Statistical Robustness in Utility Preference Robust Optimization Models

Shaoyan Guo* and Huifu Xu†

July 16, 2020

Abstract

Utility preference robust optimization (PRO) concerns decision making problems where information on decision maker's utility preference is incomplete and has to be elicited through partial information and the optimal decision is based on the worst case utility function elicited. A key assumption in the PRO models is that the true probability distribution is either known or can be recovered by real data generated by the true distribution. In data-driven optimization, this assumption may not be satisfied when perceived data differ from real data and consequently it raises a question as to whether statistical estimators of the PRO models based on perceived data are reliable. In this paper, we investigate the issue which is also known as qualitative robustness in the literature of statistics [26] and risk management [30]. By utilizing the framework proposed by Krätschmer et al. [30], we derive moderate sufficient conditions under which the optimal value and optimal solution of the PRO models are robust against perturbation of the exogenous uncertainty data, and examine how the tail behaviour of utility functions affects the robustness. Moreover, under some additional conditions on the Lipschitz continuity of the underlying functions with respect to random data, we establish quantitative robustness of the statistical estimators under the Kantorovich metric. Finally, we investigate uniform consistency of the optimal value and optimal solution of the PRO models. The results cover utility selection problems and stochastic optimization problems as special cases.

Keywords. PRO, qualitative statistical robustness, quantitative statistical robustness, uniform consistency

1 Introduction

We consider the following one-stage expected utility maximization problem

$$\max_{x\in X}\ \mathbb{E}_P[u(f(x,\xi))], \tag{1.1}$$

*School of Mathematical Sciences, Dalian University of Technology, Dalian, 116024, China. Email: syguo@dlut.edu.cn.

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

where $u : \mathbb{R} \to \mathbb{R}$ is a real-valued increasing utility function 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 feasible 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.

In practice, f may represent a financial position or the performance of an engineering design. In the well-established theory of expected utility which provides a dominant normative and descriptive model in decision making, Von Neumann and Morgenstern [45] show that any set of preferences that a decision maker may have among uncertain/risky prospects can be characterized by an expected utility function if the preferences satisfy certain reasonable axioms (i.e. completeness, transitivity, continuity and independence). Specifically, there exists a utility function $u : \mathbb{R} \to \mathbb{R}$ such that the decision maker prefers prospect A to prospect B if and only if $\mathbb{E}[u(A)] \geq \mathbb{E}[u(B)]$. A convex utility function means the decision maker is risk taking whereas a concave utility function means risk averse and an affine utility function represents risk neutral.

In the theory of expected utility, the utility function u represents the decision maker's risk attitudes or taste, P is regarded as the belief on the states of the world that the decision maker holds. The latter corresponds to the terminology of subjective judgement of probability distribution in the literature of stochastic programming. In the absence of complete information, taste and belief are often subjective in which case they may interact. The classical expected utility model of Von Neumann and Morgenstern assumes that there is no ambiguity/uncertainty in both the brief and the taste. Gilboa and Schmeidler [16] consider a situation where the decision maker's brief is uncertain and consequently propose a distributionally robust expected utility model where the preference is characterized through the most conservative belief (the worst case probability), see Maccheroni et al. [34], Marinacci [35] and Gilboa and Marinacci [17] for more recent development in this regard.

There is another important stream of research which focuses on the ambiguity of decision maker's taste, that is, the true utility function in model (1.1). Such ambiguity may arise from a lack of accurate description of human behaviour (Thurstone [43]), conginitive difficulty or incomplete information (Karmarkar [29] and Weber [46]). It could also be the case where the decision making problem involves several stakeholders who fail to reach a consensus. Parametric and non-parametric approaches have subsequently been proposed to assess the true utility function, these include discrete choice models (Train [44]), standard and paired gambling approaches for preference comparisons and certainty equivalence (Farquhar [13]), we refer readers to Hu et al. [23] for an excellent overview on this.

The focus of utility preference robust optimization (PRO in short) is on the situation where a decision maker is confronted with ambiguity in taste instead of brief, but it is possible to use partial information to elicit/construct an ambiguity set of utility functions denoted by $\mathcal{U}$ such that the true utility function which reflects precisely the decision maker's preference lies in $\mathcal{U}$ with high likelihood. Instead of attempting to obtain an estimate of the true utility function from the set, the PRO approach considers a maximin formulation

$$\text{(PRO)} \qquad \vartheta(P) := \max_{x \in X} \inf_{u \in \mathcal{U}} \; \mathbb{E}_P[u(f(x,\xi))], \tag{1.2}$$

where the optimal decision x is based on the worst u from the ambiguity set. For each fixed

$x \in X$, we highlight the inner minimization problem

$$v(x, P) := \inf_{u \in \mathcal{U}} \mathbb{E}_P[u(f(x, \xi))], \tag{1.3}$$

which is an optimal (worst case in this setup) utility selection problem in economics. The robust model aims to mitigate the risk arising from ambiguity of the true utility function albeit it could be over conservative if the ambiguity set is large.

Two main challenges have to be tackled before the model can be effectively used in decision making under uncertainty. One is construction of the ambiguity set $\mathcal{U}$ and the other is computational solvability of the model as the inner minimization problem is often infinite dimensional. Greco et al. [18] propose a framework for eliciting utility functions representing the decision maker's preferences via pairwise comparison in multiple criteria ranking even when some of the preferences are inconsistent. The elicited set of utility functions is then used to define preference relationship in ranking which is essentially about solving a utility maximization problem similar to (1.3). Hu and Stepanyan [25] propose a so-called reference-based almost stochastic dominance method for constructing a set of utility functions near a reference utility which satisfy certain stochastic dominance relationship and use the set to characterize decision maker's preference. Armbruster and Delage [2] address the challenges by adopting a number of strategies including pairwise comparison of lotteries and certainty equivalent for constructing the ambiguity set and subsequently deriving tractable formulations for the PRO models by reformulating them as linear programming problems. In their PRO models, they require utility functions in the ambiguity set to satisfy all of the decision maker's preferences elicited. Delage and Li [11] extend the research to risk management problem where the objective is a convex risk measure and the ambiguity set arises from decision maker's risk attitude. In the case when the range of $f(x, \xi)$ is bounded, Hu and Mehortra [24] consider a PRO model where the true unknown utility function is defined over a compact interval of $\mathbb{R}$ and takes values over $[0, 1]$ after some appropriate scaling. By viewing the utility function as a cumulative distribution function (cdf) of some random variable, they introduce moment-type conditions to define the ambiguity set $\mathcal{U}$ which cover a wide range of approaches such as pairwise comparison, certainty equivalent and stochastic dominance. A key condition for tractable formulation of the PRO model is that the utility function is concave. In the absence of concavity, one may reformulate the PRO model as a mixed integer programming problem, see [19, 22] for recent development in this regard.

In all these PRO models, the true probability distribution is assumed to be either known or can be recovered via real data generated by the unknown true probability distribution. In practice, this assumption may not be fulfilled either because the true probability distribution is unknown and the decision maker has to make a subjective judgement on brief or the perceived sample data are contaminated in a data-driven environment. In the absence of complete information on the taste and the belief, one may construct a robust model which bases optimal decision on the worst case utility and probability, see Haskell et al. [21]. Our research here is not about how to develop a robust mechanism to contain the risk arising from contaminated data in the PRO model, rather we study statistical impact of the potentially contaminated perceived data on the performance of the PRO model, specifically, we investigate whether a statistical estimator such as the optimal value of the PRO model based on perceived data with noise is reliable. The issue is essentially about statistical robustness in the literature of robust statistics

[26] and we believe the research is timely given that PRO models are mostly data-driven. The concept of statistical robustness may be traced back to Hampel's seminal work [20]. It has been popularized over the past few decades particularly with monographs [26, 27].

Cont et al. [8] use Hampel's classical concept to introduce a notion of qualitative robustness[1] of certain risk estimator and use the latter to examine the robustness of various risk estimators derived from empirical data. Using the definition, they demonstrate that statistical estimator of any spectral risk measure including conditional value at risk is not robust. Krätschmer et al. [30] argue that this definition of statistical robustness is in favour of value at risk and against coherent risk measures (including conditional value at risk), and consequently introduce a new concept of statistical robustness which captures tail losses to some degree in a refined topological space.

In this paper, we extend this stream of research on statistics and risk measurement to PRO models because these optimization problems are typically data driven and there is no guarantee that available perceived data do not contain noise. Our primary concern is that statistical estimators such as optimal values and optimal solutions obtained from solving the PRO models with perceived data may perform differently from those based on real data. This kind of analysis should be distinguished from the distributionally robust approach (Gilboa and Schmeidler [16]) in that we are not going to use the perceived data to construct an ambiguity set of the decision maker's belief in exogenous uncertainty. It also differs from standard stability analysis in stochastic programming from purposes to methodology and conclusions, we will come back to these later on in the relevant contexts. To the best of our knowledge, there is no statistical robustness analysis on optimization problems even in the literature of stochastic programming.

We plan to fill out this important gap by presenting a comprehensive research on statistical robustness of PRO models. Specifically, by utilizing the framework of statistical robustness in [30, 32], we derive moderate sufficient conditions under which the optimal value and optimal solution of PRO models are robust against perturbation of the exogenous uncertainty data and examine how the tail behaviour of utility functions affects such robustness. Note that Claus, Krätchmer and Schultz [7] derive similar results for a general class of risk optimization problems and apply them to two-stage mean-risk models. Whilst our analysis follows broadly the framework of analysis in [7, 30, 32], it requires more sophisticated additional mathematical treatment because PRO models involve maximin operations with respect to (w.r.t.) x and u and the utility functions are not necessarily concave. In order for us to concentrate on the fundamental issues concerning statistical robustness and broaden the coverage of our theoretical results, we undertake the analysis with abstract form of the ambiguity set $\mathcal{U}$ of increasing and continuous utility functions and refer readers to [2, 19, 24, 25] for various specifically structured ambiguity sets.

The main contributions of the paper can be summarized as follows.

- Under some growth condition of f and equi-continuity condition of utility functions in the

[1] Throughout this paper, we use terminology statistical robustness to avoid confusion with other notions of robustness.

ambiguity set $\mathcal{U}$, we show continuity of the optimal value function $\vartheta(\cdot)$ near P (Theorem 3.1) when P is perturbed under some specified topology associated with the growth order of f and the utility functions in $\mathcal{U}$. This result essentially coincides with the weak continuity of risk functionals established by Claus et al. in [7, Theorem 2.2 and Corollary 2.4] when the utility functions are increasing and concave, and controlled by a gauge function with specific polynomial growth. The novelty of this result is that it covers general utility functions which are not necessarily concave such as S-shaped utility and identifies interactions between the tail behaviours of P and $u \in \mathcal{U}$ through the topology under which the continuity is established.

- By exploiting Hampel's theorem [32, Theorem 2.4], we establish qualitative statistical robustness of estimators of the optimal value and optimal solutions when perceived data are generated by some distributions close to the true distribution P in a specified topological space (Theorem 4.1). The result covers two important specific cases: a utility selection problem (Corollary 4.1) and a stochastic programming problem (Corollary 4.2) albeit the latter can be derived straightforwardly from [7, Corollary 2.4] and Hampel's theorem. One of the interesting findings is that concavity (convexity) of the utility functions at the right (left) tail may make the PRO models more likely to be statistically robust (Example 4.1). Examples are given to illustrate how the analytical results can be applied to the existing PRO models, machine learning and utility-based shortfall risk optimization. Moreover, during the revision of the paper, we take one step further by deriving quantitative statistical robustness which allows one to explicitly derive an error bound for the discrepancy of two statistical estimators under the Kantorovich metric (Theorem 4.2) and apply the new result to utility selection problem (Corollary 4.1) and stochastic programming problem (Corollary 4.2) as special cases.

- We identify some verifiable sufficient conditions under which the statistical estimators of the optimal value and optimal solution are uniformly consistent (Theorem 6.1), which provides theoretical grounding for discrete approximation in the PRO models where the true distribution is continuously distributed. Note that discretization is vital for developing tractable numerical schemes for solving the PRO models [2, 19, 23, 24, 25].

The rest of the paper is organized as follows. Section 2 sets up the models for analysis. Section 3 discusses continuity of the optimal value function which paves the way for the analysis of qualitative and quantitative statistical robustness in Section 4. Section 5 gives some examples to illustrate the theoretical results, and Section 6 takes a step further to investigate uniform consistency of some statistical quantities in the PRO models.

2 Problem setup

In the (PRO) model [2], the true probability distribution P is either known or can be recovered from empirical data. In the latter case, the perceived empirical data may contain noise and

[2] Here and later on (PRO) model refers specifically to the maximin problem defined in (1.2) and this should be distinguished from acronym PRO for general preference robust optimization models/problems.

hence differ from real data generated by P. The discrepancy will inevitably affect the quality of statistical estimators such as the optimal value and the optimal solution obtained from solving the (PRO) model with perceived data. It is therefore vital to investigate certain insensitivity of the statistical estimators of the model to the deviation of empirical distributions from the true.

For simplicity of exposition, we assume the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is atomless. The assumption is common in the literature of finance and statistical robustness, see [15, Chapter 4] and [30]. Throughout this paper, we will use $\mathscr{P}(\mathbb{R}^k)$ to denote the set of all probability measures on $\mathbb{R}^k$.

We start by representing $v(x,P)$ in (1.3) as a composition of a real-valued function ϱ defined over $\mathscr{P}(\mathbb{R})$ and a probability measure $P \circ f(x,\cdot)^{-1} \in \mathscr{P}(\mathbb{R})$ induced by $f(x,\cdot)$, see the next proposition.

Proposition 2.1 *Let $v(x,P)$ be defined as in (1.3). Then there exists a unique function $\varrho : \mathscr{P}(\mathbb{R}) \to \mathbb{R}$ such that*

$$\varrho(\sigma) = \inf_{u\in\mathcal{U}} \int_{\mathbb{R}} u(t)\sigma(dt), \forall \sigma \in \mathscr{P}(\mathbb{R}), \tag{2.4}$$

and

$$v(x,P) = \varrho(P \circ f(x,\cdot)^{-1}), \forall x \in X \text{ and } P \in \mathscr{P}(\mathbb{R}^k). \tag{2.5}$$

Proof. The result is well known in the case when $v(x,\cdot)$ is a law invariant risk measure, see for instance discussions in [6, Section 2.1], [30, page 274] and [10, Proposition 1]. We include a similar proof for completeness as v has a different structure here.

Observe that $P = \mathbb{P} \circ \xi^{-1}$ is a probability measure on $\mathbb{R}^k$ induced by random vector ξ and $P \circ f(x,\cdot)^{-1}$ is a probability measure on $\mathbb{R}$ induced by $f(x,\xi)$. Hence

$$\mathbb{E}_P[u(f(x,\xi))] = \int_{\mathbb{R}} u(t) P \circ f(x,\cdot)^{-1}(dt). \tag{2.6}$$

Let ϱ be defined as in (2.4). Since ϱ is law invariant, this means that such a ϱ is unique. Moreover, it holds that

$$v(x,P) = \inf_{u\in\mathcal{U}} \mathbb{E}_P[u(f(x,\xi))] = \inf_{u\in\mathcal{U}} \int_{\mathbb{R}} u(t) P \circ f(x,\cdot)^{-1}(dt) = \varrho(P \circ f(x,\cdot)^{-1}).$$

The proof is complete. ■

Let $\mu(x,P) := P \circ f(x,\cdot)^{-1}$. Using Proposition 2.1, we can reformulate $v(x,P)$ as $\varrho(\mu(x,P))$ and consequently recast problem (1.2) as

$$\sup_{x\in X} \varrho(\mu(x,P)). \tag{2.7}$$

If we interpret (1.2) as a robust expected utility optimization model defined over prospect space $\{f(x,\xi) : x \in X\}$ (also known as space of acts), then (2.7) is a robust model defined over the space of probability measures $\{\mu(x,P) : x \in X\}$ (also known as space of lotteries). In the forthcoming discussions, we will use the latter formulation for analysis and let

$$S(P) := \arg\sup_{x\in X} \varrho(\mu(x,P)) \tag{2.8}$$

denote the set of optimal solutions. For any $P,Q \in \mathscr{P}(\mathbb{R}^k)$, let $\xi^1,\cdots,\xi^N$ and $\tilde{\xi}^1,\cdots,\tilde{\xi}^N$ be independent and identically distributed (iid) samples generated by P and Q respectively. Let

$$P_N := \frac{1}{N}\sum_{i=1}^N \delta_{\xi^i} \quad \text{and} \quad Q_N := \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{\xi}^i} \tag{2.9}$$

be the respective empirical distributions, where $\delta_{\tilde{\xi}}$ denotes the Dirac measure at $\tilde{\xi}$. We refer to P as the true probability distribution and Q its perturbation. In practice, samples are often obtained from perceived empirical data which contain some noise. This means the samples are not generated by P, rather they are generated by some Q (a perturbation of P). So in this setup, the samples of P are not obtainable, we define P_N only for theoretical analysis.

