Quantitative Statistical Robustness in Distributionally Robust Optimization Models *

Huifu Xu and Sainan Zhang †

December 4, 2021

Abstract

In distributionally robust optimization (DRO) models, sample data of the underlying exogenous uncertainty parameters are often used to construct an ambiguity set of plausible probability distributions. It is common to assume that the sample data do not contain noise. This assumption may not be fulfilled in some data-driven problems where the perceived data are potentially contaminated. Consequently it raises a question as to whether the statistical estimators of the optimal values obtained from solving the DRO models are statistically robust, that is, the differences between the laws of these estimators and their counterparts based on real data (without noise) are controllable. In this paper, we derive error bounds for the differences under the Kantorovich metric for two classes of DRO models with applications in machine learning and risk management.

Keywords. DRO models, moment-type conditions, ζ-balls, quantitative statistical robustness

1 Introduction

We consider the following one-stage distributionally robust optimization (DRO) problem

$$\text{(DRO)} \quad \min_{x\in X} \max_{P\in\mathcal{P}} \ \mathbb{E}_P[f(x,\xi)], \tag{1.1}$$

where $\mathcal{P}$ is an ambiguity set and $f : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ is a continuous function, x is a decision vector which is restricted to taking values over a specified compact set $X \subset \mathbb{R}^n$, $\xi : \Omega \to \mathbb{R}^k$ is a vector of random variables defined over probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $P := \mathbb{P} \circ \xi^{-1}$ is the probability measure on $\mathbb{R}^k$ induced by ξ and $\mathbb{E}_P[\cdot]$ is the mathematical expectation w.r.t. P. In this model, the true probability distribution of ξ is unknown and the optimal decision is based on the worst probability distribution from ambiguity set $\mathcal{P}$.

*This work is supported by RGC grant 14204821.

†Department of Systems Engineering and Engineering Management,The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. Email: hfxu@se.cuhk.edu.hk, snzhang@se.cuhk.edu.hk

A key component of the DRO model is the ambiguity set not only because it concerns proper use of available information for identifying the true unknown probability distribution but also its structure affects the solvability of the minimax optimization problem. In the literature of distributionally robust optimization, various approaches have been proposed for constructing ambiguity set $\mathcal{P}$ depending on available information structure, see [21, 37] for overviews. Here we consider two popular approaches. One is to use moment information such as

$$\mathcal{P} := \Big\{ P \in \mathscr{P}(\mathbb{R}^k) : \ \mathbb{E}_P[\Psi(\xi)] \in \mathcal{K} \ \Big\}, \tag{1.2}$$

where Ψ is a random mapping consisting of vector and/ or matrix-valued measurable functions, the mathematical expectation of Ψ is taken w.r.t. each component of Ψ, $\mathscr{P}(\mathbb{R}^k)$ denotes the set of all probability measures on $\mathbb{R}^k$ induced by ξ and $\mathcal{K}$ is a closed convex cone in the Cartesian product of some finite dimensional vector and/ or matrix spaces, see Xu et al. [34]. In general, Ψ depends on sample information such as sample means and sample variance. The next example explains this.

Example 1.1 (Delage and Ye [2] and So [24]) Consider the ambiguity set

$$\mathcal{P} = \mathcal{P}(\mu_N, \Sigma_N, \gamma_1, \gamma_2) := \left\{ P \in \mathscr{P}(\Xi): \quad \begin{array}{l} \mathbb{E}_P[\xi - \mu_N]^T \Sigma_N^{-1} \mathbb{E}_P[\xi - \mu_N] \le \gamma_1 \\ \mathbb{E}_P[(\xi - \mu_N)(\xi - \mu_N)^T] \preceq \gamma_2 \Sigma_N \end{array} \right\}, \tag{1.3}$$

where γ_1 and γ_2 are nonnegative constants, μ_N and Σ_N are the sample mean and sample covariance, $\Xi \subset \mathbb{R}^k$ is the support set of the true probability distribution of ξ and $\mathscr{P}(\Xi)$ is the set of all probability distributions of ξ whose support sets are contained in Ξ (or alternatively the set of all probability measures on Ξ induced by mapping ξ). The ambiguity set is considered by Delage and Ye [2], further studied by So [24] and used by many others in different DRO models. By employing the Schur complement, we can easily reformulate (1.3) in the form of (1.2) with

$$\Psi(\xi) = \Psi(\xi, \mu_N, \Sigma_N, \gamma_1, \gamma_2) := \left(\begin{array}{c} \left[\begin{array}{cc} -\Sigma_N & \mu_N - \xi \\ (\mu_N - \xi)^T & -\gamma_1 \end{array} \right] \\ (\xi - \mu_N)(\xi - \mu_N)^T - \gamma_2 \Sigma_N \end{array} \right), \tag{1.4}$$

$\mathcal{K} = \mathcal{K}_1 \times \mathcal{K}_2$, where $\mathcal{K}_1 := \mathcal{S}_-^{k+1}$, $\mathcal{K}_2 := \mathcal{S}_-^k$, and k is the dimension of ξ.

In this setup, information on the sample means and sample variances is used to identify the scope of the true unknown probability distribution. The samples are assumed to be generated by the true probability distribution of ξ, which means that they are not contaminated. The DRO model is mainly concerned with the sample size N. If the sample size can be arbitrarily large, then one can simply use samples to recover the true probability distribution of ξ. In practice particularly in data-driven problems, one may not be able to obtain a large amount of samples and subsequently use the sample mean and sample variance to construct a set of plausible probability distributions satisfying (1.3).

An important issue which is not paid adequate attention to, in the literature of distributionally robust optimization, is that the sample data may be potentially contaminated, in which case μ_N and Σ_N may not approximate the true mean value and the true covariance as N goes to

infinity when the samples are contaminated. In other words, the subsequent theoretical results of the DRO model in [2, 24, 25] are not applicable in these circumstances.

Another approach which is used in the literature of distributionally robust optimization is to use partially available information about the true probability distribution such as samples, computer simulation or subjective judgement to construct a nominal distribution and then build an ambiguity set $\mathcal{P}$ by including all distributions near the nominal in the sense of some "distance" such as Kantorovich/Wasserstein distance, semi-distance and divergence distance [5, 14, 23]. Here we give a simple example.

Example 1.2 (Pichler and Xu [20]) Let $P_N \in \mathscr{P}(\Xi)$ be a nominal distribution constructed through available sample data. Let

$$\mathcal{P} = \{P' \in \mathscr{P}(\Xi) \colon \mathsf{dl}_{\mathscr{G}}(P', P_N) \leq r_N\}, \tag{1.5}$$

where r_N is a positive number,

$$\mathsf{dl}_{\mathscr{G}}(P, Q) := \sup_{g \in \mathscr{G}} |\mathbb{E}_P[g(\xi)] - \mathbb{E}_Q[g(\xi)]| \tag{1.6}$$

is a semi-distance called a metric with ζ-structure and $\mathscr{G}$ is a family of real-valued measurable functions on Ξ. Formulation (1.6) subsumes a wide range of metrics in probability theory, see Rachev [26] or Zolotarev [38]. For the simplicity of terminology, we call it ζ-metric and the ambiguity set $\mathcal{P}$ ζ-ball. In the case that

$$\mathscr{G} = \mathscr{G}_L := \{g : \mathbb{R}^k \to \mathbb{R} : g \text{ is Lipschitz continuous with modulus being bounded by } 1\}, \tag{1.7}$$

$\mathsf{dl}_{\mathscr{G}}(P, Q)$ reduces to the Kantorovich metric, denoted by $\mathsf{dl}_{K,k}(P, Q)$, where the subscripts K, k indicate the Kantorovich metric in $\mathcal{P}(\mathbb{R}^k)$.

In the literature of DRO models, it is often assumed that P_N is constructed by samples without any noise. Under such an assumption, we know that P_N converges to the true probability distribution as $N \to \infty$ and so does the ambiguity set when $r_N \to 0$, see Esfahani and Kuhn [5] and Shapiro [23]. Like in the previous example, our concern here is that the samples to be used to construct P_N may be potentially contaminated.

Over the past decade, DRO models have found many applications including machine learning and risk management. Here we list a couple of them.

Example 1.3 (Shafieezadeh-Abadeh et al. [22]) Let $\mathbb{X} \subset \mathbb{R}^{k-1}$ be an input space (e.g., information on the frequency of certain keywords in an email) and $\mathbb{Y} \subset \mathbb{R}$ the output space (e.g., a label $+1$ (-1) if the email is likely (unlikely) to be a spam message). The relationship between an input $\boldsymbol{x} \in \mathbb{X}$ and an output $y \in \mathbb{Y}$ is described by a probability distribution P. To ease the notation, let ξ denote the input-output pair $(\boldsymbol{x}, y)$ and $\Xi := \mathbb{X} \times \mathbb{Y} \subset \mathbb{R}^{k-1} \times \mathbb{R}$ the support set of ξ.

In a supervised learning framework, the true distribution P on $\mathbb{X} \times \mathbb{Y}$ is often unknown but it is possible to obtain finite input-output data (samples) $\xi^i = (\boldsymbol{x}^i, y^i)$ for $i = 1, \cdots, N$ generated

by the true distribution P (e.g., a database of emails which have been classified by a human as legitimate or as spam messages), which are referred as the *training data*.

Given the training samples $\xi^i = (\boldsymbol{x}^i, y^i)$, $i = 1, \cdots, N$, the goal of supervised learning is to find a function $h : \mathbb{X} \to \mathbb{Y}$ to infer an unknown relationship between input $\boldsymbol{x}$ and output y, which is described by *target function* $f : \mathbb{X} \to \mathbb{Y}$ such that h solves

$$\inf_{h \in \mathbb{H}} \mathbb{E}_{P_N}[c(h(\boldsymbol{x}), y)]. \tag{1.8}$$

In this model, $P_N := \frac{1}{N} \sum_{i=1}^N \delta_{(\boldsymbol{x}^i, y^i)}$ denotes the empirical distribution, $h \in \mathbb{H}$ is known as a *hypothesis* or a *learning model*, where $\mathbb{H}$ is a hypothesis space, $c : \mathbb{R} \times \mathbb{Y} \to \mathbb{R}_+$ is the *loss function* which measures the cost of mismatch between each pair of input and output data by using the hypothesis h instead of the true target function f.

Model (1.8) is to find an optimal candidate function $h \in \mathbb{H}$ which approximates the unknown target function f such that the overall expected cost is minimized. Since the optimization process is based on training data, it is called a learning model/algorithm. When a linear hypothesis $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ with $\boldsymbol{w} \in \mathbb{R}^{k-1}$ is used, problem (1.8) only minimizes the training sample error $\mathbb{E}_{P_N}[c(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)]$ and $\boldsymbol{w}$ may still suffer from a high out-of-sample error $\mathbb{E}_P[c(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)]$ due to overfitting. The standard remedy to tackle overfitting is using regularization, such as Lasso regularization [30] and L_2-regularization to consider the regularized loss $\mathbb{E}_{P_N}[c(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)] + cR(\boldsymbol{w})$, where $cR(\boldsymbol{w})$ is an overfitting penalty.

Shafieezadeh-Abadeh et al. [22] consider a distributionally robust model which is more principled than regularization,

$$\inf_{\boldsymbol{w}} \sup_{P' \in \hat{\mathcal{B}}(P_N, r)} \mathbb{E}_{P'}[c(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)], \tag{1.9}$$

where the ambiguity set is defined by a Wasserstein ball:

$$\hat{\mathcal{B}}(P_N, r) = \{P' \in \mathscr{P}(\Xi) : d_{W,k}(P', P_N) \le r\}. \tag{1.10}$$

The particular Wasserstein distance that they consider is

$$d_{W,k}(Q, Q') := \inf_{\Pi} \left\{ \int_{\Xi^2} d(\xi, \xi') \Pi(d\xi, d\xi') : \begin{array}{c} \Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \\ \text{with marginals } Q \text{ and } Q', \text{ respectively} \end{array} \right\},$$

where d is a distance in metric space (Ξ, d) and subscripts W, k indicate the Wasserstein metric in $\mathscr{P}(\mathbb{R}^k)$. By the Kantorovich-Rubinstein theorem, $\mathsf{dl}_{W,k}$ coincides with the Kantorovich metric $\mathsf{dl}_{K,k}$ when (Ξ, d) is a metric space, see e.g. [4, page 338], [3, Theorem 11.8.2], [29, Theorem 2]. Our concern here is that how the learning model/algorithm works in the case when the training data are contaminated.

Next, we consider a DRO model in risk management.

Example 1.4 (Guo and Xu [6]) Guo and Xu [6] consider the distributionally robust shortfall risk optimization model

$$\begin{array}{ll} \min\limits_{t \in \mathbb{R}} & t \\ \text{s.t.} & \sup\limits_{P' \in \mathcal{B}(P_N, r)} \mathbb{E}_{P'}[l(-\xi - t)] \le \lambda, \end{array} \tag{1.11}$$

where ξ is a random financial position, $l : \mathbb{R} \to \mathbb{R}$ is an increasing utility loss function and λ is the maximum tolerable utility loss, the shortfall risk is the smallest amount of cash to be injected to the financial position so that the expected loss falls below the specified level. The DRO model is used because the true probability distribution of ξ is often unknown. The authors propose to use the Kantorovich ball centered at a nominal distribution P_N,

$$\mathcal{B}(P_N, r) = \{P' \in \mathscr{P}(\Xi) : \mathsf{dl}_{K,1}(P', P_N) \leq r\}, \tag{1.12}$$

for constructing the ambiguity set. Like in the previous example, our concern here is that the sample data may be contaminated.

There are potentially two ways to tackle the data contamination issue in the DRO models. One is to investigate the impact of outliers of the random data on the optimal value. This approach is well known in robust statistics where the impact is characterized by a so-called influence function and it has been applied to support vector machine by Steinwart and Christmann [28]. Robust statistics stems from Tukey [31, 32] and Hampel [12, 13] and has been popularized by others particularly the monographs by Huber [15] and Huber and Ronchetti [10]. Lecué and Lerasle [18] extend the research by proposing a so-called median-of-the mean (MOM) approach for machine learning models to reduce the impact of the outliers in the dataset. More recently, Guo and Xu [8] apply the influence function approach to stochastic generalized equations (SGE) and examine the sensitivity of the solutions of the SGE w.r.t. a single data perturbation.

