Sample Average Approximation of Stochastic Dominance Constrained Programs

Jian Hu
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
jianhu@northwestern.edu

Tito Homem-de-Mello
Department of Mechanical and Industrial Engineering
University of Illinois at Chicago
Chicago, IL 60607
thmello@uic.edu

Sanjay Mehrotra
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
mehrotra@iems.northwestern.edu

October 13, 2010

Abstract

In this paper we study optimization problems with second-order stochastic dominance constraints. This class of problems allows for the modeling of optimization problems where a risk-averse decision maker wants to ensure that the solution produced by the model dominates certain benchmarks. Here we deal with the case of multi-variate stochastic dominance under general distributions and nonlinear functions. We introduce the concept of $\mathcal{C}$-dominance, which generalizes some notions of multi-variate dominance found in the literature. We apply the Sample Average Approximation (SAA) method to this problem, which results in a semi-infinite program, and study asymptotic convergence of optimal values and optimal solutions, as well as the rate of convergence of the feasibility set of the resulting semi-infinite program as the sample size goes to infinity. We develop a finitely convergent method to find an ϵ-optimal solution of the SAA problem. An important aspect of our contribution is the construction of practical statistical lower and upper bounds for the true optimal objective value. We also show that the bounds are asymptotically tight as the sample size goes to infinity.

Key Words: Stochastic Programming, Stochastic Dominance, Sample Average Approximation, Semi-infinite Programming, Convex Programming, Cutting Plane Algorithms

1 Introduction

Stochastic dominance is used to compare the distributions of two random variables (e.g., see Shaked and Shanthikumar 1994 and Müller and Stoyan 2002), thus providing a way to measure risk. The concept of stochastic dominance is also related to utility theory (von Neumann and Morgenstern, 1947), which hypothesizes that for each rational decision maker there exists a utility function u such that the (random) outcome X is preferred to the (random) outcome Y if $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$. Often the decision maker's utility function is not known; in such cases one would say that X is preferred to Y if $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$ for all u belonging to a certain set of functions. If we have more information on the decision maker then we can restrict the set from which u is taken. In our case, we consider the situation where the decision maker is *risk-averse*; thus, X is preferred to Y if $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$ for all nondecreasing and concave utility functions u. When X and Y are unidimensional random variables, such notion is called *second order* stochastic dominance in the literature and is written $X \succeq_{(2)} Y$. This is the notion of dominance we deal with in this paper.

Dentcheva and Ruszczyński (2003, 2004) first introduced optimization problems with stochastic dominance constraints. This is an attractive approach for managing risks in an optimization setting. While pursuing expected profits, one avoids high risks by choosing options that are preferable to a random benchmark. Recently, optimization models using stochastic dominance have increasingly been the subject of theoretical considerations and practical applications in areas such as finance, energy and transportation (Karoui and Meziou, 2006; Roman et al., 2006; Dentcheva and Ruszczyński, 2006; Dentcheva et al., 2007; Drapkin and Schultz, 2007; Gollmer et al., 2007; Luedtke, 2008; Nie et al., 2009).

Much of the work on optimization with stochastic dominance has focused on the case where the underlying random quantities being compared are *unidimensional.* This is in great part due to the fact that, in that situation, it is well known that testing whether $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$ for all nondecreasing and concave utility functions u is equivalent to testing whether $\mathbb{E}[(\eta - X)_+] \leq \mathbb{E}[(\eta - Y)_+]$ for all $\eta \in \mathbb{R}$ (where $(a)_+ := \max\{a, 0\}$), a property that greatly simplifies the analysis and allows for the development of algorithms. In a recent paper, Dentcheva and Ruszczyński (2009) study a random vector space where stochastic dominance is defined using a concept of positive linear second order. For two random vectors in $\mathcal{L}_1^m$ ($\mathcal{L}_1^m$ is the space of integrable mappings from the underlying probability space to $\mathbb{R}^m$), X is said to dominate Y in positive linear second order, written $X \succeq_{(2)}^{\mathrm{Plin}} Y$, if[1]

$$v^T X \succeq_{(2)} v^T Y \qquad \text{for all } v \in \mathbb{R}_+^m.$$

Homem-de-Mello and Mehrotra (2009) further expand the definition of positive linear second order dominance to polyhedral second order dominance, written $X \succeq_{(\mathcal{P})} Y$ and called $\mathcal{P}$-dominance in short, as

$$v^T X \succeq_{(2)} v^T Y \qquad \text{for all } v \in \mathcal{P},$$

where $\mathcal{P} \subseteq \mathbb{R}_+^m$ is a polyhedron. Obviously, we have that $X \succeq_{(\mathcal{P})} Y \equiv X \succeq_{(2)}^{\mathrm{Plin}} Y$ when $\mathcal{P} = \mathbb{R}_+^m$.