With perceived empirical data, we consider the PRO model (1.2)

$$(\text{PRO-}Q_N) \qquad \vartheta(Q_N) := \max_{x\in X}\inf_{u\in\mathcal{U}} \mathbb{E}_{Q_N}[u(f(x,\xi))] \tag{2.10}$$

and examine the discrepancy between $\vartheta(Q_N)$ and $\vartheta(P_N)$ under some appropriate metric. Analogous to (2.7), we can also reformulate (2.10) as

$$\sup_{x\in X} \varrho(\mu(x,Q_N)) \tag{2.11}$$

and let

$$S(Q_N) := \arg\sup_{x\in X} \varrho(\mu(x,Q_N)). \tag{2.12}$$

denote the corresponding set of optimal solutions. Note that we may write $\hat{\vartheta}_N(\tilde{\xi}^1,\cdots,\tilde{\xi}^N)$ and $\hat{S}_N(\tilde{\xi}^1,\cdots,\tilde{\xi}^N)$ for $\vartheta(Q_N)$ and $S(Q_N)$ respectively to indicate their dependence on the samples and the product probability measures for convergence analysis in the forthcoming discussions.

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 $\hat{T}_N(\xi^1,\cdots,\xi^N)$ for $T(P_N)$. Notice that $\hat{T}_N$ 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).

Remark 2.1 *At this point, we mention that statistical robustness should be distinguished from traditional stability analysis in stochastic programming where the later focuses on either the convergence (asymptotic consistency) of $T(P_N)$ to $T(P)$ or continuity of $T(\cdot)$ near P. Statistical robustness focuses on the variations between $T(Q_N)$ and $T(P_N)$ in terms of their distributions uniformly for all Q near P, it requires not only continuity $T(\cdot)$ near P but also uniform convergence of $T(Q_N)$ to $T(Q)$ for all Q near P. Figure 1 explains roughly their relationships. By Hampel's theorem [32, Theorem 2.4], the continuity of $T(\cdot)$ near P and the uniform Glivenko-Cantelli property of the probability space imply statistical robustness. The continuity has been well investigated by Claus et al. (see [7, Theorem 2.2]) when $T(\cdot)$ is a class of convex risk functionals. Because of the difference in the focuses, statistical robustness requires a different topological structure for analysis, we will come back to this in Section 4. Note that such analysis also differs from distributionally robust optimization (DRO) approach as we don't aim to derive a robust mechanism to contain the risk arising from potential contamination of perceived data.*

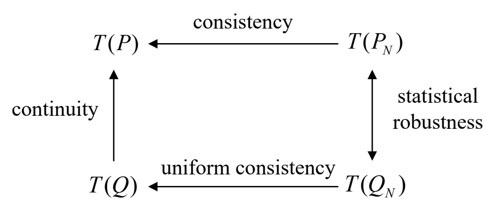

Figure 1: Relationship between stability analysis and statistical robustness

3 Continuity of the optimal value function $\vartheta(P)$

A key step to establish statistical robustness of the statistical estimator of the optimal value and optimal solutions is to show their continuity w.r.t. a small perturbation of the true probability P. This section aims to address this.

3.1 ϕ-weak topology

We start by recalling some basic definitions and results about ϕ-weak topology which are needed in the forthcoming discussions. The materials are mainly extracted from Claus [6] which gives an excellent overview on the subject, we refer readers to [6, Chapter 2], Claus et al. [7] and Föllmer and Schied [15] for more comprehensive discussions.

Definition 3.1 *Let $\phi : \mathbb{R}^k \to [0, \infty)$ be a continuous function and*

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

In the particular case when $\phi(\cdot) := \|\cdot\|^p$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^k$ and p is a positive number, write $\mathcal{M}_k^p$ for $\mathcal{M}_k^{\|\cdot\|^p}$.

Note that $\mathcal{M}_k^\phi$ defines a subset of probability measures in $\mathscr{P}(\mathbb{R}^k)$ which satisfies the generalized moment condition of ϕ. From the definition, we can see that $\mathcal{M}_k^{p_2} \subset \mathcal{M}_k^{p_1}$ for any positive numbers p_1, p_2 with $p_1 < p_2$.

Remark 3.1 *It is possible to characterize the set $\mathcal{M}_k^\phi$ in the case when $k = 1$ using the concept of* tail-index *[1]. Let $\xi : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ be a continuous random variable with cdf $F(t)$. The right tail index of ξ is a real-valued number l_0 such that $(1 - F(t))t^l$ goes to either zero or infinity respectively when l is below or above l_0 as $t \to \infty$. Likewise, the left tail index can be defined as l_0 such that $F(t)(-t)^l$ goes to either zero or infinity when l is below or above l_0 as $t \to -\infty$. Suppose that $\phi(t)$ is order $|t|^p$ when $|t|$ goes to infinity. Then $l_0 > p$ in order for $P = \mathbb{P} \circ \xi^{-1} \in \mathcal{M}_1^\phi$. In other words, $\mathcal{M}_1^\phi$ defines a set of distributions with tail index greater than p.*

Definition 3.2 (ϕ-weak topology) *Let $\phi : \mathbb{R}^k \to [0, \infty)$ be a gauge function, that is, ϕ is continuous and $\phi \geq 1$ holds outside a compact set. Define $\mathcal{C}_k^\phi$ the linear space of all continuous functions $h : \mathbb{R}^k \to \mathbb{R}$ for which there exists a positive constant c such that*

$$|h(t)| \leq c(\phi(t) + 1), \forall t \in \mathbb{R}^k. \tag{3.13}$$

The ϕ-weak topology, denoted by τ_ϕ, is the coarsest topology on $\mathcal{M}_k^\phi$ for which the mapping $g_h : \mathcal{M}_k^\phi \to \mathbb{R}$ defined by

$$g_h(P') := \int_{\mathbb{R}^k} h(t)P'(dt),\ h \in \mathcal{C}_k^\phi$$

is continuous. A sequence $\{P_l\} \subset \mathcal{M}_k^\phi$ is said to converge ϕ-weakly to $P \in \mathcal{M}_k^\phi$ written $P_l \xrightarrow{\phi} P$ if it converges w.r.t. τ_ϕ.

It is well known (see [32, Lemma 3.4]) that ϕ-weak convergence is equivalent to weak convergence, denoted by $P_l \xrightarrow{w} P$, together with

$$\int_{\mathbb{R}^k} \phi(t)P_l(dt) \to \int_{\mathbb{R}^k} \phi(t)P(dt).$$

Moreover, it follows by [30, 32] that the ϕ-weak topology on $\mathcal{M}_k^\phi$ is generated by the metric $\mathsf{dl}_\phi : \mathcal{M}_k^\phi \times \mathcal{M}_k^\phi \to \mathbb{R}$ defined by

$$\mathsf{dl}_\phi(P, Q) := \mathsf{dl}_{\mathrm{Prok}}(P, Q) + \left| \int_{\mathbb{R}^k} \phi(t)P(dt) - \int_{\mathbb{R}^k} \phi(t)Q(dt) \right|, \tag{3.14}$$

for $P, Q \in \mathcal{M}_k^\phi$, where $\mathsf{dl}_{\mathrm{Prok}} : \mathscr{P}(\mathbb{R}^k) \times \mathscr{P}(\mathbb{R}^k) \to \mathbb{R}_+$ is the Prokhorov metric:

$$\mathsf{dl}_{\mathrm{Prok}}(P, Q) := \inf\{\epsilon > 0 : P(A) \leq Q(A^\epsilon) + \epsilon, \forall A \in \mathcal{B}(\mathbb{R}^k)\}, \tag{3.15}$$

where $A^\epsilon := A + B_\epsilon(0)$ denotes the Minkowski sum of A and the open ball centred at 0 (w.r.t. Euclidean norm). Let

$$\mathsf{dl}_\phi^m(P, Q) := \left| \int_{\mathbb{R}^k} \phi(t)P(dt) - \int_{\mathbb{R}^k} \phi(t)Q(dt) \right|, \tag{3.16}$$

We can see that $\mathsf{dl}_\phi^m(P,Q)$ measures the discrepancy between P and Q in terms of their moment w.r.t. ϕ. When $\phi \equiv 1$, the second term in (3.14) disappears and consequently $\mathsf{dl}_\phi(P,Q) = \mathsf{dl}_{\text{Prok}}(P,Q)$. In that case, the ϕ-weak topology reduces to the usual topology of weak convergence (defined through bounded continuous functions). Equivalence between the two topologies may be established over a set which satisfies some uniform integration conditions, see [47, Lemma 3.4] or [31, Theorem 2.3].

Remark 3.2 *It is important to note that ϕ-weak convergence should be distinguished from convergence under some pseudo-metric. In the latter case, one often considers a class of functions $\mathcal{H}$ satisfying (3.13) and then defines a pseudo-metric by*

$$\mathsf{dl}_{\mathcal{H}}(P,Q) := \sup_{h\in\mathcal{H}} |\mathbb{E}_P[h(\xi)] - \mathbb{E}_Q[h(\xi)]|, \forall P,Q \in \mathscr{P}(\mathbb{R}^k).$$

Convergence of P_l to P under $\mathsf{dl}_{\mathcal{H}}$ implies uniform convergence

$$\sup_{h\in\mathcal{H}} \left| \int_{\mathbb{R}^k} h(t)P_l(dt) \to \int_{\mathbb{R}^k} h(t)P(dt) \right| \to 0, \tag{3.17}$$

but it does not indicate under which topology P_l converges to P. Conversely convergence under ϕ-weak topology gives a clear topological structure of how P_l converges to P but it does not guarantee the uniform convergence (3.17).

Using the notion of ϕ-weak topology, we are able to establish the continuity of ϱ defined as in (2.4).

Proposition 3.1 *Let ϱ be defined as in (2.4). Assume: (a) there is a gauge function $\psi : \mathbb{R} \to [0,+\infty)$ such that*

$$|u(s)| \le \psi(s), \forall u \in \mathcal{U}, s \in \mathbb{R}, \tag{3.18}$$

and (b) for any constant $M > 0$,

$$\mathcal{U}_{[-M,M]} := \{u_{[-M,M]}(s) := \max\{\min\{u(s), M\}, -M\} \text{ for } s \in \mathbb{R}, u \in \mathcal{U}\} \tag{3.19}$$

is equi-continuous over $\mathbb{R}$, i.e., for any $\epsilon > 0$, there exists $\delta > 0$ such that for any $u \in \mathcal{U}$,

$$|u_{[-M,M]}(s_1) - u_{[-M,M]}(s_2)| \le \epsilon, \forall s_1, s_2 \in \mathbb{R} \,\text{with}\, |s_1 - s_2| \le \delta.$$

Then $\varrho : \mathcal{M}_1^\psi \to \mathbb{R}$ is continuous and the infimum in (2.4) can be achieved when $\mathcal{U}$ is a compact set under some topology.

Proof. We first show continuity of $\varrho(\cdot)$. Since $(\mathcal{M}_1^\psi, \tau_\psi)$ is a Polish space, by [6, Theorem 2.59], it suffices to show that for each fixed $\sigma \in \mathcal{M}_1^\psi$, $\varrho(\sigma_l) \to \varrho(\sigma)$ for any sequence $\{\sigma_l\} \subset \mathcal{M}_1^\psi$ converging to σ under ψ-weak topology.

Let $\sigma_l \xrightarrow{\psi} \sigma$. By [6, Lemma 2.61], for any $\epsilon > 0$, there exist $M > 0$ and $l_0 \in \mathbb{N}$, where $\mathbb{N}$ denotes the set of all positive integers, such that

$$\int_{\mathbb{R}} \psi(s)\mathbb{1}_{(M,\infty)}(\psi(s))\sigma(ds) \le \epsilon \text{ and } \int_{\mathbb{R}} \psi(s)\mathbb{1}_{(M,\infty)}(\psi(s))\sigma_l(ds) \le \epsilon, \tag{3.20}$$

for all $l \geq l_0$. Thus

$$
\begin{aligned}
|\varrho(\sigma_l) - \varrho(\sigma)| &\leq \left| \inf_{u\in\mathcal{U}} \int_{\mathbb{R}} u\,(s)\,\sigma_l(ds) - \inf_{u\in\mathcal{U}} \int_{\mathbb{R}} u\,(s)\,\sigma(ds) \right| \\
&\leq \sup_{u\in\mathcal{U}} \left| \int_{\mathbb{R}} u\,(s)\,\sigma_l(ds) - \int_{\mathbb{R}} u\,(s)\,\sigma(ds) \right| \\
&\leq \sup_{u\in\mathcal{U}} \left| \int_{\mathbb{R}} u\,(s)\,\sigma_l(ds) - \int_{\mathbb{R}} u_{[-M,M]}\,(s)\,\sigma_l(ds) \right| \\
&\quad + \sup_{u\in\mathcal{U}} \left| \int_{\mathbb{R}} u_{[-M,M]}\,(s)\,\sigma_l(ds) - \int_{\mathbb{R}} u_{[-M,M]}\,(s)\,\sigma(ds) \right| \\
&\quad + \sup_{u\in\mathcal{U}} \left| \int_{\mathbb{R}} u_{[-M,M]}\,(s)\,\sigma(ds) - \int_{\mathbb{R}} u\,(s)\,\sigma(ds) \right|. \qquad (3.21)
\end{aligned}
$$

Let R_1, R_2, R_3 denote respectively the three terms at the right hand side of the last inequality. We will show in the next that they tend to zero as l goes to infinity. First, under condition (3.18), we obtain from (3.20) that

$$
\begin{aligned}
R_1 &\leq \sup_{u\in\mathcal{U}} \int_{\mathbb{R}} |u\,(s)\,| \mathbb{1}_{(M,\infty)}(|u(s)|)\sigma_l(ds) \qquad (3.22) \\
&\leq \sup_{u\in\mathcal{U}} \int_{\mathbb{R}} \psi(s)\mathbb{1}_{(M,\infty)}(\psi(s))\sigma_l(ds) \leq \epsilon
\end{aligned}
$$

for all $l \geq l_0$, and

$$
\begin{aligned}
R_3 &\leq \sup_{u\in\mathcal{U}} \int_{\mathbb{R}} |u\,(s)\,| \mathbb{1}_{(M,\infty)}(|u(s)|)\sigma(ds) \\
&\leq \sup_{u\in\mathcal{U}} \int_{\mathbb{R}} \psi(s)\mathbb{1}_{(M,\infty)}(\psi(s))\sigma(ds) \leq \epsilon, \qquad (3.23)
\end{aligned}
$$

where $\mathbb{1}_{(M,\infty)}(t) = 1$ if $t \in (M, \infty)$ otherwise 0. To complete the proof, we are now left to show $R_2 \to 0$ as $l \to \infty$. This is essentially about weak convergence uniformly for $u \in \mathcal{U}_{[-M,M]}$, that is,

$$
R_2 = \sup_{u\in\mathcal{U}_{[-M,M]}} \left| \int_{\mathbb{R}} u\,(s)\,\sigma_l(ds) - \int_{\mathbb{R}} u\,(s)\,\sigma(ds) \right| \leq \epsilon \qquad (3.24)
$$

for all $l \geq l_0$ and we find that [39, Theorem 3.2] aims to address this kind of convergence. The only condition that we need to verify for this effect is the equi-continuity of the integrand functions but this has been explicitly assumed in condition (b). Combining (3.21)-(3.24), we have $|\varrho(\sigma_l) - \varrho(\sigma)| \leq 3\epsilon$. The conclusion follows as ϵ can be arbitrarily small.

To see how the infimum in (2.4) can be achieved, let $w(u, \sigma) := \int_{\mathbb{R}} u(t)\sigma(dt)$. For each fixed σ, we can show $w(u, \sigma)$ is continuous w.r.t. u in the sense of weak convergence under condition (a). The rest follows with the compactness of $\mathcal{U}$. ■

Condition (a) plays an important role in the proposition. let us make a few comments about this. The condition effectively sets a bound on the growth of utility function $u(s)$ as s varies. The shape of the utility function is determined by the decision maker's preference towards gains and losses. In the well-established expected utility theory in economics, individuals are assumed to be risk averse, i.e., prefer a certain gain to any risky gains with equal expected value. Under these circumstances, the associated utility function is concave. Likewise a convex utility function may be used if individuals are risk taking. However, the results generated by experimentalists in psychology and economics find that most individuals are more risk averse in the case of gains and risk taking in the case of losses [5, 38]. S-shaped utility functions are subsequently proposed to describe this kind of phenomena, see [14, 28].

