Asymptotic Convergence Analysis for Distributional Robust Optimization and Equilibrium Problems[1]

Hailin Sun
Department of Mathematics, Harbin Institute of Technology, Harbin, 150001, China
(mathhlsun@gmail.com)

Huifu Xu
School of Engineering and Mathematical Sciences, City University of London, EC1V 0HB, London, UK
(Huifu.Xu.1@city.ac.uk)

May 2, 2013

Abstract. In this paper, we study distributional 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 worst possible distribution from that set. We consider the case when the distributional set is constructed through samples and investigate asymptotic convergence of optimal values and optimal solutions as sample size increases. The analysis provides a unified framework for asymptotic convergence of some data-driven problems and extends the classical asymptotic convergence analysis in stochastic programming. The discussion is extended to a stochastic Nash equilibrium problem where each player takes a robust action on the basis of their subjective expected objective value.

Key Words. Distributional robust minimization, asymptotic analysis, Hoffman's lemma, robust Nash equilibrium

1 Introduction

Consider the following distributional robust stochastic program (DRSP):

$$\begin{array}{ll} \min\limits_{x} \max\limits_{P\in\mathcal{P}} & \mathbb{E}_P[f(x,\xi(\omega))] \\ \text{s.t.} & x \in X, \end{array} \tag{1.1}$$

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, $\xi : \Omega \to \Xi$ is a vector of random variables defined on probability space $(\Omega, \mathcal{F}, P)$ with support set $\Xi \subset \mathbb{R}^k$, $\mathcal{P}$ is a set of distributions which contains the true probability distribution of random variable ξ, and $\mathbb{E}_P[\cdot]$ denotes the expected value with respect to probability measure $P \in \mathcal{P}$.

Differing from classical stochastic programming model, the distributional robust formulation (1.1) determines the optimal policy x on the basis of worst expected value of $f(x,\xi)$ over the distributional set $\mathcal{P}$. It reflects some practical circumstances 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 [42]. This kind of robust optimization framework can be traced back to the earlier work by Scarf [37] which was motivated to address incomplete information on the underlying uncertainty in supply chain and inventory control problems. In such problems, historical data may

[1]The research is supported by EPSRC grant EP/J014427/1 and National Natural Science Foundation of China (grant No.11171159).

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 distributional 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 [48], Dupačová [15, 16], and more recently by Shapiro and Kleywegt [43] and Shapiro and Ahmed [42]. Over the past few years, it has gained substantial popularity through further contributions by Bertsimas and Popescu [9], Betsimas et al [8], Goh and Sim [20], Zhu and Fukushima [49], Goh and Sim [20], Goldfarb and Iyengar [21], Delage and Ye [14] and Xu et al [47], 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 (1.1) reduces to worst scenario robust optimization, see monograph by Ben Tal et al [6] for the latter.

A key step in the research of DRSP (1.1) is to construct a distributional 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 [16]. 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 [43, 14, 21]. Delage and Ye [14] propose a model that describes uncertainty in both the distribution form and moments, and demonstrate that for a wide range of functions f, (1.1) 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 [39] 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 [45, 46].

Obviously there is a gap between the distributional 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/asymptotic analysis of the robust optimization problem. In the case when the distributional set reduces to a singleton, DRSP (1.1) collapses to a classical one stage stochastic optimization problem. Asymptotic convergence and/or stability analysis of the latter has been well documented, see review papers by Pflug [30], Römisch [33] and Shapiro [41].

Dupačová [16] seems to be the first to investigate stability of distributional robust optimization problem. Under some convexity and compactness conditions, she shows epi-convergence of the optimal value function based on the worst probability distribution from a distributional set defined through estimated moments. Over the past few years, there has been a few papers that address asymptotic convergence of distributional robust optimization problems where $\mathcal{P}$ is constructed through independent and identically distributed (iid for short) samples and the distributional set converges to the true probability distribution as sample size goes to infinity, see recent paper by Wang et al [45] and Wiesemann et al [46] and the references therein.

Our focus in this paper is on the case when the distributional set constructed through samples converges to a set which is not necessarily a singleton. 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 distributional set constructed through the estimated moments may converge to a set with true moments rather than a single distribution. To this end, we propose to study approximation of distributional sets under total variation metric and the pseudometric. The former

allows us to measure the convergence of the distributional 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 (1.1) as a parametric optimization problem and investigate properties of the optimal value and optimal solutions as x varies and sample size increases. Under some moderate conditions, we show that the objective function is equi-Lipschitz continuous and optimal solution set is upper semicontinuous w.r.t. x. Moreover, we demonstrate uniform convergence of the optimal value function and consequently the asymptotic convergence of robust optimal solution of (1.1) to its true counterpart as sample size increases.
- We investigate convergence of distributional sets under total variation metric as sample size increases for the cases when the distributional set is constructed through moments, mixture distribution, and moments and covariance matrix due to Delage and Ye [14] and So [39]. In the case when a distributional 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 [40] for linear conic programs and use it to establish a linear bound for the distance of two distributional sets under total variation metric.
- Finally, we extend our discussion to a distributionally robust Nash equilibrium problem where each player takes a robust action on the basis of their subjective expected objective value over a distributional set. By assuming each player constructs its distributional set through samples (e.g. historical data), we show under some moderate conditions that the sampled distributionally robust Nash equilibria converge to their true counterparts almost surely as sample size increases.

Throughout the paper, we will use the following notation. For matrices $A, B \in \mathbb{R}^{n\times n}$, $A \bullet B$ denotes 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. We write $\|A\|_F$ for the Frobenius norm of A, that is,

$$\|A\|_F := (A \bullet A)^{1/2}$$

and $\|x\|$ for the Euclidean norm of a vector x in $\mathbb{R}^n$, $\|\psi\|$ for the maximum norm of a real valued measure function $\psi : \Omega \to \mathbb{R}$ and $\|x\|_\infty$ for the infinity norm of x.

2 Data driven problem

2.1 Definition of the problem

Let Ω be a measurable space with σ-algebra $\mathcal{F}$. Let $\mathscr{P}$ be the set of all probability measures [2] the paper defined on $(\Omega, \mathcal{F})$ and $\mathcal{P}_N \subset \mathscr{P}$ be a set of probability distributions which approximate $\mathcal{P}$

[2]Throughout the paper, we use the terms measure and distribution interchangeably.

in some sense (to be specified later) as $N \to \infty$. We construct an approximation scheme for the distributional robust optimization problem (1.1) by replacing $\mathcal{P}$ with $\mathcal{P}_N$:

$$\begin{array}{ll} \min\limits_{x} \max\limits_{P\in\mathcal{P}_N} & \mathbb{E}_P[f(x,\xi(\omega))] \\ \text{s.t.} & x \in X. \end{array} \tag{2.2}$$

Typically, $\mathcal{P}_N$ may be constructed through samples. For instances, Shapiro and Ahmed [42] consider $\mathcal{P}$ being defined through moments and use empirical data (samples) to approximate the true moments. Delage and Ye [14] 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 [45] and Wiesemann et al [46] 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 asymptotic 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 $v_N(x)$ denote the optimal value of the inner maximization problem

$$\max_{P\in\mathcal{P}_N} \quad \mathbb{E}_P[f(x,\xi(\omega))] \tag{2.3}$$

and $\Phi_N(x)$ the corresponding set of optimal solutions, that is,

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

Likewise, we write $v(x)$ for the optimal value of

$$\max_{P\in\mathcal{P}} \quad \mathbb{E}_P[f(x,\xi(\omega))] \tag{2.4}$$

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

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

Consequently we can write (2.2) and (1.1) respectively as

$$\min_{x\in X} v_N(x) \tag{2.5}$$

and

$$\min_{x\in X} v(x) \tag{2.6}$$

Let ϑ_N and ϑ denote the optimal value and X_N and X^* the set of optimal solutions of (2.5) and (2.6) respectively. Our aim is to investigate convergence of ϑ_N to ϑ and X_N to X^* as $N \to \infty$. The current research has addressed this kind of convergence for some specific problems when $\mathcal{P}_N$ converges to the true probability distribution of ξ. 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, (2.2) reduces to an ordinary approximation scheme of one stage stochastic minimization problem and our proposed analysis collapses to classical stability analysis in stochastic programming [33]. From this perspective, we might regard the asymptotic 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 (1.1) and (2.2). 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 and pseudometric

We need appropriate metrics $\mathscr{P}$ to give a quantitative description of the convergence of $\mathcal{P}_N \to \mathcal{P}$ and $v_N(x) \to v(x)$. Here we consider the total variation metric for the former and pseudometrics for the latter. Both metrics are well known in probability theory and stochastic programming, see for instance [2, 33].

Let $P, Q \in \mathscr{P}$ and $\mathscr{M}$ denote the set of measurable functions defined in the probability space $(\Omega, \mathcal{F})$. Recall that the *total variation metric* between P and Q is defined as (see e.g., page 270 in [2])

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

where

$$\mathscr{M} := \{h : h : \Omega \to \mathbb{R} \text{ is } \mathcal{F} \text{ measurable}, \sup\{|h(\omega)| : \omega \in \Omega\} \leq 1\}$$

and *total variation norm* as

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

With these notions, we can define the distance from a point to a set, deviation from one set to another and Hausdorff distance between two sets in space $\mathscr{P}$. Specifically, let

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

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

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

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

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.

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

$$\lim_{N \to \infty} \int_{\Xi} h(\omega) dP_N(\omega) = \int_{\Xi} h(\omega) dP(\omega)$$

for each bounded and continuous function $h : \Omega \to \mathbb{R}$. Obviously convergence under the total variation metric implies weak convergence.

The total variation metric defined above is independent of function $f(x, \xi)$ in the distributional robust optimization problem (1.1). In what follows, we introduce another metric which is closely related to the objective function $f(x, \xi)$. Define the set of random functions:

$$\mathscr{G} := \{g(\cdot) := f(x, \xi(\cdot)) : x \in X\}. \tag{2.7}$$

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

$$\mathscr{D}(P, Q) := \sup_{g \in \mathscr{G}} \big|\mathbb{E}_P[g] - \mathbb{E}_Q[g]\big|. \tag{2.8}$$

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 $\mathscr{G}$ is sufficiently large. This type of pseudometric is widely used for stability analysis in stochastic programming; see an excellent review by Römisch [33].

Let $Q \in \mathscr{P}$ be a probability measure and $\mathcal{A}_i \in \mathscr{P}$, $i=1,2$, be a set of probability measures. With the pseudometric, we may define the distance from a point Q to a set $\mathcal{A}_1$ 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) $\mathcal{A}_2$

$$\mathscr{D}(\mathcal{A}_1,\mathcal{A}_2) := \sup_{Q\in\mathcal{A}_1} \mathscr{D}(Q,\mathcal{A}_2) \tag{2.9}$$

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

$$\mathscr{H}(\mathcal{A}_1,\mathcal{A}_2) := \max\left\{ \sup_{Q\in\mathcal{A}_1} \mathscr{D}(Q,\mathcal{A}_2), \sup_{Q\in\mathcal{A}_2} \mathscr{D}(Q,\mathcal{A}_1) \right\}. \tag{2.10}$$

Remark 2.1 There are two important cases to note.

(i) Consider the case when $\mathscr{G}$ is bounded, that is, there exists a positive number M such that

$$\sup_{g\in\mathscr{G}} \|g\| \le M.$$

Let $\tilde{\mathscr{G}} = \mathscr{G}/M$. Then

$$\mathscr{D}(P,Q) := M \sup_{\tilde{g}\in\tilde{\mathscr{G}}} \left|\mathbb{E}_P[\tilde{g}] - \mathbb{E}_Q[\tilde{g}]\right| \le M d_{TV}(P,Q). \tag{2.11}$$

(ii) Consider the case when

$$\sup_{x\in X} |f(x,\xi) - f(x,\xi')| \le c_p(\xi,\xi')\|\xi-\xi'\| : \forall \xi,\xi' \in \Xi, \tag{2.12}$$

where

$$c_p(\xi,\xi') := \max\{1, \|\xi\|, \|\xi'\|\}^{p-1}$$

for all $\xi,\xi' \in \Xi$ and $p \ge 1$.