The other is to look into the impact of all corrupted data on the statistical estimators of the DRO models rather than merely the outliers. This approach is first proposed by Cont et al. [1] for investigating qualitative robustness of statistical estimator of various risk measures derived from empirical data. An important finding out of their analysis is that statistical estimator of any spectral risk measure including conditional value at risk is not statistically robust whereas value at risk is. This is primarily because the former is more sensitive to the perturbation of tail data. Krätschmer et al. [17] find that the robustness of statistical estimators depends on the empirical data structure. They demonstrate that if the perceived (contaminated) data are "close" under some fine topology to the real data (with contamination being detached), the statistical estimators may remain qualitatively robust. In the more recent developments, Guo and Xu [7] apply the qualitative statistical robust approach to preference robust optimization models and take a step further to develop a quantitative approach by using the Kantorovich metric to measure the difference between laws of the statistical estimators of the optimal values. Wang et al. [33] extend the quantitative approach by adopting the Fortet-Mourier metric (including Kantorovich metric as a special case) and apply it to study statistical robustness of tail-dependent law invariant risk measures. Guo et al. [9] apply the latter to machine learning.

In this paper, our focus will be on the second way. This is primarily because in data-driven problems, decision makers are more likely to be confronted with a situation where all of the available data are potentially contaminated. The main contributions of this paper can be summarized as follows.

- First, we derive a general quantitative statistical robustness result for a general estimator which is globally Lipschitz continuous w.r.t. the underlying uncertainty data (Theorem

2.1) and use it as a template for describing the statistical robustness of the optimal values of various specific DRO problems.

- Second, we derive quantitative statistical robustness for the optimal value of the DRO model with ambiguity set having moment structure (1.3). This essentially requires us to take two steps. One is to demonstrate the local Lipschitz continuity of the ambiguity set as a set-valued mapping w.r.t. the change of the sample mean, sample covariance and other parameters under the Slater condition of the moment system. This step is based on the error bound for the moment system which holds uniformly for all parameter values in a neighborhood of a certain point where the Slater condition holds. The second step is to show the global Lipschitz continuity of the optimal value of the DRO model w.r.t. sample mean and sample covariance. To this end, we consider a general DRO problem with an abstract ambiguity set and derive the global Lipschitz continuity of the optimal value of the DRO problem (Lemma 3.1) under the condition that the ambiguity set is locally Lipschitz continuous w.r.t. data. We then move to discuss sufficient conditions under which the ambiguity set is locally Lipschitz continuous (Propositions 3.1 and 3.2).

- Third, by exploiting the quantitative stability of the ζ-ball in [20, Theorem 1], we establish the Lipschitz continuity of the optimal value w.r.t. independent and identically distributed (iid) samples and subsequently derive the quantitative statistical robustness of the optimal value of the DRO model with ζ-ball (Theorem 4.1). The statistical robustness result covers the DRO model with Wasserstein ball in machine learning (Proposition 4.1).

- Fourth, when the support set is compact, we identify sufficient conditions under which the statistical estimator of the DRO version of the shortfall risk with the Kantorovich ball structured ambiguity set is statistically robust against perturbation of data (Theorem 5.1). This result provides theoretical grounding for the DRO version of the shortfall risk measure to be applied in data-driven problems with contaminated data.

The rest of the paper is organized as follows. Section 2 presents the sufficient conditions for the quantitative statistical robustness of general statistical estimator. Section 3 discusses the Lipschitz continuity of the optimal value function of DRO with moment constraints which paves the way for the analysis of the quantitative statistical robustness of the corresponding statistical estimators. Section 4 derives the quantitative statistical robustness of the DRO with ζ-ball based on the Lipschitz continuity of the ζ-ball w.r.t. the center. The result is applied to the DRO model in machine learning (see Example 1.3). Section 5 focuses on Lipschitz continuity of the optimal value function of the distributionally robust shortfall risk optimization model with Kantorovich ball, which guarantees the quantitative statistical robustness.

Throughout the paper, we use the following notation. By convention, we use $\mathbb{R}^{k\times k}$, $\mathcal{S}^k$ and $\mathcal{S}^k_-$ to denote respectively the space of all $k\times k$ matrices, $k\times k$ symmetric matrices, and the cone of negative semidefinite symmetric matrices. We use $\|x\|$ to represent the Euclidean norm of a vector x in $\mathbb{R}^k$, and $\|A\| := \sqrt{\mathrm{tr}(A^TA)}$ to stand for the Frobenius norm of a matrix $A \in \mathbb{R}^{k\times k}$, where "tr" denotes the trace of a matrix and the superscript T denotes transpose. For a Banach space X, we write $\mathcal{B}$ for the closed unit ball in X. For a set $S \subseteq X$, $\mathrm{int}\, S$ denotes the interior of S, and $d(x,S) := \inf_{x'\in S}\|x'-x\|$ denotes the distance from a point $x \in X$ to a set $S \subset X$.

For two sets $S_1, S_2 \subset X$,

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

signifies the deviation of S_1 from S_2 under the metric d, and

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

denotes the Hausdorff distance between the two sets. Finally we write $\mathrm{diam}(\Xi) := \sup_{\xi, \xi' \in \Xi} \|\xi - \xi'\|$ for the diameter of Ξ and $\mathbb{N}$ for the set of positive integers.

2 Quantitative statistical robustness

Let $P \in \mathscr{P}(\mathbb{R}^k)$ be the true probability distribution of random vector ξ and $\xi^1, \cdots, \xi^N$ the iid samples generated by P (strictly speaking they are iid random variables generating iid samples). In practice, $\xi^1, \cdots, \xi^N$ may be obtained from empirical data which are potentially contaminated. Let $\tilde{\xi}^1, \cdots, \tilde{\xi}^N$ be the perceived data which contain noise. Obviously the samples are not generated by P, rather they are generated by some unknown distribution Q. If we view $\tilde{\xi}^1, \cdots, \tilde{\xi}^N$ as samples perturbed from $\xi^1, \cdots, \xi^N$, then we may regard Q as a perturbation of P. Note that neither P nor Q is known. To facilitate the discussion, we assume that $\tilde{\xi}^1, \cdots, \tilde{\xi}^N$ are also independent and identically distributed. Let

$$P_N := \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi^i} \quad \text{and} \quad Q_N := \frac{1}{N} \sum_{i=1}^{N} \delta_{\tilde{\xi}^i} \tag{2.13}$$

be the respective empirical distributions, where $\delta_{\tilde{\xi}}$ denotes the Dirac probability measure at $\tilde{\xi}$.

To explain the idea of statistical robustness, let $(\mathbb{R}^k)^{\otimes N}$ denote the Cartesian product $\mathbb{R}^k \times \cdots \times \mathbb{R}^k$ and $\mathcal{B}(\mathbb{R}^k)^{\otimes N}$ its Borel sigma algebra. Let $P^{\otimes N}$ denote the probability measure on the measurable space $\left((\mathbb{R}^k)^{\otimes N}, \mathcal{B}(\mathbb{R}^k)^{\otimes N}\right)$ with marginal P on each $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ and $Q^{\otimes N}$ with marginal Q. Consider a statistical functional $T(\cdot)$ mapping from a subset of $\mathcal{M} \subset \mathscr{P}(\mathbb{R}^k)$ to $\mathbb{R}$. For each $N \in \mathbb{N}$, we write $T_N(\xi^1, \cdots, \xi^N)$ for $T(P_N)$, where $\mathbb{N}$ is the set of positive integers. Notice that T_N maps from $(\mathbb{R}^k)^{\otimes N}$ to $\mathbb{R}$ and provides an estimator for $T(P)$. Our interest is whether $T(Q_N)$ is close to $T(P_N)$ under some appropriate metric for all N sufficiently large. Here $T(P_N)$ should be understood as the corresponding statistical estimator when the noise in the samples is detached. If $T(Q_N)$ is close to $T(P_N)$, then it is safe to use $T(Q_N)$ as an estimate of $T(P)$ (because we are unable to obtain $T(P_N)$ in practice).

Let $\phi : \mathbb{R}^k \to [0, \infty)$ be a continuous function and

$$\mathcal{M}_k^{\phi} := \left\{ P' \in \mathcal{P}(\mathbb{R}^k) : \int_{\mathbb{R}^k} \phi(t) P'(dt) < \infty \right\}.$$

$\mathcal{M}_k^{\phi}$ defines a subset of probability measures in $\mathcal{P}(\mathbb{R}^k)$ which satisfies the generalized moment condition of ϕ.

Let $P, Q \in \mathscr{P}(\mathbb{R}^k)$ be any two probability measures and $P^{\otimes N}, Q^{\otimes N} \in \mathscr{P}((\mathbb{R}^k)^{\otimes N})$, i.e., the two probability measures on $(\mathbb{R}^k)^{\otimes N}$ with marginal probabilities P and Q on $\mathbb{R}^k$ respectively.

The next lemma establishes a relationship between $\mathsf{dl}_{\mathscr{G}}((P)^{\otimes N},(Q)^{\otimes N})$ and $\mathsf{dl}_{K,k}(P,Q)$ when $\mathscr{G}$ is the set of all Lipschitz continuous functions on $(\mathbb{R}^k)^{\otimes N}$ with modulus being bounded by 1.

Lemma 2.1 ([7]) *Let $\vec{t}:=(t^1,\cdots,t^N)\in(\mathbb{R}^k)^{\otimes N}$ and $\psi:(\mathbb{R}^k)^{\otimes N}\to\mathbb{R}$ be a Lipschitz continuous function with modulus being bounded by L/N for a fixed constant $L>0$. Let Ψ denote the set of all these functions, i.e.,*

$$\Psi:=\left\{\psi:(\mathbb{R}^k)^{\otimes N}\to\mathbb{R}\,\middle|\,|\psi(\vec{\tilde{t}})-\psi(\vec{\hat{t}})|\le\frac{L}{N}\sum_{i=1}^{N}\|\tilde{t}^i-\hat{t}^i\|,\forall\vec{\tilde{t}},\vec{\hat{t}}\in(\mathbb{R}^k)^{\otimes N}\right\}.$$

Then $\mathsf{dl}_{\Psi}(P^{\otimes N},Q^{\otimes N})\le L\mathsf{dl}_{K,k}(P,Q)$, where $\mathsf{dl}_{K,k}$ is defined as in Example 1.2.

With the technical result, we are able to derive a bound for $\mathsf{dl}_{\mathscr{G}}((P)^{\otimes N}\circ T_N^{-1},(Q)^{\otimes N}\circ T_N^{-1})$ in terms of $\mathsf{dl}_{K,k}(P,Q)$.

Theorem 2.1 (Quantitative statistical robustness) *Assume that for any $N\in\mathbb{N}$*

$$|T_N(\tilde{\xi}^1,\cdots,\tilde{\xi}^N)-T_N(\hat{\xi}^1,\cdots,\hat{\xi}^N)|\le\frac{L}{N}\sum_{i=1}^{N}\|\tilde{\xi}^i-\hat{\xi}^i\|. \tag{2.14}$$

Let $P,Q\in\mathcal{M}_k^{\phi}$, where $\phi(t):=\|t\|$, $t\in\mathbb{R}^k$. Then

$$\mathsf{dl}_{K,1}\left(P^{\otimes N}\circ T_N^{-1},Q^{\otimes N}\circ T_N^{-1}\right)\le L\mathsf{dl}_{K,k}(P,Q) \tag{2.15}$$

for all $N\in\mathbb{N}$.

The result shows that when Q is close to P, the statistical estimator based on the perceived data Q_N is close to the one based on the real data P_N uniformly for all N. This kind of result is first established by Guo and Xu [7] for the statistical estimators of the optimal values of preference robust optimization problems and extended by Wang et al. [33] to risk measures where condition (2.14) is weakened to local Lipschitz continuity with polynomial rate of growth of the Lipschitz modulus. It is called quantitative statistical robustness in order to differentiate it from qualitative statistical robustness by Krätschmer et al. [17] and Cont et al. [1] where the distance between $P^{\otimes N}\circ T_N^{-1}$ and $Q^{\otimes N}\circ T_N^{-1}$ is measured by Prokhorov metric or Lévy metric and its relationship with the distance between Q and P is implicit and qualitative.

Proof. By definition

$$\begin{aligned}
&\mathsf{dl}_{K,1}\left(P^{\otimes N}\circ T_N^{-1},Q^{\otimes N}\circ T_N^{-1}\right)\\
=&\sup_{g\in\mathscr{G}_L}\left|\int_{\mathbb{R}}g(t)P^{\otimes N}\circ T_N^{-1}(dt)-\int_{\mathbb{R}}g(t)Q^{\otimes N}\circ T_N^{-1}(dt)\right|\\
=&\sup_{g\in\mathscr{G}_L}\left|\int_{(\mathbb{R}^k)^{\otimes N}}g(T_N(\vec{\xi}^N))P^{\otimes N}(d\vec{\xi}^N)-\int_{(\mathbb{R}^k)^{\otimes N}}g(T_N(\vec{\xi}^N))Q^{\otimes N}(d\vec{\xi}^N)\right|,
\end{aligned} \tag{2.16}$$

where $\mathscr{G}_L$ is defined in (1.7) and $\vec{\xi}^N := (\xi^1, \cdots, \xi^N)$. For each $g \in \mathscr{G}_L$, it follows by (2.14) that

$$\begin{aligned} |g(T_N(\tilde{\xi}^1, \cdots, \tilde{\xi}^N)) - g(T_N(\hat{\xi}^1, \cdots, \hat{\xi}^N))| &\leq |T_N(\tilde{\xi}^1, \cdots, \tilde{\xi}^N) - T_N(\hat{\xi}^1, \cdots, \hat{\xi}^N)| \\ &\leq \frac{L}{N} \sum_{i=1}^N \|\tilde{\xi}^i - \hat{\xi}^i\|, \end{aligned} \tag{2.17}$$

which means that $g(T_N(\cdot))$ is Lipschitz continuous over $(\mathbb{R}^k)^{\otimes N}$ with Lipschitz modulus being bounded by L/N. By setting $\psi = g \circ T_N$ and invoking Lemma 2.1, we have

$$\text{rhs of (2.16)} \leq L\mathsf{dl}_{K,k}(Q, P).$$

To complete the proof, it suffices to show the well-definedness of the metric $\mathsf{dl}_{K,1}$, that is, for any $g(T_N)$, both $\int_{(\mathbb{R}^k)^{\otimes N}} g(T_N(\vec{\xi}^N)) P^{\otimes N}(d\vec{\xi}^N)$ and $\int_{(\mathbb{R}^k)^{\otimes N}} g(T_N(\vec{\xi}^N)) Q^{\otimes N}(d\vec{\xi}^N)$ are well-defined. Let $\xi_0^1, \cdots, \xi_0^N$ be fixed. By (2.17),

$$|g(T_N(\xi^1, \cdots, \xi^N)) - g(T_N(\xi_0^1, \cdots, \xi_0^N))| \leq \frac{L}{N} \sum_{i=1}^N \|\xi^i - \xi_0^i\|. \tag{2.18}$$

Note that inequality (2.17) implies continuity and hence measurability of $g(T_N(\cdot))$.

