Noname manuscript No.
(will be inserted by the editor)

Quantitative Stability Analysis for Minimax Distributionally Robust Risk Optimization

Alois Pichler · Huifu Xu

January 11, 2017

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 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 it 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 centres and difference of their radius. The findings allow us to strengthen some recent convergence results on distributionally robust optimization where the centre of the Wasserstein ball is constructed by the empirical probability distribution.

Keywords: Distortion risk measure, ζ-ball, Wasserstein ball, quantitative stability analysis

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 affects the optimal value and optimal 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

Funding

A. Pichler
Fakultät für Mathematik
Technische Universität Chemnitz
Chemnitz, Germany
alois.pichler@mathematik.tu-chemnitz.de

H. Xu
School of Mathematical Sciences
University of Southampton
Southampton, United Kingdom
H.Xu@soton.ac.uk

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. In a recent monograph, Pflug and Pichler (2014, 2011) present 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 (2015) find that using the Wasserstein metric they can construct a ball in the space of (multivariate and non-discrete) probability distributions centered at the uniform distribution on the training samples, and look for decisions that perform best in view of the worst-case distribution within this Wasserstein ball. They demonstrate that, under mild assumptions, the distributionally robust optimization problems over Wasserstein balls can in fact be reformulated as a finite convex program in a number of practically interesting cases even as tractable linear programs. Similar models and analysis are presented by Zhao and Guan (2015b).

Along this direction, Gao and Kleywegt (2016) take it further to derive a dual reformulation of the corresponding distributinoally robust optimization (DRO) problem by constructing the worst-case distribution explicitly via first-order optimality conditions of the dual problem. They show that the Wasserstein ambiguity set yields a more realistic worst-case distribution with concise structure and clear interpretation. Using this structure, they prove that data-driven distributionally robust stochastic optimization problems (DRSO) can be approximated by robust programs to any accuracy, thereby many DRO problems become tractable with tools from robust optimization. Moreover, they identify necessary and sufficient conditions for the existence of a worst-case distribution, which is related to the growth rate of the objective function.

While the Wasserstein metric is unarguably an appropriate metric for quantifying the error of uncertainty approximation and its propagation in the related optimal decision-making problems, various other metrics have also been used in the literature of robust optimization. For instance, Sun and Xu (2016) use the total variation metric to quantify the approximation of the ambiguity set defined through moment conditions. A key step is to establish Hoffman's lemma for a general class of moment problems and use it to derive qualitative convergence analysis of the related distributionally robust optimization and equilibrium problem. Zhang et al. (2015a) extend the research by deriving Hoffman's lemma for a generic cone constrained moment system which is decision dependent and establish quantitative stability analysis for one stage DRO with such moment constraints.

Römisch (2003, p. 487) establishes the term *ζ-structure* in stochastic optimization for certain semi-norms, while Zhao and Guan (2015a) seem to be the first to use the ζ-metric to construct an ambiguity set in DRO. Specifically, they use iid samples of the true unknown probability distribution to construct a nominal distribution and then a ζ-ball centred at the nominal 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 $c:\mathbb{R}^n\times\mathbb{R}^k\times\mathbb{R}^m\to\mathbb{R}$ is a continuous function, $\xi:\Omega\to\Xi\subset\mathbb{R}^k$ is a vector of random variables defined on a measurable space $(\Omega,\mathcal{F})$ with support set Ξ, Y is a closed set in $\mathbb{R}^n$ and $Z:\Xi\times Y\rightrightarrows\mathbb{R}^m$ is a set-valued mapping, $\mathcal{P}$ is a set of probability measures and $\mathcal{R}_{S;P}$ is a risk measure parameterized

by S and $P \in \mathcal{P}$. 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*.

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

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

which is risk neutral.

A great deal of research in the literature of robust optimization to date is devoted to developing tractable numerical methods for solving distributionally robust formulations of one stage and two 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. Pólik and Terlaky (2007)) or dual methods, cf. Zymler et al. (2013) or Wiesemann et al. (2014). 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. (2014) for a comprehensive discussion.

Another important approach pioneered by Pflug and Wozabal (2007) 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 (2014) 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. (2015) 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. Kelley (1960)). 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 centre 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 centres and the difference of their radius, see Theorem 1.
- 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 2. 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 5). Differing from the stability results established for (DRO) which is under 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 Section 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.,

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

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

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

By convention, we use $\mathbb{R}^n$ to denote 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 $\mathscr{P}(\Xi)$. For vector $x \in \mathbb{R}^n$, we use $\|x\|$ and $\|x\|_p$ to denote the Euclidean norm, p-norm and $|x|$ the vector with each component of x being replaced by its absolute value.

Outline. The rest of the paper are organized as follows. In Section 2, we introduce the definition of ζ-balls and discuss changes of the ball as its centre 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. Sections 3–4 set out stability analysis for the DRO and DRRO models. Section 3 is focused on the DRO model under ζ-metric and Section 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 $\mathscr{B}$ be the Borel sigma algebra in $\mathbb{R}^k \cap \Xi$ 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 (2006); Gibbs and Su (2002). 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

$$\mathsf{dl}_{\mathscr{G}}(P, Q) := \sup_{g\in\mathscr{G}} \left|\mathbb{E}_P[g(\xi)] - \mathbb{E}_Q[g(\xi)]\right|. \tag{1}$$

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 (1991b) or Zolotarev (1983). 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

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

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.,

$$\mathscr{G} = \left\{ g : \sup_{\xi \in \Xi} |g(\xi)| \le 1,\ g \text{ is Lipschitz continuous and the Lipschtiz modulus } L_1(g) \le 1 \right\}, \quad (2)$$

where $L_1(g) := \sup\{|g(u) - g(v)|/d(u, v) :\ u \ne v\}$, then the resulting metric is known as *bounded Lipschitz metric*, denoted by dl_{BL}. If the boundedness of g is lifted in (2), that is,

$$\mathscr{G} = \{ g : g \text{ is Lipschitz continuous and Lipschtiz modulus } L_1(g) \le 1 \}, \quad (3)$$

then we arrive at Kantorovich metric,[1] denoted by dl_K. If we relax the Lipschitz continuity in (3), that is,

$$\mathscr{G} = \left\{ g : g \text{ is Lipschitz continuous and } L_q(g) \le 1 \right\}$$

with

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

where $\|\cdot\|$ denotes the Eucledian norm, then we obtain *Fortet-Mourier metric*, denoted by dl_{FM}. If

$$\mathscr{G} = \{ g : g := \mathbb{I}_{(-\infty, t]}(\cdot),\ t \in \mathbb{R}^n \},$$