In the case when $p=1$, $\mathscr{D}(P,Q)$ recovers the well known *Kantorovich metric* and when $p \ge 1$ 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 (both P_N and P having p-th order moments) iff it converges to P weakly and

$$\lim_{N\to\infty} \mathbb{E}_{P_N}[\|\xi\|^p] = \mathbb{E}_P[\|\xi\|^p] < \infty, \tag{2.13}$$

see [33]. It means that if f satisfies conditions (2.12), then weak convergence of P_N to $P \in \mathcal{P}$ and (2.13) imply $\mathscr{D}(P_N,P) \to 0$ and hence $\mathscr{D}(P_N,\mathcal{P}) \to 0$. If (2.13) holds for any $P_N \in \mathcal{P}_N$, then we arrive at $\mathscr{D}(\mathcal{P}_N,\mathcal{P}) \to 0$.

2.3 Well definedness of the robust problem

We need to make sure that problems (2.5) and (2.6) 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 (2.3) and (2.4) where x is treated as a parameter and probability measure P is a variable. To this end, we make the following assumptions which ensure the feasible set of P in these problems, namely $\mathcal{P}_N$ and $\mathcal{P}$, are closed and bounded.

Assumption 2.1 *Let $\mathcal{P}, \mathcal{P}_N$ be defined as in (1.1) and (2.2) respectively. There exists a compact (in weak topology) set of probability measures $\hat{\mathcal{P}} \subset \mathscr{P}$ such that the following hold.*

(a) $\mathcal{P}$ is nonempty and compact in the weak topology and $\mathcal{P} \subset \hat{\mathcal{P}}$;

(b) for each N, $\mathcal{P}_N$ is a nonempty compact set in the weak topology and $\mathcal{P}_N \subset \hat{\mathcal{P}}$ when N is sufficiently large.

Let $\mathcal{A}$ be a set of probability measures on $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is the Borel σ-algebra on Ω. Recall that $\mathcal{A}$ is said to be *tight* if for any $\epsilon > 0$, there exists a compact set $K \subset \Omega$ such that

$$\sup_{P \in \mathcal{A}} P(\xi \notin K) < \epsilon.$$

$\mathcal{A}$ is said to be *closed* (in the weak topology) if for any sequence $\{P_N\} \subset \mathcal{P}$ and $P_N \to P$ weakly, $P \in \mathcal{A}$. By Prokhorov's theorem, a closed set $\mathcal{A}$ (in the weak topology) of probability measures is compact if it is tight. In particular, if Ω is a compact metric space, then the set of all probability measures on $(\Omega, \mathcal{F})$ is weakly compact; see [40]. In Section 4, we will discuss tightness and compactness in detail when $\mathcal{P}_N$ has a specific structure.

Assumption 2.2 *Let $f(x, \xi)$ be defined as in (1.1). For each fixed $\xi \in \Xi$, $f(\cdot, \xi)$ is Lipschitz continuous on X with Lipschitz modulus being bounded by $\kappa(\xi)$, where*

$$\sup_{P \in \hat{\mathcal{P}}} \mathbb{E}_P[\kappa(\xi)] < \infty$$

and $\hat{\mathcal{P}}$ is defined as in Assumption 2.1.

The proposition below summarizes main properties of the optimal value and optimal solution set of the parametric programs (2.3) and (2.4).

Proposition 2.1 *Let Assumptions 2.1 and 2.2 hold. We have the following assertions.*

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

$$|\mathbb{E}_P[f(x, \xi)] - \mathbb{E}_Q[f(y, \xi)]| \leq \mathscr{D}(P, Q) + \sup_{P \in \hat{\mathcal{P}}} \mathbb{E}_P[\kappa(\xi)] \|x - y\| \tag{2.14}$$

for $P, Q \in \mathcal{P}_N$ and $x, y \in X$;

(ii) $\Phi_N(x) \neq \emptyset$ for all $x \in X$;

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

(iv) $v_N(\cdot)$ *is equi-Lipschitz continuous on* X *with modulus* $\sup_{P\in\hat{\mathcal{P}}}\mathbb{E}_P[\kappa(\xi)]$*, that is,*

$$|v_N(x) - v_N(y)| \le \sup_{P\in\hat{\mathcal{P}}}\mathbb{E}_P[\kappa(\xi)]\|x-y\|,\ \forall x,y\in X; \tag{2.15}$$

(v) $v(\cdot)$ *(resp.* $v_N(\cdot)$*) is Lipschitz continuous on* X*;*

(vi) $v(\cdot)$ *(resp.* $v_N(\cdot)$*) is Clarke directionally differentiable and*

$$\partial v(x) = \mathrm{cl}\left(\bigcup_{P\in\Phi(x)}\mathbb{E}_P[\nabla_x f(x,\xi)]\right), \tag{2.16}$$

(resp.

$$\partial v_N(x) = \mathrm{cl}\left(\bigcup_{P\in\Phi_N(x)}\mathbb{E}_P[\nabla_x f(x,\xi)]\right), \tag{2.17}$$

), where $\partial v(x)$ *denotes the Clarke subdifferential of* v *at* x *(see [12] for the definition), "*cl*" denotes the closure of a set.*

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

$$|\mathbb{E}_P[f(x,\xi)] - \mathbb{E}_Q[f(x,\xi)]| \le \mathscr{D}(P,Q). \tag{2.18}$$

The continuity of $\mathbb{E}_P[f(x,\xi)]$ in P and the compactness of $\mathcal{P}_N$ ensure that $\Phi_N(x)\neq\emptyset$ for every $x\in X$. This shows part (ii).

For any $x,y\in X$,

$$|\mathbb{E}_P[f(x,\xi)] - \mathbb{E}_P[f(y,\xi)]| \le \sup_{P\in\hat{\mathcal{P}}}\mathbb{E}_P[\kappa(\xi)]\|x-y\|. \tag{2.19}$$

A combination of (2.18) and (2.19) gives (2.14), that is, $\mathbb{E}_P[f(x,\xi)]$ is Lipschitz continuous w.r.t. (P,x) on $\mathcal{P}_N\times X$. This shows (i).

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). The proof is similar to [27, Theorem 1]. Let $P_N(y)\in\Phi_N(y)$. By (2.14)

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

Exchanging the role of x and y, we obtain (2.15).

Part (v). Similar to the proof of (i)-(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.

Part (vi). The result essentially follows from [38, Proposition 3.3]. Here we provide outlines of the proof for completeness. Let Z be a random variable which take value in $\mathbb{R}$ and

$$\rho(Z) := \max_{P\in\mathcal{P}}\mathbb{E}_P[Z].$$

It is easy to verify that $\rho(\cdot)$ is a convex function and directionally differentiable (indeed, it is a coherent risk measure). Let $x \in X$. Through a similar proof to Part (iv), we can show that v is Lipschitz continuous and hence $v(x)$ is finite. Moreover for fixed $d \in \mathbb{R}^n$,

$$v'(x;d) = \sup_{P\in\Phi(x)} \mathbb{E}_P[\nabla_x f(x,\xi)^T d].$$

By [38, Proposition 2.116 (ii)],

$$\partial v(x) = \partial_d v'(x;0) = \mathrm{cl}\left(\bigcup_{P\in\Phi(x)} \mathbb{E}_P[\nabla_x f(x,\xi)]\right).$$

The proof is complete. ■

Next, we discuss well definedness of parametric program (2.3) and the proposition below gives out some sufficient conditions.

Proposition 2.2 *Consider problem (2.4). Let $x \in X$ be fixed. Assume: (a) Assumption 2.1 (a) holds, (b) there exists $\alpha \in \mathbb{R}$ such that the upper level set $\mathrm{Lev}_\alpha \mathbb{E}_P[f(x,\xi)] := \{P \in \mathcal{P} : \mathbb{E}_P[f(x,\xi)] \geq \alpha\}$ is nonempty, (c) $f(x,\cdot)$ is continuous w.r.t. ξ and there exists a positive number M such that, for all $\xi \in \Xi$, $f(x,\xi) \leq M$. Then there exists $P^* \in \mathcal{P}$ which attains the maximum and the optimal value is finite.*

Proof. Let

$$\mathscr{V} := \{\mathbb{E}_P[f(x,\xi)] : P \in \mathrm{Lev}_\alpha \mathbb{E}_P[f(x,\xi)]\}.$$

It suffices to show that $\mathscr{V}$ is a nonempty compact set in $\mathbb{R}$. Observe first that under condition (c),

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi)] \leq M.$$

Moreover, since the upper level set $\mathrm{Lev}_\alpha \mathbb{E}_P[f(x,\xi)]$ is nonempty and contained in $\mathcal{P}$ under condition (a), $\mathscr{V}$ is nonempty and lower bounded by α. This shows $\mathscr{V}$ is contained in the interval $[\alpha, M]$, hence it is a bounded set in $\mathbb{R}$. In what follows, we show that $\mathscr{V}$ is closed. Let $\{v_i\} \subset \mathscr{V}$ be any sequence converging to $\hat{v}$. It suffices to show $\hat{v} \in \mathscr{V}$. Let $\{P_i\}$ be such that $\mathbb{E}_{P_i}[f(x,\xi)] = v_i$. Since $\mathcal{P}$ is compact, by taking a subsequence if necessary, we may assume that $P_i \to \hat{P}$ weakly and

$$\lim_{i\to\infty} v_i = \lim_{i\to\infty} \mathbb{E}_{P_i}[f(x,\xi)] = \mathbb{E}_{\hat{P}}[f(x,\xi)] = \hat{v}.$$

The closedness of $\mathcal{P}$ means $\hat{P} \in \mathcal{P}$. To complete the proof, we need to show that $\hat{P} \in \mathrm{Lev}_\alpha \mathbb{E}_P[f(x,\xi)]$. This is obvious because $v = \lim_{i\to\infty} v_i \geq \alpha$ and if $\hat{P} \in \mathcal{P}\backslash \mathrm{Lev}_\alpha \mathbb{E}_P[f(x,\xi)]$, then $\mathrm{Lev}_\alpha \mathbb{E}_{\hat{P}}[f(x,\xi)] < \alpha$, a contradiction! ■

Note that in some cases, $\mathcal{P}$ and/or $\mathcal{P}_N$ may not be closed, see Example 4.1. In these circumstances, we may consider the closure of the distributional sets and replace them with their closure in (1.1) and (2.2). The resulting optimal values and solutions may be more conservative in the case when there is a gap between the distributional sets and their closure and the optimum is attained at the boundary. Note that in order to solve the distributional robust optimization problem (2.3), one usually need 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 [14] and the references therein.

3 Asymptotic analysis

In this section, we investigate convergence of optimal value ϑ_N and optimal solution set X_N as $\mathcal{P}_N \to \mathcal{P}$. We will carry out the convergence analysis without referring to the specific structure of $\mathcal{P}_N$ or $\mathcal{P}$ so that the convergence results may cover a wide range of problems. To ensure the convergence implies the convergence of optimal values and optimal solutions, we need to strengthen it so that the optimal value function of (2.5) converges to that of (2.6) under pseudometrics.

Assumption 3.1 *Let $\mathcal{P}, \mathcal{P}_N$ be defined as in (1.1) and (2.2) 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 (2.10);

(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 (2.9).

Under Assumption 3.1, we are able to present one of the main asymptotic convergence results in this section.

Theorem 3.1 *Assume the setting and conditions of Proposition 2.2. Under Assumption 2.1 and Assumption 3.1 (a), the following assertions hold.*

(i) If $\mathcal{P}$ and $\mathcal{P}_N$ are convex, then $v_N(x)$ converges uniformly to $v(x)$ over X as N tends to infinity, that is,

$$\lim_{N\to\infty} \sup_{x\in X} |v_N(x) - v(x)| = 0 \tag{3.20}$$

almost surely.

