Distributionally Robust Optimization with Matrix Moment Constraints: Lagrange Duality and Cutting Plane Methods[1]

Huifu Xu
School of Mathematical Sciences, University of Southampton, SO17 1BJ, Southampton, UK
(H.Xu@soton.ac.uk)

Yongchao Liu [2]
School of Mathematical Sciences, Dalian University of Technology, Dalian, 116024, China
(lyc@dlut.edu.cn)

Hailin Sun [3]
School of Economics and Management, Nanjing University of Science and Techonolyogy,
Nanjing, 210049, China
(hlsun@njust.edu.cn)

January 24, 2017

Abstract. A key step in solving minimax distributionally robust optimization (DRO) problems is to reformulate the inner maximization w.r.t. probability measure as a semiinfinite programming problem through Lagrange dual. Slater type conditions have been widely used for strong duality (zero dual gap) when the ambiguity set is defined through moments. In this paper, we investigate effective ways for verifying the Slater type conditions and introduce other conditions which are based on lower semicontinuity of the optimal value function of the inner maximization problem. Moreover, we propose two discretization schemes for solving the DRO with one for the dualized DRO and the other directly through the ambiguity set of the DRO. In the absence of strong duality, the discretization scheme via Lagrange duality may provide an upper bound for the optimal value of the DRO whereas the direct discretization approach provides a lower bound. Two cutting plane schemes are consequently proposed: one for the discretized dualized DRO and the other for the minimax DRO with discretized ambiguity set. Convergence analysis is presented for the approximation schemes in terms of the optimal value, optimal solutions and stationary points. Comparative numerical results are reported for the resulting algorithms.

Key Words. Matrix moment constraints, Slater type conditions, lower semicontinuity conditions, strong duality, random discretization, cutting plane methods

1 Introduction

One of the most challenging issues in decision analysis is to find an optimal decision under uncertainty. The solvability of a decision problem and the quality of an optimal decision rely

[1]The research is supported by EPSRC grant EP/M003191/1.

[2]The work of this author was carried out in the School of Mathematical Sciences, University of Southampton as a postdoctoral research fellow supported by EPSRC grant EP/M003191/1.

[3]This author's work is supported in part by National Natural Science Foundation of China #11401308 and Natural Science Foundation of Jiangsu Province, China #BK20140768

heavily on the information about the underlying uncertainties which are often mathematically represented by a vector of random variables. If a decision maker has complete information on the distribution of the random variables, then he can either obtain a closed form of the integral of the random functions in the problem and then convert it into a deterministic optimization problem, or alternatively use various statistical and numerical integration approaches such as scenario method [21], Monte carlo sampling method [41] and quadrature rules [12] to develop a deterministic approximation scheme and solve this using a standard linear/nonlinear programming code. The numerical efficiency of an approximation scheme and the quality of an optimal solution obtained from it depend on the structure (both the objective and constraints) and the scale (dimensionality) of the problem.

The situation may become far more complex if the decision maker does not have complete information on the distribution of the random variables. For instance, if the decision maker does not have any information other than the range of the values of the random variables, then it might be a reasonable option to choose an optimal decision on the basis of the extreme values of the random variables in order to mitigate the risks. This kind of decision making framework is known as *robust optimization* where the decision maker is extremely risk averse or lacks information on the distribution of the underlying random variables as described above. It is useful in some decision making problems particularly in engineering design [8, 3] where a design takes into account the extreme and rare event. However, the model may incur significant economic and/or computational costs in that excessive resources are used to prevent a rare event, resulting in numerical intractability or inefficiency. Over the past two decades, numerous efforts have been made to develop approximate schemes for solving robust optimization models which balance numerical tractability and quality of an optimal solution, see monograph by Ben-Tal et al. [4].

An alternative but possibly less conservative robust optimization model, which is known as *distributionally robust optimization* (DRO), involves a decision maker who is able to construct an ambiguity set of distributions with historical data, computer simulation or subjective judgements which contains the true distribution with certain confidence. In such circumstances, it is possible to choose an optimal decision on the basis of the worst distribution from the ambiguity set. For example, if we know roughly the nature of the distribution of random variables and can observe some samples, then we may use the classical maximum likelihood method to determine the parameters of the distribution and in that way construct a set of distributions if there is an inadequacy of the sample.

This kind of robust optimization framework may be traced back to the earlier work by Scarf [39] which was motivated to address incomplete information on the underlying uncertainty in supply chain and inventory control problems. In such problems, historical data may be insufficient to estimate the future distribution either because the 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. A larger distributional set which contains the true distribution may adequately address the risk from the ambiguity of the uncertainty. DRO model has found many applications in operations research, finance and management sciences. It has been well investigated through a number of further research works by Žáčková [54], Dupačová [15], Shapiro and Ahmed [42]. Over the past few years, it has gained substantial popularity through further contributions by Bertsimas and Popescu [7], Betsimas et al. [6], Delage and Ye [14], Goh and Sim [18], Hu and

Hong [22], Goldfarb and Iyengar [19], Mehrotra and Papp [28], Pflug, Pichler and Wozabal [31], Popescu [33], Wiesemann, Kuhn and Sim [49, 50] to name a few.

In this paper, we consider the following distributionally robust optimization problem:

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

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, $\xi : \Omega \to \Xi \subset \mathbb{R}^k$ is a vector of random variables defined on measurable space $(\Omega, \mathcal{F})$ equipped with sigma algebra $\mathcal{F}$, $\mathcal{P}$ is a set of probability distributions defined as

$$\mathcal{P} := \left\{ P \in \mathscr{P} : \begin{array}{ll} \mathbb{E}_P[\Psi_i(\xi)] = 0, & \text{for } i = 1, \cdots, p \\ \mathbb{E}_P[\Psi_i(\xi)] \preceq 0, & \text{for } i = p+1, \cdots, q \end{array} \right\}. \tag{1.2}$$

Here $\Psi_i : \Xi \to \mathbb{R}^{n_i \times n_i}$, $i = 1, \cdots, q$, is a symmetric matrix or a scalar with measurable random components, and $\mathscr{P}$ denotes the set of all probability distributions/measures over space $(\Omega, \mathcal{F})$; the notation $\preceq$ means that when Ψ_i is a matrix, its expected value must be negative semidefinite. In the case when $n_i = 1$, for $i = 1, \cdots, q$, Ψ_i reduces to a scalar function and (1.2) collapses to classical moment problems. Note that if we consider $(\Xi, \mathscr{B})$ as a measurable space equipped with Borel sigma algebra $\mathscr{B}$, then $\mathcal{P}$ may be viewed as a set of probability measures defined on $(\Xi, \mathscr{B})$ induced by the random variate ξ. So we may also write $\mathscr{P}(\Xi)$ for $\mathscr{P}$. Following the terminology in the literature of robust optimization, we call $\mathcal{P}$ the *ambiguity set* which indicates ambiguity of the true probability distribution of ξ at the point of decision making. As we will see in later discussions, Ψ_i may take some specific forms. Here we consider a general form in hope that our model covers a range of interesting moment problems. To ease the notation, we will use ξ to denote either the random vector $\xi(\omega)$ or an element of $\mathbb{R}^k$ depending on the context.

An important issue concerning DRO is numerical tractability of the robust formulation. For example, Delege and Ye [14] consider the DRO problem with ambiguity in both the mean and the covariance and demonstrate how their model can be solved in polynomial time when the support set is convex and compact. Goh and Sim [18] provide a tractable approximation scheme when the DRO is applied to a class of two stage stochastic programming problems. More recently, Wiesemann, Kuhn and Sim [51] provide a unified framework for DRO problems where the ambiguity set is constructed through some probabilistic and moment constraints. Under the Slater type conditions and essential boundedness of the support set, they provide a tractable reformulation of the problems.

In a slightly different direction, the DRO approach has been applied to tackle chance constrained stochastic programming problems where there is a lack of complete information on the true probability distribution. Zymler, Kuhn and Rustem [56] consider a class of robust chance constrained optimization problems with the ambiguity set being constructed through moment conditions and reformulate the robust constraint as semiinfinite constraints. In the case when the support set of the random variable covers the whole space and the underlying functions in the chance constraint are linear w.r.t. both the decision variables and the random variables, they reformulate the semiinfinite constraints as a system of semidefinite constraints and demonstrate the resulting semidefinite program (SDP for short) is numerically tractable. In a more recent development, Yang and Xu [48] extend the research to the case where the underlying functions in the chance constraint are nonlinear. A deficiency in these robust approaches is that

they may easily cause infeasibility of the robust chance constraint in that the ambiguity set may comprise a sequence of probability measures whose probability masses near the mean value and subsequently the robust probability of the inner random constraints (in the chance constraint) is equal to 1 when the mean lies in the inner feasible set. Of course, we are less concerned by the issue if the chance constraint is focused on the tail distribution of a loss function.

Our aim in this paper is to develop numerical methods for solving problem (1.1). Differing from the mainstream research in the literature of DRO, we concentrate on practical applicability of the computational methods without paying particular attention to numerical tractability in hope that the computational schemes and the underlying theory developed in this paper can be applied to a wide range of problems. Recall that a popular method for solving minimax distributionally robust optimization problems is to reformulate the inner maximization problem as a semiinfinite programming problem thorough Lagrange dual. A key theoretical question in our context is that under what conditions, problem (1.1) and its Lagrange dual problem are equivalent, i.e., the strong duality holds. The equivalence is well known when either the support set Ξ is compact in a finite dimensional space (see [43]) or the system of equalities and/or inequalities satisfy the Slater type conditions [40]. In the latter case, since the decision variables in the inner maximization problem are probability measures, one might wish to see whether a probability measure defined by an inequality moment constraint, i.e. $\langle P, \psi(\xi)\rangle \leq 0$ (hereafter $\langle \cdot, \cdot \rangle$ is a bilinear representation of the expected value of function ψ), lies in the "interior" of the feasible set (the ambiguity set $\mathcal{P}$). Unfortunately this kind of verification may turn out to be difficult at least technically since it concerns topological structure of the ambiguity set which is a subset in the space of probability measures. Shapiro [40] proposes an alternative way to characterize the condition which requires in this context the range of $\langle \cdot, \psi(\xi)\rangle$ over the cone of positive measures generated by $\mathcal{P}$ having nonempty intersection with the interior of $\mathbb{R}_+$. While this effectively addresses the theoretical issue we have just raised, the condition is often difficult to verify particularly when ψ is a vector of random functions or matrices because in that case we would need "coordination" of the components of ψ for the expected values. Likewise, in the equality case, the condition requires 0 to lie in the range of $\langle \cdot, \psi(\xi)\rangle$ which is difficult to verify when ψ is vector-valued. It would become even more challenging when $\mathcal{P}$ is composed of both equality and inequality constraints. This motivates us to develop effective approaches for verifying the conditions and look into other complementary conditions in this paper.

Another main challenge concerning (1.1) is to develop efficient numerical methods for solving the problem. When the support set Ξ is a finite set, the Lagrange dual is an ordinary matrix optimization problem, so we may apply the available codes on matrix optimization (see i.e., [47]) to solve the latter. It is also possible to solve problem (1.1) directly as a finite dimensional minimax saddle point problem. Indeed, Pflug and Wozabal [32] propose an iterative scheme for solving distributionally robust portfolio optimization problems where the inner maximization problem and the outer minimization problem are solved in turn. Mehrotra and Papp [28] extend the approach to a general class of DRO problems and design a process which generates a "cutting surface" of the inner optimal value function at each iterate. In the case when Ξ is well structured such as polyhedral or semialgebraic and the underlying functions (f and Ψ) are quadratic or linear, one may recast the semiinfinite inequality constraints as a semidefinite constraint through the well known S-lemma [34]. We note that this kind of formulation is the most popular approach in the literature of distributionally robust optimization, see for instance

[14, 51] and the references therein. Here we concentrate on the case where Ξ is neither a finite set nor has aforementioned structure and develop some computational methods which complement the existing numerical schemes for the DRO. As far as we are concerned, the main contributions of the paper can be summarized as follows.

- We present a detailed analysis of conditions for the strong Lagrange duality of the inner maximization problem, namely the Slater type conditions and the lower semicontinuity condition. The analysis concerning the Slater type conditions is based on Shapiro's [40, Proposition 3.4] which has been widely used in the literature of distributionally robust optimization with moment constraints but rarely scrutinized in detail. In Section 2, we present detailed discussions on the Slater type condition through a few practically interesting moment problems and demonstrate how the condition may be effectively verified. We also look into the duality conditions from lower semicontinuity of the optimal value function of the perturbed inner maximization problem and derive sufficient conditions which are easy to verify (Proposition 2.3). While the conditions are restrictive in general, we find that they are satisfied in a number of important cases such as when the support set Ξ is compact or Ψ_i is bounded, and this may effectively complement the popular Slater type condition in circumstances when the latter is difficult to be verified. Indeed, we can easily find some examples where the lower semicontinuity conditions are satisfied whereas the Slater type condition fails; see Example 2.7.

- We propose a discretization scheme based on Monte Carlo sampling for approximating the semiinfinite constraints of the Lagrange dual of the inner maximization problem. The approach is in line with the randomization scheme considered by Campi and Calafiore [11] and Anderson, Xu and Zhang [1] for mathematical programs with robust convex constraints. Under some moderate conditions, we demonstrate convergence of the optimal values, the optimal solutions and the stationary points obtained from the approximate problems as sample size increases (Theorems 3.1 and 3.2). Moreover, by observing the equivalence between the Monte Carlo discretization scheme and discretization of the ambiguity set $\mathcal{P}$ under strong duality, we propose a cutting plane method for solving the approximate DRO (1.1) directly as a finite dimensional minimax optimization problem and show convergence of the approximation scheme in terms of the optimal values and optimization solutions as sample size increases (Theorem 4.2). In the absence of strong duality, we observe that the discretization scheme via Lagrange duality provides an upper bound for the optimal value of the DRO when the sample size is sufficiently large whereas the direct discretization approach provides a lower bound for any sample size.

- Based on the aforementioned approximation schemes, we propose two algorithms for solving problem (1.1): a cutting plane algorithm for solving discretized dual problem (Algorithm 3.1) and a cutting plane method for the minimax DRO with discretized ambiguity set (Algorithm 4.1). We have carried out comparative numerical tests on the two algorithms through a portfolio optimization problem (Example 5.1) and a multiproduct newsvendor problem (Example 5.2) and conclude that the former is more sensitive to the change of the number of decision variables whereas the latter is more sensitive to the change of sample size.

Throughout the paper, we use the following notation. By convention, we use $\mathcal{S}^n$, $\mathcal{S}^n_+$ and

$\mathcal{S}^n_-$ to denote the space of symmetric matrices, cone of symmetric positive semidefinite matrices and cone of symmetric negative semidefinite matrices in the $n \times n$ matrix spaces $\mathbb{R}^{n\times n}$, and $\mathbb{R}^n_+$ to denote the cone of vectors with non-negative components in $\mathbb{R}^n$. For matrices $A, B \in \mathbb{R}^{n\times n}$, we write $A \circ B$ for the Frobenius inner product, i.e., $A \circ B = \text{trace}(AB)$, and $A \preceq B$ and $A \prec B$ to indicate $A - B$ being negative semidefinite and negative definite respectively. We use $(\mathcal{Z}, d)$ to represent an abstract metric space $\mathcal{Z}$ with metric d. For a set $\mathcal{C} \subset \mathcal{Z}$, we use by convention "int $\mathcal{C}$", "cl $\mathcal{C}$" and "conv $\mathcal{C}$" to denote its interior, closure and convex hull respectively. We write $d(z, \mathcal{D}) := \inf_{z'\in\mathcal{D}} d(z, z')$ for the distance from a point z to a set $\mathcal{D}$. For two sets $\mathcal{C}$ and $\mathcal{D}$, $\mathbb{D}(\mathcal{C}, \mathcal{D}) := \sup_{z\in\mathcal{C}} d(z, \mathcal{D})$ stands for the deviation/excess of set $\mathcal{C}$ from/over set $\mathcal{D}$. For a sequence of subsets $\{\mathcal{C}_k\}$ in a metric space, we follow the standard notation [38] by using $\limsup_{k\to\infty} \mathcal{C}_k$ to denote its outer limit, that is,

$$\limsup_{k\to\infty} \mathcal{C}_k = \left\{ x : \liminf_{k\to\infty} d(x, \mathcal{C}_k) = 0 \right\}.$$

For a set-valued mapping (also called multifunction in the literature) $\mathcal{A} : X \to 2^Y$, $\mathcal{A}$ is said to be *closed* at $\bar{x}$ if $x_k \in X$, $x_k \to \bar{x}$, $y_k \in \mathcal{A}(x_k)$ and $y_k \to \bar{y}$ implies $\bar{y} \in \mathcal{A}(\bar{x})$. $\mathcal{A}$ is said to be *outer semicontinuous* at $\bar{x} \in X$ if $\limsup_{x\to\bar{x}} \mathcal{A}(x) \subseteq \mathcal{A}(\bar{x})$. When $\mathcal{A}(x)$ is compact for each x, $\mathcal{A}(x)$ is upper semicontinuous (in the sense of Berge [5]) at $\bar{x}$ if and only if for every $\epsilon > 0$, there exists a constant $\delta > 0$ such that $\mathcal{A}(\bar{x} + \delta\mathcal{B}) \subset \mathcal{A}(\bar{x}) + \epsilon\mathcal{B}$. When the set-valued mapping $\mathcal{A}(\cdot)$ is bounded, the outer semicontinuity coincides with upper semicontinuity, see [38, Theorem 5.19] for the Euclidean space and [26, Theorem 4.27] for the general Hausdorff space.

