Mathematical Programming
https://doi.org/10.1007/s10107-018-1347-4

FULL LENGTH PAPER

Series B

Quantitative stability analysis for minimax distributionally robust risk optimization

Alois Pichler[1] · **Huifu Xu**[2]

Received: 4 October 2016 / Accepted: 27 October 2018

Abstract
This paper considers distributionally robust formulations of a two stage stochastic programming problem with the objective of minimizing a distortion risk of the minimal cost incurred at the second stage. We carry out a stability analysis by looking into variations of the ambiguity set under the Wasserstein metric, decision spaces at both stages and the support set of the random variables. In the case when the risk measure is risk neutral, the stability result is presented with the variation of the ambiguity set being measured by generic metrics of ζ-structure, which provides a unified framework for quantitative stability analysis under various metrics including total variation metric and Kantorovich metric. When the ambiguity set is structured by a ζ-ball, we find that the Hausdorff distance between two ζ-balls is bounded by the distance of their centers and difference of their radii. The findings allow us to strengthen some recent convergence results on distributionally robust optimization where the center of the Wasserstein ball is constructed by the empirical probability distribution.

Keywords Distortion risk measure · ζ-ball · Wasserstein ball · Quantitative stability analysis

Mathematics Subject Classification 90C15 · 60B05 · 62P05

1 Introduction

One of the most important issues in optimization and operational research is how the underlying data in an optimization problem affect the optimal value and the optimal

✉ Alois Pichler
alois.pichler@math.tu-chemnitz.de

Huifu Xu
H.xu@soton.ac.uk

1 Fakultät für Mathematik, Technische Universität Chemnitz, Chemnitz, Germany

2 School of Mathematical Sciences, University of Southampton, Southampton, UK

Published online: 03 November 2018

decision. In stochastic programming, the underlying data are often concerned with a probability distribution of random variables because in many practical instances there is inadequate information about the true probability distribution. Over the past decade, effectively quantifying uncertainty and addressing the trade-off between using less information for approximating the true probability distribution such as samples and securing specified confidence of the resulting approximate optimal decision have been a challenging research topic in data-driven optimization problems, either because there is a limited number of available samples or it is more desirable to use fewer samples to increase the numerical tractability of the resulting optimization problem.

Research in this direction dates back to Žáčková [60] and Dupačová [12]. The recent monograph Pflug and Pichler [42] presents comprehensive discussions on approximations of probability distributions. An important technical issue which has been identified is to find an appropriate metric which can be effectively used to quantify the approximation of probability distributions. They conclude that the Wasserstein metric is most appropriate particularly in relation to (multistage) stochastic programming problems.

In an independent research on distributionally robust optimization, Esfahani and Kuhn [14] construct a ball in the space of (multivariate and non-discrete) probability distributions centered at the empirical distribution and look for decisions that perform best in view of the worst-case distribution within this Wasserstein ball. They demonstrate that, under mild assumptions, distributionally robust optimization problems over Wasserstein balls can in fact be reformulated as a finite convex program. In a number of practically interesting cases, the reformulations are even tractable linear programs, see Zhao and Guan [64] and Gao and Kleywegt [17] for further developments on this stream of research.

Rachev and Römisch [50] and Römisch [52, p. 487] establish the term ζ*-structure* in stochastic optimization for certain semi-norms, while Zhao and Guan [63] seem to be the first to use the ζ-metric to construct an ambiguity set in distributionally robust optimization (DRO). Specifically, they consider a ζ-ball centered at the distribution of independent, identically distributed (iid) samples of the true, unknown probability distribution. They establish a number of qualitative convergence results for the ζ-ball and related two stage optimization problems as the sample size increases and the radius of the ball shrinks. Moreover, they demonstrate that the resulting DRO can be easily solved by a dual formulation.

In this paper, we extend this important topic of research to a class of distributionally robust risk optimization (DRRO) problems. Specifically, we consider

$$\inf_{y\in Y} \sup_{P\in\mathcal{P}} \mathcal{R}_{S;P}\left(\inf_{z\in Z(y,\xi)} c(y,\xi,z)\right), \tag{DRRO}$$

where $c\colon \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}$ is a continuous function, $\xi\colon \Omega \to \Xi \subset \mathbb{R}^k$ is a vector of random variables defined on a measurable space $(\Omega, \mathcal{F})$ with state space Ξ, Y is a closed set in $\mathbb{R}^n$ and $Z\colon \Xi \times Y \rightrightarrows \mathbb{R}^m$ is a set-valued mapping, $\mathcal{P}$ is a set of probability measures and $\mathcal{R}_{S;P}$ is a risk measure parametrized by S and the probability measures $P \in \mathcal{P}$. The supremum is taken to immunize the risk arising

from ambiguity of the true probability distribution of ξ. The infimum with respect to z indicates that the robust risk minimization problem involves two stages of decision making processes: a choice of decision y in the first stage before realization of the uncertainty and an optimal choice of recourse action z from a feasible set $Z(y, \xi)$ in the second stage after observation of the uncertainty. Following the terminology in the literature, we call $\mathcal{P}$ *ambiguity set* and defer the specification of the parameter S to Sect. 4.

In the case when $\mathcal{R}_{S;P}(\cdot) = \mathbb{E}_P[\cdot]$ is the expectation, (DRRO) reduces to the ordinary minimax distributionally robust formulation of the two stage stochastic programming problem

$$\inf_{y\in Y} \sup_{P\in\mathcal{P}} \mathbb{E}_p \left[\inf_{z\in Z(y,\xi)} c(y, \xi, z) \right]. \tag{DRO}$$

A great deal of research in the literature of robust optimization to date has been devoted to developing tractable numerical methods for solving distributionally robust formulations of one stage stochastic optimization problems by reformulating the inner maximization problem into a semi-infinite programming problem through Lagrange dualization and further as a semi-definite programming problem via the $\mathcal{S}$-Lemma (cf. [47]) or dual methods, cf. Zymler et al. [66] or Wiesemann et al. [58]. This kind of approach requires the underlying functions in the objective and the ambiguity set to have some specific structure in terms of the variable ξ and the support set of ξ to have some polyhedral structure, see Wiesemann et al. [58] for a comprehensive discussion.

The approach can be extended to two-stage stochastic optimization problems after some appropriate approximation treatment of the second stage recourse problem through linear decision rules, k-adaptability or discretization. The decision rule approach restricts the second stage solution to a class of linear functions within the set of measurable functions in the feasible set whereas the k-adaptability approach confines the second stage feasible solutions to a set of k feasible solutions pre-determined before the realization of uncertainty. On the other hand, the discretization approach relaxes the constraints at the second stage to a finite number of scenarios. The former provides pessimistically biased solutions whereas the latter leads to optimistically biased solutions. We refer readers to Hanasusanto and Wiesemann [22,23], Hanasusanto and Kuhn [24] and references therein for this stream of research.

Another important approach pioneered by Pflug and Wozabal [44] is to discretize the ambiguity set of (DRO) and then to solve the discretized mini-max optimization problem directly as a saddle point problem in deterministic optimization. The discretization approach has received increasing attention over the past few years. For instance, Mehrotra and Papp [38] extend the approach to a general class of DRO problems and design a process which generates a *cutting surface* of the inner optimal value at each iterate. Xu et al. [59] observe that the discretization scheme is equivalent to discrete approximation of the semi-infinite constraints of the dualized inner maximization problem and apply the well known cutting plane method to solve the minimax optimization (cf. [32]). Under some moderate conditions, they show convergence of the optimal value of the discretized problem to its true counterpart as the discretization refines.

While the convergence result gives some qualitative guarantee for asymptotic consistency of the optimal value, it does not address a quantitative relationship between the sample size and the error of the optimal value. This paper aims to fill out the gap. The main contributions can be summarized as follows:

- We present a quantitative analysis for the ζ-ball by looking into how the ζ-ball evolves as its center shifts and radius changes. Under the ζ-metric, we show that the Hausdorff distance of two ζ-balls is linearly bounded by the distance of their centers and the difference of their radii, see Theorem 1 below.
- We consider the case when the ambiguity set $\mathcal{P}$ in (DRO) is constructed through a ζ-ball and investigate how variation of the ζ-ball would affect the optimal value and the optimal solution in the resulting optimization problem. Some quantitative stability results are derived under moderate conditions (see Theorem 3 below). The research provides a unified framework for the existing research on quantitative stability analysis of (DRO) under various metrics including the total variation metric and the Wasserstein metric.
- We present a detailed quantitative stability analysis for (DRRO) in terms of the optimal value and optimal solution when c is equi-Lipschitz continuous in y and z and equi-Hölder continuous in ξ (see Theorem 6). Differing from the stability results established for (DRO) under the ζ-metric, we use the Wasserstein metric due to complexity of the model arising from distortion risk measure. Some topological properties of the Wasserstein ball are also established (see Sect. 2.5).

Nomenclature Throughout the paper, we will use the following notation. For a metric space $(\mathbb{X}, d)$, we write $d(x, S)$ for the distance from a point x to a set S, $\mathbb{D}(S_1, S_1; d)$ for the *excess* of S_1 over S_2 associated with distance d, i.e.,

$$\mathbb{D}(S_1, S_2; d) = \sup_{x \in S_1} d(x, S_2) = \sup_{x \in S_1} \inf_{y \in S_2} d(x, y) \tag{1}$$

and $\mathbb{H}(S_1, S_2; d)$ for the *Hausdorff distance* between the two sets, that is,

$$\mathbb{H}(S_1, S_2; d) = \max \Big\{ \mathbb{D}(S_1, S_2; d),\ \mathbb{D}(S_2, S_1; d) \Big\}.$$

By convention, we use $\mathbb{R}^n$ to denote the n-dimensional Euclidean space and $\mathscr{P}(\Xi)$ to denote the space of probability measures over Ξ. Depending on the nature of the metric space, we will use different symbols for the metric. For instance, in a finite dimensional space $\mathbb{R}^n$, we use the ordinary letter d to denote the distance whereas $\mathsf{dl}_{\mathscr{G}}$, dl_K, d_r denote the ζ-metric, Kantorovich–Wasserstein metric and Wasserstein distance, respectively, in the space of probability measures $\mathscr{P}(\Xi)$.

Outline. The rest of the paper is organized as follows. In Sect. 2, we introduce the definition of ζ-balls and discuss changes of the ball as its center and radius vary. Particular focus is given to the Wasserstein ball. The discussion is needed to quantify the change of the ambiguity set in stability analysis of the DRO and DRRO models: Sects. 3–4 set out stability analysis for the DRO and DRRO models. Sect. 3 is focused on the DRO model under ζ-metric and Sect. 4 deals with the DRRO model under the Wasserstein metric.

2 Quantifying variation of ζ-ball and Wasserstein ball

Let Ω be a sample space and $\mathcal{F}$ be the associated sigma algebra. Let $\mathscr{P}(\Omega)$ be the set of all probability measures over the measurable space $(\Omega, \mathcal{F})$. We consider a vector valued measurable function ξ mapping from Ω to $\Xi \subset \mathbb{R}^k$. Let $\mathscr{B}(\Xi)$ be the Borel sigma algebra and $P \in \mathscr{P}(\Omega)$. For each set $A \in \mathscr{B}$, let $P^{\xi}(A) := P(\xi^{-1}(A))$. Consequently, we may focus on $\mathscr{P}(\Xi)$, the set of all probability measures defined on space $(\Xi, \mathscr{B})$ with support set contained in Ξ, where each element P^{ξ} is a probability measure on the space induced by ξ which is also known as push-forward, or image measure.

2.1 ζ-metric

In probability theory, various metrics have been introduced to quantify the distance/ difference between two probability measures; see Athreya and Lahiri [3], Gibbs and Su [18]. Here, we adopt the ζ-metric.

Definition 1 Let $P, Q \in \mathscr{P}(\Xi)$ and $\mathscr{G}$ be a family of real-valued measurable functions on Ξ. Define

$$\mathsf{dl}_{\mathscr{G}}(P, Q) := \sup_{g \in \mathscr{G}} \left| \mathbb{E}_P[g(\xi)] - \mathbb{E}_Q[g(\xi)] \right|. \tag{2}$$

The (semi-) distance defined as such is called a metric with ζ-structure and covers a wide range of metrics in probability theory, see Rachev [49] or Zolotarev [65]. For the simplicity of terminology, we call it ζ-metric throughout this paper.

It is well known that a number of important metrics in probability theory may be viewed as a special case of the ζ-metric. For instance, if we choose

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

then $\mathsf{dl}_{\mathscr{G}}(P, Q)$ reduces to the *total variation metric*, in which case we denote it specifically by dl_{TV}. If g is restricted further to be Lipschitz continuous with modulus bounded by 1, i.e.,

$$\mathscr{G} = \left\{ g \colon \sup_{\xi \in \Xi} |g(\xi)| \le 1, \ g \text{ is Lipschitz continuous and the Lipschitz modulus } L_1(g) \le 1 \right\}, \tag{3}$$

where $L_1(g) := \sup\{|g(u) - g(v)|/d(u, v) \colon u \neq v\}$, then the resulting metric is known as *bounded Lipschitz metric*, denoted by dl_{BL}. If the boundedness of g is lifted in (3), that is,

$$\mathscr{G} = \{ g \colon g \text{ is Lipschitz continuous and Lipschitz modulus } L_1(g) \le 1 \}, \tag{4}$$

then we obtain Kantorovich metric,[1] denoted by dl_K. If we relax the Lipschitz continuity in (4), that is,

$$\mathscr{G} = \big\{g\colon g \text{ is Lipschitz continuous and } L_q(g) \le 1\big\}$$

with

$$L_q(g) := \inf\Big\{L\colon |g(u)-g(v)| \le L\|u-v\| \max(1, \|u\|^{q-1}, \|v\|^{q-1}) \quad \forall \quad u, v \in \Xi\Big\},$$

where $\|\cdot\|$ denotes the Eucledian norm and $q \ge 1$, then we obtain *Fortet–Mourier metric*, denoted by dl_{FM}. Finally, if

$$\mathscr{G} = \Big\{g\colon g(\cdot) = \mathbb{1}_{(-\infty,t]}(\cdot),\ t \in \mathbb{R}^k\Big\},$$

where

$$\mathbb{1}_{(-\infty,t]}(\xi) := \begin{cases} 1 & \text{if} \quad \xi \in (-\infty, t], \\ 0 & \text{otherwise,} \end{cases}$$

then we obtain *uniform (Kolmogorov) metric*, denoted by dl_U.

