Probability Approximation Schemes for Stochastic Programs with Distributionally Robust Second Order Dominance Constraints

Shaoyan Guo*, Huifu Xu† and Liwei Zhang‡

April 6, 2016

Abstract. Since the pioneering work [7] by Dentcheva and Ruszczyński, stochastic programs with second order dominance constraints (SPSODC) have received extensive discussions over the past decade from theory of optimality to numerical schemes and practical applications. In this paper, we investigate discrete approximation of SPSODC when (a) the true probability is known but continuously distributed and (b) the true probability distribution is unknown but it lies within an ambiguity set of distributions. Differing from the well-known Monte Carlo discretization method, we propose a deterministic discrete approximation scheme due to Pflug and Pichler [20] and demonstrate that the discrete probability measure and the ambiguity set of discrete probability measures approximate their continuous counterparts under the Kantorovich metric. Stability analysis of the optimal value and optimal solutions of the resulting discrete optimization problems is presented and some comparative numerical test results are reported.

Key words. Second order dominance, probability discretization, Kantorovich metric, stability analysis

1 Introduction

Consider the following stochastic program with second order dominance constraints (SPSODC):

$$\begin{array}{ll} \min\limits_{x\in X} & f(x) \\ \text{s.t.} & G(x,\xi(\omega)) \succeq_2 Y(\xi(\omega)), \end{array} \tag{1.1}$$

where X is a nonempty convex compact set of $\mathbb{R}^n$, $\xi:\Omega\to\mathbb{R}^m$ is a vector of random variables defined on space $(\Omega,\mathcal{F},P)$ with support set Ξ, f,G,Y are continuous functions mapping from $\mathbb{R}^n$, $\mathbb{R}^n\times\mathbb{R}^m$ and $\mathbb{R}^m$ to $\mathbb{R}$. The notation $\succeq_2$ means $G(x,\xi)$ dominates $Y(\xi)$ in second order in the sense

$$\int_{-\infty}^{t} P(\{\omega\in\Omega: G(x,\xi(\omega))\le\eta\})d\eta \le \int_{-\infty}^{t} P(\{\omega\in\Omega: Y(\omega)\le\eta\})d\eta,\ \forall t\in\mathbb{R}, \tag{1.2}$$

*Institute of Operations Research and Control Theory, School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China. (syguomaths@mail.dlut.edu.cn).

†School of Mathematics, University of Southampton, Southampton, SO17 1BJ, UK. (h.xu@soton.ac.uk). Haitian Scholar, Dalian University of Technology.

‡Institute of Operations Research and Control Theory, School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China. (lwzhang@dlut.edu.cn).

or equivalently (see [6])

$$\mathbb{E}_P[(t - G(x, \xi(\omega)))_+] \leq \mathbb{E}_P[(t - Y(\omega))_+], \ \forall t \in \mathbb{R}, \tag{1.3}$$

where $(t)_+ := \max\{0, t\}$, $\mathbb{E}_P[\cdot]$ denotes the mathematical expectation with respect to probability distribution P, and this is indeed a unique feaure of the stochastic optimization problem.

The SPSODC model was first introduced by Dentcheva and Ruszczyński in their pioneering work [7] and has received wide attention over the past decade for its extensive applications particularly in portfolio optimization [8] and energy planning [5].

With (1.3), we can rewrite problem (1.1) as

$$\begin{array}{ll} \min\limits_{x \in X} & f(x) \\ \text{s.t.} & \mathbb{E}_P[H(x, t, \xi(\omega))] \leq 0, \ \forall t \in \mathbb{R}, \end{array} \tag{1.4}$$

where

$$H(x, t, \xi(\omega)) := (t - G(x, \xi(\omega)))_+ - (t - Y(\xi(\omega)))_+.$$

This is a mathematical program with the stochastic semi-infinite constraint. Note that if we consider $(\Xi, \mathscr{B})$ as a measurable space equipped with Borel sigma algebra $\mathscr{B}$, then P may be viewed as a probability measure defined on $(\Xi, \mathscr{B})$ induced by ξ. Throughout the paper, we will use $\mathscr{P}$ to denote the set of all probability measures on $(\Xi, \mathscr{B})$ and use the terms probability measure and probability distribution interchangeably. Moreover, to ease notation, we will use ξ to denote either the random vector $\xi(\omega)$ or an element of $\mathbb{R}^m$ depending on the context.

An important issue concerning problem (1.4) is that it does not satisfy the well-known Slater constraint qualification (SCQ), a condition that is often needed for deriving first order optimality conditions and developing a numerically stable method for solving the problem. Subsequently, a so-called *relaxed* form of problem (1.4) is proposed:

$$\begin{array}{ll} \min\limits_{x \in X} & f(x) \\ \text{s.t.} & \mathbb{E}_P[H(x, t, \xi)] \leq 0, \ \forall t \in T, \end{array} \tag{1.5}$$

where T is a closed interval in $\mathbb{R}$. If the support set Ξ is compact, T can be chosen as the bounded set $\{Y(\xi) : \xi \in \Xi\}$, and then problems (1.4) and (1.5) are equivalent.

In the case when ξ is discretely distributed, that is, $P(\xi = \xi^i) = p_i$ with $p_i \geq 0$ for $i = 1, \ldots, N$, and $\sum_{i=1}^N p_i = 1$, problem (1.5) can be reformulated as an ordinary nonlinear programming problem with finite number of constraints:

$$\begin{array}{ll} \min\limits_{x \in X} & f(x) \\ \text{s.t.} & \sum_{i=1}^N p_i H(x, t^k, \xi^i) \leq 0, \ \forall k = 1, \ldots, N, \end{array} \tag{1.6}$$

where $t^k = Y(\xi^k)$ for $k = 1, \ldots, N$. When f and G are linear in x, Dentcheva and Ruszczyński [8] reformulate problem (1.6) as a linear programming (LP) problem by introducing new variables which represent positive parts in each constraint of problem (1.6). The reformulation effectively tackles the nonsmoothness in the second order dominance constraint and the approach can easily be applied to the case when f and G are nonlinear.

Rudolf and Ruszczyński [26] and Fábián et al [11] propose cutting-plane methods for solving a stochastic program with second order dominance constraints. A crucial element of the method

in [11] is based on the observation that when f and G are linear with respect to x and the probability space Ω is finite, the constraint function in the second order dominance constraint is the convex envelope of finitely many linear functions, which is called cutting-plane representation and observed by Haneveld and van der Vlerk in [13]. Subsequently, an iterative scheme which exploits the fundamental idea of the classical cutting-plane method is proposed where at each iterate "cutting-plane" constraints are constructed and added. This also effectively tackles the nonsmoothness issue caused by the plus function. While the method displays strong numerical performance, it relies on discreteness of the probability space as well as the linearity of f and G. Hu, Homem-de-Mello and Mehrotra [15] and Homem-de-Mello and Mehrotra [14] also propose a cut generation algorithm for solving a sample average approximation (SAA) problem of a stochastic program with multivariate stochastic dominance constraints. Different from the cutting-plane method in [26] and [11], they reformulate every subproblem as a linear programming problem by introducing some new variables when f and G are linear. In a more recent development, Sun et al [28] propose a modified cutting-plane method for solving problem (1.5) where the underlying functions may be nonlinear.

In all these works, cutting-plane methods are applied after problem (1.5) is discretized and the discretization is based on Monte Carlo sampling of ξ over Ξ. In other words, if problem (1.6) is regarded as a discretization of problem (1.5), then $p_i = \frac{1}{N}$ for all i. This is not necessarily the best approach in terms of quality of approximation as Pflug and Pichler [20] observe because by choosing ξ^i and p_i more carefully we may achieve a better effect of approximation. This is indeed one of the main reasons motivating this work.

The purpose of this paper is twofold: (a) we propose a discretization scheme for solving problem (1.5) when P is continuously distributed; (b) if the true probability distribution P is unknown, we construct an ambiguity set of distributions which contains the true probability distribution. Consequently, we consider a robust formulation of problem (1.5) to hedge risks arising from ambiguity of the true probability distribution:

$$\begin{array}{ll} \min\limits_{x\in X} & f(x) \\ \text{s.t.} & \sup\limits_{t\in T}\sup\limits_{P\in\mathcal{P}} \mathbb{E}_P[H(x,t,\xi)] \le 0, \end{array} \tag{1.7}$$

where $\mathcal{P}$ denotes an ambiguity set containing all distributions consistent with the known partial information concerning P. This kind of robust formulation is first considered by Dentcheva and Ruszczyński [9] whose focus is on necessary and sufficient optimality conditions. Here we concentrate on numerical methods for solving problem (1.7) as we believe this is an important gap to be filled out. Again, our focus will be on a discretization scheme stemming from Pflug and Pichler [20].