2 Lagrange dual of the inner maximization problem in (1.1)

Let $x \in X$ be fixed. We consider the inner maximization problem of (1.1)

$$\begin{array}{ll} \sup\limits_{P\in\mathscr{M}_+} & \mathbb{E}_P[f(x,\xi)] \\ \text{s.t.} & \mathbb{E}_P[\Psi_i(\xi)] = 0, \text{for } i = 1, \cdots, p, \\ & \mathbb{E}_P[\Psi_i(\xi)] \preceq 0, \text{for } i = p+1, \cdots, q, \\ & \mathbb{E}_P[1] = 1, \end{array} \tag{2.3}$$

and its Lagrange dual

$$\begin{array}{ll} \inf\limits_{\lambda_0,\Lambda_1,\cdots,\Lambda_q} & \lambda_0 \\ \text{s.t.} & f(x,\xi) - \lambda_0 - \sum_{i=1}^q \Lambda_i \circ \Psi_i(\xi) \le 0, \forall\, \xi \in \Xi, \\ & \lambda_0 \in \mathbb{R}, \\ & \Lambda_i \succeq 0, \text{for } i = p+1, \cdots, q, \end{array} \tag{2.4}$$

where $\mathscr{M}_+$ denotes the positive linear space of all signed measures generated by $\mathscr{P}(\Xi)$.

As discussed in the introduction, a key step towards numerical solution of problem (1.1) is to establish equivalence between problems (2.3) and (2.4). In the literature of distributionally robust optimization, the equivalence has been well established under the circumstances where the support set Ξ is compact and $\Psi_i(\cdot)$ is continuous (see [43, page 312]), or the moment problem satisfies Slater type condition (see [40, 51] and references therein). This is underpinned by Shapiro's duality theorem ([40, Proposition 3.4]) for a general class of moment problems.

Note that in the case when the optimal value of problem (2.3) is $+\infty$, the dual problem (2.4) is infeasible. In that case, the equivalence is trivial. So our focus in this section is on the case when the optimal value of (2.3) is finite.

2.1 Slater type conditions

Let us start with the Slater type condition (STC for short). Following Shapiro's duality theory for moment problems, the condition in our context can be written as

$$(1,0) \in \text{int}\{(\langle P,1\rangle, \langle P,\Psi(\xi)\rangle) + \{0\}\times\{0\}\times\mathcal{K} : P \in \mathscr{M}_+\}, \tag{2.5}$$

where $\Psi(\xi) := (\Psi_1(\xi),\cdots,\Psi_p(\xi))$, $\langle P,\Psi(\xi)\rangle = \int_\Xi \Psi(\xi)P(d\xi)$ with the integration taken componentwise, the second 0 at the right hand side is the Cartesian product of zero matrices in the respective matrix spaces of $\mathbb{R}^{n_i\times n_i}$ corresponding to Φ_i for $i=1,\cdots,p$, and $\mathcal{K}_{q-p} := \mathcal{S}_+^{n_{p+1}}\times\cdots\times\mathcal{S}_+^{n_q}$; see [40, condition (3.12)] for general moment problems. Here we discuss how this condition may be satisfied and how it could be appropriately verified through some typical examples.

Example 2.1 (Reformulation of the STC and sufficient conditions for it) Consider the following moment problem:

$$\mathcal{P} := \left\{ P\in\mathscr{P}(\Xi): \begin{array}{ll} \mathbb{E}_P[\Psi_i(\xi)] = \mu_i, & \text{for } i=1,\cdots,p \\ \mathbb{E}_P[\Psi_i(\xi)] \preceq \mu_i, & \text{for } i=p+1,\cdots,q \end{array} \right\},$$

where $\Psi_i : \Xi\to\mathcal{S}^{n_i}$, $i=1,\cdots,q$, are continuous maps and $\mu_i\in\mathcal{S}^{n_i}$, $i=1,\cdots,q$ are constant matrices which could be either the true mean values of Ψ_i or their approximations/estimates. For the simplicity of notation we write Ψ_E for $(\Psi_1,\cdots,\Psi_p)$, Ψ_I for $(\Psi_{p+1},\cdots,\Psi_q)$, μ_E for $(\mu_1,\cdots,\mu_p)$ and μ_I for $(\mu_{p+1},\cdots,\mu_q)$. The Slater type condition in this case can be written as

$$(1,\mu_E,\mu_I) \in \text{int}\{(\langle P,1\rangle,\langle P,\Psi_E\rangle,\langle P,\Psi_I\rangle) + \mathcal{H}_1 : P\in\mathscr{M}_+\}, \tag{2.6}$$

where $\mathcal{H}_1 := \{0\}\times\{0\}\times\mathcal{K}_{q-p}$.

Proposition 2.1 *The following assertions hold.*

(i) Condition (2.6) is equivalent to

$$(\mu_E,\mu_I) \in \text{int}\{(\langle P,\Psi_E\rangle,\langle P,\Psi_I\rangle) + \mathcal{H}_2 : P\in\mathscr{P}(\Xi)\}, \tag{2.7}$$

where $\mathcal{H}_2 := \{0\}\times\mathcal{K}_{q-p}$.

(ii) Condition (2.7) is fulfilled if

$$\mu_E \in \text{int}\,\{\langle P,\Psi_E(\xi)\rangle : P\in\mathscr{P}(\Xi)\} \tag{2.8}$$

and there exists $P_E\in\mathscr{P}(\Xi)$ with $\langle P_E,\Psi_E(\xi)\rangle = \mu_E$ such that

$$0 \in \text{int}\,\{\langle P_E,\Psi_I(\xi)\rangle - \mu_I + \mathcal{K}_{q-p}\}. \tag{2.9}$$

In the case when $p = q$, i.e., there is no inequality constraint, condition (2.8) coincides with condition (2.7). Likewise, when $p = 0$, i.e., there is no equality constraint, (2.9) reduces to existence of $P \in \mathscr{P}(\Xi)$ such that

$$0 \in \text{int}\,\{\langle P, \Psi_I(\xi)\rangle - \mu_I + \mathcal{K}_{q-p}\}$$

which coincides with (2.7).

(iii) Condition (2.8) holds naturally in the case when

$$\{\langle P, \Psi_E(\xi)\rangle : P \in \mathscr{P}(\Xi)\} = \mathcal{S}^{n_1} \times \cdots \times \mathcal{S}^{n_p} \tag{2.10}$$

whereas condition (2.9) is fulfilled if there exists $P_E \in \mathscr{P}(\Xi)$ with $\langle P_E, \Psi_E(\xi)\rangle = \mu_E$ such that

$$\langle P_E, \Psi_I(\xi)\rangle - \mu_I \prec 0. \tag{2.11}$$

Proof. Part (i). Let (2.6) hold. Then there exists an open neighborhood of $\mu^* := (1, \mu_E, \mu_I)$, denoted by $\mathcal{U}$, such that $\mathcal{U} \subset \text{int}\{(\langle P, 1\rangle, \langle P, \Psi_E\rangle, \langle P, \Psi_I\rangle) + \mathcal{H}_1 : P \in \mathscr{M}_+\}$. Let $\mathcal{V} := \{P \in \mathscr{M}_+ : (\langle P, 1\rangle, \langle P, \Psi_E\rangle, \langle P, \Psi_I\rangle) \in \mathcal{U}\}$ and $P_0 \in \mathcal{V}$ such that $(\langle P_0, 1\rangle, \langle P_0, \Psi_E\rangle, \langle P_0, \Psi_I\rangle) = \mu^*$. Then

$$\begin{aligned}(\mu_E, \mu_I) &= (\langle P_0, \Psi_E\rangle, \langle P_0, \Psi_I\rangle)\\ &\in \{(\langle P, \Psi_E\rangle, \langle P, \Psi_I\rangle) : P \in \mathcal{V} \text{ with } \langle P, 1\rangle = 1\}\\ &\subset \text{int}\{(\langle P, \Psi_E\rangle, \langle P, \Psi_I\rangle) + \mathcal{H}_2 : P \in \mathscr{P}(\Xi)\}.\end{aligned}$$

Conversely, let (2.7) hold. Then for a sufficiently small positive number δ

$$\begin{aligned}(1, \mu_E, \mu_I) &\in \text{int}\{(\langle P, 1\rangle, \langle P, \Psi_E\rangle, \langle P, \Psi_I\rangle) + \mathcal{H}_1 : P \in \bigcup_{t\in(1-\delta,1+\delta)} t\mathscr{P}(\Xi)\}\\ &\subset \text{int}\{(\langle P, 1\rangle, \langle P, \Psi_E\rangle, \langle P, \Psi_I\rangle) + \mathcal{H}_1 : P \in \mathscr{M}_+\},\end{aligned}$$

which yields (2.6).

Part (ii). Conditions (2.8) and (2.9) guarantee existence of $P_E \in \mathscr{P}(\Xi)$ such that

$$\langle P_E, \Psi_E\rangle = \mu_E \in \text{int}\,\{\langle P, \Psi_E(\xi)\rangle : P \in \mathscr{P}(\Xi)\}$$

and

$$\mu_I \in \text{int}\,\{\langle P_E, \Psi_I(\xi)\rangle + \mathcal{K}_{q-p}\}$$

which imply (2.7). The equivalence statements (in the equality only case and inequality only case) are obvious.

Part (iii). Condition (2.10) implies that (2.8) holds trivially. Condition (2.11) means $\langle P_E, \Psi_I(\xi)\rangle - \mu_I \in -\text{int}\mathcal{K}_{q-p}$, which is equivalent to (2.9). ■

The proposition shows how the complex STC (2.6) can be examined through (2.7) and further through (2.9). Condition (2.11) is widely known as the Slater condition for inequality systems. The discussions show that the STC is weaker than the well known Slater condition.

We now turn to consider the case when $\mathcal{P}$ comprises a single matrix moment constraint.

Example 2.2 (STC for a single matrix moment constraint) Let

$$\Psi(\xi) = (\xi - \mu)(\xi - \mu)^T - \Sigma,$$

where ξ is a random vector with support set Ξ in $\mathbb{R}^n$, μ and Σ are either the true mean value and covariance matrix respectively or their estimates. Consider two types of moment conditions: one is inequality constrained and the other is equality constrained. The former is often used when a decision maker does not have complete information on the true mean value and/or covariance whereas the latter corresponds to the circumstance when the true covariance is known. We discuss them in sequel.

(a) With incomplete information of the mean and/or covariance, the moment problem is often written as

$$\mathbb{E}_P[(\xi - \mu)(\xi - \mu)^T] \preceq \Sigma,$$

where Σ is some positive definite matrix. Let Σ_0 denote the true covariance matrix (corresponding to the unknown true probability distribution of ξ) and assume that $\Sigma_0 \prec \Sigma$. Note that following the analysis as in Example 2.1 we can recast condition (2.5) as

$$0 \in \text{int}\{\langle P, \Psi(\xi)\rangle + \mathcal{S}_+^n : P \in \mathscr{P}(\Xi)\}. \tag{2.12}$$

It is easy to observe that condition (2.12) holds in that $\langle P_0, \Psi(\xi)\rangle \prec 0$ under the assumption $\Sigma_0 \prec \Sigma$ and any $n \times n$ positive definite matrix lies in the interior of $\mathcal{S}_+^n$.

(b) In the equality constraint case, the moment condition becomes

$$\mathbb{E}_P[(\xi - \mu)(\xi - \mu)^T] = \Sigma_0,$$

and the Slater type condition becomes $0 \in \text{int}\{\langle P, \Psi(\xi)\rangle : P \in \mathscr{P}(\Xi)\}$. The condition is fulfilled if $\Sigma_0 \in \text{int conv}\{(\xi - \mu)(\xi - \mu)^T : \xi \in \Xi\}$. The latter is automatically satisfied when $\Xi = \mathbb{R}^n$, see Proposition 2.2 below.

The example shows how condition (2.6) is verified through a different argument for equality and inequality matrix moment constraints.

Proposition 2.2 (Image of covariance mapping over $\mathscr{P}(\Xi)$) *If $\Xi = \mathbb{R}^n$, then*

$$\mathcal{S}_+^n = \left\{\mathbb{E}_P[(\xi - \mu)(\xi - \mu)^T] : P \in \mathscr{P}(\Xi)\right\}. \tag{2.13}$$

Proof. Observe that the right hand side of (2.13) is the image of the covariance mapping $\mathbb{E}_P[(\xi - \mu)(\xi - \mu)^T]$ over the space of probability measures $\mathscr{P}(\Xi)$. It suffices to show

$$\mathcal{S}_+^n \subseteq \left\{\mathbb{E}_P[(\xi - \mu)(\xi - \mu)^T] : P \in \mathscr{P}(\Xi)\right\}$$

because the opposite inclusion always holds. Let $M \in \mathcal{S}_+^n$ be any positive semidefinite matrix with eigenvalue λ_j and normalized eigenvector q_j for $j = 1, \cdots, n$. Let $\xi^j := \mu + \sqrt{n\lambda_j} q_j$ and P_j, $j = 1, \cdots, n$, denote the Dirac probability measure at ξ^j and $\hat{P} := \sum_{j=1}^n \frac{1}{n} P_j$. Then $\hat{P} \in \mathscr{P}(\Xi)$ and

$$\mathbb{E}_{\hat{P}}[(\xi - \mu)(\xi - \mu)^T] = \sum_{j=1}^n \frac{1}{n} \times n\lambda_j q_j q_j^T = \sum_{j=1}^n \lambda_j q_j q_j^T = M.$$

The conclusion follows. ■

In many practical cases, covariance constraint is often coupled by mean value constraints. Let us consider a few examples as such.

Example 2.3 (STC for matrix moments due to Delage and Ye [14] and So [45]) Consider ambiguity set

$$\mathcal{P} := \left\{ P \in \mathscr{P}(\Xi) : \quad \begin{array}{l} \mathbb{E}_P[\xi - \mu_0]^T \Sigma_0^{-1} \mathbb{E}_P[\xi - \mu_0] \preceq \gamma_1 \\ 0 \preceq \mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)^T] \preceq \gamma_2 \Sigma_0 \end{array} \right\}, \tag{2.14}$$

where γ_1 and γ_2 are nonnegative constants. The ambiguity set has first been considered by Delage and Ye [14] and further studied by So [45]. It is easy to observe that the inequality

$$\mathbb{E}_P[\xi - \mu_0]^T \Sigma_0^{-1} \mathbb{E}_P[\xi - \mu_0] \preceq \gamma_1$$

can be equivalently written as

$$\mathbb{E}_P \left[\left(\begin{array}{cc} -\Sigma_0 & \mu_0 - \xi \\ (\mu_0 - \xi)^T & -\gamma_1 \end{array} \right) \right] \preceq 0.$$

Thus $\mathcal{P}$ can be written as

$$\mathcal{P} = \left\{ P \in \mathscr{P}(\Xi) : \quad \begin{array}{l} \mathbb{E}_P \left[\left(\begin{array}{cc} -\Sigma_0 & \mu_0 - \xi \\ (\mu_0 - \xi)^T & -\gamma_1 \end{array} \right) \right] \preceq 0 \\ 0 \preceq \mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)^T] \preceq \gamma_2 \Sigma_0 \end{array} \right\}.$$

When $\gamma_i > 0$ for $i = 1, 2$, the moment constraints (2.14) satisfy the Slater type constraint qualification, see [46, Theorem 3]. However, when $\gamma_1 = 0$, the constraint qualification fails. To see this, let us note that matrix $\mathbb{E}_P \left[\left(\begin{array}{cc} -\Sigma_0 & \mu_0 - \xi \\ (\mu_0 - \xi)^T & -\gamma_1 \end{array} \right) \right]$ can never be negative definite in that by Schur complement for the matrix to be negative definite, we would need $0 - (\mu_0 - \mathbb{E}[\xi])^T (-\Sigma_0)^{-1} (\mu_0 - \mathbb{E}[\xi]) < 0$ which will never happen. Nevertheless, if we rewrite the ambiguity set as

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

then the Slater type condition holds, see [46, Theorem 3] for details.