Remark 1 It is evident that $\mathsf{dl}_{TV}(P, Q) \le 2$ and when Ξ is bounded, $\mathsf{dl}_K(P, Q) \in \big[0, \operatorname{diam}(\Xi)\big]$, see Gibbs and Su [18]. Moreover, it follows by Zhao and Guan [63, Lemmas 1–4], that $\mathsf{dl}_{BL}(P, Q) \le \max\{\mathsf{dl}_K(P, Q), \mathsf{dl}_{TV}(P, Q)\}$, $\mathsf{dl}_{FM}(P, Q) \le \max\{1, \operatorname{diam}(\Xi)^{q-1}\}\, \mathsf{dl}_K(P, Q)$ and $\mathsf{dl}_U(P, Q) \le \frac{1}{2}\mathsf{dl}_{TV}(P, Q)$.

2.2 Hömander's theorem

Based on the ζ-metric $\mathsf{dl}_{\mathscr{G}}$, we can define the distance from as single distribution to a set, the deviation from one set to another and the Hausdorff distance between two sets in the space of probability measures $\mathscr{P}(\Xi)$. We denote them respectively by $\mathsf{dl}_{\mathscr{G}}(Q, \mathcal{S})$, $\mathbb{D}(\mathcal{S}', \mathcal{S}; \mathsf{dl}_{\mathscr{G}})$ and $\mathbb{H}(\mathcal{S}', \mathcal{S}; \mathsf{dl}_{\mathscr{G}})$. It is easy to observe that $\mathbb{H}(\mathcal{S}', \mathcal{S}; \mathsf{dl}_{\mathscr{G}}) = 0$ if and only if $\mathbb{E}_P\, g(\xi) - \mathbb{E}_Q\, g(\xi) = 0$ for any $P \in \mathcal{S}'$, $Q \in \mathcal{S}$ and $g \in \mathscr{G}$.

In the theory of set-valued analysis there is a famous theorem, namely Hörmander's theorem, which establishes a relationship between the distance of two sets in Euclidean space and the maximum difference between their respective support functions over the unit ball of the same space, see Castaing and Valadier [6, Theorem II-18]. Here, we extend the theorem to the set of probability measures. One of the main reasons behind this extension is that in minimax distributionally robust optimization problems, the inner maximization of the worst expected value of a random function over an ambiguity set of probability distributions is indeed the support function of the random function over the ambiguity set. Therefore, in order to look into the difference between the worst

[1] In some references, it is called Wasserstein metric or Kantorovich–Wasserstein metric, see commentary by Villani [57]. Here we call it Kantorovich metric to distinguish it from Wasserstein metric to be defined later on.

expected values based on two ambiguity sets, it is adequate to assess the discrepancy between two support functions of the sets. We will come back to this in the next section. To this end, we need the concept of weak compactness of probability measures under the topology of weak convergence. Recall that a sequence of probability measures $\{P_N\} \subset \mathscr{P}(\Xi)$ is said to converge to $P \in \mathscr{P}(\Xi)$ *weakly*, if

$$\lim_{N\to\infty} \int_{\Xi} h(\xi) P_N(d\xi) = \int_{\Xi} h(\xi) P(d\xi)$$

for each bounded and continuous function $h\colon \Xi \to \mathbb{R}$. An important property of Kantorovich's metric is that it metrizes weak convergence of probability measures on measures with bounded r-th moment, that is, $\{P_N\}$ converges to P weakly if and only if $\mathsf{dl}_K(P_N, P) \to 0$ (cf. [18]).

For a set of probability measures $\mathcal{A}$ on $(\Xi, \mathscr{B})$, $\mathcal{A}$ is said to be *tight* if for any $\epsilon > 0$, there exists a compact set $\Xi_\epsilon \subset \Xi$ such that $\inf_{P\in\mathcal{A}} P(\Xi_\epsilon) > 1 - \epsilon$. In the case when $\mathcal{A}$ is a singleton, it reduces to the tightness of a single probability measure. $\mathcal{A}$ is said to be *closed* (under the weak topology) if for any sequence $\{P_N\} \subset \mathcal{A}$ with P_N converging to P weakly, $P \in \mathcal{A}$. $\mathcal{A}$ is said to be *weakly compact* if every sequence $\{P_N\} \subset \mathcal{A}$ contains a subsequence $\{P_{N'}\}$ and $P \in \mathcal{A}$ such that $P_{N'} \to P$ weakly; see Skorokhod [55] for the notion and Billingsley [5] for a similar notion called relative compactness. By the well-known Prokhorov's theorem (see [3]), a closed set $\mathcal{A}$ (under the weak topology) of probability measures is *compact* if it is tight. In particular, if Ξ is a compact set, then the set of all probability measures on $(\Xi, \mathscr{B})$ is compact; see Prokhorov [48, Theorem 1.12].

Proposition 1 (Cf. [30]) *Let $\mathcal{P}, \mathcal{Q} \subset \mathscr{P}(\Xi)$ be two sets of probability measures and $\mathscr{G}$ the set of all measurable functions from Ξ to $\mathbb{R}$. Suppose that $\mathcal{P}$ and $\mathcal{Q}$ are weakly compact. Then*

$$\mathbb{D}(\mathcal{P}, \mathcal{Q}; \mathsf{dl}_{\mathscr{G}}) = \sup_{h\in\mathscr{G}} \left\{ s_{\mathcal{P}}(h) - s_{\mathcal{Q}}(h) \right\}, \tag{5}$$

and

$$\mathbb{H}(\mathcal{P}, \mathcal{Q}; \mathsf{dl}_{\mathscr{G}}) = \sup_{g\in\mathscr{G}} \left| s_{\mathcal{P}}(g) - s_{\mathcal{Q}}(g) \right|, \tag{6}$$

where $s_{\mathcal{P}}(g) := \sup_{P\in\mathcal{P}} \int g \mathrm{d}P$ is a support function, $\mathbb{D}$, $\mathbb{H}$ are excess distance and Hausdorff distance associated with ζ-metric $\mathsf{dl}_{\mathscr{G}}$.

Proof By definition,

$$\mathbb{D}(\mathcal{P}, \mathcal{Q}; \mathsf{dl}_{\mathscr{G}}) = \sup_{P\in\mathcal{P}} \mathsf{dl}_{\mathscr{G}}(P, \mathcal{Q}) = \sup_{P\in\mathcal{P}} \inf_{Q\in\mathcal{Q}} \sup_{g\in\mathscr{G}} \left(\int g \mathrm{d}P - \int g \mathrm{d}Q \right). \tag{7}$$

Let $\phi(Q, g) := \int g \mathrm{d}P - \int g \mathrm{d}Q$. Then ϕ is linear in g and affine in Q. Thus it is a convex function of Q and a concave function of g. Moreover, since $\mathcal{Q}$ is a compact set in the metric space of probability measures $\mathscr{P}(\Xi)$ under the topology of weak convergence and the latter is a Hausdorff space, we may apply Fan's minimax theorem, Fan [15, Theorem 2] and obtain

$$\inf_{Q\in\mathcal{Q}} \sup_{g\in\mathcal{G}} \phi(G,g) = \sup_{g\in\mathcal{G}} \inf_{Q\in\mathcal{Q}} \phi(G,g).$$

Consequently, we have

$$\begin{aligned}\mathbb{D}(\mathcal{P},\mathcal{Q};\mathsf{dl}_{\mathcal{G}}) &= \sup_{P\in\mathcal{P}} \sup_{g\in\mathcal{G}} \inf_{Q\in\mathcal{Q}} \left(\int g\mathrm{d}P - \int g\mathrm{d}Q\right)\\ &= \sup_{P\in\mathcal{P}} \left(\sup_{g\in\mathcal{G}} \left(\int g\mathrm{d}P - \sup_{Q\in\mathcal{Q}} \int g\mathrm{d}Q\right)\right)\\ &= \sup_{g\in\mathcal{G}} \left(s_{\mathcal{P}}(g) - s_{\mathcal{Q}}(g)\right). \end{aligned} \tag{8}$$

This shows (5). Likewise, since $\mathcal{P}$ is weakly compact, we have

$$\mathbb{D}(\mathcal{Q},\mathcal{P};\mathsf{dl}_{\mathcal{G}}) = \sup_{g\in\mathcal{G}} \left(s_{\mathcal{Q}}(g) - s_{\mathcal{P}}(g)\right). \tag{9}$$

Combining (8) and (9) gives rise to (6). □

From the proposition we can see immediately that for any fixed measurable function g,

$$|s_{\mathcal{Q}}(g) - s_{\mathcal{P}}(g)| \le \mathbb{H}(\mathcal{Q},\mathcal{P};\mathsf{dl}_{\mathcal{G}}),$$

which means the difference between the maximum expected values from sets $\mathcal{Q}$ and $\mathcal{P}$ is bounded by the Hausdorff distance of the two sets under ζ-metric.

Note also that in order for us to apply Fan's minimax theorem in the proof of the proposition, we imposed weak compactness on the set $\mathcal{Q}$. In Sect. 2.5, we discuss weak compactness of the Wasserstein ball.

2.3 ζ-ball

Of particular interest is the set of probability measures defined with ball structure, that is, all probability measures within a ball centered at some probability measure with specified radius. In practice, the probability measure at the center is known as nominal distribution which may be approximated through empirical data or its smooth approximation (as kernel density approximation).

Definition 2 *(The ζ-ball)* Let $P \in \mathscr{P}(\Xi)$ and $\mathcal{G}$ be a family of real-valued bounded measurable functions on Ξ. Let r be a positive number. We call the following set of probability distributions ζ-ball:

$$\mathcal{B}(P,r) := \{P' \in \mathscr{P}(\Xi) \colon \mathsf{dl}_{\mathcal{G}}(P',P) \le r\}, \tag{10}$$

where $\mathsf{dl}_{\mathcal{G}}(\cdot,\cdot)$ is defined in (2).

In what follows, we quantify the change of the ζ-ball as its center and radius vary. To this end, we recall important properties of the ζ-distance $\mathsf{dl}_{\mathscr{G}}(P, Q)$ when Q varies over $\mathscr{P}(\Xi)$.

Proposition 2 (Convexity of the ζ-metric) *Let $P, Q_1, Q_2 \in \mathscr{P}(\Xi)$ be three probability measures and $\mathsf{dl}_{\mathscr{G}}(\cdot, \cdot)$ be defined as in (2). Then*

$$\mathsf{dl}_{\mathscr{G}}\big(P, t Q_1 + (1-t) Q_2\big) \leq t\, \mathsf{dl}_{\mathscr{G}}(P, Q_1) + (1-t)\, \mathsf{dl}_{\mathscr{G}}(P, Q_2), \quad \text{for all} \quad t \in [0, 1] \tag{11}$$

and

$$\mathsf{dl}_{\mathscr{G}}(P, Q_2) \leq \mathsf{dl}_{\mathscr{G}}(P, Q_1) + \mathsf{dl}_{\mathscr{G}}(Q_1, Q_2). \tag{12}$$

The result follows from the fact that ζ-metric is a semi-distance which satisfies all axioms of a metric except the property that $\zeta(P, Q) = 0$ if and only if $P = Q$. Equations (11) and (12) are no more than convexity and the triangle inequality of the metric which are retained by the semi-metric.

Corollary 1 *Let $P, Q_1, Q_2 \in \mathscr{P}(\Xi)$ be three probability measures and $\mathsf{dl}_{\mathscr{G}}(\cdot, \cdot)$ be the ζ-metric defined as in (2).*

For $t \in [0, 1]$, let

$$h(t) := \mathsf{dl}_{\mathscr{G}}\big(P, t Q_1 + (1-t) Q_2\big).$$

If $\max(\mathsf{dl}_{\mathscr{G}}(P, Q_1), \mathsf{dl}_{\mathscr{G}}(P, Q_2)) < \infty$, then $h(\cdot)$ is continuous on $[0, 1]$ and

$$h(t) \in \Big[0,\ \max\big(\mathsf{dl}_{\mathscr{G}}(P, Q_1),\ \mathsf{dl}_{\mathscr{G}}(P, Q_2)\big)\Big] \quad \forall \quad t \in [0, 1].$$

Proof Under the condition that $\max(\mathsf{dl}_{\mathscr{G}}(P, Q_1), \mathsf{dl}_{\mathscr{G}}(P, Q_2)) < \infty$, it follows from Proposition 2 that $h(\cdot)$ is a proper convex function. By Rockafellar [51, Corollary 10.1.1], $h(\cdot)$ is continuous over $[0, 1]$. The rest is straightforward. □

From the definition of ζ-ball and Proposition 2 we can see immediately that the ζ-ball is a convex set in the space of $\mathscr{P}(\Xi)$. However, the ball is not necessarily weakly compact. For example, if $\mathscr{G}$ is the set of all measurable functions bounded by 1, then the ζ-metric reduces to the total variation metric. The resulting ball centered at a discrete probability measure with radius smaller than 1 does not include any continuous probability measure.