Condition (b) is equivalent to the equi-continuity of $\mathcal{U}$ over any compact set in $\mathbb{R}$ when ψ defined as in (3.18) is coercive ($\psi(s) \to \infty$ as $s \to \infty$). The latter is satisfied by many utility functions which are uniformly locally Lipschitz continuous. For instance, if $\mathcal{U}$ consists of a class of increasing and concave utility functions majorized by $\psi(t) = |t|^p$ for some $p > 1$, then the functions in $\mathcal{U}$ are equi-continuous over any compact interval of $\mathbb{R}$. Note that equi-continuity reduces to normal continuity when $\mathcal{U}$ is a singleton.

In this paper, the utility functions in the ambiguity set $\mathcal{U}$ are defined over $\mathbb{R}$ and we assume that the nature of the utility functions such as concave, convex or S-shaped are known although specific structure of the ambiguity set is not given. Under these circumstances, one may be able to identify the gauge function ψ. Let us use a simple example to illustrate. Consider the case that $\mathcal{U}$ is a class of S-shaped utility functions with

$$u(t) = \begin{cases} (1 - e^{-\alpha t})/\alpha & \text{for } t \geq 0, \\ \lambda(e^{\beta t} - 1)/\beta & \text{for } t < 0, \end{cases}$$

where $\lambda \geq 1$ and $\alpha, \beta \in [1, 5]$ are parameters. Then

$$\psi(t) = \begin{cases} 1 - e^{-t} & \text{for } t \geq 0, \\ \lambda(e^{5t} - 1)/5 & \text{for } t < 0. \end{cases}$$

In the case when $\mathcal{U}$ is compact, we may use the well-known Berge's maximum theorem (see Theorem 6.2 in the Appendix) to prove the continuity of $\varrho(\cdot)$. The compactness of $\mathcal{U}$ depends on the topological structure in the space of utility functions. For instance, if $\mathcal{U}$ is a class of nondecreasing utility functions normalized with $\lim_{t\to-\infty} u(t) = 0$ and $\lim_{t\to\infty} u(t) = 1$, then each utility function may be treated a cdf of some random variable. In that case, the well-known Prokhorov theorem [37] can be exploited to guarantee that $\mathcal{U}$ is compact under usual topology of weak convergence if and only if it is closed and tight, where the tightness means for any $\epsilon > 0$, there exist numbers $M_1 < M_2$ such that $\max\{u(M_1), 1 - u(M_2)\} \leq \epsilon$ for all $u \in \mathcal{U}$. Since u is bounded, $\varrho(\cdot)$ is continuous under topology of weak convergence.

3.2 Continuity of $\vartheta(P)$

We start by deriving a gauge function which majorizes $u(f(x,t))$ for all $u \in \mathcal{U}$ and $x \in X$. To this effect, we make some appropriate assumptions on f and $\mathcal{U}$.

Assumption 3.1 (Growth condition) *Let f be defined as in the (PRO) model. There are an exponent $\gamma > 0$ and a continuous function $\eta : \mathbb{R}^n \to \mathbb{R}$ such that*

$$|f(x,t)| \leq \eta(x)(\|t\|^\gamma + 1), \forall (x,t) \in X \times \mathbb{R}^k. \tag{3.25}$$

The assumption is known as a growth condition where the right hand side of (3.25) controls the growth of $|f(x,t)|$ w.r.t. the variation of the second argument. Let $C := \sup_{x\in X} \eta(x)$. Since X is assumed to be a compact set throughout the paper, then $C < \infty$. We assume without loss of generality that $C \geq 1$. Let

$$\phi(t) := \max\left(\sup_{u\in\mathcal{U}} u(C(\|t\|^\gamma + 1)), \sup_{u\in\mathcal{U}} -u(-C(\|t\|^\gamma + 1))\right), \tag{3.26}$$

where $\mathcal{U}$ is defined as in the PRO model (1.2). Then the growth condition and monotonic increasing property of u imply that for all $(x,t) \in X \times \mathbb{R}^k$ and $u \in \mathcal{U}$,

$$u(f(x,t)) \leq u\left(C(\|t\|^\gamma + 1)\right) \leq \phi(t) \text{ for } f(x,t) \geq 0,$$
$$-u(f(x,t)) \leq -u\left(-C(\|t\|^\gamma + 1)\right) \leq \phi(t) \text{ for } f(x,t) < 0,$$

which means

$$|u(f(x,t))| \leq \phi(t), \forall (x,t) \in X \times \mathbb{R}^k, u \in \mathcal{U}. \tag{3.27}$$

Note that we assume in Proposition 3.1 that there exists a gauge function ψ such that $|u(s)| \leq \psi(s)$ for all $u \in \mathcal{U}$ and $s \in \mathbb{R}$. By letting

$$\psi(s) := \sup_{u\in\mathcal{U}} |u(s)|, \tag{3.28}$$

then inequality (3.27) can be rewritten equivalently as

$$\psi(f(x,t)) \leq \phi(t), \forall (x,t) \in X \times \mathbb{R}^k. \tag{3.29}$$

We make a blanket assumption on ψ as follows.

Assumption 3.2 *The set $\mathcal{U}$ is chosen so that $\psi(s)$ defined as (3.28) is a gauge function.*

Remark 3.3 *Assumption 3.2 is satisfied when $\mathcal{U}$ consists of an increasing continuous utility function such that $u(a) < 1$ and $u(b) > 1$ for some numbers a, b with $a < b$. Many convex, concave and S-shaped utility functions satisfy this kind of property. Under Assumption 3.2, ϕ is also a gauge function. Moreover, if $P' \xrightarrow{\phi} P$, i.e.,*

$$P' \xrightarrow{w} P \text{ and } \int_{\mathbb{R}^k} \phi(t)P'(dt) \to \int_{\mathbb{R}^k} \phi(t)P(dt),$$

then by (3.29), we have

$$\int_{\mathbb{R}^k} \psi(f(x,t))P'(dt) \to \int_{\mathbb{R}^k} \psi(f(x,t))P(dt),$$

which is equivalent to

$$\int_{\mathbb{R}} \psi(s)P' \circ f(x,\cdot)^{-1}(ds) \to \int_{\mathbb{R}} \psi(s)P \circ f(x,\cdot)^{-1}(ds).$$

Together with $P' \xrightarrow{w} P$ and the continuity of f, we obtain

$$P' \circ f(x,\cdot)^{-1} \xrightarrow{\psi} P \circ f(x,\cdot)^{-1}, \forall x \in X.$$

We are now ready to state the main technical result of this subsection which describes continuity of $\varrho(\mu(x,P))$ w.r.t. perturbation of P and x.

Theorem 3.1 (Continuity) *Let ϱ and f be defined as in Proposition 2.1 and ϕ and ψ be defined as in (3.26) and (3.28) respectively. Under Assumptions 3.1 and 3.2, the following assertions hold.*

(i) For each $(x,P) \in \mathbb{R}^n \times \mathcal{M}_k^\phi$, $\mu(x,P) = P \circ f(x,\cdot)^{-1} \in \mathcal{M}_1^\psi$, $\mu : \mathbb{R}^n \times \mathcal{M}_k^\phi \to \mathcal{M}_1^\psi$ is continuous at (x,P) w.r.t. $\tau_{\mathbb{R}^n} \otimes \tau_\phi$ and τ_ψ. Moreover, under condition (b) of Proposition 3.1, $\varrho(\mu(x,P))$ is continuous at (x,P) w.r.t. $\tau_{\mathbb{R}^n} \otimes \tau_\phi$ and $\tau_{\mathbb{R}}$, i.e.,

$$\lim_{x' \to x, P' \xrightarrow{\phi} P} \varrho(\mu(x',P')) = \varrho(\mu(x,P)). \tag{3.30}$$

(ii) $\vartheta(\cdot)$ is continuous over $\mathcal{M}_k^\phi$, i.e.,

$$\lim_{P' \xrightarrow{\phi} P} \vartheta(P') = \vartheta(P). \tag{3.31}$$

(iii) Any cluster point of the sequence $\{x(P')\} \in S(P')$ as $P' \xrightarrow{\phi} P$ is contained in $S(P)$, where $S(P)$ is defined as in (2.8).

Proof. Part (i). Since ϕ is assumed to be a gauge function, by [6, Theorem 2.59], $(\mathcal{M}_k^\phi, \tau_\phi)$ is a Polish space. Thus it suffices to show that for any sequence $\{(x_l, P_l)\} \subset \mathbb{R}^n \times \mathcal{M}_k^\phi$ converging to (x,P) w.r.t. $\tau_{\mathbb{R}^n} \otimes \tau_\phi$, $\mu(x_l, P_l)$ converges to $\mu(x,P)$ w.r.t. τ_ψ.

Since $x_l \to x$ and $P_l \xrightarrow{\phi} P$, then $\delta_{x_l} \xrightarrow{w} \delta_x$ and $P_l \xrightarrow{w} P$. By [4, Theorem 2.8], the latter implies $\delta_{x_l} \otimes P_l \xrightarrow{w} \delta_x \otimes P$. Under the continuity of f, it follows from the continuous mapping theorem [4, Theorem 2.7] that

$$(\delta_{x_l} \otimes P_l) \circ f^{-1} \xrightarrow{w} (\delta_x \otimes P) \circ f^{-1}. \tag{3.32}$$

On the other hand,

$$\begin{aligned} & \lim_{r\to\infty} \sup_{l\in\mathbb{N}} \int_{\{s:\psi(s)>r\}} \psi(s)(\delta_{x_l} \otimes P_l) \circ f^{-1}(ds) \\ = \; & \lim_{r\to\infty} \sup_{l\in\mathbb{N}} \int_{\{(x,t):\psi(f(x,t))>r\}} \psi(f(x,t))(\delta_{x_l} \otimes P_l)(d(x,t)) \\ \le \; & \lim_{r\to\infty} \sup_{l\in\mathbb{N}} \int_{\{t:\phi(t)>r\}} \phi(t) P_l(dt) = 0, \end{aligned}$$

where the last inequality holds because $P_l \xrightarrow{\phi} P$. Together with (3.32), we have from [6, Lemma 2.61] that $(\delta_{x_l} \otimes P_l) \circ f^{-1} \xrightarrow{\psi} (\delta_x \otimes P) \circ f^{-1}$, that is, $\mu(x_l, P_l) \xrightarrow{\psi} \mu(x,P)$. By combining the continuity of ϱ in Proposition 3.1, we obtain (3.30).

Parts (ii) and (iii). Since $\vartheta(P) = \sup_{x\in X} \varrho(\mu(x,P))$, X is compact and $\varrho(\mu(\cdot,\cdot))$ is continuous, then we can use Berge's maximum theorem (see Theorem 6.2 in the Appendix) to claim the continuity of $\vartheta(\cdot)$ and upper semi-continuity of $S(\cdot)$. ■

Remark 3.4 *Part (i) of the theorem gives pointwise continuity of $\mu(x, P)$ and $(\varrho \circ \mu)(x, P)$ w.r.t. variation of (x, P) in some appropriate topological space.*

Claus et al. [7] investigate ϕ-weak continuity of a class of risk functionals and derive a similar result to Theorem 3.1, see [7, Theorem 2.2 and Corollary 2.4], and also [6, Theorem 2.82]. It is natural to ask the relationship between the two results. Let

$$\vartheta(P) = \max_{x \in X} v(x, P),$$

where $v(x, P) = \mathcal{R}((\delta_x \times P) \circ f^{-1})$. The continuity of ϑ is down to derive the continuity of the composite function $v(x, P)$. In [7], the authors assume that $\mathcal{R}$ is a risk functional of a law invariant convex risk measure defined on $L^p(\Omega, \mathcal{F}, \mathbb{P})$, which ensures continuity of $\mathcal{R}$ w.r.t. the $|\cdot|^p$-weak topology. Hence their main task is to derive the continuity of $(\delta_x \times P) \circ f^{-1}$ under the growth condition of f and verify this growth condition in two-stage mean-risk models as an application.

In comparison with [7], our task here is to establish the continuity of $\mathcal{R}$ when $\mathcal{R}(\sigma) = \varrho(\sigma) := \inf_{u \in \mathcal{U}} \int_{\mathbb{R}} u(t)\sigma(dt)$, see Proposition 3.1. To this end, we need the equi-continuity of $\mathcal{U}$ and the existence of a gauge function ψ, and then derive the continuity of $\mathcal{R}$ w.r.t. ψ-weak topology. In the case that the utility functions in $\mathcal{U}$ are all increasing concave and majorized by a specified growth order $|t|^p$ ($p > 1$), i.e., $\psi(t) = |t|^p$, $\mathcal{R}$ is continuous w.r.t. the $|\cdot|^p$-weak topology. In that specific case, Theorem 3.1 is essentially covered by [7, Theorem 2.2 and Corollary 2.4].

In this paper, our interest is to consider general utility functions which are not necessarily concave such as S-shaped utilities, consequently the continuity of $\mathcal{R}$ in our context differs from that of [7, Theorem 2.2].

The continuity result implies continuity of $v(x, P)$ (see (2.5)) and consequently we may use the latter to show robustness *of objective functional v in the sense of Embrechts, Schied and Wang [12]. To explain this, let $Q \in \mathcal{M}_k^\phi$ denote some perceived data. For the fixed Q, we obtain an optimal policy $x(Q)$ by solving (PRO), i.e., $x(Q) \in \arg\max_{x \in X} v(x, Q)$ or equivalently $v(x(Q), Q) = \vartheta(Q)$. For fixed $x(Q)$, consider a small perturbation of data Q' from the perceived data. If $v(x(Q), Q')$ is close to $v(x(Q), Q)$, then v is said to be robust at Q, see [12, Definition 1]. Using the continuity result established in Part (i), we can see that v is robust at Q provided that the perturbation Q' is restricted to $\mathcal{M}_k^\phi$.*

Note also that in the case when $\mathcal{U}$ is a singleton $\{u\}$ and $u(t) = t$, the stability results of Theorem 3.1 collapse to those in stochastic programming, see i.e. Claus [6] in the context of stochastic bilevel programming.

To see how the theorem works, let us consider an example.

Example 3.1 *Consider (PRO) model with $f(x, \xi) = x\xi$, $X = [1, 2]$, $\xi : \mathbb{R} \to \mathbb{R}$, u is a S-shaped utility function:*

$$u(t) = \begin{cases} t^\alpha & \text{if } t \geq 0, \\ -(-t)^\beta & \text{otherwise}, \end{cases} \tag{3.33}$$

parameterized by α and β. Let

$$\mathcal{U} := \{u : u \,\text{is defined as in (3.33) with}\, \alpha \in [1/4, 1/2], \beta \in [1/3, 1/2]\}.$$

Let $\psi(t)$ and $\phi(t)$ be defined as in (3.28) and (3.26) respectively. In this context,

$$\psi(t) = \sup_{u \in \mathcal{U}} |u(t)| = \begin{cases} t^{1/2} & \text{if}\, t \geq 1, \\ t^{1/4} & \text{if}\, 0 \leq t < 1, \\ (-t)^{1/3} & \text{if}\, -1 \leq t \leq 0, \\ (-t)^{1/2} & \text{if}\, t < -1, \end{cases}$$

and $\phi(t) = (2|t| + 2)^{1/2}$. Let $\{P_l\} \subset \mathscr{P}(\mathrm{I\!R})$ be a sequence of probability measures such that $P_l \xrightarrow{\phi} P$. By Definition 3.2 and the follow-up remarks, we have $P_l \xrightarrow{w} P$ and

$$\lim_{l \to \infty} \int_{\mathrm{I\!R}} (2|t| + 2)^{1/2} P_l(dt) = \int_{\mathrm{I\!R}} (2|t| + 2)^{1/2} P(dt). \tag{3.34}$$

Consider a specific case that $P_l = \mathbb{1}_{\frac{1}{l}}$ and $P = \mathbb{1}_0$. Then

$$\lim_{l \to \infty} \int_{\mathrm{I\!R}} \phi(t) \mathbb{1}_{\frac{1}{l}}(dt) = \lim_{l \to \infty} \phi\left(\frac{1}{l}\right) = \lim_{l \to \infty} \left(\frac{2}{l} + 2\right)^{1/2} = \sqrt{2} = \phi(0) = \int_{\mathrm{I\!R}} \phi(t) \mathbb{1}_0(dt).$$

Together with $P_l \xrightarrow{w} P$, we have $P_l \xrightarrow{\phi} P$. Let $\{x_l\}$ be a sequence in X such that $x_l \to x_0$. Then

(i) $\lim_{l \to \infty} \varrho(\mu(x_l, P_l)) = \lim_{l \to \infty} \inf_{u \in \mathcal{U}} \int_{\mathrm{I\!R}} u(f(x_l \xi)) \mathbb{1}_{\frac{1}{l}}(dt) = \lim_{l \to \infty} \inf_{u \in \mathcal{U}} u\left(\frac{x_l}{l}\right) = 0 = \varrho(\mu(x_0, P_0))$,

(ii) $\lim_{l \to \infty} \vartheta(P_l) = \lim_{l \to \infty} \max_{x \in X} \varrho(\mu(x, P_l)) = \lim_{l \to \infty} \max_{x \in X} \inf_{u \in \mathcal{U}} u\left(\frac{x}{l}\right) = \lim_{l \to \infty} \left(\frac{2}{l}\right)^{1/4} = 0 = \vartheta(P_0)$,

(iii) for each l, $S(P_l) = \arg\max_{x \in X} \vartheta(P_l) = \{2\} \subset [1, 2] = S(P_0) = \arg\max_{x \in X} \vartheta(P_0)$.

