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
School of Mathematics Sciences, University of Southampton, SO17 1BJ, Southampton, UK
(yl2a14@soton.ac.uk)
Department of Mathematics, Dalian Maritime University, Dalian 116026, China

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

May 20, 2015

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 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 introduces other conditions which are based on lower semicontinuity of the optimal value function of the inner maximization problem. Moreover, we apply a well known random discretization scheme to approximate the semiinfinite constraints of the dual problem and demonstrate equivalence of the approach to random discretization of the ambiguity set. 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 has been 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, dual gap, random discretization, cutting plane method

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 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

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

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 [20], Monte carlo sampling method [40] and quadrature rules [13] 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 can 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 and 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 can be traced back to the earlier work by Scarf [37] which was motivated to address incomplete information on the underlying uncertainty in supply chain and inventory control problems. In such problems, historical data may be insufficient to estimate future distribution either because sample size of past demand is too small or because there is a reason to suspect that future demand will come from a different distribution that governing past history. A larger distributional set which contains the true distribution may adequately address the risk from 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 [52], Dupačová [16], Shapiro and Ahmed [41]. 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 [15], Goh and Sim [18], Hu and Hong [21], Goldfarb and Iyengar [19], Mehrotra and Papp[29], Liu and Xu [27], Pflug, Pichler

and Wozabal [30], Popescu [32], Wiesemann, Kuhn and Sim [47, 48], Wozabal [50] to name a few.

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

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

where X is a closed set of $\mathbb{R}^n$, $f : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ is a continuous function, $\xi : \Omega \to \Xi \subset \mathbb{R}^k$ 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

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

Here $\Psi_i : \Xi \to \mathbb{R}^{n_i \times n_i}$, $i = 1, \cdots, p$, is matrix and/or vector 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 matric, its expected value must be negative semidefinite. In the case when $n_i = 1$, $i = 1, \cdots, p$, 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. 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 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 [15] 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 DRO is applied to a class of two stage stochastic programming problems. More recently, Wiesemann, Kuhn and Sim [49] provide a unified framework on DRO problem 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 problem.

In a slightly different direction, DRO approach has been applied to tackle chance constrained stochastic programming problems where there is lack of complete information of the true probability distribution. Zymler et al [53] consider robust chance constrained optimization problems where the ambiguity set is constructed through moment constraints and reformulate the robust constraint as semiinfinite constraints. In the case when the support set of the random variate covers the whole space and the underlying functions in the chance constraint are linear w.r.t. both the decision vector and the random vector, they reformulate the semiinfinite constraints as a semidefinite constraint and demonstrate the resulting semidefinite program (SDP for short) is numerically tractable. In a more recent development, Yang and Xu [46] extend the research to the case where the underlying functions in the chance constraint are nonlinear. A deficiency in

these robust approach is that it may easily cause infeasibility of the robust 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, this is less concerned 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 main stream research in DRO, we concentrate on practical applicability of the 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. A main theoretical question is under what conditions, problem (1.1) and its dual problem are equivalent in the sense that they have the same optimal value. The equivalence is well known provided that either the support set Ξ is a compact set in a finite dimensional space (see [42]) or the system of equalities and/or inequalities satisfy Slater type conditions [39]. 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. Shapiro [39] 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 assesses the issue we have just raised, it is still difficult to verify the condition 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 $\langle\cdot,\psi(\xi)\rangle$ and this makes it difficult when ψ is a vector. It becomes even more challenging when $\mathcal{P}$ is composed of both equality and inequality constraints. This motivates us to develop effective ways for verifying the conditions and look into other conditions complementary in this paper.

Another main issue concerns numerical methods for solving problem (1.1) or its dual counterpart. When the support set Ξ is a finite set, the dual is an ordinary matrix optimization problem, so we may apply the available codes [45] to solve it. It is also possible to solve problem (1.1) directly as a finite dimensional minimax saddle point problem. Indeed Pflug and Wozabal [31] propose an iterative scheme for distributionally robust portfolio optimization problem where the inner maximization problem and outer maximization problem are solved in turn. Mehrotra and Papp [29] extend the approach to a general class of DRO problems and design a process which generates a "cutting surface" of the inner optimal value 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, we may recast the semiinfinite inequality as a semidefinite constraint through the well known S-lemma [33]. We note that this kind of formulation is the most popular approach in the literature of distributionally robust optimization, see for instance [15, 49] 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. As far as we concerned, the main contributions of the paper can be summarized as follows.

- We present a detailed analysis of conditions for the Lagrange duality of the inner maximization problem, namely the Slater type condition and lower semicontinuity condition. The former is based on Shapiro's [39, Proposition 3.4] which has already been widely used in the literature of distributionally robust optimization with moment constraints. In Section 2, we give a scrutiny of the Slater type condition for 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 2.2). While the conditions are restrictive in general, we find that they are satisfied in a number of important cases including the support set Ξ being compact and Ψ_i being bounded, and this may effectively complement the popular Slater type condition in circumstances when the latter is difficult to be verified. Indeed, we can find examples where the lower semicontinuity conditions are satisfied whereas the Slater type condition fails (Example 2.7).

- We propose a discretization scheme based on Monte Carlo sampling for approximating the semiinfinite constraints of the dualized inner maximization problem. The approach is in line with the randomization scheme considered by Campi and Calafiore [12] and Anderson et al [1] for a mathematical program with robust convex constraints. Under some moderate conditions, we demonstrate convergence of the optimal value, optimal solutions and stationary points 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 $\mathcal{P}$, we propose a cutting plane method for solving minimax DRO (1.1) with a randomly discretised support set and show convergence of the approximation scheme in terms of the optimal value and optimization solutions as sample size increases (Theorem 4.2).

- 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 of the two algorithms on a portfolio optimization problem and observe 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 positive semidefinite matrices and cone of negative semidefinite matrices in $\mathbb{R}^{n\times n}$, and $\mathbb{R}^n_+$ denote the cone of vectors with non-negative components in $\mathbb{R}^n$. We write $(\mathcal{Z}, d)$ for 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 $\bar{d}(z, \mathcal{D}) := \inf_{z'\in\mathcal{D}} d(z, z')$ for the distance from a point z to

a set $\mathcal{D}$. For two sets $\mathcal{C}$ and $\mathcal{D}$, $\mathbb{D}(\mathcal{C},\mathcal{D}) := \sup_{z\in\mathcal{C}} \bar{d}(z,\mathcal{D})$ 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 [36] 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 [36, 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)

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

and its Lagrange dual

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

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

As discussed in the introduction, a key step toward 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 [42, page 308]), or the moment problem satisfies Slater type condition (see [39, 49] and references therein). This is underpinned by Shapiro's duality theorem ([39, Proposition 3.4]) for a general class of moment problems.

2.1 Slater type conditions

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

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

where $\Psi(\xi) := (\Psi_1(\xi), \cdots, \Psi_p(\xi))$ and $\mathcal{S}_+ := S_+^{n_1} \times \cdots \times S_+^{n_p}$, see [39, condition (3.12)] for general moments. Here we discuss how this condition may be satisfied and how it could be appropriately verified through some examples.