Of course, the structure of problem (1.7) and the necessity of discretization depend heavily on the ambiguity set $\mathcal{P}$. In the literature of distributionally robust optimization, various ways have been proposed to construct $\mathcal{P}$ depending on availability of information on P. Here we consider a popular approach where $\mathcal{P}$ is defined through moments, that is,

$$\mathcal{P} := \{P \in \mathscr{P} : \mathbb{E}_P[\varphi(\xi)] \le 0\},$$

where $\varphi : \Xi \to \mathbb{R}^l$ is a measurable function.

Note that problem (1.7) does not satisfy the Slater constraint qualification in general, so we

may consider a relaxation of problem (1.7):

$$\begin{array}{ll} \min\limits_{x\in X} & f(x) \\ \text{s.t.} & \sup\limits_{t\in T}\sup\limits_{P\in\mathcal{P}} \mathbb{E}_P[H(x,t,\xi)] \le \tau, \end{array} \tag{1.8}$$

where τ is a small positive constant. As far as we are concerned, the main contributions of this paper can be summarized as follows.

- When the true probability distribution P is known and continuous, we propose to apply the optimal quantization scheme due to Pflug and Pichler [20] to approximate it (see Section 2.1) as opposed to Monte Carlo method. This scheme is preferable when either the range of $Y(\xi)$ is large or $G(x,\xi)$ does not have an analytic form, or the sample size is small. We establish convergence of the optimal value and optimal solutions against variation of the probability measure (Theorem 3.1).
- We consider the case that the true probability distribution is unknown but it lies in an ambiguity set of distributions defined through moment conditions. Under some moderate conditions, we derive convergence of the discretized ambiguity sets to $\mathcal{P}$ under the Kantorovich metric (Theorem 2.2), and establish a key stability result (Theorem 3.2) underpinning the approximation. Based on the approximation schemes, we apply the well-known cutting-plane method to solve the resulting discretized optimization problems (Section 4.1) and report some comparative numerical results (Section 4.2).

Throughout this paper, we use the following notations. $\mathbb{R}^n$ and $\mathbb{R}^n_+$ represent the n-dimensional Euclidean space and its nonnegative part respectively. x^Ty denotes the scalar product of two vectors x and y, $\|\cdot\|$ denotes the Euclidean norm of a vector. $d(x,A) := \inf_{x'\in A}\|x-x'\|$, denotes the distance from a point x to a set A in the Euclidean norm. For two compact sets A and B in $\mathbb{R}^n$, we write $\mathbb{D}(A,B) := \sup_{x\in A} d(x,B)$ for the deviation of A from B and $\mathbb{H}(A,B) := \max\{\mathbb{D}(A,B),\mathbb{D}(B,A)\}$ for the Hausdorff distance between A and B.

2 Discrete approximation of probability measures

In this section, we discuss discrete approximation of the true probability distribution P in problem (1.5) and the ambiguity set $\mathcal{P}$ in problem (1.8). In doing so, we may develop an approximation of the semi-infinite constraints in these two problems. To this end, we introduce two metrics in the set of probability measures $\mathscr{P}$: Kantorovich metric and pseudo-metric. The former is used for quantifying approximation of probability measures whereas the latter is used for stability analysis of discrete counterparts of problems (1.5) and (1.8).

Let $\mathscr{L}$ denote the space of all Lipschitz continuous functions $h : \Xi \to \mathbb{R}$ with Lipschitz constant no larger than 1 and $P, Q \in \mathscr{P}$ be two probability measures, the Kantorovich metric (or distance) of P and Q, denoted by $d_K(P,Q)$, is defined by

$$d_K(P,Q) := \sup_{h\in\mathscr{L}} \left\{ \int_\Xi h(\xi)P(d\xi) - \int_\Xi h(\xi)Q(d\xi) \right\}.$$

By the Kantorovich-Rubinstein theorem [16],

$$d_K(P,Q) = \inf\left\{ \int_{\Xi\times\Xi} \|\xi_1-\xi_2\|\pi(d\xi_1,d\xi_2) : \begin{array}{l} \pi \text{ is the joint distribution of } \xi_1 \text{ and } \xi_2 \\ \text{with marginals } P \text{ and } Q, \text{respectively} \end{array} \right\}.$$

The latter formulation is also known as Kantorovich formulation of Monge's transportation problem if we view P as goods spread over Ξ to be relocated with new spread Q over Ξ and $\|\xi_1 - \xi_2\|$ as unit transportation cost [23]. Using the Kantorovich metric, we can quantify the distance between two sets of probability measures. Let $\mathcal{P}, \mathcal{Q} \subset \mathscr{P}$ be two sets of probability measures, we can define

$$\mathbb{D}_K(\mathcal{P}, \mathcal{Q}) := \sup_{P \in \mathcal{P}} \inf_{Q \in \mathcal{Q}} d_K(P, Q)$$

which quantifies the deviation of $\mathcal{P}$ from $\mathcal{Q}$ and

$$\mathbb{H}_K(\mathcal{P}, \mathcal{Q}) := \max\left\{\mathbb{D}_K(\mathcal{P}, \mathcal{Q}), \mathbb{D}_K(\mathcal{Q}, \mathcal{P})\right\}$$

that quantifies the distance between $\mathcal{P}$ and $\mathcal{Q}$.

An important property of the Kantorovich metric is that it metrizes weak convergence of probability measures when the support set is bounded, that is, a sequence of probability measures $\{P_N\}$ converges to P weakly if and only if $d_K(P_N, P) \to 0$ as N tends to infinity.

Recall that $\{P_N\}$ 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 $\mathcal{A}$ on $(\Xi, \mathscr{B})$, $\mathcal{A}$ is said to be *tight* if for any $\epsilon > 0$, there exists a compact set $\Xi_\epsilon \subset \Xi$ such that $\inf_{P \in \mathcal{A}} P(\Xi_\epsilon) > 1 - \epsilon$. In the case when $\mathcal{A}$ is a singleton, it reduces to the tightness of a single probability measure. $\mathcal{A}$ is said to be *closed* (under the weak topology) if for any sequence $\{P_N\} \subset \mathcal{A}$ with P_N converging to P weakly, we have $P \in \mathcal{A}$. $\mathcal{A}$ 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 $P_{N'} \to P$ weakly; see Billingsley [2].

By the well-known Prokhorov's theorem (see [1]), a closed set $\mathcal{A}$ (under the weak topology) of probability measures is *compact* if it is tight. In particular, if Ξ is a compact metric space, then the set of all probability measures on $(\Xi, \mathscr{B})$ is compact; see [22, Theorem 1.12].

We now turn to define another metric which is needed for stability analysis later on. Define the set of functions:

$$\mathscr{G} := \{g(\cdot) := H(x, t, \cdot) : x \in X, t \in T\}.$$

The distance function for the elements in $\mathscr{P}$ is defined as

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

By definition, $\mathscr{D}(P, Q) = 0$ if and only if $\mathbb{E}_P[g] = \mathbb{E}_Q[g]$ for all $g \in \mathscr{G}$. However it does not necessarily mean that $P = Q$ unless the set $\mathscr{G}$ is sufficiently large. For this reason, $\mathscr{D}(P, Q)$ is called *pseudo-metric* in that it satisfies all other properties of a metric. This type of pseudo-metric is widely used for stability analysis in stochastic programming; see an excellent review paper by Römisch [25].

Let $P \in \mathscr{P}$ be a probability measure and $\mathcal{A}_i \subset \mathscr{P}$, $i = 1, 2$, be two sets of probability measures. With the pseudo-metric, the distance from a single probability measure P to a set of

probability measures $\mathcal{A}_1$ may be defined as $\mathscr{D}(P, \mathcal{A}_1) := \inf_{Q\in\mathcal{A}_1} \mathscr{D}(P, Q)$, the deviation (excess) of $\mathcal{A}_1$ from (over) $\mathcal{A}_2$ as

$$\mathscr{D}(\mathcal{A}_1, \mathcal{A}_2) := \sup_{P\in\mathcal{A}_1} \mathscr{D}(P, \mathcal{A}_2).$$

It is easy to verify that $\mathscr{D}(\mathcal{A}_1, \mathcal{A}_2) = 0$ when $\mathcal{A}_1 \subset \mathcal{A}_2$. We can also define the Hausdorff distance between $\mathcal{A}_1$ and $\mathcal{A}_2$ under the pseudo-metric:

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

2.1 Pflug and Pichler's optimal quantization scheme

As we discussed in the introduction, a key step towards solving problem (1.5) is to discretize the probability measure if it is continuously distributed. Let us treat problem (1.6) as an approximation regime. There are essentially three techniques which can be used for this purpose: Monte Carlo methods, Quasi-Monte Carlo methods and the optimal quantization of probability measures due to Pflug and Pichler [20].

Monte Carlo methods are based on drawing independent and identically distributed random samples $\{\xi^1, \ldots, \xi^N\}$ of ξ to construct the empirical measure $P_N := \frac{1}{N}\sum_{i=1}^N \delta_{\xi^i}$ which approximates the true distribution, Here δ_{ξ^i} denotes the Dirac probability measure at ξ^i. This has been extensively investigated in the literature so we will not discuss it in this paper. Quasi-Monte Carlo methods are based on the basic idea of replacing the random samples in Monte Carlo methods by deterministic points generated by a low-discrepance recursion; see [19]. Again, we will not focus on this method in this paper.