(ii) If, in addition, Assumption 3.1 (b) holds, then for any $\epsilon > 0$ there exist positive constants C and β such that

$$\text{Prob}\left(\sup_{x\in X} |v_N(x) - v(x)| \geq \epsilon\right) \leq Ce^{-\beta N} \tag{3.21}$$

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 3.1. Let us first show that $v_N(x) < \infty$. Under condition (b) of Proposition 2.2, there exists a constant α such that

$$\text{Lev}_\alpha \mathbb{E}_P[f(x,\xi)] := \{P \in \mathcal{P} : \mathbb{E}_P[f(x,\xi)] \geq \alpha\} \neq \emptyset.$$

By the definition of $\mathscr{H}(\mathcal{P}_N, \mathcal{P})$, for any $P \in \mathcal{P}$, there exists $Q_P \in \mathcal{P}_N$, such that

$$|\mathbb{E}_P[f(x,\xi)] - \mathbb{E}_{Q_P}[f(x,\xi)]| \leq \mathscr{H}(\mathcal{P}_N, \mathcal{P}).$$

Under Assumption 3.1 (a), there exists N_0 such that for $N \geq N_0$, $\mathscr{H}(\mathcal{P}_N, \mathcal{P}) \leq \alpha/2$. Therefore for each $P \in \mathrm{Lev}_\alpha \mathbb{E}_P[f(x,\xi)]$, there exists Q_P, such that

$$|\mathbb{E}_P[f(x,\xi)] - \mathbb{E}_{Q_P}[f(x,\xi)]| \leq \alpha/2,$$

which implies

$$\alpha/2 \leq \mathbb{E}_{Q_P}[f(x,\xi)] \leq \alpha/2 + \mathbb{E}_P[f(x,\xi)].$$

This shows

$$\mathrm{Lev}_{\alpha/2}\mathbb{E}_{P_N}[f(x,\xi)] := \{P \in \mathcal{P}_N : \mathbb{E}_P[f(x,\xi)] \geq \alpha/2\} \neq \emptyset.$$

Analogous to the proof of Proposition 2.2, we can then show that for the fixed x and N, there exists $P_N \in \mathcal{P}_N$ such that

$$v_N(x) = \mathbb{E}_{P_N}[f(x,\xi)] < \infty.$$

Part (i). Let $x \in X$ be fixed. Let

$$\mathscr{V} := \{\mathbb{E}_P[f(x,\xi)] : P \in \mathcal{P}\}$$

and

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

Since $\mathcal{P}$ and $\mathcal{P}_N$ are assumed to be compact in the weak topology, both $\mathscr{V}$ and $\mathscr{V}_N$ are compact set in $\mathbb{R}$. Let

$$a := \min_{v \in \mathscr{V}} v; \quad b := \max_{v \in \mathscr{V}} v$$

and

$$a_N := \min_{v \in \mathscr{V}_N} v; \quad b_N := \max_{v \in \mathscr{V}_N} v.$$

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

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

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

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

On the other hand, by the definition and property of the Hausdorff distance (see e.g. [23]),

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

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

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

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 (3.20).

Part (ii) follows straightforwardly from part (i) and Assumption 3.1 (b). ■

Assumption 3.1 is essential for deriving the convergence results in Theorem 3.1. 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 3.1 *Assumption 3.1 (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 C such that

$$|f(x,\xi)| \leq C, \forall (x,\xi) \in X \times \Xi.$$

(b) f satisfies condition (2.12), and for every sequence $\{P_N\} \subset \mathcal{P}_N$, $\{P_N\}$ converges to $P \in \mathcal{P}$ weakly and (2.13) holds.

Proof. Sufficiency of (a) and (b) follows from Remark 2.1 (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 3.1 reduces to the standard sufficient condition required in stability analysis of stochastic programming; see [33]. The uniform convergence established in Theorem 3.1 may be translated into the convergence of optimal values and optimal solutions of (2.5) through [32, Theorem 7.64] and Liu and Xu [28, Lemma 3.8].

Theorem 3.2 *Let X_N and X^* denote the set of optimal solutions of (2.2) and (1.1) respectively, ϑ_N and ϑ the corresponding optimal values. Assume that X is a compact set, X_N and X^* are nonempty, $\mathcal{P}$ and $\mathcal{P}_N$ are convex. Under conditions of Proposition 2.2, Assumption 2.1 and Assumption 3.1 (a),*

$$\overline{\lim_{N\to\infty}} X_N \subset X^*$$

and

$$\lim_{N\to\infty} \vartheta_N = \vartheta.$$

If, in addition, Assumption 3.1 (b) holds, then for any $\epsilon > 0$ there exist positive constants α and β (depending on ϵ) such that

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

for N sufficiently large.

Proof. By Proposition 2.1, $v(\cdot)$ and $v_N(\cdot)$ are continuous. The conclusions follow from Theorem 3.1 and [32, Theorem 7.64] or Liu and Xu [28, Lemma 3.8]. ■

4 Approximations of the distributional set

One of the key assumptions in the asymptotic 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 distributional 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 set of probability distributions $\mathcal{P}$ is defined through moment conditions:

$$\mathcal{P} := \left\{ P : \begin{array}{ll} \mathbb{E}_P[\psi_i(\omega)] = \mu_i, & \text{for } i = 1, \cdots, p \\ \mathbb{E}_P[\psi_i(\omega)] \leq \mu_i, & \text{for } i = p+1, \cdots, q \end{array} \right\}, \tag{4.22}$$

where $\psi_i : \Omega \to \mathbb{R}$, $i = 1, \cdots, q$, are measurable functions. Let

$$\mathcal{P}_N := \left\{ P : \begin{array}{ll} \mathbb{E}_P[\psi_i(\omega)] = \mu_i^N, & \text{for } i = 1, \cdots, p \\ \mathbb{E}_P[\psi_i(\omega)] \leq \mu_i^N, & \text{for } i = p+1, \cdots, q \end{array} \right\} \tag{4.23}$$

be an approximation to $\mathcal{P}$, where μ_i^N is often 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

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

and

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

It is easy to verify that both $\mathcal{P}$ and $\mathcal{P}_N$ are compact when the support set Ξ of random variable ξ is compact.

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 (4.22), 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 (which is established in a finite dimensional space).

Lemma 4.1 *(Hoffman's lemma for moment problem). Assume that $\mathcal{P}$ is closed. Then there exists a positive constant C depending on ψ such that*

$$d_{TV}(Q, \mathcal{P}) \leq C \left(\|(\mathbb{E}_Q[\psi_I(\omega)] - \mu_I)_+\| + \|\mathbb{E}_Q[\psi_E(\omega)] - \mu_E\| \right)$$

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 a probability measure $Q \in \mathscr{P}$ deviating 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 Ω is a discrete set with finite cardinality, Lemma 4.1 reduces to the classical Hoffman's lemma.

Proof of Lemma 4.1. The proof is essentially derived through Shapiro's duality theorem [40, Proposition 3.1]. Let $P \in \mathscr{P}$ and $\phi(\omega)$ be P-integrable function. Let

$$\langle P, \phi \rangle := \mathbb{E}_P[\phi(\omega)].$$

By the definition of the total variation norm (see [2]),

$$||P||_{TV} = \sup_{\|\phi\| \leq 1} \langle P, \phi \rangle.$$

Moreover, by the definition of the total variation metric

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

where the exchange is justified by [18, Theorem 1]. Under the closedness condition of $\mathcal{P}$, it follows by [40, Proposition 3.1] that

$$\inf_{P \in \{P : \mathbb{E}_P[\psi_E] = \mu_E, \mathbb{E}_P[\psi_I] \leq \mu_I\}} \langle Q - P, \phi \rangle = \sup_{\lambda \in \Lambda, \phi = \psi^T \lambda} \lambda^T (\mathbb{E}_Q[\psi(\omega)] - \mu),$$

where $\psi = (\psi_E, \psi_I)$, $\mu = (\mu_E, \mu_I)$ and

$$\Lambda := \{(\lambda_1, \cdots, \lambda_q) : \lambda_i \geq 0, \text{ for } i = p+1, \cdots, q\}.$$

Consequently

$$d_{TV}(Q, \mathcal{P}) = \sup_{\lambda \in \Lambda, \|\psi(\omega)^T \lambda\| \leq 1} \lambda^T (\mathbb{E}_Q[\psi(\omega)] - \mu). \tag{4.24}$$

Note that constraint $\|\psi(\omega)^T \lambda\| \leq 1$ means

$$\sup_{\omega \in \Omega} |\psi(\omega)^T \lambda| \leq 1,$$

which is a semi-infinite constraint (here λ is a variable). Therefore the right hand side of (4.24) is a linear semi-infinite program. To estimate its optimal value, we may relax the semi-infinite constraints by considering a sub-index set $(\omega_1, \ldots, \omega_k)$ of Ω. Let $A := (\psi(\omega_1), \ldots, \psi(\omega_k))$. It is easy to observe that

$$\{\lambda \in \Lambda : \|\psi^T \lambda\| \leq 1\} \subset \{\lambda \in \Lambda : \|A^T \lambda\|_\infty \leq 1\}$$

which implies

$$\sup_{\lambda \in \Lambda, \|\psi^T \lambda\| \leq 1} \lambda^T (\mathbb{E}_Q[\psi(\omega)] - \mu) \leq \sup_{\lambda \in \Lambda, \|A^T \lambda\|_\infty \leq 1} \lambda^T (\mathbb{E}_Q[\psi(\omega)] - \mu).$$

In what follows, we estimate the term at the right hand side of the inequality above. Indeed, this is the optimal value of a linear programming problem with polyhedral feasible set $\{\lambda \in \Lambda, \|A^T \lambda\|_\infty \leq 1\}$ in the space $\mathbb{R}^q$. Therefore the maximum is attained at a vertex of the polyhedral and it is finite because the number of vertices is finite. In other words, there exists a finite $\lambda^* \in \Lambda$ such that

$$\lambda^{*T} (\mathbb{E}_Q[\psi(\omega)] - \mu) = \sup_{\lambda \in \Lambda, \|A^T \lambda\|_\infty \leq 1} \lambda^T (\mathbb{E}_Q[\psi(\omega)] - \mu).$$

Therefore

$$\begin{aligned} d_{TV}(Q,\mathcal{P}) &\leq \lambda^{*T}(\mathbb{E}_Q[\psi(\omega)]-\mu) \\ &\leq \sum_{i=1}^{p}\lambda_i^*(\mathbb{E}_Q[\psi_i(\omega)]-\mu_i)+\sum_{i=p+1}^{q}\lambda_i^*(\mathbb{E}_Q[\psi_i(\omega)]-\mu_i) \\ &\leq \sum_{i=1}^{p}\lambda_i^*(\mathbb{E}_Q[\psi_i(\omega)]-\mu_i)_+ +\sum_{i=p+1}^{q}|\lambda_i^*||\mathbb{E}_Q[\psi_i(\omega)]-\mu_i| \\ &\leq C(\|\mathbb{E}_Q[\psi_I(\omega)]-\mu_I)_+\|+\|\mathbb{E}_Q[\psi_E(\omega)]-\mu_E\|]), \end{aligned}$$

where $C=\max_{i=1}^{p}|\lambda_i^*|$. The proof is complete. ■

Note that the closedness condition of $\mathcal{P}$ may be satisfied if there exists positive numbers τ and C such that

$$\int_\Omega |\psi_i(\omega)|^{1+\tau}P(d\omega)<C, i=1,\cdots,q, \tag{4.25}$$

for any $P\in\mathcal{P}$. Indeed, the closedness of $\mathcal{P}$ can be proved easily by applying Lemma 6.1. We omit the details. In the case when Ω is a compact subset and ψ_i, $i=1,\cdots,q$ is a continuous function, $\mathcal{P}$ is compact.

With Lemma 4.1, we are able to quantify the approximation of $\mathcal{P}_N$ to $\mathcal{P}$ under the total variation metric.