Example 2.1 Consider the following classical moment problem:

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

where $\psi_i : \Xi \to \mathbb{R}$, $i = 1, \cdots, q$, are continuous functions. Let $\psi_E = (\psi_1, \cdots, \psi_p)$, $\psi_I = (\psi_{p+1}, \cdots, \psi_q)$, $\mu_E = (\mu_1, \cdots, \mu_p)$ and $\mu_I = (\mu_{p+1}, \cdots, \mu_q)$. The Slater type condition in this case can be written as

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

where $\mathcal{K}_1 := \{0\} \times \{0_p\} \times \mathbb{R}_+^{q-p}$ and 0_p denotes the zero vector in $\mathbb{R}^p$. The condition is equivalent to

$$(\mu_E, \mu_I) \in \text{int}\{(\langle P, \psi_E\rangle, \langle P, \psi_I\rangle) + \mathcal{K}_2 : P \in \mathscr{P}\}, \tag{2.7}$$

where $\mathcal{K}_2 := \{0_p\} \times \mathbb{R}_+^{q-p}$. To see the equivalence, let (2.6) hold. Then there exists an open neighborhood of $\mu^* := (1, \mu_E, \mu_I)$, denoted by $\mathcal{U}$, such that $\mathcal{U} \subset \text{int}\{(\langle P, 1\rangle, \langle P, \psi_E\rangle, \langle P, \psi_I\rangle) + \mathcal{K}_1 : P \in \mathscr{M}_+\}$. 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

$$\begin{array}{rcl} (\mu_E, \mu_I) & = & (\langle P_0, \psi_E\rangle, \langle P_0, \psi_I\rangle) \\ & \in & \{(\langle P, \psi_E\rangle, \langle P, \psi_I\rangle) : P \in \mathcal{V} \text{ with } \langle P, 1\rangle = 1\} \\ & \subset & \text{int}\{(\langle P, \psi_E\rangle, \langle P, \psi_I\rangle) + \mathcal{K}_2 : P \in \mathscr{P}\}. \end{array}$$

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

$$\begin{array}{rcl} (1, \mu_E, \mu_I) & \in & \text{int}\{(\langle P, 1\rangle, \langle P, \psi_E\rangle, \langle P, \psi_I\rangle) + \mathcal{K}_1 : P \in \displaystyle\bigcup_{t \in (1-\delta, 1+\delta)} t\mathscr{P}\} \\ & \subset & \text{int}\{(\langle P, 1\rangle, \langle P, \psi_E\rangle, \langle P, \psi_I\rangle) + \mathcal{K}_1 : P \in \mathscr{M}_+\}, \end{array}$$

which yields (2.6).

We now discuss how condition (2.7) may be satisfied. Assume that there is a small open neighborhood of μ_E, denoted by $\mathcal{U}_E$, such that $\mathcal{U}_E \subset \text{int}\,\{\langle P, \psi_E(\xi)\rangle : P \in \mathscr{P}\}$. Let $\mathcal{P}_E := \{P \in \mathscr{P} : \langle P, \psi_E(\xi)\rangle = \mu_E\}$. If there exists $P_E \in \mathcal{P}_E$ such that

$$0_{q-p} \in \text{int}\,\{\langle P_E, \psi_I(\xi)\rangle - \mathbb{R}_-^q - \mu_I\}, \tag{2.8}$$

then condition (2.7) is satisfied. A sufficient condition for existence of $\mathcal{U}_E$ is $\{\psi_E(\xi) : \xi \in \Xi\} = \mathbb{R}^p$. The example shows that condition (2.6) may be verified through (2.7) and further through (2.8) in some special cases.

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

Example 2.2 Let

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

where ξ is a random vector with support set Ξ in $\mathbb{R}^n$, μ is either the true mean value or an estimation, Σ is the true covariance or its approximation. 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 where 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 and assume that $\Sigma_0 \prec \Sigma$. Note that following a similar analysis as in Example 2.1 we can recast condition (2.5) as

$$0 \in \operatorname{int}\{\langle P, \Psi(\xi)\rangle - \mathcal{S}_+^n : P \in \mathscr{P}\}. \tag{2.9}$$

It is easy to observe that condition (2.9) holds in that $\langle P_0, \Psi(\xi)\rangle \prec 0$ under the assumption $\Sigma_0 \prec \Sigma$ and any $n \times n$ positive 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 $0 \in \operatorname{int}\{\langle P, \Psi(\xi)\rangle : P \in \mathscr{P}\}$ because $\mathcal{S}_+ = \{0_{n\times n}\}$. The condition is fulfilled if $\Sigma_0 \in \operatorname{int}\operatorname{conv}\{(\xi - \mu)(\xi - \mu)^T : \xi \in \Xi\}$. The latter is automatically satisfied when $\Xi = \mathbb{R}^n$, see Proposition 2.1 below.

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

Proposition 2.1 *If $\Xi = \mathbb{R}^n$, then*

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

Proof. It suffices to show

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

because the opposite inclusion always holds. Let $M \in \mathcal{S}_+^n$ be any positive semidefinite matrix with eigenvalues 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 $\hat{P} \in \mathscr{P}$ 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 Consider ambiguity set

$$\mathcal{P} := \left\{ P \in \mathscr{P} : \quad \begin{array}{l} \mathbb{E}_P[\xi - \mu_0]^T \Sigma_0^{-1} \mathbb{E}_P[\xi - \mu_0] \preceq \gamma_1 \\ 0 \preceq \mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)^T] \preceq \gamma_2 \Sigma_0 \end{array} \right\}, \tag{2.10}$$

where γ_1 and γ_2 are nonnegative constants. The ambiguity is first considered by Delage and Ye [15]. 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

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

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

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

When $\gamma_i > 0$ for $i = 1, 2$, the moment constraints (2.10) satisfy the Slater type constraint qualification, see [44, Theorem 3]. However, when $\gamma_1 = 0$, the constraint qualification fails. To see this, let us note that matrix $\mathbb{E}_P \left[\begin{pmatrix} -\Sigma_0 & \mu_0 - \xi \\ (\mu_0 - \xi)^T & \gamma_1 \end{pmatrix} \right]$ can never be negative definite in that by Shur 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. Novertheless, if we rewrite the ambiguity set as

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

then the Slater type condition holds, see [44, Theorem 3] for details.

Example 2.4 Consider the following ambiguity set

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

where γ_1 and γ_2 are small positive numbers and $\|\cdot\|_2$ denotes the spectral norm of a matrix. Using the property of the norm, we can reformulate the ambiguity set as

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

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

Example 2.5 Consider the following ambiguity set

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

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 [28]. Let k be the dimension of random vector ξ, $q = \frac{k^2+3k}{2}$, $\psi_I(\xi) = \xi - \bar{\mu}$ 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

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

Analogous to Example 2.4, the Slater type 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 [39, Proposition 2.4]. To this end, we consider the following perturbation of problem (2.3)

$$\begin{array}{ll} \min\limits_{P \in \mathscr{P}} & \mathbb{E}_P[f(x, \xi)] \\ \text{s.t.} & P \in \mathcal{P}(Y), \end{array} \tag{2.11}$$

where $Y = (Y_1, \cdots, Y_p)$ and $Y_i \in \mathcal{S}^{n_i}$, $i = 1, \cdots, p$ 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

$$\mathcal{P}(Y) := \left\{ P \in \mathscr{P} : \ \mathbb{E}_P[\Psi_i(\xi)] + Y_i \preceq 0, \quad \text{for } i = 1, \cdots, p \right\}. \tag{2.12}$$

Let $v(Y)$ denote the optimal value of problem (2.11). Shapiro shows that problem (2.11) has no dual gap if and only if $v(\cdot)$ is lower semicontinuous at point 0, see [39, Proposition 2.3]. A sufficient condition in this proposition 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 $\{P_N\} \subset \mathscr{P}$, P_N is said to converge to $P \in \mathscr{P}$ *weakly* if

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

for each bounded and continuous function $h : \Xi \to \mathbb{R}$. For a set of probability measures $\mathcal{A} \subset \mathscr{P}$, $\mathcal{A}$ is said to be weakly compact 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 weak topology) 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 [2]), a closed set $\mathcal{A}$ (under the weak topology) of probability measures is *weakly compact* if it is tight. In particular, if Ξ is a compact set, then the set of all probability measures on $(\Xi, \mathscr{B})$ is weakly compact [39].

