nominal probability distribution and c is a small positive number representing the robustness of the ambiguity set. Of course, with the growth of $c, \mathcal{P}_{\mathrm{W}}$ becomes larger and hence admits a higher probability to contain the true distribution. When the nominal distribution is in the form of empirical distributions, the parameter c can be chosen appropriately by statistical methods. Suppose that the empirical distribution, denoted by P_N, is constructed through collections of historical data. If there exists a constant $\alpha > 1$ such that $\mathbb{E}_P[\exp(\|\xi\|^\alpha)] < \infty$, Esfahani and Kuhn [15] have shown that for a general k-dimension (e.g., $k > 2$) supporting space Ξ,

$$P\big(\mathsf{dl}_{\mathrm{W}}(P, P_N) \leq \theta\big) \geq 1 - C\left(\exp\big(-cN\theta^k\big)\mathbb{1}_{\{\theta\leq 1\}} + \exp\big(-cN\theta^\alpha\big)\mathbb{1}_{\{\theta>1\}}\right), \tag{3.21}$$

where N is the number of historical data, and C and c are positive constant numbers. The Eq. (3.21) provides finite sample guarantee property as well as asymptotic guarantee property. This nice feature allows us to adjust the radius θ of the Wasserstein ball such that the ambiguity set contains the true distribution P with a given probability threshold. Moreover, (3.21) implies that the ambiguity set converges to the true distribution P as the sample size N goes to infinity. See [53] for similar statistical evidence.

Under certain structural conditions, the popular tractable conic reformulation approach works for DRO problems with ambiguity set $\mathcal{P}_{\mathrm{W}}$; see, e.g., [15, 53]. Apart from the dual trackable scheme, recently, Shafieezadeh-Abadeh *et al.* [41] and Gao *et al.* [16] establish an equivalence between DRO problems with ambiguity set $\mathcal{P}_{\mathrm{W}}$ and certain regularized reformulations. This equivalence paves a new way to investigate solving DRO numerically when the regularization admits explicit expression. In particular, when the objective function follows linear structure, which gains popularity in statistics, efficient algorithms can be designed accordingly.

In this present paper, we focus on the discrete approximation method which to some extent offers a complementarity for the dual tractable approach and equivalent regularization scheme. In fact, the discrete method has been discussed in generative adversarial networks [1]. Consider the discrete approximation of $\mathcal{P}_{\mathrm{W}}$:

$$\mathcal{P}_{\mathrm{W}}^{\mathrm{N}} := \big\{Q \in \mathscr{P}(\Xi^{\mathrm{N}}) : \mathsf{dl}_{\mathrm{W}}(Q, P_0) \leq c\big\}, \tag{3.22}$$

where $\Xi^{\mathrm{N}} := \{\hat{\xi}_1, \cdots, \hat{\xi}_N\}$. The following theorem estimates the proximity from $\mathcal{P}_{\mathrm{W}}^{\mathrm{N}}$ to $\mathcal{P}_{\mathrm{W}}$.

Theorem 3.4. *Suppose Ξ is compact. Then, for any N,*

$$\mathbb{H}_W(\mathcal{P}_W^{\mathrm{N}}, \mathcal{P}_W) \leq 2\beta_{\mathrm{N}}, \tag{3.23}$$

where β_{N} is defined in (3.5).

Proof. By the definitions of $\mathcal{P}_{\rm W}^{\rm N}$ and $\mathcal{P}_{\rm W}$, $\mathcal{P}_{\rm W}^{\rm N} \subseteq \mathcal{P}_{\rm W}$ since $\Xi^{\rm N} \subset \Xi$. In order to prove (3.23), it is sufficient to show that the deviation $\mathbb{D}_{\rm W}(\mathcal{P}_{\rm W}, \mathcal{P}_{\rm W}^{\rm N}) \leq 2\beta_{\rm N}$. For any $P \in \mathcal{P}_{\rm W}$, $P_{\rm N}^r$ denotes Voronoi projection of P. Then by the triangle inequality of Wasserstein metric,

$$\mathsf{dl}_{\rm W}(P_{\rm N}^r, P_0) \leq \mathsf{dl}_{\rm W}(P_{\rm N}^r, P) + \mathsf{dl}_{\rm W}(P, P_0) \leq \beta_{\rm N} + c.$$

Denote $\lambda = \frac{\beta_{\rm N}}{\beta_{\rm N}+c}$ and $P_\lambda := (1-\lambda)P_{\rm N}^r + \lambda P_0$. By the convexity inherently obsessed by Wasserstein metric [30, Lemma 2.10],

$$\begin{aligned}
&\mathsf{dl}_{\rm W}(P_{\rm N}^r, P_\lambda) \leq (1-\lambda)\mathsf{dl}_{\rm W}(P_{\rm N}^r, P_{\rm N}^r) + \lambda \mathsf{dl}_{\rm W}(P_{\rm N}^r, P_0) \leq \lambda(\beta_{\rm N} + c) = \beta_{\rm N},\\
&\mathsf{dl}_{\rm W}(P_\lambda, P_0) \leq (1-\lambda)\mathsf{dl}_{\rm W}(P_{\rm N}^r, P_0) + \lambda \mathsf{dl}_{\rm W}(P_0, P_0) \leq (1-\lambda)(\beta_{\rm N} + c) = c.
\end{aligned}$$

Subsequently, $P_\lambda \in \mathcal{P}_{\rm W}^{\rm N}$ and

$$\mathsf{dl}_{\rm W}(P, \mathcal{P}_{\rm W}^{\rm N}) \leq \mathsf{dl}_{\rm W}(P, P_{\rm N}^r) + \mathsf{dl}_{\rm W}(P_{\rm N}^r, P_\lambda) \leq 2\beta_{\rm N}.$$

The proof is complete. □

As we may observe, the convexity admitted by the Wassertein metric plays a key role in the proof of Theorem 3.4. In particular, the convexity ensures the connectedness of the space $\mathscr{P}(\Xi)$ and hence the existence of P_λ. In fact, for other distance type ambiguity sets, the result in Theorem 3.4 remains valid once the associated distance admits the convex property. Thanks to Theorem 3.4, the discrete approximation scheme for DRO problems with distance type ambiguity set can be therefore quantified.

3.3. Mixture of moment conditions and Wasserstein metric

A friendly ambiguity set should be informative enough to include the true distribution and meanwhile precise enough to exclude pathological distributions. Although the moment type ambiguity is usually the first option, unfortunately, as aforementioned, the moment type ambiguity set does not enjoy a convergence property. The distance type ambiguity set, on the other hand, enjoys the convergence property. However, it fails to characterize distributions with desired moment information. In order to meet the needs, it is then natural for us to consider a mixture of these two types ambiguity set. To this end, we shall define ambiguity set as

$$\mathcal{Q} = \left\{ P \in \mathscr{P}(\Xi) : \begin{array}{l} \mathbb{E}_P[\psi_{\rm E}(\xi)] = \mu_{\rm E}, \quad \mathbb{E}_P[\psi_{\rm I}(\xi)] \leq \mu_{\rm I}, \\ \mathsf{dl}_{\rm W}(P, P_0) \leq c \end{array} \right\}, \tag{3.24}$$

where $\mu_{\rm E}$ and $\mu_{\rm I}$ are defined in (3.1) and c is defined in (3.20). Ambiguity set $\mathcal{Q}$ characterizes both the moment information and distance information, thus pathological distributions should be ruled out efficiently.

Ambiguity set defined in $\mathcal{Q}$ (3.24) has been studied by Gao and Kleywegt [17]. In particular, [17] establishes the dual tractable reformulation of the DRO problem with ambiguity set $\mathcal{Q}$, i.e., the DRO problem can be reformulated as a semi-definite programming when the involved objective function is piecewise linear. Moreover, the numerical experiments in [17] indicate a superior out-of-sample performance. In this part, we mainly focus on the cases where the DRO problem with ambiguity set $\mathcal{Q}$ that can not be reformulated as a semi-definite programming. For example, the support set Ξ is bounded and the objective does not have the special structure. In particular, we consider the discrete approximation schemes for the DRO problem with ambiguity set $\mathcal{Q}$. For this purpose, we first define the discrete approximation of $\mathcal{Q}$ on $\mathscr{P}(\Xi^{\text{N}})$

$$\mathcal{Q}_{\text{N}} = \left\{ P \in \mathscr{P}(\Xi^{\text{N}}) : \begin{array}{l} \mathbb{E}_P[\psi_{\text{E}}(\xi)] = \mu_{\text{E}}, \quad \mathbb{E}_P[\psi_{\text{I}}(\xi)] \le \mu_{\text{I}}, \\ \mathsf{dl}_{\text{W}}(P, P_0) \le c \end{array} \right\}. \tag{3.25}$$

In what follows, we shall quantify the converge of $\mathcal{Q}_{\text{N}}$ to $\mathcal{Q}$. Recall that

$$\mathcal{Q} = \mathcal{P} \cap \mathcal{P}_{\text{W}}, \quad \mathcal{Q}_{\text{N}} = \mathcal{P}_{\text{N}} \cap \mathcal{P}_{\text{W}}^{\text{N}},$$

where $\mathcal{P}, \mathcal{P}_{\text{N}}$ and $\mathcal{P}_{\text{W}}, \mathcal{P}_{\text{W}}^{\text{N}}$ are defined in Sections 3.1 and 3.2, respectively. The following theorem presents the quantitative convergence of the discrete approximation $\mathcal{Q}_{\text{N}}$ to $\mathcal{Q}$.