The third approach is to find a discrete probability measure which approximates P optimally under the Kantorovich metric. Compared to the other two methods, this method has the highest approximation quality with relatively fewer samples; see comprehensive discussion by Pflug and Pichler [20]. In our context, there are at least two cases that a good discrete approximation with smaller samples is preferable. One is that the range of $Y(\xi)$ is large. By adopting a small set of samples we may effectively reduce the number of constraints in problem (1.5). The other is that $G(x, \xi)$ does not have an analytic form. This may happen when $G(x, \xi)$ is the optimal value of a second stage programming problem, see Claus and Schultz [4] where they consider a two stage stochastic program with first order dominance constraints.

In what follows, we present some known results about optimal discrete approximation of probability measures, most of which are extracted from Pflug and Pichler [20].

Let $\mathcal{P}_N$ denote the set of all probability measures $\sum_{i=1}^N p_i\delta_{\xi^i}$ on $\mathbb{R}^m$ sitting on at most N points $\{\xi^1, \ldots, \xi^N\}$. The optimal probability measure, denoted by P_N, satisfies

$$d_K(P, P_N) \text{ is close to } \inf\{d_K(P, Q) : Q \in \mathcal{P}_N\}. \tag{2.9}$$

As discussed in [20], in some special cases such as when P is Laplace distribution in $\mathbb{R}$ or exponential distribution in $\mathbb{R}$, the optimal solution can be found in an analytic manner. In

general cases, if N points $\{\xi^1, \ldots, \xi^N\}$ are given, we can define a Voronoi partition $\{\Xi_1, \ldots, \Xi_N\}$ of Ξ, where Ξ_i are pairwise disjoint with

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

The possible optimal probability weights p_i for minimizing $d_K(P, \sum_{i=1}^N p_i \delta_{\xi^i})$ can then be found by

$$p = (p_1, \ldots, p_N) \text{ with } p_i = P(\Xi_i), \tag{2.10}$$

and the optimal probability weights are unique iff

$$P(y : \|y - \xi^s\| = \|y - \xi^k\|, \text{ for some } s \neq k) = 0.$$

Since P is assumed to be absolutely continuous with respect to the Lebesgue measure, then the optimal weights are unique and do not depend on the choice of a partition. However, choosing optimal N points $\{\xi^1, \ldots, \xi^N\}$ is difficult. Following [12, Lemma 3.1], it requires to solve a nonconvex optimization problem:

$$\min \left\{ D_{d_K}(z) : z \in \{\xi^1, \ldots, \xi^N\} \in (\mathbb{R}^m)^N \right\},$$

where

$$D_{d_K}(z) := \int_\Xi \min_i \|\xi - \xi^i\| dP(\xi).$$

We refer interested readers to Chapter 4 in [21] for some algorithms for solving the above problem.

Let $q_{N,d_K}(P)$ be the *N-th quantization error* of P when it is approximated by P_N, that is,

$$q_{N,d_K}(P) := \inf \left\{ d_K \left(P, \sum_{i=1}^N p_i \delta_{\xi^i} \right) : \xi^i \in \mathbb{R}^m, p_i \geq 0, \sum_{i=1}^N p_i = 1 \right\}. \tag{2.11}$$

The following theorem states the rate at which $q_{N,d_K}(P)$ converges to zero as N tends to infinity.

Theorem 2.1 *([12],[21, Corollary 4.21]) Suppose P has a density ρ with $\int_\Xi |\xi|^{1+\delta} \rho(\xi) d\xi < \infty$ for some $\delta > 0$, then the following assertions hold.*

(i)

$$\overline{q}_{d_K}(P) := \inf_N N^{1/m} q_{N,d_K}(P) = \overline{q}^{(m)}_{d_K} \Big(\int_\Xi \rho(\xi)^{\frac{m}{m+1}} d\xi \Big)^{\frac{m+1}{m}}, \tag{2.12}$$

where $\overline{q}^{(m)}_{d_K} := \inf_N N^{1/m} q_{N,d_K}(U[0,1]^m)$ and $U[0,1]^m$ is the uniform distribution on the m-dimensional unit cube $[0,1]^m$.

(ii) There exists an approximation P_N^* sitting on no more than N points such that

$$d_K(P, P_N^*) = O(N^{-\frac{1}{m}}). \tag{2.13}$$

Remark 2.1 Based on formula (2.12), the optimal asymptotic point density for ξ^i is proportional to $\rho^{\frac{m}{m+1}}$. In $\mathbb{R}^1$ it means to solve the following quantile equations:

$$\int_{-\infty}^{\xi^i} \rho^{\frac{1}{2}}(\xi)d\xi = \frac{2i-1}{2N}\int_{-\infty}^{+\infty} \rho^{\frac{1}{2}}(\xi)d\xi,$$

for $i = 1, \ldots, N$. From the result in (2.10), we know

$$p_i = \int_{\frac{\xi^{i-1}+\xi^i}{2}}^{\frac{\xi^i+\xi^{i+1}}{2}} \rho(\xi)d\xi,$$

with $\xi^0 := -\infty$ and $\xi^{N+1} := +\infty$. Then $\sum_{i=1}^N p_i\delta_{\xi^i}$ converges to P weakly; see [21, Page 148].

It can be seen from Theorem 2.1 that in order to obtain an approximating measure with distance d_K no more than ϵ, a total of at least $N = O(\epsilon^{-m})$ supporting points is needed. Suppose there exists a nondecreasing function $h : \mathbb{R}_+ \to \mathbb{R}_+\backslash\{0\}$ such that for each $x \in X$ and ξ, $\xi' \in \Xi$,

$$|G(x,\xi) - G(x,\xi')| \le h(\|x\|)\|\xi - \xi'\|, \tag{2.14}$$

then it follows from [10, Page 499] that

$$\mathscr{D}(P_N, P) \le h(\delta)d_K(P_N, P) = O(N^{-\frac{1}{m}}), \tag{2.15}$$

where $\delta := \sup_{x\in X}\|x\|$. This implies at least $N = O(\epsilon^{-m})$ supporting points are necessary to obtain an approximating measure such that the distance to the original measure with respect to the pseudo-metric is at most ϵ.

2.2 Discrete approximation of the ambiguity set

We now return to discuss discrete approximation of the ambiguity set in problem (1.8). The technique and the necessity for discretization depend largely on the structure of the ambiguity set. Here we focus on the case when $\mathcal{P}$ is defined via moments, that is,

$$\mathcal{P} := \{P \in \mathscr{P} : \mathbb{E}_P[\varphi(\xi)] \le 0\}, \tag{2.16}$$

where $\varphi : \Xi \to \mathbb{R}^l$ is continuous function and Ξ is a compact set. Let $\Xi^N := \{\xi^1, \ldots, \xi^N\} \subset \Xi$ be a subset of Ξ, we consider the discrete set of probability distributions

$$\mathcal{P}_N := \left\{\sum_{i=1}^N p_i\delta_{\xi^i} : \sum_{i=1}^N p_i\varphi(\xi^i) \le 0, \sum_{i=1}^N p_i = 1, p_i \ge 0, i = 1, \ldots, N\right\}. \tag{2.17}$$

Obviously $\mathcal{P}_N \subset \mathcal{P}$. Our purpose is to use $\mathcal{P}_N$ to approximate $\mathcal{P}$ under some metric. Of course, the approximation depends on the choice of Ξ^N: its elements can be independent and identically distributed samples or drawn in deterministic manner. We will come back to this later.

The following theorem states convergence of $\mathcal{P}_N$ to $\mathcal{P}$ under the Kantorovich metric.

Theorem 2.2 *Assume: (a) there exists a probability measure $P_0 \in \mathscr{P}$ such that $\mathbb{E}_{P_0}[\varphi(\xi)] < 0$; (b) the sequence $\{\xi^i\}_{i\in N} \subseteq \Xi$ is such that for any $\epsilon > 0$ and $\xi \in \Xi$ there exists an index $N' \in \mathbb{N}$ satisfying $\|\xi - \xi^{N'}\| \le \epsilon$. Then $\mathbb{H}_K(\mathcal{P}, \mathcal{P}_N)$ tends to zero as N tends to infinity.*

Condition (a) is the well-known Slater constraint qualification which is widely used for moment problems, for example [27, 30, 31] and references therein. Condition (b) means any point in Ξ may be approximated by a point in Ξ^N when N is sufficiently large. The approximation scheme (using $\mathcal{P}_N$ to approximate $\mathcal{P}$) is considered by Xu, Liu and Sun [30], where they propose a cutting-plane method for solving a minimax distributionally robust optimization problem directly. However, they are short of stating convergence of $\mathcal{P}_N$ to $\mathcal{P}$ explicitly. Here we fill out the gap by showing the convergence under the Kantorovich metric.