where

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

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

Remark 1 From the definition we can see immediately that the ζ-metric is the coarsest metric among all metrics of ζ-structure listed above in the sense that $\mathsf{dl}_{\mathscr{G}}(P, Q)$ is greater than $\mathsf{dl}_{TV}(P, Q)$, $\mathsf{dl}_K(P, Q)$, $\mathsf{dl}_{FM}(P, Q)$ or $\mathsf{dl}_U(P, Q)$. Moreover, it is evident that $\mathsf{dl}_{TV}(P, Q) \le 2$ and when Ξ is bounded $\mathsf{dl}_K(P, Q) \in [0, \mathrm{diam}(\Xi)]$, see Gibbs and Su (2002). Moreover, it follows by Zhao and Guan (2015a, Lemmas 1–4), $\mathsf{dl}_{BL}(P, Q) \le \max\{\mathsf{dl}_K(P, Q), \mathsf{dl}_{TV}(P, Q)\}$, $\mathsf{dl}_{FM}(P, Q) \le \max\{1, \mathrm{diam}(\Xi)^{q-1}\}\, \mathsf{dl}_K(P, Q)$ and $\mathsf{dl}_U(P, Q) \le \frac{1}{2}\mathsf{dl}_{TV}(P, Q)$.

[1] In some references, it is called Wasserstein metric or Kantorovich-Wasserstein metric, see commentary by Villani Villani (2003). Here we call it Kantorovich metric to distinguish it from Wasserstein metric to be defined later on.

2.2 Hömander's theorem

Based on the ζ-metric $\mathsf{dl}_{\mathscr{G}}$, we can define the distance from a point to a set, 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 (1977, 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 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 $h:\Xi\to\mathbb{R}$. An important property of Kantorovich's metric is that it metrizes weak convergence of probability measures when the support set is bounded, that is, $\{P_N\}$ converges to P weakly if and only if $\mathsf{dl}_K(P_N,P)\to 0$ (cf. Gibbs and Su (2002)).

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 (1989) for the notion and Billingsley (1968) for a similar notion called relative compactness. By the well-known Prokhorov's theorem (see Athreya and Lahiri (2006)), 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 (1956, Theorem 1.12).

Proposition 1 (Cf. Hörmander (1955)) *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*

$$\mathbb{D}(\mathcal{P},\mathcal{Q};\mathsf{dl}_{\mathscr{G}})=\sup_{h\in\mathscr{G}}s_{\mathcal{P}}(h)-s_{\mathcal{Q}}(h),\tag{4}$$

and

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

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 Since $\mathcal{Q}$ is weakly compact, it follows by Fan (1953, Theorem 2)

$$\begin{aligned}\mathbb{D}(\mathcal{P},\mathcal{Q};\mathsf{dl}_{\mathscr{G}}) &= \sup_{P\in\mathcal{P}} \mathsf{dl}_{\mathscr{G}}(P,\mathcal{Q}) = \sup_{P\in\mathcal{P}}\inf_{Q\in\mathcal{Q}}\sup_{g\in\mathscr{G}}\int g\mathrm{d}P - \int g\mathrm{d}Q\\ &= \sup_{P\in\mathcal{P}}\sup_{g\in\mathscr{G}}\inf_{Q\in\mathcal{Q}}\int g\mathrm{d}P - \int g\mathrm{d}Q = \sup_{P\in\mathcal{P}}\sup_{g\in\mathscr{G}}\int g\mathrm{d}P - \sup_{Q\in\mathcal{Q}}\int g\mathrm{d}Q\\ &= \sup_{g\in\mathscr{G}} s_{\mathcal{P}}(g) - s_{\mathcal{Q}}(g). \end{aligned} \tag{6}$$

This shows (4). Likewise, since $\mathcal{P}$ is weakly compact, we have

$$\mathbb{D}(\mathcal{Q},\mathcal{P};\mathsf{dl}_{\mathscr{G}}) = \sup_{g\in\mathscr{G}} s_{\mathcal{Q}}(g) - s_{\mathcal{P}}(g). \tag{7}$$

A combination of (6) and (7) gives rise to (5). □

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

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

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 Section 2.5, we discuss how the Wasserstein ball of probability measures defined in finite dimensional space may be weakly compact.

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 centred at some probability measure with specified radius. In practice, the probability measure at the centre is known as nominal distribution which may be approximated through empirical data or its smooth approximation (kernel density approximation).

Definition 2 (The ζ-ball) Let $P \in \mathscr{P}(\Xi)$ and $\mathscr{G}$ 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:

$$\mathcal{B}(P,r) := \{P' \in \mathscr{P}(\Xi) : \mathsf{dl}_{\mathscr{G}}(P',P) \le r\}, \tag{8}$$

where $\mathsf{dl}_{\mathscr{G}}(\cdot,\cdot)$ is defined in (1).

In what follows, we quantify the change of the ζ-ball as its centre and radius vary. To this end, we discuss the properties of 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 (1). Then*

$$\mathsf{dl}_{\mathscr{G}}(P, tQ_1 + (1-t)Q_2) \le t\,\mathsf{dl}_{\mathscr{G}}(P,Q_1) + (1-t)\,\mathsf{dl}_{\mathscr{G}}(P,Q_2) \qquad \forall\, t\in[0,1] \tag{9}$$

and

$$\mathsf{dl}_{\mathscr{G}}(P,Q_2) \le \mathsf{dl}_{\mathscr{G}}(P,Q_1) + \mathsf{dl}_{\mathscr{G}}(Q_1,Q_2). \tag{10}$$

Proof We only show the first inequality as the second one can be proved analogously. Since

$$\mathbb{E}_{t\,Q_1+(1-t)Q_2}[g(\xi)] = t\,\mathbb{E}_{Q_1}[g(\xi)] + (1-t)\,\mathbb{E}_{Q_2}[g(\xi)],$$

by definition

$$\begin{aligned}\mathsf{dl}_{\mathscr{G}}(P, tQ_1+(1-t)Q_2) &\le \sup_{g\in\mathscr{G}}\Big[t\,\Big|\mathbb{E}_P[g(\xi)]-\mathbb{E}_{Q_1}[g(\xi)]\Big| + (1-t)\,\Big|\mathbb{E}_P[g(\xi)]-\mathbb{E}_{Q_2}[g(\xi)]\Big|\Big]\\ &\le t\,\mathsf{dl}_{\mathscr{G}}(P,Q_1) + (1-t)\,\mathsf{dl}_{\mathscr{G}}(P,Q_2) \qquad \forall\, t\in[0,1].\end{aligned}$$

This gives rise to (9) as desired. □

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 (1). For $t\in[0,1]$, let*

$$h(t) := \mathsf{dl}_{\mathscr{G}}(P, tQ_1+(1-t)Q_2).$$

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

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

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 (1970, Corollary 10.1.1), $h(\cdot)$ is continuous over $[0,1]$. The rest are 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 centred 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 centre and radius.