In what follows, we study the quantitative stability of a ζ-ball against variation of its center and radius.

Theorem 1 (Quantitative stability of the ζ-ball) *Let $\mathcal{B}(P, r)$ be the ball defined as in (10). For every $P, Q \in \mathscr{P}(\Xi)$ and $r_1, r_2 \in \mathbb{R}_+$ it holds that*

$$\mathbb{H}\big(\mathcal{B}(P, r_1), \mathcal{B}(Q, r_2); \mathsf{dl}_{\mathcal{G}}\big) \leq \mathsf{dl}_{\mathcal{G}}(P, Q) + \big|r_2 - r_1\big|, \tag{13}$$

where $\mathbb{H}$ denotes the Hausdorff distance in $\mathscr{P}(\Xi)$ associated with the metric $\mathsf{dl}_{\mathcal{G}}$.

Proof By the definition of the Hausdorff distance, it suffices to show that

$$\mathbb{D}\big(\mathcal{B}(Q, r_2), \mathcal{B}(P, r_1); \mathsf{dl}_{\mathcal{G}}\big) \leq \mathsf{dl}_{\mathcal{G}}(P, Q) + |r_2 - r_1| \tag{14}$$

and

$$\mathbb{D}\big(\mathcal{B}(P, r_1), \mathcal{B}(Q, r_2); \mathsf{dl}_{\mathcal{G}}\big) \leq \mathsf{dl}_{\mathcal{G}}(P, Q) + |r_2 - r_1|. \tag{15}$$

We start by showing (14). The inequality holds trivially if $\mathcal{B}(Q, r_2) \subset \mathcal{B}(P, r_1)$. So we focus on the case when $\mathcal{B}(Q, r_2) \not\subset \mathcal{B}(P, r_1)$. Let $Q' \in \mathcal{B}(Q, r_2) \backslash \mathcal{B}(P, r_1)$ and set $\hat{\lambda} := r_1 / \mathsf{dl}_{\mathcal{G}}(Q', P)$. By the definition of the ball, $\hat{\lambda} \in (0, 1)$. Let $\hat{P} := \hat{\lambda} Q' + (1 - \hat{\lambda}) P$. By convexity of the distance $\mathsf{dl}_{\mathcal{G}}$,

$$\mathsf{dl}_{\mathcal{G}}(\hat{P}, P) = \mathsf{dl}_{\mathcal{G}}\big(\hat{\lambda} Q' + (1 - \hat{\lambda}) P, P\big) \leq \hat{\lambda}\, \mathsf{dl}_{\mathcal{G}}(Q', P) = r_1.$$

This shows $\hat{P} \in \mathcal{B}(P, r_1)$. Hence

$$\begin{aligned}
\mathsf{dl}_{\mathcal{G}}(Q', \mathcal{B}(P, r_1)) &\leq \mathsf{dl}_{\mathcal{G}}(Q', \hat{P}) = \mathsf{dl}_{\mathcal{G}}(Q', \hat{\lambda} Q' + (1 - \hat{\lambda}) P) \\
&\leq (1 - \hat{\lambda}) \mathsf{dl}_{\mathcal{G}}(Q', P) = \mathsf{dl}_{\mathcal{G}}(Q', P) - \hat{\lambda}\, \mathsf{dl}_{\mathcal{G}}(Q', P) \\
&= \mathsf{dl}_{\mathcal{G}}(Q', P) - r_1 \\
&\leq \big(\mathsf{dl}_{\mathcal{G}}(Q', Q) + \mathsf{dl}_{\mathcal{G}}(Q, P)\big) - r_1 \\
&\leq r_2 + \mathsf{dl}_{\mathscr{G}}(Q, P) - r_1.
\end{aligned} \tag{16}$$

This shows (14). By swapping the role of the two balls in the proof above, we obtain the formula for (15). □

The significance of Theorem 1 is that it gives a quantitative description about the Hausdorff distance of two ζ-balls. The result allows one to easily quantify the difference between a ζ-ball and its variation incurred by a perturbation of its center and/or radius. The error bound (13) is tight in the sense that the equality holds under some special cases, i.e., (i) $r_1 = 0$ and $P = Q$, and (ii) P and Q are Dirac probability measures and $\mathsf{dl}_{\mathscr{G}} = \mathsf{dl}_K$, as in that case the Kantorovich metric dl_K coincides with the Euclidean distance.

A particularly interesting case is that when $r_1 = r_2$, the error bound in (13) depends only on the distance between the centers of the balls which means any other probability measures in the balls have no impact on the bound.

Moreover, let $r_1 = 0$ and P the unknown true probability distribution while Q is an empirical distribution constructed through samples. When the sample size increases and the radius shrinks, the ζ-ball converges to the true probability distribution.

2.4 The empirical measure

For the empirical measure[2]

$$P_N(\cdot) := \frac{1}{N}\sum_{k=1}^{N}\delta_{\xi k}(\cdot) \tag{17}$$

with iid samples $(\xi_k)_{k=1}^N$, Theorem 1 reads

$$\mathsf{dl}_{\mathscr{G}}(P, \mathcal{B}(P_N, r_N)) \le \mathsf{dl}_{\mathscr{G}}(P, P_N) + r_N, \tag{18}$$

where $\mathsf{dl}_{\mathscr{G}}$ is defined as in (2).

In the literature of probability theory, there are many results concerning convergence of P_N to P. First, P_N converges to P if and only if $\mathsf{dl}_{\mathscr{G}}(P, P_N) \to 0$ under the bounded Lipschitz metric, Kantorovich metric and Fortet–Mourier metric. In particular, if there exists a positive number $\nu > 0$ such that

$$M := \int_{\Xi} \exp(\|\xi\|^{\nu}) P(d\xi) < \infty,$$

then, for any $\epsilon > 0$, there exist positive constants c and C such that

$$P^N\left(\mathsf{dl}_K(P, P_N) \ge \epsilon\right) \le C\left[\exp(-cN\epsilon^{\max(k,2)}\mathbb{1}_{\epsilon\le 1}) + \exp(-cN\epsilon^{\nu}\mathbb{1}_{\epsilon>1})\right] \tag{19}$$

for all N, where P^N is the probability measure over space $\Xi \times \cdots \times \Xi$ (N times) with Borel-sigma algebra $\mathcal{B}\otimes\cdots\otimes\mathcal{B}$, and k ($k \ne 2$) is the dimension of ξ, C, c are positive constants which depend on ν, M and k; in the case when $k = 2$, the inequality is slightly more complicated, see Fournier and Guilline [16, Theorem 2] for details.

In the case when P is a continuous probability distribution it is well-known that $\mathsf{dl}_{TV}(P, P_N) = 1$. In that case, we may use Kernel density estimation (KDE) of P_N, denoted by $\tilde{P}_N$, to replace P_N. Specifically, let h_N be a sequence of positive constants converging to zero and $\Phi(\cdot)$ be a measurable kernel function with $\Phi(\cdot) \ge 0$, $\int \Phi(\xi)d\xi = 1$. We define the KDE of P_N as

$$f_N(z) = \frac{1}{Nh_N^k}\sum_{i=1}^{N}\Phi\left(\frac{z-\xi_i}{h_N}\right), \tag{20}$$

which is the density function of the measure $\tilde{P}_N$. A simple example for $\Phi(\cdot)$ is the density function of the standard normal distribution. Under some moderate conditions, [63] established bounds for $\mathsf{dl}_{\mathscr{G}}(P, \tilde{P}_N)$ under a range of metrics with ζ-structure including dl_{TV}, dl_K, dl_{FM}, dl_{BL} and dl_U, see Zhao and Guan [63, Proposition 4].

[2] The Dirac-measure is defined by $\delta_{\xi}(A) = \mathbb{1}_A(\xi) = \begin{cases} 1 & \text{if} \quad \xi \in A, \\ 0 & \text{if} \quad \xi \notin A. \end{cases}$

Using the corollary above and the proposition, we can easily derive the rate of convergence for $\mathsf{dl}_{\mathscr{G}}\big(P, \mathcal{B}(P_N, r_N)\big)$ as N increases and r_N decreases. Note that KDE is widely used in stochastic programming, we refer readers to Shapiro et al. [54, Chapter 4], Gröwe [20] and Norkin and Keyzer [39] for more comprehensive and in depth discussions on the approach.

Note also that inequality (18) may be extended to the case when the samples are not necessarily independent. Indeed, one may use Quasi-Monte Carlo method or even a deterministic approach for developing an approximation of P, see Pflug and Pichler [41] and references therein.

2.5 Wasserstein ball

One of the most important metrics with ζ-structure is the Kantorovich metric. At this point, it might be helpful to introduce the definition of Wasserstein distance/metric and relate it to the Kantorovich metric. To this end we endow the set Ξ with a metric d and consider the Polish space (Ξ, d) in the rest of discussions.

Definition 3 *(Wasserstein distance/metric)* For probability measures P and $\tilde{P}$, the Wasserstein distance/metric of order $r \geq 1$ is

$$\mathsf{d}_r(P, \tilde{P}) = \left(\inf_{\pi} \iint d\left(\xi, \tilde{\xi}\right)^r \pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}) \right)^{1/r}, \tag{21}$$

where π is among all probability measure with marginals P and $\tilde{P}$, i.e.,

$$\begin{aligned} P(A) &= \pi(A \times \Xi), \quad A \in \mathscr{B}(\Xi) \text{ and} \\ \tilde{P}(B) &= \pi(\Xi \times B), \quad B \in \mathscr{B}(\Xi). \end{aligned} \tag{22}$$

We remind readers that the distance $\mathsf{d}_r(P, \tilde{P})$ should be distinguished from the metrics of ζ-structure discussed in the preceding subsections where we used the notation $\mathsf{dl}_{\mathscr{G}}$, dl_K and dl_{TV} etc.

One of the main results concerning the Wasserstein distance is the Kantorovich–Rubinstein Theorem [31], which establishes a relationship between the Kantorovich metric of two probability measures and the Wasserstein distance when $r = 1$, i.e.,

$$\mathsf{d}_1(P, \tilde{P}) = \mathsf{dl}_K(P, \tilde{P}), \tag{23}$$

where $\mathsf{dl}_K(P, \tilde{P})$ is defined as in Definition 1. The identity (23) recovers the metric d_1 as metric with ζ-structure, but the general Wasserstein metric with order $r > 1$ does not have the structure; indeed, the Wasserstein metric d_r is r-convex rather than convex as ζ-metric (see Shapiro et al. [54, Definition 4.7] or Pflug and Pichler [42, Lemma 2.10]).

The Wasserstein metric is a very well established concept in applied probability theory for quantifying the distance between two probability distributions. A simple and intuitive explanation of the metric is that if we regard $d(\xi, \xi')$ as a cost of moving

masses from ξ to ξ', then the metric may be interpreted as the minimal transportation cost of moving masses placed over a set of locations (represented by one probability distribution) to another set of locations (represented by another probability distribution), which is known as Kantorovich's formulation of Monge's mass transference problem (cf. [49]). The concept has found wide applications in applied probability (e.g., Gini index of dissimilarity of two random variables), partial differential equations, functional inequalities or Riemannian geometry and image processing, see commentary by Villani [57].

Here we establish a technical result which ensures weak compactness of the set of probability measures defined through the Wasserstein metric for further reference later on.

Proposition 3 *Let $\mathscr{P}(\mathbb{R}^m)$ denote the set of all probability measures on $\mathbb{R}^m$, let $d(\cdot,:) = \mathsf{d}_r(\cdot,:)$ be the Wasserstein distance of order $r \geq 1$ and let $\mathcal{P} \subseteq \mathscr{P}(\mathbb{R}^m)$ be a subset of probability measures. If $\mathcal{P}$ is tight, then the ρ-enlargement under the Wasserstein metric $\mathcal{P}_\rho := \{Q \in \mathscr{P}(\mathbb{R}^m) \colon \mathsf{d}_r(Q, P) \leq \rho \text{ for some } P \in \mathcal{P}\}$ is also tight for every $\rho \geq 0$ and $\mathcal{P}_\rho$ is weakly compact.*

Proof Let $\varepsilon \in \left(0, \frac{1}{2}\right)$ be fixed. Since $\mathcal{P}$ is tight, there is a compact set $K^\varepsilon \subset \mathrm{R}^m$ such that $P(K^\varepsilon) > 1 - \varepsilon$ for every $P \in \mathcal{P}$. This means

$$P\left(K^{\varepsilon \mathsf{c}}\right) = 1 - P\left(K^\varepsilon\right) < 1 - (1 - \varepsilon) = \varepsilon, \tag{24}$$

where we write S^{c} for the complement of a set S. Let

$$K^\varepsilon_{\rho/\varepsilon} := \left\{x \in \mathbb{R}^m \colon \ \|x - y\| \leq \rho/\varepsilon, \ y \in K^\varepsilon\right\}.$$

Then $K^\varepsilon_{\rho/\varepsilon}$ is a compact subset of $\mathbb{R}^m$. In what follows, we shall show that $Q(K^\varepsilon_{\rho/\varepsilon}) \geq 1 - 2\varepsilon$ for each $Q \in \mathcal{P}_\rho$, from which the assertion follows.