Proof of Theorem 2.2. Since $\mathcal{P}_N \subset \mathcal{P}$, we have $\mathbb{D}_K(\mathcal{P}_N, \mathcal{P}) = 0$. Thus, we only need to show $\mathbb{D}_K(\mathcal{P}, \mathcal{P}_N) \to 0$.

Since $\mathcal{P}$ is a convex set, for any fixed $P \in \mathcal{P}$ and any positive number $\lambda \in (0,1)$, $P^\lambda := \lambda P + (1-\lambda)P_0 \in \mathcal{P}$ and $\mathbb{E}_{P^\lambda}[\varphi(\xi)] < 0$. Let $\widetilde{\Xi}_1, \ldots, \widetilde{\Xi}_N$ be a Voronoi partition with each cell $\widetilde{\Xi}_i$ centered at ξ^i. Let $P_N^\lambda = \sum_{i=1}^N p_i \delta_{\xi^i}$ with $p_i := P^\lambda(\widetilde{\Xi}_i)$. Since Ξ is a compact set, condition (b) ensures the largest diameter of Voronoi cells tends to zero as N increases. Following the discussions of [20, Section 2.1], we deduce that P_N^λ converges to P^λ under the Kantorovich metric. Since convergence with respect to the Kantorovich metric implies weak convergence, we conclude that P_N^λ converges to P^λ weakly.

Next, we show that P_N^λ satisfies the moment condition in (2.17). Since $\varphi(\cdot)$ is a continuous function and Ξ is bounded, the weak convergence guarantees

$$\lim_{N\to\infty} \mathbb{E}_{P_N^\lambda}[\varphi(\xi)] = \mathbb{E}_{P^\lambda}[\varphi(\xi)].$$

Moreover, since $\mathbb{E}_{P^\lambda}[\varphi(\xi)] < 0$, the limit above ensures $\mathbb{E}_{P_N^\lambda}[\varphi(\xi)] \le 0$ for N sufficiently large, which means $P_N^\lambda \in \mathcal{P}_N$. By driving λ to one and ϵ to zero, we deduce from the discussions above that there exists a sequence $\{P_N\}$ depending on λ and ϵ with $P_N \in \mathcal{P}_N$ such that P_N converges to P under the Kantorovich metric. Since P is drawn from $\mathcal{P}$ arbitrarily, we conclude that $\mathbb{D}_K(\mathcal{P}, \mathcal{P}_N) \to 0$. ■

Note that Theorem 2.2 is established under the condition that Ξ is compact. It might be interesting to extend the result to the case where Ξ is unbounded under some tightness conditions. We leave this for our future work as it is beyond the main scope of this paper.

In order to ensure convergence of probability measures under the pseudo-metric, we need additional conditions on function G.

Assumption 2.1 *For each $\xi \in \Xi$, $G(\cdot, \xi)$ is Lipschitz continuous on X with Lipschitz modulus being bounded by $\kappa(\xi)$, where $\sup_{\xi\in\Xi} \kappa(\xi)$ is finite.*

Corollary 2.1 *Assume the setting and conditions of Theorem 2.2 hold. Under Assumption 2.1, $\mathscr{H}(\mathcal{P}_N, \mathcal{P})$ tends to zero as N tends to infinity.*

Proof. Since $\mathcal{P}_N \subset \mathcal{P}$, then it is easy to verify that $\mathscr{D}(\mathcal{P}_N, \mathcal{P}) = 0$. So it is enough to show that $\mathscr{D}(\mathcal{P}, \mathcal{P}_N)$ converges to zero as N tends to infinity. By Theorem 2.2, for any $P \in \mathcal{P}$, there exists a sequence $\{P_N\} \subset \mathcal{P}_N$ such that P_N converges to P under the Kantorovich metric. Furthermore, P_N converges to P weakly.

Next, we prove P_N converges to P under the pseudo-metric. Assume for the sake of a contradiction that there exist a positive number $\delta > 0$ and a sequence $\{x_N, t_N\} \subset X \times T$ such

that

$$\left|\mathbb{E}_{P_N}[H(x_N,t_N,\xi)]-\mathbb{E}_P[H(x_N,t_N,\xi)]\right|\geq\delta. \tag{2.18}$$

Since $X\times T$ is compact, by taking a subsequence if necessary, we may assume for the simplicity of notation that (x_N,t_N) converges to a point $(x,t)\in X\times T$. By the triangle inequality, we obtain

$$\begin{aligned}\left|\mathbb{E}_{P_N}[H(x_N,t_N,\xi)]-\mathbb{E}_P[H(x_N,t_N,\xi)]\right| \;\leq\; & \left|\mathbb{E}_{P_N}[H(x_N,t_N,\xi)]-\mathbb{E}_{P_N}[H(x,t,\xi)]\right| \\ & +\left|\mathbb{E}_{P_N}[H(x,t,\xi)]-\mathbb{E}_P[H(x,t,\xi)]\right| \\ & +\left|\mathbb{E}_P[H(x,t,\xi)]-\mathbb{E}_P[H(x_N,t_N,\xi)]\right|. \end{aligned}\tag{2.19}$$

Under Assumption 2.1,

$$\left|\mathbb{E}_{P_N}[H(x_N,t_N,\xi)]-\mathbb{E}_{P_N}[H(x,t,\xi)]\right|\leq\|x_N-x\|\sup_{\xi\in\Xi}\kappa(\xi)+2|t_N-t|\to 0. \tag{2.20}$$

The continuity and boundedness of H in ξ and the weak convergence of P_N to P means the second term at the right hand side of (2.19) goes to zero. Likewise the third term converges to zero due to continuity of H in x,t and Assumption 2.1. All these lead to a contradiction to (2.18) as desired. ■

Corollary 2.1 essentially tells us that the convergence of $\mathcal{P}_N$ to $\mathcal{P}$ under the Kantorovich metric may be translated into convergence under the pseudo-metric. It might be interesting to draw a similar conclusion for a generic class of functions $\mathcal{G}$ in the definition of the pseudo-metric. We leave this for future research.

With $\mathcal{P}_N$ being defined as in (2.17), we may consider an approximation of problem (1.8):

$$\begin{array}{ll}\min\limits_{x\in X} & f(x)\\ \text{s.t.} & \sup\limits_{t\in T}\sup\limits_{P\in\mathcal{P}_N}\mathbb{E}_P[H(x,t,\xi)]\leq\tau.\end{array} \tag{2.21}$$

3 Stability analysis

In the preceding section, we present details about discrete probability approximations for the true probability distribution P in problem (1.5) and the ambiguity set $\mathcal{P}$ in problem (1.8). In this section, we investigate the respective problems (1.6) and (2.21) where the true probability/ambiguity set is replaced by its discrete counterpart.

3.1 Program (1.6)

Let us start with problem (1.5) and regard problem (1.6) as its approximation. Let P_N be defined as in (2.9), $\mathcal{F}(P_N), S(P_N)$ and $\vartheta(P_N)$ denote the feasible set, the set of optimal solutions and the optimal value of problem (1.6) respectively. The following theorem summarizes qualitative convergence of these quantities to their true counterpart $\mathcal{F}(P), S(P)$, and $\vartheta(P)$ of problem (1.5) as N goes to infinity.

Theorem 3.1 (Stability of program (1.6)) *Suppose problem (1.5) satisfies the Slater constraint qualification, that is, there exist a positive number γ and a point $\hat{x} \in X$ such that*

$$\max_{t \in T} \mathbb{E}_P[H(\hat{x}, t, \xi)] \leq -\gamma.$$

Then the following assertions hold.

(i) The solution set $S(P)$ is nonempty and compact, and there exists $N_1 > 0$ such that $S(P_N)$ is also nonempty for $N \geq N_1$.

(ii) There exist positive constants β and N_2 such that

$$\mathbb{H}(\mathcal{F}(P), \mathcal{F}(P_N)) \leq \beta \mathscr{D}(P_N, P),$$

for $N \geq N_2$.

(iii) $\lim_{N \to \infty} \mathbb{D}(S(P_N), S(P)) = 0.$

(iv) There exist positive numbers C and N_3 such that

$$|\vartheta(P_N) - \vartheta(P)| \leq C \mathscr{D}(P_N, P),$$

for $N \geq N_3$.

Proof. The results follow straightforwardly from [18, Proposition 2.6 and Theorem 2.7]. ■

The strength of Theorem 3.1 lies in the fact that approximations of the feasible set and the optimal value are all bounded linearly by $\mathscr{D}(P_N, P)$. The latter is linearly upper bounded by $d_K(P_N, P)$ following Remark 2.1. Therefore we can plug all existing results on quantitative description of $d_K(P_N, P)$ outlined in Section 2.1 into Theorem 3.1. For instance, in order to ensure $\mathbb{H}(\mathcal{F}(P), \mathcal{F}(P_N)) \leq \epsilon$ and $|\vartheta(P_N) - \vartheta(P)| \leq \epsilon$, we need at least $N = O(\epsilon^{-m})$ supporting points.

