MATHEMATICS OF OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0364-765X | EISSN 1526-5471 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Convergence Analysis for Distributionally Robust Optimization and Equilibrium Problems*

Hailin Sun
Department of Mathematics, Harbin Institute of Technology, Harbin, 150001, China, mathhlsun@gmail.com,
School of Economics and Management, Nanjing University of Science and Technology, Nanjing, 210049, China

Huifu Xu
School of Mathematics, University of Southampton, Southampton, UK, h.xu@soton.ac.uk

In this paper, we study distributionally robust optimization approaches for a one stage stochastic minimization problem, where the true distribution of the underlying random variables is unknown but it is possible to construct a set of probability distributions which contains the true distribution and optimal decision is taken on the basis of the worst possible distribution from that set. We consider the case when the distributional set (which is also known as ambiguity set) varies and its impact on the optimal value and the optimal solutions. A typical example is when the ambiguity set is constructed through samples and we need look into impact from increase of the sample size. The analysis provides a unified framework for convergence of some problems where the ambiguity set is approximated in a process with increasing information on uncertainty and extends the classical convergence analysis in stochastic programming. The discussion is extended briefly to a stochastic Nash equilibrium problem where each player takes a robust action on the basis of the worst subjective expected objective values.

Key words: Distributionally robust minimization, total variation metric, pseudometric, convergence analysis, Hoffman's lemma, robust Nash equilibrium
MSC2000 subject classification:
OR/MS subject classification: secondary:
History:

1. Introduction.

Consider the following distributionally robust stochastic program (DRSP):

$$\begin{array}{ll} \min\limits_{x} \sup\limits_{P\in\mathcal{P}} & \mathbb{E}_P[f(x,\xi(\omega))] \\ \text{s.t.} & x\in X, \end{array} \tag{1}$$

where X is a closed set of $\mathbb{R}^n$, $f:\mathbb{R}^n\times\mathbb{R}^k\to\mathbb{R}$ is a continuous function, $\xi:\Omega\to\Xi\subset\mathbb{R}^k$ is a vector of random variables defined on measurable space $(\Omega,\mathcal{F})$ equipped with sigma algebra $\mathcal{F}$, $\mathcal{P}$ is a set of probability distributions of ξ and $\mathbb{E}_P[\cdot]$ denotes the expected value with respect to probability distribution $P\in\mathcal{P}$. If we consider $(\Xi,\mathscr{B})$ as a measurable space equipped with Borel sigma algebra $\mathscr{B}$, then $\mathcal{P}$ may be viewed as a set of probability measures defined on $(\Xi,\mathscr{B})$ induced by the random variate ξ. Following the terminology in the literature of robust optimization, we

* The research is supported by EPSRC grants EP/J014427/1, EP/M003191/1 and National Natural Science Foundation of China (grant No.11171159 and No. 11401308).

call $\mathcal{P}$ the *ambiguity set* which indicates ambiguity of the true probability distribution of ξ in this setup. To ease notation, we will use ξ to denote either the random vector $\xi(\omega)$ or an element of $\mathbb{R}^k$ depending on the context.

Differing from classical stochastic programming models, the distributionally robust formulation (1) determines the optimal policy x on the basis of the worst expected value of $f(x,\xi)$ over the ambiguity set $\mathcal{P}$. It reflects some practical situation where a decision maker does not have complete information on the distribution of ξ and has to estimate it from data or construct it using subjective judgements [37]. This kind of robust optimization framework can be traced back to the earlier work by Scarf [32] which which addresses incomplete information on the underlying uncertainty in supply chain and inventory control problems. In such problems, historical data may be insufficient to estimate future distribution either because sample size of past demand is too small or because there is a reason to suspect that future demand will come from a different distribution which governs past history in an unpredictable way. A larger ambiguity set $\mathcal{P}$ which contains the true distribution may adequately address the risk from the uncertainty.

The minimax formulation has been well investigated through a number of further research works by Žáčková [46], Dupačová [14, 15], and more recently by Riis and Andersen [29], Shapiro and Kleywegt [38] and Shapiro and Ahmed [37]. Over the past few years, it has gained substantial popularity through further contributions by Bertsimas and Popescu [7], Betsimas et al [6], Goh and Sim [17], Zhu and Fukushima [47], Goh and Sim [17], Goldfarb and Iyengar [18], Delage and Ye [13] and Xu et al [45], to name a few, which cover a wide range of topics ranging from numerical tractability to applications in operations research, finance, engineering and computer science. In the case when $\mathcal{P}$ is a set of Dirac distributions which put weights on a single point in the support set Ξ, DRSP (1) reduces to worst scenario robust optimization, see monograph by Ben Tal et al [5] for the latter.

A key step in the research of DRSP (1) is to construct the ambiguity set $\mathcal{P}$. The construction must balance between exploitation of available information on the random parameters and numerical tractability of the resulting robust optimization model [15]. One way is to use samples/empirical data to estimate moments (e.g., mean and variance) and then specify the probability distribution through sample approximated moments [38, 13, 18]. Delage and Ye [13] propose a model that describes uncertainty in both the distribution form and moments, and demonstrate that for a wide range of functions f, (1) can be solved efficiently. Moreover, by deriving a new confidence region for the mean and the covariance matrix of ξ, they provide probabilistic arguments for so called data-driven problems that heavily rely on historical data and the arguments are consolidated by So [33] under weaker moment conditions. Another way is to use Bayesian method to specify a set of parameterized distributions that make the observed data achieve a certain level of likelihood [43, 44].

Obviously there is a gap between the ambiguity set constructed through estimated moments and that constructed with true moments and this gap depends on the sample size. An important question is whether one can close up this gap with more information on data (e.g., samples) and what is the impact of the gap on the optimal decision making. The question is fundamentally down to stability/consistency analysis of the robust optimization problem. In the case when the ambiguity set reduces to a singleton, DRSP (1) collapses to a classical one stage stochastic optimization problem. Convergence and/or stability analysis of the latter has been well documented, see review papers by Pflug [25], Römisch [31] and Shapiro [36]. The importance of such analysis lies in the fact that it develops a framework which enables one to examine convergence/asymptotic behavior of statistical quantities such as optimal value and optimal solution obtained through samples (or other approximation methods). The analysis provides a basic grounding for statistical consistency and numerical efficiency of the underlying approximation scheme. This argument applies to the DRSP when we develop various approximation schemes for the problem and this paper aims to

address the underlying theoretical issues relating to convergence and rate of convergence of optimal value and solutions obtained from solving an approximated DRSP.

However, there does not seem to be much compelling research on convergence/stability analysis of DRSPs. Breton and Hachem [11, 12] and Takriti and Ahmed [42] carry out stability analysis for some DRSPs with finite discrete probability distributions. Riis and Andersen [29] extend significantly their analysis to continuous probability distributions. Dupačová [15] shows epi-convergence of the optimal value function based on the worst probability distribution from an ambiguity set defined through estimated moments under some convexity and compactness conditions. Shapiro [35] investigate consistency of risk averse stochastic programs, where the dual formulation is a distributionally robust optimization problem. In a more recent development, there emerge a few new research which addresses convergence of distributionally robust optimization problems where $\mathcal{P}$ is Monte Carlo sampling, see Xu et al [45], Wang et al [43] and Wiesemann et al [44].

Our focus in this paper is on the case when the ambiguity set is approximated by a sequence of ambiguity sets constructed through samples or other means. For instance, when $\mathcal{P}$ is defined through moment conditions, the true moments are usually unknown but they can be estimated through empirical data. The ambiguity set (e.g. constructed through the estimated moments) may converge to a set with true moments rather than a single distribution. Riis and Andersen [29] carry out convergence analysis for this kind of minimax two stage stochastic optimization problem where $\mathcal{P}$ is inner approximated by a sequence of ambiguity sets under weak topology. Here we propose to study approximation of ambiguity sets under total variation metric and the pseudometric. The former allows us to measure the convergence of the ambiguity set as sample size increases whereas the latter translate the convergence of probability measures to that of optimal values. Specifically, we have made the following contributions:

- We treat the inner maximization problem of (1) as a parametric optimization problem and investigate the impact of variation of x and ambiguity set $\mathcal{P}$ on the optimal value. Under some moderate conditions, we show that the optimal value function is equi-Lipschitz continuous and the optimal solution set is upper semicontinuous w.r.t. x. Moreover, we demonstrate uniform convergence of the optimal value function and consequently convergence of the robust optimal solution of (1) to its true counterpart as the gap between the approximated ambiguity set and its true counterpart closes up. The convergence analysis provides a framework for examining statistical consistency and numerical efficiency of a fairly general approximation scheme for the DRSP and significantly strengthens an earlier convergence result by Riis and Andersen [29] (see details at the end of Section 3).
- We investigate convergence of the ambiguity sets under total variation metric as sample size increases for the cases when the ambiguity set is constructed through moments, mixture distribution, and moments and covariance matrix due to Delage and Ye [13] and So [33]. In the case when an ambiguity set is defined through moment conditions, we derive a Hoffman type error bound for a probabilistic system of inequalities and equalities through Shapiro's duality theorem [34] for linear conic programs and use it to derive a linear bound for the distance of the two ambiguity sets under the total variation metric.
- Finally, we outline possible extension our convergence results to a distributionally robust Nash equilibrium problem where each player takes a robust action on the basis of their subjective expected objective value over an ambiguity set.

Throughout the paper, we will use $\mathbb{R}^{n\times n}$ to denote the space of all $n\times n$ matrices, and $S_+^{n\times n}$ and $S_-^{n\times n}$ the cone of positive semi-definite and negative semi-definite symmetric matrices respectively. For matrices $A, B \in \mathbb{R}^{n\times n}$, we write $A\bullet B$ for the Frobenius inner product, that is $A\bullet B := \mathrm{tr}(A^T B)$, where "tr" denotes the trace of a matrix and the superscript T denotes transpose. Moreover, we use standard notation $\|A\|_F$ for the Frobenius norm of A, that is, $\|A\|_F := (A\bullet A)^{1/2}$, $\|x\|$ for the Euclidean norm of a vector x in $\mathbb{R}^n$, $\|x\|_\infty$ for the infinity norm and $\|\psi\|_\infty$ for the maximum norm

of a real valued measure function $\psi : \Xi \to \mathbb{R}$. Finally, for a set S, we use $\mathrm{cl}\, S$ to denote the closure of S and $\mathrm{int}\, S$ the interior.

2. Problem setting

2.1. Definition of the true and approximation problems. Let $(\Omega, \mathcal{F})$ be a measurable space equipped with σ-algebra $\mathcal{F}$ and $\xi : \Omega \to \Xi \subset \mathbb{R}^k$ be a random vector with support set Ξ. Let $\mathscr{B}$ denote the sigma algebra of all Borel subsets of Ξ and $\mathscr{P}$ be the set of all probability measures of the measurable space $(\Xi, \mathscr{B})$ induced by ξ. Let $\mathcal{P} \subset \mathscr{P}$ be defined as in (1) and $\mathcal{P}_N \subset \mathscr{P}$ be a set of probability measures which approximate $\mathcal{P}$ in some sense (to be specified later) as $N \to \infty$. We construct an approximation scheme for the distributionally robust optimization problem (1) by replacing $\mathcal{P}$ with $\mathcal{P}_N$:

$$\begin{array}{ll} \min\limits_{x} \sup\limits_{P \in \mathcal{P}_N} & \mathbb{E}_P[f(x, \xi(\omega))] \\ \text{s.t.} & x \in X. \end{array} \tag{2}$$

Typically, $\mathcal{P}_N$ may be constructed through samples. For instances, Shapiro and Ahmed [37] consider $\mathcal{P}$ being defined through moments and use empirical data (samples) to approximate the true moments. Delage and Ye [13] consider the case when $\mathcal{P}$ is defined through first order and second order moments and then use iid samples to construct $\mathcal{P}_N$ which approximates $\mathcal{P}$. More recently Wang et al [43] and Wiesemann et al [44] apply the Bayesian method to construct $\mathcal{P}_N$. Note that in practice, samples of data-driven problem are usually of small size. Our focus here is on the case that sample size could be large in order for us to carry out the convergence analysis. Note also that $\mathcal{P}_N$ does not have to be constructed through samples, it may be regarded in general as an approximation to $\mathcal{P}$.

To ease the exposition, for each fixed $x \in X$, let

$$v_N(x) := \sup_{P \in \mathcal{P}_N} \mathbb{E}_P[f(x, \xi)] \tag{3}$$

denote the optimal value of the inner maximization problem, and $\Phi_N(x)$ the corresponding set of optimal solutions, that is,

$$\Phi_N(x) := \{P \in \mathrm{cl}\mathcal{P}_N : v_N(x) = \mathbb{E}_P[f(x, \xi)]\},$$

where "cl" denotes the closure of a set and the closure is defined in the sense of weak topology, see Definition 3. Likewise, we denote

$$v(x) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[f(x, \xi)] \tag{4}$$

and $\Phi(x)$ the corresponding set of optimal solutions

$$\Phi(x) := \{P \in \mathrm{cl}\mathcal{P} : v(x) = \mathbb{E}_P[f(x, \xi)]\}.$$

Consequently we can write (2) and (1) respectively as

$$\vartheta_N := \min_{x \in X} v_N(x) \tag{5}$$

and

$$\vartheta := \min_{x \in X} v(x) \tag{6}$$

where ϑ_N and ϑ denote the optimal value and X_N and X^* the set of optimal solutions of (5) and (6) respectively. Our aim is to investigate convergence of ϑ_N to ϑ and X_N to X^* as $N \to \infty$. The reason that we consider $\mathcal{P}_N \to \mathcal{P}$ rather than the true distribution is that the latter may be different from the distribution which generates the samples. This is particularly so when ξ is used to describe the future uncertainty.

In the case when $\mathcal{P}_N$ is a singleton, (2) reduces to an ordinary approximation scheme of one stage stochastic minimization problem and our proposed analysis collapses to classical stability analysis in stochastic programming [31]. From this perspective, we might regard the convergence analysis in this paper as a kind of *global* stability analysis which allows the probability measure to perturb in a wider range.

In this section, we discuss well-definedness of (1) and (2). To this end, let us introduce some metrics for the set $\mathcal{P}_N$ and $\mathcal{P}$, which are appropriate for our problems.

2.2. Total variation metric. Let $\mathscr{P}$ be defined as in the beginning of the previous subsection. We need appropriate metrics for the set in order to characterize convergence of $\mathcal{P}_N \to \mathcal{P}$ and $v_N(x) \to v(x)$ respectively. To this end, we consider total variation metric for the former and pseudometric for the latter. Both metrics are well known in probability theory and stochastic programming, see for instance [2, 31].

Definition 1. Let $P, Q \in \mathscr{P}$ and $\mathscr{M}$ denote the set of measurable functions defined in the probability space $(\Xi, \mathscr{B})$. The *total variation metric* between P and Q is defined as (see e.g., page 270 in [2])

$$d_{TV}(P,Q) := \sup_{h \in \mathscr{M}} (\mathbb{E}_P[h(\xi)] - \mathbb{E}_Q[h(\xi)]), \tag{7}$$

where

$$\mathscr{M} := \{h : \mathbb{R}^k \to \mathbb{R} | h \text{ is } \mathscr{B} \text{ measurable}, \sup_{\xi \in \Xi} |h(\xi)| \le 1\}, \tag{8}$$

and *total variation norm* as

$$||P||_{TV} = \sup_{\|\phi\|_\infty \le 1} \mathbb{E}_P[\phi(\xi)].$$

If we restrict the measurable functions in set $\mathscr{M}$ to be uniformly Lipschitz continuous, that is,

$$\mathscr{M} = \{h : \sup_{\xi \in \Xi} |h(\xi)| \le 1,\ L_1(h) \le 1\}, \tag{9}$$

where $L_1(h) = \inf\{L : |h(\xi') - h(\xi'')| \le L\|\xi' - \xi''\|, \forall \xi', \xi'' \in \Xi\}$, then $d_{TV}(P,Q)$ is known as bounded Lipschitz metric, see e.g. [26] for details.

Using the total variation norm, we can define the distance from a point to a set, deviation from one set to another and Hausdorff distance between two sets in the space of $\mathscr{P}$. Specifically, let

$$d_{TV}(Q, \mathcal{P}) := \inf_{P \in \mathcal{P}} d_{TV}(Q,P),$$

$$\mathbb{D}_{TV}(\mathcal{P}_N, \mathcal{P}) := \sup_{Q \in \mathcal{P}_N} d_{TV}(Q, \mathcal{P})$$

and

$$\mathbb{H}_{TV}(\mathcal{P}_N, \mathcal{P}) := \max\{\mathbb{D}_{TV}(\mathcal{P}_N, \mathcal{P}), \mathbb{D}_{TV}(\mathcal{P}, \mathcal{P}_N)\}.$$

Here $\mathbb{H}_{TV}(\mathcal{P}_N, \mathcal{P})$ defines Hausdorff distance between $\mathcal{P}_N$ and $\mathcal{P}$ under the total variation metric in space $\mathscr{P}$. It is easy to observe that $\mathbb{H}_{TV}(\mathcal{P}_N, \mathcal{P}) \to 0$ implies $\mathbb{D}_{TV}(\mathcal{P}_N, \mathcal{P}) \to 0$ and

$$\inf_{Q \in \mathcal{P}} \sup_{h \in \mathscr{M}} (\mathbb{E}_{P_N}[h(\xi)] - \mathbb{E}_Q[h(\xi)]) \to 0$$

for any $P_N \in \mathcal{P}_N$. We will give detailed discussions about this when $\mathcal{P}$ and $\mathcal{P}_N$ are constructed in a specific way in Section 4.