Assume for the sake of a contradiction that $Q(K^{\varepsilon\mathsf{c}}_{\rho/\varepsilon}) \geq 2\varepsilon$. For the given Q, it follows by the definition of set $\mathcal{P}_\rho$ that there exists $P \in \mathcal{P}$ such that $\mathsf{d}_r(P, Q) \leq \rho$. For this pair of P and Q, let π be a transport plan such that $\mathsf{d}_r(P, Q)^r = \iint \|y - x\|^r \, \pi(\mathrm{d}x, \mathrm{d}y)$. Existence of π is guaranteed by Villani [57, Theorem 1.3].

To proceed the proof, let us define a subset of $\mathbb{R}^m \times \mathbb{R}^m$, $A := K^\varepsilon \times K^{\varepsilon\mathsf{c}}_{\rho/\varepsilon}$. It is easy to see that $\|y - x\| > \frac{\rho}{\varepsilon}$ for very point $(x, y) \in A$. Moreover,

$$\pi(A) \geq \pi\left(\mathbb{R}^m \times K^{\varepsilon\,\mathsf{c}}_{\rho/\varepsilon}\right) - \pi\left(K^{\varepsilon\,\mathsf{c}} \times \mathbb{R}^m\right) = Q(K^{\varepsilon\,\mathsf{c}}_{\rho/\varepsilon}) - P\left(K^{\varepsilon\,\mathsf{c}}\right) \geq 2\varepsilon - \varepsilon = \varepsilon,$$

where the last inequality is due to (24). By the definition of the Wasserstein distance

$$\rho^r \geq \mathsf{d}_r(P, Q)^r \geq \iint_A \|y - x\|^r \, \pi(\mathrm{d}x, \mathrm{d}y) > \left(\frac{\rho}{\varepsilon}\right)^r \pi(A) \geq \left(\frac{\rho}{\varepsilon}\right)^r \varepsilon \geq \rho^r,$$

which is a contradiction as desired. This shows $Q(K^\varepsilon_{\rho/\varepsilon}) \geq 1 - 2\varepsilon$ for every $Q \in \mathcal{P}_\rho$ and in turn that the enlargement $\mathcal{P}_\rho$ is tight.

The weak compactness of the set $\{Q \in \mathscr{P}(\mathbb{R}^m) \colon \mathsf{d}_r(Q, P) \leq \rho \text{ for some } P \in \mathcal{P}\}$ follows from the fact that the set is closed under the topology of weak convergence and Prokhorov's theorem. □

Before concluding this section, we note that it is possible to establish a similar result to Theorem 1 for the balls defined under the Wasserstein metric. The theorem below states this.

Theorem 2 (Quantitative stability of the general Wasserstein ball) *Let* $\mathcal{B}(P, \Delta) = \{Q \colon d_r(P, Q) \leq \Delta\}$. *For every* $P, Q \in \mathscr{P}(\Xi)$ *and* $\Delta_1, \Delta_2 \in \mathbb{R}_+$ *it holds that*

$$\begin{aligned}&\mathbb{H}\big(\mathcal{B}(P, \Delta_1), \mathcal{B}(Q, \Delta_2); d_r\big)\\ &\quad\leq \left(\max\left\{(\Delta_2 + d_r(Q, P))^r - \Delta_1^r, (\Delta_1 + d_r(P, Q))^r - \Delta_2^r\right\}\right)^{\frac{1}{r}}, \end{aligned} \tag{25}$$

where $\mathbb{H}$ *denotes the Hausdorff distance in* $\mathscr{P}(\Xi)$ *associated with* d_r.

The proof is analogous to that of Theorem 1 but requires r-convexity which is elaborated in Pflug and Pichler [42, Lemma 2.10]. We omit the details.

Observe that when $r = 1$, d_r collapses to the Kantorovich metric dl_K and (25) coincides with (13). In the case when $\Delta_1 = 0$, and P lies in $\mathcal{B}(Q, \Delta_2)$, (13) reduces to

$$\mathbb{H}\big(P, \mathcal{B}(Q, \Delta_2); \mathsf{d}_r\big) \leq \Delta_2 + \mathsf{d}_r(Q, P). \tag{26}$$

This covers the case when the true probability distribution lies in some confidence region of an empirical probability measure.

3 Stability of the distributionally robust optimization problem (DRO)

With the technical results about quantitative description of the set of probability measures defined under ζ-metric and the Wasserstein metric in the preceding section, we are now ready to investigate stability of the problems (DRRO) and (DRO) in terms of the optimal value and the optimal solutions w.r.t. variation of the ambiguity set. The variation may be driven by increasingly available information about the true probability distribution or the need for numerical approximation of the distributionally robust optimization problem, see discussions in Sun and Xu [56] and Zhang et al. [61]. This kind of research may be viewed as an extension of classical stability analysis in stochastic programming (see [50,52]).

We start by considering the DRO problem in this section because (i) it is relatively easier to handle, (ii) it allows us to do the analysis under generic ζ-metric and (iii) the model is of independent interest. In the next section, we will deal with the DRRO problem which heavily relies on the Wasserstein metric.

Let us consider the perturbation

$$\inf_{y \in Y} \sup_{P \in \tilde{\mathcal{P}}} \mathbb{E}_P \left[\inf_{z \in Z(y,\xi)} c(y, \xi, z)\right] \tag{27}$$

of problem (DRO), where $\tilde{\mathcal{P}}$ is a perturbation of $\mathcal{P}$. Let $\vartheta(\tilde{\mathcal{P}})$ ($\vartheta(\mathcal{P})$, resp.) denote the optimal value of (27) ((DRO), resp.), and $Y^*(\tilde{\mathcal{P}})$ and $Y^*(\mathcal{P})$ denote the respective set of the optimal solutions. The following theorem states the relationship of these quantities.

Theorem 3 (Quantitative stability of the (DRO) problem) *Let*

$$v(y,\xi) := \inf_{z\in Z(y,\xi)} c(y,\xi,z) \tag{28}$$

be the objective of the inner problem and $\mathscr{H} := \{v(y,\cdot)\colon y \in Y\}$. *Assume that*

$$\max\left\{\sup_{P\in\mathcal{P}} \mathbb{E}_P[v(y,\xi)],\ \sup_{P\in\tilde{\mathcal{P}}} \mathbb{E}_P[v(y,\xi)]\right\} < \infty.$$

Then the following assertions hold.

(i) If $\tilde{\mathcal{P}}$ *and* $\mathcal{P}$ *are weakly compact, then*

$$|\vartheta(\tilde{\mathcal{P}}) - \vartheta(\mathcal{P})| \le \mathbb{H}(\tilde{\mathcal{P}}, \mathcal{P}; \mathsf{dl}_{\mathscr{H}}), \tag{29}$$

where $\mathbb{H}$ *is the Hausdorff distance of two sets in* $\mathscr{P}(\Xi)$ *under* ζ*-metric* $\mathsf{dl}_{\mathscr{H}}$ *associated with the class of functions* $\mathscr{H}$. *In particular, if* $\mathcal{P} = \mathcal{B}(P,r)$ *and* $\tilde{\mathcal{P}} = \mathcal{B}(\tilde{P},r')$, *then*

$$|\vartheta(\tilde{\mathcal{P}}) - \vartheta(\mathcal{P})| \le \mathsf{dl}_{\mathscr{H}}(P,\tilde{P}) + |r'-r|. \tag{30}$$

If the functions in the set $\mathscr{H}$ *are Lipschitz continuous with modulus* κ, *then* $\mathbb{H}(\tilde{\mathcal{P}},\mathcal{P};\mathsf{dl}_{\mathscr{G}}) \le \kappa\,\mathbb{H}(\tilde{\mathcal{P}},\mathcal{P};\mathsf{dl}_K)$ *and* $\mathsf{dl}_{\mathscr{G}}(P,\tilde{P}) \le \kappa\,\mathsf{dl}_K(P,\tilde{P})$, *where* dl_K *is the Kantorovich metric. If the functions in* $\mathscr{H}$ *are bounded by a positive constant* C, *then the above two inequalities hold with* κ *being replaced by* C *and* dl_K *replaced by* dl_{TV}.

(ii) If, in addition, $\sup_{P\in\mathcal{P}} \mathbb{E}_P[v(y,\xi)]$ *satisfies the second order growth condition at* $Y^*(\mathcal{P})$, *that is, there exist positive constants* C *and* υ *such that*

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P[v(y,\xi)] \ge \vartheta(\mathcal{P}) + \upsilon\, d\big(y, Y^*(\mathcal{P})\big)^2, \qquad y \in Y, \tag{31}$$

then

$$\mathbb{D}\big(Y^*(\tilde{\mathcal{P}}), Y^*(\mathcal{P})\big) \le \sqrt{\frac{3}{\upsilon}\mathbb{H}(\tilde{\mathcal{P}},\mathcal{P};\mathsf{dl}_{\mathscr{G}})}, \tag{32}$$

where $\mathbb{H}(\tilde{\mathcal{P}},\mathcal{P};\mathsf{dl}_{\mathscr{G}})$ *is the Hausdorff distance between* $\tilde{\mathcal{P}}$ *and* $\mathcal{P}$ *under* ζ*-metric.*

Proof Part (i). It is well-known that

$$|\vartheta(\tilde{\mathcal{P}}) - \vartheta(\mathcal{P})| \le \sup_{y\in Y}\left|\sup_{P\in\tilde{\mathcal{P}}} \mathbb{E}_P[v(y,\xi)] - \sup_{P\in\mathcal{P}} \mathbb{E}_P[v(y,\xi)]\right|.$$

For each y, by the definition of $\mathscr{H}$, there is a corresponding random function $h \in \mathscr{H}$ such that $h(\xi) = v(y, \xi)$ and

$$\sup_{P \in \tilde{\mathcal{P}}} \mathbb{E}_P[v(y, \xi)] = s_{\tilde{\mathcal{P}}}(h),$$

where $s_{\tilde{\mathcal{P}}}(h) = \sup_{P \in \tilde{\mathcal{P}}} \int_\Xi h(\xi) P(d\xi)$ is a support function. Thus

$$\begin{aligned} |\vartheta(\tilde{\mathcal{P}}) - \vartheta(\mathcal{P})| &\leq \sup_{y \in Y} \left| \sup_{P \in \tilde{\mathcal{P}}} \mathbb{E}_P[v(y, \xi)] - \sup_{P \in \mathcal{P}} \mathbb{E}_P[v(y, \xi)] \right| \\ &\leq \sup_{h \in \mathscr{H}} |s_{\tilde{\mathcal{P}}}(h) - s_{\mathcal{P}}(h)|. \end{aligned} \tag{33}$$

Since $\mathscr{H}$ forms a subset of measurable functions, by Proposition 1

$$\sup_{y \in Y} \left| \sup_{P \in \tilde{\mathcal{P}}} \mathbb{E}_P[v(y, \xi)] - \sup_{P \in \mathcal{P}} \mathbb{E}_P[v(y, \xi)] \right| \leq \sup_{h \in \mathscr{H}} |s_{\tilde{\mathcal{P}}}(h) - s_{\mathcal{P}}(h)| = \mathbb{H}(\tilde{\mathcal{P}}, \mathcal{P}; \mathsf{dl}_{\mathscr{H}}).$$

When the ambiguity set is structured through a ζ-ball, the conclusion follows directly from Theorem 1.

In the case when the set of functions in $\mathscr{H}$ are Lipschitz continuous with modulus κ, we can scale the set of functions by $1/\kappa$ to Lipschitz continuous with modulus being bounded by 1. This will allow us to tighten the estimation (29) by

$$\sup_{h \in \mathscr{H}} |s_{\tilde{\mathcal{P}}}(h) - s_{\mathcal{P}}(h)| \leq \kappa \sup_{g \in \mathscr{G}_1} |s_{\tilde{\mathcal{P}}}(g) - s_{\mathcal{P}}(g)| = \kappa\, \mathbb{H}(\tilde{\mathcal{P}}, \mathcal{P}; \mathsf{dl}_K), \tag{34}$$

where $\mathscr{G}_1$ denotes all Lipschitz continuous functions defined over Ξ with modulus being bounded by 1. A similar argument holds when $\mathscr{H}$ is bounded in which case we may use the definition of the total variation metric.

Part (ii). With uniform Lipschitz continuity of the function $\sup_{P \in \mathcal{P}} \mathbb{E}_P[v(y, \xi)]$ in $\mathcal{P}$ as established in (34), we obtain (32) by virtue of Liu and Xu [36, Lemma 3.8] . □

At this point it might be helpful to link the stability results to the well-established stability results in stochastic programming (see, e.g., [52]) and some recent stability results about DRO. Notice that our stability results may be viewed as a generalization of similar results, where perturbation of a single probability distribution is considered (which corresponds to $r = r' = 0$ in our setting), see for instance Römisch [52, Theorem 5]. Indeed, we can establish (29) by using the fact that

$$\left| \mathbb{E}_P[v(y, \xi)] - \mathbb{E}_Q[v(y, \xi)] \right| \leq \mathsf{dl}_{\mathscr{H}}(P, Q).$$

Our main interest here, however, is to present the error bounds for the optimal value and optimal solutions in terms of the Hausdorff distance of the two ambiguity sets $\mathcal{P}$ and $\tilde{\mathcal{P}}$ which have a ball structure under the generic ζ-metric $\mathsf{dl}_{\mathscr{H}}$ and look into the particular case when the ambiguity sets are constructed with ball structure. In the case

when $r = r'$, we find that the error bounds depend only on the distance of the centers of the two balls. Note also that instead of assuming metric regularity at the optimal solution as in Römisch [52, Theorem 5], we involve a second order growth condition which is more likely to be fulfilled in our setting.