A more general definition of stochastic dominance over random vectors is a natural extension of the concept of $\mathcal{P}$-dominance:

[1] Dentcheva and Ruszczyński (2009) define this notion as linear second order stochastic dominance; the concept is also related to the definition of positive linear convex order found in the literature, see for instance Müller and Stoyan (2002).

Definition 1.1 *Given a non-empty convex set $\mathcal{C} \subseteq \mathbb{R}^m_+$, a random vector $X \in \mathcal{L}^m_1$ dominates $Y \in \mathcal{L}^m_1$ in linear convex second order (written $X \trianglerighteq_{(\mathcal{C})} Y$ and called $\mathcal{C}$-dominance in short) with respect to $\mathcal{C}$ if*

$$v^T X \trianglerighteq_{(2)} v^T Y \qquad \textit{for all } v \in \mathcal{C}. \tag{1.1}$$

The three definitions of random-vector stochastic dominance presented above impose (unidimensional) second order dominance between certain weighted combinations of components of the two random vectors. One can think of such weights as a way to combine multiple criteria involved in a decision process. In many decision situations, however, it is difficult to specify exactly what the appropriate weights are, as the weights typically represent some subjective evaluation of the importance of each criteria. A robust optimization approach to the problem is to impose that the preference hold over a *set of weights*. For example, we can allow decision makers to suggest different weights, and then use the convex hull of those points for $\mathcal{P}$-dominance.

The specification of the set $\mathcal{P}$ as a convex hull of given weights is practical when the number of suggested weights is relatively small. However, such an approach may not be effective when the number of weights is large. This is the case, for example, in problems where the weights are obtained by questionnaire surveys (see Aretoulis et al. (2009) and Prato and Herath (2007) for examples of such problems in ecological and project management). The reasons for that are two-fold: first, some outliers are overemphasized in the construction of the convex hull, which is not desirable. Second, a great number of intermediate variables will be introduced in the formulation of the convex hull. To see that, let $v^1, \ldots, v^q \in \mathbb{R}^m_+$ be q weights recommended by a survey. One approach to formulate the convex hull is to write $\{v \in \mathbb{R}^m : v = [v^1, \ldots, v^q]\lambda,\ \|\lambda\|_1 = 1,\ \lambda \geq 0\}$ and treat λ as a q-dimensional decision vector. Clearly, if q is a large number this increases significantly the size of the program in comparison with the original dimension m.

One way to overcome the issue of a large number of weights is to view $v^1, \ldots, v^q$ as independent and identically distributed samples from an unknown distribution of weights. Then, we can use a multivariate statistical method to build a $100(1-\alpha)\%$ confident region $\{v \in \mathbb{R}^m_+ : q(v-\bar{v})^T S^{-1}(v-\bar{v}) \leq \chi^2_{\alpha,m}\}$ for the expected weight, where $\bar{v}$ is the mean of $v^1, \ldots, v^q$, S is the covariance matrix of those vectors and $\chi^2_{\alpha,m}$ is the $(1-\alpha)\%$ critical value of the chi-square distribution with m degrees of freedom (see Yang and Trewn (2004)). Note that such an approach addresses the issue of outliers and does not increase the dimensionality of the problem. In such cases it is necessary to apply the concept of $\mathcal{C}$-dominance in Definition 1.1, using the ellipsoid corresponding to the confidence region as the set $\mathcal{C}$.

We now introduce a characterization from Homem-de-Mello and Mehrotra (2009) which will be convenient for our analysis.