3.2 Program (2.21)

We now turn to investigate stability of program (2.21). Since the objective function is not affected by the discrete approximation, we concentrate our analysis on the constraint function. To facilitate the exposition, let

$$v_N(x) := \sup_{t \in T} \sup_{P \in \mathcal{P}_N} \mathbb{E}_P[H(x, t, \xi)], \tag{3.22}$$

and

$$v(x) := \sup_{t \in T} \sup_{P \in \mathcal{P}} \mathbb{E}_P[H(x, t, \xi)]. \tag{3.23}$$

Notice that the support set Ξ considered here is compact, the set T can be chosen as $\{Y(\xi) : \xi \in \Xi\}$, which is also compact under continuity of $Y(\cdot)$. Our first step is to establish the uniform convergence of $v_N(\cdot)$ to $v(\cdot)$ and Lipschitz continuity of $v(\cdot)$.

Proposition 3.1 *Under the setting and conditions of Theorem 2.2, the following assertions hold.*

(i) $v_N(x) \le v(x)$ *for all* $x \in X$ *and* $v_N(x)$ *converges uniformly to* $v(x)$ *over* X *as* N *tends to infinity, that is,*

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

(ii) If, in addition, Assumption 2.1 holds, then $v(\cdot)$ *is Lipschitz continuous on* X *with modulus being bounded by* $\sup_{P\in\mathcal{P}} \mathbb{E}_P[\kappa(\xi)]$*, that is,*

$$|v(x) - v(y)| \le \sup_{P\in\mathcal{P}} \mathbb{E}_P[\kappa(\xi)]\|x - y\|, \ \forall x, y \in X.$$

Proof. Part (i). First, the assertion $v_N(x) \le v(x)$ for all $x \in X$ holds due to the fact $\mathcal{P}_N \subset \mathcal{P}$. Now, let $x \in X$ be fixed. Define $\mathscr{V} := \{\sup_{t\in T} \mathbb{E}_P[H(x,t,\xi)] : P \in \mathcal{P}\}$ and $\mathscr{V}_N := \{\sup_{t\in T} \mathbb{E}_P[H(x,t,\xi)] : P \in \mathcal{P}_N\}$. Since Ξ is a compact set, both $\mathscr{V}$ and $\mathscr{V}_N$ are bounded subsets in $\mathbb{R}$. Let

$$a := \inf \mathscr{V}, \ b := \sup \mathscr{V}, \ a_N := \inf \mathscr{V}_N, \ b_N := \sup \mathscr{V}_N.$$

Since $\mathcal{P}_N \subset \mathcal{P}$, $\mathscr{V}_N \subseteq \mathscr{V}$, we have

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

Note that

$$b - b_N = \sup_{P\in\mathcal{P}} \sup_{t\in T} \mathbb{E}_P[H(x,t,\xi)] - \sup_{P\in\mathcal{P}_N} \sup_{t\in T} \mathbb{E}_P[H(x,t,\xi)],$$

and

$$\begin{aligned}
\mathbb{D}(\mathscr{V}, \mathscr{V}_N) = \sup_{v\in\mathscr{V}} \mathbb{D}(v, \mathscr{V}_N) &= \sup_{v\in\mathscr{V}} \inf_{v'\in\mathscr{V}_N} |v - v'| \\
&= \sup_{P\in\mathcal{P}} \inf_{Q\in\mathcal{P}_N} \left|\sup_{t\in T} \mathbb{E}_P[H(x,t,\xi)] - \sup_{t\in T} \mathbb{E}_Q[H(x,t,\xi)]\right| \\
&\le \sup_{P\in\mathcal{P}} \inf_{Q\in\mathcal{P}_N} \sup_{x\in X} \sup_{t\in T} |\mathbb{E}_P[H(x,t,\xi)] - \mathbb{E}_Q[H(x,t,\xi)]| \\
&= \mathscr{D}(\mathcal{P}, \mathcal{P}_N),
\end{aligned}$$

we obtain for any $x \in X$ that

$$\begin{aligned}
v(x) - v_N(x) &= \sup_{P\in\mathcal{P}} \sup_{t\in T} \mathbb{E}_P[H(x,t,\xi)] - \sup_{P\in\mathcal{P}_N} \sup_{t\in T} \mathbb{E}_P[H(x,t,\xi)] \\
&= b - b_N \le \mathbb{D}(\mathscr{V}, \mathscr{V}_N) \le \mathscr{D}(\mathcal{P}, \mathcal{P}_N).
\end{aligned}$$

Since x is any point in X and the right hand side of the inequality above is independent of x. By taking supremum w.r.t. x on both sides, we arrive at the conclusion.

Part (ii). The boundedness of $\mathscr{V}$ implies boundedness of $v(x)$. In what follows, we prove that $\mathscr{V}$ is closed. Let $\{v_k\} \subset \mathscr{V}$ be a sequence such that $v_k \to \hat{v}$ and $P_k \in \mathcal{P}$ with $\sup_{t\in T} \mathbb{E}_{P_k}[H(x,t,\xi)] = v_k$, we show the inclusion $\hat{v} \in \mathscr{V}$. Since $\mathcal{P}$ is weakly compact, we may assume without loss of generality that P_k converges to $P \in \mathcal{P}$ weakly. Now we claim that

$$\lim_{k\to\infty} \sup_{t\in T} |\mathbb{E}_{P_k}[H(x,t,\xi)] - \mathbb{E}_P[H(x,t,\xi)]| = 0. \tag{3.24}$$

We establish (3.24) by contradiction, suppose that there exist a positive number $\delta > 0$ and a sequence $\{t_k\} \subset T$ such that

$$|\mathbb{E}_{P_k}[H(x, t_k, \xi)] - \mathbb{E}_P[H(x, t_k, \xi)]| \geq \delta, \, \forall k. \tag{3.25}$$

Since T is compact, by taking a subsequence if necessary, we assume for the simplicity of notation that t_k converges to a point $t \in T$. By the triangle inequality,

$$\begin{aligned} \big|\mathbb{E}_{P_k}[H(x, t_k, \xi)] - \mathbb{E}_P[H(x, t_k, \xi)]\big| &\leq \big|\mathbb{E}_{P_k}[H(x, t_k, \xi)] - \mathbb{E}_{P_k}[H(x, t, \xi)]\big| \\ &\quad + \big|\mathbb{E}_{P_k}[H(x, t, \xi)] - \mathbb{E}_P[H(x, t, \xi)]\big| \\ &\quad + \big|\mathbb{E}_P[H(x, t, \xi)] - \mathbb{E}_P[H(x, t_k, \xi)]\big| \\ &\leq 4|t_k - t| + \big|\mathbb{E}_{P_k}[H(x, t, \xi)] - \mathbb{E}_P[H(x, t, \xi)]\big|, \end{aligned}$$

where the last inequality comes from the fact that H is globally Lipschitz continuous in t with modulus 2. As the second term at the right hand side of the last inequality goes to zero under the weak convergence of P_N to P, together with $t_k \to t$, a contradiction to (3.25) is obtained as desired. Therefore, we have

$$\hat{v} = \lim_{k\to\infty} v_k = \lim_{k\to\infty} \sup_{t\in T} \mathbb{E}_{P_k}[H(x, t, \xi)] = \sup_{t\in T} \mathbb{E}_P[H(x, t, \xi)] \in \mathscr{V},$$

namely $\mathscr{V}$ is closed.

For any $x \in X$, define $\Phi(x) := \{P \in \mathcal{P} : v(x) = \sup_{t\in T} \mathbb{E}_P[H(x, t, \xi)]\}$. The compactness of $\mathscr{V}$ ensures the set $\Phi(x)$ is nonempty. Let $P(y) \in \Phi(y)$, then

$$\begin{aligned} v(x) &\geq \sup_{t\in T} \mathbb{E}_{P(y)}[H(x, t, \xi)] \\ &\geq \sup_{t\in T} \mathbb{E}_{P(y)}[H(y, t, \xi)] - \big|\sup_{t\in T} \mathbb{E}_{P(y)}[H(x, t, \xi)] - \sup_{t\in T} \mathbb{E}_{P(y)}[H(y, t, \xi)]\big| \\ &\geq \sup_{t\in T} \mathbb{E}_{P(y)}[H(y, t, \xi)] - \sup_{t\in T} \big|\mathbb{E}_{P(y)}[H(x, t, \xi)] - \mathbb{E}_{P(y)}[H(y, t, \xi)]\big| \\ &\geq \sup_{t\in T} \mathbb{E}_{P(y)}[H(y, t, \xi)] - \sup_{t\in T} \mathbb{E}_{P(y)}[|(t - G(x, \xi))_+ - (t - G(y, \xi))_+|] \\ &\geq \sup_{t\in T} \mathbb{E}_{P(y)}[H(y, t, \xi)] - \mathbb{E}_{P(y)}[|G(x, \xi) - G(y, \xi)|] \\ &\geq v(y) - \sup_{P\in\mathcal{P}} \mathbb{E}_P[\kappa(\xi)]\|x - y\|. \end{aligned}$$