Example 2.4 (STC for a variation of moment system (2.14)) Consider the following ambiguity set

$$\mathcal{P} = \left\{ P \in \mathscr{P}(\Xi) : \quad \begin{array}{l} |\mathbb{E}_P[\xi - \mu_0]| \leq \gamma_1 e \\ \|\mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)^T] - \Sigma_0\|_2 \leq \gamma_2 \end{array} \right\},$$

where γ_1 and γ_2 are small positive numbers, e is a vector with components of ones, $|a|$ denotes the absolute value of a vector a with the absolute value taken componentwise, and $\|\cdot\|_2$ denotes the spectral norm of a matrix. Using the property of the norm, we can reformulate the ambiguity set as

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

If $\gamma_1 > 0$ and $\gamma_2 > 0$, then there exists a probability measure P_0 such that $\mathbb{E}_{P_0}[\xi] = \mu_0$, $\mathbb{E}_{P_0}[(\xi - \mu_0)(\xi - \mu_0)^T = \Sigma_0$, and the strict inequalities of system of moment conditions hold. Following the remark after Proposition 2.1, we conclude that the Slater type condition holds.

Example 2.5 (STC for the moment system due to [27]) Consider the following ambiguity set

$$\mathcal{P} := \left\{ P \in \mathscr{P}(\Xi) : \begin{array}{l} |\mathbb{E}_P[\xi - \mu_0]| \leq \gamma_1 e \\ \|\mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)^T] - \Sigma_0\|_{\max} \leq \gamma_2 \end{array} \right\},$$

where $\|A\|_{\max} = \max |a_{ij}|$. It is easy to verify that $\|\cdot\|_{\max}$ is a norm for the matrix but without the sub-multiplicative property. The ambiguity set is considered in [27]. Let k be the dimension of random vector ξ, $q = \frac{k^2+3k}{2}$, $\psi_I(\xi) = \xi - \mu_0$ and $\psi_J(\xi)$ denote the elements of the upper triangular of matrix $(\xi - \mu_0)(\xi - \mu_0)^T - \Sigma_0$. Then we can reformulate $\mathcal{P}$ as

$$\mathcal{P} = \left\{ P \in \mathscr{P}(\Xi) : \begin{array}{l} \Psi_I(\xi) - \gamma_1 \leq 0 \\ -\Psi_I(\xi) - \gamma_1 \leq 0 \\ \Psi_J(\xi) - \gamma_2 \leq 0 \\ -\Psi_J(\xi) - \gamma_2 \leq 0 \end{array} \right\}.$$

Analogous to Example 2.4, the Slater condition is satisfied when $\gamma_1 > 0$ and $\gamma_2 > 0$.

2.2 Lower semicontinuity condition

We now study a different condition which is fundamentally based on Shapiro's result [40, Proposition 2.4]. To this end, we consider the following perturbation of problem (2.3)

$$\begin{array}{ll} \min\limits_{P \in \mathscr{P}(\Xi)} & \mathbb{E}_P[-f(x,\xi)] \\ \text{s.t.} & P \in \mathcal{P}(Y), \end{array} \tag{2.15}$$

where $Y = (Y_1, \cdots, Y_q)$ and $Y_i \in \mathcal{S}^{n_i}$, $i = 1, \cdots, q$, is in a small neighborhood of 0. To simplify the notation, here and later on we mean 0 is in appropriate space without indicating its dimension. Let

$$\mathcal{P}(Y) := \left\{ P \in \mathscr{P}(\Xi) : \begin{array}{ll} \mathbb{E}_P[\Psi_i(\xi)] + Y_i = 0, & \text{for } i = 1, \cdots, p \\ \mathbb{E}_P[\Psi_i(\xi)] + Y_i \preceq 0, & \text{for } i = p+1, \cdots, q \end{array} \right\}. \tag{2.16}$$

Let $v(Y)$ denote the optimal value of problem (2.15). By [40, Proposition 2.3], problem (2.15) satisfies the strong duality if and only if $v(\cdot)$ is lower semicontinuous at point 0. A sufficient condition for the latter is that $\mathcal{P}(Y)$ is weakly compact for each fixed Y and $\mathcal{P}(\cdot)$ is upper semicontinuous at 0. In what follows, we develop sufficient conditions for the required property of $\mathcal{P}(\cdot)$.

Recall that for a sequence of probability measures $\{P_N\} \subset \mathscr{P}(\Xi)$, P_N is said to converge to $P \in \mathscr{P}(\Xi)$ *weakly* if

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

for each bounded and continuous function $h : \Xi \to \mathbb{R}$. For a set of probability measures $\mathcal{A} \subset \mathscr{P}(\Xi)$, $\mathcal{A}$ is said to be weakly compact w.r.t. topology of weak convergence if every

sequence $\{P_N\} \subset \mathcal{A}$ contains a subsequence $\{P_{N'}\}$ and $P \in \mathcal{A}$ such that $P_{N'} \to P$. $\mathcal{A}$ is said to be *tight* if for any $\epsilon > 0$, there exists a compact set $\Xi^\epsilon \subset \Xi$ such that $\inf_{P\in\mathcal{A}} P(\Xi^\epsilon) > 1 - \epsilon$. In the case when $\mathcal{A}$ is a singleton, it reduces to the tightness of a single probability measure. $\mathcal{A}$ is said to be *closed* (under the topology of weak convergence) if for any sequence $\{P_N\} \subset \mathcal{A}$ with $P_N \to P$ weakly, we have $P \in \mathcal{A}$.

By the well-known Prokhorov's theorem (see [35, 2]), a closed set $\mathcal{A}$ of probability measures is *weakly compact* if and only if it is tight. In particular, since Ξ is a set in $\mathbb{R}^k$, if Ξ is a compact set, then the set of all probability measures on $(\Xi, \mathscr{B})$ is weakly compact with respect to topology of weak convergence; see [29].

For two probability measures $P_1, P_2 \in \mathscr{P}(\Xi)$, the Prokhorov metric [36] is

$$\pi(P_1, P_2) := \inf\{\epsilon > 0 : P_1(A) \le P_2(A^\epsilon) + \epsilon \text{ and } P_2(A) \le P_1(A^\epsilon) + \epsilon \ \ \forall A \in \mathscr{B}\},$$

where $A^\epsilon := \bigcup_{a\in A} \epsilon\mathcal{B}(a)$ and $\mathcal{B}(a)$ denotes the unit ball centered at point a. Since Ξ is a set in $\mathbb{R}^m$, the convergence of probability measures in the Prokhorov metric is equivalent to weak convergence.

Assumption 2.1 (Sufficient conditions for tightness and closedness of $\mathcal{P}(Y)$) (a) There exists a tight subset of probability measures, denoted by $\hat{\mathcal{P}} \subset \mathscr{P}(\Xi)$, such that $\mathcal{P}(Y) \subset \hat{\mathcal{P}}$ for all Y close to 0; (b) $\Psi_i(\cdot)$, $i = 1, \cdots, q$, is continuous over Ξ and every element $\psi^i_{jt}(\xi)$ of $\Psi_i(\cdot)$ is uniformly integrable, that is,

$$\lim_{r\to\infty} \sup_{P\in\mathcal{P}} \int_{\{\xi\in\Xi, |\psi^i_{jt}(\xi)|\ge r\}} |\psi^i_{jt}(\xi)| P(d\xi) = 0$$

for $i = 1, \cdots, q; j, t = 1, \ldots, n_i$.

A sufficient condition for Assumption 2.1 (a) is that there are positive constants τ and C such that

$$\sup_{P\in\hat{\mathcal{P}}} \int_\Xi \|\xi\|^{1+\tau} P(d\xi) < C. \tag{2.17}$$

Likewise, when ψ^i_{jt} is a continuous function, a sufficient condition for Assumption 2.1 (b) is that there exists a positive constant τ such that

$$\sup_{P\in\mathcal{P}} \int_\Xi |\psi^i_{jt}(\xi)|^{1+\tau} P(d\xi) < \infty, \tag{2.18}$$

for $i = 1, \cdots, q; j, t = 1, \ldots, n_i$. Condition (2.18) holds trivially when $\psi^i_{jt}(\xi)$ is bounded.

Lemma 2.1 (Topological properties of $\mathcal{P}(Y)$) *Under Assumption 2.1, the following assertions hold.*

(i) For each fixed Y close to 0, $\mathcal{P}(Y)$ is weakly compact;

(ii) $\mathcal{P}(\cdot)$ is upper semicontinuous at 0 in the sense of Berge [5], that is, for any $\epsilon > 0$, there exists $\delta > 0$ such that $\mathcal{P}(Y) \subseteq \mathcal{P}(0) + \epsilon\mathcal{B}$ for all Y with $\|Y\| \le \delta$, where $\mathcal{B}$ denotes the unit ball in the space of $\mathcal{P}$ under Prokhorov metric.

Proof. Part (i). Let Y be fixed. Under Assumption 2.1 (a), $\mathcal{P}(Y)$ is tight because any subset of a tight set is tight. By Prokhorov's theorem, it suffices to show that $\mathcal{P}(Y)$ is closed. Let $\{P_k\} \subset \mathcal{P}(Y)$ be a sequence of probability measures such that P_k converges to P weakly. We show $P \in \mathcal{P}(Y)$. Under Assumption 2.1 (b), it follows by [46, Lemma 1],

$$\lim_{k\to\infty} \int_{\xi\in\Xi} \psi^i_{jt}(\xi) P_k(d\xi) = \int_{\xi\in\Xi} \psi^i_{jt}(\xi) P(d\xi).$$

Therefore

$$\int_{\xi\in\Xi} \Psi_i(\xi) P(d\xi) + Y = \lim_{k\to\infty} \int_{\xi\in\Xi} \Psi_i(\xi) P_k(d\xi) + Y \begin{cases} = 0 & \text{for } i = 1, \cdots, p, \\ \preceq 0 & \text{for } i = p+1, \cdots, q, \end{cases}$$

which means $P \in \mathcal{P}(Y)$.

Part (ii). Let $\{Y^k\}$ be a sequence converging to 0. By the definition of outer semicontinuity, we only need to consider the points with $\mathcal{P}(Y^k) \neq \emptyset$. Let $P_k \in \mathcal{P}(Y^k)$. By [9, Theorem 5.1], the tightness of $\hat{\mathcal{P}}$ ensures that $\{P_k\}$ has a subsequence $\{P_{k_i}\}$ such that $P_{k_i} \to P^*$ weakly. Using a similar argument to that of Part (i), we have

$$\lim_{k_i\to\infty} \int_{\xi\in\Xi} \Psi_i(\xi) P_{k_i}(d\xi) + Y^{k_i} = \int_{\xi\in\Xi} \Psi_i(\xi) P^*(d\xi) + 0 \begin{cases} = 0 & \text{for } i = 1, \cdots, p, \\ \preceq 0 & \text{for } i = p+1, \cdots, q, \end{cases} \tag{2.19}$$

which means $P^* \in \mathcal{P}(0)$. This shows the set-valued mapping $\mathcal{P}(Y)$ is outer semicontinuous. On the other hand, since Ξ is a compact set of $\mathbb{R}^k$, then by Prokhorov theorem, $\mathscr{P}(\Xi)$ is metrizable (i.e. by Prokhorov metric) and hence it is a metric space. The latter ensures $\mathscr{P}(\Xi)$ is a Hausdorff space. Subsequently, by [26, Theorem 4.27], $\mathcal{P}(\cdot)$ is upper semicontinuous at point $Y = 0$. ■

Remark 2.1 In the case when $\Psi_i(\xi)$, $i = 1, \cdots, q$, is a scalar function, Assumption 2.1 (b) may be replaced by the following in Lemma 2.1:

- (b_1') $\Psi_i(\cdot)$, $i = 1, \cdots, p$ is continuous and $\Psi_i(\cdot)$, $i = p+1, \cdots, q$ is lower continuous;
- (b_2') there exist an upper semicontinous function $l(\xi)$ and a lower semicontinous function $u(\xi)$ such that

$$l(\xi) \leq \Psi_i(\xi) \leq u(\xi), \ \ \forall \xi \in \Xi, i = 1, \cdots, p,$$
$$l(\xi) \leq \Psi_i(\xi), \ \ \forall \xi \in \Xi, i = p+1, \cdots, q$$

and for any sequence $\{P_N\} \in \hat{\mathcal{P}}$ and any accumulation point P^* of the sequence,

$$\liminf_{N\to\infty} \mathbb{E}_{P_N}[l(\xi)] \geq \mathbb{E}_{P^*}[l(\xi)] > -\infty, \quad \limsup_{N\to\infty} \mathbb{E}_{P_N}[u(\xi)] \leq \mathbb{E}_{P^*}[u(\xi)] < +\infty.$$

To see this, let $\{P_N\} \in \mathcal{P}(Y)$. Under conditions (b_1') and (b_2'), we have by [17, Theorem 4.3] and the remark following [17, Theorem 1.1]

$$\lim_{N\to\infty} \int_{\xi\in\Xi} \Psi_i(\xi) P_N(d\xi) + Y = \int_{\xi\in\Xi} \Psi_i(\xi) P^*(d\xi) + Y$$

for $i = 1, \cdots, p$ and

$$\int_{\xi\in\Xi} \Psi_i(\xi) P^*(d\xi) + Y \leq \liminf_{N\to\infty} \int_{\xi\in\Xi} \Psi_i(\xi) P_N(d\xi) + Y \leq 0,$$

for $i = p+1, \cdots, q$. The inequalities above ensure $P^* \in \mathcal{P}(Y)$ and hence closedness of $\mathcal{P}(Y)$ under topology of weak convergence. Likewise, we can derive (2.19) and hence upper semicontinuity of $\mathcal{P}(Y)$ at 0.

With Lemma 2.1, we are able to address lower semicontinuity of $v(\cdot)$.

Proposition 2.3 (Strong duality of perturbed problem (2.15)) *Let $\mathcal{P}(0) \neq \emptyset$. Under Assumption 2.1, $v(\cdot)$ is lower semicontinuous at 0 and hence there is no dual gap between problems (2.3) and (2.4).*

Proof. The claim is a direct application of [40, Proposition 2.4] to problem (2.15). We give a proof for completeness. Observe first that since $\mathcal{P}(0) \neq \emptyset$ by assumption, $v(0) < +\infty$. If $\mathcal{P}(Y) = \emptyset$ for Y close to 0, then $v(Y) = +\infty$ and hence $v(\cdot)$ is lower semicontinuous at $Y = 0$. In what follows, we consider the case when $\mathcal{P}(Y) \neq \emptyset$.

Let $S(Y)$ denote the set of optimal solutions of problem (2.15). By Lemma 2.1, $\mathcal{P}(Y)$ is weakly compact and hence $v(Y) < +\infty$. Moreover, since the objective function is continuous in P, $S(Y) \neq \emptyset$ and $S(Y)$ is weakly compact. Let

$$\mathcal{P}^*(Y) := \{P \in \mathcal{P}(Y) : \langle -f(x,\xi), P\rangle \leq v(0)\}.$$

Note that if $\mathcal{P}^*(Y) = \emptyset$, then $v(Y) \geq v(0)$. In what follows, we consider the case when $\mathcal{P}^*(Y) \neq \emptyset$. In that case $v(Y) \leq v(0)$ and $S(Y) \subset \mathcal{P}^*(Y)$.

Since $\mathcal{P}(\cdot)$ is upper semicontinuous at point $Y = 0$, it is easy to verify that $\mathcal{P}^*(Y)$ is also upper semicontinuous at point 0 in that f is continuous in ξ. Thus, for any $\epsilon > 0$ there exists a neighborhood $\mathcal{U}_s$ of $\mathcal{P}^*(0)$ such that

$$\langle -f, P\rangle \geq v(0) - \epsilon, \ \forall P \in \mathcal{U}_s. \tag{2.20}$$

By the upper semicontinuity of $\mathcal{P}^*(Y)$, there exists a neighborhood $\mathcal{U}_Y$ of $Y = 0$ such that $\mathcal{P}^*(Y) \subseteq \mathcal{U}_s$. Thus $S(Y) \subset \mathcal{U}_s$ and through (2.20) we have

$$v(Y) \geq v(0) - \epsilon, \forall Y \in \mathcal{U}_Y.$$

Since ϵ is arbitrarily chosen, we conclude that $v(Y)$ is lower semicontinuous at point $Y = 0$. ■

In what follows, we revisit some examples in the preceding subsection with Proposition 2.3. Consider Example 2.1. Assume that there exists $i_0 \in \{p+1, \cdots, q\}$ and a positive number τ such that

$$\|\xi\|^{1+\tau} \leq \psi_{i_0}(\xi), \forall \xi \in \Xi. \tag{2.21}$$

Then Assumption 2.1 (a) is satisfied with $\hat{\mathcal{P}} = \{P \in \mathscr{P}(\Xi) : \mathbb{E}_P[\psi_{i_0}(\xi)] < \infty\}$ because condition (2.21) implies condition (2.17). Moreover, if $\psi_i(\cdot)$, $i = 1, \cdots, q$ is a continuous and bounded function on Ξ, then Assumption 2.1 (b) holds.