For probability measures $P_1, P_2 \in \mathscr{P}$, the Prokhorov metric [34] 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 of $\mathbb{R}^m$, the convergence of probability measures in the Prokhorov metric is equivalent to weak convergence.

Assumption 2.1 (a) There exists a tight subset of probability measures, denoted by $\hat{\mathcal{P}} \subset \mathscr{P}$, such that $\mathcal{P}(Y) \subset \hat{\mathcal{P}}$ for all Y close to 0; (b) $\Psi_i(\cdot)$, $i = 1, \cdots, p$, 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 $i = 1, \cdots, p; j, t = 1, \ldots, n_i$.

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

$$\sup_{P\in\hat{\mathcal{P}}} \int_\Xi \|\xi\|^{1+\tau} P(d\xi) < C. \tag{2.13}$$

Likewise, a sufficient condition for Assumption 2.1 (b) is that there exists a positive constant τ such that

$$\sup_{P\in\mathcal{P}} \int_\Xi |\psi^i_{jt}(\xi)|^{1+\tau} P(d\xi) < \infty, \tag{2.14}$$

for $i = 1, \cdots, q; j, t = 1, \ldots, p$. Condition (2.14) holds trivially when $\psi^i_{jt}(\xi)$ is bounded.

Lemma 2.1 *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 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 [44, Lemma 1],

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

Therefore

$$\int_{\xi\in\Xi} \Psi_i(\xi) P(d\xi) + Y = \lim_{k\to\infty} \int_{\xi\in\Xi} \Psi_i(\xi) P_k(d\xi) + Y \preceq 0,$$

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

$$\lim_{Y^{k_i}\to 0} \int_{\xi\in\Xi} \Psi_i(\xi) P_{k_i}(d\xi) + Y^{k_i} = \int_{\xi\in\Xi} \Psi_i(\xi) P^*(d\xi) + 0 \preceq 0,$$

which means $P^* \in \mathcal{P}(0)$. This shows the set-valued mapping $\mathcal{P}(Y)$ is outer semicontinuous. Subsequently, by [26, Theorem 4.27], $\mathcal{P}(\cdot)$ is upper semicontinuous at point $Y = 0$. ■

Remark 2.1 In the case when $\Psi_i(\xi)$, $i = 1, \cdots, p$, is a scalar function, Assumption 2.1 (b) is guaranteed by the following: (b_1') $\Psi_i(\cdot)$, $i = 1, \cdots, p$ is lower semicontinuous; (b_2') for any sequence $\{P_N\} \in \hat{\mathcal{P}}$ and any accumulation point P^* of the sequence, there exists an integrable function $l(\xi)$ such that $\Psi_i(\xi) \geq l(\xi)$ for all $\xi \in \Xi$, $i = 1, \cdots, p$ and

$$\liminf_{N\to\infty} \mathbb{E}_{P_N}[l(\xi)] \geq \mathbb{E}_{P^*}[l(\xi)],$$

see [17, Theorem 4.3] for details. Condition (b_2') means that function $\Psi_i(\cdot)$ is lower bounded by a uniformly integrable function. In particular, if each component of $\Psi_i(\cdot)$ is lower bounded by a constant K, then we may choose $l(\xi) \equiv K$.

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

Proposition 2.2 *Assume the conditions of Lemma 2.1. Then $v(\cdot)$ 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 [39, Proposotion 2.4] to problem (2.11). We give a proof for completeness. By Lemma 2.1, $\mathcal{P}(0) \neq \emptyset$ and it is weakly compact. Thus $v(0)$ is finite and the corresponding optimal solution set of (2.11), denoted by $S(0)$, is nonempty and weakly compact. Let

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

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 $\langle f, P\rangle \geq v(0) - \epsilon$ for any $P \in \mathcal{U}_s$. By the upper semicontinuity of $\mathcal{P}^*(Y)$, there exists a neighborhood $\mathcal{U}_Y$ of $Y = 0$ such that $\mathcal{P}^*(Y) \subseteq \mathcal{U}_s$. Subsequently, $v(Y) \geq v(0) - \epsilon$ for any $Y \in \mathcal{U}_Y$. 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 2.2. Consider Example 2.1. Assume that there exists $i_0 \in \{p+1, \cdots, q\}$ and a positive number τ such that

$$\|\xi\|^{1+\tau} \leq \psi_{i_0}(\xi), \forall \xi \in \Xi. \tag{2.15}$$

Then Assumption 2.1 (a) is satisfied with $\hat{\mathcal{P}} = \{P \in \mathscr{P} : \mathbb{E}_P[\psi_{i_0}(\xi)] < \infty\}$ because condition (2.15) implies condition (2.13). Moreover, if $\psi_i(\cdot)$, $i = 1, \cdots, q$ is continuous and bounded function on Ξ, then Assumption 2.1 (b) holds.

Likewise, we can use Proposition 2.2 to explain no dual 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 2.2 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 Consider the following distributionally robust chance constraint

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

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

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

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 $H(x) := \{\xi \in \mathbb{R} : x\xi \leq \alpha\}$. Then $0 \in H(x)$ for any $x \in \mathbb{R}$ whenever $\alpha \geq 0$. 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 [49]), ambiguity set $\mathcal{P}$ comprises not only moment conditions but probabilistic constraints. Here we illustrate how Proposition 2.2 may be applied to such a case.

Example 2.7 Consider the following ambiguity set

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

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

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

Assumption 2.1 (a) holds if the moment constraints satisfy (2.15). Assumption 2.1 (b) holds provided that $\psi_i(\cdot)$, $i = 1, \cdots, q$ 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 satisfied, let us consider

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

where $\xi = (\xi_1, \xi_2)$ is a random vector with support set $[0, 1] \times [0, 4]$. Note that $\mathscr{P}$ is a compact set as Ξ is compact. Moreover, $\psi(\xi) := \xi_1$ is bounded and continuous. Further, $\mathbb{1}_{(0.5,1]}(\cdot)$ and $\mathbb{1}_{[0,2)}(\cdot)$ are lower semicontinuous on $[0, 1] \times [0, 4]$. 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 $P_1(\xi_1 = 1) = 0.6$).

2.3 Boundedness of Lagrange multipliers

In the last part of this section, we study boundedness of the multipliers of 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. We need the following conditions.