To see how the utility function affects the continuity, let us consider the case when $u(t) = t$. Consequently $\phi(t) = 2|t| + 2$ and the convergence of P_l to P w.r.t. τ_ϕ implies

$$\int_{\mathrm{I\!R}} (2|t| + 2) P_l(dt) \to \int_{\mathrm{I\!R}} (2|t| + 2) P(dt). \tag{3.35}$$

Comparing the convergence of sequence $\{P_l\}$ in (3.34) with that in (3.35), we can see easily that the former is implied by the latter in the sense that P_l satisfying (3.35) must satisfy (3.34) but not conversely. This is because $u(x\xi)$ goes to infinity at a slower rate than $x\xi$ as ξ goes to infinity in both directions. In other words, $\vartheta(P)$ is more likely to be continuous with the composition of the S-shaped utility function when P is perturbed locally in the space of probability measures $\mathscr{P}(\mathrm{I\!R})$ when u is convex at the left tail and concave at the right tail. In the extreme case when u is a constant, $\vartheta(P)$ is continuous over $\mathscr{P}(\mathrm{I\!R})$.

4 Statistical robustness

4.1 Qualitative statistical robustness

We now move on to discuss statistical robustness of the (PRO) model. Let's start with a formal definition of a statistical estimator $T(\cdot)$, which is based on Krätschmer et al. [30, Definition 2.11].

Definition 4.1 (Statistical robustness) *Let $\mathcal{M} \subset \mathscr{P}(\mathbb{R}^k)$ be a set of probability measures and dl_ϕ be defined as in (3.14) for some gauge function $\phi : \mathbb{R}^k \to \mathbb{R}$. A statistical estimator T is said to be statistically robust on $\mathcal{M}$ with respect to dl_ϕ and $\mathsf{dl}_{\text{Prok}}$ if for all $P \in \mathcal{M}$ and $\epsilon > 0$, there exist $\delta > 0$ and $N_0 \in \mathbb{N}$ such that*

$$Q \in \mathcal{M}, \mathsf{dl}_\phi(P,Q) \le \delta \Longrightarrow \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{T}_N^{-1}, Q^{\otimes N} \circ \hat{T}_N^{-1}\right) \le \epsilon, \text{ for } N \ge N_0,$$

where $\hat{T}_N$, $P^{\otimes N}$ and $Q^{\otimes N}$ are defined in Section 2.

The definition is based on Hampel's classical concept of statistical robustness [20] of an estimator which requires that a small change in the law of the data results in only a small change of law of the estimator. The analysis here is carried out not only for all Q in a neighborhood of P but also for all P in the set $\mathcal{M}$. To this effect, we need the following Uniform Glivenko-Cantelli property.

Definition 4.2 (Uniform Glivenko-Cantelli property) *Let $\mathcal{M}$ be a subset of $\mathscr{P}(\mathbb{R}^k)$. We say the metric space $(\mathcal{M}, \mathsf{dl})$ has the Uniform Glivenko-Cantelli (UGC) property if for every $\epsilon > 0$ and $\delta > 0$, there exists $N_0 \in \mathbb{N}$ such that for all $P \in \mathcal{M}$ and $N \ge N_0$*

$$P^{\otimes N}\left[(\xi^1, \cdots, \xi^N) : \mathsf{dl}(P, P_N) \ge \delta\right] \le \epsilon. \tag{4.36}$$

The UGC property states that for every empirical probability measure P_N generated by $P \in \mathcal{M}$, P_N is close to P under the metric dl when N is sufficiently large. The concept is a step further from weak convergence on two-fold: (a) the convergence is in the sense of dl, (b) the convergence is uniform for all $P \in \mathcal{M}$. Now we turn to establish the statistical robustness of the optimal value and the optimal solutions.

Theorem 4.1 (Statistical robustness of (PRO)) *Let $\phi(t)$ be defined as in (3.26) and*

$$\mathcal{M}_{k,\kappa}^{\phi^p} := \left\{ P \in \mathscr{P}(\mathbb{R}^k) : \int_{\mathbb{R}^k} \phi(t)^p P(dt) \le \kappa \right\} \tag{4.37}$$

for some fixed $p > 1$ and $\kappa > 0$, let $\mathcal{M} \subset \mathcal{M}_{k,\kappa}^{\phi^p}$ be a subset. Under Assumptions 3.1 and 3.2, and condition (b) of Proposition 3.1, the following assertions hold.

(i) For any $P \in \mathcal{M}$ and $\epsilon > 0$, there exist positive numbers $\delta > 0$ and $N_0 \in \mathbb{N}$ such that

$$Q \in \mathcal{M}, \mathsf{dl}_\phi(P,Q) \le \delta \Longrightarrow \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{\vartheta}_N^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}\right) \le \epsilon \tag{4.38}$$

for all $N \ge N_0$.

(ii) If, in addition, for each $Q \in \mathcal{M}$, $S(Q_N)$ and $S(Q)$ are singletons, then for any small number $\epsilon > 0$, there exist positive numbers $\delta > 0$ and $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$

$$Q \in \mathcal{M}, \mathsf{dl}_\phi(P,Q) \leq \delta \Longrightarrow \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{S}_N^{-1}, Q^{\otimes N} \circ \hat{S}_N^{-1}\right) \leq \epsilon,$$

where $S(Q_N)$ is defined as in (2.12).

The results follow straightforwardly from Theorem 3.1 (ii) and [32, Theorem 2.4]. We include a proof in the appendix for self-containing.

We make a few comments about the conditions and results of this theorem.

First, the set $\mathcal{M}_{k,\kappa}^{\phi^p}$ differs from $\mathcal{M}_k^{\phi^p}$ (see Definition 3.1) in that the former imposes a bound for the moment value uniformly for all $P \in \mathcal{M}_{k,\kappa}^{\phi^p}$ whereas the latter does not have such uniformity. This is because we need the UGC property of $(\mathcal{M}_{k,\kappa}^{\phi^p}, \mathsf{dl}_\phi)$ in order for us to apply [32, Theorem 2.4].

Second, by combining (6.87) and (6.88) in Appendix, we can obtain for any $\epsilon > 0$, there exist constants $\delta > 0$ and $N_0 \in \mathbb{N}$ such that

$$Q \in \mathcal{M}, \mathsf{dl}_\phi(P,Q) \leq \delta \Longrightarrow Q^{\otimes N}\left[\tilde{\xi} \in (\mathbb{R}^k)^N : |\vartheta(Q) - \vartheta(Q_N)| \leq \frac{\epsilon}{2}\right] \geq 1 - \frac{\epsilon}{2}$$

for $N \geq N_0$. This implies uniform convergence of $\vartheta(Q_N)$ to $\vartheta(Q)$ for all Q near P as opposed to pointwise convergence (for each fixed Q) in stochastic programming. The uniformity does not come out for free: it restricts both P and Q to the ϕ-weak topological space of probability measures.

Third, the left hand side of (4.38) restricts the perturbation of probability to the set $\mathcal{M}$. Cont, Deguest and Scandolo [8] seem to be the first to consider this kind of restriction in their study of statistical robustness whereby in the place of dl_ϕ and $\mathsf{dl}_{\text{Prok}}$ they use Lévy distance/metric and $\mathcal{M}$ is some plausible set of distributions that the underlying random variable may take. When any perturbation of the distribution is allowed, the statistical robustness corresponds to the notion of qualitative robustness (also called asymptotic robustness) as outlined by Huber in [26]. This case is not generally interesting in econometric or financial applications since requiring robustness against all perturbations of the model is quite restrictive and excludes even estimators such as the sample mean, see Cont, Deguest and Scandolo [8]. Krätschmer et al. [32] observe that this kind of robustness depends heavily on the topology of the space of probability measures that perturbation is considered and propose the current version of statistical robustness. Our result in this theorem is a direct application of their notion to PRO models.

Fourth, In practice, since P is unknown, it is difficult to identify δ for a specified ϵ. The usefulness of (4.38) should be understood as that it provides a theoretical guarantee: if the perceived data is generated by some probability distribution Q which is close to the true distribution P, and Q satisfies moment condition (4.37) (which may be examined through empirical data), then the optimal value obtained with the perceived data is close to the one with real data. There are potentially two ways to move forward the research. One is to derive quantitative statistical robustness under some additional conditions in which case the relationship between ϵ and δ may be explicitly established, we will come back to this in Section 4.2. The other is to use the perceived data to construct an ambiguity set of probability distributions and use the latter to develop a model which is robust both in preference and in brief. This will effectively create a robust mechanism to mitigate the risk arising from noise in perceived data, this kind of model is similar to Haskell et al. [21] which combines PRO with DRO if we restrict the ambiguity set of probability distributions in the latter to those within the ϕ-weak topology.

Fifth, Krätschmer et al. [32] introduce an index to measure the degree of statistical robustness of a statistical functional T. If the gauge function $\tilde{\phi}_\lambda(t) = (1 + \|\cdot\|^\gamma)^\lambda$ for $\lambda \geq 0$ and γ being taken from Assumption 3.1, then the index is defined as

$$\text{iqr}(T) := (\inf\{\lambda\gamma \in [0, \infty) : T \text{ is qualitatively robust w.r.t. } \mathsf{dl}_{\tilde{\phi}_\lambda}\})^{-1}, \tag{4.39}$$

where "iqr" is the acronym of "index of qualitative robustness". A higher index reflects a higher degree of robustness. The following examples explain this.

Example 4.1 *Consider Example 3.1 where $\mathcal{U}$ is a set of S-shaped utility functions defined as in (3.33) and $\phi(t) = (2|t| + 2)^{1/2}$. It follows from Theorem 5.2 that*

$$\begin{aligned} \text{iqr}(\vartheta) &= \left(\inf\left\{p/2 \in [0, \infty) : \vartheta \text{ is qualitatively robust w.r.t. } \mathsf{dl}_{\tilde{\phi}_{p/2}}\right\}\right)^{-1} \\ &= (1/2)^{-1} = 2. \end{aligned}$$

Likewise, if u takes a linear form, then the corresponding index for ϑ is 1. Therefore, ϑ in the former case is more robust than that in the latter.

Example 4.2 *Consider the case where the ambiguity set $\mathcal{U}$ is defined by a linear system of inequalities*

$$\mathcal{U} := \left\{u : \int_{\mathbb{R}} \psi_j(t) du(t) \leq c_j, \text{for } j = 1, \cdots, m\right\}, \tag{4.40}$$

where $\psi_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \cdots, m$ are a class of u-integrable functions and $c_j, j = 1, \cdots, m$ are some given constants. Hu and Mehrotra [24] consider this kind of conditions to characterize ambiguity set of utility functions defined over a finite interval and scaled to $[0, 1]$ and demonstrate that it covers a wide range of ambiguity sets elicited in practice including pairwise comparisons considered by Armbruster and Delage [2], certainty equivalent and ball under some metrics [25] or pseudo-metrics. Here we consider the case that u is defined over $\mathbb{R}$ and investigate how the tail behaviour of ψ_j affects the tails of u. Note that if elicitation is a lottery with two outcomes, then the corresponding ψ_j is a step function in which case ψ_j does not have a tail. Our focus here is on the case that ψ_j has a tail, which may result from pairwise comparison $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$ where X and Y are continuously distributed with unbounded ranges.

Suppose that u is differentiable and the growth order of its derivative function u' is p_0. Assume also that ψ_j has the growth order l_j for $j = 1, \cdots, m$. Then the inequalities in (4.40) imply that $1/p_0 + 1/l_j > 1$, i.e., $p_0 < \frac{l_j}{l_j - 1}$ for $j = 1, \cdots, m$. Let $\bar{l} := \min\{\frac{l_j}{l_j-1}, j = 1, \cdots, m\} + 1$.

Assuming the ambiguity set $\mathcal{U}$ defined as in (4.40) is used in the (PRO) model. We now examine how $\bar{l}$ is related to the statistical robustness of the optimal value. To this end, we go *back to (3.26) to identify the gauge function $\phi(t)$. Under the growth condition of f, $\phi(t)$ can be chosen as $\phi(t) = [C(1 + \|t\|^\gamma)]^{\bar{l}}$. With the gauge function, we can use the moment condition (4.40) to identify the set $\mathcal{M}^{\phi^p}_{k,\kappa}$, i.e., all $P \in \mathscr{P}(\mathbb{R}^k)$ satisfying*

$$\int_{\mathbb{R}^k} [C(1 + \|t\|^\gamma)]^{\bar{l}p} P(dt) \le \kappa$$

for some $p > 1$ and $\kappa > 0$, and the index for ϑ is $1/(\gamma\bar{l})$. The discussions above indicate that a lower growth order of the gauge function ϕ would imply that the proposed (PRO) model (to be used by the investor) is more likely to be statistically robust.

An interesting question is whether PRO models considered in Armbruster and Delage [2] and Hu and Mehrotra [24] are statistically robust assuming the underlying uncertainty is continuously distributed (they focus on discrete distribution case in order to derive tractable formulations of their PRO models). The answer is yes in Hu and Mehrotra's model in that the utility functions in the ambiguity set are equi-continuous under the boundedness conditions of their values and derivatives. Moreover, since the range of random outcomes is bounded, $\mathcal{M}^{\phi^p}_{k,\kappa}$ coincides with $\mathscr{P}(\mathbb{R}^k)$. The answer is not straightforward about Armbruster and Delage's model in that we can easily construct a counter example such that the ambiguity set of concave utility functions satisfying their normalization condition are not equi-continuous, which means condition (b) in Proposition 3.1 is not satisfied. However, since their worst case utility function is identified through upper envelope of linear functions, it does not affect their results if we impose a bound on the Lipschitz modulus of their utility functions, which means that our statistically robust analysis can be readily applied to their models by artificially imposing a Lipschitz boundedness condition.

In summary, statistical robustness depends on the locally Lipschitz continuity of the utility functions. When the utility functions are increasing concave and controlled by a gauge function with some specified polynomial growth, the condition is satisfied, see Remark 3.4. In general non-concave utility case, we believe decision maker's marginal utility does not change sharply at any fixed point in many practical problems, we leave a more thorough exploration on the topic for future research.

4.2 Quantitative statistical robustness

The statistical robustness in the preceding subsection is qualitative in that there is no explicit quantitative relationship between ϵ and δ in Theorem 4.1. It might be interesting to ask whether we will be able to derive some kind of quantitative relationship between these two quantities in which case we will obtain a quantitative description of the statistical robustness.

To this end, we introduce some metrics differing from the Prokhorov metric in the space $\mathscr{P}(\mathbb{R})$ to measure the difference between $P^{\otimes N} \circ \hat{\vartheta}_N^{-1}$ and $Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}$. Let $\mathscr{G}$ be a set of measurable functions defined over $\mathbb{R}^s$. For any two probability measures $P_1, P_2 \in \mathscr{P}(\mathbb{R}^s)$, define the pseudo-metric between P_1 and P_2 by

$$\mathsf{dl}_{\mathscr{G}}(P_1, P_2) := \sup_{g \in \mathscr{G}} |\mathbb{E}_{P_1}[g(\zeta)] - \mathbb{E}_{P_2}[g(\zeta)]|. \tag{4.41}$$

The pseudo-metric is widely adopted in the literature of stochastic programming (see Römisch [41]). To simplify the discussion, let us consider the case that $\mathscr{G}$ consists of all Lipschitz continuous functions $g : \mathbb{R}^s \to \mathbb{R}$ with modulus being bounded by 1. In that case $\mathsf{dl}_{\mathscr{G}}(P_1, P_2)$ reduces to the Kantorovich metric $\mathsf{dl}_{K,s}(P_1, P_2)$, where the subscript K, s indicates the Kantorovich metric in $\mathscr{P}(\mathbb{R}^s)$.

Let $P, Q \in \mathcal{P}(\mathbb{R}^k)$ 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 following lemma establishes a bound of $\mathsf{dl}_{\mathscr{G}}(P^{\otimes N}, Q^{\otimes N})$ in terms of $\mathsf{dl}_{K,k}(P, Q)$ when $\mathscr{G}$ is a specific class of Lipschitz continuous functions on $(\mathbb{R}^k)^{\otimes N}$.