Likewise, we can use Proposition 2.3 to certify the absence of a duality gap in Examples 2.3-2.5. Indeed, Assumption 2.1 (a) can be easily verified because there exists a positive constant C such that

$$\mathbb{E}_P[\|\xi\|^2] \leq C, \forall P \in \mathcal{P}.$$

Moreover, if Ξ is bounded, then Assumption 2.1 (b) is fulfilled.

Of course, the boundedness assumption is undesirable in DRO (1.1) and in fact not needed for Slater type condition, we impose the restriction just to illustrate how Proposition 2.3 could be applied in some special circumstances. However, in the application of DRO to optimization problems with chance constraint, it might be a necessity to impose boundedness of Ξ in order for the robust chance constraints to be more applicable. We illustrate this argument through the following example.

Example 2.6 (Infeasibility of robust chance constraint) Consider the following distributionally robust chance constraint

$$\sup_{P\in\mathcal{P}} P(x\xi \le \alpha) \le p^*,$$

where $x \in \mathbb{R}$, $p^* \in (0,1)$, ξ is a random variable with support set $\Xi = \mathbb{R}$,

$$\mathcal{P} = \{P \in \mathscr{P}(\Xi) : \mathbb{E}_P[\xi] = 0, \mathbb{E}_P[\xi^2] = \sigma\}$$

is an ambiguity set defined through true mean value 0 and variance σ. It is easy to show that $\sup_{P\in\mathcal{P}} P(\xi = 0) = 1$. To see this, let P_k be a discrete probability measure with

$$P_k\left(\xi = \sqrt{\frac{\sigma k}{2}}\right) = P_k\left(\xi = -\sqrt{\frac{\sigma k}{2}}\right) = \frac{1}{k}, \text{ and } P_k(\xi = 0) = 1 - \frac{2}{k},$$

where k is a positive number greater than 2. It is easy to verify that $P_k \in \mathcal{P}$ and $\sup_k P_k(\xi = 0) = 1$. Let $H(x) := \{\xi \in \mathbb{R} : x\xi \le \alpha\}$. Then $0 \in H(x)$ for any $x \in \mathbb{R}$ whenever $\alpha \ge 0$. Consequently the robust chance constraint does not have a feasible solution. The key issue here is that the unboundedness of Ξ allows the ambiguity set $\mathcal{P}$ to contain some probability measures which mass their probability near the mean value of ξ.

In a more recent development of distributionally robust optimization (see [51]), ambiguity set $\mathcal{P}$ comprises not only moment conditions but probabilistic constraints. Here we illustrate how Proposition 2.3 may be applied to such a case.

Example 2.7 (STC and the new condition for the moment conditions in [51]) Consider the following ambiguity set

$$\mathcal{P} := \left\{ P \in \mathscr{P}(\Xi) : \begin{array}{ll} \mathbb{E}_P[\psi_i(\xi)] = \mu_i, & \text{for } i = 1, \cdots, p \\ \mathbb{E}_P[\psi_i(\xi)] \le \mu_i, & \text{for } i = p+1, \cdots, q \\ P\{\xi \in \Xi_j\} \le a_j, & \text{for } j = 1, \cdots, k \end{array} \right\},$$

where Ξ_j, $j = 1, \cdots, k$ is subset of Ξ and $0 \le a_j \le 1$. Using the indicator functions, the probabilistic constraints can be rewritten as $\mathbb{E}_P[\mathbb{1}_{\Xi_j}(\xi)] \le a_j$, where

$$\mathbb{1}_{\Xi_j}(\xi) := \begin{cases} 1, & \text{for } \xi \in \Xi_j, \\ 0, & \text{otherwise.} \end{cases}$$

Assumption 2.1 (a) holds if ξ satisfies (2.21). Moreover, Assumption 2.1 (b) holds when $\psi_i(\cdot)$, $i = 1, \cdots, q$ is bounded and continuous, and $\mathbb{1}_{\Xi_j}(\cdot)$, $j = 1, \cdots, k$, is lower semicontinuous on Ξ, see Remark 2.1.

To see how these conditions could be possibly fulfilled, let us consider a more concrete setting with

$$\mathcal{P} := \{P := P_1 \times P_2 \in \mathscr{P}(\Xi) : \mathbb{E}_P[\xi_1] = 0.8, P_1(\xi_1 \in (0.5, 1]) \leq 0.6, P_2(\xi_2 \in [0, 2)) \leq 0.5\}, \quad (2.22)$$

where $\xi = (\xi_1, \xi_2)$ is a random vector with support set $[0, 1] \times [0, 4]$.

Observe first that since Ξ is compact, $\mathscr{P}(\Xi)$ is tight and so is $\mathcal{P}$ as the latter is just a subset of $\mathscr{P}(\Xi)$. Second, for any sequence $\{P^k\} \subset \mathcal{P}$ converging weakly to $\hat{P}$, the lower semicontinuity of the indicator functions $\mathbb{1}_{(0.5,1]}(\cdot)$ and $\mathbb{1}_{[0,2)}(\cdot)$ $[0, 1] \times [0, 4]$ ensures $P_1^k(\xi_1 \in (0.5, 1]) \to \hat{P}_1(\xi_1 \in (0.5, 1]) \leq 0.6$, and $P_2^k(\xi_2 \in [0, 2)) \to \hat{P}_2(\xi_2 \in [0, 2)) \leq 0.5$. On the other hand, the boundedness of ξ_1 over Ξ ensures $\mathbb{E}_{P^k}[\xi_1] \to \mathbb{E}_{\hat{P}}[\xi_1] = 0.8$. This shows $\mathcal{P}$ is closed and hence by the well known Prokhorov theorem, $\mathcal{P}$ is weakly compact. Third, by Remark 2.1, the conclusions of Lemma 2.1 hold with the above stated boundedness and lower semicontinuity, hence by Proposition 2.3, we can assert that the inner maximization problem of (1.1) with ambiguity set (2.22) satisfies the strong Lagrange duality.

On the other hand, the Slater type condition fails because P_1 is a singleton (with $P_1(\xi_1 = 0.5) = 0.4$ and $P_1(\xi_1 = 1) = 0.6$).

Note that it is possible to find an example where Assumption 2.1 fails to hold but the Slater type condition is satisfied.

Example 2.8 Let ξ be a random variable defined on $\mathbb{R}$ with σ-algebra $\mathcal{F}$. Let $\mathscr{P}(\Xi)$ denote the set of all probability measures on $(\mathbb{R}, \mathcal{F})$ and

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

It is shown in [46] that $\mathcal{P}$ is not closed. On the other hand, it is easy to verify that the Slater type condition (2.8) holds. This shows Assumption 2.1 is not necessarily strictly weaker than the Slater type condition and may be used as a condition complementary to STC.

Before concluding this section, we give a simple example where strong duality fails in the absence of STC and lower semicontinuity condition.

Example 2.9 Let ξ be a random variable with support set $\Xi = \{0, 1, 2, \cdots\}$. Let

$$a(\xi^j) = \begin{cases} 0 & \text{for } j = 0, \\ \frac{2^j - 1}{2^{j-1}} & \text{for } j = 1, 2, 3, \cdots \end{cases}$$

and

$$b(\xi^j) = \begin{cases} 0 & \text{for } j = 0, \\ \frac{j+1}{2^{j-1}} - 2 & \text{for } j = 1, 2, 3, \cdots \end{cases}$$

Consider the inner maximization problem

$$\begin{array}{ll} \inf\limits_{p_j \geq 0, j=0,1,2,3,\cdots} & \sum\limits_{j=0}^{\infty} -p_j b(\xi^j)x \\ \text{s.t.} & 2 - \sum\limits_{j=0}^{\infty} p_j a(\xi^j) \leq 0, \\ & \sum\limits_{j=0}^{\infty} p_j = 1, \end{array} \tag{2.23}$$

where $x \in [1,2]$ is fixed. For simplicity of discussion, let $x = 1$. The Lagrange dual of (2.23) is

$$\begin{array}{ll} -\inf\limits_{\lambda_0 \geq 0, \lambda_1 \in \mathbb{R}} & -2\lambda_0 + \lambda_1 \\ \text{s.t.} & a(\xi^j)\lambda_0 + b(\xi^j) - \lambda_1 \leq 0, \text{for } j = 0,1,2,\cdots \end{array} \tag{2.24}$$

It follows from [23, Example 2] that the optimal value of problem (2.24) is 2 whereas the optimal value of problem (2.23) is $+\infty$ because the feasible set of the latter is empty. Let us now consider the perturbation of problem (2.23)

$$\begin{array}{ll} \inf\limits_{p_j \geq 0, j=0,1,2,3,\cdots, \sum_{j=0}^{\infty} p_j = 1} & \sum\limits_{j=0}^{\infty} -p_j b(\xi^j)x \\ \text{s.t.} & 2 - \sum\limits_{j=0}^{\infty} p_j a(\xi^j) + y \leq 0. \end{array} \tag{2.25}$$

The optimal value $v(y)$ of (2.25) is $+\infty$ for $y \geq 0$ because the feasible set is empty in that case. When $y < 0$, we can write down its Lagrange dual

$$\begin{array}{ll} -\inf\limits_{\lambda_0 \geq 0, \lambda_1 \in \mathbb{R}} & -(2+y)\lambda_0 + \lambda_1 - y \\ \text{s.t.} & a(\xi^j)\lambda_0 + b(\xi^j) - \lambda_1 \leq 0, \text{for } j = 0,1,2,\cdots. \end{array} \tag{2.26}$$

Since the inequality constraint of problem (2.25) satisfies the STC, problem (2.26) does not have a duality gap. Analogous to [23, Example 2], we can work out the optimal value of (2.26), which is $v(y) = 2 + y(1 - r_y) - 2^{r_y}$, where r_y is either $\left[\frac{\ln(-y/\ln 2)}{\ln 2}\right]_-$ or $\left[\frac{\ln(-y/\ln 2)}{\ln 2}\right]_+$ depending on which one provides a lower value for $v(y)$. Since $v(y) \to 2$ as $y \to 0_-$, it shows that v is not lower semicontinuous at $y = 0$.

2.3 Boundedness of the Lagrange multipliers

In the last part of this section, we study existence of bounded optimal solutions to the Lagrange dual problem (2.4). This is motivated by necessity of boundedness of the set of feasible solutions to dual problem in order to carry out convergence analysis when a randomization method is applied to the Lagrange dual in Section 3. To ease the notation, we write W for the q-tuple of Lagrange multipliers $(\lambda_0, \Lambda_1, \cdots, \Lambda_q)$ which take values in $\mathbb{R} \times \mathcal{S}^{n_1} \times \cdots \times \mathcal{S}^{n_p} \times \mathcal{K}_{q-p}$. We use $\mathcal{W}(x)$ to denote the set of optimal solutions to Lagrange dual problem (2.4). We investigate conditions under which there is a positive constant η independent of x such that

$$\mathcal{W}(x) \cap \eta\mathcal{B} \neq \emptyset, \forall x \in X, \tag{2.27}$$

where $\mathcal{B}$ denotes the unit ball in the space of $\mathbb{R} \times \mathcal{S}^{n_1} \times \cdots \times \mathcal{S}^{n_p} \times \mathcal{K}_{q-p}$. The boundedness is required for the convergence in Section 3, see Assumption 3.2.

Proposition 2.4 (Existence of bounded Lagrange multipliers) *Assume: (a) the optimal value of problem (2.4) is bounded by a constant (independent of x), (b) $\sup_{x\in X,\xi\in\Xi} |f(x,\xi)| < \infty$, (c) the homogeneous system of inequalities*

$$-\sum_{i=1}^{q} \Lambda_i \circ \Psi_i(\xi) \leq 0, \forall \xi \in \Xi$$

has a unique solution 0. Then there exists a positive constant η such that (2.27) holds.

Proof. For each $x \in X$, let $\lambda_0(x)$ denote the optimal value of problem (2.4). Under condition (a), there exists a constant C such that $\lambda_0(x) \leq C$ for all $x \in X$. In what follows, we show that the components $\Lambda := (\Lambda_1, \cdots, \Lambda_q)$ of the corresponding optimal solution are also bounded. Let $\mathcal{F}(x)$ denote the set of Λ such that $(\lambda_0(x), \Lambda(x))$ is an optimal solution to problem (2.4) for given $x \in X$. It suffices to show that $\mathcal{F}(x)$ is bounded. Assume for the sake of a contradiction that there exists a sequence of $\{x^N\} \subset X$ and $\Lambda^N \in \mathcal{F}(x^N)$ such that $\|\Lambda^N\| \to \infty$ and

$$f(x^N, \xi)/\|\Lambda^N\| - \lambda_0(x^N)/\|\Lambda^N\| - \sum_{i=1}^{q} \Lambda_i^N/\|\Lambda^N\| \circ \Psi_i(\xi) \leq 0, \forall \xi \in \Xi.$$

By driving N to infinity and taking a subsequence if necessary, we may assume without loss of generality that $\Lambda^N/\|\Lambda^N\| \to \hat{\Lambda}$ with $\|\hat{\Lambda}\| = 1$ and consequently deduce

$$-\sum_{i=1}^{q} \hat{\Lambda}_i \circ \Psi_i(\xi) \leq 0, \forall \xi \in \Xi,$$

a contradiction to condition (c). ■

Note that condition (c) is implied by the Slater type condition (2.5), see [55, Remark 2.1 (iii)]. It is unclear whether the condition can be fulfilled under the lower semicontinuity condition.

3 A randomization method and convergence analysis

Having established equivalence between problem (2.4) and its primal (2.3), we are now moving on to discuss numerical methods for solving problem (1.1). For the simplicity of notation, we use Λ to denote $(\Lambda_1, \cdots, \Lambda_q)$. Let us write its dual problem as

$$\begin{array}{ll} \inf\limits_{x,\,\Lambda_1,\cdots,\Lambda_q} & v(x,\Lambda) := \sup_{\xi\in\Xi} \{f(x,\xi) - \sum_{i=1}^{q} \Lambda_i \circ \Psi_i(\xi)\} \\ \text{s.t.} & x \in X, \\ & \Lambda_i \succeq 0, \text{for } i = p+1, \cdots, q. \end{array} \tag{3.28}$$

This is an optimization problem with decision variables x and matrix variables Λ_i, $i = 1, \cdots, q$. In the case when $f(\cdot, \xi)$ is convex for every fixed ξ, the objective function is convex w.r.t. (x, Λ). Our idea here is to apply the well known cutting plane method [24] to solve (3.28). A key step of

the method is to calculate a subgradient of the objective function at each iterate. This requires us to maximize the Lagrange function w.r.t. ξ which could be numerically expensive particularly when it is not concave w.r.t. ξ.

To circumvent the difficulty, we propose a randomization approach which discretizes the space of Ξ through Monte Carlo sampling. Specifically, let $\xi^1, \cdots, \xi^N$ be independent and identically distributed samples of ξ. We consider the following

$$\begin{array}{ll} \inf\limits_{x,\,\Lambda_1,\cdots,\Lambda_q} & v_N(x,\Lambda) := \sup\limits_{j=1,\cdots,N} \left\{ f(x,\xi^j) - \sum\limits_{i=1}^{q} \Lambda_i \circ \Psi_i(\xi^j) \right\} \\ \text{s.t.} & x \in X, \\ & \Lambda_i \succeq 0, \text{for } i = p+1, \cdots, q. \end{array} \tag{3.29}$$

From practical point of view, this kind of approximation scheme is sensible in that it relies only on the samples rather than the range of the support set Ξ. This is a notable departure from the existing numerical approaches for solving distributionally robust optimization where the structure of the support is vital to develop an SDP reformulation. Of course, it might be arguable that samples obtained in practice may be contaminated, we will address this issue in a separate paper as it is not the main focus here. Unless otherwise specified, we assume the samples do not contain noise.

At this point, it might be helpful to remind readers the notation ξ. In formulation (3.28), ξ is a deterministic vector. In the randomization approach, ξ is a random vector whose distribution is unknown but it is possible to obtain its iid samples. This is similar to the "uncertain parameter" in robust convex programs considered by Campi and Calafiore [11]. Note that theoretically speaking, samples generated by any continuous distribution with support set Ξ can be used to construct a random approximation scheme (3.29) although the resulting rate of convergence may be different.