Exchanging the role of x and y, we can obtain the conclusion. ■

To ease the exposition, we rewrite problems (1.8) and (2.21) in abstract forms:

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{s.t.} & x \in \mathcal{F}, \end{array} \tag{3.26}$$

and

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{s.t.} & x \in \mathcal{F}_N, \end{array} \tag{3.27}$$

where

$$\mathcal{F} := \{x \in X : v(x) \leq \tau\} \;\text{ and }\; \mathcal{F}_N := \{x \in X : v_N(x) \leq \tau\}$$

denote the feasible sets of the two problems respectively. Since $v_N(x) \le v(x)$ for all $x \in X$ (see Proposition 3.1 (i)),

$$\mathcal{F} \subseteq \mathcal{F}_N \text{ and } \mathbb{D}(\mathcal{F}, \mathcal{F}_N) = 0.$$

Let $\vartheta := \inf\{f(x) : x \in \mathcal{F}\}$ denote the optimal value of problem (3.26) and S the corresponding set of optimal solutions, that is, $S := \{x \in \mathcal{F} : \vartheta = f(x)\}$. Likewise, let

$$\vartheta_N := \inf\{f(x) : x \in \mathcal{F}_N\} \text{ and } S_N := \{x \in \mathcal{F}_N : \vartheta_N = f(x)\}.$$

Theorem 3.2 (Stability of program (2.21)) *Suppose that Assumptions 2.1, conditions of Theorem 2.2 hold and $\mathcal{F}$ is nonempty, then*

(i) $\lim_{N\to\infty} \mathbb{H}(\mathcal{F}_N, \mathcal{F}) = 0$;

(ii) $\lim_{N\to\infty} \vartheta_N = \vartheta$;

(iii) $\lim_{N\to\infty} \mathbb{D}(S_N, S) = 0$.

Proof. Since $\mathbb{D}(\mathcal{F}, \mathcal{F}_N) = 0$ for any fixed $N \in \mathbb{N}$, it suffices to show that $\lim_{N\to\infty} \mathbb{D}(\mathcal{F}_N, \mathcal{F}) = 0$. By virtue of [29, Lemma 4.2(i)], the latter follows from uniform convergence of $v_N(\cdot)$ to $v(\cdot)$ and continuity of $v(\cdot)$ over X. This proves assertion (i).

Now we prove assertions (ii) and (iii). Since $\mathcal{F}$ is nonempty and $\mathcal{F} \subseteq \mathcal{F}_N$, we have $\mathcal{F}_N$ is also nonempty for all $N \in \mathbb{N}$. Similar to the proof of Proposition 3.1 (ii), we can easily obtain that $v_N(\cdot)$ is also continuous over X. Together with compactness of X, we have that $\mathcal{F}_N$ is nonempty compact and by [24, Theorem 1.9], the solution set S_N is nonempty.

Let $\{x^N\} \subseteq S_N$ be a sequence satisfying $\lim_{N\to\infty} x^N = \overline{x} \in \mathcal{F}$ and an arbitrary $y^* \in S$. Then for any $N \in \mathbb{N}$, we have $y^* \in \mathcal{F} \subseteq \mathcal{F}_N$, which implies $f(y^*) \ge f(x^N)$. Consequently, the continuity of f yields $f(y^*) \ge f(\overline{x})$ and hence $\overline{x} \in S$, $\lim_{N\to\infty} \vartheta_N = \lim_{N\to\infty} f(x^N) = f(\overline{x}) = \vartheta$. Therefore, we obtain assertions (iii) and (ii). ■

4 Numerical methods

In this section, we will discuss how to solve discretized problems (1.6) and (2.21). Throughout this section, we assume that f is convex and $G(x, \xi)$ is concave in x for every fixed ξ. Consequently, both problems (1.6) and (2.21) are convex optimization programs.

We plan to apply the well-known cutting-plane method [17] to solve these problems. Note that problem (1.6) is an ordinary NLP, so we use the cutting-plane method directly. However, the structure of the robust constraints of problem (2.21) is complex, we need some reformulations before applying the cutting-plane method.

To facilitate discussion, let us rewrite here the robust constraints of problem (2.21) as

$$\sup_{t\in T^N} \sup_{P\in\mathcal{P}_N} \mathbb{E}_P[H(x, t, \xi)] \le \tau, \tag{4.28}$$

where $T^N = \{t^1, \ldots, t^N\}$ with $t^k = Y(\xi^k)$ for $k = 1, \ldots, N$. For fixed $t \in T^N$, the inner maximization in P can be formulated as a LP:

$$\begin{array}{ll} \sup\limits_{(p_1,\ldots,p_N)\in\Delta_N} & \sum\limits_{i=1}^{N} p_i H(x,t,\xi^i) \\ \text{s.t.} & \sum_{i=1}^{N} p_i \varphi(\xi^i) \le 0, \end{array} \tag{4.29}$$

where $\Delta_N := \{p \in \mathbb{R}_+^N : \sum_{i=1}^N p_i = 1\}$. The dual of the LP is

$$\begin{array}{ll} \inf\limits_{\lambda\ge 0,\lambda_0} & \lambda_0 \\ \text{s.t.} & \sup\limits_{i=1,\ldots,N} H(x,t,\xi^i) - \lambda_0 - \lambda^T\varphi(\xi^i) \le 0, \end{array} \tag{4.30}$$

or equivalently,

$$\inf_{\lambda\ge 0} \sup_{i=1,\ldots,N} H(x,t,\xi^i) - \lambda^T\varphi(\xi^i). \tag{4.31}$$

Equivalence between problem (4.29) and problem (4.30) can be easily obtained under the Slater condition of the moment system defining $\mathcal{P}_N$, that is, there exists $p^* \in \Delta_N$ satisfying

$$\sum_{i=1}^{N} p_i^* \varphi(\xi^i) < 0. \tag{4.32}$$

Based on the discussions above, we can recast problem (2.21) as

$$\begin{array}{ll} \min\limits_{x\in X} & f(x) \\ \text{s.t.} & \inf\limits_{\lambda\ge 0} \sup\limits_{i=1,\ldots,N} H(x,t^k,\xi^i) - \lambda^T\varphi(\xi^i) \le \tau, \text{ for } k = 1,\ldots,N, \end{array}$$

which is equivalent to

$$\begin{array}{ll} \min\limits_{x\in X,\lambda_1,\ldots,\lambda_N\ge 0} & f(x) \\ \text{s.t.} & H(x,t^k,\xi^i) - \lambda_k^T\varphi(\xi^i) \le \tau, \text{ for } i,k = 1,\ldots,N. \end{array} \tag{4.33}$$

At this point, it might be helpful to discuss briefly sufficient conditions for the boundedness of the feasible set of problem (4.33). Note that we assume explicitly that x is restricted to a compact set X, so it is enough to discuss sufficient conditions for the boundedness of λ uniformly w.r.t. $x \in X$.

Proposition 4.1 *Assume the homogeneous system of inequalities*

$$-\lambda^T\varphi(\xi) \le 0, \xi \in \Xi^N \tag{4.34}$$

has a unique solution 0. Then the feasible set of $\lambda_1, \ldots, \lambda_N$ of problem (4.33) is bounded uniformly w.r.t. $x \in X$.

Proof. Define $\mathcal{F} : X \times [a,b] \to \mathbb{R}^l$,

$$\mathcal{F}(x,t) := \{\lambda \in \mathbb{R}^l : H(x,t,\xi) - \lambda^T\varphi(\xi) \le \tau, \forall \xi \in \Xi^N\},$$

where $[a,b]$ is a bounded set including $t^1,\ldots,t^N$. We show that $\cup_{x\in X,t\in[a,b]}\mathcal{F}(x,t)$ is compact. Assume for the sake of a contradiction that this is not true, then there exists a sequence $\{x_s,t_s\}$ with $(x_s,t_s)\to(x_0,t_0)\in X\times[a,b]$, and $\lambda^s\in\mathcal{F}(x_s,t_s)$ such that $\|\lambda^s\|\to\infty$ as $s\to\infty$. Since $\lambda^s\in\mathcal{F}(x_s,t_s)$, we know

$$[H(x_s,t_s,\xi)-(\lambda^s)^T\varphi(\xi)]/\|\lambda^s\|\le\tau/\|\lambda^s\|,\ \forall\xi\in\Xi^N.$$

By taking a subsequence if necessary, we may assume that $\lambda^s/\|\lambda^s\|\to\hat{\lambda}$ with $\|\hat{\lambda}\|=1$. Letting $s\to\infty$, we have

$$-\hat{\lambda}^T\varphi(\xi)\le 0,\ \forall\xi\in\Xi^N,$$

which contradicts the assumption of the proposition. Following the uniform compactness of $\mathcal{F}(x,t)$, we can easily obtain the conclusion. ■

Note that condition (4.34) in Proposition 4.1 is guaranteed by the Slater condition (4.32), see [31, Remark 2.1] for details.