DEFINITION 2. Let $\{P_N\} \subset \mathscr{P}$ be a sequence of probability measures. $\{P_N\}$ is said to converge to $P \in \mathscr{P}$ *under total variation metric* if

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

for each $h \in \mathscr{M}$. Let $\mathcal{A}$ be a set of probability measures on $(\Xi, \mathscr{B})$, where $\mathscr{B}$ is the Borel σ-algebra on Ξ. $\mathcal{A}$ is said to be *compact under total variation metric* if for any sequence $\{P_N\} \subset \mathcal{A}$, there exists subsequence $\{P_{N_t}\} \subset \{P_N\}$ such that $P_{N_t} \to P$ under total variation metric and $P \in \mathcal{A}$.

REMARK 1. Let $h \in \mathscr{M}$ and $\mathcal{A}$ be a set of probability measures. If $\mathcal{A}$ is compact under total variation metric, then $\mathscr{W} := \{\mathbb{E}_P[h] : P \in \mathcal{A}\}$ is a closed set. To see this, let $\{w_t\} \subset \mathscr{W}$ and $w_t \to v$ as $t \to \infty$. It suffices to show $w \in \mathscr{W}$. By definition, for each $w_t \in \mathscr{W}$, there exists a $P_t \in \mathcal{P}$ such that $w_t = \mathbb{E}_{P_t}[h]$. Since $\mathcal{P}$ is compact under the total variation metric, there exists a subsequence $\{P_{t_s}\}$ such that P_{t_s} converges to some P (under the total variation metric) with $P \in \mathcal{P}$, which means

$$w = \lim_{t_s\to\infty} w_{t_s} = \lim_{t_s\to\infty} \mathbb{E}_{P_{t_s}}[h(\xi)] = \mathbb{E}_P[h(\xi)].$$

This shows $w \in \mathscr{W}$ because $\mathbb{E}_P[h(\xi)] \in \mathscr{W}$.

DEFINITION 3. Let $\mathcal{A}$ be a set of probability measures on $(\Xi, \mathscr{B})$. $\mathcal{A}$ is said to be *tight* if for any $\epsilon > 0$, there exists a compact set $\Xi_\epsilon \subset \Xi$ such that $\inf_{P\in\mathcal{A}} P(\Xi_\epsilon) > 1 - \epsilon$. In the case when $\mathcal{A}$ is a singleton, it reduces to the tightness of a single probability measure. $\mathcal{A}$ is said to be *closed* (under the weak topology) if for any sequence $\{P_N\} \subset \mathcal{A}$ with $P_N \to P$ weakly, we have $P \in \mathcal{A}$.

DEFINITION 4. Let $\{P_N\} \subset \mathscr{P}$ be a sequence of probability measures. $\{P_N\}$ is said to converge to $P \in \mathscr{P}$ *weakly* if

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

for each bounded and continuous function $h : \Xi \to \mathbb{R}$. Let $\mathcal{A} \subset \mathscr{A}$ be a set of probability measures. $\mathcal{A}$ is said to be weakly compact if every sequence $\{A_N\} \subset \mathcal{A}$ contains a subsequence $\{A_{N'}\}$ and $A \in \mathcal{A}$ such that $A_{N'} \to A$.

Obviously if a sequence of probability measures converges under the total variation metric, then it converges weakly. Moreover, if h is a bounded continuous function defined on Ξ, then $\{\mathbb{E}_A[h] : A \in \mathcal{A}\}$ is a compact set in $\mathbb{R}$. Furthermore, by the well-known Prokhorov's theorem (see [27, 2]), a closed set $\mathcal{A}$ (under the weak topology) of probability measures is *compact* if it is tight. In particular, if Ξ is a compact metric space, then the set of all probability measures on $(\Xi, \mathscr{B})$ is compact in that Ξ is in a finite dimensional space; see [34]. In Section 4, we will discuss tightness and compactness in detail where $\mathcal{P}_N$ has a specific structure.

LEMMA 1. *Let Z be a separable metric space, P and $\{P_t\}$ be Borel probability measures on Z such that P_t converges weakly to P, let $h : Z \to \mathbb{R}$ be a measurable function with $P(D_h) = 0$, where $D_h := \{z \in Z : h$ is not continuous at $z\}$. Then it holds*

$$\lim_{t\to\infty} \int_Z h(z) P_t(dz) = \int_Z h(z) P(dz)$$

if the sequence $\{P_t h^{-1}\}$ is uniformly integrable*, i.e.,*

$$\lim_{r\to\infty} \sup_{t\in\mathcal{N}} \int_{\{z\in Z:|h(z)|\geq r\}} |h(z)| P_t(dz) = 0,$$

where $\mathcal{N}$ denotes the set of positive integers. A sufficient condition for the uniform integrability is:

$$\sup_{t\in\mathcal{N}} \int_Z |h(z)|^{1+\varepsilon} P_t(dz) < \infty \quad \text{for some } \varepsilon > 0.$$

The results of this lemma are summarized from [8, Theorem 5.4] and the preceding discussions. It is easy to observe that if h is continuous and bounded, then for all probability measure P, $P(D_h)=0$ and $\{P_t h^{-1}\}$ is uniformly integrable.

PROPOSITION 1. *Let $\mathcal{P}$ be a set of probability measures of separable measurable space Z, let $h: Z \to \mathbb{R}$ be a measurable function with $P(D_h)=0$ for all $P \in \mathcal{P}$ and $\{Ph^{-1}, P \in \mathcal{P}\}$ is uniformly integrable, where $D_h = \{z \in Z : h \text{ is not continuous at } z\}$. Let $\mathcal{V} := \{\mathbb{E}_P[h(z)] : P \in \mathrm{cl}\,\mathcal{P}\}$. If $\mathcal{P}$ is tight, then $\mathcal{V}$ is a compact set.*

Proof. Since $h(z)$ is uniformly integrable for all $P \in \mathcal{P}$, $\mathcal{V}$ is bounded. It suffices to show that $\mathcal{V}$ is closed. Let $\{v_t\} \subset \mathcal{V}$ be any sequence converging to $\hat{v}$. We show $\hat{v} \in \mathcal{V}$. Let $\{P_t\}$ be such that $\mathbb{E}_{P_t}[h(z)] = v_t$. Since $\mathcal{V}$ is bounded, by taking a subsequence if necessary, there exists $\hat{v}$ such that $v_t \to \hat{v}$. Since $\mathrm{cl}\mathcal{P}$ is compact under the weak topology, by taking a subsequence if necessary, we may assume that $P_t \to \hat{P}$ weakly. It follows by Lemma 1

$$\hat{v} = \lim_{t\to\infty} v_t = \lim_{t\to\infty} \mathbb{E}_{P_t}[h(z)] = \mathbb{E}_{\hat{P}}[h(z)]$$

The closedness of $\mathrm{cl}\mathcal{P}$ means $\hat{P} \in \mathrm{cl}\,\mathcal{P}$ and hence $\hat{v} \in \mathcal{V}$. □

2.3. Pseudometric. The total variation metric we discussed is independent of function $f(x,\xi)$ defined in the distributionally robust optimization problem (1). In what follows, we introduce another metric which is closely related to the objective function $f(x,\xi)$.

Define the set of random functions:

$$\mathcal{G} := \{g(\cdot) := f(x,\cdot) : x \in X\}. \tag{10}$$

The distance for any probability measures $P, Q \in \mathscr{P}$ is defined as:

$$\mathscr{D}(P,Q) := \sup_{g\in\mathcal{G}} \big|\mathbb{E}_P[g] - \mathbb{E}_Q[g]\big|. \tag{11}$$

Here we implicitly assume that $\mathscr{D}(P,Q) < \infty$. We will come back to this in the next subsection. We call $\mathscr{D}(P,Q)$ *pseudometric* in that it satisfies all properties of a metric except that $\mathscr{D}(P,Q)=0$ does not necessarily imply $P=Q$ unless the set of functions $\mathcal{G}$ is sufficiently large. This type of pseudometric is widely used for stability analysis in stochastic programming; see an excellent review by Römisch [31].

Let $Q \in \mathscr{P}$ be a probability measure and $\mathcal{A}_i \subset \mathscr{P}$, $i=1,2$, be a set of probability measures. With the pseudometric, we may define the distance from a single probability measure Q to a set of probability measures $\mathcal{A}_1$ as $\mathscr{D}(Q,\mathcal{A}_1) := \inf_{P\in\mathcal{A}_1} \mathscr{D}(Q,P)$, the deviation (excess) of $\mathcal{A}_1$ from (over) $\mathcal{A}_2$ as

$$\mathscr{D}(\mathcal{A}_1,\mathcal{A}_2) := \sup_{Q\in\mathcal{A}_1} \mathscr{D}(Q,\mathcal{A}_2) \tag{12}$$

and Hausdorff distance between $\mathcal{A}_1$ and $\mathcal{A}_2$ as

$$\mathscr{H}(\mathcal{A}_1,\mathcal{A}_2) := \max\left\{\sup_{Q\in\mathcal{A}_1} \mathscr{D}(Q,\mathcal{A}_2), \sup_{Q\in\mathcal{A}_2} \mathscr{D}(Q,\mathcal{A}_1)\right\}. \tag{13}$$

REMARK 2. There are two important cases to note.

(i) Consider the case when $\mathcal{G}$ is bounded, that is, there exists a positive number M such that $\sup_{g\in\mathcal{G}} \|g\| \le M$. Let $\tilde{\mathcal{G}} = \mathcal{G}/M$. Then

$$\mathscr{D}(P,Q) := M \sup_{\tilde{g}\in\tilde{\mathcal{G}}} \big|\mathbb{E}_P[\tilde{g}] - \mathbb{E}_Q[\tilde{g}]\big| \le M d_{TV}(P,Q). \tag{14}$$

(ii) Consider the set of functions

$$\mathscr{G} := \{f(x,\xi) : x \in X, \sup_{x\in X} |f(x,\xi) - f(x,\xi')| \le c_p(\xi,\xi')\|\xi - \xi'\| : \forall \xi, \xi' \in \Xi\}, \tag{15}$$

where $c_p(\xi,\xi') := \max\{1, \|\xi\|, \|\xi'\|\}^{p-1}$ for all $\xi, \xi' \in \Xi$ and $p \ge 1$. When $p = 1$, $\mathscr{D}(P,Q)$ is associated with the well-known *Kantorovich metric* and when $p \ge 1$ it is related to the p-th order *Fortet-Mourier metric* over the subset of probability measures having finite p-th order moments. It is well-known that a sequence of probability measures $\{P_N\}$ converges to P under Fortet-Mourier metric (both P_N and P having p-th order moments) iff it converges to P weakly and

$$\lim_{N\to\infty} \mathbb{E}_{P_N}[\|\xi\|^p] = \mathbb{E}_P[\|\xi\|^p] < \infty, \tag{16}$$

see [31]. It means that if f satisfies conditions (15), then weak convergence of P_N to $P \in \mathcal{P}$ and (16) imply $\mathscr{D}(P_N, P) \to 0$ and hence $\mathscr{D}(P_N, \mathcal{P}) \to 0$. If (16) holds for any $\{P_N\} \in \{\mathcal{P}_N\}$, then we arrive at $\mathscr{D}(\mathcal{P}_N, \mathcal{P}) \to 0$.

2.4. Well definedness of the robust problem. We need to make sure that problems (5) and (6) are well defined, that is, the objective functions $v_N(x)$ and $v(x)$ are finite valued and enjoy some nice properties. This requires us to investigate parametric programs (3) and (4) where x is treated as a parameter and probability measure P is a variable. To this end, we make a few assumptions. Let $\hat{\mathcal{P}} \subset \mathscr{P}$ denote a set of distributions such that

$$\mathcal{P}, \mathcal{P}_N \subset \hat{\mathcal{P}} \tag{17}$$

for N sufficiently large. Existence of $\hat{\mathcal{P}}$ is trivial as we can take the union of $\mathcal{P}$ and $\mathcal{P}_N$, for $N = 1, 2, 3, \cdots$. However, our interest is in the case where $\hat{\mathcal{P}}$ satisfies (17) and has some specified characteristics such as tightness or compactness under the weak topology. We also prefer the set to be as small as possible although we do not necessarily consider the smallest one. We will elaborate with details later on.

ASSUMPTION 1. *Let $f(x,\xi)$ be defined as in (1) and $\hat{\mathcal{P}}$ be a set of probability measures satisfying (17).*

(a) For each fixed $\xi \in \Xi$, $f(\cdot,\xi)$ is Lipschitz continuous on X with Lipschitz modulus being bounded by $\kappa(\xi)$, where $\sup_{P\in\hat{\mathcal{P}}} \mathbb{E}_P[\kappa(\xi)] < \infty$.

(b) There exists $x_0 \in X$ such that $\sup_{P\in\hat{\mathcal{P}}} \|\mathbb{E}_P[f(x_0,\xi)]\| < \infty$.

(c) X is a compact set.

Assumption 1 (a) is a standard condition in stochastic programming, see e.g. [40]; Assumption 1 (b) is also standard if $\hat{\mathcal{P}}$ is a singleton. Our interest here is the case when $\hat{\mathcal{P}}$ is a set. The condition may be satisfied when $\hat{\mathcal{P}}$ is tight and $f(x_0,\xi)$ is uniformly integrable, see Proposition 1. Assumption 1 (c) may be relaxed to the case when X is merely closed but it would then require additional conditions such as inf-compactness for the convergence analysis. We make it simpler so that we can concentrate our analysis on other important aspects.

Under Assumption 1, it is easy to verify that the pseudometric defined in (11) is bounded.

ASSUMPTION 2. *Let $\mathcal{P}, \mathcal{P}_N$ be defined as in (1) and (2) respectively.*

(a) $\mathcal{P}$ is nonempty and tight;

(b) for each N, $\mathcal{P}_N$ is a nonempty tight set;

(c) there exists a weakly compact set $\hat{\mathcal{P}} \subset \mathscr{P}$ such that (17) holds.

The assumption is rather general and may be verified when we have a concrete structure of the ambiguity sets, see Remark 3, Proposition 5 and Proposition 7. Note that in general tightness is weaker than weak compactness, the following example explains this.

EXAMPLE 1. Let ξ be a random variable defined on $\mathbb{R}$ with σ-algebra $\mathcal{F}$. Let $\mathscr{P}$ denote the set of all probability measures on $(\mathbb{R},\mathcal{F})$. Let

$$\mathcal{P}=\left\{P\in\mathscr{P}:\ \begin{array}{l}\mathbb{E}_P[\xi]=0\\ \mathbb{E}_P[\xi^2]=1\end{array}\right\}.$$

Note that $\mathcal{P}$ is not closed. To see this, we consider a special sequence of distributions $\{P_j\}$ with

$$P_j(\xi^{-1}(-\sqrt{j}))=\frac{1}{2j},\ P_j(\xi^{-1}(0))=1-\frac{1}{j}\ \text{ and }\ P_j(\xi^{-1}(\sqrt{j}))=\frac{1}{2j}.$$

It is easy to see that $P_j\in\mathcal{P}$, and P_j converges weakly to P^* with $P^*(\xi^{-1}(0))=1$, but $P^*\notin\mathcal{P}$. Note that since $\mathbb{E}_P[\xi^2]$ is bounded for all $P\in\mathcal{P}$, by Dunford-Pettis theorem (see [3, Theorem 2.4.5]), $\mathcal{P}$ is tight.

The proposition below summarizes main properties of the optimal value and the optimal solutions of parametric programs (3) and (4).

PROPOSITION 2. *Let Assumption 1 (a) and (b) hold. The following assertions hold.*

(i) Under Assumption 1 (c), $\mathbb{E}_P[f(x,\xi)]$ is Lipschitz continuous w.r.t. (P,x) on $\operatorname{int}\operatorname{cl}\mathcal{P}_N\times X$, that is,

$$|\mathbb{E}_P[f(x,\xi)]-\mathbb{E}_Q[f(y,\xi)]|\le\mathscr{D}(P,Q)+\sup_{P\in\hat{\mathcal{P}}}\mathbb{E}_P[\kappa(\xi)]\|x-y\| \tag{18}$$

for $P,Q\in\operatorname{int}\operatorname{cl}\mathcal{P}_N$ and $x,y\in X$.