For a fixed sample, we propose to apply the well known cutting plane method for solving problem (3.29). Observe that we can easily compute a subgradient of the objective function of problem (3.29). To see this, let $\mathcal{J}(x,\Lambda)$ denote the index set of $j \in \{1, \cdots, N\}$ such that

$$v^N(x,\Lambda) = f(x,\xi^j) - \sum_{i=1}^{q} \Lambda_i \circ \Psi_i(\xi^j), \text{for } j \in \mathcal{J}(x,\Lambda).$$

By the well known Danskin's theorem,

$$\partial v^N(x,\Lambda) = \text{conv } \left\{ (\nabla_x f(x,\xi^j), \Psi_i(\xi^j)) : j \in \mathcal{J}(x,\Lambda) \right\}.$$

3.1 Optimal value and optimal solution

Before going to the details of the numerical methods for problem (3.29), we derive some convergence results which theoretically justify the proposed approximation scheme. Specifically, we demonstrate convergence of the optimal value and the optimization solutions obtained from solving problem (3.29) to those of problem (3.28) as $N \to \infty$. To this end, let us first consider the following general optimization problem

$$\min_{x \in X} \quad \sup_{\xi \in \Xi} g(x,\xi) \tag{3.30}$$

where X is a compact set in $\mathbb{R}^n$, g is continuous function of (x, ξ), ξ is a parameter which takes values over $\Xi \subset \mathbb{R}^k$. By slightly abusing the notation, let us consider a random variable ξ with support set Ξ. Let $\xi^1, \cdots, \xi^N$ be independent and identically distributed samples of ξ. We consider the following approximation problem:

$$\min_{x \in X} \quad \max_{j=1,\cdots,N} g(x, \xi^j). \tag{3.31}$$

For each realization of the random variables, we solve problem (3.31) and obtain an optimal value and optimal solution. We then ask ourself convergence of these quantities as N increases and investigate conditions under which the optimal value and optimal solution converge to their counterparts of problem (3.30). In what follows, we present a detailed analysis for (3.31).

Assumption 3.1 *Denote by* $M_x(t) := \mathbb{E}\left\{e^{t(g(x,\xi)-\mathbb{E}[g(x,\xi)])}\right\}$ *the moment generating function of the random variable* $g(x, \xi) - \mathbb{E}[g(x, \xi)]$. *The following hold.*

(a) For each $x \in X$, $\sup_{y \in \Xi} g(x, y) < \infty$ *and the moment generating function* $M_x(t)$ *is finite valued for all* t *in a neighborhood of zero.*

(b) There exist a nonnegative measurable function $\kappa : \Xi \to \mathbb{R}_+$ *and constant* $\gamma > 0$ *such that*

$$|g(x', \xi) - g(x'', \xi)| \leq \kappa(\xi)\|x' - x''\|^\gamma, \forall x', x'' \in X$$

for all $\xi \in \Xi$.

(c) The moment generating function $M_\kappa(t)$ *of* $\kappa(\xi)$ *is finite valued for all* t *in a neighborhood of zero.*

Assumption 3.1 (a) means that the probability distributions of the random variables $g(x, \xi)$ and $\kappa(\xi)$ die exponentially fast in the tails. In particular, it holds if this random variables have a bounded support set.

To ease the exposition, let

$$v_N(x) := \max_{j=1,\cdots,N} g(x, \xi^j) \ \text{ and } \ v(x) := \sup_{\xi \in \Xi} g(x, \xi).$$

Let ϑ and ϑ_N denote respectively the optimal values of problem (3.30) and problem (3.31), and X^* and X^N denote the corresponding sets of optimal solutions.

Lemma 3.1 (Convergence of random discretization scheme (3.31)) *Assume: (a)* $g(x, \xi)$ *satisfies Assumption 3.1; (b) the true probability distribution of* ξ *is continuous and there exists positive constants* C_1, ν_1 *(independent of* x*) with*

$$g(x, y_1) - g(x, y_2) \leq C_1\|y_1 - y_2\|^{\nu_1}, \forall y_1, y_2 \in \Xi \tag{3.32}$$

for all $x \in X$; *and (c) there are positive constants* γ_2 *and* δ_0 *with*

$$P(\|\xi - \xi_0\| < \delta) \geq C_2\delta^{\nu_2} \tag{3.33}$$

for any fixed point $\xi_0 \in \Xi$ *and* $\delta \in (0, \delta_0)$. *Then the following assertions hold.*

(i) For any positive number ϵ, there exist positive constants $C(\epsilon)$ and $\beta(\epsilon)$ such that

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

for N sufficiently large.

(ii) Let

$$R(\epsilon) := \min_{x\in X, d(x,X^*)\geq\epsilon} \left\{ \sup_{\xi\in\Xi} g(x,\xi) \right\} - \vartheta.$$

If there exists an $\epsilon_0 > 0$ such that $R(\epsilon) > 0$ for $\epsilon \in (0, \epsilon_0)$ and $R(\cdot)$ is monotonically increasing over the interval, then $R(\epsilon) \to 0$ as $\epsilon \downarrow 0$, and

$$\mathbb{D}(X^N, X^*) \leq R^{-1}\left(3 \sup_{x\in X} |v_N(x) - v(x)|\right).$$

Proof. The thrust of the proof is to use CVaR and its sample average approximation to approximate $\sup_{\xi\in\Xi} g(x,\xi)$ of problem (3.30) which is in line with the convergence analysis carried out in [1]. However, there are a couple of important differences: (a) the convergence here is for the randomization scheme (3.31) rather than the sample average approximation of the CVaR approximation of $\sup_{\xi\in\Xi} g(x,\xi)$, (b) g is not necessarily a convex function.

Part (i). For $\beta \in (0,1)$, let

$$\text{CVaR}_\beta(g(x,\xi)) := \inf_{\eta\in\mathbb{R}} \ \eta + \frac{1}{1-\beta}\int_{\xi\in\Xi} (g(x,\xi) - \eta)_+ \rho(\xi) d\xi \tag{3.34}$$

and

$$v_\beta^N(x) := \inf_{\eta\in\mathbb{R}} \ \eta + \ \frac{1}{(1-\beta)N}\sum_{j=1}^{N} (g(x,\xi^j) - \eta)_+,$$

where $\rho(\cdot)$ denotes the density function of the random variable ξ, $(c)_+ = \max(0,c)$ for $c \in \mathbb{R}$. In the literature, $\text{CVaR}_\beta\left(g(x,\xi)\right)$ is known as conditional value at risk at a specified confidence level β and $v_\beta^N(x)$ is its sample average approximation, see [37, 1]. It is well known that the maximum w.r.t. η in the above formulation is achieved at a finite η. In other words, we may restrict the maximum w.r.t. η to be taken within a closed interval $[-a, a]$ for some sufficiently large positive number a, see [37]. It is easy to verify that

$$v_\beta^N(x) \leq v_N(x) \leq v(x). \tag{3.35}$$

We proceed the rest of the proof for this part in two steps.

Step 1. By the definition of CVaR, for any $\beta \in (0,1)$

$$\text{CVaR}_\beta\left(g(x,\xi)\right) \leq v(x).$$

Moreover, under conditions (b) and (c), it follows by [1, Proposition 1], $g(x,\xi)$ has so-called consistent tail behaviour, that is,

$$1 - G_x(\alpha) \geq K\left(g^*(x) - \alpha\right)^\tau, \text{ for all } \alpha \in (\text{CVaR}_{\beta_0}(g(x,\xi), g^*(x)), \tag{3.36}$$

where $K = \frac{C_2}{C_1}$, $\tau = \frac{\gamma_2}{\gamma_1}$ and $\beta_0 = 1 - \frac{C_2}{C_1}(C_1\delta_0)^{\gamma_2}$. By [1, Theorem 1],

$$|\mathrm{CVaR}_\beta\left(g(x,\xi)\right) - v(x)| \leq \frac{1}{K^{1/\tau}}\frac{\tau}{1+\tau}(1-\beta)^{1/\tau} \tag{3.37}$$

for all $\beta \in (\beta_0, 1)$. Therefore by driving β to 1, we have

$$\sup_{x\in X} |\mathrm{CVaR}_\beta\left(g(x,\xi)\right) - v(x)| \to 0.$$

Step 2. Using the inequalities (3.35), we have

$$\begin{aligned} |v_N(x) - v(x)| &\leq |v_\beta^N(x) - v(x)| \\ &\leq |v_\beta^N(x) - \mathrm{CVaR}_\beta\left(g(x,\xi)\right)| + |\mathrm{CVaR}_\beta\left(g(x,\xi)\right) - v(x)|. \end{aligned}$$

Let ϵ be a small positive number. By (3.37), we may set β sufficiently close to 1 such that

$$\sup_{x\in X} |\mathrm{CVaR}_\beta\left(g(x,\xi)\right) - v(x)| \leq \frac{\epsilon}{2}. \tag{3.38}$$

On the other hand, under Assumption 3.1, it follows by virtue of [44, Theorem 5.1], there exist positive constants $C(\epsilon)$ and $\alpha(\epsilon)$ such that

$$\begin{aligned} &\mathrm{Prob}\Big\{\sup_{x\in X} |v_\beta^N(x) - \mathrm{CVaR}_\beta\left(g(x,\xi)\right)| \geq \epsilon/2\Big\} \\ &\leq \mathrm{Prob}\Big\{\frac{1}{1-\beta}\sup_{x\in X}\sup_{\eta\in[-a,a]}\left|\frac{1}{N}\sum_{j=1}^N (g(x,\xi^j)-\eta)_+ - \mathbb{E}_P[(\eta - g(x,\xi))_+]\right| \geq \epsilon/2\Big\} \\ &\leq C(\epsilon)e^{-\alpha(\epsilon)N} \end{aligned} \tag{3.39}$$

for N sufficiently large. Here in the first inequality, we are using the fact that the maximum w.r.t. η is achieved in $[-a, a]$ for some appropriate positive constant a; see similar discussions in [53]. Note that $|\vartheta_N - \vartheta| \leq \sup_{x\in X} |v_N(x) - v(x)|$. Combining (3.38) and (3.39), we arrive at

$$\begin{aligned} \mathrm{Prob}\Big\{|\vartheta_N - \vartheta| \geq \epsilon\Big\} &\leq \mathrm{Prob}\Big\{\sup |v_N(x) - v(x)| \geq \epsilon\Big\} \\ &\leq \mathrm{Prob}\Big\{\sup |v_\beta^N(x) - \mathrm{CVaR}_\beta\left(g(x,\xi)\right)| \geq \epsilon/2\Big\} \\ &\leq C(\epsilon)e^{-\alpha(\epsilon)N}. \end{aligned}$$

Part (ii). Let $R(\epsilon)$ be defined as in the statement of the lemma. Let ϵ be a fixed small positive number and $\delta := R(\epsilon)/3$. Let N be such that $\sup_{x\in X} |v_N(x) - v(x)| \leq \delta$. Then for any $x \in X$ with $d(x, X^*) \geq \epsilon$, we have

$$v_N(x) - v_N(x^*) \geq v(x) - \vartheta - 2\delta \geq R(\epsilon)/3 > 0$$

which means x cannot be an optimal solution to problem (3.31), in other words, if x^N is an optimal solution to problem (3.31), then $d(x^N, X^*) < \epsilon$ when $\sup_{x\in X} |v_N(x) - v(x)| \leq R(\epsilon)/3$. The conclusion follows if we choose $\epsilon = R^{-1}\left(3\sup_{x\in X} |v_N(x) - v(x)|\right)$. The proof is complete. ■

We make a few comments in sequel about the conditions and conclusion of the lemma as the result might be of broader interest.

First, condition (a) explicitly ensures $g(x, y)$ the essential supremum of $g(x, \xi)$ being bounded for each fixed $x \in X$. Condition (b) is considered by Anderson, Xu and Zhang [1]. Inequality (3.32) is guaranteed when $g(x, \cdot)$ is Hölder continuous over Ξ. Condition (c) is fulfilled when the density function of ξ is bounded away from zero around ξ^0. A combination of (b) and (c) provide a sufficient condition for the so-call consistent tail behaviour for $g(x, \xi)$, see [1] for a detailed discussion.

Second, this kind of convergence analysis is slightly different from standard convergence analysis in stochastic programming in that here we use the largest sampled value of $g(x, \xi)$ rather than its sample average. It should also be distinguished from the convergence analysis by Campi and Calafiore [11] for a similar discretization scheme whose focus is on the feasibility of an optimal solution obtained from solving (3.31) and the number of samples needed to guarantee the feasibility with a specified confidence. Instead it is more closely related to a recent work by Esfahani, Sutter and Lygeros [16] which presents a probabilistic argument for the convergence of the optimal value of (3.31) with a similar discretization scheme. Based on Lemma 3.1, it is possible to estimate the sample size for a specified discrepancy of the optimal values ϵ through large deviation theorem; see [52] and references therein. We leave the details for the interested readers to explore as they are beyond the main focus of this paper.

With Lemma 3.1, we are ready to state convergence of problem (3.29) to problem (3.28) in terms of the optimal value and the optimal solutions. For the simplicity of notation, let

$$h(x, \Lambda, \xi) := f(x, \xi) - \sum_{i=1}^{q} \Psi_i(\xi) \circ \Lambda_i.$$

Let $\mathcal{W}_x$ be defined as in (2.27). We make the following assumption.

Assumption 3.2 *$h(x, \Lambda, \xi)$ satisfies the following conditions.*

(a) For fixed $(x, \Lambda) \in X \times \mathcal{W}_x$,

$$\sup_{\xi \in \Xi} h(x, \Lambda, \xi) < \infty.$$

(b) The true probability distribution of ξ is continuous and there exist positive constants C_1 and ν_1 (independent of x) such that

$$|h(x, \Lambda, \xi') - h(x, \Lambda, \xi'') < C_1 \|\xi' - \xi''\|^{\nu_1}, \forall \xi', \xi'' \in \Xi \tag{3.40}$$

for all $(x, \Lambda) \in X \times \mathcal{W}_x$; and condition (c) of Lemma 3.1 holds.

(c) The moment function of $h(x, \Lambda, \xi)$, denoted by

$$M_{x,\Lambda}(t) := \mathbb{E}\left[e^{t(h(x,\Lambda,\xi) - \mathbb{E}[h(x,\Lambda,\xi)])}\right],$$

is finite valued for all t in a neighborhood of zero.

(d) *Let* $\sigma(\xi) := \kappa(\xi) + \sum_{i=1}^{q} \|\Psi_i(\xi)\|$, *where* $\kappa(\xi)$ *is the Lipschitz modulus of* $f(\cdot, \xi)$. *The moment generating function of* $\sigma(\xi)$ *denoted by* $\mathbb{E}\left[e^{t(\sigma(\xi)-\mathbb{E}[\sigma(\xi)])}\right]$ *is finite valued for all* t *in a neighbourhood of zero.*

Theorem 3.1 (Convergence of random discretization scheme (3.29)) *Let* $\hat{\vartheta}_N$ *and* $\hat{\vartheta}$ *denote the optimal values of problems (3.29) and (3.28) respectively. Under Assumption 3.2, for any positive number* ϵ, *there exist positive constants* $\hat{C}(\epsilon)$ *and* $\hat{\beta}(\epsilon)$ *such that*

$$\text{Prob}(|\hat{\vartheta}_N - \hat{\vartheta}| \geq \epsilon) \leq \hat{C}(\epsilon) e^{-\hat{\beta}(\epsilon) N},$$

when N *is sufficiently large.*

Proof. The conclusion follows directly from Lemma 3.1 in that conditions (a) and (b) of Lemma 3.1 are implied by conditions (c)-(d) and (b) of Assumption 3.2 respectively. We omit the details. ■

3.2 Stationary points

In the case when $f(x, \xi)$ is not convex in x, problem (3.29) is not a convex optimization problem. In such a case, we may not be able to obtain an optimal solution by solving the problem. This motivates us to study convergence of stationary points. Let x^N be just a stationary point of problem (3.29). We look into whether any cluster point of sequence $\{x^N\}$ is a stationary point of (3.28).

To ease the exposition of analysis and maximize the potential application of the convergence results, we consider the general problems (3.30) and (3.31). Throughout this subsection, we assume g is continuously differentiable in x for every ξ. Therefore, both $v(x)$ and $v^N(x)$ are Lipschitz continuous. Let

$$\Xi^N(x) := \arg \max_{j=1,\cdots,N} g^N(x, \xi^j) \quad \text{and} \quad \Xi^*(x) := \arg \max_{\xi \in \Xi} g(x, \xi).$$

Recall that the Clarke subdifferential of a locally Lipschitz continuous function $\phi(x)$ at x, denoted by $\partial\phi(x)$, is defined as follows:

$$\partial\phi(x) := \text{conv} \left\{ \lim_{\substack{x' \in D \\ x' \to x}} \nabla\phi(x') \right\},$$

where D denotes the set of points near x at which ϕ is Fréchet differentiable, $\nabla\phi(x)$ denotes the gradient of ϕ at x. In the case when ϕ is convex, the Clarke subdifferential coincides with the convex subdifferential, see [13] for details.