Assumption 2.2 *Let $f(x, \xi)$ be defined as in DRO (1.1) and $\hat{\mathcal{P}}$ be a set of probability measures with $\mathcal{P} \subset \hat{\mathcal{P}}$.*

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

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

(c) X is a compact set.

Assumption 2.2 (a) and (b) ensure that $\mathbb{E}_P[f(x, \xi)]$ is well defined for every $x \in X$ and $P \in \mathcal{P}$. To ease the notation, we use $W \in \mathbb{R} \times \mathcal{S}_+^{n_1} \times \cdots \times \mathcal{S}_+^{n_p}$ to denote $(\lambda_0, \Lambda_1, \cdots, \Lambda_p)$. Let $\mathcal{W}_x$ denote the set of optimal solutions to Lagrange dual problem (2.4). For each fixed $x \in X$, there exists a positive constant η_x such that

$$\mathcal{W}_x \cap \eta_x \mathcal{B} \neq \emptyset, \tag{2.16}$$

where $\mathcal{B}$ denotes the unit ball in the space of $\mathbb{R} \times \mathcal{S}_+^{n_1} \times \cdots \times \mathcal{S}_+^{n_p}$, see [44, Remark 3] for details.

It might be interesting to investigate conditions under which the set of Lagrange multipliers is bounded uniformly w.r.t. x. The following theorem addresses this.

Theorem 2.1 *Let $\mathcal{F}(x)$ denote the set of feasible solutions to problem (2.4). Assume: (a)*

$$\sup_{x \in X, \xi \in \Xi} |f(x, \xi)| < \infty,$$

(b) the homogeneous system of equations

$$\lambda_0 - \sum_{i=1}^{p} \Lambda_i \circ \Psi_i(\xi) = 0, \xi \in \Xi$$

has a unique solution 0. Then there exists a positive constant C such that

$$\sup_{x \in X} \|\mathcal{F}(x)\| \leq C. \tag{2.17}$$

Proof. Observe that the feasible set is defined by a semiinfinite system of inequalities in the space of $\mathbb{R} \times \mathcal{S}_+^{n_1} \times \cdots \times \mathcal{S}_+^{n_p}$. Condition (b) ensures the null space of the linear system reduces to $\{0\}$ and it does so uniformly w.r.t. x (because the coefficients of the homogeneous system are independent of x). This shows $\mathcal{F}(x)$ is bounded for every $x \in X$.

To show (2.17), it suffices to prove that $\mathcal{F}(\cdot)$ as a set-valued mapping is outer semicontinuous (upper semicontinuous in the sense of Berge) because in that case the union of $\mathcal{F}(x)$ over the compact set X is bounded. It is easy to verify that $\mathcal{F}(x)$ is closed and hence the boundedness implies compactness for each $x \in X$. Moreover, condition (a) guarantees that $\mathcal{F}(\cdot)$ is closed at any point of X, namely, for any sequence $\{x_k\} \subset X$ with $x_k \to x$ and $\zeta_k \in \mathcal{F}(x_k)$ with $\zeta_k \to \eta$, it is an easy exercise to show under condition (a) and continuity of $f(\cdot, \xi)$, that $\zeta \in \mathcal{F}(x)$. The outer semicontinuity is evident. ■

3 A randomization method and convergence analysis

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

$$\begin{array}{ll} \inf\limits_{x, \Lambda_1, \cdots, \Lambda_p} & v(x, \Lambda) := \sup_{\xi \in \Xi} \{f(x, \xi) - \sum_{i=1}^{p} \Lambda_i \circ \Psi_i(\xi)\} \\ \text{s.t.} & x \in X, \\ & \Lambda_i \succeq 0, \text{for } i = 1, \cdots, p. \end{array} \tag{3.18}$$

This is an optimization problem with decision variables x and matrix variables Λ_i, $i = 1, \cdots, p$. In the case when $f(\cdot, \xi)$ is convex for every fixed ξ, the objective function is convex w.r.t. (x, Λ) and piecewise linear in Λ. Our idea here is to apply well known cutting plane method to solve (3.18). 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 sample average approximation scheme for problem (3.18)

$$\begin{array}{ll} \inf\limits_{x,\,\Lambda_1,\cdots,\Lambda_p} & v_N(x,\Lambda) := \sup\limits_{k=1,\cdots,N} \left\{ f(x,\xi^k) - \sum\limits_{i=1}^{p} \Lambda_i \circ \Psi_i(\xi^k) \right\} \\ \text{s.t.} & x \in X, \\ & \Lambda_i \succeq 0, \text{for } i = 1, \cdots, p. \end{array} \tag{3.19}$$

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 support set Ξ. This is a notable departure from existing numerical approaches for 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 sample does not contain noise.

For a fixed sample, we propose to apply the well known cutting plane method for solving problem (3.19). Observe that we can easily compute a subgradient of the objective function of problem (3.19). To see this, let $\mathcal{K}(x,\Lambda)$ denote the index set of $i \in \{1, \cdots, N\}$ such that

$$v(x,\Lambda) = f(x,\xi^k) - \sum_{i=1}^{p} \Lambda_i \circ \Psi_i(\xi^k), \text{for } k \in \mathcal{K}(x,\Lambda).$$

By the well known Danskin's theorem,

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

3.1 Optimal value and optimal solution

Before going to the details of numerical methods for problem (3.19), we give a theoretical justification of the proposed approximation scheme. This amounts to demonstrate convergence of the optimal value and the optimization solutions obtained from solving problem (3.19) to those of problem (3.18) as $N \to \infty$. To this end, let us first consider the following general optimization problem

$$\begin{array}{ll} \min\limits_{x} & \sup\limits_{\xi \in \Xi} g(x,\xi) \\ \text{s.t.} & x \in X, \end{array} \tag{3.20}$$

where X is a compact set in $\mathbb{R}^n$, g is continuous function, ξ is a parameter which takes values over a compact set Ξ. Here we also take ξ as a random variable with certain distributions, such as uniformly distributed on Ξ.

Let $\xi^1, \cdots, \xi^N$ be independent and identically distributed random variables of ξ. We consider

the following approximation problem:

$$\begin{array}{ll} \min\limits_{x} & \max\limits_{k=1,\cdots,N} g(x,\xi^k) \\ \text{s.t.} & x \in X. \end{array} \tag{3.21}$$

For each realization of the random variables, we solve problem (3.21) 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 (3.20). This kind of convergence analysis is slightly different from standard convergence analysis of in stochastic programming in that here we use the largest sampled value for $f(x,\xi)$ rather than its sample average. Here we present a detailed analysis for (3.21). Let ϑ and ϑ_N denote respectively the optimal values of problem (3.20) and problem (3.21). Let X^* and X^N denote the corresponding set of optimal solutions.

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 every $x \in X$ the moment generating function $M_x(t)$ is finite valued for all t in a neighborhood of zero.

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

$$|g(x',\xi) - g(x,\xi)| \le \kappa(\xi)\|x' - x\|^\gamma$$

for all $\xi \in \Xi$ and all $x', x \in X$.