Lemma 4.1 *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.,*

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

Then $\mathsf{dl}_{\Psi}(P^{\otimes N}, Q^{\otimes N}) \le L\mathsf{dl}_{K,k}(P, Q)$.

Proof. Let $\xi^{-j} := \{\xi^1, \cdots, \xi^{j-1}, \xi^{j+1}, \cdots, \xi^N\}$, $\vec{\xi^j} := \{\xi^1, \cdots, \xi^j\}$ and $\vec{\xi}^{-j} := \{\xi^{j+1}, \cdots, \xi^N\}$. For any $P_1, \cdots, P_N \in \mathscr{P}(\mathbb{R}^k)$ and any $j \in \{1, \cdots, N\}$, denote

$$P_{-j}(d\xi^{-j}) := P_1(d\xi^1) \cdots P_{j-1}(d\xi^{j-1}) P_{j+1}(d\xi^{j+1}) \cdots P_N(d\xi^N)$$

and $h_{\xi^{-j}}(\xi^j) := \int_{(\mathbb{R}^k)^{\otimes(N-1)}} \psi(\xi^{-j}, \xi^j) P_{-j}(d\xi^{-j})$. Then

$$\begin{aligned} |h_{\xi^{-j}}(\tilde{\xi}^j) - h_{\xi^{-j}}(\hat{\xi}^j)| &\le \int_{(\mathbb{R}^k)^{\otimes(N-1)}} \left| \psi(\xi^{-j}, \tilde{\xi}^j) - \psi(\xi^{-j}, \hat{\xi}^j) \right| P_{-j}(d\xi^{-j}) \\ &\le \int_{(\mathbb{R}^k)^{\otimes(N-1)}} \frac{L}{N} \|\tilde{\xi}^j - \hat{\xi}^j\| P_{-j}(d\xi^{-j}) \le \frac{L}{N} \|\tilde{\xi}^j - \hat{\xi}^j\|. \end{aligned}$$

Let $\mathcal{H}$ denote the set of functions $h_{\xi^{-j}}(\xi^j)$ generated by $\psi \in \Psi$. By the definition of dl_{Ψ} and the Kantorovich metric,

$$\begin{aligned} \mathsf{dl}_{\Psi}(P_{-j} \times \tilde{P}_j, P_{-j} \times \hat{P}_j) &= \sup_{\psi \in \Psi} \left| \int_{\mathbb{R}^k} \int_{(\mathbb{R}^k)^{\otimes(N-1)}} \psi(\xi^{-j}, \xi^j) P_{-j}(d\xi^{-j}) \tilde{P}_j(d\xi^j) \right. \\ &\quad \left. - \int_{\mathbb{R}^k} \int_{(\mathbb{R}^k)^{\otimes(N-1)}} \psi(\xi^{-j}, \xi^j) P_{-j}(d\xi^{-j}) \hat{P}_j(d\xi^j) \right| \\ &= \sup_{h_{\xi^{-j}} \in \mathcal{H}} \left| \int_{\mathbb{R}^k} h_{\xi^{-j}}(\xi^j) \tilde{P}_j(d\xi^j) - \int_{\mathbb{R}^k} h_{\xi^{-j}}(\xi^j) \hat{P}_j(d\xi^j) \right| \\ &\le \frac{L}{N} \mathsf{dl}_{K,k}(\tilde{P}_j, \hat{P}_j). \end{aligned} \tag{4.42}$$

By the triangle inequality

$$
\begin{aligned}
& \mathsf{dl}_\Psi\left(P^{\otimes N}, Q^{\otimes N}\right) \\
= & \sup_{\psi\in\Psi}\left|\int_{(\mathbb{R}^k)^{\otimes N}}\psi(\vec{\xi}^N)P^{\otimes N}(d\vec{\xi}^N)-\int_{(\mathbb{R}^k)^{\otimes N}}\psi(\vec{\xi}^N)Q^{\otimes N}(d\vec{\xi}^N)\right| \\
\leq & \sum_{j=0}^{N-1}\sup_{\psi\in\Psi}\left|\int_{(\mathbb{R}^k)^{\otimes(j)}}\int_{(\mathbb{R}^k)^{\otimes(N-j)}}\psi(\vec{\xi}^N)P^{\otimes(N-j)}(d\vec{\xi}^{N-j})Q^{\otimes(j)}(d\vec{\xi}^{-(N-j)})\right. \\
& \left.-\int_{(\mathbb{R}^k)^{\otimes(j+1)}}\int_{(\mathbb{R}^k)^{\otimes(N-j-1)}}\psi(\vec{\xi}^N)P^{\otimes(N-j-1)}(d\vec{\xi}^{N-j-1})Q^{\otimes(j+1)}(d\vec{\xi}^{-(N-j-1)})\right| \\
\leq & \sum_{j=0}^{N-1}\frac{L}{N}\mathsf{dl}_{K,k}(P,Q) = L\mathsf{dl}_{K,k}(P,Q).
\end{aligned}
$$

The last inequality is due to the fact that

$$
\begin{aligned}
& \int_{(\mathbb{R}^k)^{\otimes(j)}}\int_{(\mathbb{R}^k)^{\otimes(N-j)}}\psi(\vec{\xi}^N)P^{\otimes(N-j)}(d\vec{\xi}^{N-j})Q^{\otimes(j)}(d\vec{\xi}^{-(N-j)}) \\
& -\int_{(\mathbb{R}^k)^{\otimes(j+1)}}\int_{(\mathbb{R}^k)^{\otimes(N-j-1)}}\psi(\vec{\xi}^N)P^{\otimes(N-j-1)}(d\vec{\xi}^{N-j-1})Q^{\otimes(j+1)}(d\vec{\xi}^{-(N-j-1)}) \\
& =\int_{(\mathbb{R}^k)^{\otimes(j)}}\left[\int_{\mathbb{R}^k}\left(\int_{(\mathbb{R}^k)^{\otimes(N-j-1)}}\psi(\vec{\xi}^N)P^{\otimes(N-j-1)}(d\vec{\xi}^{N-j-1})\right)P(d\xi^j)\right]Q^{\otimes(j)}(d\vec{\xi}^{-(N-j)}) \\
& -\int_{(\mathbb{R}^k)^{\otimes(j)}}\left[\int_{\mathbb{R}^k}\left(\int_{(\mathbb{R}^k)^{\otimes(N-j-1)}}\psi(\vec{\xi}^N)P^{\otimes(N-j-1)}(d\vec{\xi}^{N-j-1})\right)Q(d\xi^j)\right]Q^{\otimes(j)}(d\vec{\xi}^{-(N-j)})
\end{aligned}
$$

and application of inequality (4.42) with $\tilde{P}_j = P$ and $\hat{P}_j = Q$. The proof is complete. ■

With the lemma, we are ready to present our main result of this subsection.

Theorem 4.2 (Quantitative statistical robustness) *Let $\phi(t)$ be defined as in (3.26) and $\mathcal{M}_k^\phi$ be defined as in Definition 3.1. Assume: (a) $u(f(x,\xi))$ is uniformly Lipschitz continuous in ξ, i.e., there exists a positive constant $L > 0$ such that*

$$|u(f(x,\xi)) - u(f(x,\xi'))| \leq L\|\xi-\xi'\|, \forall x\in X, u\in\mathcal{U};$$

(b) Assumptions 3.1 and 3.2 hold. Then for any $N\in\mathbb{N}$ and any $P,Q\in\mathcal{M}_k^\phi$

$$\mathsf{dl}_{K,1}\left(P^{\otimes N}\circ\hat{\vartheta}_N^{-1}, Q^{\otimes N}\circ\hat{\vartheta}_N^{-1}\right)\leq L\mathsf{dl}_{K,k}(P,Q). \tag{4.43}$$

Proof. By definition

$$
\begin{aligned}
& \mathsf{dl}_{K,1}\left(P^{\otimes N}\circ\hat{\vartheta}_N^{-1}, Q^{\otimes N}\circ\hat{\vartheta}_N^{-1}\right) \\
= & \sup_{g\in\mathscr{G}}\left|\int_{\mathbb{R}}g(t)P^{\otimes N}\circ\hat{\vartheta}_N^{-1}(dt)-\int_{\mathbb{R}}g(t)Q^{\otimes N}\circ\hat{\vartheta}_N^{-1}(dt)\right| \\
= & \sup_{g\in\mathscr{G}}\left|\int_{(\mathbb{R}^k)^{\otimes N}}g(\hat{\vartheta}(\vec{\xi}^N))P^{\otimes N}(d\vec{\xi}^N)-\int_{(\mathbb{R}^k)^{\otimes N}}g(\hat{\vartheta}(\vec{\xi}^N))Q^{\otimes N}(d\vec{\xi}^N)\right|,
\end{aligned}
\tag{4.44}
$$

where we write $\vec{\xi^N}$ for $(\xi^1, \cdots, \xi^N)$ and $\hat{\vartheta}(\vec{\xi^N})$ for $\hat{\vartheta}_N$ to indicate its dependence on $\xi^1, \cdots, \xi^N$. To see the well-definiteness of the pseudo-metric, we note that for each $g \in \mathscr{G}$,

$$|g(\hat{\vartheta}(\vec{\xi^N}))| \leq |g(\hat{\vartheta}(\vec{\xi_0^N}))| + |\hat{\vartheta}(\vec{\xi^N}) - \hat{\vartheta}(\vec{\xi_0^N})|, \tag{4.45}$$

where $\vec{\xi_0^N} \in (\mathbb{R}^k)^{\otimes N}$ is fixed. By the definition of $\hat{\vartheta}(\vec{\xi^N})$, we have

$$|\hat{\vartheta}(\vec{\xi^N})| \quad = \quad \left|\max_{x\in X} \inf_{u\in\mathcal{U}} \frac{1}{N}\sum_{k=1}^N u(f(x,\xi^k))\right| \leq \frac{1}{N}\sum_{k=1}^N \phi(\xi^k).$$

Moreover

$$\begin{aligned}\int_{(\mathbb{R}^k)^{\otimes N}} |\hat{\vartheta}(\vec{\xi^N})| P^{\otimes N}(d\vec{\xi^N}) &\leq \int_{(\mathbb{R}^k)^{\otimes N}} \frac{1}{N}\sum_{k=1}^N \phi(\xi^k) P^{\otimes N}(d\vec{\xi^N}) \\ &= \int_{\mathbb{R}^k} \phi(\xi)P(d\xi) < \infty, \forall P \in \mathcal{M}_k^\phi, \end{aligned} \tag{4.46}$$

where the last equality holds due to the fact that $\xi^1, \cdots, \xi^N$ are independent and identically distributed. Combining (4.45) and (4.46), we deduce that

$$\int_{(\mathbb{R}^k)^{\otimes N}} g(\hat{\vartheta}(\vec{\xi^N}))P^{\otimes N}(d\vec{\xi^N}) < \infty, \forall P \in \mathcal{M}_k^\phi.$$

The same argument can be made on $\int_{(\mathbb{R}^k)^{\otimes N}} g(\hat{v}(\vec{\xi^N}))Q^{\otimes N}(d\vec{\xi^N})$ for $Q \in \mathcal{M}_k^\phi$.

Next, we show (4.43). We do so by applying Lemma 4.1 to the right hand side of (4.44). To this end, we need to verify the condition of the lemma. Since $u(f(x,\xi))$ is uniformly Lipschitz continuous in ξ and g is also Lipschitz continuous with modulus bounded by 1, we have

$$\begin{aligned} &|g(\hat{\vartheta}(\tilde{\xi}^1, \cdots, \tilde{\xi}^N)) - g(\hat{v}(\hat{\xi}^1, \cdots, \hat{\xi}^N))| \leq |\hat{\vartheta}(\tilde{\xi}^1, \cdots, \tilde{\xi}^N) - \tilde{\vartheta}(\hat{\xi}^1, \cdots, \hat{\xi}^N)| \\ \leq\ & \frac{1}{N}\sum_{k=1}^N \sup_{x\in X, u\in\mathcal{U}} |u(f(x,\tilde{\xi}^k)) - u(f(x,\hat{\xi}^k))| \leq \frac{L}{N}\sum_{k=1}^N \|\tilde{\xi}^k - \hat{\xi}^k\|, \end{aligned}$$

which means that $g(\hat{\vartheta}(\cdot))$ is Lipschitz continuous over $(\mathbb{R}^k)^{\otimes N}$ with Lipschitz modulus bounded by L/N. The rest follows from Lemma 4.1 by setting $\psi(\xi^1, \cdots, \xi^N) = g(\hat{v}(\xi^1, \cdots, \xi^N))$ in the lemma. ■

Theorem 4.2 is a significant step forward from Theorem 4.1 in that it gives a more explicit relationship between $\mathsf{dl}_{K,1}\left(P^{\otimes N} \circ \hat{\vartheta}_N^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}\right)$ and $\mathsf{dl}_{K,k}(P,Q)$ when P and Q are restricted to $\mathcal{M}_k^\phi$. This comes out with additional condition on the uniform Lipschitz continuity of $u(f(x,\xi))$ in ξ. It might be interesting to weaken the condition by adopting other metrics such as Fortet-Mourier metric, we leave this for new research as it requires more sophisticated technical treatment. Note also that it is possible to use some standard results in parametric programming [33, 40] to derive quantitative statistical robustness of the optimal solutions under some more conditions such as second order growth conditions, we leave this for interested readers to explore because from theoretical point of view such results do no have any additional novelty and indeed these conditions may incur further questions for justifications.

4.3 Special cases

In a particular case when X is a singleton $\{x_0\}$, the PRO model (1.2) collapses to problem (1.3) for $x = x_0$. In principal, Theorem 4.1 also covers the statistical robustness of the estimator of $v(x_0, P) = \varrho(P \circ f(x_0, \cdot)^{-1})$. We describe this in the next corollary.

Corollary 4.1 (Statistical robustness of worst case utility problem (1.3)) *Let $X = \{x_0\}$ and ϕ be defined as in (3.26). Assume: (a) Assumptions 3.1 and 3.2 are fulfilled, (b) condition (b) of Proposition 3.1 is satisfied. Let $\mathcal{M} \subset \mathcal{M}_{k,\kappa}^{\phi^p}$ be a subset for some fixed $p > 1$ and $\kappa > 0$, The following assertions hold.*

(i) For any $P \in \mathcal{M}$ and $\epsilon > 0$, there exist positive numbers $\delta > 0$ and $N_0 \in \mathbb{N}$ such that

$$Q \in \mathcal{M}, \mathsf{dl}_\phi(P, Q) \le \delta \Longrightarrow \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ (\hat{v}_{x_0}^N)^{-1}, Q^{\otimes N} \circ (\hat{v}_{x_0}^N)^{-1}\right) \le \epsilon$$

for all $N \ge N_0$, where $\hat{v}_{x_0}^N(\xi^1, \cdots, \xi^N) := v(x_0, P_N)$.

(ii) If, in addition, $u(f(x_0, \xi))$ is uniformly Lipschitz continuous in ξ, i.e., there exists a positive constant $L > 0$ such that

$$|u(f(x_0, \xi)) - u(f(x_0, \xi'))| \le L\|\xi - \xi'\|, \forall u \in \mathcal{U},$$

then for any $N \in \mathbb{N}$ and any $P, Q \in \mathcal{M}_k^\phi$

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

Another important case is that $\mathcal{U}$ is a singleton with $u(t) = t$ for $t \in \mathbb{R}$. In this case, the (PRO) model collapses to an ordinary stochastic programming problem

$$\text{(SP)} \quad \max_{x \in X} \mathbb{E}_P[f(x, \xi)]. \tag{4.47}$$

The next corollary describes statistical robustness of the optimal value and optimal solution of (SP).

Corollary 4.2 (Statistical robustness of (SP)) *Let Assumption 3.1 hold and $\phi(t) = C(\|t\|^\gamma + 1)$. Let $\mathcal{M} \subset \mathcal{M}_{k,\kappa}^{\phi^p}$ be a subset for some fixed $p > 1$ and $\kappa > 0$. Then the following assertions hold.*

(i) For any $P \in \mathcal{P}$ and $\epsilon > 0$, there exist positive numbers $\delta > 0$ and $N_0 \in \mathbb{N}$ such that

$$Q \in \mathcal{M}, \mathsf{dl}_\phi(P, Q) \le \delta \Longrightarrow \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{\vartheta}_N^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}\right) \le \epsilon$$

for all $N \ge N_0$. If for each $Q \in \mathcal{M}$, the set of optimal solutions $S(Q_N)$ and $S(Q)$ are singletons, then for any $P \in \mathcal{P}$ and $\epsilon > 0$, there exist positive numbers $\delta > 0$ and $N_0 \in \mathbb{N}$ such that for all $N \ge N_0$

$$Q \in \mathcal{M}, \mathsf{dl}_\phi(P, Q) \le \delta \Longrightarrow \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{S}_N^{-1}, Q^{\otimes N} \circ \hat{S}_N^{-1}\right) \le \epsilon.$$

(ii) If, in addition, $f(x,\xi)$ is uniformly Lipschitz continuous in ξ, i.e., there exists a positive constant $L > 0$ such that

$$|f(x,\xi) - f(x,\xi')| \leq L\|\xi - \xi'\|, \forall x \in X,$$

then for any $N \in \mathbb{N}$ and any $P, Q \in \mathcal{M}_k^\phi$

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

Proof. Since $|f(x,t)| \leq C(\|t\|^\gamma + 1) = \phi(t)$ under Assumption 3.1, then the result follows from Theorems 4.1 and 4.2. ■

Note that Part (i) of the corollary can be derived straightforwardly from [7, Corollary 2.4] and Hampel's theorem [32, Theorem 2.4].