Assume, in addition, Assumption 2 holds and $\{Pf^{-1}(x,\cdot),P\in\mathcal{P}_N\}$ is uniformly integrable, i.e.,

$$\lim_{r\to\infty}\sup_{P\in\mathcal{P}_N}\int_{\{\xi\in\Xi:|f(x,\xi)|\ge r\}}|f(x,\xi)|P(d\xi)=0. \tag{19}$$

Then

(ii) $\Phi_N(x)\neq\emptyset$;

(iii) $\Phi_N(\cdot)$ is upper semicontinuous at every fixed point in X;

(iv) for all $x\in X$, $v_N(x)<\infty$. Moreover, $v_N(\cdot)$ is equi-Lipschitz continuous on X with modulus being bounded by $\sup_{P\in\hat{\mathcal{P}}}\mathbb{E}_P[\kappa(\xi)]$, that is,

$$|v_N(x)-v_N(y)|\le\sup_{P\in\hat{\mathcal{P}}}\mathbb{E}_P[\kappa(\xi)]\|x-y\|,\ \forall x,y\in X; \tag{20}$$

(v) if $\{Pf^{-1}(x,\cdot),P\in\mathcal{P}\}$ is uniformly integrable, then $v(\cdot)$ is Lipschitz continuous on X.

Proof. Under Assumption 1

$$\sup_{x\in X,P\in\hat{\mathcal{P}}}\|\mathbb{E}_P[f(x,\xi)]\|<\infty.$$

Part (i). Observe first that for every $x\in X$, $\mathbb{E}_P[f(x,\xi)]$ is continuous in P under the pseudometric $\mathscr{D}$. In fact, for any $P,Q\in\mathcal{P}_N$

$$|\mathbb{E}_P[f(x,\xi)]-\mathbb{E}_Q[f(x,\xi)]|\le\mathscr{D}(P,Q)<\infty. \tag{21}$$

For any $x,y\in X$,

$$|\mathbb{E}_P[f(x,\xi)]-\mathbb{E}_P[f(y,\xi)]|\le\sup_{P\in\hat{\mathcal{P}}}\mathbb{E}_P[\kappa(\xi)]\|x-y\|. \tag{22}$$

Combining (21) and (22), we obtain (18), that is, $\mathbb{E}_P[f(x,\xi)]$ is Lipschitz continuous w.r.t. (P,x) on $\mathcal{P}_N\times X$.

Part (ii). Let

$$\mathscr{V}_N:=\{\mathbb{E}_P[f(x,\xi)]:P\in\operatorname{cl}\mathcal{P}_N\}. \tag{23}$$

Assumption 1 ensures that $\mathscr{V}_N$ is bounded and through Prokhorov's theorem, Assumption 2 (b) guarantees that cl $\mathcal{P}_N$ is compact under weak topology. With the weak compactness, and the uniform integrability of $\{Pf^{-1}(x,\cdot), P\in\mathcal{P}_N\}$, it follows by Proposition 1 that $\mathscr{V}_N$ is compact and hence $\Phi_N(x)\neq\emptyset$.

Part (iii). Under the continuity of $\mathbb{E}_P[f(x,\xi)]$ with respect to P and the nonemptiness of $\Phi_N(x)$, it follows by [4, Theorem 4.2.1] that $\Phi_N(\cdot)$ is upper semicontinuous at every point in X.

Part (iv). Under Assumption 1, $\mathscr{V}_N$ is bounded which implies boundedness of $v_N(x)$. The proof of Lipschitz continuity is similar to [22, Theorem 1]. By Part (ii), $\Phi_N(y)$ is not empty. Let $P_N(y)\in\Phi_N(y)$. By (18)

$$\begin{aligned} v_N(x)\geq\mathbb{E}_{P_N(y)}[f(x,\xi)] &\geq \mathbb{E}_{P_N(y)}[f(y,\xi)]-|\mathbb{E}_{P_N(y)}[f(x,\xi)]-\mathbb{E}_{P_N(y)}[f(y,\xi)]| \\ &\geq v_N(y)-\sup_{P\in\hat{\mathcal{P}}}\mathbb{E}_P[\kappa(\xi)]\|x-y\|. \end{aligned}$$

Exchanging the role of x and y, we obtain (20).

Part (v). Similar to the proof of Parts (ii)-(iv), we can show that $\Phi(x)$ is nonempty for every $x\in X$, $\Phi(\cdot)$ is upper semicontinuous on X and $v(\cdot)$ is Lipschitz continuous by replacing $\mathcal{P}_N$ with $\mathcal{P}$. We omit the details. The proof is complete. □

Note that in order to solve distributionally robust optimization problem (3), one usually needs to reformulate it through Lagrangian dualization in the case when $\mathcal{P}_N$ has a specific structure. We will not go to details in this regard as it is well discussed in the literature, see [13] and the references therein.

3. Convergence analysis.

In this section, we analyze convergence of the optimal value ϑ_N and the optimal solution set X_N as $\mathcal{P}_N\to\mathcal{P}$. We will carry out the analysis without referring to a specific structure of $\mathcal{P}_N$ or $\mathcal{P}$ in order to maximize the coverage of the established in application. To ensure convergence of the ambiguity sets implies convergence of the optimal values and the optimal solutions, we need to strengthen our assumptions so that the optimal value function of (5) converges to that of (6) under pseudometrics.

ASSUMPTION 3. *Let $\mathcal{P},\mathcal{P}_N$ be defined as in (1) and (2) respectively.*
(a) $\mathscr{H}(\mathcal{P}_N,\mathcal{P})\to 0$ almost surely as $N\to\infty$, where $\mathscr{H}(\cdot,\cdot)$ is defined as in (13);
(b) for any $\epsilon>0$, there exist positive constants α and β (depending on ϵ) such that

$$\text{Prob}(\mathscr{D}(\mathcal{P}_N,\mathcal{P})\geq\epsilon)\leq\alpha e^{-\beta N}$$

for N sufficiently large, where $\mathscr{D}(\cdot,\cdot)$ is defined as in (12).

We consider the convergence of $\mathscr{D}(\mathcal{P}_N,\mathcal{P})$ in a probabilistic manner because in practice, $\mathcal{P}_N$ may be constructed in various ways associated with sampling. For instance, if $\mathcal{P}_N$ is generated through a sequence of independent and identically distributed (iid) random variables $\xi^1,\xi^2,\cdots$ with the same distribution as ξ, then the probability measure "Prob" should be understood as the product probability measure of P over measurable space $\Xi\times\Xi\cdots$ with product Borel sigma-algebra $\mathscr{B}\times\mathscr{B}\cdots$.

Assumption 3 is rather general and can be verified only when $\mathcal{P}_N$ has a structure. In Section 4, we will discuss how Assumption 3 (b) may be satisfied (note that (b) implies (a)) when the ambiguity set $\mathcal{P}_N$ is constructed in a specific way, see Corollary 1, Corollary 2 and Theorem 4. It is an open question as to whether the assumption is satisfied when $\mathcal{P}_N$ is constructed through Kullback-Leibler-divergence.

Under Assumption 3, we are able to present one of the main convergence results in this section.

THEOREM 1. *Under Assumption 1 and Assumption 3 (a), the following assertions hold.*

(i) $v_N(x)$ converges uniformly to $v(x)$ over X as N tends to infinity, that is,

$$\lim_{N\to\infty} \sup_{x\in X} |v_N(x) - v(x)| = 0 \tag{24}$$

almost surely.

(ii) If, in addition, Assumption 3 (b) holds, then for any $\epsilon > 0$ there exist positive constants α and β such that

$$\text{Prob}\left(\sup_{x\in X} |v_N(x) - v(x)| \geq \epsilon\right) \leq \alpha e^{-\beta N} \tag{25}$$

for N sufficiently large.

Part (i) of the theorem says that $v_N(\cdot)$ converges to $v(\cdot)$ uniformly over X almost surely as $N \to \infty$ and Part (ii) states that it converges in distribution at an exponential rate.

Proof of Theorem 1. Under Assumption 1, it follows from the proof of Proposition 2 (ii), $v_N(x) < \infty$.

Part (i). Let $x \in X$ be fixed. Let $\mathscr{V} := \{\mathbb{E}_P[f(x,\xi)] : P \in \text{cl}\,\mathcal{P}\}$ and $\mathscr{V}_N$ be defined as in (23). Under Assumption 1, both $\mathscr{V}$ and $\mathscr{V}_N$ are bounded sets in $\mathbb{R}$. Let

$$a := \inf_{v\in\mathscr{V}} v;\ \ b := \sup_{v\in\mathscr{V}} v;\ \ a_N := \inf_{v\in\mathscr{V}_N} v;\ \ b_N := \sup_{v\in\mathscr{V}_N} v.$$

Let "conv" denote the convex hull of a set. Then the Hausdorff distance between $\text{conv}\mathscr{V}$ and $\text{conv}\mathscr{V}_N$ can be written as follows:

$$\mathbb{H}(\text{conv}\mathscr{V}, \text{conv}\mathscr{V}_N) = \max\{|b_N - b|, |a - a_N|\}.$$

Note that

$$b_N - b = \max_{P\in\mathcal{P}_N} \mathbb{E}_P[f(x,\xi)] - \max_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)]$$

and

$$a_N - a = \min_{P\in\mathcal{P}_N} \mathbb{E}_P[f(x,\xi)] - \min_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)]$$

Therefore

$$\mathbb{H}(\text{conv}\mathscr{V}, \text{conv}\mathscr{V}_N) = \max\left\{\left|\max_{P\in\mathcal{P}_N} \mathbb{E}_P[f(x,\xi)] - \max_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)]\right|, \left|\min_{P\in\mathcal{P}_N} \mathbb{E}_P[f(x,\xi)] - \min_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)]\right|\right\}.$$

On the other hand, by the definition and property of the Hausdorff distance (see e.g. [20]),

$$\mathbb{H}(\text{conv}\mathscr{V}, \text{conv}\mathscr{V}_N) \leq \mathbb{H}(\mathscr{V}, \mathscr{V}_N) = \max(\mathbb{D}(\mathscr{V}, \mathscr{V}_N), \mathbb{D}(\mathscr{V}_N, \mathscr{V}))$$

where

$$\begin{aligned}\mathbb{D}(\mathscr{V}, \mathscr{V}_N) &= \max_{v\in\mathscr{V}} d(v, \mathscr{V}_N) = \max_{v\in\mathscr{V}} \min_{v'\in\mathscr{V}_N} \|v - v'\| = \max_{P\in\mathcal{P}} \min_{Q\in\mathcal{P}_N} |\mathbb{E}_P[f(x,\xi)] - \mathbb{E}_Q[f(x,\xi)]| \\ &\leq \max_{P\in\mathcal{P}} \min_{Q\in\mathcal{P}_N} \sup_{x\in X} |\mathbb{E}_P[f(x,\xi)] - \mathbb{E}_Q[f(x,\xi)]| = \mathscr{D}(\mathcal{P}, \mathcal{P}_N).\end{aligned}$$

Likewise, we can show $\mathbb{D}(\mathscr{V}_N, \mathscr{V}) \leq \mathscr{D}(\mathcal{P}_N, \mathcal{P})$. Therefore

$$\mathbb{H}(\text{conv}\mathscr{V}, \text{conv}\mathscr{V}_N) \leq \mathbb{H}(\mathscr{V}, \mathscr{V}_N) \leq \mathscr{H}(\mathcal{P}, \mathcal{P}_N),$$

which subsequently yields

$$|v_N(x) - v(x)| = \left|\max_{P\in\mathcal{P}_N} \mathbb{E}_P[f(x,\xi)] - \max_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)]\right| \leq \mathscr{H}(\mathcal{P}, \mathcal{P}_N).$$

Note that x is any point taken from X and the right hand side of the inequality above is independent of x. By taking supremum w.r.t. x on both sides, we arrive at (24).

Part (ii) follows straightforwardly from part (i) and Assumption 3 (b). □

Assumption 3 is essential for deriving the convergence results in Theorem 1. It would therefore be helpful to discuss how the conditions stipulated in the assumption could be possibly satisfied. The proposition below states some sufficient conditions.

PROPOSITION 3. *Assumption 3 (a) is satisfied if one of the following conditions hold.*

(a) $\mathcal{P}_N$ converges to $\mathcal{P}$ under total variation metric and $f(x,\xi)$ is uniformly bounded, that is, there exists a positive constant M_1 such that $|f(x,\xi)| \le M_1$, for all $(x,\xi) \in X \times \Xi$.

(b) $\mathscr{D}(\mathcal{P},\mathcal{P}_N) \to 0$, f satisfies condition (15), and for every sequence $\{P_N\} \subset \{\mathcal{P}_N\}$, $\{P_N\}$ converges to $P \in \mathcal{P}$ weakly and (16) holds.

Proof. Sufficiency of (a) and (b) follows from Remark 2 (i) and (ii). □

As we have commented in Section 2, our analysis collapses to classical stability analysis in stochastic programming when $\mathcal{P}_N$ reduces to a singleton. In such a case, condition (b) in Proposition 3 reduces to the standard sufficient condition required in stability analysis of stochastic programming; see [31]. The uniform convergence established in Theorem 1 may be translated into convergence of the optimal values and the optimal solutions of problem (5) through Liu and Xu [23, Lemma 3.8]. Therefore the convergence analysis presented here is a kind of global (or robust) stability analysis of the optimal value and the optimal solutions in a stochastic decision making system against the underlying system error (random data).

THEOREM 2. *Let X_N and X^* denote the set of the optimal solutions to (2) and (1) respectively, ϑ_N and ϑ the corresponding optimal values. Assume that X is a compact set, X_N and X^* are nonempty. Under Assumptions 1, 2 and 3 (a),*

$$\overline{\lim_{N\to\infty}} X_N \subset X^* \quad \textit{and} \quad \lim_{N\to\infty} \vartheta_N = \vartheta.$$

If, in addition, Assumption 3 (b) holds, then for any $\epsilon > 0$ there exist positive constants α and β (depending on ϵ) such that

$$\text{Prob}\left(|\vartheta_N - \vartheta| \ge \epsilon\right) \le \alpha e^{-\beta N}$$

for N sufficiently large.

Proof. By Proposition 2, $v(\cdot)$ and $v_N(\cdot)$ are continuous. The conclusions follow from Theorem 1 and Liu and Xu [23, Lemma 3.8]. □

To emphasize the importance of the convergence results, we conclude this section with an example suggested by a reviewer.

EXAMPLE 2. Let ξ be a random variable defined on $\mathbb{R}$ with σ-algebra $\mathcal{F}$. Let $\mathscr{P}$ denote the set of all probability measures on $(\mathbb{R},\mathcal{F})$. Let

$$\mathcal{P} := \left\{P \in \mathscr{P} : P(\xi^{-1}(0)) = 1\right\}$$

which is a singleton. We consider the following distributionally robust optimization problem

$$\min_{x\in[1,2]} \sup_{P\in\mathcal{P}} \mathbb{E}_P[x\xi^2]. \tag{26}$$

Obviously $\sup_{P\in\mathcal{P}} \mathbb{E}_P[x\xi^2] = 0$ and the optimal value is 0. Let us now consider an approximation of $\mathcal{P}$ by $\mathcal{P}_t$ which is defined as:

$$\mathcal{P}_t := \left\{ P \in \mathscr{P} : \begin{array}{l} P(\xi^{-1}(0)) = 1 - \frac{1}{t} \\ P(\xi^{-1}([\frac{\sqrt{t}}{2}, \sqrt{t}])) = \frac{1}{t} \end{array} \right\}. \tag{27}$$

The corresponding distributionally robust minimization problem is

$$\min_{x\in[1,2]} \sup_{P\in\mathcal{P}_t} \mathbb{E}_P[x\xi^2]. \tag{28}$$

Let P_t be such that

$$P_t(\xi^{-1}(0)) = 1 - \frac{1}{t} \quad \text{and} \quad P_t(\xi^{-1}(\sqrt{t})) = \frac{1}{t}.$$

Then $P_t \in \mathcal{P}_t$ and it is easy to observe that $\sup_{P\in\mathcal{P}_t} \mathbb{E}_P[x\xi^2] = \mathbb{E}_{P_t}[x\xi^2] = x$ for all $x \in [1,2]$. However, there is another issue here: with $\mathcal{P}_t$ being defined as in (27), the optimal value of problem (28) is 1 whereas the optimal value of problem (26) is 0. On the other hand, P_t converges to $\mathcal{P}$ weakly. The underlying reason for the failure of convergence of the optimal value of problem (28) to that of problem (26) is that $\mathscr{H}(\mathcal{P}_t, \mathcal{P}) = 2$, which Assumption 3 (a) fails. In other words, this assumption is a necessary condition for the desired convergence without which one can construct a specific $f(x,\xi)$ and a sequence of ambiguity sets such that $\mathscr{H}(\mathcal{P}_t, \mathcal{P}) \not\to 0$ and consequently ϑ_t fails to converge to ϑ.