(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 probability distribution of the random variable $g(x,\xi)$ (random variable $\kappa(\xi)$) dies exponentially fast in the tails. In particular, it holds if this random variable has a distribution supported on a bounded subset of $\mathbb{R}$. For simplicity, we denote in what follows:

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

Lemma 3.1 *Assume: (a) $g(x,\xi)$ satisfies Assumption 3.1; (b) ξ is continuous random variable with identical distribution to ξ^i, $i = 1,\cdots,N$ and there exist positive constants K and τ independent of x such that for each $x \in X$, there is $\alpha_0(x) < v(x) < \infty$ with*

$$1 - G_x(\alpha) \ge K\left(g^*(x) - \alpha\right)^\tau, \text{ for all } \alpha \in (\alpha_0(x), g^*(x)), \tag{3.22}$$

where G_x denotes the cumulative distribution function of $g(x,\xi)$. 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

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

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

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

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 (3.20) which is in line with the convergence analysis carried out in [1]. However, there are a few important differences: (a) the convergence here is for the randomization scheme (3.21) rather than the sample average approximation of the CVaR approximation of $\sup_{\xi\in\Xi} g(x,\xi)$, (b) g is not supposed to be convex function.

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

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

and

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

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 [35, 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 [35]. It is easy to verify that

$$v_\beta^N(x) \leq v_N(x) \leq v(x). \tag{3.23}$$

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 condition (b), it follows by [1, Theorem 2.1] that

$$|\text{CVaR}_\beta\left(g(x,\xi)\right) - v(x)| \leq \frac{1}{K^{1/\tau}}\frac{\tau}{1+\tau}(1-\beta)^{1/\tau}. \tag{3.24}$$

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 (3.23), 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 (3.24), we may set β sufficiently close to 1 such that

$$\sup_{x\in X} |\mathrm{CVaR}_\beta\left(g(x,\xi)\right) - v(x)| \leq \frac{\epsilon}{2}. \tag{3.25}$$

On the other hand, under Assumption 3.1, it follows by virtue of [43, Theorem 5.1], there exist positive constants $C(\epsilon)$ and $\alpha(\epsilon)$ such that

$$\begin{aligned} &\mathrm{Prob}\Big\{\sup_{x\in X} |v_\beta^N(x) - \mathrm{CVaR}_\beta\left(g(x,\xi)\right)| \geq \epsilon/2\Big\} \\ &\leq \mathrm{Prob}\Big\{\frac{1}{1-\beta}\sup_{x\in X}\sup_{\eta\in[-a,a]}\left|\frac{1}{N}\sum_{j=1}^N (g(x,\xi^j)-\eta)_+ - \mathbb{E}_P[(\eta - g(x,\xi))_+]\right| \geq \epsilon/2\Big\} \\ &\leq C(\epsilon)e^{-\alpha(\epsilon)N} \end{aligned} \tag{3.26}$$

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 [51]. Note that $|\vartheta_N - \vartheta| \leq \sup_{x\in X} |v_N(x) - v(x)|$. Combining (3.25) and (3.26), 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

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

which means x cannot be an optimal solution to problem (3.21), in other words, if x^N is an optimal solution to problem (3.21), 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. ■

Condition (3.22) requires the cumulative distribution function to approach 1 faster than some power of the distance to $f^*(x)$. This is a natural constraint and will be available whenever the corresponding density function is bounded as $\alpha \to f^*(x)$, (even less of a restriction than part(ii) requires). We can think of τ as related to the way in which the density function of

$f(x,\xi)$ approaches zero when α approaches its limit. If the density function approaches zero like an n'th power then we can set $\tau = n+1$. Cases where the density function is bounded away from zero in this region (for example, when $f(x,\xi)$ follows a uniform distribution) correspond to $\tau = 1$, see details in [1].

With Lemma 3.1, we are ready to state convergence of problem (3.19) to problem (3.18) in terms of the optimal value and the optimal solutions. For the simplicity of notation, let

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

Let $\mathcal{W}_x$ be defined as in (2.16). 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) There exist positive constants K and τ independent of (x,Γ) such that for each $(x,\Gamma) \in X \times W_x$, there is $\alpha_0(x) < h^*(x,\Lambda) := \max_{\xi\in\Xi} h(x,\Lambda,\xi) < \infty$ with

$$1 - H_{x,\Lambda}(\alpha) \geq K\left(h^*(x,\Lambda) - \alpha\right)^{\tau}, \text{ for all } \alpha \in (\alpha_0(x), h^*(x,\Lambda)), \tag{3.27}$$

where $H_{x,\Lambda}$ denotes the cumulative distribution function of $h(x,\Lambda,\xi)$.

(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

$$q(\xi) := \kappa(\xi) + \sum_{i=1}^{p} \|\Psi_i(\xi)\|,$$

where $\kappa(\xi)$ is the Lipschitz modulus of $f(x,\xi)$ as defined in Assumption 2.2. The moment generating function of $q(\xi)$ denoted by $\mathbb{E}\left[e^{t(q(\xi) - \mathbb{E}[q(\xi)])}\right]$ is finite valued for all t in a neighborhood of zero.

Theorem 3.1 *Let $\hat{\vartheta}_N$ and $\hat{\vartheta}$ denote the optimal values of problems (3.19) and (3.18) respectively, and Assumption 3.2 hold. Then for any positive number ϵ, there exist positive constants $\hat{C}(\epsilon)$ and $\hat{\beta}(\epsilon)$ such that*

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

for N sufficiently large.

Proof. We use Lemma 3.1 to prove the theorem. It suffices therefore to verify the conditions of the lemma. The conditions (a) and (b) of Lemma 3.1 are implied by conditions (c)-(d) and (b) of Assumption 3.2 respectively.

3.2 Stationary points

In the case when $f(x,\xi)$ is not convex in x, problem (3.19) 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. We look into whether its accumulation point of $\{x^N\}$ is a stationary point.

To ease the exposition of analysis and maximize the potential application of the convergence results, we consider the general problems (3.20) and (3.21). 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

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

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

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

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 where ϕ is convex, the Clarke subdifferential coincides with the convex subdifferential, see [14] for details.

By [14, 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 [14, Proposition 2.3.12],

$$\partial v^N(x) = \operatorname{conv}\{\nabla_x g(x,\xi^i) : \xi^i \in \Xi^N(x)\}. \tag{3.28}$$

Proposition 3.1 (Subdifferential consistency) *Let Ξ be a compact set and x^N converge to x^*. Suppose that the condition (b) 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 (3.28) that there exist positive numbers $a_i \in [0,1]$, $i = 1, \cdots, |\Xi^N(x^N)|$ such that $\sum_{i=1}^{|\Xi^N(x^N)|} a_i = 1$ and

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

Let

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

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

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

Let $\mathscr{P}$ denote the set of all probability measures over Ξ induced by ξ. Since Ξ is a compact set, then $\mathscr{P}$ 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 $P^* \in \mathscr{P}$. Since $g(x, \xi)$ is continuous and bounded on $X \times \Xi$, the weak convergence implies

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

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 ensure that $P^* \in \mathcal{P}[\Xi^*(x^*)]$. But this follows from the definition of $\mathcal{P}[\Xi^*(x^*)]$ as $\mathbb{E}_{P^*}[g(x^*, \xi)] = v(x^*)$. ■

With Proposition 3.1, we are ready to study the convergence of stationary points. We said that (x, Λ) is the stationary point of problem (3.18) if it satisfies

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

where $\mathcal{S}_+$ 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 (3.19) if it satisfies

$$0 \in \partial h^N(x, \Lambda) + \mathcal{N}_X(x) \times \mathcal{N}_{\mathcal{S}_+}(\Lambda),$$

where

$$h^N(x, \Lambda) := \frac{1}{N}\sum_{k=1}^{N}(f(x, \xi^k) - \sum_{i=1}^{p} \Psi(\xi^k) \circ \Lambda_i).$$

Theorem 3.2 *Let $\{(x^N, \Lambda^N)\}$ be a sequence of stationary points of problem (3.19) and (x^*, Λ^*) be its accumulation point. Under the conditions of Proposition 3.1, (x^*, Λ^*) is the stationary point of problem (3.18).*

Proof. Theorem 3.2 follows from the outer semicontinuity of normal cones $\mathcal{N}_X(\cdot)$ and $\mathcal{N}_{\mathcal{S}_+}(\cdot)$ 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 (3.19) 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 to solve it.