Theorem 1 (Quantitative stability of the ζ-ball) *Let $\mathcal{B}(P,r)$ be the ζ-ball defined as in (8). For every $P, Q\in\mathscr{P}(\Xi)$ and $r_1, r_2\in\mathbb{R}_+$ it holds that*

$$\mathbb{H}(\mathcal{B}(P,r_1), \mathcal{B}(Q,r_2); \mathsf{dl}_{\mathscr{G}}) \le \mathsf{dl}_{\mathscr{G}}(P,Q) + |r_2-r_1|, \tag{11}$$

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

Proof Let $P'\in\mathcal{B}(Q,r_2)\backslash\mathcal{B}(P,r_1)$ and $\hat\lambda := r_1/\mathsf{dl}_{\mathscr{G}}(P',P)$. By the definition of the ζ-ball, $\hat\lambda\in(0,1)$. Let $\hat P := \hat\lambda P' + (1-\hat\lambda)P$. By Proposition 2 and Corollary 1,

$$\begin{aligned}\mathsf{dl}_{\mathscr{G}}(\hat P, P) &= \mathsf{dl}_{\mathscr{G}}(\hat\lambda P' + (1-\hat\lambda)P, P)\\ &\le \hat\lambda\mathsf{dl}_{\mathscr{G}}(P',P) = r_1.\end{aligned}$$

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

$$\begin{aligned}\mathsf{dl}_{\mathscr{G}}(P', \mathcal{B}(P,r_1)) &\le \mathsf{dl}_{\mathscr{G}}(P',\hat P) = \mathsf{dl}_{\mathscr{G}}(P', \hat\lambda P' + (1-\hat\lambda)P)\\ &\le (1-\hat\lambda)\mathsf{dl}_{\mathscr{G}}(P',P) = \mathsf{dl}_{\mathscr{G}}(P',P) - \hat\lambda\mathsf{dl}_{\mathscr{G}}(P',P)\\ &= \mathsf{dl}_{\mathscr{G}}(P',P) - r_1\\ &\le \mathsf{dl}_{\mathscr{G}}(P',Q) + \mathsf{dl}_{\mathscr{G}}(Q,P) - r_1\\ &\le r_2 + \mathsf{dl}_{\mathscr{G}}(Q,P) - r_1.\end{aligned} \tag{12}$$

This shows

$$\mathbb{D}(\mathcal{B}(P,r_1), \mathcal{B}(Q,r_2); \mathsf{dl}_{\mathscr{G}}) \le \mathsf{dl}_{\mathscr{G}}(P,Q) + r_2 - r_1, \tag{13}$$

The conclusion follows by swapping the role of the two balls in the proof above. □

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 centre and/or radius. A particularly interesting case is when $r_1 = 0$ and P is the unknown true probability distribution whereas 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]

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

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

$$\mathsf{dl}_{\mathscr{G}}(P, \mathcal{B}(P_N, r_N)) \le \mathsf{dl}_{\mathscr{G}}(P, P_N) + r_N, \tag{14}$$

where $\mathsf{dl}_{\mathscr{G}}$ is defined as in (1).

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

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

then for any ϵ, there exist positive constants c and C such that

$$P^N\left(\mathsf{dl}_K(P, P_N) \ge \epsilon\right) \le C\left[\exp(-cN\epsilon^k \mathbb{1}_{\epsilon\le 1}) + \exp(-cN\epsilon^{\nu}\mathbb{1}_{\epsilon>1})\right] \tag{15}$$

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 is the demsion of ξ (see Fournier and Guilline (2015)).

In the case when P is a continuous probability measure, it is well known that $\mathsf{dl}_{TV}(P, P_N) = 1$. Zhao and Guan (2015a) proposed to replace P_N with its Kernel Density Estimation (KDE). For this let h_N be a sequence of positive constants converging to zero and $\Phi(\cdot)$ be a measurable kernel function with $\Phi(\cdot) \ge 0$, $\int \Phi(\xi) d\xi = 1$; the KDE of P_N is defined as

$$f_N(z) = \frac{1}{Nh_N^k}\sum_{i=1}^{N} \Phi\left(\frac{z - \xi_i}{h_N}\right). \tag{16}$$

A simple example for $\Phi(\cdot)$ is the standard normal density function. Let $\tilde{P}_N$ be a probability measure with kernel density $f_N(z)$. Under some moderate conditions, Zhao and Guan established bounds for $\mathsf{dl}_{\mathscr{G}}(P, \tilde{P}_N)$ under a range of metrics with ζ-structure including $\mathsf{dl}_{TV}, \mathsf{dl}_K, \mathsf{dl}_{FM}, \mathsf{dl}_{BL}$ and dl_U, see Zhao and Guan (2015a, Proposition 4). Using the corollary above and the proposition, we can easily derive the rate of convergence for $\mathsf{dl}_{\mathscr{G}}(P, \mathcal{B}(P_N, r_N))$ as N increases and r_N decreases.

Note that inequality (14) 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 (2011) and references therein.

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

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.

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

$$\mathsf{d}_r(P, \tilde{P}) = \left(\inf_{\pi} \iint d(\xi, \xi)^r \, \pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}) \right)^{1/r}, \tag{17}$$

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

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

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

One of the main results concerning the Wasserstein distance is the Kantorovich–Rubinshtein Theorem (Kantorovich and Rubinshtein (1958)) which characterizes the Kantorovich metric of two probability measures by the maximum difference of the expected value of two measures over the whole class of Lipschitz continuous functions with modulus being bounded by 1, that is,

$$\mathsf{d}_1(P, \tilde{P}) = \sup \mathbb{E}_P h - \mathbb{E}_{\tilde{P}} h = \mathsf{dl}_K(P, \tilde{P}), \tag{18}$$

where the supremum is taken over all Lipschitz functions h with $h(\xi) - h(\tilde{\xi}) \leq \mathsf{d}(\xi, \tilde{\xi})$. The latter recovers our definition of the Kantorovich metric in Section 2.1 and it is known as the dual representation of the metric.

The identity (18) recovers the metric d_1 as metric with ζ-structure, but the general Wasserstein metric for order $r > 1$ is not of this type.

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 it can be interpreted as the minimal transportation cost of moving good 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 Rachev (1991a). 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 (2003).

Here we establish some technical results such as weak compactness for the set of probability measures defined through the Wasserstein metric for further reference later on.

Proposition 3 *Let $\mathcal{B}_\rho = \{P \in \mathscr{P}(\mathbb{R}) : \mathsf{d}(P, P_0) \leq \rho\}$ be the Wasserstein ball of diameter ρ with center P_0 on the real line, induced by the distance $d(x, y) := |y - x|$. Then there does not exist a finite set $\mathcal{Q}$ of probability measures such that $\mathbb{D}(\mathcal{B}_\rho, \mathcal{Q}) < \rho$.*

Proof We may assume without loss of generality that $\rho = 1$ and $P_0 = \delta_0$. Suppose there were a finite set of measures $\mathcal{Q}$ such that $\mathbb{D}(\mathcal{B}_1, \mathcal{Q}) \leq r < 1$. Consider the measures $P_n := (1 - 2^{-n})\delta_0 + 2^{-n}\delta_{2^n}$ for $n = 1, 2, \ldots$ on the real line. These measures are in the unit ball of δ_0, as $\mathsf{d}(\delta_0, P_n) = 1$. Their mutual Wasserstein distance is $\mathsf{d}(P_n, P_m) = 2 - 2 \cdot 2^{-|m-n|}$.