For any $P \in \mathcal{M}_k^\phi$, by using inequality (2.18), and applying Tonelli's theorem to the integral $\int_{(\mathbb{R}^k)^{\otimes N}} g(T_N(\vec{\xi}^N)) P^{\otimes N}(d\vec{\xi}^N)$ by switching the order of integration, we obtain

$$\begin{aligned} \int_{(\mathbb{R}^k)^{\otimes N}} g(T_N(\vec{\xi}^N)) P^{\otimes N}(d\vec{\xi}^N) &\leq |g(T_N(\xi_0^1, \cdots, \xi_0^N))| + \frac{L}{N} \int_{(\mathbb{R}^k)^{\otimes N}} \sum_{i=1}^N \|\xi^i - \xi_0^i\| P^{\otimes N}(d\vec{\xi}^N) \\ &= |g(T_N(\xi_0^1, \cdots, \xi_0^N))| + \frac{L}{N} \sum_{i=1}^N \int_{\mathbb{R}^k} \|\xi^i - \xi_0^i\| P(d\xi^i) \\ &< \infty, \end{aligned}$$

where the equality is due to the fact that the integrand depends on $(\xi^1, \cdots, \xi^N)$ only and the last inequality holds because $\int_{\mathbb{R}^k} \|\xi^i - \xi_0^i\| P(d\xi^i) < \infty$ for $P \in \mathcal{M}_k^\phi$. The boundedness and measurability ensure the well-definedness of $\int_{(\mathbb{R}^k)^{\otimes N}} g(T_N(\vec{\xi}^N)) P^{\otimes N}(d\vec{\xi}^N)$ as desired. Similar conclusion can be drawn when P is replaced by Q. ■

Theorem 2.1 paves the way for the statistical robustness of the optimal values of the DRO models outlined in Section 1. Note that as in Wang et al. [33], it is possible to weaken condition (2.14) to locally Lipschitz continuity with specified rate of growth of the modulus, and subsequently derive an error bound under the Fortet-Mourier metric at the right hand side of (2.15). Here we adopt a simpler version of the quantitative statistical robustness result so that we may concentrate on other important issues concerning the DRO models.

In the forthcoming discussions (Sections 3-5), we will use Theorem 2.1 as a template to present the quantitative statistical robustness of the optimal values of the DRO problems. The basic idea is to derive Lipschitz continuity of the ambiguity sets of probability distributions w.r.t. change of sample data and subsequently demonstrate the Lipschitz continuity of the optimal value function (w.r.t. change of sample data).

3 Statistical robustness of the DRO models with moment conditions

In this section, we consider the DRO model with the ambiguity set being constructed by moment conditions:

$$\text{(DRO-moment)} \quad \min_{x\in X} \max_{P'\in\mathcal{P}(\mu_N,\Sigma_N,r_1,r_2)} \mathbb{E}_{P'}[f(x,\xi)], \tag{3.19}$$

where $\mathcal{P}(\mu_N,\Sigma_N,r_1,r_2)$ is defined as in (1.3) and is recast here as

$$\mathcal{P}(\mu_N,\Sigma_N,r_1,r_2)=\left\{P'\in\mathscr{P}(\Xi):\mathbb{E}_{P'}[\Psi(\xi,\mu_N,\Sigma_N,\gamma_1,\gamma_2)]\in\left(\begin{array}{c}\mathcal{S}_-^{k+1}\\ \mathcal{S}_-^{k}\end{array}\right)\right\}, \tag{3.20}$$

where $\Psi(\xi,\mu_N,\Sigma_N,\gamma_1,\gamma_2)$ is defined as in (1.4) and $\mathcal{S}_-^k$ denotes the cone of all $k\times k$ negative semidefinite symmetric matrices. To facilitate the forthcoming discussions, let us write down the inner maximization problem separately as

$$\begin{array}{ll}\max\limits_{P'\in\mathscr{P}(\mathbb{R}^k)} & \mathbb{E}_{P'}[f(x,\xi)]\\ \text{s.t.} & P'\in\mathcal{P}(\mu_N,\Sigma_N,r_1,r_2).\end{array} \tag{3.21}$$

Let $\vartheta(\mu_N,\Sigma_N,r_1,r_2)$ and $v(x,\mu_N,\Sigma_N,r_1,r_2)$ denote the optimal values of (3.19) and (3.21) respectively.

To derive the statistical robustness of $\vartheta(\mu_N,\Sigma_N,r_1,r_2)$, we need to show the Lipschitz continuity of ϑ w.r.t $(\xi^1,\cdots,\xi^N)$. Since $\vartheta(\mu_N,\Sigma_N,r_1,r_2)$ is the minimum of function $v(\cdot,\mu_N,\Sigma_N,r_1,r_2)$ over X, we use classical stability analysis results (see e.g. Klatte [16, Theorem 1]) to show Lipschitz continuity of $v(x,\mu_N,\Sigma_N,r_1,r_2)$. This requires us to figure out sufficient conditions for the Lipscitz continuity of $\mathcal{P}(\mu_N,\Sigma_N,r_1,r_2)$ w.r.t. (μ_N,Σ_N) in the first place, and then the Lipschitz continuity of μ_N and Σ_N w.r.t. $(\xi^1,\cdots,\xi^N)$. We do these in the sequel.

3.1 Stability of a general minimax DRO problem

We begin by presenting a stability result for an optimization problem with general minimax structure. Let $\Xi\subset\mathbb{R}^k$ be a compact set, U be a compact set in a metric space equipped with norm $\|\cdot\|_U$, and g be a continuous function mapping from $\mathbb{R}^n\times\mathscr{P}(\Xi)$ to $\mathbb{R}$. We consider the following general parametric minimax problem:

$$\min_{x\in X}\max_{P\in\mathcal{P}(u)} g(x,P), \tag{3.22}$$

where $u\in U$ is a fixed parameter, $X\subset\mathbb{R}^n$ is compact set. Let $v(x,u):=\max_{P\in\mathcal{P}(u)} g(x,P)$ denote the optimal value function of the maximization in (3.22) and $\vartheta(u):=\min_{x\in X} v(x,u)$ be the optimal value function of problem (3.22). Assume $\mathcal{P}(u)\neq\emptyset$ for each u. Since Ξ is compact, then $\mathscr{P}(\Xi)$ is weakly compact, and consequently by the continuity of g and compactness of X, $v(x,u)$ and $\vartheta(u)$ are finite valued.

We quantify the impact of arbitrary perturbation of parameter u on the optimal value ϑ. To this end, we first investigate the case that parameter u is perturbed in a neighborhood of $\bar{u}$.

Lemma 3.1 *Let $\bar{u}$ be fixed. Assume: (a) $g(x,P)$ is globally Lipschitz continuous in (x,P) over $X \times \mathscr{P}(\Xi)$, i.e., there exists a positive constant σ_1 such that*

$$|g(\tilde{x},\tilde{P}) - g(\hat{x},\hat{P})| \le \sigma_1(\|\tilde{x}-\hat{x}\| + \mathsf{dl}_{K,k}(\tilde{P},\hat{P})),\ \forall \tilde{x},\hat{x} \in X, \tilde{P},\hat{P} \in \mathscr{P}(\Xi); \tag{3.23}$$

(b) $\mathcal{P}(u)$ is locally Lipschitz continuous under the Kantorovich metric at $\bar{u}$, i.e., there exist positive constants $\bar{\delta}$ and σ_2 such that

$$\mathbb{H}(\mathcal{P}(\tilde{u}),\mathcal{P}(\hat{u});\mathsf{dl}_{K,k}) \le \sigma_2\|\tilde{u}-\hat{u}\|_U,\ \ \forall \tilde{u},\hat{u} \in \mathcal{N}(\bar{u},\bar{\delta}), \tag{3.24}$$

where $\mathcal{N}(\bar{u},\bar{\delta}) := \{u' \in U : \|\bar{u}-u'\|_U \le \bar{\delta}\}$. Then

(i) the optimal value function $\vartheta(u)$ is Lipschitz continuous at $\bar{u}$, i.e., there exists a positive constant $\sigma > 0$ such that

$$|\vartheta(\tilde{u}) - \vartheta(\hat{u})| \le \sigma\|\tilde{u}-\hat{u}\|_U,\ \forall \tilde{u},\hat{u} \in \mathcal{N}(\bar{u},\bar{\delta});$$

(ii) if, in addition, (c) $\mathcal{P}(u)$ is locally Lipschitz continuous under Kantorovich metric at every $u \in U$, i.e., there exist δ^u and $\sigma_2' > 0$ such that

$$\mathbb{H}(\mathcal{P}(\tilde{u}),\mathcal{P}(\hat{u});\mathsf{dl}_{K,k}) \le \sigma_2'\|\tilde{u}-\hat{u}\|_U,\ \ \forall \tilde{u},\hat{u} \in \mathcal{N}(u,\delta^u), \tag{3.25}$$

then the optimal value function $\vartheta(u)$ is Lipschitz continuous in u, i.e., there exists a positive constant $\sigma' > 0$ such that

$$|\vartheta(\tilde{u}) - \vartheta(\hat{u})| \le \sigma'\|\tilde{u}-\hat{u}\|_U,\ \forall \tilde{u},\hat{u} \in U.$$

It might be helpful to give some comments about the conditions and the results of the lemma before presenting a proof. Condition (a) requires $g(x,P)$ to be globally Lipschitz continuity w.r.t. x and P. This condition is fulfilled when g is the expected value of a random function, i.e., $\mathbb{E}_P[f(x,\xi)]$ when (i) $f(x,\xi)$ is globally Lipschitz continuous in x for almost every ξ and the Lipschitz modulus is integrably bounded, and (ii) $f(x,\xi)$ is globally Lipschitz continuous uniformly for all $x \in X$, see [20] and references therein. Condition (b) requires local Lipschitz continuity of the ambiguity set mapping w.r.t. variation of parameters. This kind of result may be derived when the ambiguity set takes a specific structure, see [19, 20, 25]. We will come back to this for problem (3.19) in the next subsection.

Lemma 3.1 (i) resembles [36, Theorem 3.1] although the stability result here is derived for a general ambiguity set $\mathcal{P}(u)$. Lemma 3.1 (ii) is about global Lipschitz continuity of the optimal value function. Pichler and Xu [20] derive a similar result to Lemma 3.1 (ii) when the ambiguity set is constructed through ζ-ball, see [20, Theorem 3]. Here we will use Lemma 3.1 to derive global Lipschitz continuity and subsequently quantitative statistical robustness of the optimal value of problem (3.22) when the objective function takes a specific structure, see details in the next subsection.

Proof. The stability results are essentially based on the classical stability results for parametric programming (e.g. Klatte [16], Zhang et al. [35]). Part (i). By the definition of $v(x,u)$, we have

$$\begin{aligned} v(\tilde{x},\tilde{u}) - v(\hat{x},\hat{u}) &= \sup_{\tilde{P}\in\mathcal{P}(\tilde{u})} g(\tilde{x},\tilde{P}) - \sup_{\hat{P}\in\mathcal{P}(\hat{u})} g(\hat{x},\hat{P}) \\ &= \sup_{\tilde{P}\in\mathcal{P}(\tilde{u})} g(\tilde{x},\tilde{P}) - \sup_{\hat{P}\in\mathcal{P}(\hat{u})} g(\tilde{x},\hat{P}) + \sup_{\hat{P}\in\mathcal{P}(\hat{u})} g(\tilde{x},\hat{P}) - \sup_{\hat{P}\in\mathcal{P}(\hat{u})} g(\hat{x},\hat{P}) \\ &\le \sup_{\tilde{P}\in\mathcal{P}(\tilde{u})} \inf_{\hat{P}\in\mathcal{P}(\hat{u})} |g(\tilde{x},\tilde{P}) - g(\tilde{x},\hat{P})| + \sup_{\hat{P}\in\mathcal{P}(\hat{u})} |g(\tilde{x},\hat{P}) - g(\hat{x},\hat{P})| \\ &\le \sigma_1 \sup_{\tilde{P}\in\mathcal{P}(\tilde{u})} \inf_{\hat{P}\in\mathcal{P}(\hat{u})} \mathsf{dl}_{K,k}(\tilde{P},\hat{P}) + \sigma_1\|\tilde{x}-\hat{x}\| \\ &= \sigma_1(\mathbb{D}(\mathcal{P}(\tilde{u}),\mathcal{P}(\hat{u});\mathsf{dl}_{K,k}) + \|\tilde{x}-\hat{x}\|), \end{aligned}$$

where the second inequality is due to (3.23). Likewise, we can obtain

$$v(\hat{x},\hat{u}) - v(\tilde{x},\tilde{u}) \le \sigma_1(\mathbb{D}(\mathcal{P}(\hat{u}),\mathcal{P}(\tilde{u});\mathsf{dl}_{K,k}) + \|\tilde{x}-\hat{x}\|).$$

Combining the above two inequalities and the Lipschitz continuity of $\mathcal{P}(\cdot)$ in (3.24), we have

$$\begin{aligned} |v(\tilde{x},\tilde{u}) - v(\hat{x},\hat{u})| &\le \sigma_1(\mathbb{H}(\mathcal{P}(\tilde{u}),\mathcal{P}(\hat{u});\mathsf{dl}_{K,k}) + \|\tilde{x}-\hat{x}\|) \\ &\le \max\{\sigma_1\sigma_2,\sigma_1\}(\|\tilde{u}-\hat{u}\|_U + \|\tilde{x}-\hat{x}\|). \end{aligned}$$

The conclusion follows by setting $\sigma := \max\{\sigma_1\sigma_2,\sigma_1\}$. The Lipschitz continuity of $\vartheta(\cdot)$ follows from the classical stability results in [16, Theorem 1].