Theorem 3.5. *Suppose (a) the conditions of Theorems* 3.2 *and* 3.4*, (b) for any given N, $P_0 \in \mathcal{Q}_{\text{N}}$. Then, for any N,*

$$\mathbb{H}_{\text{W}}(\mathcal{Q}_{\text{N}}, \mathcal{Q}) \le \kappa_3 \beta_{\text{N}}, \tag{3.26}$$

where β_{N} is defined in (3.5) *and $\kappa_3 := (2\kappa_1 + 3)$ with κ_1 is defined as in* (3.11).

Proof. By the definitions of $\mathcal{Q}$ and $\mathcal{Q}_{\text{N}}$, $\mathcal{Q}_{\text{N}} \subset \mathcal{Q}$ in that $\Xi^{\text{N}} \subset \Xi$. It is sufficient to show (3.26) holds for the deviation $\mathbb{D}_{\text{W}}(\mathcal{Q}, \mathcal{Q}_{\text{N}})$. Denotes Voronoi projection of $P \in \mathcal{Q}$ as P_{N}^r. If $P_{\text{N}}^r \in \mathcal{Q}_{\text{N}}$, then

$$\mathsf{dl}_{\text{W}}(P, \mathcal{Q}_{\text{N}}) \le \mathsf{dl}_{\text{W}}(P, P_{\text{N}}^r) \le \beta_{\text{N}},$$

where the second inequality follows from (3.7). Thus, we are left with the case with $P_{\text{N}}^r \notin \mathcal{Q}_{\text{N}}$.

Denote Q_1 as the projection of P_{N}^r on $\mathcal{P}_{\text{N}}$ under the $\mathsf{dl}_{\text{W}}(\cdot)$. Then

$$\mathsf{dl}_{\text{W}}(P_{\text{N}}^r, Q_1) = \mathsf{dl}_{\text{W}}(P_{\text{N}}^r, \mathcal{P}_{\text{N}}) \le \kappa_1 \beta_{\text{N}},$$

where the second inequality follows from Theorem 3.2. If $Q_1 \in \mathcal{Q}_{\text{N}}$,

$$\mathsf{dl}_{\text{W}}(P, \mathcal{Q}_{\text{N}}) \le \mathsf{dl}_{\text{W}}(P, P_{\text{N}}^r) + \mathsf{dl}_{\text{W}}(P_{\text{N}}^r, Q_1) \le (\kappa_1 + 1)\beta_{\text{N}}.$$

If $Q_1 \notin \mathcal{Q}_{\text{N}}$,

$$c < \mathsf{dl}_{\text{W}}(Q_1, P_0) \le \mathsf{dl}_{\text{W}}(Q_1, P) + \mathsf{dl}_{\text{W}}(P, P_0) \le (\kappa_1 + 1)\beta_{\text{N}} + c,$$

where c is the parameter in (3.25).

Denote

$$\lambda = \frac{(\kappa_1 + 1)\beta_{\text{N}}}{(\kappa_1 + 1)\beta_{\text{N}} + c}$$

and $Q_\lambda := (1-\lambda)Q_1 + \lambda P_0$. Since Q_1 and P_0 are contained in $\mathcal{P}_{\rm N}$ and $\mathcal{P}_{\rm N}$ is a convex set, $Q_\lambda \in \mathcal{P}_{\rm N}$. Moreover, by the convex property of Wasserstein metric [30, Lemma 2.10],

$$\begin{aligned}\mathsf{dl}_{\rm W}(Q_1, Q_\lambda) &\le (1-\lambda)\mathsf{dl}_{\rm W}(Q_1, Q_1) + \lambda \mathsf{dl}_{\rm W}(Q_1, P_0)\\ &\le \lambda\big((\kappa_1+1)\beta_{\rm N} + c\big) = (\kappa_1+1)\beta_{\rm N},\\ \mathsf{dl}_{\rm W}(Q_\lambda, P_0) &\le (1-\lambda)\mathsf{dl}_{\rm W}(Q_1, P_0) + \lambda \mathsf{dl}_{\rm W}(P_0, P_0)\\ &\le (1-\lambda)\big((\kappa_1+1)\beta_{\rm N} + c\big) = c.\end{aligned}$$

Then $Q_\lambda \in \mathcal{Q}_{\rm N}$ and

$$\begin{aligned}\mathsf{dl}_{\rm W}(P, \mathcal{Q}_{\rm N}) &\le \mathsf{dl}_{\rm W}(P, P_{\rm N}^r) + \mathsf{dl}_{\rm W}(P_{\rm N}^r, Q_1) + \mathsf{dl}_{\rm W}(Q_1, Q_\lambda)\\ &\le (2\kappa_1 + 3)\beta_{\rm N}.\end{aligned}$$

The proof is complete. □

Compared to the moment type ambiguity set $\mathcal{P}$ and distance type ambiguity set $\mathcal{P}_{\rm W}$, $\mathcal{Q}$ is more informative by nature. The mixture of moment information and probability distance helps to exclude pathological distributions efficiently, unfortunately, it prevents the application of the tractable conic reformulation scheme. The validity of the dual tractable approach usually requires sophisticated problem structures. The discrete approximation method, on the other hand, works under less restrictive conditions. Therefore, it seems for solving DRO problems with ambiguity set $\mathcal{Q}$, the discrete approximation method serves as a reliable numerical scheme. In order to quantify the proximity of the approximation method through sample size in theory, Theorem 3.5 presents a quantitative description for the difference between two ambiguity sets $\mathcal{Q}_{\rm N}$ and $\mathcal{Q}$ in terms of Wasserstein metric.

Remark 3.3. The proof of Theorem 3.5 depends on two facts:

(a) the error bound conditions on $\mathcal{P}_{\rm N}$ and $\mathcal{P}_{\rm W}^{\rm N}$ share the same residual $\beta_{\rm N}$,

(b) the convexity of the Wasserstein metric [30, Lemma 2.10].

According to error bound theory, in general, the intersection of finitely many sets cannot guarantee the existence of a local error bound even if each single set admits a error bound. Usually, certain regularity condition, for instance, the bounded linear regularity condition[‡] should be imposed. In fact, if $\mathcal{Q}_{\rm N}$ ($\mathcal{Q}_{\rm N} = \mathcal{P}_{\rm N} \cap \mathcal{P}_{\rm W}^{\rm N}$) satisfies the bounded linear regularity conditions, we may arrive at (3.10) through Theorems 3.2 and 3.4 directly. However, it is difficult to verify the bounded linear regularity condition even under the case that $\mathcal{P}_{\rm N}$ is a polyhedral set and $P_0 \in \mathcal{P}_{\rm N} \cap \mathcal{P}_{\rm W}^{\rm N}$. The underlying reason is that the Wasserstein metric is not a standard metric in Euclidean space. This feature prevents us from applying the established results in Banach space.

[‡] Let $C_1, \cdots, C_k$ be closed subsets with a non-empty intersection C. We say that the collection $C_1, \cdots, C_k$ is boundedly linearly regular if for every bounded subset B, there exists a constant κ such that

$$d(c, C) \le \kappa \max d(c, C_i), \quad \forall c \in B.$$

A sufficient condition of the bounded linear regularity is that $C_i, i = 1, m$ are polyhedral and $\cap_{i=1}^m C_i \cap_{i=1}^m {\rm ri} C_i \ne \emptyset$.

4. Quantitative stability analysis for DRO problems

With established quantification of the difference between $\mathcal{P}_{\rm N}$ and $\mathcal{P}$ in the preceding section, we are now able to investigate how this quantitative difference propagates in DRO problems. In particular, we trace the changes of optimal values, optimal solutions as the underlying probability measure varies under appropriate metrics. Indeed, the stability analysis technique has been intensively adopted for standard stochastic programs; see [12, 37, 38] and reference therein. We next employ the stability analysis technique from stochastic programming problems to DRO problems. For this purpose, we focus on a general form of DRO problem

$$\begin{array}{ll} \min\limits_{x\in X} & \sup\limits_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)] \\ \text{s.t.} & \sup\limits_{P\in\mathcal{P}} \mathbb{E}_P[g(x,\xi)] \le 0, \end{array} \tag{4.1}$$

and its discrete approximation

$$\begin{array}{ll} \min\limits_{x\in X} & \sup\limits_{P\in\mathcal{P}_{\rm N}} \mathbb{E}_P[f(x,\xi)] \\ \text{s.t.} & \sup\limits_{P\in\mathcal{P}_{\rm N}} \mathbb{E}_P[g(x,\xi)] \le 0, \end{array} \tag{4.2}$$

where X is a nonempty compact subset of $\mathbb{R}^n$, $f : \mathbb{R}^n \times \Xi \to \mathbb{R}$ and $g : \mathbb{R}^n \times \Xi \to \mathbb{R}^d$ are Lipschitz continuous functions, and for every $\xi \in \Xi$, $g(\cdot,\xi) : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable. We shall build the relationship between optimal values and optimal solutions to the approximation problem (4.2) and optimal values and optimal solutions to the true problem (4.1). Before we can present our main theory, we shall need the following assumptions.

Assumption 4.1. Assume that problem (4.1) satisfies the following error bound condition:

$$d(x,\mathcal{F}) \le \kappa_{\rm F} \left\| \left(\sup_{P\in\mathcal{P}} \mathbb{E}_P[g(x,\xi)] \right)_+ \right\|_1, \quad \forall x \in X, \tag{4.3}$$

where $\mathcal{F}$ denotes the feasible set to problem (4.1) and $\kappa_{\rm F} > 0$ is the error bound constant.