5 Applications

In this section, we take the opportunity to look into statistical robustness of a simple machine learning model due to its structural similarity to the (PRO) model, and we also include an application of statistically robust analysis in robust utility-based shortfall risk measure and associated optimization problem recently considered by Delage et al. [9].

5.1 Machine learning

Let $\Theta \subset \mathbb{R}^n$ be the input space and $\Xi \subset \mathbb{R}$ the output space. The relation between an input $\theta \in \Theta$ and an output $\xi \in \Xi$ is described by a probability distribution $P(\theta,\xi)$. Let $Z \subset \mathbb{R}^n \times \mathbb{R}$ denote the product space $\Theta \times \Xi$. For each input $\theta \in \Theta$, output $\xi \in \Xi$ and $z = (\theta,\xi)$, let $c(z, f(\theta))$ denote the loss caused by the use of f as a model for the unknown process producing ξ from θ. If P is known, then the problem is down to solve

$$\vartheta(P) = \min_{f \in \mathcal{F}} \mathbb{E}_P[c(z, f(\theta))] := \int_Z c(z, f(\theta))P(dz), \tag{5.48}$$

where $\mathcal{F}$ denotes the set of feasible models.

In practice, the true probability distribution P is unknown but it is possible to obtain iid samples $z^i = (\theta^i, \xi^i)$ for $i = 1, \cdots, N$ generated by P, which is known as training data. Given the sample $\{z^i\}_{i=1}^N$, the goal of machine learning is to find a function $f_z : \Theta \to \Xi$ such that f_z solves

$$\vartheta(P_N) = \min_{f \in \mathcal{F}} \mathbb{E}_{P_N}[c(z, f(\theta))] := \frac{1}{N}\sum_{i=1}^N c(z^i, f(\theta^i)), \tag{5.49}$$

where $P_N := \frac{1}{N}\sum_{i=1}^N \delta_{z^i}$ denotes the empirical probability distribution. For each $N \in \mathbb{N}$, denote by $\hat{\vartheta}(z_1, \cdots, z_N) := \vartheta(P_N)$. Let $f_N(P_N)$ denote an optimal solution of the sample average approximation problem (5.49). Then $f_N(P_N)$ is called an estimator and the framework generating f_N is called a learning algorithm.

The nature of functions f and f_N in (5.48) and (5.49) need to be specified. Let $\mathcal{H}(\Theta)$ denote a class of functions $f : \Theta \to \Xi$. $\mathcal{H}(\Theta)$ is called hypotheses space if $\mathcal{F}$ is restricted to $\mathcal{H}(\Theta)$. This is because the choice of $\mathcal{H}(\Theta)$ is based on hypotheses of the structure of these functions.

Let $\mathcal{H}(\Theta)$ be a Hilbert space of functions with inner product $\langle \cdot, \cdot \rangle$ and $K(\cdot, \cdot) : \Theta \times \Theta \to \mathbb{R}$ be a kernel. $\mathcal{H}(\Theta)$ is said to be a reproducing kernel Hilbert space (RKHS for short) if there is a kernel function $K(\cdot, \cdot) : \Theta \to \mathbb{R}$ such that: (a) $K(\cdot, \theta) \in \mathcal{H}(\Theta)$ for all $\theta \in \Theta$ and (b) $f(\theta) = \langle f, K(\cdot, \theta) \rangle$ for all $f \in \mathcal{H}(\Theta)$ and $\theta \in \Theta$.

The goal of learning theory is to find an approximation f_N of an unknown target function $f^* : \Theta \to \Xi$ through a set of samples $\{(\theta^i, \xi^i)\}_{i=1}^N$ generated by the unknown true probability distribution P on $\Theta \times \Xi$. In practice, the perceived sample data may be contaminated which means that they are not real data generated by P, rather they are generated by some probability distribution Q which is a perturbation of P. This motivates us to investigate statistical robustness of model (5.49). To faciliate the discussion, we focus on the least square model, i.e., $c(z, f(\theta)) = |\xi - f(\theta)|^2$. It remains to explain the connection between the machine learning model (5.48) and the (PRO) model that we consider in (1.2). To this end, let us fix x in the PRO model (1.2). Then the maximin problem reduces to

$$\vartheta(P) := \inf_{u \in \mathcal{U}} \mathbb{E}_P[u(f(x_0, \xi))].$$

Let $g(u, \xi) := u(f(x_0, \xi))$. Then we can recast the problem above as

$$\vartheta(P) := \inf_{u \in \mathcal{U}} \mathbb{E}_P[g(u, \xi)]. \tag{5.50}$$

It is evident that (5.50) is in a similar form to the machine learning problem (5.48). In the rest of the section, we derive conditions under which $\vartheta(P_N)$ is statistically robust. To this effect, we first establish continuity of $\vartheta(\cdot)$ near P.

Proposition 5.1 *Assume: (a) there is a gauge function $\phi(z)$ such that*

$$|\xi - f(\theta)|^2 \leq \phi(z), \forall z \in Z \text{ and } f \in \mathcal{F}, \tag{5.51}$$

where $\phi(z) \to \infty$ as $\|z\| \to \infty$; (b) there is a positive constant β such that $\|f\|_K \leq \beta$ for all $f \in \mathcal{F}$, where $\|\cdot\|_K$ is a norm in $\mathcal{H}(\Theta)$ induced by the inner product; (c) $K(\cdot, \theta)$ is equicontinuous, that is, for any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\|K(\cdot, \theta') - K(\cdot, \theta)\|_K \leq \epsilon, \forall \|\theta' - \theta\| \leq \delta,$$

where $\|\cdot\|$ is the Euclidean norm; (d) Z is a convex set. Then

$$\lim_{P' \xrightarrow{\phi} P} \vartheta(P') = \vartheta(P). \tag{5.52}$$

Proof. Let $\mathscr{G} := \{g : g(z) := |\xi - f(\theta)|^2 \text{ for } f \in \mathcal{F}\}$ and

$$\mathscr{G}_M := \{g_M : g_M(z) := \min\{g(z), M\} \text{ for } g \in \mathscr{G}\}.$$

We first prove that $\mathscr{G}_M$ is equi-continuous over Z for any fixed constant $M > 0$. By the reproducing property of the kernel $K(\cdot,\cdot)$, i.e., $f(\theta) = \langle f, K(\cdot,\theta)\rangle$, we have

$$\begin{aligned} |f(\theta') - f(\theta)| &= |\langle f, K(\cdot,\theta')\rangle - \langle f, K(\cdot,\theta)\rangle| \leq \|f\|_K \|K(\cdot,\theta') - K(\cdot,\theta)\|_K \\ &\leq \beta \|K(\cdot,\theta') - K(\cdot,\theta)\|_K. \end{aligned} \tag{5.53}$$

Let $z, z' \in Z$ be such that $g_M(z) = g(z)$ and $g_M(z') = g(z')$. Then

$$\begin{aligned} |g_M(z) - g_M(z')| &= |g(z) - g(z')| \\ &\leq |\xi' - f(\theta') + \xi - f(\theta)|(|\xi' - \xi| + |f(\theta') - f(\theta)|) \\ &\leq 2\sqrt{M}(|\xi' - \xi| + \beta\|K(\cdot,\theta') - K(\cdot,\theta)\|_K). \end{aligned} \tag{5.54}$$

In the case that $g(z)$ goes above M at one of the two points, i.e., at z', we can find a new point $\hat{z}$ on the line segment connecting z, z' such that $\hat{z} \in Z$ (due to convex structure of Z), $\|z - \hat{z}\| \leq \|z - z'\|$ and $g_M(\hat{z}) = g(\hat{z})$. Consequently, we have from (5.54)

$$\begin{aligned} |g_M(z) - g_M(z')| &= |g(z) - g(\hat{z})| \\ &\leq 2\sqrt{M}(|\hat{\xi} - \xi| + \beta\|K(\cdot,\hat{\theta}) - K(\cdot,\theta)\|_K). \end{aligned} \tag{5.55}$$

When $g(z)$ goes above M at both z and z', $|g_M(z) - g_M(z')| = 0$. Combining (5.53)-(5.55), we can see that the equi-continuity of $K(\cdot,\theta)$ implies the equi-continuity of $\mathscr{G}_M$ over Z for each fixed M.

Next, we show (5.52). The proof is similar to that of Proposition 3.1. We include a sketch.

Since $(\mathcal{M}^\phi_{n+1}, \tau_\phi)$ is a Polish space, then it suffices to show (5.52) for any sequence $\{P_l\} \subset \mathcal{M}^\phi_{n+1} := \left\{P \in \mathscr{P}(\mathbb{R}^{n+1}) : \int_{\mathbb{R}^{n+1}} \phi(z)P(dz) < \infty\right\}$ such that $P_l \xrightarrow{\phi} P \in \mathcal{M}^\phi_{n+1}$. Since

$$\begin{aligned} |\vartheta(P_l) - \vartheta(P)| &\leq \sup_{f\in\mathcal{F}} \left|\int_Z |\xi - f(\theta)|^2 P_l(dz) - \int_Z |\xi - f(\theta)|^2 P(dz)\right| \\ &= \sup_{g\in\mathscr{G}} \left|\int_Z g(z)P_l(dz) - \int_Z g(z)P(dz)\right|, \end{aligned}$$

we only need to show that

$$\lim_{l\to\infty} \sup_{g\in\mathscr{G}} \left|\int_Z g(z)P_l(dz) - \int_Z g(z)P(dz)\right| = 0. \tag{5.56}$$

Since $P_l \xrightarrow{\phi} P$, then $P_l \xrightarrow{w} P$ and $\lim_{l\to\infty} \int_Z \phi(z)P_l(dz) = \int_Z \phi(z)P(dz)$. Moreover, it follows by [6, Lemma 2.61],

$$\lim_{M\to\infty} \int_Z \phi(z)\mathbb{1}_{(M,\infty)}(\phi(z))P(dz) = 0, \tag{5.57}$$

$$\lim_{M\to\infty} \sup_{l\in\mathbb{N}} \int_Z \phi(z)\mathbb{1}_{(M,\infty)}(\phi(z))P_l(dz) = 0. \tag{5.58}$$

Under the growth condition (5.51), (5.57) and (5.58) imply

$$\begin{aligned} & \lim_{M\to\infty} \int_Z g(z)\mathbb{1}_{(M,\infty)}(g(z))P(dz) \\ \leq \quad & \lim_{M\to\infty} \int_Z \phi(z)\mathbb{1}_{(M,\infty)}(\phi(z))P(dz) = 0 \end{aligned} \tag{5.59}$$

and

$$\begin{aligned} & \lim_{M\to\infty} \sup_{l\in\mathbb{N}} \int_Z g(z)\mathbb{1}_{(M,\infty)}(g(z))P_l(dz) \\ \leq \quad & \lim_{M\to\infty} \sup_{l\in\mathbb{N}} \int_Z \phi(z)\mathbb{1}_{(M,\infty)}(\phi(z))P_l(dz) = 0. \end{aligned} \tag{5.60}$$

On the other hand, since $\mathscr{G}_M$ is equi-continuous over Z and (5.51) holds, then by [39, Theorem 3.2],

$$\lim_{l\to\infty} \sup_{g\in\mathscr{G}} \left| \int_Z g_M(z)P_l(dz) - \int_Z g_M(z)P(dz) \right| = 0. \tag{5.61}$$

Combing (5.59)-(5.61), we obtain (5.56) as desired. The proof is complete. ■

Note that Proposition 5.1 is parallel to Theorem 3.1. We provide a separate proof for two reasons: (a) the settings are different, here problem (5.48) is no longer a maximin problem; (b) $\mathcal{F}$ is in general not a compact set and hence we cannot apply the Berge's maximum theorem as in the proof of Theorem 3.1.

With Proposition 5.1, we are ready to address statistical robustness of the machine learning model.

Corollary 5.1 (Statistical robustness) *Let*

$$\mathcal{M}_{n+1,\kappa}^{\phi^p} := \left\{ P \in \mathscr{P}(\mathbb{R}^{n+1}) : \int_{\mathbb{R}^{n+1}} \phi(z)^p P(dz) \leq \kappa \right\}$$

for some fixed positive numbers $p > 1$ and $\kappa > 0$, where $\phi(z)$ is defined as in (5.51). Suppose conditions of Proposition 5.1 hold. Then for any $P \in \mathcal{M}_{n+1,\kappa}^{\phi^p}$ and $\epsilon > 0$, there exist positive numbers $\delta > 0$ and $N_0 \in \mathbb{N}$ such that when $Q \in \mathcal{M}_{n+1,\kappa}^{\phi^p}$ with $\mathsf{dl}_\phi(P,Q) \leq \delta$, $\mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{\vartheta}_N^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}\right) \leq \epsilon$ for all $N \geq N_0$.

We omit the proof as it is in the same spirit of Corollary 4.1.

5.2 Utility-based shortfall risk optimization

Delage et al. [9] consider a robust version of the well-known utility-based shortfall risk measure to address a situation where the decision maker is ambiguous about their risk preference, that is,

$$\text{(PRSR)} \quad \text{SR}_L^P(Z) := \inf \left\{ t : \sup_{l\in L} \mathbb{E}_P[l(-Z-t)] - l(0) \leq 0 \right\},$$

where Z represents a financial position, and L is an ambiguity set of loss functions $l : \mathbb{R} \to \mathbb{R}$ which are convex and non-decreasing. They apply the proposed robust risk measure to an optimal decision making problem

$$\text{(PRSR-Opt)} \qquad \begin{array}{rll} \vartheta(P) := \min\limits_{x \in X} & \inf\limits_{t \in \mathbb{R}} & t \\ & \text{s.t.} & \sup\limits_{l \in L} \mathbb{E}_P[l(-c(x,\xi) - t)] - l(0) \le 0, \end{array} \tag{5.62}$$

where $c(x,\xi) : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ is a continuous function representing a financial gain, ξ is a random vector mapping from $(\Omega, \mathcal{F}, \mathbb{P})$ to $\Xi \subseteq \mathbb{R}^k$ and $P := \mathbb{P} \circ \xi^{-1}$ is the probability measure on $\mathbb{R}^k$ induced by ξ. Here we investigate statistical robustness of (PRSR-Opt). To this end, we write the inner minimization problem equivalently as

$$\text{(PRSR-inner)} \quad \inf_t \sup_{\lambda \ge 0} \left\{ t + \lambda \left[\sup_{l \in L} \mathbb{E}_P[l(-c(x,\xi) - t)] - l(0) \right] \right\}. \tag{5.63}$$

The Lagrangian dual of (5.63) is

$$\text{(PRSR-inner-Dual)} \quad \sup_{\lambda \ge 0} \inf_t \left\{ t + \lambda \left[\sup_{l \in L} \mathbb{E}_P[l(-c(x,\xi) - t)] - l(0) \right] \right\}. \tag{5.64}$$

Assume that $c(x,\xi)$ is concave in x for almost every ξ and the inequality in (5.62) satisfies the well-known Slater condition, i.e., for each fixed $x \in X$, there exists t_0 such that

$$\sup_{l \in L} \mathbb{E}_P[l(-c(x,\xi) - t_0)] - l(0) < 0.$$

Then (PRSR-inner) and (PRSR-inner-Dual) are equivalent, and the set of the Lagrange multipliers at the optimum is bounded. The latter means that we can restrict λ to take values from an interval $[0, K]$ for some sufficiently large $K > 0$. On the other hand, since (PRSR-inner) can be written as

$$\inf_t \sup_{\lambda \ge 0, l \in L} \{ t + \lambda [\mathbb{E}_P[l(-c(x,\xi) - t)] - l(0)] \},$$

we can then reformulate (PRSR-Opt) as

$$\min_{x \in X, t} \sup_{\lambda \ge 0, l \in L} \{ t + \lambda [\mathbb{E}_P[l(-c(x,\xi) - t)] - l(0)] \}. \tag{5.65}$$

Moreover, under the Slater condition, the optimum in t is finite-valued and therefore we can restrict t to a compact interval $[-T, T]$ for some sufficiently large $T > 0$, see [9, Section 5] for the argument. Note that x is restricted to a compact set X by assumption. Let

$$\rho(x, t, P) := \sup_{\lambda \in [0,K], l \in L} \{ t + \lambda [\mathbb{E}_P[l(-c(x,\xi) - t)] - l(0)] \}. \tag{5.66}$$

Based on the above above, we can reformulate (5.65) as

$$\vartheta(P) := \min_{x \in X, t \in [-T,T]} \rho(x, t, P). \tag{5.67}$$

In what follows, we establish the continuity of ϑ under some topology. To this effect, we make the following assumptions.