Part (ii). For any $\tilde{u},\hat{u}\in U$, since U is compact, we can construct a δ-net $\{u^1,\cdots,u^{\hat{N}}\}$ in the compact set U such that

$$\big\{u(\lambda) := (1-\lambda)\tilde{u} + \lambda\hat{u} : \lambda\in[0,1]\big\} \subset \bigcup_{j=1}^{J} \mathcal{N}(u^j,\delta^{u^j}),$$

with $J\le\hat{N}$ and $U\subset\cup_{j=1}^{\hat{N}}\mathcal{N}(u^j,\delta^{u^j})$. Specifically, we can select an increasing sequence $\{\lambda_j\}_{j=1}^{J+1}\subset[0,1]$ with $\lambda_1=0$ and $\lambda_{J+1}=1$, such that for $u_{1_j} := (1-\lambda_j)\tilde{u}+\lambda_j\hat{u}$ and $u_{2_j} := (1-\lambda_{j+1})\tilde{u}+\lambda_{j+1}\hat{u}$, $j=1,\cdots,J$, we have

$$u_{2_j} = u_{1_{j+1}} \text{ for } j\in[J-1], \quad \bigcup_{j=1}^{J}[u_{1_j},u_{2_j}] = [\tilde{u},\hat{u}] \quad \text{and} \quad [u_{1_j},u_{2_j}]\subset\mathcal{N}(u^j,\delta^{u^j}) \text{ for } j\in[J],$$

where we write $[J]$ for $\{1,\cdots,J\}$ and $[a,b]$ for the line segment connecting a and b. By Part (i), there exist positive constants σ^{u^j}, $j\in[J]$ such that

$$\big|\vartheta(u_{1_j}) - \vartheta(u_{2_j})\big| \le \sigma^{u^j}\|u_{1_j}-u_{2_j}\|_U, \text{ for } j\in[J].$$

Let $\sigma' := \max_{j\in[\hat{N}]}\sigma^{u^j}$. Then we have

$$\begin{aligned} |\vartheta(\tilde{u}) - \vartheta(\hat{u})| &\le \sum_{j=1}^{J}|\vartheta(u_{1_j}) - \vartheta(u_{2_j})| \le \sum_{j=1}^{J}\sigma^{u^j}\|u_{1_j}-u_{2_j}\|_U \\ &= \sum_{j=1}^{J}\sigma^{u^j}\|((1-\lambda_j)\tilde{u}+\lambda_j\hat{u}) - ((1-\lambda_{j+1})\tilde{u}+\lambda_{j+1}\hat{u})\|_U \\ &\le \max_{j\in[\hat{N}]}\sigma^{u^j}\|\tilde{u}-\hat{u}\|_U = \sigma'\|\tilde{u}-\hat{u}\|_U \end{aligned} \tag{3.26}$$

for all $\tilde{u}, \hat{u} \in U$. ■

3.2 Statistical robustness of problem (3.19)

We now return to discuss statistical robustness of problem (3.19). To ease the exposition, let $u_N := (\mu_N, \Sigma_N, \gamma_1, \gamma_2)$. Let $u^0 := (\mu^0, \Sigma^0, \gamma_1^0, \gamma_2^0) \in \mathbb{R}^k \times \mathbb{R}^{k\times k} \times \mathbb{R} \times \mathbb{R}$ be fixed and $\delta > 0$ be a positive number. Define

$$\mathcal{N}(u^0, \delta) := \left\{ u_N \in \mathbb{R}^k \times \mathbb{R}^{k\times k} \times \mathbb{R} \times \mathbb{R} : \|u_N - u^0\| \leq \delta \right\}, \tag{3.27}$$

where we write succinctly $\|u_N - u^0\|$ for $\|\mu_N - \mu^0\| + \|\Sigma_N - \Sigma^0\| + |\gamma_1 - \gamma_1^0| + |\gamma_2 - \gamma_2^0|$.

Next, we derive the Lipschitz continuity of the optimal value function $\vartheta(u_N) = \min_{x\in X} v(x, u_N)$ of Problem (3.19). We do so by applying Lemma 3.1 to the specific problem (3.19). To this end, we require the Lipschitz continuity of $\mathbb{E}_P[f(x,\xi)]$ and the feasible set-valued mapping $\mathcal{P}(\cdot)$.

Proposition 3.1 *Let $u^0 = (\mu^0, \Sigma^0, \gamma_1^0, \gamma_2^0) \in \mathbb{R}^k \times \mathbb{R}^{k\times k} \times \mathbb{R} \times \mathbb{R}$ be fixed. Assume: (a) Ξ is a compact set, and (b) the Slater condition for the constraint in (3.20) holds for u^0, i.e., there exists $P_{u^0} \in \mathscr{P}(\Xi)$ such that*

$$\mathbb{E}_{P_{u^0}}[\Psi(\xi, u^0)] \in \operatorname{int} \begin{pmatrix} \mathcal{S}_-^{k+1} \\ \mathcal{S}_-^{k} \end{pmatrix}, \tag{3.28}$$

*where "*int*" denotes the interior of a set; (c) $f(x,\xi)$ is globally Lipschitz continuous in (x,ξ), i.e. there exists a positive constant $L_1 > 0$ such that*

$$|f(x,\xi) - f(x',\xi')| \leq L_1(\|x - x'\| + \|\xi - \xi'\|),\ \forall x, x' \in X,\ \xi, \xi' \in \Xi. \tag{3.29}$$

Then the optimal value function $\vartheta(u_N)$ is Lipschitz continuous in u_N, i.e., there exist a positive constant $C_1^{u^0} > 0$ and $\delta_1 > 0$ such that

$$|\vartheta(\tilde{u}_N) - \vartheta(\hat{u}_N)| \leq C_1^{u^0} \|\tilde{u}_N - \hat{u}_N\|$$

for all $\tilde{u}_N, \hat{u}_N, \in \mathcal{N}(u^0, \delta_1)$ with $\delta_1 > 0$.

Proof. We use Lemma 3.1 to prove the result. Thus, it is enough to verify the conditions of the lemma.

First, $\mathbb{E}_P[f(x,\xi)]$ satisfies property (3.23). By the Lipschitz continuity of f in ξ, we have $\frac{1}{L_1} f(x,\cdot) \in \mathscr{G}_L$ for all $x \in X$, and consequently

$$\begin{aligned} |\mathbb{E}_{\tilde{P}}[f(\tilde{x},\xi)] - \mathbb{E}_{\hat{P}}[f(\hat{x},\xi)]| &= |\mathbb{E}_{\tilde{P}}[f(\tilde{x},\xi)] - \mathbb{E}_{\tilde{P}}[f(\hat{x},\xi)]| + |\mathbb{E}_{\tilde{P}}[f(\hat{x},\xi)] - \mathbb{E}_{\hat{P}}[f(\hat{x},\xi)]| \\ &\leq \mathbb{E}_{\tilde{P}}[|f(\tilde{x},\xi) - f(\hat{x},\xi)|] + L_1 \sup_{g\in\mathscr{G}_L} |\mathbb{E}_{\tilde{P}}[g(\xi)] - \mathbb{E}_{\hat{P}}[g(\xi)]| \\ &\leq L_1\|\tilde{x} - \hat{x}\| + L_1 \mathsf{dl}_{K,k}(\tilde{P}, \hat{P}), \end{aligned}$$

where the last inequality is due to (3.29), and this verifies (3.23).

Second, $\mathcal{P}(u_N)$ satisfies the property (3.24). Since Ξ is compact, then $\mathscr{P}(\Xi)$ is weakly compact. By the Kantorovich-Rubinstein theorem, $\mathsf{dl}_{W,k}$ coincides with the Kantorovich metric $\mathsf{dl}_{K,k}$ when (Ξ, d) is a metric space. It follows by [11, Theorem 4] that

$$\mathsf{dl}_{K,k}(\tilde{P}, \hat{P}) \leq \operatorname{diam}(\Xi)\mathsf{dl}_{TV}(\tilde{P}, \hat{P}),\ \forall \tilde{P}, \hat{P} \in \mathscr{P}(\Xi), \tag{3.30}$$

where $\operatorname{diam}(\Xi) := \sup_{\xi,\xi' \in \Xi} \|\xi - \xi'\|$, and $\mathsf{dl}_{TV}(\tilde{P}, \hat{P}) := \sup_{g \in \mathscr{G}} |\mathbb{E}_{\tilde{P}}[g(\xi)] - \mathbb{E}_{\hat{P}}[g(\xi)]|$, where

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

Next, we estimate $\mathsf{dl}_{TV}(\tilde{P}, \hat{P})$. We do so by utilizing [36, Theorem 2.1]. For this, we need to show the Lipschitz continuity of Ψ w.r.t. parameters $(\mu_N, \Sigma_N, \gamma_1, \gamma_2)$. Let $\tilde{u}_N, \hat{u}_N \in \mathcal{N}(u^0, \delta)$. Observe that

$$\begin{aligned}
\|\Psi(\xi, \tilde{u}_N) - \Psi(\xi, \hat{u}_N)\| &= \left\| \begin{pmatrix} \begin{bmatrix} -\tilde{\Sigma}_N & \tilde{\mu}_N - \xi \\ (\tilde{\mu}_N - \xi)^T & -\tilde{\gamma}_1 \end{bmatrix} \\ (\xi - \tilde{\mu}_N)(\xi - \tilde{\mu}_N)^T - \tilde{\gamma}_2 \tilde{\Sigma}_N \end{pmatrix} - \begin{pmatrix} \begin{bmatrix} -\hat{\Sigma}_N & \hat{\mu}_N - \xi \\ (\hat{\mu}_N - \xi)^T & -\hat{\gamma}_1 \end{bmatrix} \\ (\xi - \hat{\mu}_N)(\xi - \hat{\mu}_N)^T - \hat{\gamma}_2 \hat{\Sigma}_N \end{pmatrix} \right\| \\
&\leq \left\| \begin{bmatrix} -\tilde{\Sigma}_N & \tilde{\mu}_N - \xi \\ (\tilde{\mu}_N - \xi)^T & -\tilde{\gamma}_1 \end{bmatrix} - \begin{bmatrix} -\hat{\Sigma}_N & \hat{\mu}_N - \xi \\ (\hat{\mu}_N - \xi)^T & -\hat{\gamma}_1 \end{bmatrix} \right\| \\
&\quad + \|(\xi - \tilde{\mu}_N)(\xi - \tilde{\mu}_N)^T - \tilde{\gamma}_2 \tilde{\Sigma}_N - (\xi - \hat{\mu}_N)(\xi - \hat{\mu}_N)^T + \hat{\gamma}_2 \hat{\Sigma}_N\| \\
&\leq \| - \tilde{\Sigma}_N + \hat{\Sigma}_N\| + 2\|\tilde{\mu}_N - \hat{\mu}_N\| + | - \tilde{\gamma}_1 + \hat{\gamma}_1| \\
&\quad + 2\|\xi\|\|\tilde{\mu}_N - \hat{\mu}_N\| + \|\tilde{\mu}_N \tilde{\mu}_N^T - \hat{\mu}_N \hat{\mu}_N^T\| + \| - \tilde{\gamma}_2 \tilde{\Sigma}_N + \hat{\gamma}_2 \hat{\Sigma}_N\| \\
&\leq \| - \tilde{\Sigma}_N + \hat{\Sigma}_N\| + 2\|\tilde{\mu}_N - \hat{\mu}_N\| + | - \tilde{\gamma}_1 + \hat{\gamma}_1| \\
&\quad + 2\|\xi\|\|\tilde{\mu}_N - \hat{\mu}_N\| + \|\tilde{\mu}_N\|\|\tilde{\mu}_N - \hat{\mu}_N\| + \|\hat{\mu}_N\|\|\tilde{\mu}_N - \hat{\mu}_N\| \\
&\quad + |\tilde{\gamma}_2|\|\tilde{\Sigma}_N - \hat{\Sigma}_N\| + \|\hat{\Sigma}_N\||\tilde{\gamma}_2 - \hat{\gamma}_2| \\
&\leq \max\left\{ |\tilde{\gamma}_2| + 1, 4 + \|\tilde{\mu}_N\| + \|\hat{\mu}_N\|, \|\hat{\Sigma}_N\| \right\} \max\{1, \|\xi\|\} \|\tilde{u}_N - \hat{u}_N\| \\
&\leq C_{u^0} \max\{1, \|\xi\|\} \|\tilde{u}_N - \hat{u}_N\|,
\end{aligned} \tag{3.31}$$

where C_{u^0} depends on u_0 and δ, the last inequality follows from the fact that $|\tilde{\gamma}_2|, \|\tilde{\mu}_N\|, \|\hat{\mu}_N\|, \|\hat{\Sigma}_N\| \leq \|u_0\| + \delta$. By inequality (3.31) and the Slater condition (condition (b)), we have, by [36, Theorem 2.1], that there exist positive constants $\bar{C}_{u_0} > 0$ and $\delta_1 > 0$ such that

$$\mathbb{H}(\mathcal{P}(\tilde{u}_N), \mathcal{P}(\hat{u}_N); \mathsf{dl}_{TV}) \leq \bar{C}_{u_0} \|\tilde{u}_N - \hat{u}_N\|,\ \forall \tilde{u}_N, \hat{u}_N, \in \mathcal{N}(u^0, \delta_1). \tag{3.32}$$

Combining (3.30) and (3.32), we have

$$\mathbb{H}(\mathcal{P}(\tilde{u}_N), \mathcal{P}(\hat{u}_N); \mathsf{dl}_{K,k}) \leq \operatorname{diam}(\Xi) \bar{C}_{u_0} \|\tilde{u}_N - \hat{u}_N\|, \tag{3.33}$$

which shows that $\mathcal{P}(u_N)$ satisfies the property (3.24). The conclusion follows by Lemma 3.1 with $C_1^{u_0} := \max\{\operatorname{diam}(\Xi)\bar{C}_{u_0} L_1, L_1\}$. ■

The Slater condition (3.28) plays a crucial role in the derivation of error bound (3.33). Unfortunately the error bound holds only in a neighborhood of u^0 which is inadequate for us to establish the global Lipschitz continuity of $v(u)$. To address the issue, we impose a stronger condition in the following proposition.

Proposition 3.2 *Assume:(a) Ξ is a compact set; (b) the Slater condition for the constraint in (3.20) holds for every $u \in \Xi \times \{\xi\xi^T : \xi \in \Xi\} \times \{\gamma_1\} \times \{\gamma_2\}$, i.e., there exists $P_u \in \mathscr{P}(\Xi)$ such that*

$$\mathbb{E}_{P_u}[\Psi(\xi, u)] \in \text{int} \begin{pmatrix} \mathcal{S}_-^{k+1} \\ \mathcal{S}_-^{k} \end{pmatrix}, \forall u \in \Xi \times \{\xi\xi^T : \xi \in \Xi\} \times \{\gamma_1\} \times \{\gamma_2\}, \tag{3.34}$$

where $\mathscr{P}(\Xi)$ is the set of all probability distributions of ξ whose support sets are contained in Ξ (or alternatively the set of all probability measures on Ξ induced by mapping ξ); (c) the condition (c) in Proposition 3.1. Then the optimal value function $\vartheta(u_N)$ is globally Lipschitz continuous in u_N, i.e., there exists a positive constant $C_2 > 0$ such that

$$|\vartheta(\tilde{u}_N) - \vartheta(\hat{u}_N)| \leq C_2 \|\tilde{u}_N - \hat{u}_N\|$$

for all $\tilde{u}_N, \hat{u}_N \in \Xi \times \{\xi\xi^T : \xi \in \Xi\} \times \{\gamma_1\} \times \{\gamma_2\}$.