Before concluding this section, we note that Riis and Andersen [29] carry out similar convergence analysis when the minimax distributionally robust formulation is applied to a two-stage stochastic program with recourse. A key condition in their analysis is that there exists a sequence of probability measures $\{P_N\} \subset \mathcal{P}_N$ such that P_N converges to $P \in \mathcal{P}$ weakly (see [29, Proposition 2.1]). This condition implicitly requires the underlying random function $f(x,\xi)$ to be bounded and continuous w.r.t. ξ over $\mathbb{R}^n$ in some circumstances; see condition (iii) [29, Proposition 2.1]. Our analysis is carried out under the pseudometric which is based on the uniform integrability of f over its support set Ξ. Another important assumption in [29] is that $\mathcal{P}_N$ must be contained in $\mathcal{P}$. Here, we don't require $\mathcal{P}$ to be inner approximated by $\mathcal{P}_N$ because it may not be satisfied in some important instances, see Section 4.

4. Approximations of the ambiguity set. One of the key assumptions in the convergence analysis is the convergence of $\mathcal{P}_N$ to $\mathcal{P}$. In this section, we look into details as to how such convergence may be obtained. In the literature of robust optimization, there have been various ways to construct the ambiguity set $\mathcal{P}_N$. Here we review some of them and present a quantitative convergence analysis of $\mathcal{P}_N$ to $\mathcal{P}$ under total variation metric.

4.1. Moment problems. Let us first consider the case when the ambiguity set $\mathcal{P}$ is defined through moment conditions:

$$\mathcal{P} := \left\{ P \in \mathscr{P} : \begin{array}{l} \mathbb{E}_P[\psi_i(\xi)] = \mu_i, \text{ for } i = 1,\cdots,p \\ \mathbb{E}_P[\psi_i(\xi)] \le \mu_i, \text{ for } i = p+1,\cdots,q \end{array} \right\}, \tag{29}$$

where $\xi : \Omega \to \Xi$ is defined as in (1), $\psi_i : \Xi \to \mathbb{R}$, $i = 1,\cdots,q$, are measurable functions and $\mathscr{P}$ is the set of all probability measures on Ξ induced by ξ as defined at the beginning of Section 2.1. In other words, the mathematical expectations are taken with respect to *some* probability distribution of the random vector ξ. Let

$$\mathcal{P}_N := \left\{ P \in \mathscr{P} : \begin{array}{l} \mathbb{E}_P[\psi_i(\xi)] = \mu_i^N, \text{ for } i = 1,\cdots,p \\ \mathbb{E}_P[\psi_i(\xi)] \le \mu_i^N, \text{ for } i = p+1,\cdots,q \end{array} \right\} \tag{30}$$

be an approximation to $\mathcal{P}$, where μ_i^N is constructed through samples. To simplify the notation, let $\psi_E = (\psi_1,\cdots,\psi_p)^T$, $\psi_I = (\psi_{p+1},\cdots,\psi_q)^T$, where the subscripts E and I indicates the components corresponding equality constraints and inequality constraints respectively. Likewise, let $\mu_E = (\mu_1,\cdots,\mu_p)^T$ and $\mu_I = (\mu_{p+1},\cdots,\mu_q)^T$. Then we can rewrite $\mathcal{P}$ and $\mathcal{P}_N$ as

$$\mathcal{P} = \{P \in \mathscr{P} : \mathbb{E}_P[\psi_E(\xi)] = \mu_E, \mathbb{E}_P[\psi_I(\xi)] \le \mu_I\}$$

and

$$\mathcal{P}_N = \{P \in \mathscr{P} : \mathbb{E}_P[\psi_E(\xi)] = \mu_E^N, \mathbb{E}_P[\psi_I(\xi)] \le \mu_I^N\}.$$

In what follows, we investigate approximation of $\mathcal{P}_N$ to $\mathcal{P}$ when μ_i^N converges to μ_i. By viewing $\mathcal{P}$ as a set of solutions to the system of equalities and inequalitie defined by (29), we may derive an error bound for a probability measure deviating from set $\mathcal{P}$. This kind of result may be regarded as a generalization of classical Hoffman's lemma in a finite dimensional space (see [39, Theorem 7.11]).

Let us write $\langle P, \psi_E(\xi)\rangle$ for $\mathbb{E}_P[\psi_E(\xi)]$ and $\langle P, \psi_I(\xi)\rangle$ for $\mathbb{E}_P[\psi_I(\xi)]$. Let int (S) denote the interior of set S. Let $\mathscr{M}_+$ denote the space of positive measures generated by $\mathscr{P}$.

LEMMA 2. *(Hoffman's lemma for moment problem). Assume the regularity condition*

$$(1, \mu_E, \mu_I) \in \text{int}\,[(\langle P, 1\rangle, \langle P, \psi_E(\xi)\rangle, \langle P, \psi_I(\xi)\rangle) - \{0\} \times \{0_p\} \times \mathbb{R}_-^{q-p} : P \in \mathscr{M}_+] \tag{31}$$

holds and $\mathcal{P}$ is tight. Then there exists a positive constant C_1 depending on ψ such that

$$d_{TV}(Q, \mathcal{P}) \le C_1 \left(\|(\mathbb{E}_Q[\psi_I(\xi)] - \mu_I)_+\| + \|\mathbb{E}_Q[\psi_E(\xi)] - \mu_E\|\right), \tag{32}$$

where $(a)_+ = \max(0, a)$ for $a \in \mathbb{R}$, and the maximum is taken componentwise when a is a vector, $\|\cdot\|$ denotes the Euclidean norm.

The lemma says that the deviation of a probability measure $Q \in \mathscr{P}$ from $\mathcal{P}$ under the total variation metric is linearly bounded by the residual of the system of equalities and inequalities defining $\mathcal{P}$. In the case when Ξ is a discrete set with finite cardinality, Lemma 2 reduces to the classical Hoffman's lemma (see [39, Theorem 7.11]).

Proof of Lemma 2. The proof is essentially derived through Shapiro's duality theorem [34, Proposition 3.1]. We proceed it in three steps.

Step 1. Let $P \in \mathscr{P}$ and $\phi(\xi)$ be P-integrable function. Let $\langle P, \phi\rangle := \mathbb{E}_P[\phi(\xi)]$. By the definition of the total variation norm (see [2]), $||P||_{TV} = \sup_{\|\phi\|_\infty \le 1} \langle P, \phi\rangle$. Moreover, by the definition of the total variation metric

$$\begin{aligned} d_{TV}(Q, \mathcal{P}) &= \inf_{P \in \mathcal{P}} d_{TV}(Q, P) \\ &= \inf_{P \in \{P : \mathbb{E}_P[\psi_E] = \mu_E, \mathbb{E}_P[\psi_I] \le \mu_I\}} \sup_{\|\phi(\xi)\|_\infty \le 1} \langle Q - P, \phi\rangle \\ &= \sup_{\|\phi(\xi)\|_\infty \le 1} \inf_{P \in \{P : \mathbb{E}_P[\psi_E] = \mu_E, \mathbb{E}_P[\psi_I] \le \mu_I\}} \langle Q - P, \phi\rangle, \end{aligned}$$

where the exchange is justified by [16, Theorem 1] because the closure of $\mathcal{P}$ is weakly compact under the tightness condition. It is easy to observe that $d_{TV}(P, \mathcal{P}) \le 2$. Moreover, under the regularity condition (31), it follows by [34, Proposition 3.4] that

$$\begin{aligned} \inf_{P \in \{P : \mathbb{E}_P[\psi_E] = \mu_E, \mathbb{E}_P[\psi_I] \le \mu_I\}} \langle Q - P, \phi\rangle &= \sup_{\lambda \in \Lambda, \lambda_0} \inf_{P \in \mathscr{M}_+} \langle Q - P, \phi\rangle + \lambda^T(\langle P, \psi\rangle - \mu) + \lambda_0(\langle P, 1\rangle - 1) \\ &= \sup_{\lambda \in \Lambda, \lambda_0} \inf_{P \in \mathscr{M}_+} \langle Q - P, \phi - \lambda^T\psi - \lambda_0\rangle + \langle Q, \lambda^T\psi + \lambda_0\rangle - \lambda^T\mu - \lambda_0, \end{aligned} \tag{33}$$

where $\psi = (\psi_E, \psi_I)$, $\mu = (\mu_E, \mu_I)$ and $\Lambda := \{(\lambda_1, \cdots, \lambda_q) : \lambda_i \ge 0, \text{ for } i = p+1, \cdots, q\}$. If there exists some ω_0 such that $\phi(\xi(\omega_0)) - \lambda^T\psi(\xi(\omega_0)) - \lambda_0 > 0$, then the right hand side of (33) is $+\infty$ because we can choose $P = \alpha\delta_{\xi(\omega_0)}(\cdot)$, where $\delta_{\xi(\omega_0)}(\cdot)$ denotes the Dirac probability measure at $\xi(\omega_0$, and drive α to $+\infty$. Thus we are left to consider that case with

$$\phi(\xi(\omega)) - \lambda^T\psi(\xi(\omega)) - \lambda_0 \le 0, \text{for a.e. } \omega \in \Omega.$$

Consequently we can rewrite (33) as

$$\begin{aligned}\inf_{P\in\{P:\mathbb{E}_P[\psi_E]=\mu_E,\mathbb{E}_P[\psi_I]\le\mu_I\}}\langle Q-P,\phi\rangle &= \inf_{P\in\mathscr{M}_+}\langle Q-P,\phi-\lambda^T\psi-\lambda_0\rangle+\langle Q,\lambda^T\psi+\lambda_0\rangle-\lambda^T\mu-\lambda_0\\ &= \langle Q,\phi-\lambda^T\psi-\lambda_0\rangle+\langle Q,\lambda^T\psi+\lambda_0\rangle-\lambda^T\mu-\lambda_0\\ &= \langle Q,\phi\rangle-\lambda^T\mu-\lambda_0.\end{aligned}$$

The second inequality is due to the fact that the optimum is attained at $P=0$. Summarizing the discussions above, we arrive at

$$\begin{aligned}d_{TV}(Q,\mathcal{P}) = &\sup_{\|\phi(\xi)\|_\infty\le 1}\ \sup_{\lambda\in\Lambda,\lambda_0}\ \langle Q,\phi\rangle-\lambda^T\mu-\lambda_0\\ &\quad\text{s.t.}\quad \phi(\xi(\omega))-\lambda^T\psi(\xi(\omega))-\lambda_0\le 0,\quad \text{a.e. } \omega\in\Omega\\ = &\sup_{\lambda\in\Lambda,\lambda_0}\ \langle Q,\min\{\lambda^T\psi(\xi(\omega))+\lambda_0,1\}\rangle-\lambda^T\mu-\lambda_0\\ &\quad\text{s.t.}\quad -1-\lambda^T\psi(\xi(\omega))-\lambda_0\le 0,\ \text{a.e. } \omega\in\Omega.\end{aligned}\tag{34}$$

Step 2. We show that the optimization problem at the right hand side of (34) has a bounded optimal solution. Let

$$\mathcal{F}:=\{(\lambda,\lambda_0)\in\Lambda\times\mathbb{R}:-1-\lambda^T\psi(\xi(\omega))-\lambda_0\le 0,\quad \text{a.e. } \omega\in\Omega\}$$

denote the feasible set of (34) and

$$\mathcal{C}:=\{(\lambda,\lambda_0)\in\Lambda\times\mathbb{R}:\|\lambda\|+|\lambda_0|=1,-\lambda^T\psi(\xi(\omega))-\lambda_0\le 0,\ \text{a.e } \omega\in\Omega\}.\tag{35}$$

Then $\mathcal{F}$ may be represented as

$$\mathcal{F}=\mathcal{F}_0+\{t\mathcal{C}:t\ge 0\}.$$

where $\mathcal{F}_0$ is a bounded convex set and the addition is in the sense of Minkowski. If $\mathcal{C}=\emptyset$, then $\mathcal{F}$ is bounded. Consequently we have

$$\begin{aligned}&\sup_{\lambda\in\Lambda,\lambda_0}\ \langle Q,\min\{\lambda^T\psi(\xi(\omega))+\lambda_0,1\}\rangle-\lambda^T\mu-\lambda_0\\ &\quad\text{s.t.}\quad -1-\lambda^T\psi(\xi(\omega))-\lambda_0\le 0,\ \text{a.e. } \omega\in\Omega\\ \le &\sup_{\lambda\in\Lambda,\lambda_0}\ \langle Q,\lambda^T\psi(\xi(\omega))\rangle-\lambda^T\mu\\ = &\sup_{\lambda\in\Lambda,\lambda_0}\ \lambda^T(\langle Q,\psi(\xi(\omega))\rangle-\mu).\end{aligned}\tag{36}$$

In what follows, we consider the case when $\mathcal{C}\neq\emptyset$. From (35) we immediately have

$$-\lambda^T\langle P,\psi(\xi(\omega))\rangle-\lambda_0\langle P,1\rangle\le 0,\forall(\lambda,\lambda_0)\in\mathcal{C},P\in\mathscr{M}_+.\tag{37}$$

On the other hand, by the Slater type condition (31),

$$(0,0_q)\in\text{int}\,[(\langle P,1\rangle-1,\langle P,\psi(\xi)\rangle-\mu,-\{0\}\times\{0_p\}\times\mathbb{R}_-^{q-p}:P\in\mathscr{M}_+].$$

Therefore there exists a closed neighborhood of $(0,0_q)$, denoted by $\mathcal{W}$, such that

$$\mathcal{W}\subset\text{int}\,[(\langle P,1\rangle-1,\langle P,\psi(\xi)\rangle-\mu),-\{0\}\times\{0_p\}\times\mathbb{R}_-^{q-p}:P\in\mathscr{M}_+].$$

Let ϵ be a small positive number and $(\lambda,\lambda_0)\in\mathcal{C}$. Then there exists a vector $\tilde{w}\in\mathcal{W}$ (depending on ϵ and (λ,λ_0)) such that $\langle w,(\lambda,\lambda_0)\rangle\le-\epsilon$. In other words, there exist $\tilde{P}\in\mathscr{M}_+$ and $\eta\in\mathbb{R}_-^{q-p}$ such that

$$\lambda_0(\langle\tilde{P},1\rangle-1)+\lambda^T(\langle\tilde{P},\psi(\xi)\rangle-\mu)-\eta^T\lambda_I\le-\epsilon.\tag{38}$$

Since $-\eta^T\lambda_I \geq 0$, we deduce from (37) and (38) that $-\lambda_0 - \lambda^T\mu \leq -\epsilon$. The inequality holds for every $(\lambda,\lambda_0)\in\mathcal{C}$. Thus, for any $(\lambda,\lambda_0)\in\mathcal{F}$, we may write it in the form

$$(\lambda,\lambda_0) = (\hat{\lambda},\hat{\lambda}_0) + t(\tilde{\lambda},\tilde{\lambda}_0),$$

where $(\hat{\lambda},\hat{\lambda}_0)\in\mathcal{F}_0$, $(\tilde{\lambda},\tilde{\lambda}_0)\in\mathcal{C}$ and $t\geq 0$. Observe that $\langle Q, \min\{\lambda^T\psi(\xi(\omega))+\lambda_0, 1\}\rangle \in [-1,1]$ and

$$-\lambda^T\mu - \lambda_0 = -\mu^T\hat{\lambda} - \hat{\lambda}_0 - t(\mu^T\tilde{\lambda}+\tilde{\lambda}_0) \leq -\mu^T\hat{\lambda} - \hat{\lambda}_0 - t\epsilon.$$

To ensure the optimal value of the optimization problem at the right hand side of (34) to be positive, we must have $-1-\mu^T\hat{\lambda}-\hat{\lambda}_0 - t\epsilon > 0$ or equivalently $t < \frac{1}{\epsilon}(1+\mu^T\hat{\lambda}+\hat{\lambda}_0)$. A sufficient condition is $t < \frac{1}{\epsilon}\left(1+\sup_{(\hat{\lambda},\hat{\lambda}_0)\in\mathcal{F}_0}(\mu^T\hat{\lambda}+\hat{\lambda}_0)\right)$. Let

$$t_1 = \frac{1}{\epsilon}\left(1+\sup_{(\hat{\lambda},\hat{\lambda}_0)\in\mathcal{F}_0}(\|\mu\|\|\hat{\lambda}\|+|\hat{\lambda}_0|)\right),$$

and $\mathcal{F}_1 := \mathcal{F}_0 + \{t\mathcal{C} : t_1 \geq t \geq 0\}$. Based on the discussions above, we conclude that the optimization problem at the right hand side of (34) has and optimal solution in $\mathcal{F}_1$.