Fix $Q \in \mathcal{Q}$ and choose n such that $\mathsf{d}(Q, P_n) \leq r$. By the reverse triangle inequality it holds that $\mathsf{d}(Q, P_m) \geq \mathsf{d}(P_n, P_m) - \mathsf{d}(P_n, Q) \geq 2 - 2^{1-|m-n|} - r$. Whenever m is large enough it holds that $\mathsf{d}(Q, P_{m'}) \geq 1 > r$ for infinitely many $m' \geq m$. One may repeat this procedure now for every $Q \in \mathcal{Q}$ with the result that there remains $m \in \mathbb{N}$ such that $\mathsf{d}(Q, P_{m'}) \geq r$ for infinitely many $m' \geq m$ and all $Q \in \mathcal{Q}$. This contradicts the assumption, a finite set $\mathcal{Q}$ with the desired property $\mathbb{D}(\mathcal{B}_\rho, \mathcal{Q}) < \rho$ hence does not exist. □

Proposition 4 *Let $\mathscr{P}(\mathbb{R}^m)$ denote the set of all probability measures on $\mathbb{R}^m$ and let $d(\cdot,:) = \mathsf{d}_r(\cdot,:)$ be the Wasserstein distance of order $r \geq 1$. If $\mathcal{P}$ is tight, then the ρ-enlargement under the Wasserstein metric $\mathcal{P}_\rho := \{Q \in \mathscr{P}(\mathbb{R}^m) : \mathsf{d}_r(Q,\mathcal{P}) < \rho\}$ is tight as well for every $\rho \geq 0$ and $\{Q \in \mathscr{P}(\mathbb{R}^m) : \mathsf{d}_r(Q,\mathcal{P}) \leq \rho\}$ is weakly compact.*

Proof Let $K^\varepsilon \subset \mathbb{R}^m$ be a compact set such that $P(K^\varepsilon) > 1-\varepsilon$ for every $P \in \mathcal{P}$. Define the set $K^\varepsilon_{\rho/\varepsilon} := \{x \in \mathbb{R}^m : \|x-y\| < \rho/\varepsilon \text{ for some } y \in K^\varepsilon\}$ and note that the enlargement $K^\varepsilon_{\rho/\varepsilon}$ is compact. We shall show that $Q(K^\varepsilon_{\rho/\varepsilon}) > 1-2\varepsilon$ for every $Q \in \mathcal{P}_\rho$ by contraposition.

Indeed, assume that $Q(K^{\varepsilon\,\complement}_{\rho/\varepsilon}) \geq 2\varepsilon$ ($K^{\varepsilon\,\complement}_{\rho/\varepsilon}$ is the complement of $K^\varepsilon_{\rho/\varepsilon}$). Pick $P \in \mathcal{P}$ such that $\mathsf{d}_r(P,Q) < \rho$ and let π be a transport plan for P and Q such that $\mathsf{d}(P,Q)^r = \iint \|y-x\|^r \,\pi(\mathrm{d}x,\mathrm{d}y)$. Consider the set $A := K^\varepsilon \times K^{\varepsilon\,\complement}_{\rho/\varepsilon}$ and note that $\|y-x\| > \frac{\rho}{\varepsilon}$ whenever $(x,y) \in A$. Further it holds that

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

The Wasserstein distance then satisfies

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

which is a contradiction. Hence it holds that $Q(K^\varepsilon_{\rho/\varepsilon}) > 1-2\varepsilon$ for every $Q \in \mathcal{P}_\rho$ and $\mathcal{P}_\rho$ thus is tight.

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

Corollary 2 *The Wasserstein-ball $B_\rho(P) := \{Q \in \mathscr{P}(\mathbb{R}^m) : \mathsf{d}_r(P,Q) < \rho\}$ is tight and $\{Q : \mathsf{d}_r(P,Q) \leq \rho\}$ is weakly compact.*

Cf. Villani (2003, Theorem 7.12 (ii)).

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 increasing information about the true probability distribution or need for numerical approximation of the distributionally robust optimization problem, see discussions in Sun and Xu (2016) and Zhang et al. (2015a). This kind of research may be viewed as an extension of classical stability analysis in stochastic programming (see Römisch (2003)).

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

$$\inf_{y\in Y} \sup_{P\in\tilde{\mathcal{P}}} \mathbb{E}_P\left[\inf_{z\in Z(y,\xi)} c(y,\xi,z)\right] \tag{19}$$

of problem (DRO), where $\tilde{\mathcal{P}}$ is a perturbation of $\mathcal{P}$. Let $\vartheta(\tilde{\mathcal{P}})$ and $\vartheta(\mathcal{P})$ denote respectively the optimal value of (19) and (DRO), and $Y^*(\tilde{\mathcal{P}})$ and $Y^*(\mathcal{P})$ denote the respective set of optimal solutions. The following theorem states the relationship of these quantities.

Theorem 2 (Quantitative stability of the DRO problem) *Let*

$$v(y,\xi) := \inf_{z\in Z(y,\xi)} c(y,\xi,z) \tag{20}$$

and $\mathscr{H} := \{v(y,\cdot) : y \in Y\}$. *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) $|\vartheta(\tilde{\mathcal{P}}) - \vartheta(\mathcal{P})| \le \mathbb{H}(\tilde{\mathcal{P}}, \mathcal{P}; \mathsf{dl}_{\mathscr{H}})$, *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*

$$|\vartheta(\tilde{\mathcal{P}}) - \vartheta(\mathcal{P})| \le \mathsf{dl}_{\mathscr{H}}(P,\tilde{P}) + |r' - r|. \tag{21}$$

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*

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

then

$$\mathbb{D}(Y^*(\tilde{\mathcal{P}}), Y^*(\mathcal{P})) \le \sqrt{\frac{3}{\upsilon}\mathbb{H}(\tilde{\mathcal{P}}, \mathcal{P}; \mathsf{dl}_{\mathscr{G}})}, \tag{22}$$

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

$$\begin{aligned} |\vartheta(\tilde{\mathcal{P}}) - \vartheta(\mathcal{P})| &\le \sup_{y\in Y}\left|\sup_{P\in\tilde{\mathcal{P}}} \mathbb{E}_P[v(y,\xi)] - \sup_{P\in\mathcal{P}} \mathbb{E}_P[v(y,\xi)]\right| \\ &\le \sup_{h\in\mathscr{H}} |s_{\tilde{\mathcal{P}}}(h) - s_{\mathcal{P}}(h)|. \end{aligned} \tag{23}$$