Note that (4.3) is an error bound condition for the system of inequalities in the constraints of problem (4.1). The error bound theory has been well studied, see survey papers [3, 29]. Under Assumption 4.1, we present the quantitative stability of the feasible sets of the problems (4.1) and (4.2). For simplicity, we consider the case $d = 1$, that is, $g(x,\xi)$ is a scalar function, Note that our result can be extended to vector function case easily.

Theorem 4.1. *Let X be a compact set. Denote $\mathcal{F}$ and $\mathcal{F}_{\rm N}$ as the feasible sets to problems* (4.1) *and* (4.2)*, respectively. Suppose $g(x,\cdot)$ is Lipschitz continuous in ξ with bounded Lipschitz modulus L^* for any $x \in X$. Then* *(i)* $\lim\limits_{N\to\infty} \mathcal{F}_{\mathrm{N}} = \mathcal{F}$,

(ii) if Assumption 4.1 *additionally holds, then we have*

$$\mathscr{H}\left(\mathcal{F}_{\mathrm{N}}, \mathcal{F}\right) \leq 2L^{*}\kappa_{\mathrm{F}}\kappa\beta_{\mathrm{N}},$$

where κ_{F} is defined in (4.1) *and $\kappa := \kappa_1$ for moment type ambiguity set* (3.1)*, κ_2 for distance type ambiguity set* (3.20) *or κ_3 for mixed type ambiguity set* (3.24)*.*

Proof. Part (i). For any $x \in \mathcal{F}$,

$$0 \geq \sup_{P\in\mathcal{P}} \mathbb{E}_P[g(x,\xi)] \geq \sup_{P\in\mathcal{P}_{\mathrm{N}}} \mathbb{E}_P[g(x,\xi)],$$

where the second inequality follows from the fact that $\mathcal{P}_{\mathrm{N}} \subseteq \mathcal{P}$. Then $\mathcal{F} \subset \mathcal{F}_{\mathrm{N}}$ and we only need to show $\limsup_{N\to\infty} \mathcal{F}_{\mathrm{N}} \subseteq \mathcal{F}$. Note that

$$\left| \sup_{P\in\mathcal{P}_{\mathrm{N}}} \langle P, g(x,\xi)\rangle - \sup_{P\in\mathcal{P}} \langle P, g(x,\xi)\rangle \right| \leq 2L^{*}\mathbb{H}_{\mathrm{W}}(\mathcal{P}_{\mathrm{N}}, \mathcal{P}), \tag{4.4}$$

where the inequality follows from the Lipschitz continuity of $g(x,\cdot)$. Let x^* be an accumulation point of the sequence $\{x_N\}$ with $x_N \in \mathcal{F}_{\mathrm{N}}$. It is easy to show that

$$\lim_{N\to\infty} \sup_{P\in\mathcal{P}_{\mathrm{N}}} \langle P, g(x_N,\xi)\rangle = \sup_{P\in\mathcal{P}} \langle P, g(x^*,\xi)\rangle \leq 0.$$

Then we have $\limsup_{N\to\infty} \mathcal{F}_{\mathrm{N}} \subseteq \mathcal{F}$.

Part (ii). We only have to estimate the deviation $\mathscr{D}(\mathcal{F}_{\mathrm{N}}, \mathcal{F})$ as $\mathcal{F} \subseteq \mathcal{F}_{\mathrm{N}}$. By Assumption 4.1, for any $x \in \mathcal{F}_{\mathrm{N}}$, we have

$$\begin{aligned}
d(x,\mathcal{F}) &\leq \kappa_{\mathrm{F}} \left| \left(\sup_{P\in\mathcal{P}} \mathbb{E}_P[g(x,\xi)] \right)_+ \right| \\
&= \kappa_{\mathrm{F}} \left(\left(\sup_{P\in\mathcal{P}} \mathbb{E}_P[g(x,\xi)] \right)_+ - \left(\sup_{P\in\mathcal{P}_{\mathrm{N}}} \mathbb{E}_P[g(x,\xi)] \right)_+ \right) \\
&\leq \kappa_{\mathrm{F}} \left(\sup_{P\in\mathcal{P}} \mathbb{E}_P[g(x,\xi)] - \sup_{P\in\mathcal{P}_{\mathrm{N}}} \mathbb{E}_P[g(x,\xi)] \right) \\
&\leq 2L^{*}\kappa_{\mathrm{F}}\mathbb{D}_{\mathrm{W}}(\mathcal{P}, \mathcal{P}_{\mathrm{N}}),
\end{aligned}$$

where the second inequality follows from the fact that $\sup_{P\in\mathcal{P}_{\mathrm{N}}} \mathbb{E}_P[g(x,\xi)] \leq 0$, and the last inequality follows from (4.4). Then, by Theorem 3.2, we have

$$\mathscr{H}(\mathcal{F}_{\mathrm{N}}, \mathcal{F}) \leq 2L^{*}\kappa_{\mathrm{F}}\kappa\beta_{\mathrm{N}},$$

where κ_{F} and κ are defined in (4.3) and (3.11), respectively. □

With the quantitative stability results of the feasible sets, we are ready to present the stability of the optimal values.

Theorem 4.2. *Assume the conditions of Theorem* 4.1. *Assume also that* $f(\cdot)$ *is Lipschitz with modulus* L_f. *Denote* $S_{\rm N}$ *and* S *as the sets of optimal solutions to problems* (4.2) *and* (4.1)*, respectively,* θ_N *and* θ *be the corresponding optimal values. Then the following stability properties hold:*

(i)

$$|\theta - \theta_N| \leq (1 + 2L^* \kappa_F) L_f \kappa \beta_{\rm N}, \tag{4.5}$$

where L_f *is the Lipschitz modulus of* $f(\cdot)$*,* L^* *is the Lipschitz modulus of* $g(x,\cdot)$ *for any* $x \in X$*,* κ_F*,* κ *and* $\beta_{\rm N}$ *are defined in* (4.1)*,* (3.11) *and* (3.5) *respectively,*

(ii) $\lim\sup_{N\to\infty} S_{\rm N} \subseteq S$,

(iii) *if in addition,* $\sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)]$ *satisfies the growth condition, that is, there exists a positive constant* γ *such that*

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)] - \theta \geq \gamma d(x, S), \quad x \in X,$$

then

$$\mathscr{D}(S_{\rm N}, S) \leq \frac{2}{\gamma}(1 + L^* \kappa_F) L_f \kappa \beta_{\rm N},$$

where $L_f, L^*, \kappa_F, \kappa$ *and* $\beta_{\rm N}$ *are presented in part (i).*

Proof. With the convergence of feasible solutions, it is easy to study the stability of the optimal values and optimal solutions. We sketch the proof for completeness.

Part (i). Let $x_1 \in S_{\rm N}$ and $x_2 \in S$. Denote the projections of x_1 and x_2 on the sets $\mathcal{F}$ and $\mathcal{F}_{\rm N}$ by x_1^* and x_2^*, respectively. If $\theta_N \geq \theta$, then we have

$$\begin{aligned}
|\theta - \theta_N| &= \sup_{P\in\mathcal{P}_{\rm N}} \mathbb{E}_P[f(x_1,\xi)] - \sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x_2,\xi)] \\
&\leq \sup_{P\in\mathcal{P}_{\rm N}} \mathbb{E}_P[f(x_2^*,\xi)] - \sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x_2,\xi)] \\
&\leq \left| \sup_{P\in\mathcal{P}_{\rm N}} \mathbb{E}_P[f(x_2^*,\xi)] - \sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x_2^*,\xi)] \right| \\
&\quad + \left| \sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x_2^*,\xi)] - \sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x_2,\xi)] \right| \\
&\leq L_f \kappa \beta_{\rm N} + 2L_f L^* \kappa_{\rm F} \kappa \beta_{\rm N} \\
&= (1 + 2L^* \kappa_{\rm F}) L_f \kappa \beta_{\rm N},
\end{aligned}$$

where the first inequality follows from the fact that $x_2^* \in \mathcal{F}_{\rm N}$, the first item of the third inequality follows from the definition of $\mathsf{dl}_{\rm W}(\cdot)$ and the Lipschitz continuity of $f(x,\xi)$ with bounded modulus L_f, the second item of the third inequality follows from part (ii) of Theorem 4.1 and the Lipschitz continuity of $f(x,\xi)$ with bounded modulus L_f. On the other hand, if $\theta \geq \theta_N$, then we have

$$|\theta - \theta_N| = \sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x_2,\xi)] - \sup_{P\in\mathcal{P}_{\rm N}} \mathbb{E}_P[f(x_1,\xi)]$$

$$
\begin{aligned}
&\le \sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x_1^*,\xi)] - \sup_{P\in\mathcal{P}_{\rm N}} \mathbb{E}_P[f(x_1,\xi)]\\
&\le \Big|\sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x_1^*,\xi)] - \sup_{P\in\mathcal{P}_{\rm N}} \mathbb{E}_P[f(x_1^*,\xi)]\Big|\\
&\quad + \Big|\sup_{P\in\mathcal{P}_{\rm N}} \mathbb{E}_P[f(x_1^*,\xi)] - \sup_{P\in\mathcal{P}_{\rm N}} \mathbb{E}_P[f(x_1,\xi)]\Big|\\
&= (1+2L^*\kappa_{\rm F})L_f\kappa\beta_{\rm N}.
\end{aligned}
$$

Summarizing the discussion above, we arrive at (4.5).