Proposition 4.1 *Suppose that $\mathcal{P}$ and $\mathcal{P}_N$ are closed. Then there exists a positive constant C depending on $\phi(\omega)$ such that*

$$\mathbb{H}_{TV}(\mathcal{P}_N,\mathcal{P})\leq C[\max(\|(\mu_I^N-\mu_I)_+\|,\|(\mu_I-\mu_I^N)_+\|)+\|\mu_E^N-\mu_E\|], \tag{4.26}$$

where C is defined as in Lemma 4.1. Moreover, if there is a positive constant M such that $||g||\leq M$ for all $g\in\mathscr{G}$, where $\mathscr{G}$ is defined in (2.7), then

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

Proof. Let $Q\in\mathcal{P}_N$. By Lemma 4.1, there exists a positive constant C such that

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

Therefore

$$\begin{aligned} \mathbb{D}_{TV}(\mathcal{P}_N,\mathcal{P}) &= \sup_{Q\in\mathcal{P}_N} d_{TV}(Q,\mathcal{P}) \\ &= \sup_{Q\in\mathcal{P}_N}\inf_{Q\in\mathcal{P}} d_{TV}(P,Q) \\ &\leq C(\|(\mu_I^N-\mu_I)_+\|+\|\mu_E^N-\mu_E\|). \end{aligned}$$

On the other hand, by applying Lemma 4.1 to the moment system defining $\mathcal{P}_N$, we have

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

and hence

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

Combining the inequalities above, we have

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

For $\mathscr{H}(\mathcal{P}_N, \mathcal{P})$, it follows by Remark 2.1 that

$$\frac{1}{M}\mathscr{D}(P, Q) \leq d_{TV}(P, Q),$$

which implies $\mathscr{H}(\mathcal{P}_N, \mathcal{P}) \leq M\mathbb{H}_{TV}(\mathcal{P}_N, \mathcal{P})$. Since $\mathbb{H}_{TV}(\mathcal{P}_N, \mathcal{P}) \leq C\|\mu_N - \mu\|$, $\mathscr{H}(\mathcal{P}_N, \mathcal{P}) \leq CM\|\mu_N - \mu\|$. The proof is complete. ■

In the case when μ is constructed from independent and identically distributed samples of ω, 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 4.1 *Let ω^j, $j = 1, \cdots, N$ be independent and identically distributed sampling of ω and*

$$\mu_N := \frac{1}{N}\sum_{j=1}^{N} \psi(\omega^j).$$

Assume that Ω is a compact subset of $\mathbb{R}^k$ and ψ_i, $i = 1, \cdots, q$, is continuous on Ω. 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 3.1, 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 (4.26) and Proposition 3.1 noting that $\mathcal{P}_N$ is a compact set. ■

By Corollary 4.1, we can easily derive uniform exponential convergence of $v_N(x)$ to $v(x)$ in the case when the support set Ξ is compact. Note that this kind of convergence results may be obtained through a dual formulation of (2.3) which is a semi-infinite programming problem (see e.g. [44, Section 6.6]):

$$\begin{array}{ll} \max\limits_{x, \lambda_0, \lambda_1, \cdots, \lambda_q} & \lambda_0 + \sum_{j=1}^{q} \lambda_j \mu_j^N \\ \text{s.t.} & x \in X, \\ & \lambda_j \geq 0, \ \text{ for } j = p+1, \cdots, q, \\ & f(x, \xi(\omega)) \leq \lambda_0 + \sum_{j=1}^{q} \lambda_j \psi_j(\omega), \ \text{for a.e. } \omega \in \Omega. \end{array} \tag{4.27}$$

Indeed, if the dual gap is zero and the set of optimal solutions of (4.27) is uniformly bounded, then the convergence of optimal value of the dual program can be easily established as $\mu_N \to \mu$. We omit the details.

Note that Dupačová [16] recently investigates stability of one stage distributional robust optimization problem. She derives asymptotic convergence of optimal value of distributional robust

minimization problems where the distributional set is constructed by sample averaged 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 asymptotic convergence of optimal solutions and optimal values, see [16, 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 [16, Assumption 2.5] to continuity of $f(\cdot, \xi)$, we may recover [16, Theorem 3.3] through Theorem 3.2 without convexity of the feasible set X. In that case, the distributional set $\mathcal{P}$ and $\mathcal{P}_N$ are 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 [22] and Peel and McLachlan [29]. More recently, Zhu and Fukushima [49] 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 4.2 *Assume that P_l, $l = 1, \cdots, L$, is tight. Then $\mathcal{P}$ (resp. $\mathcal{P}_N$) is compact.*

Proof. $\mathcal{P}$ is a convex hull of a finite set $\mathcal{P}_v := \{P_l, l = 1, \ldots, L\}$, which is an image of the set under continuous mapping $F : (P_l, \cdots, P_L; t_1, \cdots, t_L) \to \sum_{l=1}^{L} \alpha_l P_l$. 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_i^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 [10, Theorem 1.4] that $P_i^L \in \mathcal{P}_v$ is tight.

Proposition 4.3 *Assume that*

$$\|P_l^N - P_l\|_{TV} \to 0, \text{for } l = 1, \cdots, L$$

as $N \to \infty$. Then for N sufficiently large

$$\mathbb{H}_{TV}(\mathcal{P}_N, \mathcal{P}) \leq \max\{\|P_l^N - P_l\|_{TV} : l = 1, \cdots, L\}. \tag{4.28}$$

Proof. Let

$$\widetilde{\mathcal{P}} := \{P_l : l = 1, \cdots, L\}$$

and

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

By [23, Proposition 2.1]

$$\mathbb{H}_{TV}(\mathcal{P}_N, \mathcal{P}) \leq \mathbb{H}_{TV}(\widetilde{\mathcal{P}}_N, \widetilde{\mathcal{P}}).$$

It suffices to show that

$$\mathbb{H}_{TV}(\widetilde{\mathcal{P}}_N, \widetilde{\mathcal{P}}) \leq \max\{\|P_l^N - P_l\|_{TV} : i = 1, \cdots, L\}. \tag{4.29}$$

Let ϵ denote the minimal distance between each pair of probability measures in $\widetilde{\mathcal{P}}$ under total variation metric, that is,

$$\epsilon := \min\{\|P_i - P_j\|_{TV} : i, j = 1, \cdots, L, i \neq j\}.$$

Let N_0 be sufficiently large such that for $N \geq N_0$,

$$\max\{\|P_l^N - P_l\|_{TV} : i = 1, \cdots, L\} \leq \frac{\epsilon}{8}. \tag{4.30}$$

Note that, for any l,

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

By above inequality and (4.30), we have

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

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

$$\mathbb{D}_{TV}(\widetilde{\mathcal{P}}_N, \widetilde{\mathcal{P}}) = \max\{\|P_l^N - P_l\|_{TV} : l = 1, \cdots, L\}. \tag{4.31}$$

On the other hand, for any l

$$\|P_l - P_m^N\|_{TV} \geq \|P_l - P_m\|_{TV} - \|P_m^N - P_m\|_{TV} \geq \frac{5}{8}\epsilon, \ \forall m = 1, \cdots, L, m \neq l.$$

Therefore

$$\mathbb{D}_{TV}(P_l, \widetilde{\mathcal{P}}_N) = \|P_l^N - P_l\|_{TV}, \text{for } l = 1, \cdots, L$$

and hence

$$\mathbb{D}_{TV}(\widetilde{\mathcal{P}}, \widetilde{\mathcal{P}}_N) = \max\{\|P_l^N - P_l\|_{TV} : l = 1, \cdots, L\}. \tag{4.32}$$

Combining (4.31) and (4.32), we obtain (4.28). ■

Corollary 4.2 *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.*

4.3 Distributional set due to Delage and Ye [14] and So [39]

Delage and Ye [14] propose to construct a distributional set through moment conditions which consist of the mean and covariance matrix. Specifically they consider the following distributional set:

$$\mathcal{P}(\mu_0, \Sigma_0, \gamma_1, \gamma_2) := \left\{ P \in \mathscr{P} : \quad \begin{array}{l} \mathbb{E}_P[\xi - \mu_0]^T \Sigma_0^{-1} \mathbb{E}_P[\xi - \mu_0] \leq \gamma_1 \\ 0 \preceq \mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)^T] \preceq \gamma_2 \Sigma_0 \end{array} \right\}, \tag{4.33}$$

where $\mu_0 \in \mathbb{R}^n$ is the true mean vector, $\Sigma_0 \in \mathbb{R}^{n \times n}$ 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. Note that in [14], a condition on the support set is explicitly imposed in the definition of the distribution 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, with or 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 [14] propose to construct an approximation of $\mathcal{P}$ with the distributional 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

$$\text{Prob}\{(\xi - \mu_0)^T \Sigma^{-1} (\xi - \mu_0) \leq \hat{R}^2\} = 1, \tag{4.34}$$

they proved that the true distribution of ξ lies in set $\mathcal{P}(\mu_N, \Sigma_N, \gamma_1^N, \mathbf{r}_2^N)$ with probability $1 - \delta$, see [14, Corollary 4]. The condition implies the support set of ξ is bounded. So [39] observes that the condition may be weakened to the following moment growth condition:

$$\mathbb{E}_P[\|\Sigma_0^{1/2}(\xi - \mu_0)\|_2^p] \leq (cp)^{p/2}. \tag{4.35}$$

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 $p \geq 1$, he shows that the true distribution of ξ lies in $\mathcal{P}_N$ with probability $1 - \delta$ for N is sufficiently large, where

$$\mathcal{P}_N := \mathcal{P}(\mu_N, \Sigma_N, \gamma_1^N, \gamma_2^N). \tag{4.36}$$

See [39, Theorem 9]. The significance of So's new results lies not only in the fact condition (4.35) is strictly weaker than (4.34) but the parameters γ_1^N and γ_2^N depend merely on the sample size N rather than the sample as in [14]. The latter will simplify our discussions later on.

Note that as 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 distributional set w.p.1:

$$\mathcal{P} : = \left\{ P \in \mathscr{P} : \quad \begin{array}{l} \mathbb{E}_P[\xi - \mu_0]^T \Sigma_0^{-1} \mathbb{E}_P[\xi - \mu_0] \leq 0 \\ 0 \preceq \mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)^T] \preceq \Sigma_0 \end{array} \right\}. \tag{4.37}$$

Observe first that since Σ_0^{-1} is positive definite, constraint

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

is equivalent to $\mathbb{E}_P[\xi - \mu_0] = 0$. Consequently we can write $\mathcal{P}$ as

$$\mathcal{P} \quad = \quad \left\{ P \in \mathscr{P} : \quad \begin{array}{l} \mathbb{E}_P[\xi] - \mu_0 = 0 \\ 0 \preceq \mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)^T] \preceq \Sigma_0 \end{array} \right\}. \tag{4.38}$$

A clear benefit of formulation (4.38) is that the system equalities and inequalities in brackets is linear w.r.t. operation $\mathbb{E}_P[\cdot]$. 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 4.4 *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) There exists a compact set $\mathcal{S} \subset \mathbb{R}^k$ such that $\mathrm{Prob}\{\xi \in \mathcal{S}\} = 1$ for all $P \in \mathcal{P}_N$ (resp. $P \in \mathcal{P}_N$).

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

$$\sup_{P \in \mathcal{P}_N} \int_{\Xi} \|\xi\|^{2+\varepsilon} P(d\xi) < \infty. \tag{4.39}$$

Proof. We only prove the conclusion for $\mathcal{P}_N$ as the proof for $\mathcal{P}$ is similar. We first show the tightness of set $\mathcal{P}_N$ in the weak topology.

Tightness. The first 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 6.1

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

Therefore

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

which means $\mathcal{P}_N$ is tight.

Closedness. The conclusion follows straightforwardly from Lemma 6.1 under condition (a). Let us consider condition (b). Let $\{P_k\} \in \mathcal{P}_N$ and $P_k \to P^*$ weakly. Under the bounded integral

condition (4.39), it follows from Lemma 6.1 that $\{P_k h(\cdot)\}$ is uniformly integrable, where $h(\cdot)$ denotes the inverse mapping of $\|\xi\|^2$. Indeed

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

The uniform integrability and weak convergence yield

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

which ensures

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

and

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

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

Proposition 4.4 gives sufficient conditions for compactness of the distributional sets $\mathcal{P}$ and $\mathcal{P}_N$. In the case when neither condition (i) nor (ii) is satisfied, the distributional set may not be compact. The example below illustrates that a distributional set similar to $\mathcal{P}$ is tight but not compact. In that case, we may consider the closure of the distributional set.