By [13, Theorem 2.8.2], the Clarke subdifferential of $v(x)$ can be written as

$$\partial v(x) = \{\mathbb{E}_P[\nabla_x g(x, \xi)] : P \in \mathcal{P}[\Xi^*(x)]\},$$

where $\mathcal{P}[S]$ signifies the collection of probability measures supported on S. Likewise, by [13, Proposition 2.3.12],

$$\partial v^N(x) = \text{conv}\{\nabla_x g(x, \xi^j) : \xi^j \in \Xi^N(x)\}. \tag{3.41}$$

Proposition 3.1 (Subdifferential consistency) *Let Ξ be a compact set and $\{x^N\}$ converge to x^*. Suppose that conditions (b) and (c) of Lemma 3.1 holds. Then*

$$\lim_{N\to\infty} \mathbb{D}(\partial v^N(x^N), \partial v(x^*)) = 0.$$

Proof. Let $\eta_N \in \partial v^N(x^N)$ be any element of the subdifferential. By the definition of $\mathbb{D}$, it suffices to show that every accumulation point of sequence $\{\eta_N\}$ lies in $\partial v(x^*)$. By taking a subsequence if necessary, we may assume without loss of generality that $\eta_N \to \eta^*$. Let $|\Xi^N(x^N)|$ denote the cardinality of set $\Xi^N(x^N)$. By relabeling the samples, we may assume

$$\Xi^N(x^N) = \{\xi^1, \cdots, \xi^{|\Xi^N(x^N)|}\}.$$

Using the property of the Clarke subdifferential, we deduce from (3.41) that there exist positive numbers $a_j \in [0,1]$, $j = 1, \cdots, |\Xi^N(x^N)|$ such that $\sum_{j=1}^{|\Xi^N(x^N)|} a_j = 1$ and

$$\eta_N = \sum_{j=1}^{|\Xi^N(x^N)|} a_j \nabla_x g(x^N, \xi^j).$$

Let

$$P_N(\xi) := \begin{cases} a_j, & \text{for } \xi = \xi^j, j = 1, \cdots, |\Xi^N(x^N)|, \\ 0, & \text{otherwise.} \end{cases}$$

Then we may view P_N as a probability distribution of ξ over the support set $\Xi^N(x^N)$ and consequently write η^N as

$$\eta^N = \mathbb{E}_{P_N}[\nabla_x g(x^N, \xi)].$$

Let $\mathscr{P}(\Xi)$ denote the set of all probability measures over Ξ induced by ξ. Since Ξ is a compact set, then $\mathscr{P}(\Xi)$ is weakly compact, which means $\{P_N\}$ must have a weakly convergent subsequence. Assume for simplicity of notation that $P_N \to P^*$ weakly. Then $P^* \in \mathscr{P}(\Xi)$. Since $g(x,\xi)$ is continuous and bounded on $X \times \Xi$, the weak convergence and conditions (b) of Lemma 3.1 ensure

$$\lim_{N\to\infty} v^N(x^N) = \lim_{N\to\infty} \mathbb{E}_{P_N}[g(x^N, \xi)] = \mathbb{E}_{P^*}[g(x^*, \xi)].$$

Moreover, since Ξ is compact, all conditions of Lemma 3.1 are fulfilled. Thus $v_N(x)$ converges to $v(x)$ uniformly over X as $N \to \infty$. Likewise

$$\lim_{N\to\infty} \eta^N = \lim_{N\to\infty} \mathbb{E}_{P_N}[\nabla_x g(x^N, \xi)] = \mathbb{E}_{P^*}[\nabla_x g(x^*, \xi)] = \eta^*.$$

To complete the proof, we need to show that $P^* \in \mathcal{P}[\Xi^*(x^*)]$. But this follows from the definition of $\mathcal{P}[\Xi^*(x^*)]$ in that the uniform convergence of $v_N(x)$ to $v(x)$ ensures $\mathbb{E}_{P^*}[g(x^*, \xi)] = v(x^*)$. ■

With Proposition 3.1, we are ready to study the convergence of stationary points. We call (x, Λ) a stationary point of problem (3.28) if it satisfies

$$0 \in \partial v(x, \Lambda) + \mathcal{N}_X(x) \times \mathcal{N}_{\{0\}\times\mathcal{K}}(\Lambda),$$

where $\{0\} \times \mathcal{K}$ is defined as in (2.5), and $\mathcal{N}_Z(z)$ denotes the Clarke normal cone to Z at z, that is, for $z \in Z$,

$$\mathcal{N}_Z(z) = \left\{\zeta \in \mathcal{V} : \zeta^T t \le 0, \forall t \in \mathcal{T}_Z(z)\right\},$$

$$\mathcal{T}_Z(z) = \liminf_{t\to 0,\ Z\ni z'\to z} \frac{1}{t}(Z - z')$$

and $\mathcal{N}_Z(z) = \emptyset$ when $z \notin Z$. Likewise, we say (x, Λ) is a stationary point of problem (3.29) if it satisfies

$$0 \in \partial v^N(x, \Lambda) + \mathcal{N}_X(x) \times \mathcal{N}_{\{0\}\times\mathcal{K}}(\Lambda).$$

Theorem 3.2 (Convergence of the stationary point of (3.29)) *Let $\{(x^N, \Lambda^N)\}$ be a sequence of stationary points of problem (3.29) and (x^*, Λ^*) be its accumulation point. Under the conditions of Proposition 3.1, (x^*, Λ^*) is a stationary point of problem (3.28).*

Proof. Theorem 3.2 follows from the outer semicontinuity of normal cones $\mathcal{N}_X(\cdot)$ and $\mathcal{N}_{\{0\}\times\mathcal{K}}(\cdot)$ and the consistency of the subdifferential of Proposition 3.1. ■

3.3 Cutting plane method

We now turn to discuss numerical methods for solving problem (3.29) with a fixed sample. This is a deterministic convex program when $f(x, \xi)$ is convex in x for every ξ. We propose to apply the well known cutting plane method for solving the problem.

Algorithm 3.1 (Cutting plane method for problem (3.28)) *Let M be a large positive number. Set $t := 0$, and*

$$\mathbb{F}_0 := X \times [-M, M] \times \mathcal{S}^{n_1} \times \cdots \times \mathcal{S}^{n_p} \times \mathcal{S}_+^{n_{p+1}} \times \cdots \times \mathcal{S}_+^{n_q}.$$

Step 1. Solve the linear semidefinite programming problem:

$$\begin{array}{ll} \inf\limits_{x,\lambda_0,\Lambda_1,\cdots,\Lambda_q} & \lambda_0 \\ \text{s.t.} & (x, \lambda_0, \Lambda_1, \cdots, \Lambda_q) \in \mathbb{F}_t. \end{array} \tag{3.42}$$

Let $(x^t, \lambda_0^t, \Lambda_1^t, \cdots, \Lambda_q^t)$ denote the optimal solution.

Step 2. Find j_t^* such that

$$j_t^* \in \operatorname{argmax}\left\{ f(x^t, \xi^j) - \lambda_0^t - \sum_{i=1}^q \Lambda_i^t \circ \Psi_i(\xi^j) : j = 1, \ldots, N \right\}.$$

Step 3. If $f(x^t, \xi^{j_t^*}) - \lambda_0^t - \sum_{i=1}^q \Lambda_i^t \circ \Psi_i(\xi^{j_t^*}) \le 0$, stop, return $(x^t, \lambda_0^t, \Lambda_1^t, \cdots, \Lambda_q^t)$ as the optimal solution. Otherwise, construct a feasibility cut

$$\Upsilon_t(x, \lambda_0, \Lambda_1, \cdots, \Lambda_q) = \nabla_x f(x^t, \xi^{j_t^*})^T (x - x^t) + f(x^t, \xi^{j_t^*}) - \lambda_0 - \sum_{i=1}^q \Lambda_i \circ \Psi_i(\xi^{j_t^*})$$

and set

$$\mathbb{F}_{t+1} := \mathbb{F}_t \cap \{(x, \lambda_0, \Lambda_1, \cdots, \Lambda_q) : \Upsilon_t(x, \lambda_0, \Lambda_1, \cdots, \Lambda_q) \le 0\}.$$

Go to Step 1 with $t := t + 1$.

The algorithmic procedures follow the classical cutting plane method by Kelley [24]. The only minor difference is that our problem (3.29) involves some matrix variables and problem (3.42) has to be solved by an SDP solver. Convergence of the algorithm can be easily established similar to Kelley [24], we omit the details.

4 Discretization of the ambiguity set

The randomization scheme (3.29) may be investigated from a different perspective. Let $\Xi^N := \{\xi^1, \cdots, \xi^N\}$. If we restrict the ambiguity set $\mathcal{P}$ in (1.2) to the discrete probability measures with support set Ξ^N, then we have

$$\mathcal{P}_N = \left\{ (p_1, \cdots, p_N) : \sum_{j=1}^N p^j \Psi(\xi^j) \preceq 0, \sum_{j=1}^N p_j = 1, p_j \geq 0, j = 1, \cdots, N \right\}.$$

Here instead of writing $\mathcal{P}$, we use $\mathcal{P}_N$ to indicate that the set depends on Ξ^N. Obviously $\mathcal{P}_N \subset \mathcal{P}$ in the sense that for every $(p_1, \cdots, p_N) \in \mathcal{P}_N$, $P_N := \sum_{j=1}^N p_j \delta_{\xi^j}(\xi) \in \mathcal{P}$, where $\delta_{\xi^j}(\xi)$ denotes the Dirac probability measure over Ξ with probability mass at ξ^j. Consequently the distributionally robust optimization problem (1.1) can be written as

$$\begin{array}{lll} \min\limits_{x \in X} \max\limits_{(p_1, \cdots, p_N) \in \mathbb{R}^N_+} & \sum_{j=1}^N p_j f(x, \xi^j) \\ \text{s.t.} & \sum_{j=1}^N p^j \Psi(\xi^j) \preceq 0, \\ & \sum_{j=1}^N p_j = 1. \end{array} \tag{4.43}$$

It is easy to verify that the Lagrange dual of the inner maximization problem can be written as

$$\begin{array}{ll} \inf\limits_{x, \lambda_0, \Lambda_1, \cdots, \Lambda_p} & \lambda_0 \\ \text{s.t.} & x \in X, \lambda_0 \in \mathbb{R}, \\ & \Lambda_i \succeq 0, \text{for } i = 1, \cdots, q, \\ & f(x, \xi^j) - \lambda_0 - \sum_{i=1}^p \Lambda_i \circ \Psi_i(\xi^j) \leq 0, \ j = 1, \cdots, N, \end{array} \tag{4.44}$$

which is equivalent to (3.29). This means the randomization scheme in Section 4 is equivalent to the discretization scheme (4.43). From numerical point of view, the difference between (4.43) and (4.44) lies in the fact that the latter is a single minimization problem whereas the former a finite dimensional min-max optimization problem. When $f(x, \xi)$ is convex in x for every ξ, (4.43) becomes a saddle point problem. In the previous section, we have developed a numerical method for solving (4.44). Here our focus is on a numerical scheme which solves (4.43) directly for fixed Ξ^N.

Our idea is based on the classical cutting plane method to be applied to the convex function $v_N(x) := \sup_{P \in \mathcal{P}_N} \mathbb{E}_P[f(x, \xi)]$ over the compact set X, which can be described as follows: we start by selecting a probability $\mathbf{p}^0 \in \mathcal{P}_N$ and $x^0 \in X$. Let $l_0(x) := \mathbb{E}_{\mathbf{p}^0}[f(x^0, \xi)] + \mathbb{E}_{\mathbf{p}^0}[\nabla_x f(x^0, \xi)]^T (x - x^0)$ and find a minimizer of $l_0(x)$ over X. Note that $l_0(x) \leq v_N(x)$ for all $x \in X$ but it is not necessarily a cutting plane of $v_N(x)$ at x^0 unless $v_N(x^0) = \mathbb{E}_{\mathbf{p}^0}[f(x_0, \xi)]$. Let x^1 denote the optimal solution of $l_0(x)$. Next, evaluate $v_N(x)$ at x^1. We do so by solving the inner maximization problem, that is, maximization of $\mathbb{E}_P[f(x^1, \xi)]$ w.r.t. P over $\mathcal{P}_N$. Let $\mathbf{p}^1$

denote the optimal solution. Then $v_N(x^1)$ is the corresponding optimal value. If $v_N(x^1) \leq \sigma^1$, stop. Otherwise, let $l_1(x) := \mathbb{E}_{\mathbf{p}^1}[f(x^1,\xi)] + \mathbb{E}_{\mathbf{p}^1}[\nabla_x f(x^0,\xi)]^T(x - x^1)$ and find minimizer of $\max(l_0(x), l_1(x))$. In this way, we generate a sequence of cutting planes of $v_N(x)$ and a sequence of approximate optimal solutions $\{x^t\}$.

Algorithm 4.1 (Direct cutting plane method for problem (4.43)) Let $\mathbf{p}^t := (p_1^t, \cdots, p_N^t)$ and $\mathbf{p}^0 \in \mathcal{P}_N$. Let $\mathcal{P}^0 := \{\mathbf{p}^0\}$ and $x^0 \in X$. Set $t := 0$.

Step 1. Solve outer minimization problem

$$\begin{array}{ll} \min\limits_{x,\sigma} & \sigma \\ \text{s.t.} & x \in X, \\ & \sum_{j=1}^N p_j^t[f(x^t,\xi^j) + \nabla_x f(x^t,\xi^j)^T(x - x^t)] \leq \sigma, \text{ for } \mathbf{p}^t \in \mathcal{P}^t. \end{array} \tag{4.45}$$

Let x^t and σ^t denote the optimal solution and optimal value respectively.

Step 2. Solve the inner maximization problem

$$\begin{array}{ll} \max\limits_{(p_1,\cdots,p_N)\in\mathbb{R}_+^N} & \sum_{j=1}^N p_j f(x^t,\xi^j) \\ \text{s.t.} & \sum_{j=1}^N p_j \Psi(\xi^j) \preceq 0, \\ & \sum_{j=1}^N p_j = 1. \end{array} \tag{4.46}$$

Let $\mathbf{p}^t$ and v^t denote the optimal solution and optimal value. If $v^t \leq \sigma^t$, then stop.

Step 3. Let $\mathcal{P}^{t+1} := \mathcal{P}^t \bigcup \{\mathbf{p}^t\}$ and $t := t + 1$, go to Step 1.

Algorithm 4.1 is inspired by a similar algorithm proposed by Pflug and Wozabal [32] for solving a distributionally robust portfolio problem and cutting surface method by Mehrotra and Papp [28] for a general class of moment robust optimization. Compared to the cutting surface method, our algorithm is not particularly aimed at finding a finite number of "points" in Ξ such that the inner maximum w.r.t. P is achieved at these points, i.e., it is ordinary cutting surface method based on the fundamental idea of cutting plane method.

In comparison with Algorithm 3.1, a notable difference is that Algorithm 4.1 builds up cutting planes in the space of decision variables whereas Algorithm 3.1 construct cutting planes in the space of decision variables and Lagrange multipliers. The difference affects applicability of the algorithms in different circumstances. We will come back to this in Section 5 after conducting some comparative numerical tests of the two algorithms.

Following convergence of classical cutting plane method (see [24]), we can assert the convergence of Algorithm 4.1.

Theorem 4.1 (Convergence of Algorithm 4.1) *Let $\{x^t\}$ be a sequence generated by Algorithm 4.1. Then x^t converges to an optimal solution of problem (4.43).*

Note that Algorithm 4.1 is proposed for solving the discretized minimax problem (4.43) for fixed sample size N. It might be interesting to ask whether the optimum obtained from the sampling scheme converges to the optimum of the original DRO (1.1) as N increases. The following theorem addresses this.