Part (ii). Let $\{x_N\}$ be a sequence of optimal solutions to the problem (4.2). Taking a subsequence if necessary, we may assume that $\lim_{N\to\infty} x_N = x^*$. By Theorem 4.1, $x^* \in \mathcal{F}$. Moreover, note that $f(x_N) \le \theta$ as $\mathcal{F} \subseteq \mathcal{F}_{\rm N}$, then $f(x^*) = \theta$ which means $x^* \in S$. By the definition of outer limit of set, $\limsup_{N\to\infty} S_{\rm N} \subseteq S$.

Part (iii). Let $x_N \in S_{\rm N}$ be an optimal solution to the problem (4.2). Then we have

$$
\begin{aligned}
\gamma d(x_N,S) &\le \sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x_N,\xi)] - \theta\\
&\le \Big|\sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x_N,\xi)] - \sup_{P\in\mathcal{P}_{\rm N}} \mathbb{E}_P[f(x_N,\xi)]\Big|\\
&\quad + \Big|\sup_{P\in\mathcal{P}_{\rm N}} \mathbb{E}_P[f(x_N,\xi)] - \theta\Big|\\
&\le L_f\kappa\beta_{\rm N} + (1+2L^*\kappa_{\rm F})L_f\kappa\beta_{\rm N}\\
&= 2(1+L^*\kappa_{\rm F})L_f\kappa\beta_{\rm N},
\end{aligned}
$$

where the first inequality follows from the growth condition, and the third inequality follows from the Lipschitz continuity of $f(x,\xi)$ with bounded modulus L_f and part (i). Subsequently, we have

$$
d(x_N,S) \le \frac{2}{\gamma}(1+L^*\kappa_{\rm F})L_f\kappa\beta_{\rm N}.
$$

The rest follows from the arbitrariness of $x_N \in S_{\rm N}$. The proof is complete. □

We close this section with two illustrative examples; one is the standard one stage DRO problem while the other is the stochastic problem with the distributionally robust stochastic second order dominance constraints.

Example 4.1. Consider the DRO problem:

$$
\min_{x\in X}\max_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)],
$$

where X is a nonempty compact subset of $\mathbb{R}^n$, $f:\mathbb{R}^n\times\Xi\to\mathbb{R}$ is Lipschitz continuous functions, and $\mathcal{P}$ is defined as in Section 3. By introducing an auxiliary variable t, we may reformulate the DRO problem above as

$$
\begin{aligned}
\min_{x\in X,t\in\mathbb{R}} \quad & t\\
\text{s.t.} \quad & \max_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)] - t \le 0.
\end{aligned}
$$

Under moderate conditions, for example, the Lipschitz continuity of $f(x,\xi)$ on compact set $X\times\Xi$, the Slater constraint qualification holds. According to [50], Assumption 4.1 automatically holds, and Theorem 4.2 can be applied to quantify the error proximity from discrete approximation problems to the original problem in terms of optimal values and optimal solutions.

Example 4.2. Dentcheva and Ruszczyński [13] propose the following stochastic problem with the distributionally robust stochastic second order dominance constraints[§]:

$$\begin{array}{ll} \min\limits_{x\in X} & \sup\limits_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)] \\ \text{s.t} & \mathbb{E}_P[(t-G(x,\xi))_+] \le \mathbb{E}_P[(t-Y(\xi))_+], \quad \forall t\in T_0, \quad P\in\mathcal{P}, \end{array} \tag{4.6}$$

where X is a nonempty compact subset of $\mathbb{R}^n$, $f:\mathbb{R}^n\times\Xi\to\mathbb{R}$ and $G:\mathbb{R}^n\times\Xi\to\mathbb{R}$ are Lipschitz continuous functions, and for every $\xi\in\Xi$, $G(\cdot,\xi):\mathbb{R}^n\to\mathbb{R}$ is concave and continuously differentiable, $Y(\xi)$ is a random variable which is usually taken as the benchmark. If problem (4.6) satisfies the uniform robust dominance condition (see [13, Definition 3] for the details), it is easy to show that Assumption 4.1 holds.[¶] Then the requirements of Theorem 4.2 are met, and hence the discrete approximation approach can be employed to design algorithms.

5. Numerical implementation

In this section, we focus on the numerical implementation of the discrete approximation scheme. We are motivated by a very recent paper [26] which appears to be the first work applying primal-dual type methods to solve the DRO problems. In [26], the suggested approach gains its strength from the fact that the DRO problem is approximated by a sequence of min-max subproblem in a finite Euclidean space, with the ambiguity set $\mathcal{P}$ replaced by a set of discrete distributions. For the development on numerical schemes for min-max subproblems, there is a vast set of literature. Following this thought, [26] suggests applying the discretization technique to approximate the DRO problem, and then consider the lifting technique to further reformulate the discretized DRO subproblem as a min-max problem with certain separable structure. In particular, [26] considers the following DRO problem which is a special case of (4.1)

$$\min_{x\in X}\max_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)]. \tag{5.1}$$

Denote $\Xi^{\mathrm{N}} := \{\hat{\xi}_1,\cdots,\hat{\xi}_N\}$ and restrict the ambiguity set $\mathcal{P}$ to $\mathscr{P}(\Xi^{\mathrm{N}})$, that is, $\mathcal{P}_{\mathrm{N}} := \mathcal{P}\cap\mathscr{P}(\Xi^{\mathrm{N}})$. By introducing an auxiliary variable t, the discrete approximation of (5.1) reads as

$$\min_{(x,t)\in S}\max_{P\in\mathcal{P}_{\mathrm{N}}} \langle P,t\rangle, \tag{5.2}$$

[§] The constraints in problem (4.6) is a relaxation of distributionally robust stochastic second order dominance constraints as the index set T is a subset of $\mathbb{R}$.

[¶] The norm $\|\cdot\|_1$ in Assumption 4.1 should be replaced by the norm $\|\cdot\|_\infty$ with respect to t.

where

$$S := \left\{ (x,t) : \begin{cases} x \in X, & |t_i| \le t_{\max}, \\ f(x, \hat{\xi}_i) \le t_i, & i = 1, \ldots, N \end{cases} \right\}$$

with $t_{\max} := \max_{x\in X, \xi\in\Xi} |f(x,\xi)|$. Then, [26] implements the primal-dual hybrid gradient (PDGH) proposed in [7] to the reformulated min-max subproblem (5.2). That is, starting with some given initial point $(x_0, t_0; P_0)$, PDHG generates a sequence $(x_k, t_k; P_k)$ via the scheme

$$\begin{cases} (x_{k+1}, t_{k+1}) = \operatorname{argmin}_{(x,t)\in S} \left\{ \frac{1}{2\tau} \|t - (t_k - \tau P_k)\|^2 \right\}, \\ \hat{P}_{k+1} = P_k + \sigma(2t_{k+1} - t_k), \\ P_{k+1} = \operatorname{argmin}_{P\in\mathcal{P}_{\text{N}}} \left\{ \frac{1}{2\sigma} \|P - \hat{P}_{k+1}\|^2 \right\}, \end{cases}$$

where $\tau > 0, \sigma > 0$ and $\sigma\tau < 1$.

However, in actual applications of DRO, the discrete approximation likely requires more than thousands of samples $\mathcal{P}_{\text{N}}$ for the discrete approximation to perform properly. When the sample size N increases large, this PDHG scheme for solving min-max subproblems becomes conceptual, and not really a “true” algorithm, in the sense that it suffers from (at least) two main drawbacks.

- First, each step k requires exact minimization of two large-scale convex optimization problems in updating (x_{k+1}, t_{k+1}) and P_{k+1}. Some parallel computation techniques can be adopted for updating (x_{k+1}, t_{k+1}), see, e.g., [9, Algorithm 1]. However, the updating for P_{k+1} cannot be implemented in parallel as the components of variable P is by nature not completely separable. Thus exact minimization of at least one large-scale convex optimization problem is inevitable, see, e.g., step 6 of [9, Algorithm 1]. Consequently, the total computation load in each iteration k cannot be truly reduced to a friendly level by the parallel technique considered in [9].

- Secondly, it is a nested scheme which implies two nontrivial issues: (i) accumulations of computational errors in each step k, and (ii) how and when to stop each step k before passing to the next iteration $k+1$.

5.1. Separable reformulation and stochastic primal-dual type method

Seeking to address the above issues, in this work, instead of solving the discretized min-max subproblem directly, we reformulated the approximation subproblem into saddle point problem form with the help of Fenchel conjugate. As well known, a very popular algorithm to solve standard saddle point problems is the PDHG. Its popularity stems from two facts: First, it is easy to implement. Second, it involves only simple operations like matrix-vector multiplications and evaluations of proximal operations which are usually in closed form. However, for large problems even these simple operations might be still too expensive to perform too often. The reformulated saddle point subproblems of our interest are usually large as the sample size increases. Under some problem data assumptions, we shall observe an interesting fact that such saddle point problems always admit completely separable structures, see, e.g., Examples 5.1-5.3. Thanks to the separable structures, these operations of standard PDHG can be implemented in a parallel fashion. In this paper, we may adopt a stochastic extension of the PDHG for separable saddle point problem where not all but only a few of these operations are performed in each iteration.

We next illustrate how to reformulate the discrete approximation problem (5.2) into completely separable saddle point problems by three types of examples. We shall need the structural problem data assumption.

Assumption 5.1. For each fixed ξ, the objective function $f(x,\xi)$ is piecewise linear.