Compared to the existing results on stability analysis for DRO problems (see, e.g., [61]), Theorem 3 exhibits something new in that (i) the stability results are established under any metric of ζ-structure including the total variation metric and Kantorovich metric when $\mathscr{H}$ is bounded or uniformly Lipschitz continuous and (ii) when the ambiguity set is structured via ζ-ball, the variation of the optimal value is bounded by the distance of the centers of the balls and the difference of their radii. In a particular case when $r = 0$ and P is the true unknown probability measure of ξ and P' is constructed through empirical data P_N, we can use Theorem 3 and inequality (18) to estimate the rate of convergence of $\vartheta(\mathcal{B}(P_N, r_N))$ as the sample size increases, see (19). Of course, this estimation is not particularly good when k, the dimension of the random vector ξ, is large, because the bound in (19) depends on k (the curse of dimensionality). Note also that when the functions in the set $\mathscr{H}$ are Lipschitz continuous with modulus κ, $\mathsf{dl}_{\mathscr{H}}(P, P_N) \leq \kappa\, \mathsf{dl}_K(P, P_N)$. Through (19) and (30), we may obtain a confidence interval for the true unknown optimal value $\vartheta(P)$; we refer readers to a similar result in Guo and Xu [21, Corollary 1]

Finally, we note that it is possible to derive some verifiable sufficient conditions for the second order growth condition (31). Consider the case that $v(y, \xi)$ is strongly convex in y for every ξ, that is, there exist an integrable vector-valued function $\eta(\xi)$ and a positive integrable function $\alpha(\xi)$ such that for any fixed $y \in Y$,

$$v(y', \xi) \geq v(y, \xi) + \eta(\xi)^\top (y' - y) + \alpha(\xi)\|y' - y\|^2, \quad \forall \ \ y' \in Y, \xi \in \Xi, \tag{35}$$

where $a^T b$ denotes the scalar product of two vectors a and b, $\alpha(\xi)$ is a positive function with $\inf_{P\in\mathcal{P}} \mathbb{E}_P[\alpha(\xi)] > 0$. Let

$$\phi(y') := \sup_{P\in\mathcal{P}} \left(\mathbb{E}_P[v(y, \xi)] + \mathbb{E}_P[\eta(\xi)]^\top (y' - y)\right)$$

and $\nu := \inf_{P\in\mathcal{P}} \mathbb{E}_P[\alpha(\xi)]$. Then

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P[v(y', \xi)] \geq \phi(y') + \nu\|y' - y\|^2, \quad \forall \ \ y' \in Y. \tag{36}$$

Moreover, $\phi(\cdot)$ is a convex function and $\phi(y) = \sup_{P\in\mathcal{P}} \mathbb{E}_P\,[v(y, \xi]$. Thus there exists some deterministic vector $\hat{\eta}$ (depending on y) such that

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P[v(y', \xi)] \geq \sup_{P\in\mathcal{P}} \mathbb{E}_P[v(y, \xi)] + \hat{\eta}^\top (y' - y) + \nu\|y' - y\|^2, \quad \forall \ \ y' \in Y. \tag{37}$$

Since the inequality holds for any y, this shows $\sup_{P\in\mathcal{P}} \mathbb{E}_P[v(y', \xi)]$ is strongly convex and hence $Y^*(\mathcal{P})$ is a singleton, we denote it by $\{y^*(\mathcal{P})\}$. This immediately implies the second order growth condition (31) because $\sup_{P\in\mathcal{P}} \mathbb{E}_P[v(y^*(\mathcal{P}), \xi)] = \vartheta(\mathcal{P})$ and we can choose $\hat{\eta} = 0$ at y^*.

3.1 Robust optimization

In a particular case when the ambiguity set $\mathcal{P} = \mathscr{P}(\Xi)$, the DRO model collapses to the robust optimization problem

$$\inf_{y \in Y} \sup_{\xi \in \Xi} \inf_{z \in Z(y,\xi)} c(y, \xi, z), \tag{RO}$$

where the optimal decision on the first stage is based on the worst scenario of ξ. There is a vast literature on robust optimization, see the monograph [4] for a comprehensive overview of the model, numerical methods and applications. Unfortunately, Theorem 3 does not cover this important case. Below, we make a separate statement about stability of the problem (RO) for the case when both Y and Ξ are perturbed. Perturbation of Y may stem from change of problem data in the constraints of the first stage decision variables and/or removal of some constraints of first stage decision variables such as integer constraints whereas perturbation of Ξ may result from discretization of ξ in the minimax optimization, see for instance Xu et al. [59, Section 3] and Chen et al. [7].

Theorem 4 (Quantitative stability of problem (RO)) *Let $v(y, \xi)$ be defined as in (28). Assume that there exist positive constants L_Ξ, L_Y such that*

$$v(y, \xi) - \tilde{v}(\tilde{y}, \tilde{\xi}) \leq L_Y \cdot d(y, \tilde{y}) + L_\Xi \cdot d(\xi, \tilde{\xi}) \quad \forall \quad \tilde{\xi} \in \tilde{\Xi}, \tilde{y} \in \tilde{Y}. \tag{38}$$

Then

$$\inf_{y \in Y} \sup_{\xi \in \Xi} v(y, \xi) - \inf_{\tilde{y} \in \tilde{Y}} \sup_{\tilde{\xi} \in \tilde{\Xi}} \tilde{v}(y, \tilde{\xi}) \leq L_\Xi \cdot \mathbb{D}(\Xi, \tilde{\Xi}) + L_Y \cdot \mathbb{D}(\tilde{Y}, Y). \tag{39}$$

If, in addition, v is Lipschitz continuous in both y and ξ, i.e.,

$$\left| v(y, \xi) - v(\tilde{y}, \tilde{\xi}) \right| \leq L_Y \cdot d(y, \tilde{y}) + L_\Xi \cdot d(\xi, \tilde{\xi}), \tag{40}$$

then

$$\left| \inf_{y \in Y} \sup_{\xi \in \Xi} v(y, \xi) - \inf_{\tilde{y} \in \tilde{Y}} \sup_{\tilde{\xi} \in \tilde{\Xi}} v(y, \tilde{\xi}) \right| \leq L_Y \cdot \mathbb{H}(Y, \tilde{Y}) + L_\Xi \cdot \mathbb{H}(\Xi, \tilde{\Xi}). \tag{41}$$

Proof By taking the infimum in (38) with respect to $\tilde{\xi} \in \tilde{\Xi}$ it follows that

$$v(y, \xi) - \sup_{\tilde{\xi} \in \tilde{\Xi}} \tilde{v}(\tilde{y}, \tilde{\xi}) \leq L_Y \cdot d(y, \tilde{y}) + L_\Xi \cdot \inf_{\tilde{\xi} \in \tilde{\Xi}} d(\xi, \tilde{\xi}),$$

and consequently

$$\sup_{\xi \in \Xi} v(y, \xi) - \sup_{\tilde{\xi} \in \tilde{\Xi}} \tilde{v}(\tilde{y}, \tilde{\xi}) \leq L_Y \cdot d(y, \tilde{y}) + L_\Xi \cdot \sup_{\xi \in \Xi} \inf_{\tilde{\xi} \in \tilde{\Xi}} d(\xi, \tilde{\xi}).$$

By taking infimum w.r.t. $y \in Y$, it yields

$$\inf_{y\in Y}\sup_{\xi\in\Xi} v(y,\xi) - \sup_{\tilde{\xi}\in\tilde{\Xi}} \tilde{v}(\tilde{y},\tilde{\xi}) \le L_Y \cdot \inf_{y\in Y} d(y,\tilde{y}) + L_\Xi \cdot \mathbb{D}(\Xi,\tilde{\Xi})$$

and a further operation of supremum w.r.t. $\tilde{y} \in \tilde{Y}$ gives rise to

$$\begin{aligned}\inf_{y\in Y}\sup_{\xi\in\Xi} v(y,\xi) - \inf_{\tilde{y}\in\tilde{Y}}\sup_{\tilde{\xi}\in\tilde{\Xi}} \tilde{v}(\tilde{y},\tilde{\xi}) &\le L_Y \cdot \sup_{\tilde{y}\in\tilde{Y}}\inf_{y\in Y} d(y,\tilde{y}) + L_\Xi \cdot \mathbb{D}(\Xi,\tilde{\Xi})\\ &= L_Y \cdot \mathbb{D}(\tilde{Y},Y) + L_\Xi \cdot \mathbb{D}(\Xi,\tilde{\Xi}),\end{aligned}$$

which is (39), the first assertion.

The second assertion is immediate by interchanging the roles of (y,ξ) and $(\tilde{y},\tilde{\xi})$ as the distances d on Y and Ξ are symmetric. □

The condition on Lipschitz continuity of $v(y,\xi)$ w.r.t. (y,ξ) is essential in deriving the stability result. In what follows, we give a sufficient condition for this.

Consider the feasible set-valued mapping $Z\colon Y \times \Xi \rightrightarrows \mathbb{R}^m$ at the second stage. Let $(y_0,\xi_0) \in Y \times \Xi$ be fixed, $z_0 \in Z(y_0,\xi_0)$. We say Z is *pseudo-Lipschitzian* at (y_0,ξ_0), if there are neighbourhoods

V of z_0, U of (y_0,ξ_0) and positive constant L_Z such that

$$Z(y,\xi)\cap V \subset Z(y_0,\xi_0) + L_Z d((y,\xi),(y_0,\xi_0))\mathcal{B}$$

and

$$Z(y_0,\xi_0)\cap V \subset Z(y,\xi) + L_Z d((y,\xi),(y_0,\xi_0))\mathcal{B}$$

for all (y,ξ) in U, where $\mathcal{B}$ denotes the unit ball in the space $\mathbb{R}^m \times \mathbb{R}^k$. In the case when the feasible set at the second stage is defined by a cone constrained system, a sufficient condition for the desired Lipschitzian property is the Slater condition w.r.t. the variable z, see Zhang et al. [62, Lemma 2.2]. Here we make a generic assumption on the desired property of $Z(y,\xi)$ rather than look into its concrete structure so that we can focus on the fundamental issues about stability.

Proposition 4 (Lipschitz continuity of $v(y,\xi)$) *Assume: (i) $Z(y,\xi)$ is pseudo-Lipschitzian at every pair of $(z_0,(y_0,\xi_0)) \in Z(y_0,\xi_0)\times\{(y_0,\xi_0)\}$, (ii) there exists a positive constants L_c and β such that*

$$|c(y,\xi,z) - c(y_0,\xi_0,z_0)| \le L_c[d(y,y_0) + d(\xi,\xi_0)^\beta + d(z,z_0)] \tag{42}$$

for $(y,\xi) \in U$ and $z \in V$. Then there exists a positive constant L such that

$$|v(y,\xi) - v(y_0,\xi_0)| \le L\left[d(y,y_0) + d(\xi,\xi_0) + d(\xi,\xi_0)^\beta\right] \tag{43}$$

for $(y,\xi) \in U$.

Proof The conclusion follows directly from Klatte [33, Theorem 1]. □

It is important to note that the error bound in (43) is determined by the term $d(\xi, \xi_0)$ when ξ is close to ξ_0 and the term $d(\xi, \xi_0)^\beta$ otherwise (in the case when $\beta > 1$).

Note that in the case when $\tilde{Y} = Y$, (41) reduces to

$$\left| \inf_{y\in Y} \sup_{\xi\in\Xi} v(y,\xi) - \inf_{y\in Y} \sup_{\tilde{\xi}\in\tilde{\Xi}} v(y,\tilde{\xi}) \right| \le L_\Xi \cdot \mathbb{H}(\Xi, \tilde{\Xi})$$

when v is uniformly Lipschitz continuous in ξ, i.e.,

$$\left| v(y,\xi) - v(y,\tilde{\xi}) \right| \le L_\Xi \cdot d(\xi,\tilde{\xi}) \qquad \forall \quad y \in Y. \tag{44}$$

In that case, we regard $v(y, \tilde{\xi})$ as a perturbation of $v(y, \xi)$.

4 Stability of the problem (DRRO)

We now move on to discuss stability of the distributionally robust risk optimization problem (DRRO). To this end we need to give a detailed description about the risk measure $\mathcal{R}_{S;P}$ in problem (DRRO).

4.1 Risk functionals

Let X be a random variable. Recall that the value at risk at level $\alpha \in [0, 1)$ is defined as

$$\mathsf{V@R}_\alpha(X) := \inf\{x \in \mathbb{R} \colon P(X \le x) \ge \alpha\}.$$

It is well known that the $\mathsf{V@R}_\alpha(X)$ is a lower semicontinuous quantile function of α over $[0, 1)$. The average value at risk is an upper average value at risk defined as

$$\mathsf{AV@R}_\alpha(X) := \frac{1}{1-\alpha} \int_\alpha^1 \mathsf{V@R}_t(X) dt.$$

Obviously, $\mathsf{AV@R}_0(X) = \mathbb{E}_P[X]$.