Example 4.1 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 $(\mathbb{R}, \mathcal{F})$. Consider the distributional set:

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

Let $\{P_k\}_{k\ge1}$ be a sequence of probability measures such that

$$P_k(\xi^{-1}(0)) = 1-\frac{1}{k} \text{ and } P_k(\xi^{-1}(\sqrt{k})) = \frac{1}{k}.$$

Let P^* be a probability measure which masses at 0, i.e., $P^*(\xi^{-1}(0))=1$. It is easy to observe that P_k converges to P^* weakly. Moreover

$$\mathbb{E}_{P_k}[\xi] = \frac{1}{\sqrt{k}} \text{ and } \mathbb{E}_{P_k}[\xi^2] = 1$$

for $k=1,\cdots$, $\mathbb{E}_{P^*}[\xi^2]=0$. Therefore

$$\lim_{k\to\infty} \mathbb{E}_{P_k}[\xi^2] \ne \mathbb{E}_{P^*}[\xi^2],$$

which means $P^* \notin \tilde{\mathcal{P}}$ and hence $\tilde{\mathcal{P}}$ is not closed. On the other hand, since $\mathbb{E}_P[\xi^2]$ is bounded for all $P\in\tilde{\mathcal{P}}$, by Dunford-Pettis theorem [3, Theorem 2.4.5], $\tilde{\mathcal{P}}$ is tight.

4.3.1 Error bound

In what follows, we estimate $\mathbb{H}_{TV}(\mathcal{P}, \mathcal{P}_N)$. To this end, we intend to express $\mathcal{P}_N$ through a linear system of $\mathbb{E}_P[\cdot]$ as well. But this seems to be impossible because the left hand side of the inequality

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

is nonlinear (indeed the inequality defines an ellipsoid in the space of $\mathbb{R}^k$ if we view $\mathbb{E}_P[\xi]$ as a variable)! In what follows, we try to approximate this ellipsoid constraint with two set of linear constraints representing an outer box which contains the ellipsoid and an inner box to be contained in the ellipsoid.

Let $\{\lambda_i\}_{i=1}^n$ denote the eigenvalues of Σ_N^{-1}, $\bar{\lambda} := \max_{i=1,\ldots,n} \lambda_i$ and $\underline{\lambda} := \min_{i=1,\ldots,n} \lambda_i$. Note that λ_i depends on N, so do $\bar{\lambda}$ and $\underline{\lambda}$. Define

$$\mathcal{P}_N^1 := \left\{ P \in \mathscr{P} : \begin{array}{l} -r_1^N \leq \mathbb{E}_P[\xi] - \mu_N \leq r_1^N, \\ \mathbb{E}_P[(\xi - \mu_N)(\xi - \mu_N)^T] \preceq \gamma_2^N \Sigma_N. \end{array} \right\}, \tag{4.40}$$

and

$$\mathcal{P}_N^2 := \left\{ P \in \mathscr{P} : \begin{array}{l} -r_2^N \leq \mathbb{E}_P[\xi] - \mu_N \leq r_2^N, \\ \mathbb{E}_P[(\xi - \mu_N)(\xi - \mu_N)^T] \preceq \gamma_2^N \Sigma_N. \end{array} \right\}, \tag{4.41}$$

where $r_1^N := (\bar{\lambda}\gamma_1^N)^{1/2} e$, $r_2^N := (\underline{\lambda}\gamma_1^N)^{1/2} e$, $e := (1, \ldots, 1)^T$. Note that r_1^N, r_2^N are vectors. We claim that

$$\mathcal{P}_N^2 \subset \mathcal{P}_N \subset \mathcal{P}_N^1. \tag{4.42}$$

To see this, let us show the second inclusion in (4.42). It suffices to show that inequality

$$\mathbb{E}_P[\xi - \mu_N]^T \Sigma_N^{-1} \mathbb{E}_P[\xi - \mu_N] \leq \gamma_1^N,$$

implies

$$|\mathbb{E}_P[\xi] - \mu_N| \leq r_1^N.$$

Since Σ_N^{-1} is symmetric and positive definite matrix, there exists an orthogonal matrix Q such that $\Sigma_N^{-1} = Q\Lambda Q^{-1}$, where $\Lambda := \mathrm{diag}\{\lambda_1^{-1}, \ldots, \lambda_n^{-1}\}$. Let $w^T := \mathbb{E}_P[\xi - \mu_N]^T Q \in \mathbb{R}^n$. Then $\|w\|_2 = \|\mathbb{E}_P[\xi - \mu_N]\|_2$ and $\mathbb{E}_P[\xi - \mu_N]^T \Sigma_N^{-1} \mathbb{E}_P[\xi - \mu_N] = w^T \Lambda w$, thereby

$$\|\mathbb{E}_P[\xi - \mu_N]\|_2^2 = \|w\|_2^2 \leq \bar{\lambda}\gamma_1^N.$$

This means $\|\mathbb{E}_P[\xi - \mu_N]\|_\infty \leq \sqrt{\bar{\lambda}\gamma_1^N}$ and hence $|\mathbb{E}_P[\xi] - \mu_N| \leq r_1^N$.

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 $A^T B$. The Frobenius norm of A, denoted by $\|A\|_F$, to is the square root of the trace of $A^T A$. 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, where Q is an orthogonal matrix. We define

$$M_+ := Q\mathrm{diag}\left\{\max\{\iota_1, 0\}, \ldots, \max\{\iota_n, 0\}\right\} Q^T$$

and

$$M_- := M - M_+.$$

The underlying purpose of this definition is to measure 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 4.2 *Let $A, B \in \mathbb{R}^{n\times n}$ be two symmetric matrices and $A \succeq 0$. The following assertion hold.*

(i) $\mathrm{tr}(AB) \leq \mathrm{tr}(AB_+)$.

(ii) $\|(A+B)_+\|_F \leq \|A_+\|_F + \|B_+\|_F$.

Proof. Part (i). Since $B_+ - B \succeq 0$ and $A \succeq 0$, by [11, 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 $M \in \mathbb{R}^{n\times n}$ be a symmetric matrix. It is well known that for $X \preceq 0$, $\|X - M\|_F$ attaints its minimum when $X = M_-$, see [24]. Using this argument, we have

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

where the first inequality is due to the fact that $A_- + B_- \preceq 0$. ■

Theorem 4.1 *Let $\mathcal{P}$ be defined as in (4.37) and $\mathcal{P}_N^1$ and $\mathcal{P}_N^2$ by (4.40) and (4.41) respectively. Assume that $\mathcal{P}, \mathcal{P}_1^N$ and $\mathcal{P}_2^N$ are closed. Then the following assertions hold.*

(i) There exists a positive constant C_1 depending on $\mathcal{P}$ such that

$$d_{TV}(Q, \mathcal{P}) \leq C_1(\|(\mathbb{E}_Q[(\xi-\mu_0)(\xi-\mu_0)^T] - \Sigma_0)_+\|_F + ||\mathbb{E}_Q[\xi] - \mu_0||) \tag{4.43}$$

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

(ii) There exists a positive constant C_2 such that

$$\begin{aligned} d_{TV}(Q, \mathcal{P}_N^1) \leq\ & \tilde{C}_2(||(\mathbb{E}_Q[\xi] - \mu_N - r_1^N)_+|| + ||(-\mathbb{E}_Q[\xi] + \mu_N - r_1^N)_+|| \\ & + \|(\mathbb{E}_Q[(\xi-\mu_0)(\xi-\mu_0)^T] - \gamma_2^N \Sigma_N)_+\|_F + ||\mu_0 - \mu_N||) \end{aligned} \tag{4.44}$$

for every $Q \in \mathscr{P}$ with $\|\mathbb{E}_Q[\xi]\| \leq \eta$, where η is a positive number and $\tilde{C}_2 = \eta C_2$.

(iii) There exists a positive constant C_3 such that

$$\begin{aligned} d_{TV}(Q, \mathcal{P}_N^2) \leq\ & \tilde{C}_3(||(\mathbb{E}_Q[\xi] - \mu_N - r_2^N)_+|| + ||(-\mathbb{E}_Q[\xi] + \mu_N - r_2^N)_+|| \\ & + \|(\mathbb{E}_Q[(\xi-\mu_0)(\xi-\mu_0)^T] - \gamma_2^N \Sigma_N)_+\|_F + ||\mu_0 - \mu_N||) \end{aligned}$$

for every $Q \in \mathscr{P}$ with $\|\mathbb{E}_Q[\xi]\| \leq \eta$, where η is a positive number and $\tilde{C}_3 = \eta C_3$.

Theorem 4.1 gives bounds for a probability measure Q deviating from $\mathcal{P}$ and $\mathcal{P}_N^1$ deviating from $\mathcal{P}_N^2$ in terms of the residuals of the linear inequality systems defining the distributional sets. We defer the proof to the appendix as it is long and technical. With the error bounds established in Theorem 4.1, we are ready to give an upper bound for the Hausdorff distance between $\mathcal{P}$ and $\mathcal{P}_N$ under total variation metric.

Theorem 4.2 *In the setting of Theorem 4.1, the following assertions hold.*

(i) There exist positive constants $\bar{C}_1$ such that for any small $\epsilon > 0$

$$\begin{aligned}\mathbb{H}_{TV}(\mathcal{P}, \mathcal{P}_N^1) \quad &\leq \quad \bar{C}_1(\max\left\{\|(\gamma_2^N \Sigma_N + C\|\mu_N - \mu_0\| I - \Sigma_0)_+\|_F, \|(\Sigma_0 - \gamma_2^N \Sigma_N)_+\|_F\right\} \\ &\quad +\|\mu_0 - \mu_N\| + \|r_1^N\|) \end{aligned} \tag{4.45}$$

for N sufficiently large. Here and in the next statements, C is a positive constant and I denotes the identity matrix.

(ii) There exist positive constants $\bar{C}_2$ such that

$$\begin{aligned}\mathbb{H}_{TV}(\mathcal{P}, \mathcal{P}_N^2) \quad &\leq \quad \bar{C}_2(\max\left\{\|(\gamma_2^N \Sigma_N + C\|\mu_N - \mu_0\| I - \Sigma_0)_+\|_F, \|(\Sigma_0 - \gamma_2^N \Sigma_N)_+\|_F\right\} \\ &\quad +\|\mu_0 - \mu_N\| + \|r_2^N\|) \end{aligned}$$

for N sufficiently large.

(iii)

$$\begin{aligned}\mathbb{H}_{TV}(\mathcal{P}_N, \mathcal{P}) \quad &\leq \quad \max(\bar{C}_1, \bar{C}_2)(\max\left\{\|(\gamma_2^N \Sigma_N + C\|\mu_N - \mu_0\| I - \Sigma_0)_+\|_F, \|(\Sigma_0 - \gamma_2^N \Sigma_N)_+\|_F\right\} \\ &\quad +\|\mu_0 - \mu_N\| + \|r_1^N\| + \|r_2^N\|) \end{aligned}$$

for N sufficiently large.

Theorem 4.2 indicates that $\mathbb{H}_{TV}(\mathcal{P}_N, \mathcal{P}) \to 0$ as $N \to \infty$ in that by definition $r_1^N \to 0$, $r_2^N \to 0$, $\mu_N \to \mu_0$, $\Sigma_N \to \Sigma_0$ and $\gamma_2^N \to 1$. The rate of convergence is exponential because by law of large numbers $\mu_N \to \mu_0$ and $\Sigma_N \to \Sigma_0$ converge at the exponential rate while all other quantities converge in a deterministic manner (r_1^N, r_2^N and γ_2^N depend only on sample size rather than samples).