Proof. We use Lemma 3.1 (ii) to prove the result. It suffices to show $\mathcal{P}(u_N)$ is locally Lipschitz continuous under Kantorovich metric at every $u \in \Xi \times \{\xi\xi^T : \xi \in \Xi\} \times \{\gamma_1\} \times \{\gamma_2\}$. This property is guaranteed by Proposition 3.1 under conditions (a)-(c). Thus the conclusion follows from Lemma 3.1 (ii). ■

Under conditions in Proposition 3.2, we are able to establish the main result of this section. Let us write $\vec{\xi}^N$ for $(\xi^1, \cdots, \xi^N)$, $\hat{\vartheta}(\vec{\xi}^N)$ for $\vartheta(u_N)$ and $\hat{\vartheta}(\vec{\tilde{\xi}}^N)$ for $\vartheta(\tilde{u}_N)$ to indicate their dependence on $\vec{\xi}$ and $\vec{\tilde{\xi}}$ respectively. Then $\hat{\vartheta}(\vec{\xi})$ and $\hat{\vartheta}(\vec{\tilde{\xi}})$ are two statistical estimators of the optimal value of (DRO-moment) and we are interested in the difference between laws of the two estimators, that is, the difference between $Q^{\otimes N} \circ \hat{\vartheta}(\vec{\xi})^{-1}$ and $P^{\otimes N} \circ \hat{\vartheta}(\vec{\tilde{\xi}})^{-1}$, where Q and P are the probability measures of ξ and $\tilde{\xi}$ respectively. The next theorem addresses this.

Theorem 3.1 (Quantitative statistical robustness of model (3.19)) *Assume the settings and conditions in Proposition 3.2. Let γ_1, γ_2 be fixed and $P, Q \in \mathcal{M}_k^{\phi}$, where $\phi(t) := \|t\|$. Then there exists a constant $C_3 > 0$ such that*

$$\mathsf{dl}_{K,1}\left(P^{\otimes N} \circ \hat{\vartheta}^{-1}, Q^{\otimes N} \circ \hat{\vartheta}^{-1}\right) \leq C_3 \mathsf{dl}_{K,k}(P, Q) \tag{3.35}$$

for all $N \in \mathbb{N}$.

Proof. Based on the Lipschitz continuity of $\vartheta(\cdot)$ in Proposition 3.2, we can obtain the statistical robustness of the estimator $\hat{\vartheta}(\cdot)$ in the following result. By the definition of $\mathsf{dl}_{K,1}$,

$$\begin{aligned} & \mathsf{dl}_{K,1}(P^{\otimes N} \circ \hat{\vartheta}^{-1}, Q^{\otimes N} \circ \hat{\vartheta}^{-1}) \\ = & \sup_{g \in \mathscr{G}_L} \left| \int_{\mathbb{R}} g(t) P^{\otimes N} \circ \hat{\vartheta}^{-1}(dt) - \int_{\mathbb{R}} g(t) Q^{\otimes N} \circ \hat{\vartheta}^{-1}(dt) \right| \\ = & \sup_{g \in \mathscr{G}_L} \left| \int_{(\Xi)^{\otimes N}} g(\hat{\vartheta}(\vec{\xi}^N)) P^{\otimes N}(d\vec{\xi}^N) - \int_{(\Xi)^{\otimes N}} g(\hat{\vartheta}(\vec{\xi}^N)) Q^{\otimes N}(d\vec{\xi}^N) \right|, \end{aligned} \tag{3.36}$$

where $\mathscr{G}_L$ is defined in (1.7). To show (3.35), it suffices to show the Lipschitz continuity of $g(\hat{\vartheta}(\vec{\xi^N}))$ and well-definedness of the integrals. Let $R_1 := \|\tilde{\mu}_N - \hat{\mu}_N\|$ and $R_2 := \|\tilde{\Sigma}_N - \hat{\Sigma}_N\|$. Then

$$R_1 = \left\| \frac{1}{N}\sum_{i=1}^N \tilde{\xi}^i - \frac{1}{N}\sum_{i=1}^N \hat{\xi}^i \right\| \le \frac{1}{N}\sum_{i=1}^N \|\tilde{\xi}^i - \hat{\xi}^i\| \tag{3.37}$$

and

$$\begin{aligned}
R_2 &= \left\| \frac{1}{N}\sum_{i=1}^N (\tilde{\xi}^i - \tilde{\mu}_N)(\tilde{\xi}^i - \tilde{\mu}_N)^T - \frac{1}{N}\sum_{i=1}^N (\hat{\xi}^i - \hat{\mu}_N)(\hat{\xi}^i - \hat{\mu}_N)^T \right\| \\
&\le \frac{1}{N}\sum_{i=1}^N \left\| (\tilde{\xi}^i - \tilde{\mu}_N)(\tilde{\xi}^i - \tilde{\mu}_N)^T - (\hat{\xi}^i - \hat{\mu}_N)(\hat{\xi}^i - \hat{\mu}_N)^T \right\| \\
&= \frac{1}{N}\sum_{i=1}^N \left\| \tilde{\xi}^i(\tilde{\xi}^i)^T - \tilde{\xi}^i\tilde{\mu}_N^T - \tilde{\mu}_N(\tilde{\xi}^i)^T + \tilde{\mu}_N\tilde{\mu}_N^T - \hat{\xi}^i(\hat{\xi}^i)^T - \hat{\xi}^i\hat{\mu}_N^T - \hat{\mu}_N(\hat{\xi}^i)^T + \hat{\mu}_N\hat{\mu}_N^T \right\| \\
&= \frac{1}{N}\sum_{i=1}^N \Big(\|\tilde{\xi}^i\|\|\tilde{\xi}^i - \hat{\xi}^i\| + \|\hat{\xi}^i\|\|\tilde{\xi}^i - \hat{\xi}^i\| + 2\|\tilde{\xi}^i\|\|\tilde{\mu}_N - \hat{\mu}_N\| + 2\|\hat{\mu}_N\|\|\tilde{\xi}^i - \hat{\xi}^i\| \\
&\qquad + \|\tilde{\mu}_N\|\|\tilde{\mu}_N - \hat{\mu}_N\| + \|\hat{\mu}_N\|\|\tilde{\mu}_N - \hat{\mu}_N\| \Big) \\
&\le \frac{3}{N} \sup_{i=1,\cdots,N} \Big\{ 2\|\tilde{\xi}^i\|, \|\hat{\xi}^i\|, \|\tilde{\mu}_N\|, 2\|\hat{\mu}_N\| \Big\} \sum_{i=1}^N \Big(\|\tilde{\xi}^i - \hat{\xi}^i\| + \|\tilde{\mu}_N - \hat{\mu}_N\| \Big) \\
&\le \frac{6}{N} \sup_{i=1,\cdots,N} \Big\{ 2\|\tilde{\xi}^i\|, \|\hat{\xi}^i\|, \|\tilde{\mu}_N\|, 2\|\hat{\mu}_N\| \Big\} \sum_{i=1}^N \|\tilde{\xi}^i - \hat{\xi}^i\|.
\end{aligned} \tag{3.38}$$

Combining inequalities (3.37)-(3.38), we have

$$R_1 + R_2 \le \left(\frac{1}{N} + \frac{6}{N} \sup_{i=1,\cdots,N} \Big\{ 2\|\tilde{\xi}^i\|, \|\hat{\xi}^i\|, \|\tilde{\mu}_N\|, 2\|\hat{\mu}_N\| \Big\} \right) \sum_{i=1}^N \|\tilde{\xi}^i - \hat{\xi}^i\|. \tag{3.39}$$

Since $f(x,\xi)$ is uniformly Lipschitz continuous in ξ and g is also Lipschitz continuous with modulus being bounded by 1, it follows by (3.39) that

$$\begin{aligned}
|g(\hat{\vartheta}(\tilde{\xi}^1, \cdots, \tilde{\xi}^N)) - g(\hat{\vartheta}(\hat{\xi}^1, \cdots, \hat{\xi}^N))| &\le |\hat{\vartheta}(\tilde{\xi}^1, \cdots, \tilde{\xi}^N) - \hat{\vartheta}(\hat{\xi}^1, \cdots, \hat{\xi}^N)| \\
&= |\vartheta(\tilde{u}_N) - \vartheta(\hat{u}_N)| \\
&\le C_2(R_1 + R_2) \\
&\le \frac{C_3}{N}\sum_{i=1}^N \|\tilde{\xi}^i - \hat{\xi}^i\|,
\end{aligned}$$

where $C_3 := C_2 + 12C_2(\|\xi_0\| + \mathrm{diam}(\Xi))$, $\hat{C}_2$ is defined as in Proposition 3.2 and ξ_0 is some fixed element in Ξ.

This means that $g(\hat{\vartheta}(\cdot))$ is Lipschitz continuous over $(\Xi)^{\otimes N}$ with Lipschitz modulus bounded by C_3, and thus the well-definedness of $\int_{(\Xi)^{\otimes N}} g(\hat{\vartheta}(\vec{\xi^N}))P^{\otimes N}(d\vec{\xi^N})$ and $\int_{(\Xi)^{\otimes N}} g(\hat{\vartheta}(\vec{\xi^N}))Q^{\otimes N}(d\vec{\xi^N})$ can be deduced from Theorem 2.1. The rest follows from Theorem 2.1 by setting $T_N(\xi^1, \cdots, \xi^N) = \hat{\vartheta}(\xi^1, \cdots, \xi^N)$ in the theorem. ■

4 Statistical robustness of the DRO models with ζ-ball

In this section, we consider a DRO model where the ambiguity set is constructed by ζ-ball centered at Q_N with some fixed radius r:

$$\text{(DRO-ball)} \quad \min_{x\in X} \max_{P'\in\mathcal{B}_{\mathscr{H}_1}(Q_N,r)} \mathbb{E}_{P'}[f(x,\xi)], \tag{4.40}$$

where

$$\mathcal{B}_{\mathscr{H}_1}(Q_N,r) := \{P'\in\mathscr{P}(\Xi) : \mathsf{dl}_{\mathscr{H}_1}(P',Q_N)\le r\}$$

and $\mathscr{H}_1 := \{f(x,\cdot) : x\in X\}$. This corresponds to Example 1.2 with $\mathscr{G} := \mathscr{H}_1$. Let $\vartheta(Q_N)$ denote the optimal value of problem (4.40) and $\vartheta(P_N)$ the one with Q_N be replaced by P_N, that is,

$$\begin{aligned}\vartheta(Q_N) &:= \min_{x\in X} \max_{P'\in\mathcal{B}_{\mathscr{H}_1}(Q_N,r)} \mathbb{E}_{P'}[f(x,\xi)],\\ \vartheta(P_N) &:= \min_{x\in X} \max_{P'\in\mathcal{B}_{\mathscr{H}_1}(P_N,r)} \mathbb{E}_{P'}[f(x,\xi)].\end{aligned}$$

Then $\vartheta(Q_N)$ and $\vartheta(P_N)$ are two statistical estimators of the optimal value of (DRO-ball) and we are interested in the difference between laws of the two estimators, that is, the difference between $Q^{\otimes N}\circ\vartheta(Q_N)^{-1}$ and $P^{\otimes N}\circ\vartheta(P_N)^{-1}$. We write $\vec{\xi}^N$ for $(\xi^1,\cdots,\xi^N)$ and $\hat{\vartheta}(\vec{\xi}^N)$ for $\vartheta(P_N)$ to indicate its dependence on $\xi^1,\cdots,\xi^N$.

Theorem 4.1 (Quantitative statistical robustness of model (4.40)) *Assume that $f(x,\xi)$ is continuous in x for each fixed ξ and is globally Lipschitz continuous in ξ uniformly for $x\in X$, i.e. there exists a positive constant $L_1>0$ such that*

$$\sup_{x\in X}|f(x,\xi)-f(x,\xi')|\le L_1\|\xi-\xi'\|,\ \forall\xi,\xi'\in\Xi.$$

Let $P,Q\in\mathcal{M}_k^\phi$, where $\phi(t)=\|t\|$. Then

$$\mathsf{dl}_{K,1}\left(P^{\otimes N}\circ\hat{\vartheta}_N^{-1}, Q^{\otimes N}\circ\hat{\vartheta}_N^{-1}\right)\le L_1\mathsf{dl}_{K,k}(P,Q) \tag{4.41}$$

for all $N\in\mathbb{N}$.

Proof. We will show (4.41) by applying Theorem 2.1. Thus it suffices to verify the Lipschitz continuity of $\hat{\vartheta}(\tilde{\xi}^1,\cdots,\tilde{\xi}^N)$ w.r.t. $(\tilde{\xi}^1,\cdots,\tilde{\xi}^N)$. Since $f(x,\xi)$ is uniformly Lipschitz continuous in ξ and then $\frac{\mathscr{H}_1}{L_1}\subset\mathscr{G}_L$, which implies that

$$\mathsf{dl}_{\mathscr{H}_1}(\tilde{Q}_N,\hat{Q}_N)\le L_1\sup_{h\in\mathscr{G}_L}|\mathbb{E}_{\tilde{Q}_N}[h(\xi)]-\mathbb{E}_{\hat{Q}_N}[h(\xi)]|,$$

where $\tilde{Q}_N := \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{\xi}^i}$ and $\hat{Q}_N := \frac{1}{N}\sum_{i=1}^N \delta_{\hat{\xi}^i}$. We have

$$\begin{aligned}
\hat{\vartheta}(\tilde{\xi}^1,\cdots,\tilde{\xi}^N) - \hat{\vartheta}(\hat{\xi}^1,\cdots,\hat{\xi}^N) &= \min_{x\in X} \sup_{\tilde{P}\in\mathcal{B}_{\mathscr{H}_1}(\tilde{Q}_N,r)} \mathbb{E}_{\tilde{P}}[f(x,\xi)] - \min_{x\in X} \sup_{\hat{P}\in\mathcal{B}_{\mathscr{H}_1}(\hat{Q}_N,r)} \mathbb{E}_{\hat{P}}[f(x,\xi)] \\
&\le \sup_{x\in X}\left(\sup_{\tilde{P}\in\mathcal{B}_{\mathscr{H}_1}(\tilde{Q}_N,r)} \mathbb{E}_{\tilde{P}}[f(x,\xi)] - \sup_{\hat{P}\in\mathcal{B}_{\mathscr{H}_1}(\hat{Q}_N,r)} \mathbb{E}_{\hat{P}}[f(x,\xi)]\right) \\
&\le \sup_{x\in X} \sup_{\tilde{P}\in\mathcal{B}_{\mathscr{H}_1}(\tilde{Q}_N,r)} \inf_{\hat{P}\in\mathcal{B}_{\mathscr{H}_1}(\hat{Q}_N,r)} \left|\mathbb{E}_{\tilde{P}}[f(x,\xi)] - \mathbb{E}_{\hat{P}}[f(x,\xi)]\right| \\
&\le \sup_{x\in X} \sup_{\tilde{P}\in\mathcal{B}_{\mathscr{H}_1}(\tilde{Q}_N,r)} \inf_{\hat{P}\in\mathcal{B}_{\mathscr{H}_1}(\hat{Q}_N,r)} \sup_{h\in\mathscr{H}_1} \left|\mathbb{E}_{\tilde{P}}[h(\xi)] - \mathbb{E}_{\hat{P}}[h(\xi)]\right| \\
&= \sup_{x\in X} \mathbb{D}(\mathcal{B}_{\mathscr{H}_1}(\tilde{Q}_N,r), \mathcal{B}_{\mathscr{H}_1}(\hat{Q}_N,r); \mathsf{dl}_{\mathscr{H}_1}) \\
&\le \sup_{x\in X} \mathbb{H}(\mathcal{B}_{\mathscr{H}_1}(\tilde{Q}_N,r), \mathcal{B}_{\mathscr{H}_1}(\hat{Q}_N,r); \mathsf{dl}_{\mathscr{H}_1}) \\
&\le \sup_{x\in X} \mathsf{dl}_{\mathscr{H}_1}(\tilde{Q}_N,\hat{Q}_N) \\
&\le L_1 \sup_{h\in\mathscr{G}_L} |\mathbb{E}_{\tilde{Q}_N}[h(\xi)] - \mathbb{E}_{\hat{Q}_N}[h(\xi)]| \\
&= L_1 \sup_{h\in\mathscr{G}_L} \left|\frac{1}{N}\sum_{i=1}^N h(\tilde{\xi}^i) - \frac{1}{N}\sum_{i=1}^N h(\hat{\xi}^i)\right| \\
&\le \frac{L_1}{N}\sum_{i=1}^N \|\tilde{\xi}^i - \hat{\xi}^i\|,
\end{aligned}$$