Proposition 1.1 *Let $C \subseteq \mathbb{R}^m_+$ be a convex set. Then* (1.1) *holds if and only if $v^T X \trianglerighteq_{(2)} v^T Y$ for all $v \in \widetilde{\mathcal{C}} := \mathrm{cl}(\mathrm{cone}(\mathcal{C})) \cap \Delta$, where* cl *denotes the closure of a set,* cone *denotes the conical hull, and $\Delta := \{v \in \mathbb{R}^m_+ : \|v\|_1 \leq 1\}$.*

Using this concept, we build an optimization model with stochastic dominance constraints as follows:

$$\min f(z) \tag{SD}$$

$$\text{s.t. } H(z, X) \trianglerighteq_{(\widetilde{\mathcal{C}})} Y, \tag{1.2}$$

$$z \in Z.$$

where $Z \subseteq \mathbb{R}^n$ represents a deterministic feasible region, $f : \mathbb{R}^n \to \mathbb{R}$ represents the objective to be minimized. We denote by $\Xi \subseteq \mathbb{R}^{d+m}$ the support of the probability distribution of joint random vector (X, Y). Further, let Ξ_X be the projection of Ξ on the space $\mathbb{R}^d$ for random vector X and Ξ_Y be the projection of Ξ on the space $\mathbb{R}^m$ for random vector Y. The function $H : \mathbb{R}^n \times \Xi_X \to \mathbb{R}^m$ is a given constraint mapping. We also assume that $H(z, X) \in \mathcal{L}_1^m$ for all $z \in Z$. Using the properties of second order dominance and the definition of $\mathcal{C}$-dominance, we translate (1.2) into the equivalent representation

$$E[(\eta - v^T H(z, X))_+] \ \le \ E[(\eta - v^T Y)_+] \quad \text{for all } (\eta, v) \in \mathbb{R} \times \widetilde{\mathcal{C}}. \tag{1.3}$$

For some of the results in the paper we shall need the following assumption:

(A0) The random vector Y has bounded support (i.e., the set Ξ_Y is bounded).

From the compactness of $\widetilde{\mathcal{C}}$ and boundedness of Ξ_Y, it follows that $v^T Y$ is uniformly bounded for all $v \in \widetilde{\mathcal{C}}$. In other words, there exists a closed interval $\mathcal{A} \subset \mathbb{R}$ such that $v^T Y \in \mathcal{A}$ for all $v \in \widetilde{\mathcal{C}}$ a.e.. The proposition below shows that, in such case, it is not necessary to check inequality (1.3) for all $\eta \in \mathbb{R}$.

Proposition 1.2 *Suppose that Assumption (A0) holds, and let $\mathcal{A} = [a, b]$ be an interval such that $v^T Y \in \mathcal{A}$ for all $v \in \widetilde{\mathcal{C}}$ a.e.. Then,* (1.3) *is equivalent to*

$$E[(\eta - v^T H(z, X))_+] \ \le \ E[(\eta - v^T Y)_+] \quad \textit{for all } (\eta, v) \in \mathcal{A} \times \widetilde{\mathcal{C}}. \tag{1.4}$$

Proof: If (1.3) holds, trivially (1.4) must be satisfied. Suppose now that (1.4) holds for some $z \in Z$. Since $E[(a - v^T H(z, X))_+] \le E[(a - v^T Y)_+] = 0$, $z \in Z$ satisfies $v^T H(z, X))_+ \ge a$ for all $v \in \widetilde{\mathcal{C}}$ a.e. and thus $E[(\eta - v^T H(z, X))_+] = 0$ for all $\eta \le a$. On the other hand, we have that $E[(b - v^T H(z, X))_+] \le E[(b - v^T Y)_+] = E[b - v^T Y]$. Therefore, for any $\eta \ge b$ we have

$$\begin{aligned} E[(\eta - v^T H(z, X))_+] \ &\le \ E[\eta - b + (b - v^T H(z, X))_+] \ \le \ E[\eta - b + (b - v^T Y)_+] \\ &= \ E[(\eta - v^T Y)_+]. \end{aligned}$$

Altogether, we can conclude that (1.3) holds if and only if (1.4) does. □

We can see that problem (SD) can be formulated as a stochastic program with *uncountably many* expected-value constraints. Dentcheva and Ruszczyński (2009) study models of the form (SD) where the constraints correspond to the notion of positive second order linear dominance. Many useful theoretical results are derived, but no algorithms are proposed. Homem-de-Mello and Mehrotra (2009) investigate the case of $\mathcal{P}$-dominance constraints for a linear function, $H(z, X) = Xz$, where X is a $m \times n$ random coefficient matrix with *finite* support. A key result in that paper is the proof that the set of constraints in (1.3) can be represented by finitely many deterministic linear inequalities, i.e., the feasible set is a polyhedron. Using that property, the authors develop a cutting surface algorithm where the constraints are generated one at a time. That approach, however, assumes that one can enumerate all possible scenarios of the problem and therefore evaluate expectations exactly.