Step 3. Let $C_1 := \max_{(\lambda,\lambda_0)\in\mathcal{F}_1}\|\lambda\|$. Then

$$\begin{aligned}
d_{TV}(Q,\mathcal{P}) &= \sup_{(\lambda,\lambda_0)\in\mathcal{F}_1} \langle Q, \min\{\lambda^T\psi(\xi(\omega))+\lambda_0,1\}\rangle - \lambda^T\mu - \lambda_0 \\
&\leq \sup_{(\lambda,\lambda_0)\in\mathcal{F}_1} \langle Q, \lambda^T\psi(\xi(\omega))+\lambda_0\rangle - \lambda^T\mu - \lambda_0 \\
&= \lambda^T(\mathbb{E}_Q[\psi(\xi)]-\mu) \\
&\leq \sum_{i=1}^{p}\lambda_i(\mathbb{E}_Q[\psi_i(\xi)]-\mu_i) + \sum_{i=p+1}^{q}\lambda_i(\mathbb{E}_Q[\psi_i(\xi)]-\mu_i) \\
&\leq \sum_{i=1}^{p}|\lambda_i||\mathbb{E}_Q[\psi_i(\xi)]-\mu_i| + \sum_{i=p+1}^{q}\lambda_i(\mathbb{E}_Q[\psi_i(\xi)]-\mu_i)_+ \\
&\leq C_1(\|(\mathbb{E}_Q[\psi_E(\xi)]-\mu_E)\| + \|(\mathbb{E}_Q[\psi_I(\xi)]-\mu_I)_+\|).
\end{aligned}$$

The inequality also holds for the case when $\mathcal{C}=\emptyset$ because $\mathcal{F}_0\subset\mathcal{F}_1$. The proof is complete. □

REMARK 3. It might be helpful to make a few comments on the proof and conditions of the lemma.

(i) An important argument we have made in the proof is that for a linear semi-infinite program with finite optimal value, there exists a bounded set of optimal solutions where the bound is determined by the feasible set.

(ii) The regularity condition (31) is a kind of Slater constraint qualification which is widely used in the literature of distributionally robust optimization. It means the underlying functions ψ in the definition of the ambiguity set through moment conditions cannot be arbitrary.

(iii) $\mathcal{P}$ is tight if there are positive constraints τ and ϱ such that

$$\sup_{P\in\mathcal{P}}\int_{\Xi}\|\xi\|^{1+\tau}P(d\xi) < \varrho. \tag{39}$$

To see this, let ϵ be any fixed small positive number and $r>1$ be sufficiently large. Then by (39)

$$\sup_{P\in\mathcal{P}}\int_{\{\xi\in\Xi:\|\xi\|\geq r\}}P(d\xi) \leq \frac{1}{r}\sup_{P\in\mathcal{P}}\int_{\{\xi\in\Xi:\|\xi\|\geq r\}}r^{1+\tau}P(d\xi) \leq \frac{1}{r}\sup_{P\in\mathcal{P}}\int_{\{\xi\in\Xi:\|\xi\|\geq r\}}\|\xi\|^{1+\tau}P(d\xi) \leq \frac{\varrho}{r} \leq \epsilon$$

Note that (39) holds trivially when Ξ is a compact set.

(iv) In the case when $\mathcal{P}$ is tight, then a sufficient condition for $\mathcal{P}$ to be closed is

$$\sup_{P\in\mathcal{P}}\int_{\Xi}|\psi_i(\xi)|^{1+\tau}P(d\xi)<\varrho, i=1,\cdots,q, \tag{40}$$

for some strictly positive number τ and positive number ϱ. To show this, let $\{P_t\}\subset\mathcal{P}$ be a sequence which converges to P weakly. Under condition (40), it follows by Lemma 1,

$$\mu_i=\lim_{t\to\infty}\mathbb{E}_{P_t}[\psi_i(\xi)]=\mathbb{E}_P[\psi_i(\xi)]$$

for $i=1,\cdots,p$. Likewise

$$\mu_i\geq\lim_{t\to\infty}\mathbb{E}_{P_t}[\psi_i(\xi)]=\mathbb{E}_P[\psi_i(\xi)]$$

for $i=p+1,\cdots,q$, which shows $P\in\mathcal{P}$.

With Lemma 2, we are able to quantify the approximation of $\mathcal{P}_N$ to $\mathcal{P}$ under the total variation metric.

Proposition 4. *Suppose that (31) holds and μ_I^N and μ_E^N converge to μ_I and μ_E respectively as $N\to\infty$. Then*

$$\mathbb{H}_{TV}(\mathcal{P}_N,\mathcal{P})\leq C_1[\max(\|(\mu_I^N-\mu_I)_+\|,\|(\mu_I-\mu_I^N)_+\|)+\|\mu_E^N-\mu_E\|], \tag{41}$$

where C_1 is defined as in Lemma 2. Moreover, if there exists a positive constant M_2 such that $||g||\leq M_2$ for all $g\in\mathcal{G}$, where $\mathcal{G}$ is defined in (10), then

$$\mathscr{H}(\mathcal{P}_N,\mathcal{P})\leq C_1M_2[\max(\|(\mu_I^N-\mu_I)_+\|,\|(\mu_I-\mu_I^N)_+\|)+\|\mu_E^N-\mu_E\|].$$

Proof. Let $Q\in\mathcal{P}_N$. By Lemma 2, there exists a positive constant C_1 such that

$$\begin{aligned}d_{TV}(Q,\mathcal{P})&\leq C_1(\|(\mathbb{E}_Q[\psi_I(\xi(\omega))]-\mu_I)_+\|+\|\mathbb{E}_Q[\psi_E(\xi(\omega))]-\mu_E\|)\\&\leq C_1(\|(\mathbb{E}_Q[\psi_I(\xi(\omega))]-\mu_I^N)_+\|+\|\mathbb{E}_Q[\psi_E(\xi(\omega))]-\mu_E^N\|+\|(\mu_I^N-\mu_I)_+\|+\|\mu_E^N-\mu_E\|)\\&=C_1(\|(\mu_I^N-\mu_I)_+\|+\|\mu_E^N-\mu_E\|).\end{aligned}$$

Therefore, $\mathbb{D}_{TV}(\mathcal{P}_N,\mathcal{P})=\sup_{Q\in\mathcal{P}_N}d_{TV}(Q,\mathcal{P})\leq C_1(\|(\mu_I^N-\mu_I)_+\|+\|\mu_E^N-\mu_E\|)$. On the other hand, since μ_I^N and μ_E^N converge to μ_I and μ_E, a regularity condition similar to (31) for the system defining $\mathcal{P}_N$ holds when N is sufficiently large. By applying Lemma 2 to the moment system defining $\mathcal{P}_N$, we have that for all $P\in\mathcal{P}$

$$\begin{aligned}d_{TV}(P,\mathcal{P}_N)&\leq C_1(\|(\mathbb{E}_P[\psi_I(\xi)]-\mu_I^N)_+\|+\|\mathbb{E}_P[\psi_E(\xi)]-\mu_E^N\|)\\&\leq C_1[\|(\mathbb{E}_P[\psi_I(\xi)]-\mu_I)_+\|+\|\mathbb{E}_P[\psi_E(\xi)]-\mu_E\|+\|(\mu_I-\mu_I^N)_+\|+\|\mu_E-\mu_E^N\|]\\&=C_1(\|(\mu_I-\mu_I^N)_+\|+\|\mu_E-\mu_E^N\|).\end{aligned}$$

and hence

$$\mathbb{D}_{TV}(\mathcal{P},\mathcal{P}_N)\leq C_1(\|(\mu_I-\mu_I^N)_+\|+\|\mu_E-\mu_E^N\|).$$

Combining the inequalities above, we have

$$\mathbb{H}_{TV}(\mathcal{P}_N,\mathcal{P})\leq C_1(\max(\|(\mu_I^N-\mu_I)_+\|,\|(\mu_I-\mu_I^N)_+\|)+\|\mu_E^N-\mu_E\|).$$

For $\mathscr{H}(\mathcal{P}_N,\mathcal{P})$, it follows by Remark 2 that $\frac{1}{M_2}\mathscr{D}(P,Q)\leq d_{TV}(P,Q)$, which implies $\mathscr{H}(\mathcal{P}_N,\mathcal{P})\leq M_2\mathbb{H}_{TV}(\mathcal{P}_N,\mathcal{P})$. Since $\mathbb{H}_{TV}(\mathcal{P}_N,\mathcal{P})\leq C_1\|\mu_N-\mu\|$,

$$\mathscr{H}(\mathcal{P}_N,\mathcal{P})\leq C_1M_2(\max(\|(\mu_I^N-\mu_I)_+\|,\|(\mu_I-\mu_I^N)_+\|)+\|\mu_E^N-\mu_E\|).$$

The proof is complete. □

In the case when μ is constructed from independent and identically distributed samples of ξ, we can show $\mathbb{H}_{TV}(\mathcal{P}_N,\mathcal{P})$ converges to zero at an exponential rate with the increase of sample size N.

COROLLARY 1. *Let ξ^j, $j=1,\cdots,N$ be independent and identically distributed sampling of ξ, $\mu_N := \frac{1}{N}\sum_{j=1}^N \psi(\xi^j)$. Assume that the conditions of Proposition 4 hold, Ξ is a compact subset of $\mathbb{R}^k$ and ψ_i, $i=1,\cdots,q$, is continuous on Ξ. Then for any $\epsilon>0$, there exist positive numbers α and β such that*

$$\text{Prob}(\mathbb{H}_{TV}(\mathcal{P}_N,\mathcal{P})\geq\epsilon)\leq \alpha e^{-\beta N}$$

for N sufficiently large. If $f(x,\xi)$ satisfies one of the conditions in Proposition 3, then

$$\text{Prob}(\mathscr{H}(\mathcal{P}_N,\mathcal{P})\geq\epsilon)\leq \alpha e^{-\beta N}$$

for N sufficiently large.

Proof. The conclusion follows from classical large deviation theorem being applied to the sample average of ψ. The rest follows from (41) and Proposition 3 since both $\mathcal{P}$ and $\mathcal{P}_N$ are compact. □

Note also that Dupačová [15] recently investigates stability of one stage distributionally robust optimization problem. She derives convergence of optimal value of distributionally robust minimization problems where the ambiguity set is constructed by sample average approximated moments and the underlying objective function is lower semicontinuous and convex w.r.t. decision variables. Assuming the random variable is defined in a finite dimensional space with compact support set, Dupačová establishes convergence of the optimal solutions and the optimal values, see [15, Theorem 2.6, Theorem 3.1 and Theorem 3.3]. It is possible to relate the results to what we have established in this paper. Indeed, if we strengthen the fourth condition in [15, Assumption 2.5] to continuity of $f(\cdot,\xi)$, we may recover [15, Theorem 3.3] through Theorem 2 without convexity of the feasible set X. Note that when the support set is compact, the ambiguity set $\mathcal{P}$ and $\mathcal{P}_N$ are weakly compact.

4.2. Mixture distribution Let $P_1,\cdots,P_L$ be a set of probability measures and

$$\mathcal{P}:=\left\{\sum_{l=1}^L \alpha_l P_l : \sum_{l=1}^L \alpha_l=1, \alpha_l\geq 0, l=1,\cdots,L\right\}.$$

In this setup, we assume that probability distributions P_l, $l=1,\cdots,L$, are known and the true probability distribution is in the convex hull of them. Robust optimization under mixture probability distribution can be traced back to Hall et al [19] and Peel and McLachlan [24]. More recently, Zhu and Fukushima [47] studied robust optimization of CVaR of a random function under mixture probability distributions.

Assume that for each P_l, one can construct P_l^N to approximate it (e.g. through samples). Let

$$\mathcal{P}_N:=\left\{\sum_{l=1}^L \alpha_l P_l^N : \sum_{l=1}^L \alpha_l=1, \alpha_l\geq 0, l=1,\cdots,L\right\}.$$

We investigate the convergence of $\mathcal{P}_N$ to $\mathcal{P}$.

PROPOSITION 5. *Assume that $\{P_l\}$ (resp. $\{P_l^N\}$), $l=1,\cdots,L$, is tight. Then $\mathcal{P}$ (resp. $\mathcal{P}_N$) is compact.*

Proof. Observe that $\mathcal{P}$ is the convex hull of a finite set $\mathcal{P}_v:=\{P_l, l=1,\ldots,L\}$, which is the image of the set under continuous mapping $F:(P_l,\cdots,P_L;\alpha_1,\cdots,\alpha_L)\to\sum_{l=1}^L\alpha_l P_l$. The image of a compact set under continuous mapping is compact. Therefore it is adequate to show that $\mathcal{P}_v$ is compact. However, the compactness of $\mathcal{P}_v$ is obvious under the tightness of P_l and finite cardinality of the set. □

Note that in the case when the random variable ξ is defined in finite dimensional space, it follows by [8, Theorem 1.4] that $P_l\in\mathcal{P}_v$ is tight.

PROPOSITION 6. *Assume that* $\|P_l^N - P_l\|_{TV} \to 0$, for $l = 1, \cdots, L$ *as* $N \to \infty$. *Then for* N *sufficiently large*

$$\mathbb{H}_{TV}(\mathcal{P}_N, \mathcal{P}) \le \max\{\|P_l^N - P_l\|_{TV} : l = 1, \cdots, L\}. \tag{42}$$

Proof. Let

$$\widetilde{\mathcal{P}} := \{P_l : l = 1, \cdots, L\} \quad \text{and} \quad \widetilde{\mathcal{P}}_N := \left\{P_l^N : l = 1, \cdots, L\right\}.$$

By [20, Proposition 2.1], $\mathbb{H}_{TV}(\mathcal{P}_N, \mathcal{P}) \le \mathbb{H}_{TV}(\widetilde{\mathcal{P}}_N, \widetilde{\mathcal{P}})$. It suffices to show that

$$\mathbb{H}_{TV}(\widetilde{\mathcal{P}}_N, \widetilde{\mathcal{P}}) \le \max\{\|P_l^N - P_l\|_{TV} : i = 1, \cdots, L\}. \tag{43}$$

Let ϵ denote the minimal distance between each pair of probability measures in $\widetilde{\mathcal{P}}$ under total variation metric, that is, $\epsilon := \min\{\|P_i - P_j\|_{TV} : i, j = 1, \cdots, L, i \ne j\}$. Let N_0 be sufficiently large such that for $N \ge N_0$,

$$\max\{\|P_l^N - P_l\|_{TV} : i = 1, \cdots, L\} \le \frac{\epsilon}{8}. \tag{44}$$

Note that, for any l,

$$\|P_l^N - P_m\|_{TV} \ge \|P_l - P_m\|_{TV} - \|P_l^N - P_l\|_{TV} \ge \frac{7}{8}\epsilon, \ \forall m = 1, \cdots, L, m \ne l.$$

By above inequality and (44), we have

$$\mathbb{D}_{TV}(P_l^N, \widetilde{\mathcal{P}}) = \min_{m \in \{1, \ldots, L\}} \|P_l^N - P_m\|_{TV} = \|P_l^N - P_l\|_{TV}$$

for $l = 1, \cdots, L$. Therefore

$$\mathbb{D}_{TV}(\widetilde{\mathcal{P}}_N, \widetilde{\mathcal{P}}) = \max\{\|P_l^N - P_l\|_{TV} : l = 1, \cdots, L\}. \tag{45}$$

Likewise, we can show

$$\mathbb{D}_{TV}(\widetilde{\mathcal{P}}, \widetilde{\mathcal{P}}_N) = \max\{\|P_l^N - P_l\|_{TV} : l = 1, \cdots, L\}. \tag{46}$$

Combining (45) and (46), we obtain (42). □

COROLLARY 2. *If* P_l^N *converges to* P_l *at an exponential rate for* $l = 1, \cdots, L$, *then* $\mathcal{P}_N$ *converges to* $\mathcal{P}$ *at the same exponential rate under the total variation metric.*

Note that convergence under the total variation metric may be too strong for some approximations. For example, if P_l^N is an empirical probability measure, then Pflug and Pichler have shown that we may end up with $\|P_l^N - P_l\|_{TV} = 1$, see [26]. The fundamental reason is that the metric is too restrictive. If we restrict h in (7) to a class of Lipschitz continuous functions with bounded modulus, then we will get a weaker metric which is known as *bounded Lipschitz metric* under which an empirical measure P_l^N would converge to P_l, see [26].The latter is related to Kantorovich metric and Fortet-Mourier metric that we discussed in Remark 2. All our technical results in the section hold under bounded Lipschitz metric.

4.3. Ambiguity set due to Delage and Ye [13] and So [33]. Delage and Ye [13] propose to construct an ambiguity set through moment conditions which consist of the mean and covariance matrix. Specifically they consider the following ambiguity set:

$$\mathcal{P}(\mu_0,\Sigma_0,\gamma_1,\gamma_2):=\left\{P\in\mathscr{P}:\ \begin{array}{l}\mathbb{E}_P[\xi-\mu_0]^T\Sigma_0^{-1}\mathbb{E}_P[\xi-\mu_0]\le\gamma_1\\ 0\preceq\mathbb{E}_P[(\xi-\mu_0)(\xi-\mu_0)^T]\preceq\gamma_2\Sigma_0\end{array}\right\},\tag{47}$$

where $\mu_0\in\mathbb{R}^k$ is the true mean vector, $\Sigma_0\in S_+^{k\times k}$ is the true covariance matrix, and γ_i, $i=1,2$ are parameters. The parameters are introduced in that the true mean value and covariance may be estimated through empirical data in data-driven problems and in these circumstances one may not be entirely confident in these estimates. The formulation allows one to construct an ambiguity set where the true mean and covariance do not have to be matched precisely and this particularly helpful when μ_0 and Σ_0 are estimated through empirical data. Note that in [13], a condition on the support set is explicitly imposed in the definition of the ambiguity set, that is, there is a closed convex set in $\mathbb{R}^k$, denoted by $\mathcal{S}$ such that $\text{Prob}\{\xi\in\mathcal{S}\}=1$. We remove this constraint as it complicates the presentation of error bounds to be discussed later on. Moreover, without this constraint, the main results in this subsection will not change.