Algorithm 3.1 (Cutting plane method for problem (3.18)) *Let M be a large positive number. Set $t := 0$, $\mathbb{S}_0 := X \times [-M, M] \times \mathcal{S}_+^{n_1} \times \cdots \times \mathcal{S}_+^{n_p}$.*

Step 1. Solve the linear programming problem:

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

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

Step 2. Find j_t^* such that

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

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

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

and set

$$\mathbb{S}_{t+1} := \mathbb{S}_t \cap \left\{ (x,\lambda_0,\Lambda_1,\cdots,\Lambda_p) \in X \times \mathbb{R} \times \mathcal{S}_+^{n_1} \times \cdots \times \mathcal{S}_+^{n_p} : \Upsilon_t(x,\lambda_0,\Lambda_1,\cdots,\Lambda_p) \le 0 \right\}.$$

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

The algorithmic procedures follow the classical cutting plane method by Kelley [23]. The only minor difference is that our problem (3.19) involves some matrix variable and the resulting linear programming problem at each iteration have to be solved by an SDP solver. Convergence of the algorithm can be easily established similar to Kelley [23], we omit the details.

4 Discretization of the ambiguity set

The randomization scheme (3.19) 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

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

Here instead of writing $\mathcal{P}$, we use $\mathcal{P}_N$ to indicate that the set depends on Ξ^N. Consequently the distributionally robust optimization problem (1.1) can be written as

$$\begin{array}{lll} \min\limits_{x} & \max\limits_{(p_1,\cdots,p_N)\in\mathbb{R}^N_+} & \sum_{j=1}^N p_j f(x,\xi^j) \\ & \text{s.t.} & x\in X, \\ & & \sum_{j=1}^N p^j \Psi(\xi^j) \preceq 0, \\ & & \sum_{j=1}^N p_j = 1. \end{array} \tag{4.29}$$

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

$$\begin{array}{ll} \inf\limits_{x,\lambda_0,\Lambda_1,\cdots,\Lambda_p} & \lambda_0 \\ \text{s.t.} & x\in X, \lambda_0\in\mathbb{R}, \\ & \Lambda_i \succeq 0, \text{for } i=1,\cdots,p, \\ & f(x,\xi^j) - \lambda_0 - \sum_{i=1}^p \Lambda_i \circ \Psi_i(\xi^j) \le 0,\ j=1,\cdots,N, \end{array} \tag{4.30}$$

which is equivalent to (3.19). This means the randomization scheme in Section 4 is equivalent to the discretization scheme (4.29). From numerical point of view, the difference between (4.29) and (4.30) 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, (4.29) becomes a saddle point problem. In the previous section, we have developed a numerical method for solving (4.30). Here our focus is on a numerical scheme which solves (4.29) directly for fixed Ξ^N.

Our idea is based on classical cutting plane method to be applied to the convex function $\sup_{P\in\mathcal{P}_N}\mathbb{E}_P[f(x,\xi)]$ 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 solve the outer minimization problem w.r.t. x for $\mathbb{E}_{\mathbf{p}^0}[f(x^1,\xi)]$ over X. Let x^1 and σ^1 denote the optimal solution and optimal value respectively. For fixed x^1, we solve 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. If $\mathbb{E}_{\mathbf{p}^1}[f(x^1,\xi)]=\sigma^1$, stop. Otherwise, we solve the outer maximization problem for $\max_{P\in\{\mathbf{p}^0,\mathbf{p}^1\}}\mathbb{E}_P[f(x,\xi)]$. In this way, we generate a sequence of approximate optimal solutions $\{x^t\}$ and a sequence of "cutting planes" [2] $\{\mathbb{E}_{P^i}[f(\cdot,\xi)]\}$ whose maximum forms a lower approximation of $\sup_{P\in\mathcal{P}_N}\mathbb{E}_P[f(x,\xi)]$. Indeed, $\sup_{P\in\mathcal{P}_N}\mathbb{E}_P[f(x,\xi)]$ is the lower envelope of the set of functions $\{\mathbb{E}_P[f(x,\xi): P\in\mathcal{P}_N\}$. The iterative process identifies a function $\mathbb{E}_P[f(\cdot,\xi)]$ whose epigraph supports the epigraph of $\sup_{P\in\mathcal{P}_N}\mathbb{E}_P[f(\cdot,\xi)]$ at point $(x^t,\mathbb{E}_P[f(x^t,\xi)])$.

Algorithm 4.1 (Cutting plane method for problem (4.29)) Let $\mathbf{p}^t:=(p_1^t,\cdots,p_N^t)$ and $\mathbf{p}^0\in\mathcal{P}_N$. Let $\mathcal{P}^0:=\{\mathbf{p}^0\}$. Set $t:=0$.

Step 1. Solve outer minimization problem

$$\begin{array}{ll} \min\limits_{x,\sigma} & \sigma \\ \text{s.t.} & x\in X, \\ & \sum_{j=1}^N p_j^t f(x,\xi^j)\le\sigma, \text{for } \mathbf{p}^t\in\mathcal{P}^t. \end{array} \tag{4.31}$$

[2] In the case when f is nonlinear in x, they are not really planes, here we call them cutting planes to simplify the terminology.

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

Step 2. Solve the inner maximization problem

$$\begin{array}{ll} \max\limits_{(p_1,\cdots,p_N)\in\mathbb{R}_+^N} & \sum_{j=1}^N p_j f(x^t,\xi^j) \\ \text{s.t.} & \sum_{j=1}^N p_j \Psi(\xi^j) \preceq 0, \\ & \sum_{j=1}^N p_j = 1. \end{array} \tag{4.32}$$

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

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

It might be helpful to add a note on the stopping criterion. Let $v_N(x) := \sup_{P\in\mathcal{P}_N} \mathbb{E}_P[f(x,\xi)]$ and $v_N^t(x) := \sup_{P\in\mathcal{P}^t} \mathbb{E}_P[f(x,\xi)]$. Then $v_N^t(x) \le v_N(x)$ for all $x \in X$. Moreover, since $\sigma^t = \min_{x\in X} v_N^t(x)$ (with corresponding minimizer x^t) and $v^t = v_N(x^t)$, then $\sigma^t \le v^t$. The algorithm terminates when $\sigma^t = v^t$. If this happens in a finite number of iterations for some t_0, then $v_N(x)$ must achieve its minimum over X at x_{t_0}. This can be seen easily in that there was $x' \in X$ with $v_N(x') < v_N(x_{t_0})$, then we would have $v_N^t(x') \le v_N(x') < v_N(x_{t_0})$.