where the last third inequality is due to [20, Theorem 1]. By swapping the roles of $\hat{\vartheta}(\tilde{\xi}^1,\cdots,\tilde{\xi}^N)$ and $\hat{\vartheta}(\hat{\xi}^1,\cdots,\hat{\xi}^N)$, we obtain

$$\begin{aligned}
\hat{\vartheta}(\hat{\xi}^1,\cdots,\hat{\xi}^N) - \hat{\vartheta}(\tilde{\xi}^1,\cdots,\tilde{\xi}^N) &= \sup_{x\in X} \mathbb{D}(\mathcal{B}_{\mathscr{H}_1}(\hat{Q}_N,r), \mathcal{B}_{\mathscr{H}_1}(\tilde{Q}_N,r); \mathsf{dl}_{\mathscr{H}_1}) \\
&\le \sup_{x\in X} \mathbb{H}(\mathcal{B}_{\mathscr{H}_1}(\tilde{Q}_N,r), \mathcal{B}_{\mathscr{H}_1}(\hat{Q}_N,r); \mathsf{dl}_{\mathscr{H}_1}) \\
&\le \frac{L_1}{N}\sum_{i=1}^N \|\tilde{\xi}^i - \hat{\xi}^i\|,
\end{aligned}$$

and thus

$$|\hat{\vartheta}(\tilde{\xi}^1,\cdots,\tilde{\xi}^N) - \hat{\vartheta}(\hat{\xi}^1,\cdots,\hat{\xi}^N)| \le \frac{L_1}{N}\sum_{i=1}^N \|\tilde{\xi}^i - \hat{\xi}^i\|.$$

This means that $\hat{\vartheta}(\cdot)$ is Lipschitz continuous over $(\mathbb{R}^k)^{\otimes N}$ with Lipschitz modulus being bounded by L_1/N. The well-definedness of $\int_{(\Xi)^{\otimes N}} g(\hat{\vartheta}(\vec{\xi}^N))P^{\otimes N}(d\vec{\xi}^N)$ and $\int_{(\Xi)^{\otimes N}} g(\hat{\vartheta}(\vec{\xi}^N))Q^{\otimes N}(d\vec{\xi}^N)$ can be subsequently deduced as in the proof of Theorem 2.1. The rest follows from Theorem 2.1 by setting $T_N(\xi^1,\cdots,\xi^N) = g(\hat{\vartheta}(\xi^1,\cdots,\xi^N))$ in the theorem. ■

As an application [1] we consider the DRO model in machine learning as described in Exam-

[1]This is not a direct application as the ambiguity set is defined in a different manner. However, the DRO model in machine is essentially under the DRO framework where the ambiguity set is of ζ-structure.

ple 1.3:

$$\inf_{\boldsymbol{w}} \sup_{P' \in \mathcal{B}(P_N, r)} \mathbb{E}_{P'}[c(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)], \tag{4.42}$$

where $\boldsymbol{x} \in \mathbb{X} \subset \mathbb{R}^{k-1}$ and $y \in \mathbb{Y} \subset \mathbb{R}$ and

$$\mathcal{B}(P_N, r) = \{P' \in \mathscr{P}(\Xi) : \mathsf{dl}_{K,k}(P', P_N) \leq r\}.$$

Note that the Kantorovich ball coincides with the Wasserstein ball $\hat{\mathcal{B}}(P_N, r)$ defined in (1.10) and it is a special ζ-ball [20].

In practice, the perceived sample data may be contaminated which means that they are not real data generated by the true distribution P, rather they are generated by some distribution Q which is a perturbation of P. This motivates us to investigate statistical robustness the optimal value of the DRO problem (4.42). Let

$$\begin{aligned} \vartheta(P_N) &:= \inf_{\boldsymbol{w}} \sup_{P' \in \mathcal{B}(P_N, r)} \mathbb{E}_{P'}[c(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)], \\ \vartheta(Q_N) &:= \inf_{\boldsymbol{w}} \sup_{P' \in \mathcal{B}(Q_N, r)} \mathbb{E}_{P'}[c(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)]. \end{aligned}$$

We are interested in the difference between laws of the two estimators, that is, the difference between $Q^{\otimes N} \circ \vartheta(Q_N)^{-1}$ and $P^{\otimes N} \circ \vartheta(P_N)^{-1}$. We can write $\vec{\xi}^N$ for $(\xi^1, \cdots, \xi^N)$ and write $\hat{\vartheta}(\vec{\xi}^N)$ for $\vartheta(P_N)$ and $\hat{\vartheta}(\vec{\tilde{\xi}}^N)$ for $\vartheta(Q_N)$ to indicate their dependence on $\vec{\xi}$ and $\vec{\tilde{\xi}}$ respectively. Let

$$\hat{c}(\xi, \boldsymbol{w}) := c(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y), \text{ where } \xi = (\boldsymbol{x}, y),$$

and define $\mathscr{H}_2 := \{\hat{c}(\cdot, \boldsymbol{w}) : \boldsymbol{w} \in \mathbb{R}^{k-1}\}$.

Proposition 4.1 (Quantitative statistical robustness of model (4.42)) *Assume that $\hat{c}(\xi, \boldsymbol{w})$ is globally Lipschitz continuous in ξ uniformly for $\boldsymbol{w} \in \mathbb{R}^{k-1}$, i.e. there exists a positive constant $L_2 > 0$ such that*

$$\sup_{\boldsymbol{w} \in \mathbb{R}^{k-1}} |\hat{c}(\xi, \boldsymbol{w}) - \hat{c}(\xi', \boldsymbol{w})| \leq L_2 \|\xi - \xi'\|, \ \forall \xi, \xi' \in \Xi = \mathbb{X} \times \mathbb{Y}. \tag{4.43}$$

Let $P, Q \in \mathcal{M}_k^{\phi}$, where $\phi(t) = \|t\|$, $t \in \mathbb{R}^k$. Then

$$\mathsf{dl}_{K,1}\left(P^{\otimes N} \circ \hat{\vartheta}^{-1}, Q^{\otimes N} \circ \hat{\vartheta}^{-1}\right) \leq L_2 \mathsf{dl}_{K,k}(P, Q) \tag{4.44}$$

for all $N \in \mathbb{N}$.

Proof. We show (4.44) by applying Theorem 2.1, where the Lipschitz continuity of $\hat{\vartheta}(\tilde{\xi}^1, \cdots, \tilde{\xi}^N)$ w.r.t. $(\tilde{\xi}^1, \cdots, \tilde{\xi}^N)$ is needed. Since $\hat{c}(\xi, \boldsymbol{w})$ is Lipschitz continuous in ξ uniformly for $\boldsymbol{w}$ in (4.43), we have $\frac{\mathscr{H}_2}{L_2} \subset \mathscr{G}_L$, which implies that for any $\tilde{P}, \hat{P} \in \mathscr{P}(\Xi)$,

$$\mathsf{dl}_{\mathscr{H}_2}(\tilde{P}, \hat{P}) \leq L_2 \sup_{h \in \mathscr{G}_L} |\mathbb{E}_{\tilde{P}}[h(\xi)] - \mathbb{E}_{\hat{P}}[h(\xi)]|. \tag{4.45}$$

Analogous to the proof of Theorem 4.1, we have

$$\begin{aligned}
\hat{\vartheta}(\tilde{\xi}^1,\cdots,\tilde{\xi}^N)-\hat{\vartheta}(\hat{\xi}^1,\cdots,\hat{\xi}^N) &= \inf_{\boldsymbol{w}}\sup_{\tilde{P}\in\mathcal{B}(\tilde{Q}_N,r)}\mathbb{E}_{\tilde{P}}[c(\langle\boldsymbol{w},\boldsymbol{x}\rangle,y)]-\inf_{\boldsymbol{w}}\sup_{\hat{P}\in\mathcal{B}(\hat{Q}_N,r)}\mathbb{E}_{\hat{P}}[c(\langle\boldsymbol{w},\boldsymbol{x}\rangle,y)]\\
&\le \sup_{\boldsymbol{w}}\left(\sup_{\tilde{P}\in\mathcal{B}(\tilde{Q}_N,r)}\mathbb{E}_{\tilde{P}}[c(\langle\boldsymbol{w},\boldsymbol{x}\rangle,y)]-\sup_{\hat{P}\in\mathcal{B}(\hat{Q}_N,r)}\mathbb{E}_{\hat{P}}[c(\langle\boldsymbol{w},\boldsymbol{x}\rangle,y)]\right)\\
&= \sup_{\boldsymbol{w}}\sup_{\tilde{P}\in\mathcal{B}(\tilde{Q}_N,r)}\inf_{\hat{P}\in\mathcal{B}(\hat{Q}_N,r)}(\mathbb{E}_{\tilde{P}}[c(\langle\boldsymbol{w},\boldsymbol{x}\rangle,y)]-\mathbb{E}_{\hat{P}}[c(\langle\boldsymbol{w},\boldsymbol{x}\rangle,y)])\\
&\le \sup_{\boldsymbol{w}}\sup_{\tilde{P}\in\mathcal{B}(\tilde{Q}_N,r)}\inf_{\hat{P}\in\mathcal{B}(\hat{Q}_N,r)}\sup_{h\in\mathscr{H}_2}|\mathbb{E}_{\tilde{P}}[h(\xi)]-\mathbb{E}_{\hat{P}}[h(\xi)]|\\
&\le \sup_{\tilde{P}\in\mathcal{B}(\tilde{Q}_N,r)}\inf_{\hat{P}\in\mathcal{B}(\hat{Q}_N,r)}L_2\sup_{h\in\mathscr{G}_L}|\mathbb{E}_{\tilde{P}}[h(\xi)]-\mathbb{E}_{\hat{P}}[h(\xi)]|\\
&= L_2\mathbb{D}(\mathcal{B}(\tilde{Q}_N,r),\mathcal{B}(\hat{Q}_N,r);\mathsf{dl}_{K,k})\\
&\le L_2\mathbb{H}(\mathcal{B}(\tilde{Q}_N,r),\mathcal{B}(\hat{Q}_N,r);\mathsf{dl}_{K,k})\\
&\le L_2\mathsf{dl}_{K,k}(\tilde{Q}_N,\hat{Q}_N)\\
&= L_2\sup_{h\in\mathscr{G}_L}\left|\frac{1}{N}\sum_{i=1}^N h(\tilde{\xi}^i)-\frac{1}{N}\sum_{i=1}^N h(\hat{\xi}^i)\right|\\
&\le \frac{L_2}{N}\sum_{i=1}^N\|\tilde{\xi}^i-\hat{\xi}^i\|,
\end{aligned}$$

where the second last inequality is due to the Lipschitz continuity of ζ-ball $\mathcal{B}(\cdot,r)$ in [20, Theorem 1], and the third inequality is due to (4.45). By swapping the positions of $\hat{\vartheta}(\tilde{\xi}^1,\cdots,\tilde{\xi}^N)$ and $\hat{\vartheta}(\hat{\xi}^1,\cdots,\hat{\xi}^N)$, we obtain

$$\begin{aligned}
\hat{\vartheta}(\hat{\xi}^1,\cdots,\hat{\xi}^N)-\hat{\vartheta}(\tilde{\xi}^1,\cdots,\tilde{\xi}^N) &\le \mathbb{D}(\mathcal{B}(\hat{Q}_N,r),\mathcal{B}(\tilde{Q}_N,r);\mathsf{dl}_{K,k})\\
&\le L_2\mathbb{H}(\mathcal{B}(\tilde{Q}_N,r),\mathcal{B}(\hat{Q}_N,r);\mathsf{dl}_{K,k})\\
&\le \frac{L_2}{N}\sum_{i=1}^N\|\tilde{\xi}^i-\hat{\xi}^i\|.
\end{aligned}$$

Summarizing the discussions above, we have

$$|\hat{\vartheta}(\tilde{\xi}^1,\cdots,\tilde{\xi}^N)-\hat{\vartheta}(\hat{\xi}^1,\cdots,\hat{\xi}^N)|\le\frac{L_2}{N}\sum_{i=1}^N\|\tilde{\xi}^i-\hat{\xi}^i\|.$$

The above relations mean that $\hat{\vartheta}(\cdot)$ is Lipschitz continuous over $(\Xi)^{\otimes N}$ with Lipschitz modulus bounded by L_2/N, and thus the well-definedness of $\int_{(\Xi)^{\otimes N}}g(\hat{\vartheta}(\vec{\xi}^N))P^{\otimes N}(d\vec{\xi}^N)$ and $\int_{(\Xi)^{\otimes N}}g(\hat{\vartheta}(\vec{\xi}^N))Q^{\otimes N}(d\vec{\xi}^N)$ can be deduced from Theorem 2.1. The rest follows from Theorem 2.1 by setting $T_N(\xi^1,\cdots,\xi^N)=\hat{\vartheta}(\xi^1,\cdots,\xi^N)$ in the theorem. ■

5 Statistical robustness of the distributionally robust shortfall risk optimization model

We consider distributionally robust shortfall risk optimization model where the ambiguity set is constructed by Kantorovich ball centered at a nominal distribution P_N with some fixed radius r

$$\begin{array}{ll} \min\limits_{t\in\mathbb{R}} & t \\ \text{s.t.} & \sup\limits_{P'\in\mathcal{B}(P_N,r)} \mathbb{E}_{P'}[l(-\xi-t)] \le \lambda, \end{array} \tag{5.46}$$

where $\xi \in \mathbb{R}$, and

$$\mathcal{B}(P_N, r) := \{P' \in \mathscr{P}(\Xi) : \mathsf{dl}_{K,1}(P', P_N) \le r\}.$$

In order to investigate the statistical robustness of model (5.46), let $\vartheta(P_N)$ denote the optimal value of problem (5.46) and $\vartheta(Q_N)$ the one with P_N be replaced by Q_N, that is,

$$\begin{aligned} \vartheta(P_N) &:= \min_{t\in\mathbb{R}} \left\{ t : \sup_{P'\in\mathcal{B}(P_N,r)} \mathbb{E}_{P'}[l(-\xi-t)] \le \lambda \right\}, \\ \vartheta(Q_N) &:= \min_{t\in\mathbb{R}} \left\{ t : \sup_{P'\in\mathcal{B}(Q_N,r)} \mathbb{E}_{P'}[l(-\xi-t)] \le \lambda \right\}. \end{aligned}$$

Then $\vartheta(P_N)$ and $\vartheta(Q_N)$ are two statistical estimators of the optimal value of (5.46) and we are interested in the difference between laws of the two estimators, that is, the difference between $Q^{\otimes N} \circ \vartheta(Q_N)^{-1}$ and $P^{\otimes N} \circ \vartheta(P_N)^{-1}$. To this end, we need to make the following assumption.