Let $\{\xi^i\}_{i=1}^N$ be a set of N samples generated independently at random according to the distribution of ξ. Let

$$\mu_N:=\frac{1}{N}\sum_{i=1}^N\xi^i\quad\text{and}\quad\Sigma_N:=\frac{1}{N}\sum_{i=1}^N(\xi^i-\mu_N)(\xi^i-\mu_N)^T.$$

Delage and Ye [13] propose to construct an approximation of $\mathcal{P}$ with the ambiguity set $\mathcal{P}(\mu_N,\Sigma_N,\gamma_1^N,\gamma_2^N)$ by replacing the true mean and covariance μ_0 and Σ_0 with their sample average approximation μ_N and Σ_N, where γ_1^N and γ_2^N are some positive constants depending on the sample. By assuming that there exists a positive number $\hat{R}<\infty$ such that

$$P\{(\xi-\mu_0)^T\Sigma_0^{-1}(\xi-\mu_0)\le\hat{R}^2\}=1,\tag{48}$$

they prove that the true distribution of ξ lies in set $\mathcal{P}(\mu_N,\Sigma_N,\gamma_1^N,\gamma_2^N)$ with probability at least $1-\delta$, see [13, Corollaries 3 and 4]. The condition implies the support set of ξ is bounded. So [33] observes that the condition may be weakened to the following moment growth condition:

$$\mathbb{E}_P[\|\Sigma_0^{1/2}(\xi-\mu_0)\|_2^p]\le(cp)^{p/2},\tag{49}$$

where c and p are some parameters defined in [33]. Specifically, by setting

$$\gamma_1^N:=\frac{t_m^N}{1-t_c^N-t_m^N},\quad\text{and}\quad\gamma_2^N:=\frac{1+t_m^N}{1-t_c^N-t_m^N},$$

where

$$t_m^N:=\frac{4ce^2\ln^2(2/\delta)}{N},\quad t_c^N:=\frac{4c'(2e/3)^{3/2}\ln^{3/2}(4h/\delta)}{\sqrt{N}},$$

$\delta\in(0,2e^{-3})$, c is a constant and $p\ge1$, he shows that the true distribution of ξ lies in $\mathcal{P}_N$ with probability $1-\delta$ for N is sufficiently large, where

$$\mathcal{P}_N:=\mathcal{P}(\mu_N,\Sigma_N,\gamma_1^N,\gamma_2^N).\tag{50}$$

See [33, Theorem 9]. The significance of So's new results lies not only in the fact condition (49) is strictly weaker than (48) but the parameters γ_1^N and γ_2^N depend merely on the sample size N rather than the sample as in [13]. The latter will simplify our discussions later on.

Note that as the sample size $N \to \infty$, it follows from law of large numbers that the iid sampling ensures $\mu_N \to \mu_0$, $\Sigma_N \to \Sigma_0$, $\gamma_1^N \downarrow 0$ and $\gamma_2^N \downarrow 1$ w.p.1. We predict that $\mathcal{P}_N$ converges to the following ambiguity set w.p.1:

$$\mathcal{P} := \left\{ P \in \mathscr{P} : \begin{array}{l} \mathbb{E}_P[\xi - \mu_0]^T \Sigma_0^{-1} \mathbb{E}_P[\xi - \mu_0] \le 0 \\ 0 \preceq \mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)^T] \preceq \Sigma_0 \end{array} \right\}. \tag{51}$$

Before proceeding to convergence analysis of $\mathcal{P}_N$ to $\mathcal{P}$, we note that both sets are compact in weak topology under some circumstances.

PROPOSITION 7. *Both $\mathcal{P}_N$ and $\mathcal{P}$ are tight. Moreover, they are closed (and hence compact in the weak topology) if one of the following conditions holds.*

(a) For $\mathcal{P}_N$ (resp. $\mathcal{P}$), there exists a positive number ε such that

$$\sup_{P \in \mathcal{P}_N} \int_\Xi \|\xi\|^{2+\varepsilon} P(d\xi) < \infty. \tag{52}$$

(b) There exists a compact set $\mathcal{S} \supseteq \Xi$ such that $P\{\xi \in \mathcal{S}\} = 1$ for all $P \in \mathcal{P}_N$ (resp. $P \in \mathcal{P}$).

Proof. We only prove the conclusion for $\mathcal{P}_N$ as the proof for $\mathcal{P}$ is identical.
Tightness. The second inequality in the definition of $\mathcal{P}_N$ implies

$$\sup_{P \in \mathcal{P}_N} \int_\Xi \|\xi\|^2 P(d\xi) < \infty$$

which yields, through Lemma 1, that

$$\lim_{r \to \infty} \sup_{P \in \mathcal{P}_N} \int_{\{\xi \in \Xi : \|\xi\| \ge r\}} \|\xi\| P(d\xi) = 0.$$

Therefore

$$0 \le \lim_{r \to \infty} \sup_{P \in \mathcal{P}_N} \int_{\{\xi \in \Xi : \|\xi\| \ge r\}} P(d\xi) \le \lim_{r \to \infty} \sup_{P \in \mathcal{P}_N} \int_{\{\xi \in \Xi : \|\xi\| \ge r\}} \|\xi\| P(d\xi) = 0.$$

By [2, Definition 9.2.2], $\mathcal{P}_N$ is tight.
Closedness. Let us prove the closedness under condition (a) first. Let $\{P_t\} \in \mathcal{P}_N$ and $P_t \to P^*$ weakly. Under the bounded integral condition (52), it follows from Lemma 1 that $\{P_t h(\cdot)\}$ is uniformly integrable, where $h(\cdot)$ denotes the inverse mapping of $\|\xi\|^2$. Indeed

$$0 \le \lim_{r \to \infty} \sup_{P \in \mathcal{P}_N} \int_{\{\xi \in \Xi : \|\xi\| \ge r\}} \|\xi\|^2 P(d\xi) \le \lim_{r \to \infty} \frac{1}{r^\varepsilon} \sup_{P \in \mathcal{P}_N} \int_{\{\xi \in \Xi : \|\xi\| \ge r\}} \|\xi\|^{2+\varepsilon} P(d\xi) = 0.$$

The uniform integrability and the weak convergence yield

$$\lim_{t \to \infty} \int_\Xi \|\xi\|^2 P_t(d\xi) = \int_\Xi \|\xi\|^2 P^*(d\xi),$$

which ensures

$$\lim_{t \to \infty} \mathbb{E}_{P_t}[\xi - \mu_N]^T \Sigma_N^{-1} \mathbb{E}_{P_t}[\xi - \mu_N] = \mathbb{E}_{P^*}[\xi - \mu_N]^T \Sigma_N^{-1} \mathbb{E}_{P^*}[\xi - \mu_N] \le \gamma_1^N$$

and

$$\lim_{t \to \infty} \mathbb{E}_{P_t}[(\xi - \mu_N)(\xi - \mu_N)^T] = \mathbb{E}_{P^*}[(\xi - \mu_N)(\xi - \mu_N)^T] \preceq \Sigma_N.$$

This shows $P^* \in \mathcal{P}_N$ and hence the closedness of $\mathcal{P}_N$.

Now we prove the closedness under condition (b). This is obvious because the compactness of $\mathcal{S}$ implies (52). □

Proposition 7 gives sufficient conditions for weak compactness of the ambiguity sets $\mathcal{P}$ and $\mathcal{P}_N$. It is possible to derive some weaker conditions which ensure the closedness, e.g., by adjusting the values of some parameters in the definition of the ambiguity sets. In the case when neither condition (i) nor condition (ii) is satisfied, the ambiguity set is not necessarily weakly compact. In these circumstances (the lack of closedness), we may consider the closure of the ambiguity set.

4.3.1. Error bound. We now turn to estimate $\mathbb{H}_{TV}(\mathcal{P},\mathcal{P}_N)$. The first step is to express $\mathcal{P}$ and $\mathcal{P}_N$ through a linear system of $\mathbb{E}_P[\cdot]$. To this end, we note that inequality

$$\mathbb{E}_P[\xi-\mu_0]^T\Sigma_0^{-1}\mathbb{E}_P[\xi-\mu_0]\le 0$$

is equivalent to

$$\mathbb{E}_P\left[\begin{pmatrix}-\Sigma_0 & \mu_0-\xi\\ (\mu_0-\xi)^T & 0\end{pmatrix}\right]\preceq 0,$$

where notation $M\preceq 0$ means matrix M is negative semidefinite (we will come back to this shortly). Likewise

$$\mathbb{E}_P[\xi-\mu_N]^T\Sigma_N^{-1}\mathbb{E}_P[\xi-\mu_N]\le \gamma_1^N \tag{53}$$

can be written as

$$\mathbb{E}_P\left[\begin{pmatrix}-\Sigma_N & \mu_N-\xi\\ (\mu_N-\xi)^T & -\gamma_1^N\end{pmatrix}\right]\preceq 0.$$

Consequently, we can rewrite $\mathcal{P}$ and $\mathcal{P}_N$ as

$$\mathcal{P}=\left\{P\in\mathscr{P}:\ \begin{array}{l}\mathbb{E}_P\left[\begin{pmatrix}-\Sigma_0 & \mu_0-\xi\\ (\mu_0-\xi)^T & 0\end{pmatrix}\right]\preceq 0\\ \mathbb{E}_P[(\xi-\mu_0)(\xi-\mu_0)^T]\preceq\Sigma_0\end{array}\right\} \tag{54}$$

and

$$\mathcal{P}_N=\left\{P\in\mathscr{P}:\ \begin{array}{l}\mathbb{E}_P\left[\begin{pmatrix}-\Sigma_N & \mu_N-\xi\\ (\mu_N-\xi)^T & -\gamma_1^N\end{pmatrix}\right]\preceq 0\\ \mathbb{E}_P[(\xi-\mu_N)(\xi-\mu_N)^T]\preceq\gamma_2^N\Sigma_N\end{array}\right\} \tag{55}$$

respectively.

We need some preliminary notation and results in matrix to proceed the rest of discussions. Recall that for two real matrices $A,B\in\mathbb{R}^{n\times n}$, the Frobenius product of A and B is defined as the trace of A^TB. The Frobenius norm of A, denoted by $\|A\|_F$, is the square root of the trace of A^TA. Let $M\in\mathbb{R}^{n\times n}$ be a real symmetric matrix and $\{\iota_i\}_{i=1}^n$ be the set of eigenvalue of M. Let $Q\mathrm{diag}\{\iota_1,\ldots,\iota_n\}Q^T$ be the spectral decomposition of M with Q being an orthogonal matrix. We define $M_+:=Q\mathrm{diag}\{\max\{\iota_1,0\},\ldots,\max\{\iota_n,0\}\}Q^T$ and $M_-:=M-M_+$. This is motivated by the need to quantify the violation of the semidefinite constraint $M\preceq 0$, where $M\preceq 0$ means matrix M is negative semidefinite and $M\succeq 0$ means M is positive semidefinite. Clearly, if M is a negative semidefinite matrix, then $M_+=0$. Moreover, it is easy to observe that $M\preceq M_+$.

LEMMA 3. *Let $A,B\in S^{n\times n}$ and $A\succeq 0$. The following assertion hold.*
(i) $\mathrm{tr}(AB)\le\mathrm{tr}(AB_+)$.
(ii) $\|(A+B)_+\|_F\le\|A_+\|_F+\|B_+\|_F$.

Proof. Part (i). Since $B_+ - B \succeq 0$ and $A \succeq 0$, by [10, Example 2.24], $\mathrm{tr}(A(B_+ - B)) \geq 0$ and

$$\mathrm{tr}(AB_+) - \mathrm{tr}(AB) = \mathrm{tr}(A(B_+ - B)) \geq 0.$$

The conclusion follows.

Part (ii). Let $M \in S^{n\times n}$. It is well known that for $X \preceq 0$, $\|X - M\|_F$ attaints its minimum when $X = M_-$, see [21]. Using this argument, we have

$$\begin{aligned}\|(A+B)_+\|_F &= \|A+B-(A+B)_-\|_F \leq \|A+B-(A_-+B_-)\|_F \\ &= \|A_+ + B_+\|_F \leq \|A_+\|_F + \|B_+\|_F.\end{aligned}$$

This completes the proof. □

We are now ready to return to our discussion on estimation of $\mathbb{H}_{TV}(\mathcal{P}, \mathcal{P}_N)$. We do so by deriving an error bound for $d_{TV}(Q, \mathcal{P})$ and $d_{TV}(Q, \mathcal{P}_N)$ in the first place.

THEOREM 3. *Let $\mathcal{P}$ be defined as in (54) and $\mathcal{P}_N$ by (55) respectively. Assume that the true distribution of ξ is continuous with convex support set Ξ. Then the following assertions hold.*

(i) There exists a positive constant C_2 depending on $\mathcal{P}$ such that

$$d_{TV}(Q, \mathcal{P}) \leq C_2 \left(\|\mathbb{E}_Q[(\xi - \mu_0)]\| + \|(\mathbb{E}_Q[(\xi - \mu_0)(\xi - \mu_0)^T] - \Sigma_0)_+\|_F \right) \tag{56}$$

for every $Q \in \mathscr{P}$.

(ii) There exists a positive constant C_3 such that for every $Q \in \mathscr{P}$ with $\|\mathbb{E}_Q[\xi]\| \leq \eta$

$$\begin{aligned} d_{TV}(Q, \mathcal{P}_N) \leq C_3 \Bigg(&\left\| \left(\mathbb{E}_Q \left[\begin{pmatrix} -\Sigma_0 & \mu_0 - \xi \\ (\mu_0 - \xi)^T & 0 \end{pmatrix} \right] \right)_+ \right\|_F + ||\Sigma_0 - \Sigma_N||_F + \gamma_1^N \\ &+ \|(\mathbb{E}_Q[(\xi - \mu_0)(\xi - \mu_0)^T] - \gamma_2^N \Sigma_N)_+\|_F + 2||\mu_0 - \mu_N|| \Bigg) \end{aligned} \tag{57}$$

with probability approaching one when $N \to \infty$.

Proof. Part (i). To ease the notation, let

$$\psi_1(\xi) := \xi - \mu_0 \quad \text{and} \quad \psi_2(\xi) := (\xi - \mu_0)(\xi - \mu_0)^T.$$

Then the ambiguity set defined in (54) can be equivalently written as

$$\mathcal{P} = \left\{ P \in \mathscr{P} : \begin{array}{l} \mathbb{E}_P[\psi_1(\xi)] = 0 \\ \mathbb{E}_P[\psi_2(\xi)] \preceq \Sigma_0 \end{array} \right\}.$$

Let $\psi(\xi) := (1, \psi_1(\xi), \psi_2(\xi))$ and $u = (1, 0, \Sigma_0)$. Since the true distribution of ξ is continuous with convex support set, then the mean value of ξ is located in the interior of Ξ, the support set of ξ. Using the definition of positive definiteness of a matrix, we can easily show that $\Sigma_0 \succ 0_{k\times k}$. Let $\mathscr{M}_+$ denote the set of all positive measures of Ξ and $\mathcal{K} := \{0\} \times \{0_k\} \times S_-^{k\times k}$. We want to show

$$(0, 0_k, 0_{k\times k}) \in \mathrm{int}\, \{\mathbb{E}_P[1] - 1, \mathbb{E}_P[\psi(\xi)] - u - \mathcal{K} : P \in \mathscr{M}_+\}, \tag{58}$$

where for two sets A and B, $A - B$ stands for the Minkowski difference. For $\xi \in \Xi$, let $y := \xi - \mu_0$ and P_y denote the Dirac probability measure at ξ, that is, $P_y(\{\xi\}) = 1$. Then $\mathbb{E}_{P_y}[\psi(\xi)] = (1, y, yy^T)$. Let λ_i, $i = 1, \cdots, n$, denote the i-th eigenvalue of Σ_0 and q_i the corresponding eigenvector. For each i

$$(1, 0, \lambda_i q_i q_i^T) \in \mathrm{cone}\, \{(1, y, yy^T) : y \in \Xi - \mu_0\}$$

and hence $u = (1, 0, \Sigma_0) = (1, 0, \sum_{i=1}^n \lambda_i q_i q_i^T) \in \text{int cone}\,\{(y, yy^T) : y \in \Xi - \mu_0\}$. On the other hand, it is easy to show $\{\mathbb{E}_P[\psi(\xi)] : P \in \mathscr{M}_+\} = \text{cone}\{(y, yy^T) : y \in \Xi - \mu_0\}$. Therefore

$$u = (0, \Sigma_0) \in \text{int}\,\{\mathbb{E}_P[\psi(\xi)] : P \in \mathscr{M}_+\},$$

which implies (58) because $\text{int}\,\{\mathbb{E}_P[\psi(\xi)] : P \in \mathscr{M}_+\} \subset \text{int}\,(\{\mathbb{E}_P[\psi(\xi)] : P \in \mathscr{M}_+\} - \mathcal{K})$.