Let $\sigma : [0, 1) \to \mathbb{R}_+$ be a nonnegative, nondecreasing function with $\int_0^1 \sigma(t)\mathrm{d}t = 1$. We call

$$\mathcal{R}_\sigma(X) := \int_0^1 \sigma(\alpha)\mathsf{V@R}_\alpha(X)\mathrm{d}\alpha \tag{45}$$

the *distortion risk measure* of X associated with the *distortion functional* σ. Clearly, $\mathcal{R}_\sigma(X)$ is a weighted average of the value at risk and the average value at risk is a special distortion risk measure because

$$\mathsf{AV@R}_\alpha(X) := \int_0^1 \mathsf{V@R}_t(X)\sigma_\alpha(t)dt$$

with

$$\sigma_\alpha(t) = \begin{cases} 0 & \text{if} \quad t \in [0, \alpha], \\ \frac{1}{1-\alpha} & \text{if} \quad t \in (\alpha, 1]. \end{cases} \tag{46}$$

The distortion risk measure coincides with the *spectral risk measure* introduced by Denneberg [8], Acerbi [1] whereby σ is called *risk spectrum*. The risk measure is coherent and law invariant in that it satisfies monotonicity, positive homogeineity, subadditivity and translation invariance, and it depends only on the distribution of X. The non-decreasing property of σ is vital to ensure the coherence property as "it assigns bigger weights to worse cases" [1]. For more concrete examples of σ and the resulting distortion/spectral risk measures, see Dowd et al. [11].

For a set S of distortion functionals, we can define

$$\mathcal{R}_{S;P}(X) := \sup_{\sigma \in S} \mathcal{R}_\sigma(X),$$

where P is the probability measure. By employing $\mathcal{R}_{S;P}$ in the model (DRRO) we are concerned not only with ambiguity of the true probability distribution of the underlying random variables but also with ambiguity of the risk profile σ to be used in the definition of risk measure $\mathcal{R}_\sigma(X)$. The latter may be related to ambiguity of a decision maker's risk preference [2,25].

The following result is a combination of the well-known Kusuoka representation theorem Kusuoka [35] and its implication in terms of the connection with the distortion risk measure, see Pflug and Pichler [42, Theorem 3.13, Corollary 3.14]. For further examples of risk measures and respective Kusuoka representations we refer readers to Dentcheva et al. [9].

Theorem 5 (Kusuoka representation theorem) *Let* $(\Omega, \mathcal{F}, P)$ *be a nonatomic probability space and* $\mathscr{L}^\infty$ *be a set of random variables mapping from* Ω *to* $\mathbb{R}$. *Let* $\mathcal{R}\colon \mathscr{L}^\infty \to \mathbb{R}$ *be a law invariant coherent risk measure. Then there exists a set of probability measures* $\mathcal{M}$ *on* $[0, 1)$ *equipped with Borel sigma algebra such that*

$$\mathcal{R}(X) = \sup_{\mu \in \mathcal{M}} \int_0^1 \mathsf{AV@R}_\alpha(X)d\mu(\alpha).$$

Moreover, $\mathcal{R}$ *has the representation*

$$\mathcal{R}(X) = \sup_{\sigma \in S} \mathcal{R}_\sigma(X), \tag{47}$$

where S *is a set of continuous and bounded distortion densities.*

The theorem means that any law invariant coherent risk measure can be represented as the supremum of a class of the distortion risk measures. Without loss of generality,

we may assume from here on that the risk measure $\mathcal{R}_{S;P}$ in problem (DRRO) is defined as in (47) and hence it is a law invariant conherent risk measure.

Note that for each fixed $\sigma \in S$, it follows by the convex duality (cf. [46,53] or [45] for the appropriate space) that the risk measure $\mathcal{R}_\sigma(\cdot)$ has a dual representation which is given by

$$\begin{aligned}\mathcal{R}_\sigma(X) = \sup\Bigg\{ & \mathbb{E}_P[X\zeta]\colon \zeta \geq 0,\ \mathbb{E}_P[\zeta] = 1, \\ & \mathsf{AV@R}_\alpha(\zeta) \leq \int_\alpha^1 \sigma(u)\mathrm{d}u \quad \text{for all} \quad \alpha \in (0,1)\Bigg\}\end{aligned} \tag{48}$$

and hence

$$\begin{aligned}\mathcal{R}_{S,P}(X) = \sup\Bigg\{ & \mathbb{E}_P[X\zeta]\colon \zeta \geq 0,\ \mathbb{E}_P[\zeta] = 1, \\ & \mathsf{AV@R}_\alpha(\zeta) \leq \int_\alpha^1 \sigma(u)\mathrm{d}u \quad \text{for all} \quad \alpha \in (0,1), \sigma \in S\Bigg\}.\end{aligned} \tag{49}$$

We will use the latter in the forthcoming stability analysis.

4.2 Stability analysis

We proceed the analysis in a slightly different manner from what we did in the previous section by considering a variation not only of the ambiguity set but also of the space of the decision variables and the support set of the random variables. This will allow our results to be applicable to a broader class of problems including multistage stochastic programming problems. We need the following intermediate result, which allows comparing risk functionals, evaluated for different random variables and probability measures.

Proposition 5 *Let $X, \tilde{X}\colon \Xi \to \mathbb{R}$ be real valued random variables and S be a compact set of distortion functionals in Lebesgue space $\mathscr{L}^q$. Assume that there are positive constants $L > 0$ and $\beta \in (0,\ 1]$ such that*

$$X(\xi) - \tilde{X}(\tilde{\xi}) \leq L \cdot d(\xi, \tilde{\xi})^\beta. \tag{50}$$

Then

$$\mathcal{R}_{S;P}(X) - \mathcal{R}_{S;\tilde{P}}(\tilde{X}) \leq L \sup_{\sigma \in S} \|\sigma\|_q \ \mathsf{d}_{\beta p}(P, \tilde{P})^\beta, \tag{51}$$

where $\mathcal{R}_{S;P}$ is the risk functional induced by S as defined in (47), $q \geq 1$ is the Hölder conjugate exponent to $p \geq 1$, $\frac{1}{p} + \frac{1}{q} = 1$ and d_r is the Wasserstein distance of order $r \geq 1$ (see 21). The bound in (51) is tight in the case that σ is uniformly bounded, $\beta = 1$ and $p = 1$, i.e., $\mathsf{d}_{\beta p}(P, \tilde{P})^\beta$ reduces to the Kantorovich metric.

Before providing a proof, we make some comments on condition (50). In practice, $X(\xi)$ may be regarded as the optimal value of the original second stage problem

whereas $\tilde{X}(\tilde{\xi})$ is the optimal value of the approximate second stage problem where the approximation may result from discretization, application of decision rules or k-adaptability. Such approximation will usually affect the space of the second stage decision variables. This motivates us to write the optimal value $\tilde{X}(\tilde{\xi})$ rather than $X(\tilde{\xi})$.

Proof We shall employ the dual representation (49) to prove the result. Let $\zeta, \tilde{\zeta} \in \mathscr{L}^1$ be the dual variables so that

$$\mathcal{R}_{\mathcal{S};P}(X) = \mathbb{E}_P[X\zeta] \quad \text{and} \quad \mathcal{R}_{\mathcal{S};\tilde{P}}(\tilde{X}) = \mathbb{E}_{\tilde{P}}[\tilde{X}\tilde{\zeta}].$$

Let π be a bivariate probability measure with marginals P and $\tilde{P}$. Note that X, ζ and $\tilde{X}$, $\tilde{\zeta}$ are coupled in a comonotone manner so that we can derive by Hoeffding's Lemma (cf. [28])

$$\begin{aligned}\mathcal{R}_{S;P}(X) - \mathcal{R}_{S;\tilde{P}}(\tilde{X}) = \mathbb{E}_P[X\zeta] - \mathbb{E}_{\tilde{P}}[\tilde{X}\tilde{\zeta}] &= \iint [X(\xi)\zeta(\xi) - \tilde{X}(\tilde{\xi})\tilde{\zeta}(\tilde{\xi})]\pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}) \\ &\leq \iint \Big(X(\xi) - \tilde{X}(\tilde{\xi})\Big)\zeta(\xi)\pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}) \\ &\leq L \iint \zeta(\xi) d(\xi, \tilde{\xi})^{\beta} \pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}).\end{aligned}$$

The first inequality is due to the fact that $\mathbb{E}_{\tilde{P}}[\tilde{X}\tilde{\zeta}] \geq \mathbb{E}_{\tilde{P}}[\tilde{X}\zeta]$ because $\tilde{\zeta}$ gives the maximum value of $\mathcal{R}_{\cdot;\tilde{P}}(\tilde{X})$ over S, see (47). By applying Hölder's inequality, we obtain

$$\mathcal{R}_{S;P}(X) - \mathcal{R}_{S;\tilde{P}}(\tilde{X}) \leq L \cdot \left(\int \zeta^q \mathrm{d}\pi\right)^{\frac{1}{q}} \left(\int d(\xi, \tilde{\xi})^{\beta p} \pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi})\right)^{\frac{1}{p}}.$$

Moreover, by taking infimum with respect to all probability measures π with marginals P and $\tilde{P}$, we deduce

$$\mathcal{R}_{S;P}(X) - \mathcal{R}_{S;\tilde{P}}(\tilde{\zeta}) \leq L \cdot \|\zeta\|_q \mathsf{d}_{\beta p}(P, \tilde{P})^{\beta},$$

from which the conclusion follows, as the cumulative distribution function of ζ is σ for some $\sigma \in S$. □

We are now ready to state the main stability result of this section.

Theorem 6 (Quantitative stability of the problem (DRRO)) *Let $v(y, \xi)$ be defined as in (28),*

$$\vartheta := \inf_{y \in Y} \sup_{P \in \mathcal{P}} \mathcal{R}_{S;P}\big(v(y, \xi)\big) \quad \text{and} \quad \tilde{\vartheta} := \inf_{\tilde{y} \in \tilde{Y}} \sup_{P \in \tilde{\mathcal{P}}} \mathcal{R}_{S;P}\big(v(\tilde{y}, \xi)\big).$$

The following assertions hold.

(i) If there are positive constants L_Ξ, L_Y and L_Z and $\beta \in (0, 1]$ such that

$$|c(y, \xi, z) - \tilde{c}(\tilde{y}, \tilde{\xi}, \tilde{z})| \leq L_\Xi \, d(\xi, \tilde{\xi})^\beta + L_Y \, d(y, \tilde{y}) + L_Z \, d(z, \tilde{z}), \tag{52}$$

for all $(y, \xi, z), (\tilde{y}, \tilde{\xi}, \tilde{z}) \in Y \times \Xi \times \hat{Z}$, where $\hat{Z}$ is a set of $\mathbb{R}^m$ containing $Z(y, \xi)$ for all $(y, \xi) \in Y \times \Xi$, and the feasible set-valued mapping Z is pseudo-Lipschitzian at $(z, (y, \xi)) \in Z(y, \xi) \times \{(y, \xi)\}$ for every $(y, \xi) \in Y \times \Xi$, then there exists a positive constant L such that

$$|\tilde{\vartheta} - \vartheta| \leq L \cdot \left[\sup_{\sigma \in S} \|\sigma\|_q \left(\mathbb{H}\left(\mathcal{P}, \tilde{\mathcal{P}}; \mathsf{d}_p\right) + \mathbb{H}\left(\mathcal{P}, \tilde{\mathcal{P}}; \mathsf{d}_{p\beta}\right)^\beta \right) + \mathbb{H}(\tilde{Y}, Y) \right], \tag{53}$$

where p and q are Hölder conjugate exponents, i.e., $\frac{1}{p} + \frac{1}{q} = 1$;

(ii) let $Y^*(\mathcal{P})$ denote the set of optimal solutions of the problem (DRRO). If the function $\sup_{P \in \mathcal{P}} \mathcal{R}_{S,P}\,(v(y, \xi))$ satisfies the second order growth condition at $Y^*(\mathcal{P})$, that is, there exists a positive constant υ such that

$$\sup_{P \in \mathcal{P}} \mathcal{R}_{S;P}\,(v(y, \xi)) \geq \vartheta(\mathcal{P}) + \upsilon d(y, Y^*(\mathcal{P}))^2, \quad \forall \quad y \in Y, \tag{54}$$

then

$$\mathbb{D}(Y^*(\tilde{\mathcal{P}}), Y^*(\mathcal{P})) \leq \sqrt{\frac{3}{\upsilon} \mathbb{H}\left(\mathcal{P}, \tilde{\mathcal{P}}; \mathsf{d}_{p\beta}\right)}; \tag{55}$$

(iii) in the case when $\mathcal{P} = \{P\}$, where P is the true probability distribution and $\mathcal{P} = \mathcal{B}(P_N, r_N)$ (where $\mathcal{B}(P_N, r_N)$ is defined as in inequality (18),

$$\mathbb{H}\left(\mathcal{P}, \tilde{\mathcal{P}}; \mathsf{d}_p\right) \leq \mathsf{d}_p(P_N, P) + r_N.$$

Proof Part (i). Observe first that under condition (52), it follows by Proposition 4 that there exists a positive constant L such that

$$|v(y, \xi) - v(\tilde{y}, \tilde{\xi})| \leq L[d(y, \tilde{y}) + d(\xi, \tilde{\xi}) + d(\xi, \tilde{\xi})^\beta]. \tag{56}$$

For fixed y and $\tilde{y}$, define random variables

$$X(\xi) := v(y, \xi) \quad \text{and} \quad \tilde{X}(\tilde{\xi}) := v(\tilde{y}, \tilde{\xi}) + Ld(y, \tilde{y}).$$