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

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

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 $\frac{1}{\kappa}$ to Lipschitz continuous with modulus being bounded by 1. This will allow us to tighten the estimation in (24) by

$$\sup_{h\in\mathscr{H}} |s_{\tilde{\mathcal{P}}}(h) - s_{\mathcal{P}}(h)| \le \kappa \sup_{g\in\mathscr{G}_1} |s_{\tilde{\mathcal{P}}}(g) - s_{\mathcal{P}}(g)| = \kappa\, \mathbb{H}(\tilde{\mathcal{P}}, \mathcal{P}; \mathsf{dl}_K), \tag{24}$$

where $\mathscr{G}_1$ denotes all Lipschitz continuous functions defined over Ξ with modulus being bounded by 1. 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 (24), we obtain (22) by virtue of (Liu and Xu, 2013, Lemma 3.8). □

Compared to the existing results on stability analysis for DRO problems (see, e.g., Zhang et al. (2015a)), Theorem 2 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 and (ii) when the ambiguity set is structured via ζ-ball, the variation of the optimal value is bounded by the distance of the centres of the balls and the difference of their radius. 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 2 and inequality (14) to derive the rate of convergence of $\vartheta(\mathcal{B}(P_N, r_N))$ as the sample size increases, see (15).

As for the stability of optimal solutions, we note that the second order growth condition may be fulfilled if there exists a positive function $\alpha(\xi)$ with $\inf_{P\in\mathcal{P}} \mathbb{E}_P[\alpha(\xi)] > 0$ such that

$$v(y,\xi) \ge \vartheta(\mathcal{P}) + \alpha(\xi) d(y, Y^*(\mathcal{P}))^2 \qquad \forall\, y \in Y, \xi \in \Xi. \tag{25}$$

A sufficient condition for the inequality above is that for any fixed $y \in Y$,

$$v(y',\xi) \ge v(y,\xi) + \alpha(\xi)\|y' - y\|^2 \qquad \forall\, y' \in Y, \xi \in \Xi. \tag{26}$$

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

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

There is a vast literature on robust optimization, see the monograph Ben-Tal et al. (2009) for a comprehensive overview of the model, numerical methods and applications. Unfortunately, Theorem 2 does not cover this important case. Below, we make a separate statement about stability of the problem (RO) by comparing the value, even for different objectives v and $\tilde{v}$.

Theorem 3 (Quantitative stability of problem (RO)) *Assume that*

$$v(y,\xi) - \tilde{v}(\tilde{y},\tilde{\xi}) \le L_Y \cdot d(y,\tilde{y}) + L_\Xi \cdot d(\xi,\tilde{\xi}). \tag{27}$$

Then the robust optimization problem is bounded with respect to the decision sets and the support, i.e.,

$$\inf_{y\in Y} \sup_{\xi\in\Xi} v(y,\xi) - \inf_{\tilde{y}\in\tilde{Y}} \sup_{\tilde{\xi}\in\tilde{\Xi}} \tilde{v}(y,\tilde{\xi}) \le L_\Xi \cdot \mathbb{D}(\Xi,\tilde{\Xi}) + L_Y \cdot \mathbb{D}(\tilde{Y}, Y). \tag{28}$$

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

$$\left|v(y,\xi) - v(\tilde{y},\tilde{\xi})\right| \le L_Y \cdot d(y,\tilde{y}) + L_\Xi \cdot d(\xi,\tilde{\xi}), \tag{29}$$

then

$$\left|\inf_{y\in Y}\sup_{\xi\in\Xi} v(y,\xi) - \inf_{\tilde{y}\in\tilde{Y}}\sup_{\tilde{\xi}\in\tilde{\Xi}} v(y,\tilde{\xi})\right| \le L_Y\cdot\mathbb{H}(Y,\tilde{Y}) + L_\Xi\cdot\mathbb{H}(\Xi,\tilde{\Xi}). \tag{30}$$

Proof By taking the infimum in (27) with respect to $\tilde{\xi}\in\tilde{\Xi}$ it follows that

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

and consequently

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

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 (28), 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 deriving the stability result. In what follows, we give a sufficient condition for this.

Consider the feasible set-valued mapping $Z: Y\times\Xi\rightrightarrows\mathbb{R}^m$ 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 Slater condition w.r.t. the variable z, see Zhang et al. (2015b, 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 5 (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*

$$|c(y,\xi,z) - c(y_0,\xi_0,z_0)| \le L_c[d(y,y_0) + d(\xi,\xi_0)^\beta + d(z,z_0)] \tag{31}$$

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

$$|v(y,\xi) - v(y_0,\xi_0)| \le L\left[d(y,y_0) + d(\xi,\xi_0) + d(\xi,\xi_0)^\beta\right] \tag{32}$$

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

Proof The conclusion follows directly from (Klatte, 1987, Theorem 1). □

It is important to note that the error bound in (32) 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.

Note that in the case when $\tilde{Y} = Y$, (30) 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.,

$$\left| v(y,\xi) - v(y,\tilde{\xi}) \right| \le L_\Xi \cdot d(\xi,\tilde{\xi}) \qquad \forall\, y \in Y. \tag{33}$$

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

4 Stability of the DRRO problem

We now move on to discuss stability of 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 Y be a random variable. Recall that the value at risk at level $\alpha \in [0,1)$ is defined as

$$\mathsf{V@R}_\alpha(Y) := \inf\{y \in \mathbb{R} : P(Y \le y) \ge \alpha\},$$

where P is the probability distribution of Y. It is well known that the $\mathsf{V@R}_\alpha$ 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

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

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

Definition 4 Let $\sigma : \mathbb{R} \to \mathbb{R}_+$ be a nonnegative, nondecreasing function with

$$\int_0^1 \sigma(u)\mathrm{d}u = 1.$$

We call

$$\mathcal{R}_\sigma(Y) := \int_0^1 \sigma(u)\mathsf{V@R}_\alpha(Y)\mathrm{d}\alpha \tag{34}$$

the *distortion risk measure* of Y associated with the *distortion functional* σ.

Clearly $\mathcal{R}_\sigma(Y)$ is a weighted average of the value at risk and the average value at risk is a special distortion risk measure because

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

with

$$\sigma_\alpha(t) = \begin{cases} 0 & \text{if } t \in [0,\alpha], \\ \frac{1}{1-\alpha} & \text{if } t \in [\alpha,1]. \end{cases} \tag{35}$$

For a set S of distortion functionals, we can define

$$\mathcal{R}_{S;P}(Y) := \sup_{\sigma\in S} \mathcal{R}_\sigma(Y), \tag{36}$$

where P is the probability measure.