Assumption 5.1 *Assume that*

(a) Ξ *is compact;*

(b) there exists a point $z_0 < 0$ *such that* l *is strictly increasing over* $[z_0, \infty)$*;*

(c) $\lambda \in \text{int range } l$*, where* $\text{range } l := \{l(t) : t \in \mathbb{R}\}$.

Under Assumption 5.1, inequality in (5.46) satisfies the Slater condition, i.e., there exists a point $t_P \in \mathbb{R}$ such that

$$\sup_{P'\in\mathcal{B}(P,r)} \mathbb{E}_{P'}[l(-\xi-t_P)] - \lambda < 0. \tag{5.47}$$

To see this, notice that since $\lambda \in \text{int range } l$, there exists a constant $\varepsilon > 0$ such that $\lambda + (-\varepsilon, \varepsilon) \subset \text{range } l$. We can then choose a positive constant $\varepsilon_1 \in (0, \varepsilon)$ such that $(\lambda - \varepsilon, \lambda - \varepsilon_1) \subset \text{range } l$. Consequently, we can find $\bar{t} \in \mathbb{R}$ such that $l(\bar{t}) \le \lambda - \varepsilon_1$. Let $t_P := -\operatorname{ess\,inf} \xi - \bar{t}$. Then

$$\sup_{P'\in\mathcal{B}(P,r)} \mathbb{E}_{P'}[l(-\xi-t_P)] - \lambda \le \sup_{\xi\in\Xi} l(-\xi-t_P) - \lambda \le l(-\operatorname{ess\,inf} \xi - t_P) - \lambda = l(\bar{t}) - \lambda \le -\varepsilon_1 < 0,$$

which shows (5.47).

Under Assumption 5.1, the optimal value $\vartheta(Q_N)$ of problem (5.46) has a lower bound uniformly for all Q_N. To see this, we note that since $l(\cdot)$ is increasing,

$$l\left(-\operatorname{ess\,sup}\xi - t\right) = \sup_{P' \in \mathcal{B}(Q_N, r)} \mathbb{E}_{P'}\left[l\left(-\operatorname{ess\,sup}\xi - t\right)\right] \leq \sup_{P' \in \mathcal{B}(Q_N, r)} \mathbb{E}_{P'}[l(-\xi - t)].$$

Let $t_1 := \min\left\{t \in \mathbb{R} : l\left(-\operatorname{ess\,sup}\xi - \mathrm{t}\right) \leq \lambda\right\}$. Then the inequality above implies that $\vartheta(Q_N) \geq t_1$ for all Q_N. We assert that $t_1 > -\infty$. Indeed, by Assumption 5.1 (c), there exists a constant $\alpha > 0$ such that

$$l\left(-\operatorname{ess\,sup}\xi - t_P\right) = \sup_{P' \in \mathcal{B}(P, r)} \mathbb{E}_{P'}\left[l\left(-\operatorname{ess\,sup}\xi - t_P\right)\right] \leq \sup_{P' \in \mathcal{B}(P, r)} \mathbb{E}_{P'}[l(-\xi - t_P)] \leq \lambda - \alpha.$$

Since $l(\cdot)$ is strictly increasing over $[z_0, \infty)$, we have

$$\lim_{t \to -\infty} l\left(-\operatorname{ess\,sup}\xi - t\right) = +\infty,$$

which implies $t_1 > -\infty$. This shows

$$\vartheta(Q_N) \geq t_1 > -\infty, \ \ \forall Q_N. \tag{5.48}$$

Proposition 5.1 *Let r be fixed. Assume: (a) Assumption 5.1 holds, (b) $l(\cdot)$ is a convex function over $\mathbb{R}$, (c) $l(\cdot)$ is Lipschitz continuous over the interval $I := [-\operatorname{ess\,sup}\xi - t_P, -\operatorname{ess\,inf}\xi - t_1]$, i.e., there exists a positive constant $L_3 > 0$ such that*

$$|l(z_1) - l(z_2)| \leq L_3 |z_1 - z_2|, \ \forall z_1, z_2 \in I. \tag{5.49}$$

Then the optimal value function of problem (5.46) is Lipschitz continuous, i.e., there exist constants $\delta_2 > 0$ and $\hat{C}_P$ such that

$$|\vartheta(\tilde{Q}_N) - \vartheta(\hat{Q}_N)| \leq \hat{C}_P \mathsf{dl}_{K,1}(\tilde{Q}_N, \hat{Q}_N) \tag{5.50}$$

for all $\tilde{Q}_N, \hat{Q}_N \in \mathcal{B}(P, \delta_2)$.

Proof. We first show that the optimal value $\vartheta(Q_N)$ also has a uniform upper bound for all Q_N near P. Note that it follows from [20, Theorem 1] that

$$\mathbb{H}\big(\mathcal{B}(\tilde{Q}_N, r), \mathcal{B}(\hat{Q}_N, r); \mathsf{dl}_{K,1}\big) \leq \mathsf{dl}_{K,1}(\tilde{Q}_N, \hat{Q}_N). \tag{5.51}$$

Let $\mathscr{H}_3 := \{l(-\xi - t) : t \in [t_1, t_P]\}$. Then by the Lipschitz continuity of loss function $l(\cdot)$ (see (5.49)), we have $\frac{\mathscr{H}_3}{L_3} \subset \mathscr{G}_L$. For any $t \in [t_1, t_P]$,

$$\begin{aligned}
& \sup_{\tilde{P} \in \mathcal{B}(\tilde{Q}_N, r)} \mathbb{E}_{\tilde{P}}[l(-\xi - t)] - \sup_{\hat{P} \in \mathcal{B}(\hat{Q}_N, r)} \mathbb{E}_{\hat{P}}[l(-\xi - t)] \\
= & \sup_{\tilde{P} \in \mathcal{B}(\tilde{Q}_N, r)} \inf_{\hat{P} \in \mathcal{B}(\hat{Q}_N, r)} \left(\mathbb{E}_{\tilde{P}}[l(-\xi - t)] - \mathbb{E}_{\hat{P}}[l(-\xi - t)]\right) \\
\leq & \sup_{\tilde{P} \in \mathcal{B}(\tilde{Q}_N, r)} \inf_{\hat{P} \in \mathcal{B}(\hat{Q}_N, r)} \sup_{l \in \mathscr{H}_3} \left|\mathbb{E}_{\tilde{P}}[l(-\xi - t)] - \mathbb{E}_{\hat{P}}[l(-\xi - t)]\right| \\
\leq & L_3 \sup_{\tilde{P} \in \mathcal{B}(\tilde{Q}_N, r)} \inf_{\hat{P} \in \mathcal{B}(\hat{Q}_N, r)} \sup_{h \in \mathscr{G}_L} \left|\mathbb{E}_{\tilde{P}}[h(\xi)] - \mathbb{E}_{\hat{P}}[h(\xi)]\right| \\
= & L_3 \mathbb{D}(\mathcal{B}(\tilde{Q}_N, r), \mathcal{B}(\hat{Q}_N, r); \mathsf{dl}_{K,1}),
\end{aligned} \tag{5.52}$$

where $\mathscr{G}_L$ is defined in (1.7). Swapping the positions of $\hat{Q}_N$ and $\tilde{Q}_N$, we have

$$\sup_{\hat{P}\in\mathcal{B}(\hat{Q}_N,r)} \mathbb{E}_{\hat{P}}[l(-\xi-t)] - \sup_{\tilde{P}\in\mathcal{B}(\tilde{Q}_N,r)} \mathbb{E}_{\tilde{P}}[l(-\xi-t)] \le L_3\mathbb{D}(\mathcal{B}(\hat{Q}_N,r),\mathcal{B}(\tilde{Q}_N,r);\mathsf{dl}_{K,1}). \quad (5.53)$$

Combining inequalities (5.52) and (5.53), we obtain for any $t\in[t_1,t_P]$,

$$\left|\sup_{\tilde{P}\in\mathcal{B}(\tilde{Q}_N,r)} \mathbb{E}_{\tilde{P}}[l(-\xi-t)] - \sup_{\hat{P}\in\mathcal{B}(\hat{Q}_N,r)} \mathbb{E}_{\hat{P}}[l(-\xi-t)]\right| \le L_3\mathbb{H}(\mathcal{B}(\tilde{Q}_N,r),\mathcal{B}(\hat{Q}_N,r);\mathsf{dl}_{K,1}). \quad (5.54)$$

A combination of (5.51) and (5.54) yields

$$\left|\sup_{\tilde{P}\in\mathcal{B}(\tilde{Q}_N,r)} \mathbb{E}_{\tilde{P}}[l(-\xi-t)] - \sup_{\hat{P}\in\mathcal{B}(\hat{Q}_N,r)} \mathbb{E}_{\hat{P}}[l(-\xi-t)]\right| \le L_3\mathsf{dl}_{K,1}(\tilde{Q}_N,\hat{Q}_N) \quad (5.55)$$

for all $t\in[t_1,t_P]$. It follows from (5.55) and the Salter condition that

$$\sup_{P'\in\mathcal{B}(Q_N,r)} \mathbb{E}_{P'}[l(-\xi-t_P)] \le \sup_{P'\in\mathcal{B}(P,r)} \mathbb{E}_{P'}[l(-\xi-t_P)] + L_3\mathsf{dl}_{K,k}(Q_N,P) \le \lambda-\alpha+L_3\mathsf{dl}_{K,k}(Q_N,P),$$

and thus

$$\sup_{P'\in\mathcal{B}(Q_N,r)} \mathbb{E}_{P'}[l(-\xi-t_P)] \le \lambda - \frac{\alpha}{2} < \lambda \quad (5.56)$$

for $Q_N\in\mathcal{B}(P,\delta_2)$ with $\delta_2=\frac{\alpha}{2L_3}$, which means t_P is a feasible solution of problem (5.46) with P_N being replaced by $Q_N\in\mathcal{B}(P,\delta_2)$. This shows

$$\vartheta(Q_N)\le t_P,\ \ \forall Q_N\in\mathcal{B}(P,\delta_2). \quad (5.57)$$

Combining (5.48) and (5.57), we have

$$\vartheta(Q_N)\in[t_1,t_P],\ \ \forall Q_N\in\mathcal{B}(P,\delta_2),$$

and thus problem (5.46) with parameter $Q_N\in\mathcal{B}(P,\delta_2)$ can be written equivalently as

$$\begin{aligned}\min_{t\in[t_1,t_P]}\ & t\\ \text{s.t.}\ & \sup_{P'\in\mathcal{B}(Q_N,r)} \mathbb{E}_{P'}[l(-\xi-t)]\le\lambda.\end{aligned} \quad (5.58)$$

Next, we investigate the Lipschitz continuity of the restricted feasible set

$$\mathcal{F}(Q_N)\cap T := \left\{t\in[t_1,t_P] : \sup_{P'\in\mathcal{B}(Q_N,r)} \mathbb{E}_{P'}[l(-\xi-t)]\le\lambda\right\}.$$

Note that the convexity of l ensures that $\sup_{\hat{P}\in\mathcal{B}(\hat{Q}_N,r)}\mathbb{E}_{\hat{P}}[l(-\xi-t)-\lambda]$ is convex in t. Since $t\in\mathcal{F}(\tilde{Q}_N)\cap T$ is equivalent to $\sup_{\tilde{P}\in\mathcal{B}(\tilde{Q}_N,r)}\mathbb{E}_{\tilde{P}}[l(-\xi-t)]-\lambda\le 0$, it follows by Robinson's error bound for convex system of inequality (see [27, Section 3]) that

$$\begin{aligned} d(t,\mathcal{F}(\hat{Q}_N)\cap T) &\le \kappa_P\max\left\{0,\ \sup_{\hat{P}\in\mathcal{B}(\hat{Q}_N,r)} \mathbb{E}_{\hat{P}}[l(-\xi-t)]-\lambda\right\}\\ &\le \kappa_P\max\left\{0,\ \sup_{\hat{P}\in\mathcal{B}(\hat{Q}_N,r)} \mathbb{E}_{\hat{P}}[l(-\xi-t)]-\lambda-\left(\sup_{\tilde{P}\in\mathcal{B}(\tilde{Q}_N,r)} \mathbb{E}_{\tilde{P}}[l(-\xi-t)]-\lambda\right)\right\}\end{aligned}$$

for all $t \in \mathcal{F}(\tilde{Q}_N) \cap T$, where $\kappa_P = 2(t_P - t_1)/\alpha$, $d(t', t'') := |t' - t''|$, $d(t', A) := \inf_{t'' \in A} |t' - t''|$, and thus

$$\mathbb{D}(\mathcal{F}(\tilde{Q}_N) \cap T, \mathcal{F}(\hat{Q}_N) \cap T; d) \leq \kappa_P \sup_{t \in T} \left| \sup_{\tilde{P} \in \mathcal{B}(\tilde{Q}_N, r)} \mathbb{E}_{\tilde{P}}[l(-\xi - t)] - \sup_{\hat{P} \in \mathcal{B}(\hat{Q}_N, r)} \mathbb{E}_{\hat{P}}[l(-\xi - t)] \right|$$

for all $\tilde{Q}_N, \hat{Q}_N \in \mathcal{B}(P, \delta_2)$. Similarly, we have

$$\mathbb{D}(\mathcal{F}(\hat{Q}_N) \cap T, \mathcal{F}(\tilde{Q}_N) \cap T; d) \leq \kappa_P \sup_{t \in T} \left| \sup_{\tilde{P} \in \mathcal{B}(\tilde{Q}_N, r)} \mathbb{E}_{\tilde{P}}[l(-\xi - t)] - \sup_{\hat{P} \in \mathcal{B}(\hat{Q}_N, r)} \mathbb{E}_{\hat{P}}[l(-\xi - t)] \right|$$

for all $\tilde{Q}_N, \hat{Q}_N \in \mathcal{B}(P, \delta_2)$. Combining the above two inequalities and (5.55), we have

$$\mathbb{H}(\mathcal{F}(\tilde{Q}_N) \cap T, \mathcal{F}(\hat{Q}_N) \cap T; d) \leq \hat{C}_P \mathsf{dl}_{K,1}(\tilde{Q}_N, \hat{Q}_N) \tag{5.59}$$

for all $\tilde{Q}_N, \hat{Q}_N \in \mathcal{B}(P, \delta_2)$, where $\hat{C}_P = \kappa_P L_3$. It follows by [16, Theorem 1] that the optimal value function of problem (5.46) is Lipschitz continuous, i.e.,

$$|\vartheta(\tilde{Q}_N) - \vartheta(\hat{Q}_N)| \leq \hat{C}_P \mathsf{dl}_{K,1}(\tilde{Q}_N, \hat{Q}_N)$$

for all $\tilde{Q}_N, \hat{Q}_N \in \mathcal{B}(P, \delta_2)$. ■

The Slater condition (5.47) plays a crucial role in the Lipschitz continuity of the restricted feasible set-valued mapping $\mathcal{F}(\cdot) \cap T$ in (5.59). Unfortunately the latter holds only in a neighborhood of the true probability measure P which is inadequate for us to establish "global" Lipschitz continuity of $\vartheta(\cdot)$. To address the issue, we impose a stronger Slater condition for the inequality in (5.46). Specifically, we require that for every $Q \in \mathscr{P}(\Xi)$, there exists a point $t_Q \in \mathbb{R}$ such that

$$\sup_{P' \in \mathcal{B}(Q, r)} \mathbb{E}_{P'}[l(-\xi - t_Q)] - \lambda < 0. \tag{5.60}$$

It is easy to observe that this kind of Slater condition is guaranteed by Assumption 5.1.