The rest of proof is similar to that of Lemma 2 except that ψ_2 is a matrix. Let $P \in \mathscr{P}$ and $\phi(\xi)$ be P-integrable componentwise. Recall that $\langle P, \phi\rangle = \mathbb{E}_P[\phi(\xi)]$ and $||P||_{TV} = \sup_{\|\phi\|_\infty \le 1}\langle P, \phi\rangle$. Through a similar argument to that of Step 1 in Lemma 2, we have

$$\begin{aligned} d_{TV}(Q, \mathcal{P}) &= \inf_{P\in\mathcal{P}} d_{TV}(Q, P) = \inf_{P\in\{P:\mathbb{E}_P[\psi(\xi)]\preceq u\}} \sup_{\|\phi(\xi)\|_\infty \le 1} \langle Q - P, \phi\rangle \\ &= \sup_{\|\phi(\xi)\|_\infty \le 1} \inf_{P\in\{P:\mathbb{E}_P[\psi(\xi)]\preceq u\}} \langle Q - P, \phi\rangle. \end{aligned}$$

Let $\lambda_1 \in \mathbb{R}^k$ and $\Gamma_2 \in S_+^{k\times k}$ denote the dual variables of constraints $\mathbb{E}_P[\psi(\xi)] \preceq u$, let

$$\Lambda := \{(\lambda_1, \Gamma_2) : \lambda_1 \in \mathbb{R}^k,\, \Gamma_2 \in S_+^{k\times k}\}.$$

Under the regularity condition (58), it follows by [34, Proposition 3.4] that

$$\inf_{P\in\{P:\mathbb{E}_P[\psi(\xi)]\preceq u\}} \langle Q - P, \phi\rangle = \sup_{(\lambda_1,\Gamma_2)\in\Lambda, \phi = \Gamma\bullet\psi(\xi)} [\Gamma \bullet (\mathbb{E}_Q[\psi(\xi)] - u)],$$

where $\Gamma \bullet \psi(\xi) := \lambda_1^T\psi_1(\xi) + \Gamma_2 \bullet \psi_2(\xi)$ and

$$\Gamma \bullet (\mathbb{E}_Q[\psi(\xi)] - u) = \lambda_1^T\mathbb{E}_Q[\psi_1(\xi)] + \Gamma_2 \bullet (\mathbb{E}_Q[\psi_2(\xi)] - \Sigma_0).$$

Consequently we arrive at

$$d_{TV}(Q, \mathcal{P}) = \sup_{(\lambda_1,\Gamma_2)\in\Lambda, \|\Gamma\bullet\psi(\xi)\|_\infty \le 1} \Gamma \bullet (\mathbb{E}_Q[\psi(\xi)] - u). \tag{59}$$

The rest of the proof amounts to estimate (59). Since the Frobenius inner product of two matrices is the sum of the scalar product of vectors, problem (59) is a linear program with linear semi-infinite constraints. Using a similar argument to that in the proof of Lemma 2, it is not difficult to show that the optimum is achieved in a bounded set. Let $(\lambda_1^*, \Gamma_2^*)$ denote the corresponding optimal solution. Then

$$d_{TV}(Q, \mathcal{P}) = \lambda_1^{*T}\mathbb{E}_Q[(\xi - \mu_0)] + \Gamma_2^* \bullet (\mathbb{E}_Q[(\xi - \mu_0)(\xi - \mu_0)^T] - \Sigma_0).$$

Let C_2 denote the maximum F-norm of all Γ from the bounded set. Then the right hand side of the equation above is bounded by $C_2\,(\|\mathbb{E}_Q[\xi] - \mu_0\| + \|(\mathbb{E}_Q[(\xi - \mu_0)(\xi - \mu_0)^T] - \Sigma_0)_+\|_F)$ and hence (56) follows.

Part (ii). Consider $\mathcal{P}_N$. Let

$$\tilde{\psi}_1(\xi) := \begin{pmatrix} -\Sigma_0 & \mu_0 - \xi \\ (\mu_0 - \xi)^T & 0 \end{pmatrix} \quad \text{and} \quad \tilde{\psi}_2(\xi) := (\xi - \mu_0)(\xi - \mu_0)^T.$$

Let

$$\tau_N^1 := \begin{pmatrix} \Sigma_0 - \Sigma_N & \mu_N - \mu_0 \\ (\mu_N - \mu_0)^T & -\gamma_1^N \end{pmatrix}$$

and

$$\tau_N^2(Q) := (\mathbb{E}_Q[\xi] - \mu_0)(\mu_0 - \mu_N)^T + (\mu_0 - \mu_N)(\mathbb{E}_Q[\xi] - \mu_N)^T.$$

Then

$$\mathbb{E}_Q\left[\begin{pmatrix} -\Sigma_N & \mu_N-\xi \\ (\mu_N-\xi)^T & -\gamma_1^N \end{pmatrix}\right] = \mathbb{E}_Q[\tilde{\psi}_1(\xi)] + \tau_N^1$$

and

$$\mathbb{E}_Q[(\xi-\mu_N)(\xi-\mu_N)^T] = \mathbb{E}_Q[\tilde{\psi}_2(\xi)] + \tau_N^2(Q).$$

Note that by assumption $\|\mathbb{E}_Q[\xi]\| \le \eta$. Therefore there exists a positive constant C_4 such that

$$\|\tau_N^1\|_F \le C_4(\|\mu_0-\mu_N\| + \|\Sigma_0-\Sigma_N\| + \gamma_1^N)$$

and w.p.1 $\|\tau_N^2(Q)\|_F \le C_4\|\mu_0-\mu_N\|$. The second inequality implicitly uses the law of large numbers to ensure μ_N converges to μ_0 almost surely as N goes to infinity and hence μ_N is bounded w.p.1. Let $\tilde{\psi}(\xi) := (1, \tilde{\psi}_1(\xi), \tilde{\psi}_2(\xi))$, $\bar{u}_1^N = -(\tau_N^1)$, $\bar{u}_2^N = \gamma_2^N\Sigma_N - \tau_N^2(Q)$ and $\bar{u}^N := (1, \bar{u}_1^N, \bar{u}_2^N)$. Based on the discussions above, we can present $\mathcal{P}_N$ as $\mathcal{P}_N = \{P \in \mathscr{P} : \mathbb{E}_P[\psi(\xi)] \preceq \bar{u}^N\}$. A clear benefit of the presentation is that in the system $\mathbb{E}_P[\psi(\omega)] \preceq \bar{u}^N$, only the right hand side depends on N and this will facilitate us to derive an error bound analogous to Lemma 2.

We first prove the Slater type regularity condition. Note that $\mu_N \to \mu_0$ and $\Sigma_N \to \Sigma_0$. Thus when N is sufficiently large, $\mu_N \in \operatorname{int}\Xi$ and $\Sigma_N \succ 0_{k\times k}$, which implies $\Sigma_N \in \operatorname{int} S_+^{k\times k}$. For fixed N, let $\sum_{j=1}^k \lambda_j q_j q_j^T$ be the spectral decomposition of Σ_N. Let σ be a positive number such that $\xi_i := \mu_N + \sigma\sqrt{\lambda_i}q_i \in \Xi$ for $i = 1, \cdots, k$, and $\xi_i := \mu_N - \sigma\sqrt{\lambda_{i-k}}q_{i-k} \in \Xi$ for $i = k+1, \cdots, 2k$. Let P_i denote the Diract probability measure at ξ_i and $\hat{P} := \frac{1}{2\sigma^2}\sum_{i=1}^{2k} P_i$. We can make the following three claims. (a) $\hat{P} \in \mathscr{M}_+$; (b) $\mathbb{E}_{\hat{P}}[\tilde{\psi}_2(\xi)] - \bar{u}_2^N = (1-\gamma_2^N)\Sigma_N \prec 0$ and (c) $\mathbb{E}_{\hat{P}}[\tilde{\psi}_1(\xi)] - \bar{u}_1^N \prec 0$ bcause $-\Sigma_N \prec 0$, $\mathbb{E}_{\hat{P}}[\mu_0 - \xi] = 0$, and hence its Schur complement, namely $-\gamma_1^N + \mathbb{E}_{\hat{P}}[\mu_N-\xi]^T\Sigma_N^{-1}\mathbb{E}_{\hat{P}}[\mu_N - \xi] = -\gamma_1^N < 0$. The claims immediately indicate

$$(0, 0_{(k+1)\times(k+1)}, 0_{k\times k}) \in \operatorname{int}\left\{\mathbb{E}_P[\tilde{\psi}(\xi)] - \mu_N + \{0\} \times S_+^{(k+1)\times(k+1)} \times S_+^{k\times k} : P \in \mathscr{M}_+\right\}. \tag{60}$$

The rest of proof is similar to part (i). Specifically

$$d_{TV}(Q, \mathcal{P}_N) = \sup_{(\Gamma_1,\Gamma_2)\in\Lambda_1, \|\tilde{\psi}_{(\Gamma_1,\Gamma_2)}(\xi)\|_\infty \le 1} \Gamma \bullet (\mathbb{E}_Q[\tilde{\psi}(\xi)] - \bar{u}^N), \tag{61}$$

where

$$\Gamma \bullet (\mathbb{E}_Q[\tilde{\psi}(\xi)] - \bar{u}^N) = \Gamma_1 \bullet (\mathbb{E}_Q[\tilde{\psi}_1(\xi)] + \tau_N^1)) + \Gamma_2 \bullet (\mathbb{E}_Q[\tilde{\psi}_2(\xi)] - \gamma_2^N\Sigma_N + \tau_N^2(Q)),$$

$$\psi_{(\Gamma_1,\Gamma_2)}(\xi) := \tilde{\psi}_1(\xi) \bullet \Gamma_1 + \tilde{\psi}_2(\xi) \bullet \Gamma_2,$$

and

$$\Lambda_1 := \{(\Gamma_1, \Gamma_2) : \Gamma_1 \in S_+^{(k+1)\times(k+1)}, \Gamma_2 \in S_+^{k\times k}\}.$$

Following a similar argument to Part (i) of the proof, we can show that there exists $\tilde{C}_2 \ge \|\Gamma^*\|_F$ where Γ^* is an optimal solution of (61) and (57) holds with $C_3 := \tilde{C}_2 C_4$. □

Theorem 3 gives a bound for a probability measure Q deviating from $\mathcal{P}$ and $\mathcal{P}_N$ respectively in terms of the residual of the linear inequality systems defining the ambiguity sets. With the error bounds established in Theorem 3, we are ready to give an upper bound for the Hausdorff distance between $\mathcal{P}$ and $\mathcal{P}_N$ under the total variation metric.

THEOREM 4. *Under the setting and conditions of Theorem 3, the following assertions hold.*
(i) There exists a positive constant C_5 such that

$$\begin{aligned}\mathbb{H}_{TV}(\mathcal{P}, \mathcal{P}_N) \le C_5(&\max\{\|(\gamma_2^N\Sigma_N - \Sigma_0)_+\|_F, \|(\Sigma_0 - \gamma_2^N\Sigma_N)_+\|_F\} + 2\|\mu_0-\mu_N\| \\ &+\gamma_1^N + (\gamma_1^N)^{1/2} + \|\Sigma_0-\Sigma_N\|)\end{aligned} \tag{62}$$

with probability approaching 1 when $N \to \infty$.

(ii) Let

$$M_i(t) := \mathbb{E}\left[e^{t[\xi_i - \mathbb{E}[\xi_i]]}\right] \text{ and } M_{ij}(t) := \mathbb{E}\left[e^{t(\xi_i\xi_j - \mathbb{E}[\xi_i\xi_j])}\right]$$

denote the moment generating function of the random variable $\xi_i - \mathbb{E}[\xi_i]$, for $i = 1, \cdots, k$, and random variable $\xi_i\xi_j - \mathbb{E}[\xi_i\xi_j]$, for $i, j = 1, \ldots, k$. If $M_i(t)$ and $M_{ij}(t)$ are finite valued for all t close to 0, then for any $\epsilon > 0$, there exist positive constants $\alpha(\epsilon)$ and $\beta(\epsilon)$ such that

$$\text{Prob}(\mathbb{H}_{TV}(\mathcal{P}, \mathcal{P}_N) \geq \epsilon) \leq \alpha(\epsilon)e^{-\beta(\epsilon)N} \tag{63}$$

when N is sufficiently large.

Proof. Part (i). By the definition of $\mathbb{H}_{TV}$, we show that

$$\max\left\{\sup_{Q\in\mathcal{P}_N} d_{TV}(Q, \mathcal{P}), \sup_{Q\in\mathcal{P}} d_{TV}(Q, \mathcal{P}_N)\right\}$$

is bounded by the term at the right hand side of (62). Let $\bar{\lambda}_N$ and $\bar{\lambda}_0$ be the maximize eigenvalue of Σ_N and Σ_0,

$$\tau_Q^N := (\mu_N - \mu_0)(\mathbb{E}_Q[\xi] - \mu_N)^T + (\mathbb{E}_Q[\xi] - \mu_0)(\mu_N - \mu_0)^T.$$

Moreover, since $\mu_N \to \mu_0$, $\Sigma_N \to \Sigma_0$, $\bar{\lambda}_N \to \bar{\lambda}_0$ and γ_1^N is close to 0 with probability approaching 1 at exponential rate as $N \to \infty$, it follows via inequality (53) that both $\mathbb{E}_Q[\xi] - \mu_N$ and $\mathbb{E}_Q[\xi] - \mu_0$ are bounded w.p.1 for $Q \in \mathcal{P}_N$. then we claim that there exist positive numbers C_6 and ϱ such that $\|(\tau_Q^N)_+\|_F \leq C_6\|\mu_N - \mu_0\|$ and $\bar{\lambda}_N \leq \bar{\lambda}_0 + \varrho \leq \frac{1}{k^2}C_6^2$ with probability close to 1 when $N \to \infty$. Next, using the notation τ_Q^N, we have, through a simple rearrangement

$$\mathbb{E}_Q[\xi - \mu_0] = \mathbb{E}_Q[\xi - \mu_N] + (\mu_N - \mu_0)$$

and

$$\mathbb{E}_Q[(\xi - \mu_0)(\xi - \mu_0)^T] = \mathbb{E}_Q[(\xi - \mu_N)(\xi - \mu_N)^T] + \tau_Q^N.$$

Let $Q \in \mathcal{P}_N$. Since $\bar{\lambda}_N \leq \frac{1}{k^2}C_6^2$ for N sufficiently large, $\mathbb{E}_Q[(\xi - \mu_0)]^T\Sigma_N^{-1}\mathbb{E}_Q[(\xi - \mu_0)] \leq \gamma_1^N$ implies $\|\mathbb{E}_Q[\xi - \mu_N]\| \leq k((\lambda_0 + \varrho)\gamma_1^N)^{1/2} \leq C_6(\gamma_1^N)^{1/2}$. By Theorem 3 (i) and Lemma 3 (ii),

$$\begin{aligned} d_{TV}(Q, \mathcal{P}) &\leq C_2\left(\left\|(\mathbb{E}_Q[\xi - \mu_N] + (\mu_N - \mu_0))_+\right\|\right. \\ &\quad \left. + \|(\mathbb{E}_Q[(\xi - \mu_N)(\xi - \mu_N)^T] + \tau_Q^N - \gamma_2^N\Sigma_N)_+\|_F + \|(\gamma_2^N\Sigma_N - \Sigma_0)_+\|_F\right) \\ &\leq C_2((C_6 + 1)\|\mu_N - \mu_0\| + C_6(\gamma_1^N)^{1/2} + \|\Sigma_0 - \Sigma_N\|_F + \|(\gamma_2^N\Sigma_N - \Sigma_0)_+\|_F). \end{aligned} \tag{64}$$

The second inequality holds because the second term at the right hand side of the first inequality are bounded by $\|(\tau_Q^N)_+\|_F$. Likewise, for $Q \in \mathcal{P}$, it follows by Theorem 3 (ii) and Lemma 3 (ii)

$$\begin{aligned} d_{TV}(Q, \mathcal{P}_N) &\leq C_3\left(\left\|\left(\mathbb{E}_Q\left[\begin{pmatrix} -\Sigma_0 & \mu_0 - \xi \\ (\mu_0 - \xi)^T & 0 \end{pmatrix}\right]\right)_+\right\|_F + ||\Sigma_0 - \Sigma_N||_F + \gamma_1^N\right. \\ &\quad \left. + \|(\mathbb{E}_Q[(\xi - \mu_0)(\xi - \mu_0)^T] - \gamma_2^N\Sigma_N)_+\|_F + 2||\mu_0 - \mu_N||\right) \\ &\leq C_3(||\Sigma_0 - \Sigma_N||_F + \gamma_1^N + \|(\mathbb{E}_Q[(\xi - \mu_0)(\xi - \mu_0)^T] - \Sigma_0 + \Sigma_0 - \gamma_2^N\Sigma_N)_+\|_F \\ &\quad + 2\|(\mu_0 - \mu_N)\|) \\ &\leq C_3(||\Sigma_0 - \Sigma_N||_F + \gamma_1^N + \|(\Sigma_0 - \gamma_2^N\Sigma_N)_+\|_F + 2\|(\mu_0 - \mu_N)\|), \end{aligned} \tag{65}$$

where the second inequality follows from the fact that $\mathbb{E}_Q[\xi - \mu_0]^T\Sigma_0^{-1}\mathbb{E}_Q[\xi - \mu_0] = 0$ implies

$$\mathbb{E}_Q\left[\begin{pmatrix} -\Sigma_0 & \mu_0 - \xi \\ (\mu_0 - \xi)^T & 0 \end{pmatrix}\right] \preceq 0.$$

Combining (64) and (65), we obtain (62) with $C_5 := \max\{C_2(C_6 + 1), C_3\}$.