The following result is a combination of the well known Kusuoka representation theorem and its implication in terms of the connection with the distortion risk measure, see Pflug and Pichler (2014, Theorem 3.13, Corollary 3.14).

Theorem 4 (Kusuoka representation theorem) *Let $(\Omega, \mathcal{F}, P)$ be a nonatomic probability space and L^∞ be a set of random variables mapping from Ω to $\mathbb{R}$. Let $\rho : L^\infty \to \mathbb{R}$ be a law invariant coherent risk measure. Then there exists a set of probability measures $\mathcal{M}$ on $[0, 1)$ equiped with Borel sigma algebra such that*

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

Moreover ρ has the representation

$$\rho(Y) = \sup_{\sigma\in S} \mathcal{R}_\sigma(Y),$$

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

The theorem means that any law invariant coherent risk measure is the maximum of distortion risk measure over a certain set of distortion risk functionals. In what follows, we assume that the risk measure $R_{S;P}$ in problem (DRRO) is given by (36).

By Shapiro (2013); Pichler and Shapiro (2015) (cf. also Pichler (2013, Theorem 12)), the risk measure has a dual representation, that is, there exists a dual variable Z such that

$$\mathcal{R}_{S;P}(Y) = \mathbb{E}_P YZ. \tag{37}$$

It is important to note that Z depends on P, therefore we cannot replace $\sup_{P\in\mathcal{P}} \mathcal{R}_{S;P}(Y)$ with $\sup_{P\in\mathcal{P}} \mathbb{E}_P YZ$, which means our stability results on DRO in the previous section cannot be directly applied to the DRRO model.

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 the ambiguity set but also 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.

Proposition 6 *Let Y and $\tilde{Y}$ be two random variables. Assume that there are positive constants L and β such that*

$$Y(\xi) - \tilde{Y}(\tilde{\xi}) \le L \cdot \left(d(\xi, \tilde{\xi}) + d(\xi, \tilde{\xi})^\beta \right). \tag{38}$$

Then

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

where $\mathcal{R}_{S;P}$ is the risk functional induced by S as defined in (36), $q \ge 1$ is the Hölder conjugate exponent to $p \ge 1$, $\frac{1}{p} + \frac{1}{q} = 1$ and d_r is the Wasserstein distance of order $r \ge 1$ (see (17)).

Proof We shall employ the dual representation of risk measures, which can be derived from Shapiro (2013) (cf. also Pichler (2013, Theorem 12)) . Let Z ($\tilde{Z}$, resp.) be dual variables so that $\mathcal{R}_{S;P}(Y) = \mathbb{E}_P YZ$ ($\mathcal{R}_{S;\tilde{P}}(\tilde{Y}) = \mathbb{E}_{\tilde{P}}\tilde{Y}\tilde{Z}$, resp.). Let π be a bivariate probability measure with marginals P and $\tilde{P}$. Note that Y and Z, as well as $\tilde{Y}$ and $\tilde{Z}$ are coupled in a comonotone way, so that we have by Hoeffding's Lemma (cf. Hoeffding (1940))

$$\begin{aligned}\mathcal{R}_{S;P}(Y) - \mathcal{R}_{S;\tilde{P}}(\tilde{Y}) = \mathbb{E}_P YZ - \mathbb{E}_{\tilde{P}}\tilde{Y}\tilde{Z} &= \iint [Y(\xi)Z(\xi) - \tilde{Y}(\tilde{\xi})\tilde{Z}(\tilde{\xi})]\pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}) \\ &\le \iint \left(Y(\xi) - \tilde{Y}(\tilde{\xi})\right) Z(\xi)\pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}) \\ &\le L \cdot \iint Z(\xi)\left[d(\xi,\tilde{\xi}) + d(\xi,\tilde{\xi})^\beta\right]\pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}).\end{aligned}$$

The first inequality is due to the fact that $\mathbb{E}_{\tilde{P}}\tilde{Y}\tilde{Z} \ge \mathbb{E}_{\tilde{P}}\tilde{Y}Z$ noting that $\tilde{Z}$ gives the maximum value of $\mathcal{R}_{\cdot;\tilde{P}}(\tilde{Y})$ over S, see (36). By applying Hölder inequality we obtain that

$$\begin{aligned}\mathcal{R}_{S;P}(Y) - \mathcal{R}_{S;\tilde{P}}(\tilde{Y}) &\le L \cdot \left(\int Z^q \mathrm{d}\pi\right)^{\frac{1}{q}} \left(\int \left[d(\xi,\tilde{\xi}) + d(\xi,\tilde{\xi})^\beta\right]^p \pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi})\right)^{\frac{1}{p}} \\ &\le L \cdot \left(\int Z^q \mathrm{d}\pi\right)^{\frac{1}{q}} \left[\left(\int d(\xi,\tilde{\xi})^p \pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi})\right)^{\frac{1}{p}} + \left(\int d(\xi,\tilde{\xi})^{\beta p} \pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi})\right)^{\frac{1}{p}}\right],\end{aligned}$$

and by taking infimum with respect to all probability measures π with marginals P and $\tilde{P}$ we arrive at

$$\mathcal{R}_{S;P}(Y) - \mathcal{R}_{S;\tilde{P}}(\tilde{Y}) \le L \cdot \|Z\|_q \left(\mathsf{d}_p(P,\tilde{P} + \mathsf{d}_{\beta p}(P,\tilde{P})^\beta\right),$$

from which the conclusion follows, as the cdf. of Z is σ for some $\sigma \in S$. □

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

Theorem 5 (Quantitative stablity of the problem (DRRO)**)** *Let $v(y,\xi)$ be defined as in* (20)*,*

$$\vartheta := \inf_{y\in Y} \sup_{P\in\mathcal{P}} \mathcal{R}_{S;P}\left(v(y,\xi)\right)$$

and

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

The following assertions hold.