We write $\vec{\xi}^N$ for $(\xi^1, \cdots, \xi^N)$ and write $\hat{\vartheta}(\vec{\tilde{\xi}}^N)$ for $\vartheta(\tilde{Q}_N)$ and $\hat{\vartheta}(\vec{\hat{\xi}}^N)$ for $\vartheta(\hat{Q}_N)$ to indicate their dependence on $\vec{\tilde{\xi}}$ and $\vec{\hat{\xi}}$ respectively.

Theorem 5.1 (Quantitative statistical robustness of model (5.46)) *Assume: (a) Assumption 5.1 holds, (b) $l(\cdot)$ is a convex function on $\mathbb{R}$, (c) $l(\cdot)$ is Lipschitz continuous over a compact set, i.e., there exists a positive constant $L_3 > 0$ such that (5.49) holds. Then there exists a positive constant $\hat{C} > 0$ such that for $P, Q \in \mathcal{M}_k^\phi$ with $\phi(t) := |t|$, $t \in \mathbb{R}$,*

$$\mathsf{dl}_{K,1}\left(P^{\otimes N} \circ \hat{\vartheta}^{-1}, Q^{\otimes N} \circ \hat{\vartheta}^{-1}\right) \leq \hat{C} \mathsf{dl}_{K,k}(P, Q). \tag{5.61}$$

for all $N \in \mathbb{N}$.

Proof. Under Assumptions 5.1 and conditions (b) and (c), for any $\tilde{Q}_N, \hat{Q}_N \in \mathscr{P}(\Xi)$, since $\mathscr{P}(\Xi)$ is a weakly compact set under metric $d_{K,1}$, we can construct a δ_2-net $\{Q^1, \cdots, Q^{\tilde{N}}\}$ in $\mathscr{P}(\Xi)$ such that

$$\left\{ Q_N(\lambda) := (1 - \lambda)\tilde{Q}_N + \lambda \hat{Q}_N : \lambda \in [0, 1] \right\} \subset \bigcup_{j=1}^{\tilde{J}} \mathcal{B}(Q^j, \delta_2^j)$$

with $\tilde{J} \leq \tilde{N}$ and $\mathscr{P}(\Xi) \subset \cup_{j=1}^{\tilde{N}} \mathcal{B}(Q^j, \delta_2^j)$. Similar to the proof of Lemma 3.1 (ii), we can select an increasing sequence $\{\lambda\}_{j=1}^{\tilde{J}+1} \subset [0,1]$, with $\lambda_1 = 0$, $\lambda_{\tilde{J}+1} = 1$, $Q_{1_j} := (1-\lambda_j)\tilde{Q}_N + \lambda_j \hat{Q}_N$ and $Q_{2_j} := (1-\lambda_{j+1})\tilde{Q}_N + \lambda_{j+1}\hat{Q}_N$, $j = 1, \cdots, \tilde{J}$ such that

$$Q_{2_j} = Q_{1_{j+1}}, \text{ for } j \in [\tilde{J}-1], \quad Q_{1_j}, Q_{2_j} \in \mathcal{B}(Q^j, \delta_2^j) \text{ for } j \in [\tilde{J}],$$

where we write $[\tilde{J}]$ for $\{1, \cdots, \tilde{J}\}$. It follows by Proposition 5.1 that there exist positive constants $\hat{C}_j$, $j \in [\tilde{J}]$ such that

$$|\vartheta(Q_{1_j}) - \vartheta(Q_{2_j})| \leq \hat{C}_j \mathsf{dl}_{K,1}(Q_{1_j}, Q_{2_j}), \text{ for} j \in [\tilde{J}].$$

Let $\hat{C} := \max_{j \in [\tilde{N}]} \{\hat{C}_j\}$. Consequently, we have

$$\begin{aligned}
|\vartheta(\tilde{Q}_N) - \vartheta(\hat{Q}_N)| &\leq \sum_{j=1}^{\tilde{J}} |\vartheta(Q_{1_j}) - \vartheta(Q_{2_j})| \leq \sum_{j=1}^{\tilde{J}} \hat{C}_j \mathsf{dl}_{K,1}(Q_{1_j}, Q_{2_j}) \\
&= \sum_{j=1}^{\tilde{J}} \hat{C}_j \mathsf{dl}_{K,1}((1-\lambda_j)\tilde{Q}_N + \lambda_j \hat{Q}_N, (1-\lambda_{j+1})\tilde{Q}_N + \lambda_{j+1}\hat{Q}_N) \\
&= \sum_{j=1}^{\tilde{J}} \hat{C}_j (\lambda_{j+1} - \lambda_j) \mathsf{dl}_{K,1}(\tilde{Q}_N, \hat{Q}_N) \\
&\leq \tilde{N} \max_{j \in [\tilde{J}]} \hat{C}_j \mathsf{dl}_{K,1}(\tilde{Q}_N, \hat{Q}_N) =: \hat{C} \mathsf{dl}_{K,1}(\tilde{Q}_N, \hat{Q}_N) \qquad (5.62)
\end{aligned}$$

for all $\tilde{Q}_N, \hat{Q}_N \in \mathscr{P}(\Xi)$. We show (5.61) by applying Theorem 2.1, where the Lipschitz continuity of $\hat{\vartheta}(\tilde{\xi}^1, \cdots, \tilde{\xi}^N)$ w.r.t. $(\tilde{\xi}^1, \cdots, \tilde{\xi}^N)$ is needed. Note that

$$\begin{aligned}
|\hat{\vartheta}(\tilde{\xi}^1, \cdots, \tilde{\xi}^N) - \tilde{\vartheta}(\hat{\xi}^1, \cdots, \hat{\xi}^N)| &= \left|\vartheta(\tilde{Q}_N) - \vartheta(\hat{Q}_N)\right| \\
&\leq \hat{C} \mathsf{dl}_{K,1}(\tilde{Q}_N, \hat{Q}_N) \\
&= \hat{C} \sup_{g \in \mathscr{G}_L} |\mathbb{E}_{\tilde{Q}_N}[g(\xi)] - \mathbb{E}_{\hat{Q}_N}[g(\xi)]| \\
&= \hat{C} \sup_{g \in \mathscr{G}_L} \left| \frac{1}{N} \sum_{i=1}^N g(\tilde{\xi}^i) - \frac{1}{N} \sum_{i=1}^N g(\hat{\xi}^i) \right| \\
&\leq \frac{\hat{C}}{N} \sum_{i=1}^N \|\tilde{\xi} - \hat{\xi}\|,
\end{aligned}$$

where the first inequality is due to (5.62).

The above relations mean that $\hat{\vartheta}(\cdot)$ is Lipschitz continuous over $(\Xi)^{\otimes N}$ with Lipschitz modulus bounded by $\hat{C}/N$, and thus the well-definedness of $\int_{(\Xi)^{\otimes N}} g(\hat{\vartheta}(\vec{\xi}^N)) P^{\otimes N}(d\vec{\xi}^N)$ and $\int_{(\Xi)^{\otimes N}} g(\hat{\vartheta}(\vec{\xi}^N)) Q^{\otimes N}(d\vec{\xi}^N)$ can be deduced from Theorem 2.1. The rest follows from Theorem 2.1 by setting $T_N(\xi^1, \cdots, \xi^N) = \hat{\vartheta}(\xi^1, \cdots, \xi^N)$ in the theorem. ■

6 Concluding remarks

In this paper, we study quantitative statistical robustness in distribution robust optimization models and demonstrate under some moderate conditions that it is safe to use the DRO models as long as the topological structure of perceived data does not deviate significantly from that of the real data. Our theoretical results are presented for the optimal values of the DRO models, it might be interested to extend them to the optimal solutions. Moreover, the sample data are assumed to be iid, it might be interesting to explore the case when the same data are not iid. Finally, it might be interesting to carry out some numerical tests to verify the established theoretical results, we leave all these for interested readers to explore.

References

[1] R. Cont, R. Deguest and G. Scandolo, Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, 10: 593-606, 2010.

[2] E. Delage and Y. Ye, Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58: 592-612, 2010.

[3] R. M. Dudley, *Real Analysis and Probability*, Wadsworth & Brooks/Cole, Belmont, CA, 2004.

[4] D. A. Edwards, On the kantorovich–rubinstein theorem. *Expositiones Mathematicae*, 29: 387-398, 2011.

[5] P. M. Esfahani and D. Kuhn, Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171: 115-166, 2018.

[6] S. Guo and H. Xu, Distributionally robust shortfall risk optimization model and its approximation. *Mathematical Programming*, 174: 473-498, 2019.

[7] S. Guo and H. Xu, Statistical robustness in utility preference robust optimization models. *Mathematical Programming Series A*, 2020, https://doi.org/10.1007/s10107-020-01555-5.

[8] S. Guo and H. Xu, Data perturbations in stochastic generalized equations: statistical robustness in static and sample average approximated models, preprint, 2021.

[9] S. Guo, H. Xu and L. Zhang, Statistical robustness of empirical risks in machine learning. arXiv preprint, arXiv:2005.08458v, 2020.

[10] P. J. Huber and E. M. Ronchetti, *Robust Statistics*. 2nd Edition, John Wiley & Sons, New Jersey, 2009.

[11] A. L. Gibbs and F. E. Su, On choosing and bounding probability metrics. *International Statistical Review*, 70: 419-435, 2002.

[12] F. R. Hampel, *Contribution to The Theory of Robust Estimation.* Ph. D. Thesis, University of California, Berkeley, 1968.

[13] F. R. Hampel, A general statistical definition of robustness. *The Annals of Mathematical Statistics*, 42: 1887-1896, 1971.

[14] Z. Hu and J. L. Hong, Kullback-leibler divergence constrained distributionally robust optimization. Preprint Optimization Online, 2012.

[15] P. J. Huber, *Robust Statistics.* 3rd Edition, John Wiley & Sons, New York, 1981.

[16] D. Klatte, A note on quantitative stability results in nonlinear optimization. *Seminarbericht Nr. 90*, Sektion Mathematik, Humboldt-Universität zu Berlin, Berlin, pp. 77-86, 1987.

[17] V. Krätschmer, A. Schied and H. Zähle, Comparative and qualitative robustness for law-invariant risk measures. *Finance and Stochastics*, 18: 271-295, 2014.

[18] G. Lecué and M. Lerasle, Robust machine learning by median-of-means: theory and practice. arXiv preprint arXiv:1711.10306, 2017.

[19] Y. Liu, A. Pichler and H. Xu, Discrete approximation and quantification in distributionally robust optimization. *Mathematics of Operations Research*, 44: 19-37, 2019.

[20] A. Pichler and H. Xu, Quantitative stability analysis for minimax distributionally robust risk optimization. *Mathematical Programming Series B*, 2019.

[21] H. Rahimian and S. Mehrotra, Distributionally robust optimization: A review. arXiv:1908.05659 [math.OC], 2019.

[22] S. Shafieezadeh-Abadeh, D. Kuhn and P. M. Esfahani, Regularization via mass transportation. *Journal of Machine Learning Research*, 20: 1-68, 2019.

[23] A. Shapiro, Distributionally Robust Stochastic Programming, *SIAM J. Optimization*, 27: 2258–2275, 2017.

[24] A. M. C. So, Moment inequalities for sums of random matrices and their applications in optimization. *Mathematical Programming*, 130: 125-151, 2011.

[25] H. Sun, H. Xu, Convergence analysis for distributionally robust optimization and equilibrium problems. *Mathematics of Operations Research*, 41: 377-401, 2016.

[26] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models.* Wiley, West Sussex, 1991.

[27] S. M. Robinson, An application of error bounds for convex programming in a linear space. *SIAM Journal on Control*, 13: 271-273, 1975.

[28] I. Steinwart and A. Christmann, *Support Vector Machines.* Springer Science & Business Media, 2008.

[29] A. Szulga, On minimal metrics in the space of random variables. *Theory of Probability & Its Applications*, 27: 424-430, 1983.

[30] R. Tibshirani, Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58: 267-288, 1996.

[31] J. W. Tukey, A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, 2: 448-485, 1960.

[32] J. W. Tukey, The future of data analysis. *The Annals of Mathematical Statistics*, 33: 1-67, 1962.

[33] W. Wang , H. Xu and T. Ma, Quantitative statistical robustness for tail-dependent law invariant risk measures. *Quantitative Finance*, March 2021.

[34] H. Xu, Y. Liu and H. Sun, Distributionally robust optimization with matrix moment constraints: lagrange duality and cutting plane methods. *Mathematical Programming Series A*, 169: 489-529, 2018.

[35] J. Zhang, H. Xu and L.W. Zhang, Quantitative stability analysis of stochastic quasi-variational inequality problems and applications. *Mathematical Programming Series B*, 165: 433-470, 2017.

[36] J. Zhang, H. Xu and L. Zhang, Quantitative stability analysis for distributionally robust optimization with moment constraints. *SIAM Journal on Optimization*, 26: 1855-1882, 2016.

[37] J. Zhen, D. Kuhn and W. Wiesemann, Mathematical foundations of robust and distributionally robust optimization, arXiv:2105.00760 [math.OC], 2021.

[38] V. M. Zolotarev, Probability metrics. *Teoriya Veroyatnostei i ee Primeneniya*, 28: 264-287, 1983.