Assumption 5.1 *Problem (5.65) satisfies the following conditions: (a) There are an exponent $\gamma > 0$ and a continuous function $\eta : \mathbb{R}^n \to \mathbb{R}$ such that*

$$|c(x, z)| \le \eta(x)(\|z\|^\gamma + 1), \forall (x, z) \in \mathbb{R}^n \times \mathbb{R}^k. \tag{5.68}$$

(b) For any constant $M > 0$,

$$L_{[-M,M]} := \left\{ l_{[-M,M]}(s) := \max\{\min\{l(s), M\}, -M\right\} : s \in \mathbb{R}, l \in L\} \tag{5.69}$$

is equi-continuous over $\mathbb{R}$. (c) Let $\psi(s) := \sup_{l \in L} |l(s)|$. The set L is chosen so that $\psi(s)$ is a gauge function.

Note that condition (a) in Assumption 5.1 corresponds to the growth condition in Assumption 3.1 and condition (b) is similar to equi-continuity condition (b) in Proposition 3.1 which is explicitly considered in Delage et al. [9, Section 5.3], condition (c) corresponds to the existence of gauge function in Assumption 3.2. In [9], Delage et al. consider a class of loss functions L where each $l \in L$ is assumed to be non-constant increasing convex function with $l(-1) = -1$, obviously such loss functions fulfil condition (c).

Proposition 5.2 *Under Assumption 5.1, $\vartheta(\cdot)$ is continuous over $\mathcal{M}_k^\phi$, i.e.,*

$$\lim_{P' \xrightarrow{\phi} P} \vartheta(P') = \vartheta(P), \tag{5.70}$$

where $\phi(z) := \max\left(\sup_{l \in L} l((C + T)(\|z\|^\gamma + 1)), \sup_{l \in L} -l(-(C + T)(\|z\|^\gamma + 1))\right)$ for some positive constants C and T.

The thrust of the proof is to show that $\rho(x, t, P)$ is continuous in (x, t, P) which is down to the continuity of $\sup_{l \in L} \mathbb{E}_P[l(-c(x, \xi) - t)]$. The latter can be proved by writing the term as $\varrho(P \circ f(x, t, \cdot)^{-1})$ where $f(x, t, \xi) := -c(x, \xi) - t$ and

$$\varrho(\sigma) := \sup_{l \in L} \int_{\mathbb{R}} l(s)\sigma(ds), \forall \sigma \in \mathscr{P}(\mathbb{R}).$$

The rest follows from a similar analysis to the proof of Theorem 3.1, we omit these details.

Following a similar analysis to that of Theorem 4.1, we can establish the statistical robustness of the estimator $\hat{\vartheta}_N(\xi^1, \cdots, \xi^N)$ of $\vartheta(P_N)$ defined as in (5.67).

Corollary 5.2 *Let $\phi(t)$ defined as in Proposition 5.2 be a gauge function and*

$$\mathcal{M}_{k,\kappa}^{\phi^p} := \left\{ P \in \mathscr{P}(\mathbb{R}^k) : \int_{\mathbb{R}^k} \phi(t)^p P(dt) \le \kappa \right\}$$

for some fixed $p > 1$ and $\kappa > 0$, let $\mathcal{M} \subset \mathcal{M}_{k,\kappa}^{\phi^p}$ be a subset. Under Assumption 5.1, the following assertions hold.

(i) *For any* $P \in \mathcal{M}$ *and* $\epsilon > 0$*, there exist positive numbers* $\delta > 0$ *and* $N_0 \in \mathbb{N}$ *such that*

$$Q \in \mathcal{M}, \mathsf{dl}_\phi(P, Q) \le \delta \Longrightarrow \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{\vartheta}_N^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}\right) \le \epsilon$$

for all $N \ge N_0$.

(ii) *If, in addition, for each* $Q \in \mathcal{M}$*,* $S(Q_N)$ *and* $S(Q)$ *are singletons, then for any* $P \in \mathcal{M}$ *and* $\epsilon > 0$*, there exist positive numbers* $\delta > 0$ *and* $N_0 \in \mathbb{N}$ *such that for all* $N \ge N_0$

$$Q \in \mathcal{M}, \mathsf{dl}_\phi(P, Q) \le \delta \Longrightarrow \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{S}_N^{-1}, Q^{\otimes N} \circ \hat{S}_N^{-1}\right) \le \epsilon,$$

where $S(Q)$ *is the optimal value of (PRSR-Opt) when* P *is replaced by* Q *and* $\hat{S}_N(\tilde{\xi}^1, \cdots, \tilde{\xi}^N)$ *for* $S(Q_N)$ *.*

Note that following an analysis analogous to Section 4.2, in the case when $l(-c(x,\xi) - t)$ is uniformly Lipschitz continuous in ξ, we can derive the quantitative statistical robustness of the estimator, we leave this for readers as an exercise.

6 Uniform consistency

In Section 4, we introduce the UGC property on a set of probability measures in the metric space $\mathscr{P}(\mathbb{R}^k)$ and use it to derive statistical robustness of $\vartheta(P)$ and associated optimal solutions. In this section, we investigate the uniform consistency of $\varrho(\mu(x, P_N))$ to $\varrho(\mu(x, P))$ for all P in a subset of $\mathscr{P}(\mathbb{R}^k)$, and its uniformity w.r.t. x, the latter leads to uniform consistency of $\vartheta(P_N)$ to $\vartheta(P)$. The convergence result is important because in the literature of PRO models, all tractable formulations rely heavily on discrete approximation of P by P_N when P is continuously distributed and convergence of such approximation has not been explicitly covered by the results in the preceding sections.

To establish the uniform consistency results, we need the following technical assumption.

Assumption 6.1 *Let* $\mathcal{G} := \{g(\cdot) := u(f(x, \cdot)) : \mathbb{R}^k \to \mathbb{R} | x \in X, u \in \mathcal{U}\}$ *and*

$$\mathcal{G}_{[-M,M]} := \{g_{[-M,M]}(t) := u_{[-M,M]}(f(x,t)) | x \in X, u_{[-M,M]} \in \mathcal{U}_{[-M,M]}\}$$

where $\mathcal{U}_{[-M,M]}$ *is defined as in (3.19) for any fixed constant* $M > 0$*. For any sequence* $\{P_l\} \subset \mathscr{P}(\mathbb{R}^k)$ *with* $P_l \xrightarrow{w} P$*,*

$$\lim_{l \to \infty} \sup_{g_{[-M,M]} \in \mathcal{G}_{[-M,M]}} \left| \int_{\mathbb{R}^k} g_{[-M,M]}(t) P_l(dt) - \int_{\mathbb{R}^k} g_{[-M,M]}(t) P(dt) \right| = 0. \tag{6.71}$$

The following proposition states some sufficient conditions for the assumption.

Proposition 6.1 *Assume: (a)* $\mathcal{U}_{[-M,M]}$ *is equi-continuous over* $\mathbb{R}$ *for any* $M > 0$*; (b)* $f(x,t)$ *is continuous in* t *uniformly w.r.t.* $x \in X$*. Then Assumption 6.1 is satisfied.*

Proof. Under conditions (a)-(b), we can show that $\mathcal{G}_{[-M,M]}$ is equi-continuous on any $\mathbb{R}^k$, that is, for any $\epsilon > 0$, there exists a $\delta > 0$ such that for every $g_{[-M,M]} \in \mathcal{G}_{[-M,M]}$, $|g_{[-M,M]}(t') - g_{[-M,M]}(t'')| < \epsilon$ for all $t', t'' \in \mathbb{R}^k$ with $\|t' - t''\| < \delta$. Moreover $|g_{[-M,M]}(t)| \leq M$ for all $t \in \mathbb{R}$. By [39, Theorem 3.2], (6.71) holds for any sequence $\{P_l\} \subset \mathscr{P}(\mathbb{R}^k)$ with $P_l \xrightarrow{w} P$. ■

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

Theorem 6.1 (Uniform consistency) *Let $\phi(t)$ be a continuous gauge function defined as in (3.26) and $\mathcal{M}^{\phi^p}_{k,\kappa}$ be defined as in (4.37) for some fixed $p > 1$ and $\kappa > 0$, let $\mathcal{M}$ be a compact subset of $\mathcal{M}^{\phi^p}_{k,\kappa}$. Under Assumptions 3.1 and 6.1, the following assertions hold.*

(i) For each $P \subset \mathcal{M}^{\phi}_k$,

$$\lim_{P' \xrightarrow{\phi} P} \sup_{g \in \mathcal{G}} \left| \int_{\mathbb{R}^k} g(t) P'(dt) - \int_{\mathbb{R}^k} g(t) P(dt) \right| = 0, \tag{6.72}$$

and moreover

$$\lim_{P' \xrightarrow{\phi} P} \sup_{x \in X} \left| \varrho(\mu(x, P')) - \varrho(\mu(x, P)) \right| = 0. \tag{6.73}$$

(ii) For every $\epsilon > 0$ and $\delta > 0$, there exists $N_0 \in \mathbb{N}$ such that

$$P^{\otimes N} \left[(\xi^1, \cdots, \xi^N) : \sup_{x \in X} |\varrho(\mu(x, P_N)) - \varrho(\mu(x, P))| \geq \delta \right] \leq \epsilon, \forall P \in \mathcal{M} \tag{6.74}$$

and hence

$$P^{\otimes N} \left[(\xi^1, \cdots, \xi^N) : |\vartheta(P_N) - \vartheta(P)| \geq \delta \right] \leq \epsilon, \forall P \in \mathcal{M} \tag{6.75}$$

for all $N \geq N_0$.

(iii) If, in addition, $\varrho(\mu(x, P))$ satisfies the second order growth condition at $S(P)$, i.e., there exists a positive constant β such that

$$\max_{x \in X} \varrho(\mu(x, P)) \geq \varrho(\mu(x, P)) + \beta d(x, S(P))^2, \forall x \in X, P \in \mathcal{M}, \tag{6.76}$$

then

$$P^{\otimes N} \left[(\xi^1, \cdots, \xi^N) : \mathbb{D}(S(P_N), S(P)) \geq \sqrt{3\delta/\beta} \right] \leq \epsilon, \forall P \in \mathcal{M} \tag{6.77}$$

for all $N \geq N_0$, where $S(P)$ is defined as in (2.12) with P, $d(x, S) := \min_{x' \in S} \|x - x'\|$, and $\mathbb{D}(S', S) := \sup_{x' \in S'} d(x', S)$.

Proof. Part (i). Since

$$\sup_{x \in X} \left| \varrho(\mu(x, P')) - \varrho(\mu(x, P)) \right| \leq \sup_{g \in \mathcal{G}} \left| \int_{\mathbb{R}^k} g(t) P'(dt) - \int_{\mathbb{R}^k} g(t) P(dt) \right|,$$

we only need to prove (6.72). Since $(\mathcal{M}_\kappa^\phi, \mathsf{dl}_\phi)$ is a Polish space, thus it suffices to show that for any sequence $\{P_l\} \subset \mathcal{M}_k^\phi$ with $P_l \xrightarrow{\phi} P \in \mathcal{M}_k^\phi$,

$$\lim_{l\to\infty} \sup_{g\in\mathcal{G}} \left| \int_{\mathbb{R}^k} g(t)P_l(dt) - \int_{\mathbb{R}^k} g(t)P(dt) \right| = 0. \tag{6.78}$$

From the definition of the set $\mathcal{G}$ and the gauge function ϕ,

$$g(t) \le \phi(t), \forall t \in \mathbb{R}^k, g \in \mathcal{G}.$$

Then by $P_l \xrightarrow{\phi} P \in \mathcal{M}_k^\phi$ and [6, Lemma 2.61], for any $\epsilon > 0$ we may set M large enough so that

$$\sup_{g\in\mathcal{G}} \int_{\mathbb{R}^k} |g(t)| \mathbb{1}_{(M,\infty)}(|g(t)|)P(dt) \le \int_{\mathbb{R}^k} \phi(t)\mathbb{1}_{(M,\infty)}(\phi(t))P(dt) \le \epsilon, \tag{6.79}$$

and

$$\begin{aligned} \sup_{g\in\mathcal{G}} \sup_{l\in\mathbb{N}} \int_{\mathbb{R}^k} |g(t)| \mathbb{1}_{(M,\infty)}(|g(t)|)P_l(dt) &\le \sup_{l\in\mathbb{N}} \int_{\mathbb{R}^k} \phi(t)\mathbb{1}_{(M,\infty)}(\phi(t))P_l(dt) \\ &\le \epsilon. \end{aligned} \tag{6.80}$$

Moreover, by Assumption 6.1,

$$\sup_{g_{[-M,M]}\in\mathcal{G}_{[-M,M]}} \left| \int_{\mathbb{R}^k} g_{[-M,M]}(t)P_l(dt) - \int_{\mathbb{R}^k} g_{[-M,M]}(t)P(dt) \right| \le \epsilon \tag{6.81}$$

for l sufficiently large. Combining (6.79), (6.80) and (6.81), we obtain

$$\begin{aligned} & \sup_{g\in\mathcal{G}} \left| \int_{\mathbb{R}^k} g(t)P_l(dt) - \int_{\mathbb{R}^k} g(t)P(dt) \right| \\ \le\ & \sup_{g\in\mathcal{G}} \left| \int_{\mathbb{R}^k} g(t)P_l(dt) - \int_{\mathbb{R}^k} g_{[-M,M]}(t)P_l(dt) \right| \\ & + \sup_{g\in\mathcal{G}} \left| \int_{\mathbb{R}^k} g_{[-M,M]}(t)P_l(dt) - \int_{\mathbb{R}^k} g_{[-M,M]}(t)P(dt) \right| \\ & + \sup_{g\in\mathcal{G}} \left| \int_{\mathbb{R}^k} g_{[-M,M]}(t)P(dt) - \int_{\mathbb{R}^k} g(t)P(dt) \right| \le 3\epsilon. \end{aligned} \tag{6.82}$$

Hence the continuity result (6.78) holds as ϵ can be arbitrarily small.

Part (ii). We only prove inequality (6.74) as inequality (6.75) follows straightforwardly from the former. For fixed $\bar{P}$, by the continuity result (6.73) in Part (i), for any $\delta > 0$, there exists a positive constant $\eta_1 > 0$ such that

$$\sup_{x\in X} \left| \varrho(\mu(x,Q)) - \varrho(\mu(x,\bar{P})) \right| < \delta, \tag{6.83}$$

for each Q satisfying $d_\phi(Q,\bar{P}) < \eta_1$. It follows by [32, Corollary 3.5] that $(\mathcal{M}_\kappa^{\phi^p}, \mathsf{dl}_\phi)$ has the UGC property for all $p > 1$ and $\kappa > 0$, then for any $\epsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that

$$P^{\otimes N}\left[\mathsf{dl}_\phi(P_N, P) \ge \eta_1\right] \le \epsilon, \forall P \in \mathcal{M}_\kappa^{\phi^p}$$

for all $N \geq N_0$. Thus, by (6.83) for any $\epsilon > 0$ and $\delta > 0$, there exists N_0 such that

$$\bar{P}^{\otimes N}\left[\sup_{x\in X}\left|\varrho(\mu(x,\bar{P}_N)) - \varrho(\mu(x,\bar{P}))\right| \geq \delta\right] \leq \bar{P}^{\otimes N}\left[\mathsf{dl}_\phi(\bar{P}_N,\bar{P}) \geq \eta_1\right] \leq \epsilon$$

for all $N \geq N_0$. Therefore, (6.74) holds when P is fixed at $\bar{P}$.