Example 5.1. Consider the discrete approximation of DRO with moment type ambiguity set:

$$\min_{x\in X}\sup_{P\in\mathcal{P}_K}\mathbb{E}_P[f(x,\xi)], \tag{5.3}$$

where

$$\mathcal{P}_K := \left\{P\in\mathscr{P}(\Xi^{\rm K}) : \mathbb{E}_P[\psi_{\rm E}(\xi)]=0, \mathbb{E}_P[\psi_{\rm I}(\xi)]\le 0\right\},$$

and $\Xi^{\rm K}=\{\hat{\xi}_1,\cdots,\hat{\xi}_K\}$. Reformulating the inner max problem of (5.3) by the Lagrange dual method, we arrive at the following minimization problem:

$$\begin{aligned} &\inf_{x\in X,\,\lambda_0,\lambda_1,\cdots,\lambda_p,\lambda_{p+1}\ge 0,\cdots,\lambda_q\ge 0} && \lambda_0\\ &\text{s.t.} && f(x,\hat{\xi}_j)-\sum_{i=1}^{q}\lambda_i\psi_i(\hat{\xi}_j)\le\lambda_0,\\ & && j=1,\ldots,K. \end{aligned} \tag{5.4}$$

If $f(x,\xi)$ satisfies Assumption 5.1, i.e.

$$f(x,\xi) := \max_{t=1,\ldots,T_0}\left(\xi^T A_t x + b_t\right),$$

problem (5.4) can be further rewritten as:

$$\begin{aligned} &\inf_{x\in X,\,\lambda_0,\lambda_1,\cdots,\lambda_p,\lambda_{p+1}\ge 0,\cdots,\lambda_q\ge 0} && \lambda_0\\ &\text{s.t.} && \hat{\xi}_j^T A_t x + b_t-\sum_{i=1}^{q}\hat{\lambda}_i\psi_i(\hat{\xi}_j)\le\lambda_0,\\ & && j=1,\ldots,K,\quad t=1,\ldots,T_0. \end{aligned} \tag{5.5}$$

The number of constraints of problem (5.5) depends on the discrete set $\Xi^{\rm K}$, which can be huge when the sample size increases. This results in ill-conditioned coefficient matrix of the constraints in (5.5), and therefore causes difficulty for numerical schemes. It is then often advantageous to reformulate problem (5.5) as a saddle point problem by Fenchel conjugate, i.e.,

$$\inf_{x\in X,\,\lambda\in\Lambda}\max_{\mu\geq 0}\lambda_0+\sum_{j=1}^{K}\sum_{t=1}^{T_0}\mu_{j,t}\left(\hat{\xi}_j^T A_t x+b_t-\sum_{i=1}^{q}\lambda_i\psi_i(\hat{\xi}_j)-\lambda_0\right), \tag{5.6}$$

where

$$\lambda:=(\lambda_0,\cdots,\lambda_q)^T\in\mathbb{R}^{q+1},\quad \Lambda:=\left\{\lambda\in\mathbb{R}^{q+1}:\lambda_{p+1}\geq 0,\cdots,\lambda_q\geq 0\right\}$$

and

$$\mu_j\in\mathbb{R}^{T_0},\quad j=1,\ldots,K,\quad \mu:=\left(\mu_1^T,\cdots,\mu_K^T\right)^T\in\mathbb{R}^{KT_0}.$$

Denote

$$A_j=\begin{bmatrix}\hat{\xi}_j^T A_1 & -\psi_1(\hat{\xi}_j) & \cdots & -\psi_q(\hat{\xi}_j) & -1\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \hat{\xi}_j^T A_1 & -\psi_1(\hat{\xi}_j) & \cdots & -\psi_q(\hat{\xi}_j) & -1\end{bmatrix}\in\mathbb{R}^{T_0\times(n+q+1)},$$

$$B=\ [b_1,\cdots,b_{T_0}]^T\in\mathbb{R}^{T_0},\quad z=\left[x^T,\lambda_1,\cdots,\lambda_q,\lambda_0\right]^T\in\mathbb{R}^{(n+q+1)}.$$

Eq. (5.6) can be reformulated as

$$\inf_{x\in X,\,\lambda\in\Lambda}\max_{\mu\geq 0}\lambda_0+\sum_{j=1}^{K}\left(\mu_j^T A_j z+B^T\mu_j\right). \tag{5.7}$$

Example 5.2. Consider the discrete approximation of DRO with distance type ambiguity set

$$\min_{x\in X}\sup_{P\in\mathcal{P}_{\mathrm{W}}^K}\mathbb{E}_P[f(x,\xi)], \tag{5.8}$$

where

$$\mathcal{P}_{\mathrm{W}}^K:=\left\{Q\in\mathscr{P}(\Xi^{\mathrm{K}}):\mathsf{dl}_{\mathrm{W}}(P,P_0)\leq c\right\},$$

$\Xi^{\mathrm{K}}:=\{\hat{\xi}_1,\cdots,\hat{\xi}_K\}$ and the nominal distribution P_0 is an empirical probability distribution generated by sample $\Xi_N:=\{\bar{\xi}_1,\cdots,\bar{\xi}_N\}$. By employing Lagrange dual method, we may reformulated problem (5.8) as

$$\begin{aligned}\inf_{x\in X,\,\lambda_0\geq 0,\lambda_1,\cdots,\lambda_N}\quad & \lambda_0 c+\frac{\lambda_1+\cdots+\lambda_N}{N}\\ \text{s.t.}\quad & f(x,\hat{\xi}_j)-\lambda_0\|\hat{\xi}_j-\bar{\xi}_i\|\leq\lambda_i,\\ & j=1,\ldots,K,\quad i=1,\ldots,N.\end{aligned} \tag{5.9}$$

Consider the case that $f(x,\xi)$ is piecewise linear function,

$$f(x,\xi):=\max_{t=1,\ldots,T_0}\xi^T A_t x+b_t.$$

Problem (5.9) can be reformulate as

$$\begin{array}{rl} \inf\limits_{x\in X,\, \lambda_0\geq 0, \lambda_1,\cdots,\lambda_N} & \lambda_0 c + \dfrac{\lambda_1+\cdots+\lambda_N}{N} \\ \text{s.t.} & \hat{\xi}_j^T A_t x + b_t - \lambda_0 \|\hat{\xi}_j - \bar{\xi}_i\| \leq \lambda_i, \\ & j=1,\ldots,K, \quad t=1,\ldots,T_0, \quad i=1,\ldots,N. \end{array} \tag{5.10}$$

We next reformulate problem (5.10) as a saddle point problem by Fenchel conjugate, i.e.,

$$\begin{aligned} \inf_{x\in X,\, \lambda\in\Lambda} \max_{\mu\geq 0} \lambda_0 c &+ \frac{\lambda_1+\cdots+\lambda_N}{N} \\ &+ \sum_{j=1}^{K}\sum_{t=1}^{T_0}\sum_{i=1}^{N} \mu_{j,t,i}\left(\hat{\xi}_j^T A_t x + b_t - \lambda_0\|\hat{\xi}_j - \bar{\xi}_i\| - \lambda_i\right), \end{aligned} \tag{5.11}$$

where

$$\lambda = (\lambda_0,\lambda_1,\cdots,\lambda_N)^T \in \mathbb{R}^{1+N}, \quad \Lambda := \left\{\lambda\in\mathbb{R}^{1+N} : \lambda_0 \geq 0\right\},$$

and

$$\begin{aligned} &\mu_j = (\mu_{1,1},\mu_{1,2},\cdots,\mu_{T_0,N}) \in \mathbb{R}^{T_0 N}, \quad j=1,\ldots,K, \\ &\mu := \left(\mu_1^T,\cdots,\mu_K^T\right)^T \in \mathbb{R}^{T_0 K N}. \end{aligned}$$

Denote

$$\begin{aligned} A_j =& \begin{bmatrix} \hat{\xi}_j^T A_1 & -\|\hat{\xi}_j - \bar{\xi}_1\| & -1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\xi}_j^T A_1 & -\|\hat{\xi}_j - \bar{\xi}_N\| & 0 & \cdots & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\xi}_j^T A_{T_0} & -\|\hat{\xi}_j - \bar{\xi}_1\| & -1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\xi}_j^T A_{T_0} & -\|\hat{\xi}_j - \bar{\xi}_N\| & 0 & \cdots & -1 \end{bmatrix} \in \mathbb{R}^{T_0 N\times(n+1+N)}, \\ B =& \ [b_1,\cdots,b_1,\cdots,b_{T_0},\cdots,b_{T_0}]^T \in \mathbb{R}^{T_0 N}, \\ z =& \ \left[x^T,\lambda_0,\lambda_1,\cdots,\lambda_N\right]^T \in \mathbb{R}^{n+1+N}. \end{aligned}$$

We may rewrite the problem (5.11) into a compact form:

$$\inf_{x\in X,\, \lambda\in\Lambda} \max_{\mu\geq 0} \lambda_0 c + \frac{\lambda_1+\cdots+\lambda_N}{N} + \sum_{j=1}^{K}\left(\mu_j A_j z + B^T \mu_j\right). \tag{5.12}$$

Example 5.3. Consider the discrete approximation DRO with moment and distance mixture type ambiguity set

$$\min_{x\in X}\sup_{P\in\mathcal{Q}_{\mathrm{N}}}\mathbb{E}_P[f(x,\xi)], \tag{5.13}$$

where

$$\mathcal{Q}_{\mathrm{N}}=\left\{P\in\mathscr{P}(\Xi^{\mathrm{K}}):\begin{array}{ll}\mathbb{E}_P[\psi_{\mathrm{E}}(\xi)]=0, & \mathbb{E}_P[\psi_{\mathrm{I}}(\xi)]\le 0,\\ \mathsf{dl}_{\mathsf{W}}(P,P_0)\le c & \end{array}\right\},$$