Proof of Theorem 4.2. Part (i). It suffices to show that

$$\max\left\{\sup_{Q \in \mathcal{P}_N^1} d_{TV}(Q, \mathcal{P}), \sup_{Q \in \mathcal{P}} d_{TV}(Q, \mathcal{P}_N^1)\right\}$$

is bounded by the right hand side of (4.45). Let

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

For $Q \in \mathcal{P}_N^1$, $\mathbb{E}_Q[\xi]$ is bounded by $\|r_1^N\|$. The latter is only dependent of N and is bounded when N is sufficiently large. Consequently we may write τ_Q^N as $O(\mu_0 - \mu_N)$, meaning that it is a matrix bounded by $C\|\mu_N - \mu_0\| I$ for some positive constant C when $\|\mu_0 - \mu_N\|$ is close to 0. Through a simple rearrangement, we have

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

By Theorem 4.1 (i) and Lemma 4.2 (ii),

$$\begin{aligned} d_{TV}(Q,\mathcal{P}) &\leq C_1(\|(\mathbb{E}_Q[(\xi-\mu_N)(\xi-\mu_N)^T]+\tau_Q^N-\gamma_2^N\Sigma_N-C\|\mu_N-\mu_0\|I)_+\|_F \\ &\quad +\|(\gamma_2^N\Sigma_N+C\|\mu_N-\mu_0\|I-\Sigma_0)_+\|_F+||\mathbb{E}_Q[\xi]-\mu_0||) \\ &= C_1(\|(\gamma_2^N\Sigma_N+C\|\mu_N-\mu_0\|I-\Sigma_0)_+\|_F+||\mu_N-\mu_0||+\|r_1^N\|). \end{aligned} \tag{4.46}$$

The equality holds because the first term at the right hand side of the inequality is equal to zero. Likewise, for $Q\in\mathcal{P}$, Theorem 4.1 (ii) and Lemma 4.2 (ii)

$$\begin{aligned} d_{TV}(Q,\mathcal{P}_N^1) &\leq C_2(||(\mu_0-\mu_N-r_1^N)_+||+||(-\mu_0+\mu_N-r_1^N)_+|| \\ &\quad +\|(\mathbb{E}_Q[(\xi-\mu_0)(\xi-\mu_0)^T]-\Sigma_0)_+\|_F+\|(\Sigma_0-\gamma_2^N\Sigma_N)_+\|_F \\ &\quad +||\mu_0-\mu_N||) \\ &\leq 3C_2(||(\mu_0-\mu_N)||+||r_1^N||+\|(\Sigma_0-\gamma_2^N\Sigma_N)_+\|_F). \end{aligned} \tag{4.47}$$

Combining (4.46) and (4.47), we obtain (4.45).

Part (ii). Part (ii) is similar to part (i), we omit the details.

Part (iii). By (4.42) and the definition of $\mathbb{D}_{TV}$

$$\mathbb{D}_{TV}(\mathcal{P}_N,\mathcal{P})\leq\mathbb{D}_{TV}(\mathcal{P}_N^1,\mathcal{P})$$

and

$$\mathbb{D}_{TV}(\mathcal{P},\mathcal{P}_N)\leq\mathbb{D}_{TV}(\mathcal{P},\mathcal{P}_N^2).$$

Therefore

$$\mathbb{H}_{TV}(\mathcal{P}_N,\mathcal{P})\leq\max\{\mathbb{D}_{TV}(\mathcal{P}_N^1,\mathcal{P}),\mathbb{D}_{TV}(\mathcal{P},\mathcal{P}_N^2)\}\leq\max\{\mathbb{H}_{TV}(\mathcal{P}_N^1,\mathcal{P}),\mathbb{H}_{TV}(\mathcal{P},\mathcal{P}_N^2)\},$$

which yields (4.46) through Parts (i) and (ii). ■

It is important to note that Theorem 4.2 does not require condition (4.35). Indeed, the theorem does not indicate whether the true distribution is located in $\mathcal{P}_N$ albeit $\mathcal{P}_N$ may be arbitrarily close to $\mathcal{P}$. However, with the condition, we are guaranteed that $\mathcal{P}_N$ contains the true distribution with a probability at least $1-\delta$.

5 Robust equilibrium problem

In this section, we extend the asymptotic analysis of distributional robust optimization to robust equilibrium problems arising from robust games. 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 aims to minimize its expected disutility with the expectation being taken with respect to its *subjective* probability distribution of uncertainty in future. Mathematically, we can formulate an individual player i's problem as follows:

$$\vartheta_i(y_i,y_{-i}):=\min_{y_i\in Y_i}\max_{P_i\in\mathcal{P}_i}\mathbb{E}_{P_i}[f_i(y_i,y_{-i},\xi(\omega))], \tag{5.48}$$

where Y_i is a closed subset of $\mathbb{R}^{n_i}$, y_i denotes i's decision vector and y_{-i} the decision vectors of its rivals. The uncertainty is described by random variable ξ defined in space $(\Omega,\mathcal{F})$ and its true distribution is unknown. However, player i believes that the true distribution of ξ is in a distributional set, denoted by $\mathcal{P}_i$, and the mathematical expectation on the disutility function f_i

is taken with respect to $P_i \in \mathcal{P}_i$. Note that in this game, players face the same uncertainty but different player may have different subjective distributional set which relies heavily on availability of information of the uncertainty to them. The robust operation means that due to limited information on the future uncertainty, player i takes a conservative view on its expected disutility. To simplify our discussion, we assume that player i's feasible solution set Y_i is deterministic and independent of its competitor's action.

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

$$y_i^* \in \arg\min_{y_i \in Y_i} \max_{P_i \in \mathcal{P}_i} \mathbb{E}_{P_i}[f_i(y_i, y_{-i}^*, \xi(\omega))], \ \text{ for } \ i = 1, \cdots, m. \tag{5.49}$$

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; see also [25, 26] and references therein. More recently, Qu and Goh [31] propose a distributional robust version of the finite game where each player uses a distributional robust approach to deal with incomplete information of uncertainty. Our model may be viewed as an extension to continuous games.

Our focus here is on the case when an individual player builds its distributional set $\mathcal{P}_i$ through samples. We analyze convergence of the sample based robust equilibrium as sample size increases. 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 $y^N := (y_1^N, \cdots, y_m^N) \in Y_1 \times \cdots \times Y_m$ such that

$$y_i^N \in \arg\min_{y_i \in Y_i} \max_{P_i \in \mathcal{P}_i^N} \mathbb{E}_{P_i}[f_i(y_i, y_{-i}^N, \xi(\omega))], \ \text{ for } \ i = 1, \cdots, m. \tag{5.50}$$

To ease the exposition and notation, let

$$v_i(y) := \max_{P_i \in \mathcal{P}_i} \mathbb{E}_{P_i}[f_i(y_i, y_{-i}, \xi(\omega))] \tag{5.51}$$

and

$$v_i^N(y) := \max_{P_i \in \mathcal{P}_i^N} \mathbb{E}_{P_i}[f_i(y_i, y_{-i}, \xi(\omega))] \tag{5.52}$$

for $i = 1, \cdots, m$. Then (5.49) and (5.50) can be written as

$$\vartheta_i(y_i^*, y_{-i}^*) := \min_{y_i \in Y_i} v_i(y_i, y_{-i}), \ \text{ for } \ i = 1, \cdots, m$$

and

$$\vartheta_i(y_i^N, y_{-i}^N) := \min_{y_i \in Y_i} v_i^N(y_i, y_{-i}), \ \text{ for } \ i = 1, \cdots, m.$$

Let us define the set of random functions:

$$\mathscr{G}_E := \{g(\cdot) := f_i(y, \xi(\cdot)) : y \in Y, i = 1, \cdots, m\}, \tag{5.53}$$

where the subscript E indicates $\mathscr{G}_E$ is defined for the underlying functions of the equilibrium problems in order to differentiate it from a similar notation used in Section 3 for the robust optimization problems. Let $\mathscr{P}$ denote the set of all probability measures defined on $(\Omega, \mathcal{F})$ and $P, Q \in \mathscr{P}$. The pseudo-distance between P and Q is defined as:

$$\mathscr{D}_E(P, Q) := \sup_{g \in \mathscr{G}_E} \big|\mathbb{E}_P[g] - \mathbb{E}_Q[g]\big|. \tag{5.54}$$

For the set of probability measures $\mathcal{A}_k \subset \mathscr{P}$, $k = 1, 2$, let

$$\mathscr{D}_E(Q, \mathcal{A}_1) := \inf_{P \in \mathcal{A}_1} \mathscr{D}_E(Q, P), \quad \mathscr{D}_E(\mathcal{A}_1, \mathcal{A}_2) := \sup_{Q \in \mathcal{A}_1} \mathscr{D}_E(Q, \mathcal{A}_2)$$

and

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

We need assumptions parallel to Assumption 2.1 and Assumption 2.2.

Assumption 5.1 *Let $\mathcal{P}_i, \mathcal{P}_i^N$, $i = 1, \cdots, m$, be defined as in (5.48) and (5.50). There exists a weakly compact set $\mathscr{P}_i$ such that*

(a) $\mathcal{P}_i$ is compact in the weak topology and $\mathcal{P}_i \subset \hat{\mathcal{P}}_i$;

(b) for each N, $\mathcal{P}_i^N$ is a nonempty compact set in the weak topology and $\mathcal{P}_i^N \subset \hat{\mathcal{P}}_i$ when N is sufficiently large.

Assumption 5.2 *Let $f_i(y_i, y_{-i}, \xi)$ be defined as in (5.48) for $i = 1, \cdots, m$. For each fixed $\xi \in \Xi$, $f_i(y_i, y_{-i}, \xi)$ is Lipschitz continuous w.r.t. y_{-i} on Y_{-i} with Lipschitz modulus being bounded by $\kappa_{-i}(\xi)$, where*

$$\sup_{P \in \hat{\mathcal{P}}_i} \mathbb{E}_P[\kappa_{-i}(\xi)] < \infty \text{ for } i = 1, \cdots, m$$

and $\hat{\mathcal{P}}_i$ is given by Assumption 5.1.

Proposition 5.1 *Consider problem (5.51). Let $y_i \in \mathcal{Y}_i$ be fixed. Assume: (a) $\mathcal{P}_i$ is a compact set in the weak topology, (b) there exists $\alpha \in \mathbb{R}$ such that the upper level set $\mathrm{Lev}_\alpha \mathbb{E}_P[f(y_i, y_{-i}, \xi)] := \{P \in \mathcal{P} : \mathbb{E}_P[f(y_i, y_{-i}, \xi)] \geq \alpha\}$ is nonempty, (c) there exists a positive number M such that, for all $\xi \in \Xi$, $f(y_i, y_{-i}, \xi) \leq M$. Then there exists $P_i^* \in \mathcal{P}_i$ which attains the optimum and the maximum is finite.*

Proof. The proof can be obtained as in Proposition 2.2. We omit the details. ■

Proposition 5.2 *Assume: (a) $f_i(y_i, y_{-i}, \xi)$ is convex w.r.t. y_i on Y_i for each fixed y_{-i} and ξ, and Y_i is convex for $i = 1, \cdots, m$, (b) Assumptions 5.1 and 5.2 hold. Then problem (5.49) has an equilibrium.*

Proof. We use [34, Theorem 1] to prove the claim. It suffices to show that $v_i(y_i, y_{-i})$ is continuous and convex w.r.t. y_i on Y_i. The continuity can be shown similar to Proposition 2.1 (i) under Assumptions 5.1 and 5.2 whereas convexity is preserved under max operation w.r.t. the probability measure. ■

Assumption 5.3 *Let $\mathscr{P}$ be the set of probability measures and $\mathcal{P}_i, \mathcal{P}_i^N \subset \mathscr{P}$ are closed for all $i = 1, \ldots, m$. The following hold.*

(a) $\mathscr{H}(\mathcal{P}_i^N, \mathcal{P}_i) \to 0$ almost surely as $N \to \infty$;

(b) for any $\epsilon > 0$, there exist positive constants α and β (depending on ϵ) such that

$$\text{Prob}(\mathscr{D}_E(\mathcal{P}_i^N, \mathcal{P}_i) \geq \epsilon) \leq \alpha e^{-\beta N}, \ \forall i = 1, \ldots, m,$$

for N sufficiently large.

Theorem 5.1 *Let $\{y^N\}$ be a sequence of approximate robust equilibrium obtained from solving (5.50). Let Assumptions 5.1, 5.2 and 5.3 (a) hold. Under conditions of Proposition 5.2,*

$$\overline{\lim_{N\to\infty}} \, y^N \subset Y^*,$$

where Y^* denotes the set of robust Nash equilibria of (5.49). If, in addition, Assumption 5.3 (b) holds, then there exists positive constants α_1 and β_1 such that

$$\text{Prob}(d(y^N, Y^*) \geq \epsilon) \leq \alpha_1 e^{-\beta_1 N}, \ \forall i = 1, \ldots, m,$$

for N sufficiently large.