(i) If there are positive constants L_Ξ, L_Y and L_Z such that

$$|c(y,\xi,z) - \tilde{c}(\tilde{y},\tilde{\xi},\tilde{z})| \le L_\Xi\, d(\xi,\tilde{\xi})^\beta + L_Y\, d(y,\tilde{y}) + L_Z\, d(z,\tilde{z}), \tag{39}$$

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-Lipschitizan 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

$$|\tilde{\vartheta} - \vartheta| \le L \cdot \left\{\sup_{\sigma\in S} \|\sigma\|_q \left[\mathbb{H}\left(\mathcal{P},\tilde{\mathcal{P}};\mathsf{d}_p\right) + \mathbb{H}\left(\mathcal{P},\tilde{\mathcal{P}};\mathsf{d}_{p\beta}\right)\right] + \mathbb{H}(\tilde{Y},Y)\right\}, \tag{40}$$

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 $\sup_{P\in\mathcal{P}} \mathcal{R}_{S,P}\left(v(y,\xi)\right)$ satisfies the second order growth condition at $Y^*(\mathcal{P})$, that is, there exists a positive constant υ such that*

$$\sup_{P\in\mathcal{P}} \mathcal{R}_{S;P}\left(v(y,\xi)\right) \geq \vartheta(\mathcal{P}) + \upsilon d(y, Y^*(\mathcal{P}))^2, \forall y \in Y, \tag{41}$$

then

$$\mathbb{D}(Y^*(\tilde{\mathcal{P}}), Y^*(\mathcal{P})) \leq \sqrt{\frac{3}{\upsilon}\left(\mathbb{H}\left(\mathcal{P},\tilde{\mathcal{P}};\mathsf{d}_p\right) + \mathbb{H}\left(\mathcal{P},\tilde{\mathcal{P}};\mathsf{d}_{p\beta}\right)\right)}; \tag{42}$$

(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 (14),*

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

Proof Part (i). It follows by Proposition 5 that there exists a positive constant L such that

$$|v(y,\xi) - v(\tilde{y},\tilde{\xi})| \leq L[d(y,\tilde{y}) + d(\xi,\tilde{\xi}) + d(\xi,\tilde{\xi})^\beta]. \tag{43}$$

Let $Y(\xi) := v(y,\xi)$ and

$$\tilde{Y}(\tilde{\xi}) := v(\tilde{y},\tilde{\xi}) + L\, d(y,\tilde{y}).$$

Then

$$Y(\xi) - \tilde{Y}(\tilde{\xi}) \leq L\, d(\xi,\tilde{\xi}) + d(\xi,\tilde{\xi})^\beta.$$

By Proposition 6, we have

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

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

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

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

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

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

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

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 2. We omit the details of the proof.

Part (iii) follows from Theorem 1. □

Theorem 5 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 2, Theorem 5 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 5 recovers part of Theorem 2 (with Kantorovich-Wasserstein metric).
- In (43), 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 5. When $\beta = 1$, (40) simplifies to

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

- 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. (2016) and references therein.
- An important case that Theorem 5 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 dimensionnal space, consequently we can apply some existing numerical methods in the literature such as the cutting plane method in Xu, Liu and Sun (see Xu et al. (2015, Algorithm 4.1)) to solve the problem. In Liu et al. (2016), Liu, Pichler and Xu derived an error bound for such ambiguity set and its discretization, see Liu et al. (2016, Section 3) for details. Our Theorem 5 applies to such a case when the ambiguity set in problem (DRRO) is defined and discretized in that manner.

5 An example

In this section, we present an example which illustrates our established stability results in the preceding section.

Consider a linear two stage recourse minimization problem:

$$\begin{array}{ll} \min\limits_{y \in \mathbb{R}^n} & c^\top y + \mathbb{E}_P[v(y,\xi)] \\ \text{s.t.} & Ay = b,\ y \ge 0, \end{array} \tag{44}$$

where $v(y,\xi)$ is the optimal value function of the second stage problem

$$\begin{array}{ll} \min\limits_{z \in \mathbb{R}^m} & q(\xi)^\top z \\ \text{s.t.} & T(\xi)y + Wz = h(\xi),\ z \ge 0, \end{array} \tag{45}$$

where $W \in \mathbb{R}^{r\times m}$ is a fixed recourse matrix, $T(\xi) \in \mathbb{R}^{r\times n}$ is a random matrix, and $h(\xi) \in \mathbb{R}^r$ and $q(\xi) \in \mathbb{R}^m$ are random vectors. We assume that $T(\cdot)$, $h(\cdot)$ and $q(\cdot)$ are affine functions of ξ and that Ξ is a polyhedral subset of $\mathbb{R}^s$ (for example, $\Xi = \mathbb{R}^s$). Let $Y = \{y \in \mathbb{R}^n : Ay = b, y \ge 0\}$. The following result is established in Y. Liu and Xu (2014).

Proposition 7 *Let* $\mathcal{M}(q(\xi)) := \{\pi \in \mathbb{R}^r : W^\top \pi \le q(\xi)\}$ *be nonempty for every* $\xi \in \Xi$ *and* $\phi_P(y) := \mathbb{E}_P v(y,\xi)$. *Then there exists a constant* $\hat{L}$ *such that* v *satisfies the local Lipschitz continuity property*

$$|v(y,\xi) - v(\tilde{y},\tilde{\xi})| \le \hat{L}(\max\{1, \|\xi\|, \|\tilde{\xi}\|\}^2 \|\tilde{y} - y\| + \max\{1, \|y\|, \|\tilde{y}\|\} \max\{1, \|\xi\|, \|\tilde{\xi}\|\}\|\tilde{\xi} - \xi\|) \quad (46)$$

for all pairs $(y,\xi), (\tilde{y},\tilde{\xi}) \in (Y \cap \operatorname{dom}\phi_P) \times \Xi$, *where* dom *denotes domain of a function, i.e., the set of points where the function is finite valued. Moreover,* $v(\cdot,\xi)$ *is convex for every* $\xi \in \Xi$.

To fit the example entirely into the framework of the model (DRO) and the model (DRRO), we may incorporate the term $c^T y$ into the optimal value function of the second stage problem, that is,

$$v(y,\xi) := c^T y + v(y,\xi).$$

Obviously v enjoys the same property as what we described in Proposition 7. Moreover, if Ξ and Y are bounded, then $v(y,\xi)$ is uniformly Lipschitz continuous in ξ. In that case there exists a positive constant such that the set $\mathscr{H}$ of functions defined in Theorem 2 are Lipschitz continuous with bounded modulus and hence all conditions in part (i) of the theorem are fulfilled.

To see how the second order growth condition in Part (ii) of the theorem can be fulfilled, we note that if W is invertible and $W^{-1}T(\xi)$ is uniformly positive definite, then $v(\cdot,\xi)$ is uniformly strongly convex and hence condition (26), the latter implies the desired second order condition. It might be interesting to explore weaker conditions, i.e., W is not invertible, we leave this to interested readers.

Let us now move on to verify conditions of Theorem 5. Conditions in Part (i) do not involve the risk measure. Indeed, we can bypass these conditions to get Lipschitz continuity of v in (y,ξ) from Proposition 7 when Ξ and Y are bounded.

The second order growth condition requires a bit more scrutiny as it involves the risk function $R_{S;P}$. Let us consider the case when $R_{S;P}$ is AVaR. Assume that ξ is continuously distributed with density function $\rho(u)$ defined over Ξ. Let

$$\Phi_\beta(y,\eta) := \eta + \frac{1}{1-\beta}\int_{u\in\Xi} (v(y,u) - \eta)_+ \rho(u)du = \eta + \mathbb{E}_P[(v(y,\xi) - \eta)_+], \quad (47)$$

where $(t)_+ = \max(0,t)$ and P is the probability measure of ξ. It is well known (see Rockafellar and Uryasev (2000)) that

$$\mathrm{AVaR}_\beta(v(y,\xi)) = \min_{\eta\in\mathbb{R}} \Phi_\beta(y,\eta). \quad (48)$$

Assume that the ambiguity set is constructed through ζ-ball. From our discussions in Section 4.1, we know that $\mathrm{AVaR}_\beta(v(y,\xi))$ is a special example of $R_{S;P}$ with $S = \{\sigma_\alpha\}$ where σ_α is defined as in (35).