$\Xi^{\mathrm{K}}:=\{\hat{\xi}_1,\cdots,\hat{\xi}_K\}$ and the nominal distribution P_0 is an empirical probability distribution generated by sample $\{\bar{\xi}_1,\cdots,\bar{\xi}_N\}$. Similar to the analysis in Examples 5.1 and 5.2, by taking the Lagrange dual of the inner maximum problem of (5.13), we arrive at

$$\begin{array}{cl}\inf\limits_{x\in X,\,\gamma\in\Gamma,\lambda\in\Lambda} & \lambda_0 c+\dfrac{\lambda_1+\cdots+\lambda_N}{N}\\ \text{s.t.} & f(x,\hat{\xi}_j)-\sum\limits_{\varsigma=1}^{q}\gamma_\varsigma\psi_\varsigma(\hat{\xi}_j)-\lambda_0\|\hat{\xi}_j-\bar{\xi}_i\|\le\lambda_i,\\ & j=1,\dots,K,\quad i=1,\dots,N,\end{array} \tag{5.14}$$

where $\gamma:=(\gamma_1,\cdots,\gamma_q)^T$, $\lambda:=(\lambda_0,\lambda_1,\cdots,\lambda_N)^T$ and

$$\Gamma=\big\{\gamma\in\mathbb{R}^q:\gamma_{p+1}\ge 0,\cdots,\gamma_q\ge 0\big\},\quad \Lambda:=\big\{\lambda\in\mathbb{R}^{N+1}:\lambda_0\ge 0\big\}.$$

Suppose that $f(x,\xi)$ is piecewise linear function, i.e.

$$f(x,\xi):=\max_{t=1,\dots,T_0}\left(\xi^T A_t x+b_t\right).$$

Problem (5.14) is further expressed as

$$\begin{array}{cl}\inf\limits_{x\in X,\,\gamma\in\Gamma,\lambda\in\Lambda} & \lambda_0 c+\dfrac{\lambda_1+\cdots+\lambda_N}{N}\\ \text{s.t.} & \hat{\xi}_j^T A_t x+b_t-\sum\limits_{\varsigma=1}^{q}\lambda_\varsigma\psi_\varsigma(\hat{\xi}_j)-\lambda_0\|\hat{\xi}_j-\bar{\xi}_i\|\le\lambda_i,\\ & j=1,\dots,K,\quad t=1,\dots,T_0,\quad i=1,\dots,N.\end{array} \tag{5.15}$$

In order to address the issue caused by the ill-conditioned coefficient matrix, we reformulate problem (5.14) as a saddle point problem, i.e.,

$$\begin{aligned}\inf_{x\in X,\,\gamma\in\Gamma,\lambda\in\Lambda}\max_{\mu\ge 0}\lambda_0 c+\frac{\lambda_1+\cdots+\lambda_N}{N}+\sum_{j=1}^{K}\sum_{t=1}^{T_0}\sum_{i=1}^{N}\mu_{j,t,i}\Bigg(&\hat{\xi}_j^T A_t x\\ &+b_t-\sum_{\varsigma=1}^{q}\lambda_\varsigma\psi_\varsigma(\hat{\xi}_j)-\lambda_0\|\hat{\xi}_j-\bar{\xi}_i\|-\lambda_i\Bigg),\end{aligned} \tag{5.16}$$

where

$$\begin{aligned}&\mu_j=(\mu_{1,1},\mu_{1,2},\cdots,\mu_{T_0,N})\in\mathbb{R}^{T_0N},\quad j=1,\dots,K,\\ &\mu:=\left(\mu_1^T,\cdots,\mu_K^T\right)^T\in\mathbb{R}^{T_0KN}.\end{aligned}$$

Denote

$$A_j = \begin{bmatrix} \hat{\xi}_j^T A_1 & -\psi_1(\hat{\xi}_j) & \cdots & -\psi_q(\hat{\xi}_j) & -\|\hat{\xi}_j - \bar{\xi}_1\| & -1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\xi}_j^T A_1 & -\psi_1(\hat{\xi}_j) & \cdots & -\psi_q(\hat{\xi}_j) & -\|\hat{\xi}_j - \bar{\xi}_N\| & 0 & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\xi}_j^T A_{T_0} & -\psi_1(\hat{\xi}_j) & \cdots & -\psi_q(\hat{\xi}_j) & -\|\hat{\xi}_j - \bar{\xi}_1\| & -1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\xi}_j^T A_{T_0} & -\psi_1(\hat{\xi}_j) & \cdots & -\psi_q(\hat{\xi}_j) & -\|\hat{\xi}_j - \bar{\xi}_N\| & 0 & 0 & -1 \end{bmatrix} \in \mathbb{R}^{T_0 N \times (n+q+1+N)},$$

$$B = [b_1, \cdots, b_1, \cdots, b_{T_0}, \cdots, b_{T_0}]^T \in \mathbb{R}^{T_0 N},$$

$$z = \left[x^T, \gamma_1, \cdots, \gamma_q, \lambda_0, \lambda_1, \cdots, \lambda_N\right]^T \in \mathbb{R}^{(n+q+1+N)}.$$

Eq. (5.16) can be reformulated as

$$\inf_{x \in X, \gamma \in \Gamma, \lambda \in \Lambda} \max_{\mu \geq 0} \lambda_0 c + \frac{\lambda_1 + \cdots + \lambda_N}{N} + \sum_{j=1}^{K} \left(\mu_j^T A_j z + B^T \mu_j\right). \tag{5.17}$$

Observation on structures of the reformulated subproblems (5.7), (5.12) and (5.17) inspires the following a class of convex-concave saddle point problems with separable structures in variable t:

$$\min_s \max_t \phi_1(s) + \sum_{m=1}^{M} \langle A_m s, t_m \rangle + \phi_2(t), \tag{5.18}$$

where A_m are bounded linear operators and $\phi_2(\cdot)$ are block separable, i.e.

$$\phi_2(t) := \sum_{m=1}^{M} \theta_m(t_m).$$

In general, the saddle point problem (5.18) is numerically trackable by the PDHG. Specifically, the proximal operators involved in the PDHG are simple and in closed-form for subproblems (5.7), (5.12) and (5.17) of our interest, see, e.g., Tables 1-3.

In Tables 1-3, given a closed set $C, Proj_C(a)$ denotes the projection of point a on C. Given a convex function g, proximal operator (or proximity/resolvent operator) is defined as

$$Prox_g(y^*) = \mathrm{argmin}_{y \in Y}\, g(y) + \frac{1}{2}\|y - y^*\|^2.$$

Tables 1-3 reveal that, when implementing the PDHG on (5.7), (5.12) and (5.17), variables μ whose dimension is determined by the sample size K, can be updated in parallel. This parallel operation may address the computation cost issue when the sample size K is large. In this present paper, we are more interested in the case where

Table 1: Closed form of proximal operators: moment type.

s	(x, λ)
t	μ
$\phi_1(s)$	$\lambda_0 + \delta_{X \times \Lambda}(x, \lambda)$
$\phi_2(t)$	$\sum\limits_{j=1}^{K} \left(B^T \mu_j\right)$
$\text{Prox}_{\phi_1}(s^*)$	$Proj_X(x^*) \times (\lambda_0^* - 1, \lambda_2^* \cdots, \lambda_p^*) \times \max\{(\lambda_{p+1}^*, \cdots, \lambda_q^*), 0\}$
$\text{Prox}_{\phi_2}(t^*)$	$\mu_j^* - B^T,\ j = 1, \ldots, K$

Table 2: Closed form of proximal operators: distance type.

s	(x, λ)
t	μ
$\phi_1(s)$	$\lambda_0 c + \frac{\lambda_1 + \cdots + \lambda_N}{N} + \delta_{X \times \Lambda}(x, \lambda)$
$\phi_2(t)$	$\sum\limits_{j=1}^{K} \left(B^T \mu_j\right)$
$\text{Prox}_{\phi_1}(s^*)$	$Proj_X(x^*) \times \max\{\lambda_0^* - c, 0\} \times (\lambda_1^* - \frac{1}{N}, \cdots, \lambda_N^* - \frac{1}{N})$
$\text{Prox}_{\phi_2}(t^*)$	$\mu_j^* - B^T,\ j = 1, \ldots, K$

Table 3: Closed form of proximal operators: mixture type.

s	(x, γ, λ)
t	μ
$\phi_1(s)$	$\lambda_0 c + \frac{\lambda_1 + \cdots + \lambda_N}{N} + \delta_{X \times \Gamma \times \Lambda}(x, \gamma, \lambda)$
$\phi_2(t)$	$\sum\limits_{j=1}^{K} \left(B^T \mu_j\right)$
$\text{Prox}_{\phi_1}(s^*)$	$Proj_X(x^*) \times (\gamma_1^*, \cdots, \gamma_p^*) \times \max\{(\gamma_{p+1}^*, \cdots, \gamma_q^*), 0\}$ $\times \max\{\lambda_0^* - c, 0\} \times (\lambda_1^* - \frac{1}{N}, \cdots, \lambda_N^* - \frac{1}{N})$
$\text{Prox}_{\phi_2}(t^*)$	$\mu_j^* - B^T,\ j = 1, \ldots, K$

the sample size K is huge. To this end, we adopt the PDHG algorithm in a stochastic setting proposed in [8] where in each iteration we update a random subset of the sampled variables. Denoting $A = (A_1; \cdots; A_M)$, the iteration scheme of stochastic PDHG for (5.18) reads in Algorithm 5.1. Chambolle *et al* [8] has studied the convergence and linear convergence of the stochastic PDHG (under some strongly convex condition) and shows that stochastic PDHG significantly outperform the PDHG variant on a variety of imaging tasks.