Part (ii). Under the moment conditions, it follows from Cramér's Large Deviation Theorem that $\|\frac{1}{N}\sum_{l=1}^{N}\xi_i^l - \mathbb{E}[\xi_i]\| \to 0$ at an exponential rate with increase of sample size N for $i = 1, \cdots, k$ and hence $\|\mu_N - \mu_0\|$ converges to 0 at the exponential rate. Likewise, $\|\frac{1}{N}\sum_{l=1}^{N}\xi_i^l\xi_j^l - \mathbb{E}[\xi_i\xi_j]\| \to 0$ at an exponential rate with increase of sample size N for $i, j = 1, \cdots, k$, which imply that $\|\Sigma_N - \Sigma_0\| \to 0$ and $\|\Sigma_N\|_F \le 2\|\Sigma_0\|_F$ with probability approaching one when N is sufficiently large. Observe that $\gamma_1^N = O(1/N)$ and $\gamma_2^N - 1 = O(1/\sqrt{N})$. On the other hand, it follows by Lemma 3 (b)

$$\|(\gamma_2^N\Sigma_N - \Sigma_0)_+\|_F \le |\gamma_2^N - 1|\|(\Sigma_N)_+\|_F + \|(\Sigma_N - \Sigma_0)_+\|_F$$

and

$$\|(\Sigma_0 - \gamma_2^N\Sigma_N)_+\|_F \le \|(\Sigma_0 - \Sigma_N)_+\|_F + |\gamma_2^N - 1|\|(\Sigma_N)_+\|_F$$

Combing the inequalities with the discussions above, we conclude that the right hand side of (62) converges to zero at an exponential rate as $N \to \infty$. We omit the details as it is a standard argument. □

Remark 4. A few comments about Theorem 4 are in sequel. (i) In this theorem, $\mathcal{P}$ is not a singleton in general which means an increase of the sample size N will not eventually reduce ambiguity of the true probability distribution. (ii) Theorem 4 does not require condition (49). Indeed, the theorem does not indicate whether the true distribution is located in $\mathcal{P}_N$ or not although $\mathcal{P}_N$ may be arbitrarily close to $\mathcal{P}$. However, under the condition, we are guaranteed that $\mathcal{P}_N$ contains the true distribution with a probability at least $1 - \delta$. (iii) The moment conditions in Part (ii) imply that the probability distribution of random variables ξ_i and $\xi_i\xi_j$ die exponentially fast in the tails. In particular, it holds if Ξ is bounded; see comments before [40, Theorem 5.1]. The exponential rate of convergence can also be proved through a combination of [13, Corollary 2] and [13, Theorem 1] despite there is a small disadvantage where the support set Ξ needs to be bounded; see [13, Assumption 5]. (iv) It is possible to strengthen Part (ii) of Theorem 4 by presenting a kind of asymptotic convergence, that is, for any positive number $\epsilon < 1$ there exists a positive constant $\beta(N, \epsilon)$ such that

$$\text{Prob}\left(\sqrt{N}\mathbb{H}_{TV}(\mathcal{P}, \mathcal{P}_N) \le \beta(N, \epsilon)\right) \ge 1 - \epsilon$$

when N is sufficiently large. Here $\beta(N, \epsilon) \to 0$ as $N \to \infty$. This can be achieved by utilizing asymptotic convergence of μ_N to μ_0 and Σ_N to Σ_0 established by Delage and Ye in [13, Corollary 1] and [13, Theorem 2] respectively, and then by following a similar analysis to that of [13, Section 3], demonstrating that the right hand side of (62) converges to zero at a rate of $O(1/\sqrt{N})$ with probability $1 - \epsilon$, see also [25]. Such a result will indicate the rate of convergence of $\mathbb{H}_{TV}(\mathcal{P}, \mathcal{P}_N)$ ($O(1/\sqrt{N})$) but requires some delicate analysis to derive the constant $\beta(N, \epsilon)$, we leave this to interested readers.

We conclude this section with a note that the error bounds derived in this section are about approximation of $\mathcal{P}_N$ to $\mathcal{P}$ under the pseudometric and the total variation metric when the ambiguity sets are constructed in a specific manner. The results can be immediately plugged into Assumption 3 and consequently translated into convergence of the optimal value and the optimal solutions through Theorems 1 and 2 when $f(x, \xi)$ satisfies Assumption 1.

5. An extension to distributionally robust equilibrium problems. In this section, we sketch some ideas of potential extension of the convergence analysis in the preceding sections to robust equilibrium problems. Let us consider a stochastic game where m players compete to provide a homogenous goods or service for future. Players need to make a decision at the present before realization of uncertainty. Each player does not have complete information on the underlying uncertainties but he is able to use partial information (e.g. samples) or subjective judgement to construct a set of distributions (ambiguity set) which contains the true distribution. We consider an

equilibrium where each player takes a robust action, that is, calculating his expected disutlity based on the worst distribution from his ambiguity set. Mathematically, we can formulate an individual player i's problem as follows:

$$\vartheta_i(y_i, y_{-i}) := \min_{y_i \in Y_i} \sup_{P_i \in \mathcal{P}_i} \mathbb{E}_{P_i}[f_i(y_i, y_{-i}, \xi(\omega))], \tag{66}$$

where Y_i is a closed subset of $\mathbb{R}^{n_i}$, y_i denotes player i's decision vector and y_{-i} the decision vectors of his rivals. The uncertainty is described by random variable $\xi : \Omega \to \Xi$ defined in space $(\Omega, \mathcal{F})$ and his true distribution is unknown. However, player i believes that the true distribution of ξ lies in an ambiguity set, denoted by $\mathcal{P}_i$, and the mathematical expectation on the disutility function f_i is taken with respect to $P_i \in \mathcal{P}_i$. If we consider $(\Xi, \mathscr{B})$ as a measurable space equipped with Borel sigma algebra $\mathscr{B}$, then $\mathcal{P}_i$ may be viewed as a set of probability measures defined on $(\Xi, \mathscr{B})$ induced by the random variate ξ. Note that in this game, players face the same uncertainty but different players may have a different ambiguity set which relies heavily on availability of information on the underlying uncertainty to them. The robust operation means that due to incomplete information on the future uncertainty, player i takes a conservative view on his expected disutility to hedge the risks. To simplify our discussion, we assume that player i's feasible solution set Y_i is deterministic and independent of his competitor's action.

Assuming the players compete under Nash conjecture, we may consider the following one stage *distributionally robust Nash equilibrium* problem: find $y^* := (y_1^*, \cdots, y_m^*) \in Y_1 \times \cdots \times Y_m =: Y$ such that

$$y_i^* \in \arg\min_{y_i \in Y_i} \sup_{P_i \in \mathcal{P}_i} \mathbb{E}_{P_i}[f_i(y_i, y_{-i}^*, \xi(\omega))], \text{ for } i = 1, \cdots, m. \tag{67}$$

Aghassi and Bertsimas [1] apparently are the first to investigate robust games. They consider a distribution-free model of incomplete-information *finite* games, both with and without private information, in which the players use a robust optimization approach to contend with payoff uncertainty. More recently, Qu and Goh [28] propose a distributionally robust version of the finite game where each player uses a distributionally robust approach to deal with incomplete information of uncertainty. Our model (67) may be viewed as an extension of [28] to continuous games.

If we look at a situation where each individual player builds his ambiguity set $\mathcal{P}_i$ in a process (of accumulation of sample information), then we may analyze convergence of the resulting robust equilibrium as the process goes to a limit. Specifically, let $\mathcal{P}_i^N$ denote an approximation of $\mathcal{P}_i$ for $i = 1, \cdots, m$. We consider the *approximate robust Nash equilibrium* problem: find $y^N \in Y$ such that

$$y_i^N \in \arg\min_{y_i \in Y_i} \sup_{P_i \in \mathcal{P}_i^N} \mathbb{E}_{P_i}[f_i(y_i, y_{-i}^N, \xi(\omega))], \text{ for } i = 1, \cdots, m. \tag{68}$$

Here y^N may be interpreted as a robust Nash equilibrium on the basis of each player's perception of uncertainty at stage N. Our interest here is convergence of such equilibrium as each player gathers more information to build up his ambiguity set. Through a similar analysis in Section 3, we may establish convergence of y^N to y^* under some appropriate conditions, we omit the details as the analysis is fundamentally analogous to the minimax distributionally robust optimization problem. We refer interested readers to our earlier report [41].

Acknowledgments. We would like to thank Werner Römisch, Alexander Shapiro, Yi Xu, Chao Ding, Anthony So, Daniel Kuhn and Yongchao Liu for valuable discussions on a number of technical details. We are also grateful to two anonymous referees for careful reading and constructive comments which help us significantly improve the quality of this paper.

References

[1] M. Aghassi and D. Bertsimas, Robust game theory, *Mathematical Programming*, Vol. 107, pp. 231-273, 2006.

[2] K. B. Athreya and S. N. Lahiri, *Measure Theory and Probability Theory*, Springer, NewYork, 2006.

[3] H. Attouch, G. Buttazzo and G. Michaille, *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*, SIAM, Philadelphia, PA, 2005.

[4] B. Bank, J. Guddat, D. Klatte, D. Kummer and K. Tammer, *Nonlinear Parametric Optimization*, Academic Verlag, Berlin, 1982.

[5] A. Ben-Tal, L. El Ghaoui, A. Nemirovski, *Robust Optimization*, Princeton University Press, Princeton, NJ, 2009.

[6] D. Bertsimas, X. V. Doan, K. Natarajan, and C.-P. Teo, Models for minimax stochastic linear optimization problems with risk aversion, *Mathematics of Operations Research*, Vol. 35, pp. 580-602, 2010.

[7] D. Bertsimas and I. Popescu, Optimal inequalities in probability theory: A convex optimization approach, *SIAM Journal on Optimization*, Vol. 15, pp. 780-804, 2005.

[8] P. Billingsley, *Convergence and Probability Measures*, Wiley, New York, 1968.

[9] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[11] M. Breton, S. E. Hachem, Algorithms for the solution of stochastic dynamic minimax problems, *Computational Optimization and Applications*, Vol. 4, pp. 317-345, 1995.

[12] M. Breton, S.E. Hachem, A scenario aggregation algorithm for the solution of stochastic dynamic minimax problems, *Stochastics and Stochastic Reports*, Vol. 53, pp. 305-322, 1995.

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

[14] J. Dupačová, The minimax approach to stochastic programming and an illustrative application, *Stochastics*, Vol. 20, pp. 73-88, 1987.

[15] J. Dupačová, Uncertanities in minimax stochastic programs, *Optimization*, Vol. 60, pp. 10-11, 2011.

[16] K. Fan, Minimax theorems, *Proceedings of National Academy of Sciences of the United States of America*, Vol. 39, pp. 42-47, 1953.

[17] J. Goh and M. Sim, Distributionally robust optimization and its tractable approximations, *Operations Research*, Vol. 58, pp. 902-917, 2010.

[18] D. Goldfarb and G. Iyengar, Robust portfolio selection problems, *Mathematics of Operations Research*, Vol. 28, pp. 1-38, 2003.

[19] J. A. Hall, B. W. Brorsen, and S. H. Irwin, The distribution of futures prices: a test of stable Paretian and mixture of normals hypotheses, *Journal of Financial and Quantitative Analysis*, Vol. 24, pp. 105-116, 1989.

[20] C. Hess, Conditional expectation and marginals of random sets, *Pattern Recognition*, Vol. 32, pp. 1543-1567, 1999.

[21] N. Higham, Computing a nearest symmetric positive semidefinite matrix, *Linear Algebra and its Applications*, Vol. 103, pp. 103-118, 1988.

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

[23] Y. Liu and H. Xu, Stability and sensitivity analysis of stochastic programs with second order dominance constraints. *Mathematical Programming*, Vol. 142, pp. 435-460, 2013.

[24] D. Peel and G. J. McLachlan, Robust mixture modelling using t distribution, *Statistics and Computing*, Vol. 10, pp. 339-348, 2000.

[25] G. Ch. Pflug, Stochastic optimization and statistical inference, A. Rusczyński and A. Shapiro, eds. *Stochastic Program.*, Handbooks in OR & MS, Vol. 10, North-Holland Publishing Company, Amsterdam, 2003.

[26] G. Ch. Pflug and A. Pichler, Approximations for Probability Distributions and Stochastic Optimization Problems, M. Bertocchi, G. Consigli and M. A. H. Dempster, eds. *Stochastic Optimization Methods in Finance and Energy*, Vol. 163, Springer, Berlin, 2011

[27] Y. V. Prokhorov, Convergence of random processes and limit theorems in probability theory, *Theory of Probabilty and Its Applications*, Vol. 1, pp. 157-214, 1956.

[28] S. J. Qu and M. Goh, Distributionally robust games with an application to supply chain, Harbin Institute of Technology, 2012.

[29] M. Riis and K. A. Andersen, Applying the minimax criterion in stochastic recourse programs, *European Journal of Operational Research*, Vol. 165, pp. 569-584, 2005.

[30] R. T. Rockafellar and R. J-B. Wets, *Variational analysis*, Springer, Berlin, 1998.

[31] W. Römisch, Stability of stochastic programming problems, in *Stochastic Programming*, A. Ruszczyński and A. Shapiro, eds., Elsevier, Amsterdam, pp. 483-554, 2003.

[32] H. Scarf, A min-max solution of an inventory problem. K. S. Arrow, S. Karlin and H. E. Scarf., eds. *Studies in the Mathematical Theory of Inventory and Production*, Stanford University Press, Stanford, CA, pp. 201-209, 1958.

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

[34] A. Shapiro, On duality theory of conic linear problems, Miguel A. Goberna and Marco A. López, eds., *Semi-Infinite Programming: Recent Advances, Kluwer Academic Publishers*, pp. 135-165, 2001.

[35] A. Shapiro, Consistency of sample estimates of risk averse stochastic programs, *Journal of Applied Probability*, Vol. 50, pp. 533-541, 2013.

[36] A. Shapiro, Monte Carlo sampling methods, A. Rusczyński and A. Shapiro, eds. *Stochastic Program.*, Handbooks in OR & MS, Vol. 10, North-Holland Publishing Company, Amsterdam, 2003.

[37] A. Shapiro and S. Ahmed, On a class of minimax stochastic programs, *SIAM Journal on Optimization*, Vol. 14, pp. 1237-1249, 2004.

[38] A. Shapiro and A. J. Kleywegt, Minimax analysis of stochastic problems, *Optimization Methods and Software*, Vol. 17, pp. 523-542, 2002.

[39] A. Shapiro, D. Dentcheva and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, SIAM, Philadelphia, PA, 2009.

[40] A. Shapiro and H. Xu, Stochastic mathematical programs with equilibrium constraints, modeling and sample average approximation, *Optimization*, Vol. 57, pp. 395-418, 2008.

[41] H. Sun and H. Xu, Asymptotic Convergence Analysis for Distributional Robust Optimization and Equilibrium Problems, Optimization online, May 2013,

[42] S. Takriti, S. Ahmed, Managing short-term electricity contracts under uncertainty: A minimax approach. *http://www.isye.gatech.edu/ sahmed*

[43] Z. Wang, P. W. Glynn and Y. Ye, Likelihood robust optimization for data-driven Newsvendor problems, manuscript, 2012.

[44] W. Wiesemann, D. Kuhn, and B. Rustem, Robust Markov decision process, *Mathematics of Operations Research* Vol. 38, pp. 153-183, 2013.

[45] H. Xu, C. Caramanis and S. Mannor, A distributional interpretation of robust optimization, *Mathematics of Operations Research*, Vol. 37, pp. 95-110, 2012.

[46] J. Žáčková, On minimax solution of stochastic linear programming problems, *Časopis pro Pěstování Matematiky*, Vol. 91, pp. 423-430, 1966.

[47] S. Zhu and M. Fukushima, Worst-case conditional Value-at-Risk with application to robust portfolio management, *Operations Research*, Vol. 57, pp. 1155-1156, 2009.