It is easy to see that the function $\eta + (x-\eta)_+$ is increasing and convex in x for each fixed η. Thus, if $v(y,\xi)$ is uniformly strongly convex in y, then $\eta + (v(y,\xi) - \eta)_+$ is also uniformly strongly convex in y. The expectation w.r.t. probability measure P and minimization in η preserve the strong convexity. This means the second order growth condition in part (ii) of the theorem may be fulfilled when $v(y,\xi)$ is uniformly strongly convex in y.

6 Acknowledgement

We would like to thank Jie Zhang for an initial proof of Theorem 1 and Shaoyan Guo for careful reading of the paper.

References

K. B. Athreya and S. N. Lahiri. *Measure theory and probability theory*. Springer Science & Business Media, 2006. 4, 6

A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, Princeton, 2009. 13

P. Billingsley. *Convergence of Probability Measures*. John Wiley, New York, 1968. 6

C. Castaing and M. Valadier. *Convex Analysis and Measurable Multifunctions. Lecture Notes in Mathematics*. Springer, Berlin, 1977. 6

P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. preprint arXiv:1505.05116, 2015. 2

K. Fan. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39(1):42, 1953. 7

N. Fournier and A. Guilline. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015. doi:0.1007/s00440-014-0583-7. 9

R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. preprint arXiv:1604.02199, 2016. 2

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International statistical review*, 70:419–435, 2002. 4, 5, 6

W. Hoeffding. Maßstabinvariante Korrelationstheorie. *Schriften des Mathematischen Instituts der Universität Berlin*, 5:181–233, 1940. German. 17

L. Hörmander. Sur la fonction d'appui des ensembles convexes dans un espace localement convexe. *Arkiv för matematik*, 3(2):181–186, 1955. doi:10.1007/BF02589354. In French. 6

L. V. Kantorovich and G. S. Rubinshtein. On a space of totally additive functions. *Vestnik Leningradskogo Universiteta*, 13:52–59, 1958. 10

J. E. Kelley. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8:703–712, 1960. 3

D. Klatte. A note on quantitative stability results in nonlinear optimization. *Seminarbericht, Sektion Mathematik, Humboldt-Universität zu Berlin, Berlin*, 90:77–86, 1987. 15

Y. Liu and H. Xu. Stability and sensitivity analysis of stochastic programs with second order dominance constraintss. *Mathematical Programming Series A*, 142:435–460, 2013. 13

Y. Liu, A. Pichler, and H. Xu. Discrete approximation and quantification in distributionally robust optimization. manuscript, 2016. 19

S. Mehrotra and D. Papp. A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. *SIAM Journal on Optimization*, 24:1670–1697, 2014. doi:10.1137/130925013. 3

G. Ch. Pflug and A. Pichler. Approximations for probability distributions and stochastic optimization problems. In M. Bertocchi, G. Consigli, and M. A. H. Dempster, editors, *Stochastic Optimization Methods in Finance and Energy*, volume 163 of *International Series in Operations Research & Management Science*, chapter 15, pages 343–387. Springer, New York, 2011. ISBN 978-1-4419-9586-5. doi:10.1007/978-1-4419-9586-5. 2, 9

G. Ch. Pflug and A. Pichler. *Multistage Stochastic Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2014. doi:10.1007/978-3-319-08843-3. 2, 16

G. Ch. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007. doi:10.1080/14697680701455410. 3

A. Pichler. Premiums and reserves, adjusted by distortions. *Scandinavian Actuarial Journal*, 2015(4): 332–351, sep 2013. doi:10.1080/03461238.2013.830228. 16, 17

A. Pichler and A. Shapiro. Minimal representations of insurance prices. *Insurance: Mathematics and Economics*, 62:184–193, 2015. ISSN 0167-6687. doi:10.1016/j.insmatheco.2015.03.011. 16

I. Pólik and T. Terlaky. A survey of the s-lemma. *SIAM REview*, 49:371–418, 2007. 3

Y. V. Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.*, 1:157–214, 1956. 6

S. T. Rachev. *Probability metrics and the stability of stochastic models*. John Wiley& Sons, 1991a. 10

S. T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. John Wiley and Sons, West Sussex, England, 1991b. URL http://books.google.com/books?id=5grvAAAAMAAJ. 4

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970. 8

R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2:21–41, 2000. 20

W. Römisch. Stability of stochastic programming problems. In A. Ruszczyński and A. Shapiro, editors, *Stochastic Programming, Handbooks in Operations Research and Management Science*, volume 10, chapter 8. Elsevier, Amsterdam, 2003. 2, 11

A. Shapiro. On Kusuoka representation of law invariant risk measures. *Mathematics of Operations Research*, 38(1):142–152, 2013. doi:10.1287/moor.1120.0563. 16, 17

A. V. Skorokhod. *Basic principles and applications of probability theory*. Springer, New York, 1989. 6

H. Sun and H. Xu. Convergence analysis for distributionally robust optimization and equilibrium problems. *Mathematics of Operations Research*, 2016. to appear. 2, 11

C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. ISBN 0-821-83312-X. doi:10.1090/gsm/058. URL http://books.google.com/books?id=GqRXYFxe0l0C. 5, 10, 11

W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62:1358–1376, 2014. 3

H. Xu, Y. Liu, and H. Sun. Distributionally robust optimization with matrix moment constraints: Lagrange duality and cutting plane method. Technical report, University of Southampton, 2015. 3, 19

W. R. Y. Liu and H. Xu. Quantitative stability analysis of stochastic generalized equation. *SIAM J. Optimization*, 24:467–497, 2014. 19

J. Zhang, H. Xu, and L. W. Zhang. Quantitative stability analysis for distributionally robust optimization with moment constraints. Optimization Online, 2015a. 2, 11, 13

J. Zhang, H. Xu, and L. W. Zhang. Quantitative stability analysis of stochastic quasi-variational inequality problems and applications. Optimization Online, 2015b. 14

C. Zhao and Y. Guan. Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics. Optimization Online, 2015a. 2, 5, 9

C. Zhao and Y. Guan. Data-driven risk-averse stochastic optimization with wasserstein metrics. Optimization Online, 2015b. 2

V. M. Zolotarev. Probability metrics. *Teoriya Veroyatnostei i ee Primeneniya*, 28:264–287, 1983. 4

S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137:167–198, 2013. 3