5.2. Portfolio optimization problem

In this section, we consider the DRO formulation of a portfolio optimization problem. The stochastic PDHG presented in Algorithm 5.1 to the discretized reformulation

Algorithm 5.1 Stochastic PDHG method for problem (5.18)

```
**Require**: $s_0$, $t_0$, $\epsilon > 0$, $\tau > 0$, $\sigma > 0$ and $\sigma\tau < 1$
**for** $k = 0, 1, \dots$ **do**
    $$\begin{cases} s^{k+1} = \mathrm{Prox}_{\phi_1}\left(s^k - \tau A^T \bar{t}^k\right), \\ \text{select subset } I^{k+1} \text{ of } \{1, \dots, M\}, \\ t^{k+1} = \begin{cases} \mathrm{Prox}_{\phi_2}\left(t_i^k - \sigma A_i s^{k+1}\right), & \text{if} \quad i \in I^{k+1}, \\ t_i^k, & \text{if} \quad i \notin I^{k+1}, \end{cases} \\ \bar{t}^{k+1} = t^{k+1} - \theta(t^{k+1} - t^k). \end{cases}$$
**end for**
```

of the DRO model is implemented. We are interested in maximizing the expected utility while minimizing the risk which is characterized by the conditional value at risk (CVaR for short)

$$\min_{x \in X} \max_{P \in \mathcal{P}} \mathbb{E}_P[-f(x,\xi)] + \mathrm{CVaR}_\alpha^P\big(-f(x,\xi)\big), \tag{5.19}$$

where

$$f(x,\xi) := r_1 x_1 + \cdots + r_k x_k,$$

and CVaR_α^P is short for conditional value-at-risk and 'α' is the confidence level of α under the distribution P. For simplicity, we assume no trading fee. By employing the reformulation of CVaR [34], problem (5.19) can be reformulated as

$$\min_{x \in X, \tau \in \mathbb{R}} \max_{P \in \mathcal{P}} \mathbb{E}_P\big[\max\{a_1 f(x,\xi) + b_1\tau, a_2 f(x,\xi) + b_2\tau\}\big], \tag{5.20}$$

where

$$a_1 = -1, \quad a_2 = -1 - \frac{\rho}{\alpha}, \quad b_1 = \rho b_2 = \rho(1 - \frac{1}{\alpha}),$$

see [15, 34] for details.

Obviously, when the ambiguity sets in (5.20) are constructed as introduced in Examples 5.1-5.3, (5.20) can be reformulated as completely separable saddle point problem (5.18). We then implement the stochastic PDHG to find an optimal portfolio selection. Particularly, in the numerical test, the three types of ambiguity sets are defined as follows.

$$\text{Moment type:} \quad \mathcal{P}_{\mathrm{M}} := \left\{ P \in \mathscr{P}(\Xi) : \begin{array}{l} -\epsilon \leq (\mathbb{E}_P[\xi] - \bar{\mu})_i \leq \epsilon, \quad i = 1, \dots, m, \\ \left\| \mathbb{E}_P\left[(\xi - \bar{\mu})(\xi - \bar{\mu})^T\right] - \bar{\Sigma} \right\|_* \leq \sigma \end{array} \right\},$$

where $\|A\|_* = \max |a_{ij}|$, $\bar{\mu}$ and $\bar{\Sigma}$ are sample mean and sample covariance. See [45] for the motivation and more discussion on $\mathcal{P}_M$.

$$\text{Distance type:} \quad \mathcal{P}_{\mathrm{D}} := \big\{ Q \in \mathscr{P}(\Xi) : \mathsf{dl}_{\mathsf{W}}(P, P_0) \leq c \big\},$$

where P_0 is the empirical distributions [15, 53].

$$\text{Mixture type:} \quad \mathcal{P}_J := \big\{Q \in \mathscr{P}(\Xi) : \mathcal{P}_D \cap \mathcal{P}_M\big\}.$$

The parameters ϵ, σ, c are fixed as $0.05, 0.0025, 0.01,$ respectively. Moreover, the maximal iteration limitation is set to 10^7.

We collect the following ten stocks: Aberdeen Asset Management plc, Admiral Group PLC, AMEC PLC, Anglo American PLC, Antofagasta PLC, AstraZeneca PLC, Aviva PLC, Babcock PL (`http://finance.google.com`) (from 13th Apr 2013 to 18th Nov 2013) with a total of 150 datasets. Similar to the work [11], to ensure that the sample is independent and it follows the same distribution, we use 50 days from the most recent history to assign the portfolio. We have carried out out-of-sample tests with a rolling window of 50 days: use the first 50 data to construct the ambiguity set $\mathcal{P}$ and calculate the optimal portfolio strategy for the 51-th day and then move on a rolling basis.

We then implement the stochastic PDHG for different sample size which varies from 200 to 20000. The optimal values are reported in Table 4. We increase the sample size constantly until an obvious convergence trend regarding the optimal values is observable. In particular, the objective values is actually monotonically increasing when the sample size is growing. When the sample size excesses 8000, the changes concerning the optimal values become relatively quite small. The discrete approximation scheme is regarded as convergent.

Table 4: Convergence with increasing sample size.

Ambiguity type	200	500	1000	2000	4000	6000	8000	10000	20000
Moment	0.0132	0.0167	0.0187	0.0191	0.0190	0.0192	0.0192	0.0194	0.0199
Distance	0.0063	0.0085	0.0100	0.0101	0.0102	0.0103	0.0104	0.0104	0.0106
Mixture	0.0063	0.0085	0.0100	0.0101	0.0103	0.0103	0.0104	0.0104	0.0106

Note that when the constraint X of problem (5.20) is polyhedral convex with explicit expression, the portfolio problem can be formulated as a linear programming problem, see, e.g., [15, Section 7]. Under this special circumstance, some linear programming solvers in Python can be called as benchmark methods to justify the efficiency of the stochastic PDHG algorithm. In particular, we shall compare the optimal values returned by the stochastic PDHG method and the linear programming solver in Python. Setting sample size as 10000, in Table 5, we report 5 groups of tests on moment type ambiguity sets. As observed, the stochastic PDHG method returns “nearly” optimal values, where relative error rate is smaller than 1%. The stochastic PDHG method usually needs more time to find the solution than the linear programming solver for the tests in Table 5. However, for problems with the distance type and mixture type ambiguity sets, the linear programming solver in Python reports ’ memory error’ when sample size is setting as 10000.

Table 6 summarizes daily returns generated by three types of portfolio problems with moment type, distance type and mixture type ambiguity sets, where “L”, “H” and

Table 5: Benchmark testing.

Optimal Value Method	test1	test2	test3	test4	test5
SPDHG	0.0335	0.0315	0.0314	0.0321	0.0273
LP Solver	0.0338	0.0317	0.0315	0.0323	0.0274
error rate	0.9%	0.6%	0.3%	0.6%	0.3%

Table 6: Daily return.

Our DRO Model	L	H	A	Down	Up
Moment	0.9855	1.0202	1.0006	44	56
Distance	0.9853	1.0198	1.0006	43	57
Mixture	0.9853	1.0198	1.0006	44	56
EW	0.9815	1.0240	0.9999	54	46

"A" denote respectively the lowest, the highest and average returns and 'EW' stands for equally weighted strategy. We record the number of days when the overall portfolio return falls below 1 and exceeds (or equals to) 1, and denote them respectively by "Down" and "Up". We can see that the distributionally robust optimization with different type of ambiguity sets achieves comparable average daily return and displays more stable performance with a narrower range between the best and the worst return curves. The Fig. 1 indicates that the wealth curves generated by the DRO with three different ambiguity sets have the same tangency in going up or going down. They

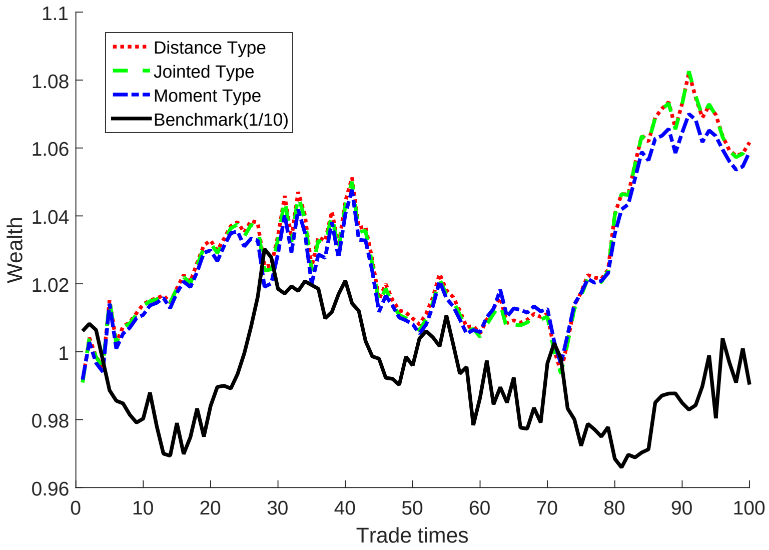

Figure 1: Sensitivity-Wealth evolution with the trading times.

outperform the benchmark wealth curve (equally weighted strategy) lying well above the black benchmark curve. As summarized in Table 6, the decisions returned by DRO model make a profit (going up) but the equally weighted strategy may experience loss (going down) in some trading days. At the end of the time horizon, the total wealth from our PRO models are around 1.06 compared to 0.99 in equally weighted strategy case. Another interesting observation on Fig. 1 is that the distance type ambiguity set and the mixture type ambiguity set have almost the same wealth curves. The underlying reason may be that the distance condition plays a key role in the definition of the mixture type ambiguity set.