Algorithm 4.1 is inspired by a similar algorithm proposed by Pflug and Wozabal [31] for solving a distributionally robust portfolio problem and cutting surface method by Mehrotra and Papp [29] 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/surfaces 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 a similar proof to a proposition in [31], we can easily establish the convergence of Algorithm 4.1.

Theorem 4.1 *Let $\{x^t, P^t\}$ be a sequence generated by Algorithm 4.1. Then x^t converges to the optimal solution of problem (4.29).*

Proof. Observe first that $\mathcal{P}_N$ is a convex and compact set in $\mathbb{R}^N$. The compactness is because Ξ^N is a finite set and the convexity is due to the linearity of the moment constraints.

Second, the sequence $\{\mathcal{P}_N^t\}$ is monotonically increasing in the sense that $\mathcal{P}_N^t \subset \mathcal{P}_N^{t+1}$ and $\mathcal{P}_N^t \subset \mathcal{P}_N$ for all t. Therefore there exists $\mathcal{P}_N^* \subset \mathcal{P}_N$ such that

$$\lim_{t\to\infty} \mathcal{P}_N^t = \mathcal{P}_N^*. \tag{4.33}$$

Third, the algorithm procedures indicate that (x^t, P^t) is a saddle point of the following minimax optimization problem

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

i.e.

$$\max_{P\in\mathcal{P}_N^t}\langle P, f(x^t,\xi)\rangle \le \langle P^t, f(x^t,\xi)\rangle \le \min_{x\in X}\langle P^t, f(x,\xi)\rangle. \tag{4.34}$$

Since X and $\mathcal{P}_N$ are compact, we may assume without loss of generality $(x^t, P^t) \to (x^*, P^*)$ as $t \to \infty$. Moreover, by (4.33) and [10, Proposition 4.4],

$$\lim_{t\to\infty}\max_{P\in\mathcal{P}_N^t}\langle P, f(x^t,\xi)\rangle = \max_{P\in\mathcal{P}_N^*}\langle P, f(x^*,\xi)\rangle \tag{4.35}$$

and

$$\lim_{t\to\infty}\max_{P\in\mathcal{P}_N}\langle P, f(x^t,\xi)\rangle = \max_{P\in\mathcal{P}_N}\langle P, f(x^*,\xi)\rangle. \tag{4.36}$$

Through (4.34) and (4.35), we arrive at

$$\max_{P\in\mathcal{P}_N^*}\langle P, f(x^*,\xi)\rangle \le \langle P^*, f(x^*,\xi)\rangle \le \lim_{t\to\infty}\min_{x\in X}\langle P^t, f(x,\xi)\rangle = \min_{x\in X}\langle P^*, f(x,\xi)\rangle. \tag{4.37}$$

The last equality follows from the classical stability result in parametric programming, see e.g., [24, Theorem 1]. Since

$$\max_{P\in\mathcal{P}_N}\langle P, f(x^t,\xi)\rangle = \max_{P\in\mathcal{P}_N^{t+1}}\langle P, f(x^t,\xi)\rangle$$

for all t, then by (4.36), we have

$$\begin{aligned}\max_{P\in\mathcal{P}_N}\langle P, f(x^*,\xi)\rangle &= \lim_{t\to\infty}\max_{P\in\mathcal{P}_N}\langle P, f(x^t,\xi)\rangle\\ &= \lim_{t\to\infty}\max_{P\in\mathcal{P}_N^{t+1}}\langle P, f(x^t,\xi)\rangle\\ &= \lim_{t\to\infty}\max_{P\in\mathcal{P}_N^{t}}\langle P, f(x^t,\xi)\rangle = \max_{P\in\mathcal{P}_N^*}\langle P, f(x^*,\xi)\rangle.\end{aligned}$$

Combining the above equation and (4.37), we arrive at

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

This shows that (x^*, P^*) is a saddle point of problem

$$\min_{x\in X}\max_{P\in\mathcal{P}_N}\langle P, f(x,\xi)\rangle.$$

The proof is complete. ■

Note that Algorithm 4.1 is proposed for solving the discretized minimax problem (4.29) for fixed sample. It might be interesting to ask whether the optimum obtained from the sampling scheme converges to the optimum of the original DRO (1.1). The following theorem addresses this.

Theorem 4.2 *Let x_N be the optimal solution of problem (4.29). Assume: (a) there exists probability measure P_0 such that $\langle P_0, \Psi_i(\xi)\rangle \prec 0$, $i = 1, \cdots, p$, (b) for any $\epsilon > 0$ and $\xi \in \Xi$, there exists $\xi' \in \Xi^N$ such that $\|\xi - \xi'\| \leq \epsilon$ almost surely as N sufficiently large. 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 (4.29), there exists P_N such that (x_N, P_N) is a saddle point of $\max_{P\in\mathcal{P}_N}\langle P, f(x,\xi)\rangle$, i.e.,

$$\max_{P\in\mathcal{P}_N}\langle P, f(x^N,\xi)\rangle \leq \langle P_N, f(x^N,\xi)\rangle \leq \min_{x\in X}\langle P_N, f(x,\xi)\rangle. \tag{4.38}$$

By taking a subsequence if necessary, we may assume that $x_N \to x^*$ and $P_N \to P^*$ weakly. From the second inequality above, we obtain $\langle P^*, f(x^*,\xi)\rangle \leq \min_{x\in X}\langle P^*, f(x,\xi)\rangle$. In what follows, we show

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

In doing so, we will arrive at

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

which means (x^*, P^*) is a saddle point of $\max_{P\in\mathcal{P}}\langle P, f(x,\xi)\rangle$.

Assume for the sake of a contradiction that there exists $\hat{P} \in \mathcal{P}$ such that

$$\langle \hat{P}, f(x^*,\xi)\rangle > \langle P^*, f(x^*,\xi)\rangle. \tag{4.39}$$

By the definition, $\mathcal{P}$ is a convex set hence for any $0 < \lambda < 1$, $\lambda\hat{P} + (1-\lambda)P_0 \in \mathcal{P}$ and

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

For fixed λ, there exists $\hat{P}_N^\lambda \in \mathcal{P}_N$ such that $\hat{P}_N^\lambda$ converges in distribution to $P_\lambda := \lambda\hat{P}+(1-\lambda)P_0$. To see this, we note that for any positive number ϵ, we may construct a partition of support set Ξ, denoted by $\{\Xi_1, \cdots, \Xi_k\}$, and a probability measure P_k such that int $\Xi_s \neq \emptyset$, for $s = 1, \cdots, k$, $P_k(\Xi_s) = p_s$, $\sum_{s=1}^k p_s = 1$ and

$$\sum_{s=1}^k |P_\lambda(\Xi_s) - p_s| \leq \epsilon.$$

Under condition (b), for each Ξ_s, there exists $\xi^{N_s} \in \Xi_N$ such that $\xi^{N_s} \in \Xi_s$ almost surely. Let $\hat{P}_N^\lambda = \sum_{s=1}^k p_s \mathbb{1}_{\xi^{N_s}}$, where $\mathbb{1}_{\xi^{N_s}}$ denotes the Dirac probability measure at ξ^{N_s}. Since $P_\lambda \in$ int $\mathcal{P}_N$, then $\hat{P}_N^\lambda \in \mathcal{P}_N$ by setting a small ϵ in the first place. The convergence of $\hat{P}_N^\lambda$ to P_λ is evident as ϵ can be arbitrarily small.