Then $X(\xi) - \tilde{X}(\tilde{\xi}) \leq L[d(\xi, \tilde{\xi}) + d(\xi, \tilde{\xi})^\beta]$, which enables us to use Proposition 5 to derive

$$\mathcal{R}_P(X) - \mathcal{R}_{\tilde{P}}(\tilde{X}) \leq L \sup_{\sigma \in \mathcal{S}} \|\sigma\|_q \left(\mathsf{d}_p(P, \tilde{P}) + \mathsf{d}_{\beta p}(P, \tilde{P})^\beta \right).$$

Moreover, by exploiting the property of translation invariance of the risk measure, we obtain

$$\mathcal{R}_P\left(v(y,\xi)\right)-\mathcal{R}_{\tilde{P}}\left(v(\tilde{y},\tilde{\xi})\right)\leq L\left[\sup_{\sigma\in\mathcal{S}}\|\sigma\|_q\left(\mathsf{d}_p(P,\tilde{P})+\mathsf{d}_{\beta p}(P,\tilde{P})^{\beta}\right)+d(y,\tilde{y})\right].$$

Taking infimum w.r.t. $\tilde{P}\in\tilde{\mathcal{P}}$ and supremum w.r.t. $P\in\mathcal{P}$, we obtain

$$\begin{aligned}
&\sup_{P\in\mathcal{P}}\mathcal{R}_P\left(v(y,\xi)\right)-\sup_{\tilde{P}\in\tilde{\mathcal{P}}}\mathcal{R}_{\tilde{P}}\left(v(\tilde{y},\tilde{\xi})\right)\\
&\qquad\leq L\left[\sup_{\sigma\in\mathcal{S}}\|\sigma\|_q\sup_{P\in\mathcal{P}}\inf_{\tilde{P}\in\tilde{\mathcal{P}}}\left(\mathsf{d}_p(P,\tilde{P})+\mathsf{d}_{\beta p}(P,\tilde{P})^{\beta}\right)+d(y,\tilde{y})\right]\\
&\qquad= L\left[\sup_{\sigma\in\mathcal{S}}\|\sigma\|\left(\mathbb{D}\left(\mathcal{P},\tilde{\mathcal{P}};\mathsf{d}_p\right)+\mathbb{D}\left(\mathcal{P},\tilde{\mathcal{P}};\mathsf{d}_{\beta p}\right)^{\beta}\right)+d(y,\tilde{y})\right].
\end{aligned}$$

Finally, taking infimum with respect to $y\in Y$ and then the supremum with respect to $\tilde{y}\in\tilde{Y}$, we arrive at

$$\begin{aligned}
&\inf_{y\in Y}\sup_{P\in\mathcal{P}}\mathcal{R}_P\left(v(y,\xi)\right)-\inf_{\tilde{y}\in\tilde{Y}}\sup_{\tilde{P}\in\tilde{\mathcal{P}}}\mathcal{R}_{\tilde{P}}\left(v(\tilde{y},\tilde{\xi})\right)\\
&\quad\leq L\left[\sup_{\sigma\in\mathcal{S}}\|\sigma\|\left(\mathbb{D}\left(\mathcal{P},\tilde{\mathcal{P}};\mathsf{d}_p\right)+\mathbb{D}\left(\mathcal{P},\tilde{\mathcal{P}};\mathsf{d}_{\beta p}\right)^{\beta}\right)+\sup_{\tilde{y}\in\tilde{Y}}\inf_{y\in Y}d(y,\tilde{y})\right]\\
&\quad= L\left[\sup_{\sigma\in\mathcal{S}}\|\sigma\|\left(\mathbb{D}\left(\mathcal{P},\tilde{\mathcal{P}};\mathsf{d}_p\right)+\mathbb{D}\left(\mathcal{P},\tilde{\mathcal{P}};\mathsf{d}_{\beta p}\right)^{\beta}\right)+\mathbb{D}(\tilde{Y},Y)\right].
\end{aligned}$$

The conclusion follows by swapping the position between y, $\mathcal{P}$ and $\tilde{y}$ and $\tilde{\mathcal{P}}$.

Part (ii) follows from a similar argument to Part (ii) of Theorem 3. We omit the details of the proof.

Part (iii) follows from Theorem 1. □

Theorem 6 gives a quantitative description on the impact of the optimal value of the problem (DRRO) upon the change of the ambiguity set $\mathcal{P}$ and the space of the first stage decision variables Y. It might be helpful to give a few comments about this result.

- As far as we are concerned, this is the first stability result for the distributionally robust risk optimization model. Compared to Theorem 3, Theorem 6 requires additional condition on uniform Lipschitz/Hölder continuity of $v(y,\xi)$ in ξ. The condition allows us to use a less tighter metric than ζ-metric. In the case when the set S of distortion functionals consists of a unique function which takes constant value 1 and $\beta p=1$, Theorem 6 recovers part of Theorem 3 (with Kantorovich–Wasserstein metric).
- In (56), the term $d(\xi,\tilde{\xi})$ arises from pseudo-Lipschitzian continuity of the feasible set of the second stage problem $Z(y,\xi)$ w.r.t. ξ whereas the term $d(\xi,\tilde{\xi})^{\beta}$ arises

from Hölder continuity of the cost function c w.r.t. ξ, see Proposition 4. When $\beta = 1$, (53) simplifies to

$$|\tilde{\vartheta} - \vartheta| \leq L \cdot \left\{ 2 \sup_{\sigma \in S} \|\sigma\|_q \cdot \mathbb{H}\left(\mathcal{P}, \tilde{\mathcal{P}}; \mathsf{d}_p\right) + \mathbb{H}(\tilde{Y}, Y) \right\}.$$

- The variation of decision variables y, z and ξ in the stability results allows one to apply the result to multistage decision making process where change of the underlying uncertainty arises not only from probability distribution at leaves of the random process but also the tree structure (filtration). In that case, the variation of ξ must be distinguished from that of $\mathcal{P}$. The former will also affect the structure of the decision variable via nonanticipativity conditions, see Liu et al. [37] and references therein.
- An important case that Theorem 6 may cover is when $\mathcal{P}$ is defined by some prior moment conditions whereas $\tilde{\mathcal{P}}$ is its discretization. The discretization is important because when $\mathcal{P}$ is a set of discrete probability measures, the problem (DRRO) becomes an ordinary minimax optimization problem in finite dimensional space, consequently we can apply some existing numerical methods in the literature such as the cutting plane method in Xu et al. [59, Algorithm 4.1] to solve the problem. Liu et al. [37] derived an error bound for such ambiguity set and its discretization, see Liu et al. [37, Section 3] for details. Our Theorem 6 applies to such a case when the ambiguity set in problem (DRRO) is defined and discretized in that manner.
- Dentcheva et al. [10] considered a class of composite risk measures and presented some asymptotic convergence results when the underlying probability is approximated by the empirical measure. Our results do not apply to composite risk measures as they are not law invariant and consequently do not have Kusuoka representation.

5 Applications

In this section, we outline potential applications of the stability results established in the preceding sections. We focus on (DRRO) as similar conclusions can be drawn for (DRO).

5.1 Distretization of (DRRO) through empirical probability distribution

Let us consider problem (DRRO) with the ambiguity set $\mathcal{P}$ being defined by Kantorovich-ball centered at the true probability distribution P with radius Δ. Consider a perturbation of the problem where P is replaced by the empirical distribution P_N defined in (17) and Δ is replaced by Δ_N (depending on the data). We denote the resulting ambiguity set by $\mathcal{P}_N$ and the corresponding distributionally robust risk minimization problem by (DRRO').

This kind of perturbation is often considered for data-driven problems where the true probability distribution is unknown but it is possible to use empirical data to

construct an empirical probability distribution, see Esfahani and Kuhn [14]. When $\Delta = 0$, problem (DRRO) reduces to an ordinary two-stage risk minimization problem. In that case, we may interpret $\mathcal{P}_N$ as a confidence interval of the empirical probability distribution. We refer readers to Dentcheva et al. [10] for asymptotic convergence of this kind of approach applied to a class of composite risk functionals.

Our focus here is on the difference between the optimal values of (DRRO) and (DRRO'). Under the conditions of Theorem 6, we are able to obtain from part (i) of the theorem and Corollary 1 that

$$|\tilde{\vartheta} - \vartheta| \leq L \cdot \left[\sup_{\sigma \in S} \|\sigma\|_q \left(\mathbb{H}\left(\mathcal{P}, \mathcal{P}_N; \mathsf{d}_p\right) + \mathbb{H}\left(\mathcal{P}, \mathcal{P}_N; \mathsf{d}_{p\beta}\right)^{\beta} \right) + \mathbb{H}(\tilde{Y}, Y) \right]. \quad (57)$$

In the case when $\tilde{Y} = Y$ and $\beta = 1$, we have from (57) and (25)

$$\begin{aligned} |\tilde{\vartheta} - \vartheta| &\leq 2L \cdot \sup_{\sigma \in S} \|\sigma\|_q \, \mathbb{H}\left(\mathcal{P}, \mathcal{P}_N\right) \\ &\leq 2L \cdot \sup_{\sigma \in S} \|\sigma\|_q \max\left(\Delta_N + \mathsf{d}_r(P_N, P))^r - \Delta^r, (\Delta + \mathsf{d}_r(P, P_N))^r - \Delta_N^r \right)^{\frac{1}{r}}. \end{aligned} \quad (58)$$

Two sub-cases might be of special interest. One is that $\Delta = 0$ and $\mathsf{d}_r(P, P_N) \leq \Delta_N$; in this case the above inequality yields

$$|\tilde{\vartheta} - \vartheta| \leq 2L \cdot \sup_{\sigma \in S} \|\sigma\|_q \left(\Delta_N + \mathsf{d}_r(P_N, P)\right). \quad (59)$$

This is the case when true probability distribution P lies in the confidence region of P_N. The other is the case when $\Delta = \Delta_N = 0$. In that case, (DRRO) reduces to the classical risk minimization problem and its perturbation is no more than the well known sample average approximation. Consequently, (58) can be simplified as

$$|\tilde{\vartheta} - \vartheta| \leq 2L \cdot \sup_{\sigma \in S} \|\sigma\|_q \, \mathsf{d}_r(P_N, P). \quad (60)$$

5.2 Optimal quantization

Dupačová et al. [13] propose scenario reduction methods, which aim at simplifying a given discrete probability measure. The underlying techniques are based on heuristics involving the Wasserstein distance and the methods have been developed further by Heitsch and Römisch [26,27] and other authors. The approach typically reduces the number of scenarios by cutting atoms from a discrete probability distribution with small probability or by merging neighboring probability distributions. The heuristics thus successively produce new probability measures, which are distinct, but close in the Wasserstein distance to the preceding measure. Some of the algorithms allow monitoring the Wasserstein distance to the genuine probability measure throughout.

In case that monitoring the distance is not possible or just bounds are available, the distance $\mathsf{d}_r\left(P, \tilde{P}\right)$ between the genuine measure $P_N = \sum_{i=1}^{N} p_i \delta_{\xi_i}$ (say) and its

approximation $\tilde{P}_N = \sum_{i=1}^{\tilde{N}} \tilde{p}_j \delta_{\tilde{\xi}_j}$ can be computed explicitly, it is the objective of the linear program

$$\begin{aligned} \min_{\pi} & \sum_{i=1}^{N} \sum_{j=1}^{\tilde{N}} \pi_{i,j} \cdot d\big(\xi_i, \tilde{\xi}_j\big)^r \\ \text{s.t.} & \sum_{j=1}^{\tilde{N}} \pi_{i,j} = p_i, \quad i = 1, \dots N, \qquad (61) \\ & \sum_{i=1}^{N} \pi_{i,j} = \tilde{p}_j, \quad j = 1, \dots, \tilde{N} \text{ and} \\ & \pi_{i,j} \geq 0, \qquad (62) \end{aligned}$$

which is the discrete equivalent of (21)–(22). In this formulation the supporting points ξ_i and $\tilde{\xi}_j$ possibly differ for both measures P_N and $P_{\tilde{N}}$.

For a fixed number $\tilde{N}$ of approximating scenarios, the problem (61)–(62) can also be considered as optimization problem with variables $\pi_{i,j}$ and ξ_j *simultaneously* ($i = 1, \dots, N$, $j = 1, \dots \tilde{N}$).

The problem of finding optimal locations $\tilde{\xi}_j$ is occasionally referred to as the *facility location* problem. This extended problem is nonlinear and combinatorial, but many algorithms are discussed in the literature to find optimal, or nearly optimal locations or quantizers. We refer readers to the work Pagès [40] or to the monograph Graf and Luschgy [19] for a comprehensive survey. Efficient techniques also employ stochastic approximation. Kovacevic and Pichler [34] generalize the techniques to stochastic processes, whilst Pflug and Pichler [43] also propose probabilistic approaches.