Acknowledgements.

We would like to thank Shu Gong for doing the numerical tests. The research was supported by the NSFC (Grants 11971090, 11971220), the General Research Fund from Hong Kong Research Grants Council (Grant 12313516), the Fundamental Research Funds for the Central Universities (Grant DUT19LK24) and the Guangdong Basic and Applied Basic Research Foundation (Grant 2019A1515011152).

References

[1] M. Arjovsky, S. Chintala and L. Bottou, *Wasserstein Generative Adversarial Networks*, ICML, 2017.

[2] K. B. Athreya and S. N. Lahiri, *Measure Theory and Probability Theory*, Springer Science & Business Media, 2006.

[3] D. Azé, *A survey on error bounds for lower semicontinuous functions*, ESAIM: Proc., 13 (2003) 1–17.

[4] D. Bertsimas, X. V. Doan, K. Natarajan and C.-P. Teo, *Models for minimax stochastic linear optimization problems with risk aversion,* Math. Oper. Res., 35 (2010) 580–602.

[5] D. Bertsimas and I. Popescu, *On the relation between option and stock prices: an optimization approach*, Oper. Res., 50 (2002) 358–374.

[6] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Series in Operational Research, Springer, 2000.

[7] A. Chambolle, T. Pock, *A first-order primal-dual algorithms for convex problem with applications to imaging,* J. Math. Imaging Vis., 40 (2011) 120–145.

[8] A. Chambolle, M. J. Ehrhardt, P. Richtárik and C-B. Schönlieb, *Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications,* SIAM J. Optim., 28(4), 2783–2808.

[9] Y. Chen, H. Sun and H. Xu, *Decomposition methods for solving two-stage distributionally robust optimization problems,* Optimization-online, 2018.

[10] F. H. Clarke, *Optimization and Nonsmooth Analysis,* Wiley, 1983.

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

[12] D. Dentcheva and W. Römisch, *Stability and sensitivity of stochastic dominance constrained optimization models*, SIAM J. Optim., 23 (2013) 1672–1688.

[13] D. DENTCHEVA AND A. RUSZCZYŃSKI, *Robust stochastic dominance and its application to risk averse optimization*, Math. Prog., 123 (2010) 85–100.

[14] J. DUPAČOVÁ, *Uncertanities in minimax stochastic programs*, Optimization, 60 (2011) 1235–1250.

[15] P. M. ESFAHANI, D. KUHN, *Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations*, Math. Prog., 2017.

[16] R. GAO, X. CHEN AND A. J. KLEYWEGT, *Distributional robustness and regularization in statistical learning*, arXiv preprint, 2017.

[17] R. GAO AND A. J. KLEYWEGT, *Distributionally robust stochastic optimization with dependence structure*, arXiv:1701.04200v1, 2017.

[18] A. L. GIBBS AND F. E. SU, *On choosing and bounding probability metrics*, Int. stat. Eev., 70 (2002) 419–435.

[19] D. GOLDFARB AND G. IYENGAR, *Robust portfolio selection problems*, Math. Oper. Res., 28 (2003) 1–38.

[20] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Res. National Bureau Stand., 49 (1952) 263–265.

[21] Z. HU AND J. HONG, *Kullback-Leibler divergence constrained distributionally robust optimization*, Optimization-online, 2012.

[22] R. JIANG, Y. GUAN, *Data-driven chance constrained stochastic program*, Math. Prog., 158 (2016) 291–327.

[23] H. LEVY, *Stochastic dominance and expected utility: survey and analysis*, Manag. Sci., 38 (1992) 555–593.

[24] W. LI, *The sharp Lipschitz constants for feasible and optimal solutions of a perturbed linear program*, Lin. Algebra Appl., 187 (1993) 15–40.

[25] Y. LIU, A. PICHLER AND H. XU, *Discrete approximation and quantification in distributionally robust optimization*, Math. Oper. Res., Vol. 44(1), 2019.

[26] Y. LIU, X. M. YUAN, S. Z. ZENG AND J. ZHANG, *Primal-dual hybrid gradient method for distributionally robust optimization problems*, Oper. Res. Lett., 45 (2017) 625–630.

[27] S. MEHROTRA AND D. PAPP, *A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization*, SIAM J. Optim., 24 (2014) 1670–1697.

[28] A. MÜLLER AND M. SCARSINI, Eds., *Stochastic Orders and Decision Under Risk*, Institute of mathematical statistics, Hayward, CA, 1991.

[29] J.-S. PANG, *Error bounds in mathematical programming*, Math. Program., 79 (1997) 299–332.

[30] G. CH. PFLUG AND A. PICHLER, *Multistage Stochastic Optimization*, Series in Operations Research and Financial Engineering, Springer, 2014.

[31] G. CH. PFLUG AND D. WOZABAL, *Ambiguity in portfolio selection*, Quant. Finan., 7 (2007) 435–442.

[32] A. PICHLER AND H. XU, *Quantitative stability analysis for minimax distributionally robust risk optimization*, Optimization-online, 2017.

[33] S. T. RACHEV AND W. RÖMISCH, *Quantitative stability in stochastic programming: The method of probability metrics*, Math. Oper. Res., 27 (2002) 792–818.

[34] R. T. ROCKAFELLAR AND S. URYASEV, *Optimization of conditional value-at-risk*, J. Risk, Vol. 2 (2000) 21–42.

[35] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Variational Analysis*, Springer, 1998.

[36] W. RÖMISCH, *On discrete approximations in stochastic programming*, in: K. Lommatzsch (Ed.), Proceedings 13. Jahrestagung Mathematische Optimierung, Humboldt-Universität

Berlin, Sektion Mathematik. Seminarbericht Nr. 39, 166–175, 1981.

[37] W. Römisch, *Stability of stochastic programming problems*, in: Stochastic Programming, Handbooks in Operations Research and Management Science (A. Ruszczynski and A. Shapiro Eds.), Vol. 10, 483–554, Elsevier, 2003.

[38] W. Römisch and R. Schultz, *Distribution sensitivity in stochastic programming*, Math. Prog., 50 (1–3) (1991) 197–226.

[39] A. Ruszczyński and A. Shapiro, *Stochastic Programming, Handbook in Operations Research and Management Science*, Elsevier, 2003.

[40] H. Scarf, *A min-max solution of an inventory problem*, Studies in the Mathematical Theory of Inventory and Production (K. S. Arrow, S. Karlin, H. E. Scarf Eds.), Stanford University Press, 201–209, 1958.

[41] S. Shafieezadeh-Abadeh, D. Kuhn and P. M. Esfahani, *Regularization via mass transportation*, J. Mach Learn Res., 20 (2019) 1–68.

[42] A. Shapiro and S. Ahmed, *On a class of minimax stochastic programs*, SIAM J. Optim., 14 (2004) 1237–1249.

[43] A. H. Soubra and E. Bastidas-Arteaga, *Functions of random variables*, ALERT Doctoral School 2014 - Stochastic Analysis and Inverse Modelling, Michael A. Hicks, Cristina Jommi, 43-52, 2014.

[44] H. Sun and H. Xu, *Convergence analysis for distributional robust optimization and equilibrium problems*, Math. Oper. Res., 41 (2016) 377–401.

[45] R. H. Tütüncü and M. Koenig, *Robust asset allocation,* Annal. Oper. Res., 132 (2004) 157–187.

[46] W. Wiesemann, D. Kuhn and B. Rustem, *Robust resource allocations in temporal networks*, Math. Prog., 135 (2012) 437–471.

[47] W. Wiesemann, D. Kuhn and B. Rustem, *Robust Markov decision process*, Math. Oper. Res., 38 (2013) 153–183.

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

[49] H. Xu, Y. Liu and H. Sun, *Distributionally robust optimization with matrix moment constraints: Lagrange duality and cutting plane methods*, Math. Prog., 169 (2018) 489–529.

[50] J. J. Ye, *Constraint qualifications and necessary optimality conditions for optimization problems with variational inequality constraints*, SIAM J. Optim., 10 (2000) 943–962.

[51] J. Žáčková, *On minimax solution of stochastic linear programming problems*, Časopis pro Pěstování Matematiky, 91 (1966) 423–430.

[52] J. Zhang, H. Xu and L. W. Zhang, *Quantitative stability analysis for distributionally robust optimization with moment constraints*, SIAM J. Optim., 26 (2016) 1855–1882.

[53] C. Zhao, Y. Guan, *Data-driven risk-averse stochastic optimization with Wasserstein metric*, Oper. Res. Lett., 2018.

[54] A. Zhigljavsky and A. Žilinskas, *Stochastic Global Optimization*, Springer, 2008.

[55] S. Zymler, D. Kuhn and B. Rustem, *Distributionally robust joint chance constraints with second-order moment information*, Math. Prog., 137 (2013) 167–198.