Proof. Let

$$\rho(x, y) := \sum_{i=1}^m \vartheta_i(x_i, y_{-i})$$

and

$$\hat{\rho}^N(x, y) := \sum_{i=1}^m \hat{\vartheta}_i^N(x_i, y_{-i}).$$

It is well-known (see e.g. [34]) that $y^* \in Y$ is a Nash equilibrium of the true problem (5.49) if and only if y^* solves the following minimization problem

$$\min_{x\in Y} \rho(x, y^*).$$

Likewise $y^N \in Y$ is a Nash equilibrium of the problem (5.50) if and only if y^N solves the following minimization problem

$$\min_{x\in Y} \hat{\rho}^N(x, y^N).$$

Assume without loss of generality (by taking a subsequence if necessary) that $\{y^N\}$ converges to y^* w.p.1. We show that w.p.1 $\hat{\rho}^N(x, y^N)$ converges to $\rho(x, y^*)$ uniformly with respect to x. Let us consider

$$\hat{\rho}^N(x, y^N) - \rho(x, y^*) = \hat{\rho}^N(x, y^N) - \hat{\rho}^N(x, y^*) + \hat{\rho}^N(x, y^*) - \rho(x, y^*).$$

Since $f_i(y_i, y_{-i}, \xi)$ is Lipschitz with respect to y_{-i} with modulus $\kappa_{-i}(\xi)$, we have

$$\begin{aligned}
|\hat{\rho}^N(x, y^N) - \hat{\rho}^N(x, y^*)| &\leq \sum_{i=1}^m |\vartheta_i^N(x_i, y_{-i}^N) - \vartheta_i^N(x_i, y_{-i}^*)| \\
&\leq \sum_{i=1}^m \left| \min_{x_i\in Y_i} \max_{P_i\in\mathcal{P}_i^N} \mathbb{E}_{P_i}[f_i(x_i, y_{-i}^N, \xi(\omega))] - \min_{x_i\in Y_i} \max_{P_i\in\mathcal{P}_i^N} \mathbb{E}_{P_i}[f_i(x_i, y_{-i}^*, \xi(\omega))] \right| \\
&\leq \sum_{i=1}^m \sup_{x_i\in Y_i} \max_{P_i\in\mathcal{P}_i^N} \left| \mathbb{E}_{P_i}[f_i(x_i, y_{-i}^N, \xi(\omega))] - \mathbb{E}_{P_i}[f_i(x_i, y_{-i}^*, \xi(\omega))] \right| \\
&\leq \sum_{i=1}^m \sup_{x_i\in Y_i} \max_{P_i\in\mathcal{P}_i^N} \mathbb{E}_{P_i}[\kappa_{-i}(\xi)] \|y_{-i}^N - y_{-i}^*\|.
\end{aligned}$$

The last term tends to 0 uniformly with respect to y when $N \to \infty$ because, by Assumptions 5.1-5.2, $\max_{P\in\hat{\mathcal{P}}_i} \mathbb{E}_P[\kappa_{-i}(\xi)] < \infty$ and $\mathcal{P}_i^N \subset \hat{\mathcal{P}}_i$ when N sufficiently large.

On the other hand,

$$\begin{aligned}
|\hat{\rho}^N(x,y^*) - \rho(x,y^*)| &\leq \sum_{i=1}^m |\vartheta_i^N(x_i, y^*_{-i}) - \vartheta_i(x_i, y^*_{-i})| \\
&\leq \sum_{i=1}^m \left| \min_{x_i\in Y_i} \max_{P_i\in\mathcal{P}_i^N} \mathbb{E}_{P_i}[f_i(x_i, y^*_{-i}, \xi(\omega))] - \min_{x_i\in Y_i} \max_{P_i\in\mathcal{P}_i} \mathbb{E}_{P_i}[f_i(x_i, y^*_{-i}, \xi(\omega))] \right| \\
&\leq \sum_{i=1}^m \sup_{x_i\in Y_i} \left| \max_{P_i\in\mathcal{P}_i^N} \mathbb{E}_{P_i}[f_i(x_i, y^*_{-i}, \xi(\omega))] - \max_{P_i\in\mathcal{P}_i} \mathbb{E}_{P_i}[f_i(x_i, y^*_{-i}, \xi(\omega))] \right| \\
&\leq \sum_{i=1}^m \mathscr{H}(\mathcal{P}_i^N, \mathcal{P}_i).
\end{aligned}$$

The last inequality follows from Assumption 5.3. It is well-known (see for instance [35, Theorem A1][3]) that the uniform convergence implies that the limit of the global minimizer of $\hat{\rho}^N(x, y^N)$ over set Y is a global minimizer of $\rho(x, y^*)$ over Y, which means y^* is a global minimizer of $\rho(x, y^*)$ over Y, hence a Nash equilibrium of the true robust Nash equilibrium problem (5.48). In the case when $\mathcal{P}_i^N$ converges to $\mathcal{P}_i$ at exponential rate under the pseudometric, we may apply [35, Theorem A1] to obtain exponential convergence of y^N to Y^* albeit the constants α_1 and β_1 may be different from α and β. We omit the details. ■

Before concluding this section, we outline some numerical schemes for solving (5.50). Our general idea is to reformulate the problem through dualization of the inner maximization problem and then consider the first order optimality conditions of the minimax problem. The dualization depends heavily on the structure of the distributional set $\mathcal{P}_i^N$. Here we consider the simplest case when the distributional set is built through first order moments:

$$\mathcal{P}_i^N := \left\{ P: \begin{array}{ll} \mathbb{E}_P[\psi_j(\xi(\omega))] = \mu_j^N, & \text{for } j = 1, \cdots, p \\ \mathbb{E}_P[\psi_j(\xi(\omega))] \leq \mu_j^N, & \text{for } j = p+1, \cdots, q \end{array} \right\}, i = 1, \cdots, m. \tag{5.55}$$

Following the discussions in [5, 44], we can obtain the Lagrangian dual of inner maximization of (5.50) as

$$\begin{array}{ll} \min\limits_{\lambda_0, \lambda_1, \cdots, \lambda_q} & \lambda_0 + \sum_{j=1}^q \lambda_j \mu_j^N \\ \text{s.t.} & \lambda_j \geq 0, \quad \text{for } j = p+1, \cdots, q, \\ & f_i(y_i, y_{-i}, \xi(\omega)) \geq \lambda_0 + \sum_{j=1}^q \lambda_j \psi_j(\xi(\omega)), \text{ a.e. } \omega \in \Omega, \end{array}$$

and hence the dual formulation of (5.50) as

$$\begin{array}{ll} \min\limits_{y_i, \lambda_0, \lambda_1, \cdots, \lambda_q} & \lambda_0 + \sum_{j=1}^q \lambda_j \mu_j^N \\ \text{s.t.} & y_i \in Y, \\ & \lambda_j \geq 0, \quad \text{for } j = p+1, \cdots, q, \\ & f_i(y_i, y_{-i}, \xi(\omega)) \geq \lambda_0 + \sum_{j=1}^q \lambda_j \psi_j(\xi(\omega)), \text{ a.e. } \omega \in \Omega. \end{array} \tag{5.56}$$

This is a semi-infinite programming problem. Under some standard constraint qualifications (see [7]), we can write down the first order optimality conditions of (5.56). We omit the detail as it is

[3] In the theorem, the convergence of $\bar{v}_N$ to v^* was proved under the condition that v^* is a unique global minimizer of $l(v)$ but the conclusion can be easily extended to the case when $l(v)$ has multiple minimizers in which case one can prove that $d(\bar{v}^N, V^*) \to 0$ where V^* denotes the set of global minimizers of $l(v)$.

not the main focus of this paper. In the case when the support set Ξ comprises of a finite number of points, i.e., $\Xi := \{\xi_1, \ldots, \xi_d\}$, the first order optimality conditions of problem (5.56) can be written as

$$\begin{cases} 0 \in \sum_{k=1}^d \tau_k \nabla_{y_i} f_i(y_i, y_{-i}, \xi_k) + \mathcal{N}_Y(y_i) \\ 0 = 1 - \sum_{k=1}^d \tau_k, \\ 0 = \mu_j^N - \sum_{k=1}^d \tau_k \psi_j(\xi_k),\ j = 1, \ldots, p, \\ 0 = \mu_j^N - \sum_{k=1}^d \tau_k \psi_j(\xi_k) + \varsigma_j,\ j = p+1, \ldots, q, \\ 0 = \varsigma_j \lambda_j,\ \lambda_j \geq 0,\ \varsigma_j \geq 0,\ j = p+1, \ldots, q, \\ 0 = \tau_k (f_i(y_i, y_{-i}, \xi_k) - \lambda_0 - \sum_{j=1}^q \lambda_j \psi_j(\xi_k)), \tau_k \geq 0,\ k = 1, \ldots, d, \\ 0 \leq f_i(y_i, y_{-i}, \xi_k) - \lambda_0 - \sum_{j=1}^q \lambda_j \psi_j(\xi_k),\ k = 1, \ldots, d. \end{cases} \tag{5.57}$$

By combining the m optimality conditions, we obtain the first order equilibrium conditions for (5.50), which is a deterministic variational inequality problem. We refer readers to monograph [17] about various numerical methods for the latter.

Acknowledgements. We would like to thank Werner Römisch, Alexander Shapiro, Anthony So and Daniel Kuhn for valuable discussions on a number of technical details. The first author would like to thank Yi Xu and Chao Ding for helpful discussions about Lemma 4.2. An earlier version of this paper was presented at Workshop II on Optimization Under Uncertainty at the Institute of Mathematical Sciences, National University of Singapore during 10-14 December 2012. We gratefully acknowledge insightful comments from the audience including Shabbir Ahmed, Darinka Dentcheva, Andrzej Ruszczyński, Melvyn Sim and Wolfram Wiesemann.

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 texts in statistics, Springer, NewYork, 2006.

[3] H. Attouch, G. Buttazzo and G. Michaille, *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*, MPS-SIAM series on optimization, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2005.

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

[5] M. Bertocchi, M. Vespucci and H. Xu, Robust approaches for two stage stochastic programs with applications in capacity expansion planning in energy industry, Manuscript, 2012.

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

[7] D. Bertsimas and M. Sim, The price of robustness, *Operations Research*, Vol. 52, pp. 35-53, 2004.

[8] 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.

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

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

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

[12] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

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

[14] 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.

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

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

[17] F. Facchinei and J-S Pang, *Finite-dimensional Variational Inequalities and Complementarity Problems*, Springer, New York, 2003.

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

[19] L. El Ghaoui and H. Lebret, Robust solutions to least-squares problems with uncertain data, *SIAM Journal on Matrix Analysis and Applications*, Vol. 18, pp. 1035-1064, 1997.

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

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

[22] 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.

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

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

[25] H. Jiang, S. Netessine and S. Savin, Robust newsvendor competition under asymmetric information, *Operations Research*, Vol. 59, pp. 254-261, 2011.

[26] E. Kardes, F. Ordonez and E. W. Hall, Discounted robust stochastic games and an application to queuing control, *Operations Research*, Vol. 59, pp. 365-382, 2011.

[27] 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.

[28] Y. Liu and H. Xu, Stability and sensitivity analysis of stochastic programs with second order dominance constraints. To appear in *Mathematical Programming Series A*, 2013.

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

[30] 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.

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

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

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

[34] J. B. Rosen, Existence and uniqueness of equilibrium points for concave N-person games, *Econometrica*, Vol. 33, pp. 520-534, 1965.

[35] R. Y. Rubinstein and A. Shapiro, *Discrete Events Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Methods*, John Wiley and Sons, New York, 1993.

[36] A. Ruszczynski and A. Shapiro, Optimization of convex risk functions, *Mathematics of Operations Research*, Vol. 31, pp. 433-452, 2006.

[37] 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.

[38] A. Ruszczynski and A. Shapiro, Conditional risk mappings, *Mathematics of Operations Research*, Vol. 31, pp. 544-561, 2006.

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

[40] 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.

[41] 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.

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

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

[44] A. Shapiro, D. Dentcheva and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, MPS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2009.

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

[46] W. Wiesemann, D. Kuhn, and B. Rustem, Robust Markov decision process, *Optimization Online*, 2012.

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

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

[49] 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.