We now turn to look into the case when P_N is discrete and of the form $P_N = \sum_{i=1}^{N} p_i \delta_{\xi_i}$ obtained through an optimal quantization method as outlined above, or obtained through scenario reduction or a quasi-Monte Carlo method (cf. [29]). For the support $\Xi^N = \{\xi_1, \dots, \xi_N\}$ of P_N, let

$$\beta_N := \max_{\xi \in \Xi} \min_{1 \leq i \leq N} d(\xi, \xi_i). \qquad (63)$$

Since $\Xi^N \subset \Xi$, it is easy to see that β_N is indeed the Hausdorff distance between Ξ and Ξ^N. By Pflug and Pichler [42, Lemma 4.9],

$$\mathsf{dl}_K(P, P_N) = \int \min_{1 \leq i \leq N} d(\xi, \xi_i) P(d\xi) \leq \beta_N. \qquad (64)$$

This upper bound can be employed directly in inequality (57), which establishes a comparison of the objectives of the stochastic programs in the case when $p = 1$ and $\beta = 1$ by

$$
\begin{aligned}
|\tilde{\vartheta} - \vartheta| &\le L \cdot \Big[\sup_{\sigma \in S} \|\sigma\|_\infty \left(\mathsf{dl}_K(P, P_N) + |\Delta_N - \Delta| \right. \\
&\quad \left. + (d_K(P, P_N) + |\Delta_N - \Delta|) \right) + \mathbb{H}(\tilde{Y}, Y) \Big] \\
&= L \cdot \left[2 \sup_{\sigma \in S} \|\sigma\|_\infty \left(\beta_N + |\Delta_N - \Delta| \right) + \mathbb{H}(\tilde{Y}, Y) \right].
\end{aligned}
$$

In the case when $\Delta_N = \Delta = 0$, the right hand side of the inequality reduces to

$$
L \cdot \left[2 \sup_{\sigma \in S} \|\sigma\|_\infty \beta_N + \mathbb{H}(\tilde{Y}, Y) \right],
$$

which gives rise to an error bound for the ordinary risk minimization problem.

Acknowledgements We would like to thank Jie Zhang for an initial proof of Theorem 1 and Shaoyan Guo for careful reading of the paper. We would also like to thank the guest editor and the two anonymous referees for insightful comments which help us significantly strengthen this paper.

References

1. Acerbi, B.: Spectral measures of risk: a coherent representation of subjective risk aversion. J. Bank. Finance **26**, 1505–1518 (2002)
2. Armbruster, B., Delage, E.: Decision making under uncertainty when preference information is incomplete. Manag. Sci. **61**, 111–128 (2015)
3. Athreya, K.B., Lahiri, S.N.: Measure Theory and Probability Theory. Springer, Berlin (2006)
4. Ben-Tal, A., Ghaoui, L.E., Nemirovski, A.: Robust Optimization. Princeton University Press, Princeton (2009)
5. Billingsley, P.: Convergence of Probability Measures. Wiley, New York (1968)
6. Castaing, C., Valadier, M.: Convex Analysis and Measurable Multifunctions, Lecture Notes in Mathematics. Springer, Berlin (1977)
7. Chen, X., Sun, H., Xu, H.: Discrete approximation of two-stage stochastic and distributionally robust linear complementarity problems. Math. Program. (2018). https://doi.org/10.1007/s10107-018-1266-4
8. Denneberg, D.: Distorted probabilities and insurance premiums. Methods Oper. Res. **63**, 21–42 (1990)
9. Dentcheva, D., Penev, S., Ruszczyński, A.: Kusuoka representation of higher order dual risk measures. Ann. Oper. Res. **181**, 325–335 (2010). https://doi.org/10.1007/s10479-010-0747-5
10. Dentcheva, D., Penev, S., Ruszczynski, A.: Statistical estimation of composite risk functionals and risk optimization problems. Ann. Inst. Stati. Math. **69**, 737–760 (2017)
11. Dowd, K., Cotter, J., Sorwar, G.: Spectral risk measures: properties and limitations. J. Finance Serv. Res. **34**, 61–75 (2008)
12. Dupačová, J.: Stability in stochastic programming with recourse contaminated distributions. In: Prékopa, A., Wets, R.J.B. (eds.) Stochastic Programming 84 Part I, pp. 133–144. Springer, Berlin (2009). https://doi.org/10.1007/bfb0121117
13. Dupačová, J., Gröwe-Kuska, N., Römisch, W.: Scenario reduction in stochastic programming. Math. Program. Ser. A **95**(3), 493–511 (2003). https://doi.org/10.1007/s10107-002-0331-0
14. Esfahani, P.M., Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Math. Program. **171**, 115–166 (2018)
15. Fan, K.: Minimax theorems. Proc. Natl. Acad. Sci. U. S. A. **39**(1), 42 (1953)
16. Fournier, N., Guilline, A.: On the rate of convergence in Wasserstein distance of the empirical measure. Probab. Theory Relat. Fields **162**, 707–738 (2015). https://doi.org/10.1007/s00440-014-0583-7
17. Gao, R., Kleywegt, A.: Distributionally robust stochastic optimization with wasserstein distance. (2016) Preprint arXiv:1604.02199

18. Gibbs, A.L., Su, F.E.: On choosing and bounding probability metrics. Int. Stat. Rev. **70**, 419–435 (2002)
19. Graf, S., Luschgy, H.: Foundations of Quantization for Probability Distributions, Volume 1730 of Lecture Notes in Mathematics. Springer, Berlin (2000). https://doi.org/10.1007/BFb0103945
20. Gröwe, N.: Estimated stochastic programs with chance constraints. Eur. J. Oper. Res. **101**(2), 285–305 (1997). https://doi.org/10.1016/S0377-2217(96)00398-0
21. Guo, S., Xu, H.: Distributionally Robust Shortfall Risk Optimization Model and Its Approximation. Mathematical Programming Series B. Springer, Berlin (2018). https://doi.org/10.1007/s10107-018-1307-z
22. Hanasusanto, D.K., A., G., Wiesemann, W.: K-adaptability in two-stage robust binary programming. Oper. Res. **63**, 877–891 (2015)
23. Hanasusanto, D.K., A., G., Wiesemann, W.: K-adaptability in two-stage distributionally robust binary programming. Oper. Res. Lett. **44**, 6–11 (2016)
24. Hanasusanto GA, Kuhn D (2018) Conic programming reformulations of two-stage distributionally robust linear programs over wasserstein balls. Oper. Res. https://doi.org/10.1287/opre.2017.1698
25. Haskell, W.B., Huang, W., Xu, H.: Preference elicitation and robust optimization with multi-attribute quasi-concave choice functions (2018). Preprint arXiv:1805.06632
26. Heitsch, H., Römisch, W.: Scenario reduction algorithms in stochastic programming. Comput. Optim. Appl. Stoch. Program. **24**(2–3), 187–206 (2003). https://doi.org/10.1023/A:1021805924152
27. Heitsch, H., Römisch, W.: A note on scenario reduction for two-stage stochastic programs. Oper. Res. Lett. **6**, 731–738 (2007)
28. Hoeffding, W.: Maßstabinvariante Korrelationstheorie. Schr. Math. Inst. Univ. Berlin **5**, 181–233 (1940). German
29. Homem de Mello, T.: On rates of convergence for stochastic optimization problems under non-iid sampling. SIAM J. Optim. **19**, 524–551 (2008)
30. Hörmander, L.: Sur la fonction d'appui des ensembles convexes dans un espace localement convexe. Ark. Mat. **3**(2), 181–186 (1955). https://doi.org/10.1007/BF02589354. In French
31. Kantorovich, L.V., Rubinshtein, G.S.: On a space of totally additive functions. Vestnik Leningr. Univ. **13**, 52–59 (1958)
32. Kelley, J.E.: The cutting-plane method for solving convex programs. J. Soc. Ind. Appl. Math. **8**, 703–712 (1960)
33. Klatte, D.: A note on quantitative stability results in nonlinear optimization. Seminarbericht, Sektion Mathematik, Humboldt-Universität zu Berlin, Berlin **90**, 77–86 (1987)
34. Kovacevic, R.M., Pichler, A.: Tree approximation for discrete time stochastic processes: a process distance approach. Ann. Oper. Res. **235**, 395–421 (2015). https://doi.org/10.1007/s10479-015-1994-2
35. Kusuoka, S.: Chapter 4: on law invariant coherent risk measures. In: Kusuoka, S., Maruyama, T. (eds.) Advances in Mathematical Economics, vol. 3, pp. 83–95. Springer, Berlin (2001). https://doi.org/10.1007/978-4-431-67891-5
36. Liu, Y., Xu, H.: Stability and sensitivity analysis of stochastic programs with second order dominance constraintss. Math. Program. Ser. A **142**, 435–460 (2013)
37. Liu, Y., Pichler, A., Xu, H.: Discrete approximation and quantification in distributionally robust optimization Math. Oper. Res. (2017). https://doi.org/10.1287/moor.2017.0911
38. Mehrotra, S., Papp, D.: A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. SIAM J. Optim. **24**, 1670–1697 (2014). https://doi.org/10.1137/130925013
39. Norkin, V.I., Keyzer, M.A.: On convergence of kernel learning estimators. SIAM J. Optim. **20**(3), 1205–1223 (2009). https://doi.org/10.1137/070696817
40. Pagès, G.: A space quantization method for numerical integration. J. Comput. Appl. Math. **89**(1), 1–38 (1998). https://doi.org/10.1016/S0377-0427(97)00190-8. **(ISSN 0377-0427)**
41. Pflug, G., Pichler, A.: Approximations for probability distributions and stochastic optimization problems. In: Bertocchi, M., Consigli, G., Dempster, M.A.H. (eds.) Stochastic Optimization Methods in Finance and Energy Volume 163 of International Series in Operations Research & Management Science, chapter 15, pp. 343–387. Springer, New York (2011). https://doi.org/10.1007/978-1-4419-9586-5. **(ISBN 978-1-4419-9586-5)**
42. Pflug, Ch. G., Pichler. A.: Multistage Stochastic Optimization. Springer Series in Operations Research and Financial Engineering. Springer, Berlin (2014). https://doi.org/10.1007/978-3-319-08843-3. https://books.google.com/books?id=q_VWBQAAQBAJ. **(ISBN 978-3-319-08842-6)**

43. Pflug, G.C., Pichler, A.: From empirical observations to tree models for stochastic optimization: convergence properties. SIAM J. Optim. **26**(3), 1715–1740 (2016). https://doi.org/10.1137/15M1043376
44. Pflug, Ch G., Wozabal, D.: Ambiguity in portfolio selection. Quant. Finance **7**(4), 435–442 (2007). https://doi.org/10.1080/14697680701455410
45. Pichler, A.: The natural Banach space for version independent risk measures. Insur. Math. Econ. **53**(2), 405–415 (2013). https://doi.org/10.1016/j.insmatheco.2013.07.005
46. Pichler, A., Shapiro, A.: Minimal representations of insurance prices. Insur. Math. Econ. **62**, 184–193 (2015). https://doi.org/10.1016/j.insmatheco.2015.03.011
47. Pólik, I., Terlaky, T.: A survey of the S-lemma. SIAM Rev. **49**, 371–418 (2007). https://doi.org/10.2307/20453987
48. Prokhorov, Y.V.: Convergence of random processes and limit theorems in probability theory. Theory Probab. Appl. **1**, 157–214 (1956)
49. Rachev, S.T.: Probability Metrics and the Stability of Stochastic Models. Wiley, West Sussex (1991). http://books.google.com/books?id=5grvAAAAMAAJ
50. Rachev, S.T., Römisch, W.: Quantitative stability in stochastic programming: the method of probability metrics. Math. Oper. Res. **27**(4), 792–818 (2002). https://doi.org/10.1287/moor.27.4.792.304
51. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
52. Römisch, W.: Stability of stochastic programming problems. In: Ruszczyński, A., Shapiro, A. (eds.) Stochastic Programming, Handbooks in Operations Research and Management Science, volume 10, chapter 8. Elsevier, Amsterdam (2003)
53. Shapiro, A.: On Kusuoka representation of law invariant risk measures. Math. Oper. Res. **38**(1), 142–152 (2013). https://doi.org/10.1287/moor.1120.0563
54. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming. MOS-SIAM Series on Optimization, 2nd edn. SIAM, Philadelphia (2014). https://doi.org/10.1137/1.9780898718751
55. Skorokhod, A.V.: Basic Principles and Applications of Probability Theory. Springer, New York (1989)
56. Sun, H., Xu, H.: Convergence analysis for distributionally robust optimization and equilibrium problems. Math. Oper. Res. **41**, 377–401 (2015)
57. Villani, C.: Topics in Optimal Transportation, volume 58 of Graduate Studies in Mathematics. American Mathematical Society, Providence (2003). https://doi.org/10.1090/gsm/058. **(ISBN 0-821-83312-X)**
58. Wiesemann, W., Kuhn, D., Sim, M.: Distributionally robust convex optimization. Oper. Res. **62**, 1358–1376 (2014)
59. Xu, H., Liu, Y., Sun, H.: Distributionally robust optimization with matrix moment constraints: Lagrange duality and cutting plane method (2017)
60. Žáčková, J.: On minimax solutions of stochastic linear programming problems. Časopis pro pěstování mathematiky **91**, 423–430 (1966)
61. Zhang, J., Xu, H., Zhang, L.W.: Quantitative stability analysis for distributionally robust optimization with moment constraints (2016)
62. Zhang, J., Xu, H., Zhang, L.W.: Quantitative stability analysis of stochastic quasi-variational inequality problems and applications (2017)
63. Zhao, C., Guan, Y.: Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics. Optim. Online. http://www.optimization-online.org/DB_HTML/2015/07/5014.html
64. Zhao, C., Guan, Y.: Data-driven risk-averse stochastic optimization with wasserstein metrics. Optimization Online. http://www.optimization-online.org/DB_HTML/2015/05/4902.html
65. Zolotarev, V.M.: Probability metrics. Teoriya Veroyatnostei i ee Primeneniya **28**, 264–287 (1983)
66. Zymler, S., Kuhn, D., Rustem, B.: Distributionally robust joint chance constraints with second-order moment information. Mathem. Program. **137**, 167–198 (2013)