Let $\lambda_N \to 1$. Then $\hat{P}_N^{\lambda_N} \in \mathcal{P}_N$ and $\hat{P}_N^{\lambda_N} \to \hat{P}$ (in distribution). Since $f(x,\xi)$ is continuous in (x,ξ), by (4.39), for N sufficiently large $\langle \hat{P}_N^{\lambda_N}, f(x^*,\xi)\rangle > \langle P^*, f(x^*,\xi)\rangle$ and further that

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

which contradicts the first inequality of (4.38). ■

Note that condition (b) means that Ξ^N may be iid samples generated by any continuous distribution with support set Ξ or constructed in a deterministic manner. Condition (a) implies that the Lagrange dual of the inner maximization problem of (1.1) does not have a dual gap hence the convergence of the optimal value is covered by Theorem 3.1. Theorem 4.2 complements Theorem 3.1 by ensuring convergence of the optimal solution.

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. To simplify the discussions, we ignore the transaction fee, therefore the total value of portfolio is

$$f(x, \xi) = \xi_1 x_1 + r_2 x_2 + \cdots + \xi_n x_n,$$

where ξ_j denotes the random return rate of asset j.

In implementing the methods, we use the ambiguity set defined as in (2.10) with $\gamma_1 = 0.1$ and $\gamma_2 = 1.1$, where μ_0 and Σ_0 are computed through historical data. 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" [45].

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.

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 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 CUP 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 (3.19) but it does affect the size of inner maximization problem of (4.32).

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 (4.32) which is a key step of Algorithm 4.1, and its impact on outer minimization problem (4.31) is limited because the latter is an LP without any matrix variables.

Our conclusion is that if we have a decision problems with moderate number of decision variables but large samples, we may apply Algorithm 3.1, in the case when we have high number of decision variables but small size of sample data, then we may go with Algorithm 4.1.

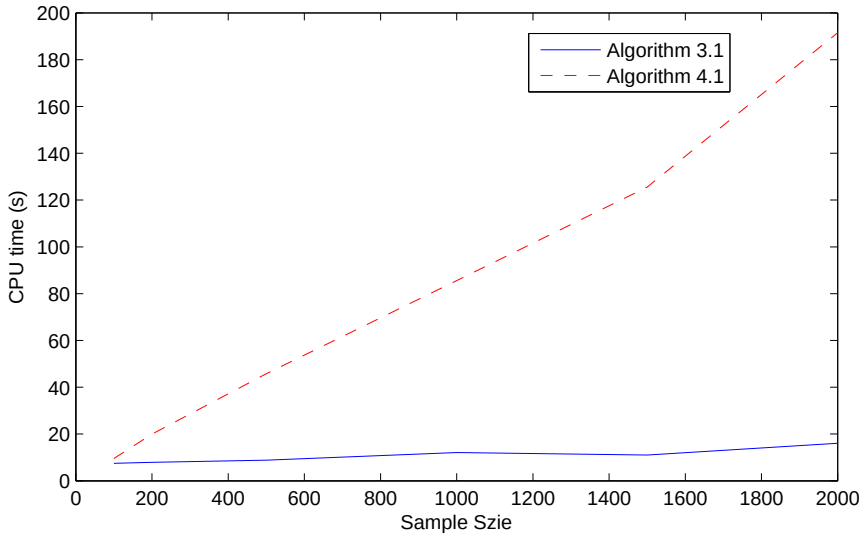

Figure 1: CUP time w.r.t sample size

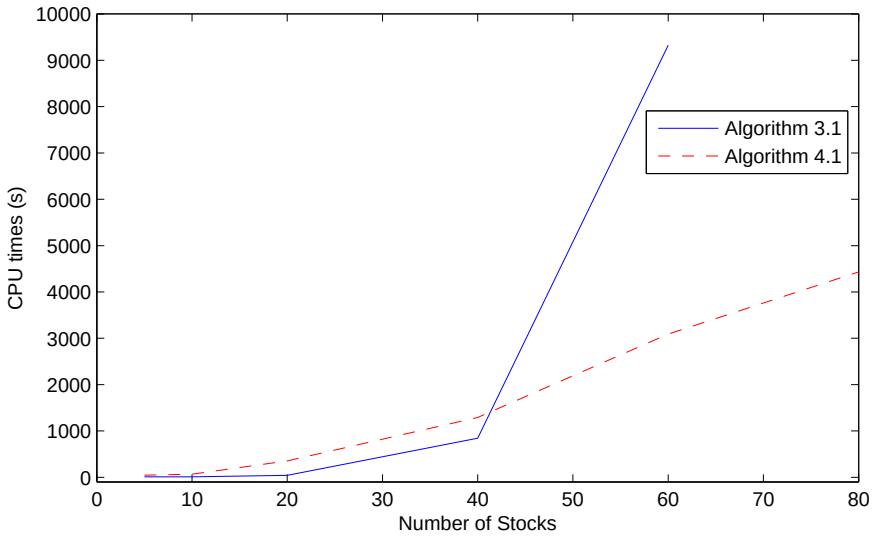

Figure 2: CUP time w.r.t the number of portfolios

References

[1] E. Anderson, H. Xu and D. Zhang, Confidence levels for CVaR risk measures and minimax limits, Optimization online, 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. C. Calafiore, Ambiguous risk measures and optimal robust portfolios, *SIAM J. Optim.*, 18: 853-877, 2007.

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

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

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

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

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

[17] E. A. Feinberg, P. O. Kasyanov and N. V. Zadoianchuk, Fatou's Lemma for weakly converging probabilities, *Theor. 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] H. Heitsch and W. Römisch, Scenario reduction algorithms in stochastic programming, *Comput. optim. Appl.*, 24: 187-206, 2003.

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

[22] R. Jiang and Y. Guan, Data-driven chance constrained stochastic program, Optimzation online, 2013.

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

[24] 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.

[25] M. Kocvara and M. Stingl, On the solution of large-scale SDP problems by the modified barrier method using iterative solvers, *Math. Prog.*, 109: 413-444, 2007.

[26] I. Kupka and V. Toma, A mannuscript, see http: //hore.dnom.fmph.uniba.sk/ s-vana/veb/preklady/TK/ch4.pdf

[27] Y. Liu and H. Xu, Entropic approximation for mathematical programs with robust equilibrium constraints, *SIAM J. Optim.*, 24: 933-958, 2014

[28] Y. Liu, W. Römisch and H. Xu, Quantitative stability analysis of stochastic generalized equations, *SIAM J. Optim.*, 24: 467-497, 2014

[29] 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.

[30] 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.

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

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

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

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

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

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

[37] 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.

[38] K. Schmüdgen, The k-moment problem for compact semialgebraic sets, *Math. Annal.*, 289: 203-206, 1991.

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

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

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

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

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

[44] H. Sun and H. Xu, Convergence analysis for distributional robust optimization and equilibrium problems, *Math. Oper. Res.*, to appear.

[45] 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.

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

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

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

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

[50] D. Wozabal, Robustifying convex risk measures for linear portfolios: a nonparametric approach, *Oper. Res.*, 62: 1302-1315, 2014.

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

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

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