4.1 A cutting-plane method

We now turn to discuss the cutting-plane method for solving problem (4.33). By introducing a new variable y, we can write (4.33) in an epigraphical form:

$$\begin{array}{ll}\min\limits_{x\in X,y\in Y,\lambda_1,\ldots,\lambda_N\ge 0} & y\\ \text{s.t.} & \psi_{i,k}(x,\lambda_1,\ldots,\lambda_N)\le\tau,\ \text{for } i,k=1,\ldots,N,\\ & f(x)-y\le 0,\end{array}\tag{4.35}$$

where

$$\psi_{i,k}(x,\lambda_1,\ldots,\lambda_N):=(t^k-G(x,\xi^i))_+-(t^k-Y(\xi^i))_+-{\lambda_k}^T\varphi(\xi^i)$$

for $i,k=1,\ldots,N$, Y is a compact and convex set including $\{f(x):x\in X\}$. Existence of Y is due to the fact that $f(\cdot)$ is continuous and X is a compact set. We apply the classical cutting-plane method to both $f(x)-y$ and $\psi_{i,k}(x,\lambda_1,\ldots,\lambda_N)$. For convenience, let $\Lambda:=(\lambda_1,\ldots,\lambda_N)\in\mathbb{R}^{lN}$.

Algorithm 4.1 (Cutting-plane method) Set $t:=0$, $S_0:=X\times Y\times Z$ with $Z\subset\mathbb{R}^{lN}_+$.

Step 1. Solve the following convex optimization problem:

$$\begin{array}{ll}\min\limits_{x,y,\Lambda} & y\\ \text{s.t.} & (x,y,\Lambda)\in S_t,\end{array}\tag{4.36}$$

and let (x_t,y_t,Λ_t) denote the optimal solution. If problem (4.36) is infeasible, stop: the original problem is infeasible.

Step 2. Find $\{i_t^*,k_t^*\}$ such that

$$\{i_t^*,k_t^*\}:=\text{argmax}\{\psi_{i,k}(x_t,\Lambda_t),i,k=1,\ldots,N\}.$$

Step 3. If $\psi_{i_t^*,k_t^*}(x_t,\Lambda_t)\le\tau$ and $f(x_t)-y_t\le 0$, stop, return (x_t,y_t,Λ_t) as an optimal solution. Otherwise, construct feasible cuts

$$\nabla f(x_t)^Tx-y\le\nabla f(x_t)^Tx_t-f(x_t),$$

and

$$\zeta_t(x) + w_t(\Lambda) \le \zeta_t(x_t) + w_t(\Lambda_t) - \psi_{i_t^*,k_t^*}(x_t, \Lambda_t) + \tau$$

with $(\zeta_t, w_t) \in \partial\psi_{i_t^*,k_t^*}(x, \Lambda)$, where $\partial\psi$ denotes subdifferential of a convex function ψ. Set

$$S_{t+1} := S_t \bigcap \left\{ (x, y, \Lambda) : \begin{array}{l} \nabla f(x_t)^T x - y \le \nabla f(x_t)^T x_t - f(x_t), \\ \zeta_t(x) + w_t(\Lambda) \le \zeta_t(x_t) + w_t(\Lambda_t) - \psi_{i_t^*}(x_t, \Lambda_t) + \tau \end{array} \right\}.$$

Proceed with iteration $t + 1$.

Let us make a few comments on the subdifferential operation in Step 3 of Algorithm 4.1. Let $\theta(z) := \max\{0, z\}$ for $z \in \mathbb{R}$. It is well known that the subdifferential of $\theta(z)$ can be written as

$$\partial_z \theta(z) = \begin{cases} [0, 1], & \text{if } z = 0, \\ 1, & \text{if } z > 0, \\ 0, & \text{if } z < 0. \end{cases}$$

By [3, Proposition 2.3.6 and Theorem 2.3.10], we have

$$\partial_x \theta(t - G(x, \xi)) = \begin{cases} \nabla_x G(x, \xi)^T [0, 1], & \text{if } t - G(x, \xi) = 0, \\ \nabla_x G(x, \xi)^T, & \text{if } t - G(x, \xi) > 0, \\ 0, & \text{if } t - G(x, \xi) < 0. \end{cases}$$

The following theorem states convergence of Algorithm 4.1 which can be easily established similarly to Kelley [17], we omit the details here.

Theorem 4.1 *Let $\{x_t, y_t, \Lambda_t\}$ be the sequence generated by Algorithm 4.1. Let*

$$S := \{(x, y, \Lambda) \in X \times Y \times Z : f(x) - y \le 0, \psi_{i,k}(x, \Lambda) \le \tau, \text{for } i, k = 1, \ldots, N\}.$$

Assume: (a) f(x) is continuously differentiable and convex, $G(x, \xi)$ is continuously differentiable and concave w.r.t x for almost every $\xi \in \Xi$; (b) $X \times Y \times Z$ is a compact set; (c) there exists a positive constant L such that the Lipschitz moduli of $f(\cdot)$ and $\psi_{i,k}(\cdot, \Lambda)$ are bounded by L on X; (d) S is nonempty. Then $\{(x_t, y_t, \Lambda_t)\}$ contains a subsequence converging to a point $(x^*, y^*, \Lambda^*) \in S$, where (x^*, y^*, Λ^*) is the optimal solution and y^* is the optimal value of problem (4.35).

4.2 Numerical experiments

We have carried out some numerical experiments on the cutting-plane method for solving problem (1.6) and Algorithm 4.1 for solving problem (4.35) and report some preliminary results. The tests are carried out in MATLAB 8.5 installed on a Dell-PC with Windows 7 operating system and Intel Core i7-3770 processor.

Example 4.1 Consider problem (1.5) with $f(x) = -\mathbb{E}_P[x\xi]$, $G(x, \xi) = x\xi - \frac{1}{2}x^2$, $Y(\xi) = \xi - \frac{1}{2}$, $X = [0, 20]$, where the true distribution of ξ is uniform distribution over $[2, 3]$. Here we compare the optimal quantization scheme with the Monte Carlo method (where $p_i = \frac{1}{N}$) in terms of

the optimal solutions and optimal value. Specifically, the approximation problem (1.6) can be presented as

$$\begin{array}{ll}\min\limits_{x\in X} & -\sum_{i=1}^{N} p_i x\xi^i \\ \text{s.t.} & \sum_{i=1}^{N} p_i (Y(\xi^k) - x\xi^i + \frac{1}{2}x^2)_+ - (Y(\xi^k) - Y(\xi^i)_+) \le 0,\ \forall k = 1,\dots,N.\end{array} \tag{4.37}$$

Here we need to point out that the optimal discrete distribution P_N can be obtained from Remark 2.1, that is, $\xi^i = 2 + \frac{2i-1}{2N}$ for $i = 1,\dots,N$, and the corresponding probability is $p_i = \frac{1}{N}$ for all i. Note that in this problem any point in the interval $[1,3]$ is feasible, and $x = 3$ is the optimal solution with corresponding optimal value -7.5. The results are depicted in Figure 1. As we can see clearly that the optimal quantization scheme displays faster convergence as N increases.

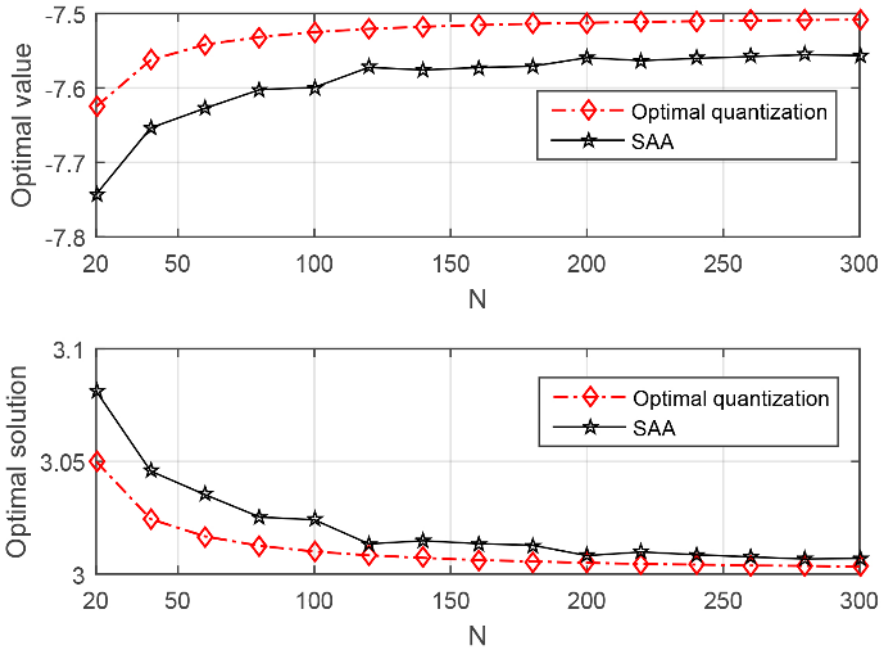

Figure 1: Optimal value and solution w.r.t N.