6 Appendix

Lemma 6.1 *Let X be a separable metric space, P and P_n be Borel probability measures on X such that $\{P_n\}$ converges weakly to P, and $h : X \to \mathbb{R}$ measurable with $P(D_h) = 0$, where $D_h = \{x \in X : h \text{ is not continuous at } x\}$. Then it holds*

$$\lim_{n\to\infty} \int_X h(x) P_n(dx) = \int_X h(x) P(dx)$$

if the sequence $(P_n h^{-1})$ is uniformly integrable*, i.e.,*

$$\lim_{r\to\infty} \sup_{n\in\mathcal{N}} \int_{\{x\in X:|h(x)|\geq r\}} |h(x)| P_n(dx) = 0,$$

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

$$\sup_{n\in\mathcal{N}} \int_X |h(x)|^{1+\varepsilon} P_n(dx) < \infty \quad \text{for some } \varepsilon > 0.$$

The results of this lemma are summarized from [10, Theorem 5.4] and the preceding discussions.

Proof of Theorem 4.1. Part (i). Let $\psi_1(\omega) = \xi(\omega)$ and $\psi_2(\omega) = (\xi(\omega) - \mu_0)(\xi(\omega) - \mu_0)^T$. Then

$$\mathcal{P} = \left\{ P \in \mathscr{P} : \quad \begin{array}{l} \mathbb{E}_P[\psi_1(\omega)] = \mu_0 \\ \mathbb{E}_P[\psi_2(\omega)] \preceq \Sigma_0 \end{array} \right\}.$$

Let $\psi_E(\omega) := \psi_1(\omega), \psi_I(\omega) := \psi_2(\omega)^T$ and $u_E = \mu_0, u_I := \Sigma_0$. The proof is similar to that of Lemma 4.1. The only thing we need to explain is how to deal with the matrix constraint. Let $P \in \mathscr{P}$ and $\phi(\omega)$ be a P-integrable function. Recall that we write $\langle P, \phi\rangle = \mathbb{E}_P[\phi(\omega)]$ and $||P||_{TV} = \sup_{\|\phi\|\leq 1}\langle P, \phi\rangle$, where $\|\cdot\|$ denotes the maximum norm for function $\phi(\omega)$. Following a similar argument to that of Lemma 4.1, we have

$$\begin{aligned} d_{TV}(Q, \mathcal{P}) &= \inf_{P\in\mathcal{P}} d_{TV}(Q, P) \\ &= \inf_{P\in\{P:\mathbb{E}_P[\psi_E]=u_E, \mathbb{E}_P[\psi_I]\preceq u_I\}} \sup_{\|\phi(\omega)\|\leq 1} \langle Q - P, \phi\rangle \\ &= \sup_{\|\phi(\omega)\|\leq 1} \inf_{P\in\{P:\mathbb{E}_P[\psi_E]=u_E, \mathbb{E}_P[\psi_I]\preceq u_I\}} \langle Q - P, \phi\rangle. \end{aligned}$$

Let $\lambda_1 \in \mathbb{R}^k$ and $\Gamma_1 \in \mathbb{R}^{k\times k}_+$ denote the dual variables of constraints $\mathbb{E}_P[\psi_1(\omega)] = \mu_0$ and $\mathbb{E}_P[\psi_2(\omega)] \preceq \Sigma_0$, let

$$\Lambda := \{(\lambda_1, \Gamma_1) : \lambda_1 \in \mathbb{R}^k,\ \Gamma_1 \in \mathbb{R}^{k\times k}_+ \text{ is a symmetric matrix}\}.$$

By the duality theorem in conic linear problems [40],

$$\inf_{P\in\{P:\mathbb{E}_P[\psi_E]=u_E,\mathbb{E}_P[\psi_I]\preceq u_I\}} \langle Q-P,\phi\rangle \quad = \quad \sup_{(\lambda_1,\Gamma_1)\in\Lambda,\phi=\lambda^T\psi(\omega)} [\lambda \bullet (\mathbb{E}_Q[\psi(\omega)] - u)],$$

where $\lambda \bullet \psi(\omega) := \lambda_1^T\psi_1(\xi(\omega)) + \psi_2(\xi(\omega)) \bullet \Gamma_1$ and

$$\lambda \bullet (\mathbb{E}_Q[\psi(\omega)] - u) = (\mathbb{E}_Q[\psi_1(\omega)] - \mu_0)^T\lambda_1 + (\mathbb{E}_Q[\psi_2(\omega)] - \Sigma_0) \bullet \Gamma_1.$$

Consequently we arrive at

$$d_{TV}(Q,\mathcal{P}) = \sup_{(\lambda_1,\Gamma_1)\in\Lambda,\|\lambda^T\psi\|\le 1} \lambda \bullet (\mathbb{E}_Q[\psi(\omega)] - u). \tag{6.58}$$

The rest of the proof amounts to estimate (6.58) and this can be achieved by relaxing the semi-infinite, semidefinite constraint $\|\lambda^T\psi\| \le 1$ to a finite semidefinite constraints by considering a subindex set $(\omega_1,\ldots,\omega_n)$ of Ω. The resulting problem is a linear semi-definite programming problem and it is not difficult to show that the optimal value is bounded. Using a similar argument to that in the proof of Lemma 4.1, we can show that there exist λ_1^* and Γ_1^* such that

$$d_{TV}(Q,\mathcal{P}) = \lambda_1^{*T}(\mathbb{E}_Q[\xi] - \mu_0) + (\mathbb{E}_Q[(\xi-\mu_0)(\xi-\mu_0)^T] - \Sigma_0) \bullet \Gamma_1^*$$

and hence there exists a positive constant C_1 depending on $\mathcal{P}$ such that

$$\begin{aligned} d_{TV}(Q,\mathcal{P}) &\le C_1(\|(\mathbb{E}_Q[(\xi-\mu_0)(\xi-\mu_0)^T] - \Sigma_0)_+\|_F + ||\mathbb{E}_Q[\xi] - \mu_0||) \\ &\le C_1(\|(\mathbb{E}_Q[(\xi-\mu_0)(\xi-\mu_0)^T] - \Sigma_0)_+\|_F + ||\mathbb{E}_Q[\xi] - \mu_0||). \end{aligned}$$

Part (ii). To ease the notation, let

$$\psi_1^1(\omega) := \xi(\omega),\ \psi_2^1(\omega) := -\xi(\omega)$$

and

$$\psi_3^1(\omega) := (\xi(\omega)-\mu_0)(\xi(\omega)-\mu_0)^T.$$

Let

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

Then we can write

$$\mathbb{E}_Q[(\xi-\mu_N)(\xi-\mu_N)^T] = \mathbb{E}_Q[\psi_3^1(\omega)] + \tau_Q^N.$$

By assumption, $\|\mathbb{E}_Q[\xi]\| \le \eta$. Hence when $\|\mu_0 - \mu_N\|$ is small (close to 0), we may write τ_Q^N as $O(\mu_0 - \mu_N)$, meaning that it is a matrix bounded by $2\eta\|\mu_N - \mu_0\|I$ when $\|\mu_0 - \mu_N\|$ is close to 0. Let $\psi(\omega) := (\psi_1^1(\omega), \psi_2^1(\omega), \psi_3^1(\omega))^T$ and

$$\bar{u}_N = (r_1^N + \mu_N, r_1^N - \mu_N, \gamma_2^N\Sigma_N - O(\mu_0 - \mu_N))^T.$$

Based on the discussions above, we can present set $\mathcal{P}_N^1$ as follows:

$$\mathcal{P}_N^1 = \{P \in \mathscr{P} : \mathbb{E}_P[\psi(\omega)] \le \bar{u}_N\}.$$

A clear benefit of the presentation is that in the system $\mathbb{E}_P[\psi(\omega)] \leq \bar{u}_N$, only the right hand side depends on N and this will facilitate us to derive an error bound analogous to Lemma 4.1. That is,

$$d_{TV}(Q, \mathcal{P}_N^1) \quad = \quad \sup_{(\lambda_1,\lambda_2,\Gamma_3)\in\Lambda_1, \|\psi^1_{(\lambda_1,\lambda_2,\Gamma_3)}\|\leq 1} \lambda \bullet (\mathbb{E}_Q[\psi(\omega)] - \bar{u}_N), \tag{6.59}$$

where

$$\psi^1_{(\lambda_1,\lambda_2,\Gamma_3)}(\omega) := \lambda_1^T \psi_1^1(\xi(\omega)) + \lambda_2^T \psi_2^1(\xi(\omega)) + \psi_3^1(\xi(\omega)) \bullet \Gamma_3$$

and

$$\Lambda_1 := \{(\lambda_1, \lambda_2, \Gamma_3) : \lambda_i \in \mathbb{R}^n_+ \text{ for } i = 1, 2,\ \Gamma_3 \in \mathbb{R}^{k\times k}_+\}.$$

Following a similar argument to the proof of Lemma 4.1, we can show that there exist a finite $\lambda^* \in \Lambda_1$ such that

$$d_{TV}(Q, \mathcal{P}) \quad \leq \quad \lambda^* \bullet (\mathbb{E}_Q[\psi(\omega)] - \bar{u}_N). \tag{6.60}$$

The right hand side of the inequality above comprises three terms:

$$(\mathbb{E}_Q[\psi_1^1(\omega)] - \mu_N - r_1^N))^T \lambda_1^*,\ (\mathbb{E}_Q[\psi_2^1(\omega)] + \mu_N - r_1^N))^T \lambda_2^*$$

and

$$\Gamma_3^* \bullet (\mathbb{E}_Q[\psi_3^1(\omega)] - \gamma_2^N \Sigma_N + O(\mu_0 - \mu_N)).$$

In what follows, we estimate them. First, since $\lambda_1^* \in \mathbb{R}^n_+$,

$$\begin{aligned}(\mathbb{E}_Q[\psi_1^1(\omega)] - \mu_N - r_1^N))^T \lambda_1^* &= ((\mathbb{E}_Q[\psi_1^1(\omega)] - \mu_0 - r_1^N) + (\mu_0 - \mu_N))^T \lambda_1^* \\ &\leq ((\mathbb{E}_Q[\psi_1^1(\omega)] - \mu_0 - r_1^N)_+ + (\mu_0 - \mu_N))^T \lambda_1^* \\ &\leq \|\lambda_1^*\|(\|(\mathbb{E}_Q[\psi_1^1(\omega)] - \mu_0 - r_1^N)_+\| + \|\mu_0 - \mu_N\|) \end{aligned} \tag{6.61}$$

Likewise,

$$(\mathbb{E}_Q[\psi_2^1(\omega)] + \mu_N - r_1^N)^T \lambda_2^* \leq \|\lambda_2^*\|(\|(\mathbb{E}_Q[\psi_2^1(\omega)] + \mu_0 - r_1^N)_+\| + \|\mu_0 - \mu_N\|). \tag{6.62}$$

Next, using the definition and properties of Frobenius product, we have

$$\begin{aligned} & \Gamma_3^* \bullet (\mathbb{E}_Q[\psi_3^1(\omega)] - \gamma_2^N \Sigma_N + O(\mu_0 - \mu_N)) \\ = \ & \text{tr}\left(\Gamma_3^{*T}(\mathbb{E}_Q[\psi_3^1(\omega)] - \gamma_2^N \Sigma_N)\right) + \text{tr}\left(\Gamma_3^{*T} O(\mu_0 - \mu_N)\right) \\ \leq \ & \text{tr}\left(\Gamma_3^{*T}(\mathbb{E}_Q[\psi_3^1(\omega)] - \gamma_2^N \Sigma_N)_+\right) + \text{tr}\left(\Gamma_3^{*T} O(\mu_0 - \mu_N)\right) \\ \leq \ & 2\eta \|\Gamma_3^*\|_F(\|(\mathbb{E}_Q[\psi_3^1(\omega)] - \gamma_2^N \Sigma_N)_+\| + \|(\mu_0 - \mu_N)\|), \end{aligned} \tag{6.63}$$

Combining (6.60)-(6.63), we obtain (4.44).

Part (iii). Let ψ^N be same as in Part (ii) and $\bar{u}_N = (r_2^N + \mu_N, r_2^N - \mu_N, \gamma_2^N \Sigma_N - O(\mu_0 - \mu_N), O(\mu_0 - \mu_N))^T$. Then the proof of Part (iii) is similar to that of Part (ii). ■