Theorem 4.2 (Convergence of discretization scheme (4.43)) *Let x_N be the optimal solution of problem (4.43). Assume: (a) for each $P \in \mathcal{P}$, there exists a sequence $\{P_N\} \subset \mathcal{P}_N$ such that P_N converges to P weakly; (b) Ξ is a compact set. Then w.p.1 an accumulation point of $\{x_N\}$ is an optimal solution of problem (1.1).*

Proof. Since x_N is an optimal solution of problem (4.43), there exists $P_N \in \mathcal{P}_N$ such that (x_N, P_N) is a saddle point of $\min_{x\in X} \max_{P\in\mathcal{P}_N} \langle P, f(x,\xi)\rangle$, i.e.,

$$\max_{P\in\mathcal{P}_N} \langle P, f(x^N,\xi)\rangle = \langle P_N, f(x^N,\xi)\rangle = \min_{x\in X} \langle P_N, f(x,\xi)\rangle. \tag{4.47}$$

On the other hand, since Ξ is a compact set in Euclidean space, by [35, Theorem 1.12] $\mathscr{P}(\Xi)$ is weakly compact under the topology of weak convergence. The latter guarantees every sequence in $\mathcal{P}$ contains a convergent subsequence, see Rachev [35, 36]. By taking a subsequence if necessary, we may assume that $x_N \to x^*$ and $P_N \to P^*$ weakly. By the second equality of (4.47), we obtain $\langle P^*, f(x^*,\xi)\rangle \le \min_{x\in X}\langle P^*, f(x,\xi)\rangle$. In what follows, we show

$$\max_{P\in\mathcal{P}} \langle P, f(x^*,\xi)\rangle \le \langle P^*, f(x^*,\xi)\rangle,$$

which will then enable us to claim

$$\max_{P\in\mathcal{P}} \langle P, f(x^*,\xi)\rangle \le \langle P^*, f(x^*,\xi)\rangle \le \min_{x\in X} \langle P^*, f(x,\xi)\rangle,$$

and hence (x^*, P^*) is a saddle point of $\min_{x\in X} \max_{P\in\mathcal{P}} \langle P, f(x,\xi)\rangle$. Assume for the sake of a contradiction that there exists $\hat{P} \in \mathcal{P}$ such that

$$\langle \hat{P}, f(x^*,\xi)\rangle > \langle P^*, f(x^*,\xi)\rangle. \tag{4.48}$$

Since $f(x,\xi)$ is continuous in (x,ξ), by (4.48), for N sufficiently large $\langle \hat{P}, f(x_N,\xi)\rangle > \langle P_N, f(x_N,\xi)\rangle$. Moreover, under condition (a), there exists a sequence $\{\hat{P}_N\} \subset \mathcal{P}_N$ converging to $\hat{P}$ weakly such that

$$\langle \hat{P}_N, f(x_N,\xi)\rangle > \langle P_N, f(x_N,\xi)\rangle,$$

which contradicts the first equality of (4.47). ■

Corollary 4.1 *Consider problem (1.1). Assume: (a) the moment system in the definition of the ambiguity set $\mathcal{P}$ (see (1.2)) does not have equality constraints, i.e., $p = 0$; (b) there exists probability measure P_0 such that*

$$\langle P_0, \Psi_i(\xi)\rangle \prec 0, \text{for } i = 1, \cdots, q;$$

(c) for any $\epsilon > 0$ and $\xi \in \Xi$, there exists $\xi' \in \Xi^N$ such that $\|\xi - \xi'\| \le \epsilon$ almost surely as N sufficiently large; (d) Ξ is a compact set. Then w.p.1 every accumulation point of $\{x_N\}$ is an optimal solution of problem (1.1).

Proof. Let $\hat{P}$ be defined as in the proof of Theorem 4.2. Let $\lambda \in (0,1)$ be a constant and P_0 be defined as in condition (b), let $P_\lambda := \lambda\hat{P} + (1-\lambda)P_0$. Since $\mathcal{P}$ is a convex set, $P_\lambda \in \mathcal{P}$ and

$$\langle P_\lambda, \Psi(\xi)\rangle = \lambda\langle \hat{P}, \Psi(\xi)\rangle + (1-\lambda)\langle P_0, \Psi(\xi)\rangle \prec 0. \tag{4.49}$$

For fixed λ, there exists $\hat{P}_N^\lambda \in \mathcal{P}_N$ such that $\hat{P}_N^\lambda$ converges weakly to P_λ. To see this, let $\{\Xi_1, \cdots, \Xi_N\}$ be a Voronoi partition, that is, Ξ_i, $i = 1, \cdots, N$ are pairwise disjoint sets with

$$\Xi_i \subseteq \left\{ y : \|y - \xi^i\| = \min_k \|y - \xi^k\| \right\}.$$

Let $\hat{P}_N^\lambda = \sum_{i=1}^N p_i \mathbb{1}_{\xi^i}$, where $p_i = P_\lambda(\Xi_i)$ and $\mathbb{1}_{\xi_i}$ denotes the Dirac probability measure at ξ_i. Under condition (b), the largest diameter of the Voronoi cells goes to zero as N increases. Consequently, we deduce by [30, Lemma 4.9] that $\hat{P}_N^\lambda$ converges to P_λ under the Wasserstein/Kantorovish metric as $N \to \infty$. The latter guarantees weak convergence of $\hat{P}_N^\lambda$ to P_λ because Wasserstein/Kantorovish metric metrizes weak convergence; see [29, Section 2.1]. Let $\lambda \to 1$. The discussion above shows that there exists a sequence $\{\hat{P}_N^\lambda\} \subset \mathcal{P}_N$ converging to $\hat{P}$ weakly as $N \to \infty$. The rest of the proof are similar to that of Theorem 4.2. ■

It might be helpful to make a few comments about the conditions of Theorem 4.2 and Corollary 4.1.

First, from the proof of the corollary, we can see that conditions (b) in the theorem can be replaced by conditions (b) and (c) in the corollary when the moment system in the definition of $\mathcal{P}$ does not involve an equality constraint. It is an open question as to whether this is correct when the moment system involves an equality constraint, we leave this for our future research. We prefer conditions (b) and (c) in the corollary to condition (b) of the theorem in that the former are more verifiable. Moreover, since condition (b) in the corollary is a Slater condition, it ensures strong duality for the inner maximization problem of (1.1) whereas condition (b) of the theorem does not have such a guarantee. Further, under conditions (b) and (c) of the corollary, convergence of the optimal value of problem (4.43) can be drawn directly from Theorem 3.1, and in that case Theorem 4.2 may be understood as complementing Theorem 3.1 by ensuring convergence of the optimal solution. In contrast, under condition (b) of the theorem, it is unclear whether Theorem 3.1 would also give us a guarantee of convergence of the optimal value of (3.29) to that of problem (1.1) without the Slater condition (although we may verify the lower semicontinuity condition derived in section 2). Overall, we conclude that the discretization scheme (4.43) is a bit safer than scheme (3.29) in the absence of strong duality for the inner maximization problem (1.1).

Second, condition (c) of the corollary means that Ξ^N may be iid samples generated by any continuous distribution with support set Ξ or constructed in a *deterministic* manner.

Third, in the absence of strong duality, the optimal value of the discretized minimax optimization problem (4.43) provides a *lower bound* for the optimal value of the original distributionally robust optimization problem (1.1) because the discretized ambiguity set $\mathcal{P}_N$ is smaller than $\mathcal{P}$. In contrast, the optimal value of problem (3.28) may provide an *upper bound* for problem (1.1) as it is formulated through the Lagrange dual of the inner maximization problem. The follow-up discretization scheme (3.29) gives a lower bound for the optimal value of problem (3.28). Overall, in the absence of strong duality, we can conclude via Theorem 3.1 that the optimal value of problem (3.29) provides an upper bound for problem (1.1) when N is sufficiently large.

5 Numerical tests

In this section, we investigate the numerical performance of Algorithms 3.1 and 4.1 by carrying out some comparative analysis. We do so by applying them to a portfolio optimization problem and a multiproduct newsvendor problem. In the implementation of the algorithms, we use the ambiguity set defined as in (2.14) with $\gamma_1 = 0.1$ and $\gamma_2 = 1.1$. The mean and covariance matrix μ_0 and Σ_0 are calculated through samples which are either obtained from historical data (in the first example) or generated by computer (in the second example).

The tests are carried out in MATLAB 8.0 installed on a Thinkpad T430 notebook computer with Windows 7 operating system and Intel Core i5 processor. The SDP subproblems in Algorithms 4.1 and 3.1 are solved by Matlab solver "SDPT3-4.0" [47].

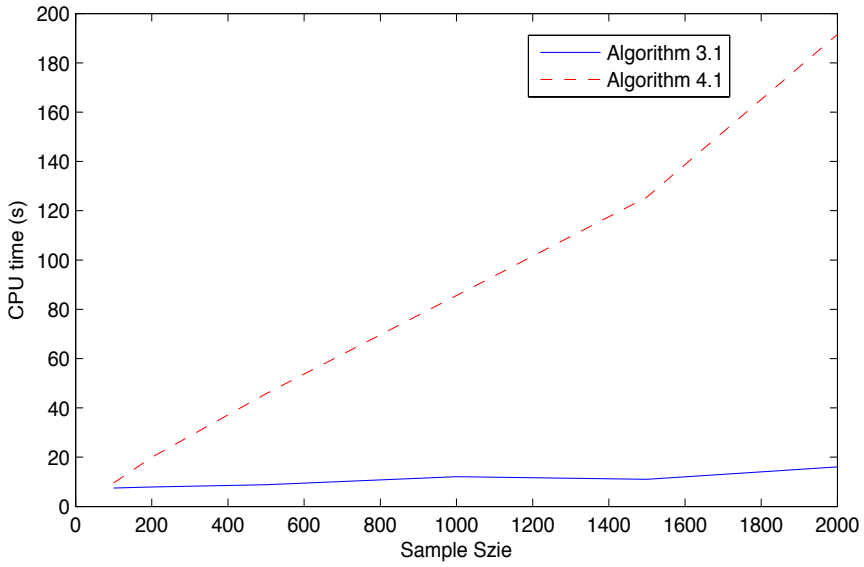

Figure 1: CPU time v.s. sample size, Example 5.1.

Example 5.1 We consider a portfolio optimization problem where the investor makes an optimal decision using historical return rate of 80 stocks between May 2009 and April 2015 from National Association of Securities Deal Automated Quotations (NASDAQ) index. The sample size is 2000. To simplify the discussions, we ignore the transaction fee, therefore the total value of portfolio is

$$f(x, \xi) = \xi_1 x_1 + \xi_2 x_2 + \cdots + \xi_n x_n,$$

where ξ_j denotes the random return rate of asset j.

The investor wants to choose several stocks from NASDAQ index with highest average return rates and make an optimal decision based on them, where the average return rates in the selection rule are calculated by taking average from all historical rates. In order to compare the two algorithms, we have carried out two sets of experiments. One is for the fixed number of portfolios as 5, we examine the performance of the algorithms in terms of CPU time with different sample sizes. This is to investigate sensitivity of the algorithms w.r.t. the change of

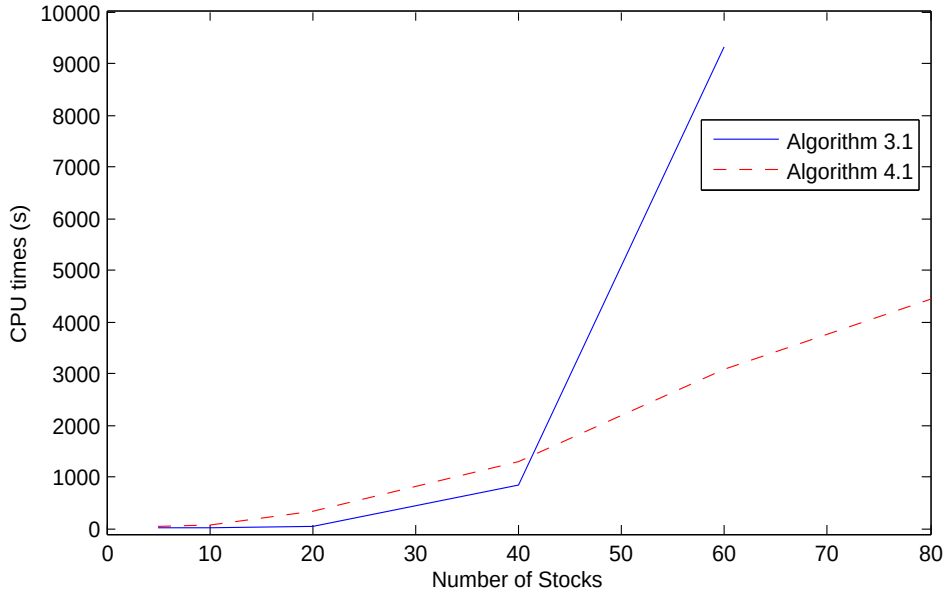

Figure 2: CPU time v.s. the number of portfolios, Example 5.1.

sample size. The other is for fixed sample size 500, we test the performance of the algorithms as problem size increases from 5 to 80.

The results are depicted in Figures 1 and 2 which show the relationships between CPU time and sample size and CPU time and portfolio size. In Figure 1, we can see that the CPU time of Algorithm 4.1 increases rapidly at a linear rate as sample size increases whereas Algorithm 3.1 is not sensitive to the change of sample size. The underlying reason is that increase of sample size does not impact on the problem size of (3.29) but it does affect the size of inner maximization problem of (4.46).

Figure 2 displays an opposite performance of the two algorithms where we fix up the sample size to 500 but increase the portfolio size. The phenomena can be interpreted by the fact that Algorithm 3.1 is sensitive to the increase of portfolio size (number of variables of x) because the cutting planes are constructed in higher dimensional vector and matrix spaces. With the matrix variables in place, any increase of the number of variables of x will significantly affect the overall problem size and hence the effectiveness of the cutting plane method. In contrast, the change of portfolio size does not have any impact on the size of problem (4.46) which is a key step of Algorithm 4.1, and its impact on outer minimization problem (4.45) is limited because the latter is an LP without any matrix variables.

In Example 5.1, the objective function is linear in x, so we don't need linearization at Step 1 of Algorithm 4.1. In what follows, we consider the case when the objective function is nonlinear.

Example 5.2 (Multiproduct newsvendor problem varied from Wiesemann et al. [51]) Assume that a newsvendor trades in $i = 1, \cdots, n$ products. Before observing the uncertain demands ξ_i, the newsvendor orders x_i units of product i at the wholesale price c_i. Once ξ_i is observed, she can sell the quantity $\min(x_i, \xi_i)$ at the retail price v_i. Any unsold stock $(x_i - \xi_i)_+$ is cleared at the salvage price g_i, and any unsatisfied demand $(\xi_i - x_i)_+$ is lost.

We can describe the newsvendor's total loss as a function of the order decision $x := (x_1, \cdots, x_n)^T$:

$$L(x, \xi) = c^T x - v^T \min(x, \xi) - g^T (x - \xi)_+ = (c - v)^T x + (v - g)^T (x - \xi)_+,$$

where the minimum and nonnegativity operator are applied componentwise. We study the risk-averse variant of the multiproduct newsvendor problem:

$$\min_x \sup_{P \in \mathcal{P}} \mathbb{E}_P[U(L(x, \xi))], \tag{5.50}$$

where $U(y) := e^{y/10}$ is an exponential disutility function. In order to compare performance of the two algorithms, we have carried out three sets of experiments. The first one is for the fixed number of products as 7, we examine the performance of the algorithms in terms of CPU times with different sample sizes from 400 to 900, the results are depicted in Figure 3. The second one is for fixed sample size 100, we test the performance of the algorithms as problem size (number of products) increases from 9 to 27, the results are displayed in Figure 4. The third one is for fixed number of products as 2, we investigate the performance of the optimal values from the two algorithms when the sample size increases from 100 to 900. We generate 20 groups of samples for each sample size, calculate the optimal value by the two algorithms for each group and show the convergence in Figures 5 and 6.

The data are generated as follows: for ith product, the wholesale, retail and savage prices are set $c_i = 0.1(5 + i - 1)$, $v_i = 0.15(5 + i - 1)$ and $g_i = 0.05(5 + i - 1)$; the vector of the product demands ξ is characterized by a multivariate log-normal distribution with the mean $\mu = (\mu_1, \cdots, \mu_n)$, $\mu_i = 2, i = 1, \cdots, n$, and covariance $\Sigma = (\sigma_{ij})$, $\sigma_{ii} = 0.35 + 0.01 * (i - 1)$ and $\sigma_{ij} = 0.01$ for $i \neq j$, $i, j = 1, \cdots, n$.