Now we show (6.74) holds for all $P \in \mathcal{M}$. Assume for the sake of a contradiction that there exist some positive numbers ϵ_0 and δ_0 such that for any $s \in \mathbb{N}$, there exist $s' > s$, $P_{s'} \in \mathcal{M}$ and some $N_{s'} \geq s$ such that

$$P_{s'}^{\otimes N_{s'}}\left[\sup_{x\in X}\left|\varrho(\mu(x,P_{N_{s'}})) - \varrho(\mu(x,P_{s'}))\right| \geq \delta_0\right] > \epsilon_0. \tag{6.84}$$

Let s increase. Then we obtain a sequence of $\{P_{s'}\}$ which satisfies (6.84). Since $\mathcal{M}$ is compact under the ϕ-weak topology, then $\{P_{s'}\}$ has a converging subsequence. Assume without loss of generality that $P_{s'} \xrightarrow{\phi} P_* \in \mathcal{M}$. Since (6.73) holds for P_*, then there exists $\eta_2 > 0$ such that

$$\sup_{x\in X}|\varrho(\mu(x,Q)) - \varrho(\mu(x,P_*))| < \delta_0/2 \tag{6.85}$$

for P satisfying $\mathsf{dl}_\phi(Q,P_*) < \eta_2$. Since $P_{s'} \xrightarrow{\phi} P_*$, there exists s_0' such that $\mathsf{dl}_\phi(P_{s'},P_*) < \eta_2/2$ for $s' \geq s_0'$, and then $\sup_{x\in X}|\varrho(\mu(x,P_{s'})) - \varrho(\mu(x,P_*))| < \delta_0/2$. On the other hand, by the UGC property

$$\begin{aligned} P_s^{\otimes N_s}(\mathsf{dl}_\phi(P_{N_{s'}},P_*) \geq \eta_2) &\leq P_{s'}^{\otimes N_{s'}}(\mathsf{dl}_\phi(P_{N_{s'}},P_{s'}) + \mathsf{dl}_\phi(P_{s'},P_*) \geq \eta_2) \\ &= P_{s'}^{\otimes N_{s'}}(\mathsf{dl}_\phi(P_{N_{s'}},P_{s'}) \geq \eta_2 - \mathsf{dl}_\phi(P_{s'},P_*)) \\ &\leq P_{s'}^{\otimes N_{s'}}(\mathsf{dl}_\phi(P_{N_{s'}},P_{s'}) \geq \eta_2/2) \leq \epsilon_0 \end{aligned}$$

for sufficiently large $N_{s'}$. Therefore, by (6.85)

$$P_{s'}^{\otimes N_{s'}}\left[\sup_{x\in X}\left|\varrho(\mu(x,P_{N_{s'}})) - \varrho(\mu(x,P_*))\right| \geq \delta_0/2\right] \leq \epsilon_0,$$

and then

$$\begin{aligned} & P_{s'}^{\otimes N_{s'}}\left[\sup_{x\in X}\left|\varrho(\mu(x,P_{N_{s'}})) - \varrho(\mu(x,P_{s'}))\right| \geq \delta_0\right] \\ \leq\ & P_{s'}^{\otimes N_{s'}}\left[\sup_{x\in X}\left|\varrho(\mu(x,P_{N_{s'}})) - \varrho(\mu(x,P_*))\right| + \sup_{x\in X}|\varrho(\mu(x,P_{s'})) - \varrho(\mu(x,P_*))| \geq \delta_0\right] \\ \leq\ & P_{s'}^{\otimes N_{s'}}\left[\sup_{x\in X}\left|\varrho(\mu(x,P_{N_{s'}})) - \varrho(\mu(x,P_*))\right| \geq \delta_0/2\right] \leq \epsilon_0, \end{aligned}$$

which leads to a contradiction with (6.84) as desired.

Part (iii). Under the second-order growth condition (6.76), it follows from [33, Lemma 3.8] that

$$D(S(P_N),S(P)) \leq \sqrt{\frac{3}{\beta}\sup_{x\in X}|\varrho(\mu(x,P_N)) - \varrho(\mu(x,P))|},$$

and hence (6.77) is deduced from (6.74). The proof is complete. ■

Theorem 6.1 states asymptotic consistency of $\varrho(\mu(x, P_N))$ to $\varrho(\mu(x, P))$ uniformly w.r.t. $x \in X$ and $P \in \mathcal{M}$ and subsequently asymptotic consistency of $\vartheta(P_N)$ to $\vartheta(P)$ uniformly w.r.t. $P \in \mathcal{M}$. The uniform consistency and the continuity of $\vartheta(\cdot)$ established in Theorem 3.1 imply statistical robustness. In other words, Theorem 4.1 can be derived from Theorems 3.1 and 6.1. However, the former can be established under weaker conditions and this is why it is presented in Section 4. As we commented at the beginning of this section, uniform consistency is of independent interest which goes beyond statistical robustness. The second order growth condition is widely used in parametric programming for deriving stability of optimal solutions, see [40] for a slightly more general form.

Remark 6.1 *In the case when the support of ξ is compact, $\mathcal{M}^{\phi^p}_{k,\kappa} = \mathscr{P}(\mathbb{R}^k)$. Consequently the uniform convergence in (6.75) holds for all $P \in \mathscr{P}(\mathbb{R}^k)$. Note also that the convergence results should be distinguished from similar convergence results in the literature of sample average approximation where a single P rather than a set $\mathcal{M}$ in (6.75) is considered, see [42, Chapter 7] and [9, Section 5] in the context of preference robust shortfall optimization. Moreover, our convergence is qualitative rather than quantitative in that Theorem 6.1 does not indicate the rate of convergence. It is possible to strengthen the convergence but this is beyond the main focus of this paper.*

Acknowledgments

The authors would like to thank Volker Krätschmer and Matthias Claus for valuable discussions on the subject of UGC property during preparation of the paper.

References

[1] E. J. Anderson, *Business Risk Management: Models and Analysis*, John Wiley & Sons, Chichester, 2014.

[2] B. Armbruster and E. Delage, Decision making under uncertainty when preference information is incomplete, *Management Science*, 61: 111–128, 2015.

[3] C. Berge, *Topological spaces: including a treatment of multi-valued functions, vector spaces, and convexity*, Courier Corporation, 1997.

[4] P. Billingsley, *Convergence of Probability Measures*, 2nd Edition, John Wiley & Sons, New York, 1999.

[5] C. Camerer, Individual Decision Making, in J. H. Kagel and A. Roth (eds.), *The Handbook of Experimental Economics*, Princeton University Press, Princeton, 587–703, 1995.

[6] M. Claus, *Advancing Stability Analysis of Mean-Risk Stochastic Programs: Bilevel and Two-Stage Models*, PhD Thesis, Universität Dusburg-Essen, 2016.

[7] M. Claus, V. Krätschmer and R. Schultz, Weak continuity of risk functionals with applications to stochastic programming, *SIAM Journal on Optimization*, 27: 91–109, 2017.

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

[9] E. Delage, S. Guo and H. Xu, Shortfall risk models when information of loss function is incomplete, working paper, 2017.

[10] E. Delage, D. Kuhn, W. Wiesemann, "Dice"-sion making under uncertainty: when can a random decision reduce risk? *Management Science*, 65: 3282–3301, 2019.

[11] E. Delage and J. Y. Li, Minimizing risk exposure when the choice of a risk measure is ambiguous, *Management Science*, 64: 327–344, 2018.

[12] P. Embrechts, A, Schied and R. Wang, Robustness in the optimization of risk measures, arXiv:1809.09268, 2018.

[13] P. H. Farquhar, Utility assessment methods, *Management Science*, 30: 1283–1300, 1984.

[14] P. Fishburn and G. Kochenberger, Two-piece von Neumann-Morgenstern utility functions, *Decision Sciences*, 10: 503–518, 1979.

[15] H. Föllmer and A. Schied, *Stochastic Finance: An Introduction in Discrete Time*, 3rd Edition, de Gruyter, Berlin, 2011.

[16] I. Gilboa, and D. Schmeidler, Maxmin expected utility with non-unique prior, *Journal of Mathematical Economics*, 18: 141–153, 1989.

[17] I. Gilboa and M. Marinacci, Ambiguity and the bayesian paradigm, In H. Arlò–CCosta, V. F. Hendricks, & J. van Benthem (Eds.), *Readings in Formal Epistemology*, Springer, 2016.

[18] S. Greco, V. Mousseau and R. Słowiński, Ordinal regression revisited: multiple criteria ranking using a set of additive vaulr functions, *European Journal of Operational Research*, 191: 415–435, 2008.

[19] S. Guo and H. Xu, Utility preference robust optimization: piecewise linear approximation, error bounds and stability, manuscript, 2018.

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

[21] W. Haskell, L. Fu and M. Dessouk, Ambiguity in risk preferences in robust stochastic optimization, *European Journal of Operational Research*, 254: 214–225, 2016.

[22] B. Haskell, W. Huang and H. Xu, Preference elicitation and robust optimization with quasi-concave choice functions in multi-attribute decision making, arXiv:1805.06632, 2018.

[23] J. Hu, M. Bansal and S. Mehrotra, Robust decision making using a general utility set, *European Journal of Operational Research*, 269: 699–714, 2018.

[24] J. Hu and S. Mehrotra, Robust decision making over a set of random targets or risk-averse utilities with an application to portfolio optimization, *IIE Transaction*, 47: 358–372, 2015.

[25] J. Hu and G. Stepanyan, Optimization with reference-based robust preference constraints, *SIAM Journal on Optimization*, 27: 2230–2257, 2017.

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

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

[28] D. Kahneman and A. Tversky, Prospect Theory: An analysis of decisions under risk, *Econometrica*, 47: 263–291, 1979.

[29] U. S. Karmarkar, Subjectively weighted utility: A descriptive extension of the expected utility model, *Organizational Behavior and Human Performance*, 21: 61–72, 1978.

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

[31] V. Krätschmer, A. Schied and H. Zähle, Domains of weak continuity of statistical functionals with a view toward robust statistics, *Journal of Multivariate Analysis*, 158: 1–19, 2017.

[32] V. Krätschmer, A. Schied and H. Zähle, Qualitative and infinitesimal robustness of tail-dependent statistical functionals, *Journal of Multivariate Analysis*, 103: 35–47, 2012.

[33] Y. Liu and H. Xu, Stability analysis of stochastic programs with second order dominance constraints, *Mathematical Programming*, 142: 435–460, 2013.

[34] F. Maccheroni, M. Marinacci and A. Rustichini, Ambiguity aversion, robustness, and the variational representation of preferences, *Econometrica*, 74: 1447–1498, 2006.

[35] M. Marinacci, Model uncertainty, *Journal of the European Economic Association*, 13: 1022–1100, 2015.

[36] A. Müller and D. Stoyan, *Comparison Methods for Stochastic Models and Risks*, John Wiley & Sons, 2002.

[37] Y. V. Prokhorov, Convergence of random processes and limit theorems in probability theory, *Theory of Probability & Its Applications*, 1: 157–214, 1956.

[38] M. Rabin, Psychology and economies, *Journal of Economic Literature*, 36: 11–46, 1998.

[39] R. Ranga Rao, Relations between weak and uniform convergence of measures with applications, *The Annals of Mathematical Statistics*, 33: 659-680, 1962.

[40] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.

[41] W. Römisch, Stability of stochastic programming problems. In: Ruszczy??nski, A., Shapiro, A. (eds.) Stochastic Programming, Handbooks in Operations Research and Management Science, volume 10, chapter 8. Elsevier, Amsterdam, 2003.

[42] A. Shapiro, D. Dentcheva and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, 2nd Edition, SIAM, Philadelphia, 2014.

[43] L. L. Thurstone, A law of comparative judgement, *Psychological Review*, 34: 278–286, 1927.

[44] K. Train, *Discrete choice methods with simulation*, New York: Cambridge University Press, 2009.

[45] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, 1947.

[46] M Weber, Decision making with incomplete information, *European Journal of Operational Research*, 28: 44–57, 1987.

[47] H. Zähle, A definition of statistical robustness for general point estimators, and examples, *Journal of Multivariate Analysis*, 143: 12–31, 2016.

Appendix

Theorem 6.2 (Berge's maximum theorem, Page 116 [3]) *Let ϕ be a continuous numerical function in a topological space $X \times Y$ and Γ be a continuous mapping of X into Y such that for each x, $\Gamma(x) \neq \emptyset$, then the numerical function M defined by $M(x) := \max\{\phi(x,y) : y \in \Gamma(x)\}$ is continuous in X and the mapping Φ defined by $\Phi(x) := \{y | y \in \Gamma(x), \phi(y) = M(x)\}$ is a upper semicontinuous mapping of X into Y.*

Proof of Theorem 4.1. Part (i). It follows from Theorem 3.1 (ii) that ϑ is continuous at P w.r.t. ϕ-weak topology, that is, $|\vartheta(P') - \vartheta(P)| \to 0$ as $\mathsf{dl}_\phi(P', P) \to 0$. Together with the UGC property of $(\mathcal{M}_{k,\kappa}^{\phi^p}, \mathsf{dl}_\phi)$ (see [32, Corallary 3.5]), we can obtain the conclusion by applying [32, Theorem 2.4]. Here we include a sketch of proof for completeness. By triangle inequality

$$\begin{aligned}\mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{\vartheta}_N^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}\right) &\leq \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{\vartheta}_N^{-1}, \delta_{\vartheta(P)}\right) \\ &\quad + \mathsf{dl}_{\text{Prok}}\left(\delta_{\vartheta(P)}, Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}\right),\end{aligned}$$

it suffice to show that for any small number $\epsilon > 0$ there exist positive numbers $\delta > 0$ and $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$

$$\mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{\vartheta}_N^{-1}, \delta_{\vartheta(P)}\right) \leq \frac{\epsilon}{2}, \mathsf{dl}_{\text{Prok}}\left(\delta_{\vartheta(P)}, Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}\right) \leq \frac{\epsilon}{2}$$

when $\mathsf{dl}_\phi(P, Q) \leq \delta$. By the Strassen's theorem [27, Theorem 2.13], the above two inequalities are guaranteed respectively by

$$P^{\otimes N}\left[\xi \in (\mathbb{R}^k)^N : |\vartheta(P) - \vartheta(P_N)| \leq \frac{\epsilon}{2}\right] \geq 1 - \frac{\epsilon}{2}, \tag{6.86}$$

and

$$Q^{\otimes N}\left[\tilde{\xi} \in (\mathbb{R}^k)^N : |\vartheta(P) - \vartheta(Q_N)| \leq \frac{\epsilon}{2}\right] \geq 1 - \frac{\epsilon}{2}, \tag{6.87}$$

where P_N and Q_N are defined in (2.9). Since ϑ is continuous at P by Theorem 3.1, for any $\epsilon > 0$, there exists a constant $\delta > 0$ such that

$$\mathsf{dl}_\phi(P, P') \le 2\delta \Longrightarrow |\vartheta(P) - \vartheta(P')| \le \epsilon/2 \tag{6.88}$$

By plugging P_N into the position of P', we obtain

$$P^{\otimes N}\left[\xi \in (\mathbb{R}^k)^N : \mathsf{dl}_\phi(P, P_N) \le 2\delta\right] \le P^{\otimes N}\left[\xi \in (\mathbb{R}^k)^N : |\vartheta(P) - \vartheta(P_N)| \le \frac{\epsilon}{2}\right].$$

Moreover, by the UGC property of $(\mathcal{M}^{\phi^p}_{k,\kappa}, \mathsf{dl}_\phi)$, there exists $N_0 \in \mathbb{N}$ such that

$$P^{\otimes N}\left[\xi \in (\mathbb{R}^k)^N : \mathsf{dl}_\phi(P, P_N) \le \delta\right] \ge 1 - \epsilon/2$$

for $N \ge N_0$, a combination of the two inequalities gives rise to (6.86). Next, we prove (6.87). For any fixed Q satisfying $\mathsf{dl}_\phi(P, Q) \le \delta$, we have $\mathsf{dl}_\phi(P, Q_N) \le \mathsf{dl}_\phi(P, Q) + \mathsf{dl}_\phi(Q, Q_N)$, By the UGC property

$$\begin{aligned}
1 - \frac{\epsilon}{2} &\le Q^{\otimes N}\left[\tilde{\xi} \in (\mathbb{R}^k)^N : \mathsf{dl}_\phi(Q, Q_N) \le \delta\right] \\
&\le Q^{\otimes N}\left[\tilde{\xi} \in (\mathbb{R}^k)^N : \mathsf{dl}_\phi(P, Q_N) \le \mathsf{dl}_\phi(P, Q) + \delta\right] \\
&\le Q^{\otimes N}\left[\tilde{\xi} \in (\mathbb{R}^k)^N : \mathsf{dl}_\phi(P, Q_N) \le 2\delta\right] \\
&\le Q^{\otimes N}\left[\tilde{\xi} \in (\mathbb{R}^k)^N : |\vartheta(P) - \vartheta(Q_N)| \le \frac{\epsilon}{2}\right],
\end{aligned}$$

which implies (6.87).

Part (ii). If for each $Q \in \mathcal{M}$, $S(Q)$ and $S(Q_N)$ are singletons, it follows from Theorem 3.1 (iii) that $S(\cdot)$ is continuous at P w.r.t. ϕ-weak topology. The rest holds based on similar analysis in Part (i). ■