Next, we report our experiments on Algorithm 4.1 for a portfolio optimization problem.

Example 4.2 Consider the portfolio optimization problem with robust second order dominance constraints (1.8):

$$\begin{array}{ll}\min\limits_{x\in X} & -\mathbb{E}_P[\xi]^T x \\ \text{s.t.} & \sup\limits_{t\in T, P\in\mathcal{P}} \mathbb{E}_P[(t - \xi^T x)_+ - (t - Y(\xi))_+] \le \tau,\end{array} \tag{4.38}$$

where $X = \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1, x_i \ge 0, i = 1,\dots,n\}$. Here, the ambiguity set is defined by

$$\mathcal{P} := \left\{P \in \mathscr{P} : \mathbb{E}_P[\xi] = \mu, \|\mathbb{E}_P[(\xi-\mu)(\xi-\mu)^T - \Sigma]\|_* \le \delta\right\},$$

where μ and Σ are estimated from empirical data, and $\|A\|_* := \max |a_{ij}|$ for matrix $A = (a_{ij})$. Note that by setting $\mathbb{E}_P[\xi] = \mu$, we make the objective function deterministic so that this test problem fits into our robust model. We collect historical data of 5 assets (Admiral Group PLC, Anglo American PLC, Antofagasta PLC, AstraZeneca PLC and Aviva PLC) over a time horizon of 3 years (from 26th Nov 2010 to 18th Nov 2013) with a total 750 records on the historical stock returns (these are obtained from http://finance.google.com with adjustment for stock splitting). We have carried out out-of-sample tests with a rolling window of 400 days, that is, we use first 400 data to calculate the optimal portfolio strategy for day 401 and move on a rolling basis.

In implementing the numerical scheme, we use the equally weighted portfolio as a benchmark strategy $Y(\xi)$ and set positive numbers $\tau = 0.001$, $\delta = 2$. We compare the portfolio returns between model (4.33) and stochastic programming model (1.5) with sample average approximation over investment period of 350 days. Figure 2 depicts the performance of the three models/strategies: the robust model, the stochastic model with SAA and the benchmark. It shows that the robust model displays slightly better performance in comparison with the stochastic model although at this point, we do not have theoretical guarantee for this phenomena. We envisage that when the data contains significant fluctuations, the robust model may display a more stable performance and we will continue our research on this in our future work.

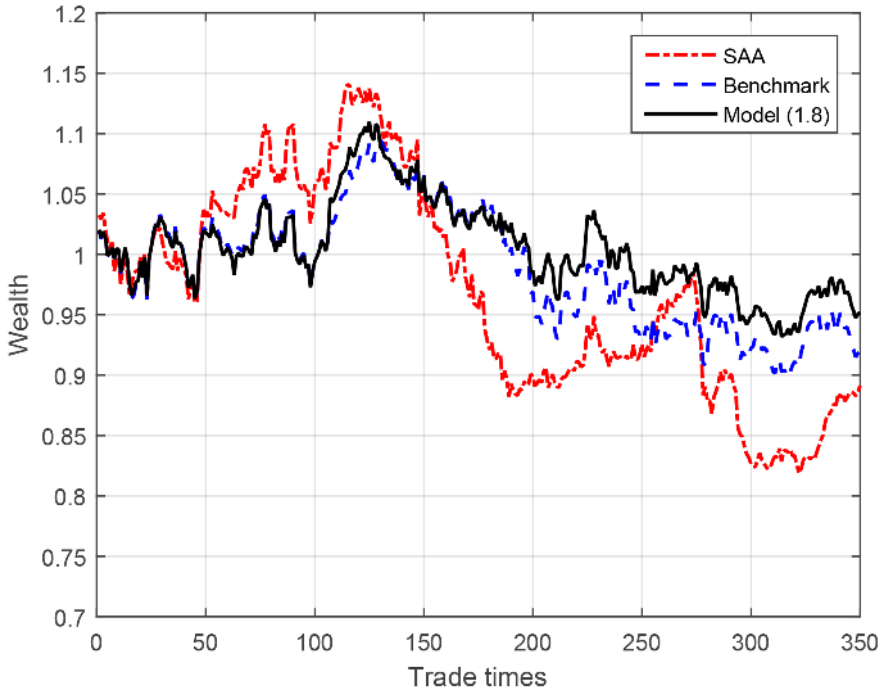

Figure 2: Wealth evolution w.r.t the trading times.

Acknowledgement. We are grateful to an anonymous referee whose insightful comments have helped us significantly strengthen the paper.

References

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

[2] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

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

[4] M. Claus and R. Schultz, Lipschitzian properties and stability of a class of first-order stochastic dominance constraints, *SIAM J. Optim.*, 25: 396-415, 2015.

[5] A. Conejo, M. Carrin and J. Morales, *Decision making under uncertainty in electricity markets*, Internation Series in Operations Research & Management Science, Springer, New York, 2010.

[6] D. Dentcheva and A. Ruszczyński, Optimality and duality theory for stochastic optimization with nonlinear dominance constraints, *Math. Program.*, 99: 329-350, 2004.

[7] D. Dentcheva and A. Ruszczyński, Optimization with stochastic dominance constraints, *SIAM J. Optim.*, 14: 548-566, 2003.

[8] D. Dentcheva and A. Ruszczyński, Portfolio optimization with stochastic dominance constraints, *J. Banking Financ.*, 30: 433-451, 2006.

[9] D. Dentcheva and A. Ruszczyński, Robust stochastic dominance and its application to risk-averse optimization, *Math. Program.*, 123: 85-100, 2010.

[10] J. Dupačová, N. Gröwe-Kuska and W. Römisch, Scenario reduction in stochastic programming: An approach using probability metrics, *Math. Program.*, 95: 493-511, 2003.

[11] C. Fábián, G. Mitra and D. Roman, Processing second-order stochastic dominance models using cutting-plane representations, *Math. Program.*, 130: 33-57, 2011.

[12] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, Lecture Notes in Math. 1730, Springer, Berlin, 2000.

[13] W. K. Klein Haneveld and M. H. van der Vlerk, Integrated chance constraints: reduced forms and an algorithm, *Comput. Manag. Sci.,* 3: 245-269, 2006.

[14] T. Homem-de-Mello and S. Mehrotra, A cutting surface method for uncertain linear programs with polyhedral stochastic dominance constraints, *SIAM J. Optim.*, 20: 1250-1273, 2009.

[15] J. Hu, T. Homen-De-Mello and S. Mehrotra, Sample average approximation of stochastic dominance constrained programs, *Math. Program.*, 133: 171-201, 2012.

[16] L. V. Kantorovich and G. S. Rubinshtein, On a space of totally additive functions, Vestnik Leningradskogo Universiteta, 13: 52-59, 1958.

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

[18] Y. Liu and H. Xu, Stability analysis of stochastic programs with second order dominance constraints, *Math. Program.*, 142: 435-460, 2013.

[19] H. Niederreiter, *Random Number Generation and Quasi Monte Carlo Methods*, SIAM, Philadelphia, 1992.

[20] G. C. Pflug and A. Pichler, *Approximations for probability distributions and stochastic optimization problems*, International Series in Operations Research & Management Science, Springer, New York, 163: 343-387, 2011.

[21] G. C. Pflug and A. Pichler, *Multistage stochastic optimization*, Springer International Publishing Switzerland, 2014.

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

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

[24] R. T. Rockafellar and R. J. B. Wets, *Variational analysis*, Springer, New York, 1998.

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

[26] G. Rudolf and A. Ruszczyński, Optimization problems with second order stochastic dominance constraints: duality, compact formulations, and cut generation methods, *SIAM J. Optim.*, 19: 1326-1343, 2008.

[27] H. Sun and H. Xu, Convergence analysis for distributionally robust optimization and equilibrium problems, *Math. Oper. Res.*, 2015, DOI 10.1287/moor.2015.0732.

[28] H. Sun, H. Xu, R. Meskarian and Y. Wang, Exact penalization, level function method and modified cutting-plane method for stochastic programs with second order stochastic dominance constraints, *SIAM J. Optim.*, 23: 602-631, 2013.

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

[30] H. Xu, Y. Liu and H. Sun, Distributionally robust optimization with matrix moment constraints: lagrange duality and cutting-plane methods, *Optim. Online*, 2015.

[31] J. Zhang, H. Xu and L.W. Zhang, Quantitative stability analysis for distributionally robust optimization with moment constraints, *Optim. Online*, 2015.