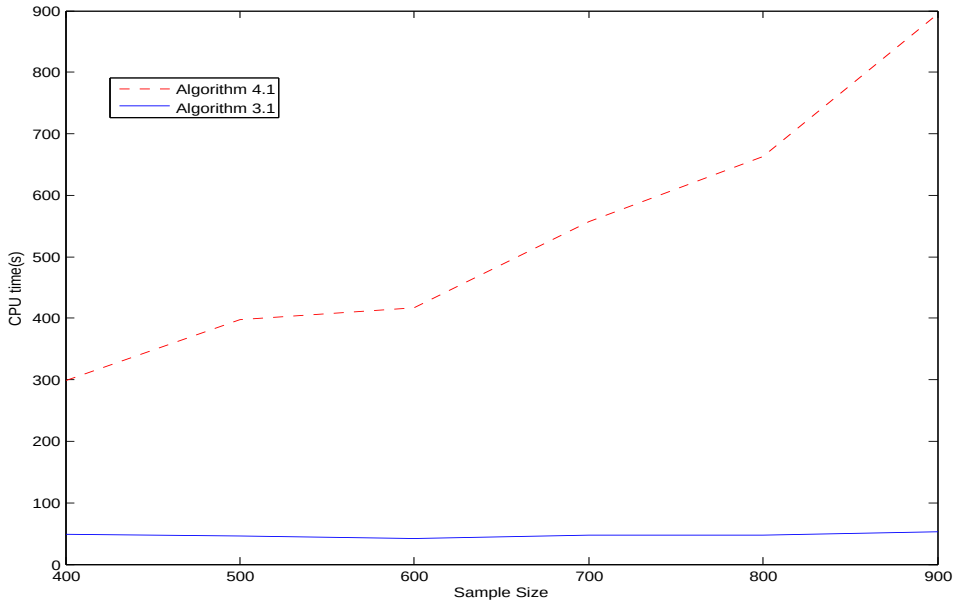

Figure 3: CPU time v.s. sample size, Example 5.2.

Figures 3 and 4 display similar patterns to what we observed in Example 5.1 about change of CPU times against variation of the sample size and the number of products (the problem size). Figures 5 and 6 display the same trend of convergence of the optimal values obtained

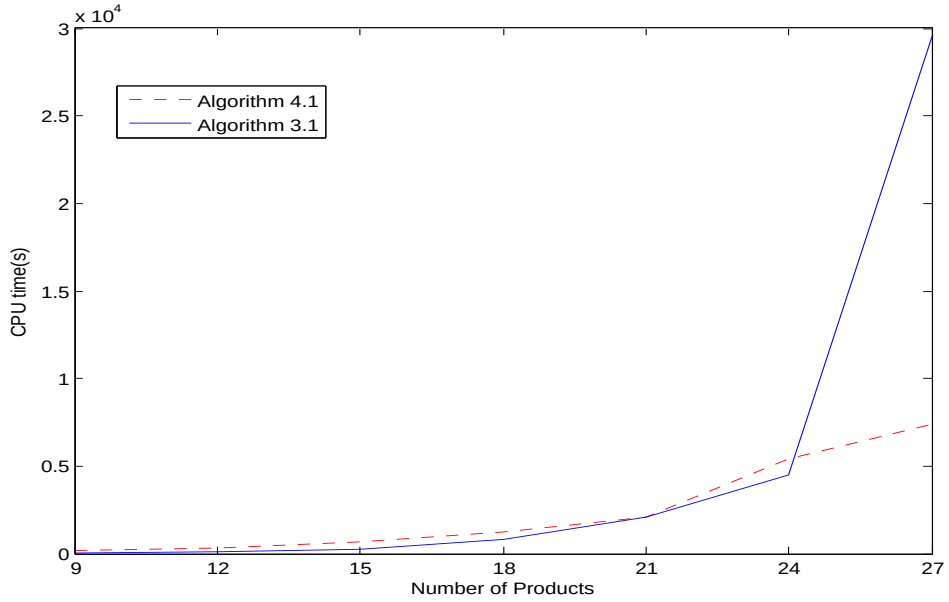

Figure 4: CPU time v.s. the number of products, Example 5.2.

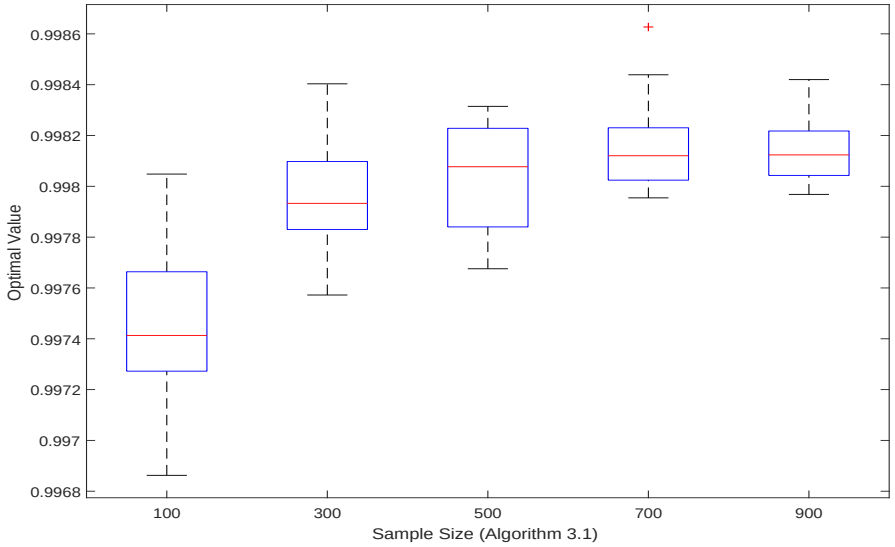

Figure 5: Convergence of optimal values from Algorithm 3.1, Example 5.2.

from the two algorithms as the sample size increases through boxplot. We can see roughly that the optimal values (or the range of the optimal values) converge relatively quickly when the sample size less than 500 and the convergence slows down when the sample size reaches 700. The observation is consistent with our established exponential convergence results. Note that no gap is observed as the strong duality holds in this case.

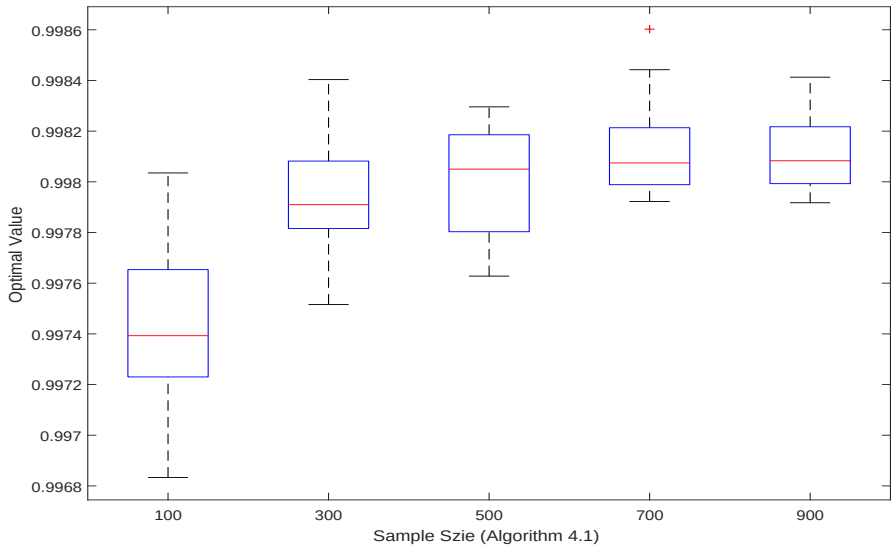

Figure 6: Convergence of optimal values from Algorithm 4.1, Example 5.2.

6 Conclusion

The paper explores conditions for strong duality in distributionally robust optimization with moment constraints and discrete approximation schemes for solving such problems. For the moment problems with only inequality constraints, Slater condition is often satisfied and in this paper we show how it can be verified for some specific moment problems. For the moment problems with only equality and/or inequality constraints, the strong duality often requires the Slater type conditions which are relatively difficult to fulfil and verify. In the absence of the Slater type conditions, it is discovered that a new condition based on lower semicontinuity of the perturbed inner maximization may be used.

We propose two discrete approximation schemes for solving (1.1): one through the well known Lagrange dual formulation and the other through discretization of the ambiguity set which is effectively a kind of direct discretization of the minimax optimization problem. In terms of the optimal value, the dual based discretization scheme tends to give an upper bound whereas the direct discretization gives rise to a lower bound in the absence of strong duality. We then apply the well known cutting plane method to solve the respective discretized problems. In view of numerical efficiency, the preliminary tests show that the dual based discretization scheme is more sensitive to the increase of decision variables whereas the direct discretization scheme is more sensitive to the increase of the sample size. Neither of the schemes requires any specific structure of the underlying functions in the moment problems, in the objective or specific structure of the support set of the random variable, hence they provide an alternative to the mainstream SDP based approaches in the literature.

There is a prospect of applying the discretization schems to distributionally robust optimization problems with objective of minimizing risks. For example, in formulation (1.1), if we replace the expected loss $\mathbb{E}_P[f(x,\xi)]$ with CVaR of $f(x,\xi)$ as defined in (3.34), then the objective

becomes minimization of the worst CVaR. By exchanging the operation of minimization w.r.t. η and maximization w.r.t. probability measure, we end up with the standard formulation (1.1) with an auxiliary "decision variable" η. Similar reformulation can be applied to the case when the objective is a convex composition of $\mathbb{E}_P[f(x,\xi)]$ through Fenchel duality. Thus both the theoretical results in Section 2 and the numerical schemes in Sections 3-4 apply to a large class of distributionally robust optimization problems with moment constraints.

Acknowledgements. We would like to thank Daniel Kuhn, Wolfram Wiesemann and Shaoyan Guo for their valuable comments which helped us significantly strengthen the paper.

References

[1] E. Anderson, H. Xu and D. Zhang, Varying confidence levels for CVaR risk measures and minimax limits manuscript, 2014.

[2] K. B. Athreya and S. N. Lahiri, *Measure theory and probability theory*, Springer texts in statistics, Springer, NewYork, 2006.

[3] A. Ben-Tal and A. Nemirovski, Robust truss topology design via semidefinite programming, *SIAM J. Optim.*, 7: 991-1016, 1997.

[4] A. Ben-Tal, L. El Ghaoui and A. Nemirovski, *Robust optimization*, Princeton University Press, NJ, 2009.

[5] C. Berge, *Espaces topologiques et fonctions multivoques*, Dunod, Paris, 1959.

[6] D. Bertsimas, X. V. Doan, K. Natarajan and C.-P. Teo, Models for minimax stochastic linear optimization problems with risk aversion, *Math. Oper. Res.*, 35: 580-602, 2010.

[7] D. Bertsimas and I. Popescu, Optimal inequalities in probability theory: A convex optimization approach, *SIAM J. Optim.*, 15: 780-804, 2005.

[8] H.-G. Beyer and B. Sendhoff, Robust optimization -a comprehensive survey, *Comp. Appl. Mech. Engin.*, 196: 3190-3218, 2007.

[9] P. Billingsley, *Convergence of probability measures*, Wiley, 1999,

[10] J. F. Bonnans and A. Shapiro, *Perturbation analysis of optimization problems*, Springer, New York, 2000.

[11] G. Calafiore and M. C. Campi, Uncertain convex programs: randomized solutions and confidence levels, *Math. Prog.*, 102: 25-46, 2005.

[12] M. Chen and S. Mehrotra, Epi-convergent scenario generation method for stochastic problems via sparse grid, *Stochastic Programming E-Print*, 2008.

[13] F. H. Clarke, *Optimization and nonsmooth analysis*, Wiley, New York, 1983.

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

[15] J. Dupačová, Uncertanities in minimax stochastic programs, *Optimization*, 60: 1235-1250, 2011.

[16] P. Mohajerin Esfahani, T. Sutter, and J. Lygeros, Performance bounds for the scenario approach and an extension to a class of non-convex programs, *IEEE T Automat. Contr.*, 2015.

[17] E. A. Feinberg, P. O. Kasyanov and N. V. Zadoianchuk, Fatou's Lemma for weakly converging probabilities, *Theory Probab. Appl.*, 58: 683-689, 2014.

[18] J. Goh and M. Sim, Distributionally robust optimization and its tractable approximations, *Oper. Res.*, 58: 902-917, 2010.

[19] D. Goldfarb and G. Iyengar, Robust portfolio selection problems, *Math. Oper. Res.*, 28: 1-38, 2003.

[20] S. Guo, H. Xu and L. Zhang, Stability analysis for mathematical programs with distributionally robust chance constraint, manuscript, 2015.

[21] H. Heitsch and W. Römisch, Scenario reduction algorithms in stochastic programming, *Comput. optim. Appl.*, 24: 187-206, 2003.

[22] Z. Hu and J. Hong, Kullback-Leibler divergence constrained distributionally robust optimization, manuscript, 2012.

[23] D. F. Karney, Duality gaps in semi-infinite liner programming – an approximation problem, *Mathematical Programming*, Vol. 20, pp. 129–143, 1981.

[24] J. E. Kelley, The cutting-plane method for solving convex programs, *SIAM J. Appl. Math.*, 8: 703-712, 1960.

[25] 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.

[26] I. Kupka and V. Toma, Mannuscript of some known results about multifunctions, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, http:// hore.dnom.fmph.uniba.sk/ svana/veb/preklady/TK/ch4.pdf

[27] Y. Liu, R. Meskarian and H. Xu, A Semi-infinite programming approach for distributionally robust reward-risk ratio optimization with matrix moments constraints, manuscript, 2015.

[28] S. Mehrotra and D. Papp, A cutting surface algorithm for semiinfinite convex programming with an application to moment robust optimization, *SIAM J. Optim.*, 24: 1670-1697, 2014.

[29] G. Ch. Pflug, A. Pichler, Approximations for probability distributions and stochastic optimization problems, *Stochastic Optimization Methods in Finance and Energy*, Springer New York, 163:343-387, 2011.

[30] G. Ch. Pflug and A. Pichler, *Multistage Stochastic Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2014.

[31] G. Ch. Pflug, A. Pichler and D. Wozabal, The 1/N investment strategy is optimal under high model ambiguity, *J. Bank. Financ.*, 36: 410-417, 2012.

[32] G. Ch. Pflug and D. Wozabal. Ambiguity in portfolio selection, *Quant. Financ.*, 7: 435-442, 2007.

[33] I. Popescu, Robust mean-covariance solutions for stochastic optimization, *Oper. Res.*, 55: 98-112, 2007.

[34] I. Pólik and T. Terlaky, A Survey of the S-Lemma, *SIAM Rev.*, 49: 371-418, 2007.

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

[36] S. T. Rachev, *Probability metrics and the stability of stochastic models*, John Wiley& Sons Ltd, 1991.

[37] R. T. Rockafellar and S. Uryasev, Optimization of conditional value-at-risk, *J. risk*, 2: 21-42, 2000.

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

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

[40] A. Shapiro, *On duality theory of conic linear problems*, Miguel et al. Eds., SemiInfinite Programming: Recent Advances, 135-165, 2001.

[41] A. Shapiro, *Monte Carlo sampling methods*, A. Rusczyński and A. Shapiro, eds. Stochastic Programming, Handbooks in OR & MS, 10, 2003.

[42] A. Shapiro and S. Ahmed, On a class of minimax stochastic programs, *SIAM J. Optim.*, 14: 1237-1249, 2004.

[43] A. Shapiro, D. Dentcheva and A. Ruszczynski, *Lectures on stochastic programming: modeling and theory*, SIAM, Philadelphia, 2009.

[44] A. Shapiro and H. Xu, Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation, *Optimization*, 57: 395-418, 2008.

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

[46] H. Sun and H. Xu, Convergence analysis for distributional robust optimization and equilibrium problems, *Math. Oper. Res.*, Vol. 41, pp. 377-401, 2016.

[47] K. C. Toh, M. J. Todd and R. H. Tütüncü, SDPT3 -a Matlab software package for semidefinite programming, *Optim. Meth. Soft.*, 11: 545-581, 1999.

[48] W. Yang and H. Xu, Distributionally robust chance constraints for non-Linear uncertainties, *Math. Prog.*, to appear.

[49] W. Wiesemann, D. Kuhn and B. Rustem, Robust resource allocations in temporal networks, *Math. Prog.*, 135: 437-471, 2012.

[50] W. Wiesemann, D. Kuhn and B. Rustem, Robust Markov decision process, *Math. Oper. Res.* 38: 153-183, 2013.

[51] W. Wiesemann, D. Kuhn and M. Sim, Distributionally robust convex optimization, *Oper. Res.*, 62: 1358-1376, 2014.

[52] H. Xu, Uniform exponential convergence of sample average random functions under general sampling with applications in stochastic programming, *J. Math. Anal. Appl.*, 368: 692-710, 2010.

[53] H. Xu and D. Zhang, Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications, *Math. Prog.*, 119: 371-401, 2009.

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

[55] J. Zhang, H. Xu and L.W. Zhang, Quantitative stability analysis for distributionally robust optimization With moment constraints, Optimization-online, September 2015.

[56] S. Zymler, D. Kuhn and B. Rustem, Distributionally robust joint chance constraints with second-order moment information, *Math. Prog.*, 137: 167-198, 2013.