In the literature, stochastic programming problems with a finite number of expected-value constraints have been widely investigated (Vogel, 1994; O'Brien, 2000; Atlason et al., 2004; Wang and

Ahmed, 2008). These papers demonstrate the difficulty in evaluating exactly an expected-value constraint, which results from the need to compute the mean of a random function by multi-dimensional integration. One way to circumvent the problem is to use the Sample Average Approximation (SAA) method (see e.g. Kleywegt et al. (2001) and references therein) which substitutes the mean with the average of Monte Carlo samples to formulate an approximation of the original program. Vogel (1994) and Shapiro (2003) study in detail the stability and consistency of the SAA optimal value and solutions. Under some mild conditions, Wang and Ahmed (2008) prove that the feasible region of the approximation model converges exponentially fast to true counterpart in probability as the sample size N increases. With a suitable sample size, good approximations can be obtained for the true optimal solution and objective value with high confidence. However, in our case the sample approximation of (1.3) is a semi-infinite program, so solving the problem — both theoretically and algorithmically — requires additional results.

In this paper, we study an approach to solve (SD) based on the SAA method. First, we show the asymptotic convergence of the approach by adjusting the right-hand-side of (1.3) with $\pm\epsilon$. The resulting restriction or relaxation of the original feasible region provides a convenient way to probabilistically measure the quality of the approximation. As in the case of finitely many expected value constraints (Wang and Ahmed, 2008), by using large deviations analysis we show that the probability that the feasible region of the semi-infinite SAA problem is close to the original feasible region converges to one exponentially fast with the sample size. Next, we consider how to solve the SAA problem with infinitely many constraints. In the case of $\mathcal{P}$-dominance requirement, we can adapt the cutting surface algorithm in Homem-de-Mello and Mehrotra (2009) that terminates after generating a finite number of cuts, although we need to overcome certain technical difficulties arising due to the fact that in a cutting surface method the expected value constraints are not available in an explicit formulation. We extend this algorithm to solve general models with $\mathcal{C}$-dominance constraints using an outer approximation approach. Finally, we propose and analyze methods for computing statistical bounds for the optimal value of the original problem. Such bounds are crucial for a practical use of the algorithm, as they provide concrete optimality gaps that can be used to determine whether the sample size is large enough. In related work we present computational results for the algorithms and methods developed in this paper for a homeland security problem (Hu et al. (2010)).

2 Notions, Assumptions and Basic Propositions

We start by introducing a reformulation of the problem and some notions and assumptions that will be used in the sequel. As in Dentcheva and Ruszczyński (2003), to overcome some technical difficulties associated with the dominance constraint (more specifically, satisfaction of the Slater condition assumed later in the paper), we consider a relaxed version of (1.3). In Dentcheva and Ruszczyński (2003) such a relaxation is imposed by restricting the set of η's in (1.3) to a specific set satisfying a certain assumption; in our case, we relax that inequality by a constant $\iota > 0$. The optimization model is changed to

$$\begin{aligned} &\min f(z) && \text{(RSD)} \\ &\text{s.t. } E[(\eta - v^T H(z,X))_+ - (\eta - v^T Y)_+] \ \leq \ \iota \quad \text{for all } (\eta, v) \in \mathcal{A} \times \widetilde{\mathcal{C}} && (2.1) \\ &\qquad z \in Z, \end{aligned}$$

where $\mathcal{A}$ is a compact interval. As shown in Proposition 1.2, when Assumption (A0) holds there is no loss of generality in restricting η to an interval $\mathcal{A}$ that covers the support of $v^T Y$ for all $v \in \widetilde{\mathcal{C}}$. Still, the results in this section and the next are valid for an arbitrary compact interval $\mathcal{A}$ and do not require boundedness of Y.

We write the expected-value function in (2.1) as

$$g(z, \eta, v) := E[G(z, \eta, v, X, Y)]$$

where the integrand is

$$G(z, \eta, v, X, Y) := (\eta - v^T H(z,X))_+ - (\eta - v^T Y)_+ - \iota.$$

We use Monte Carlo sampling to generate N sample pairs $\{(X^1, Y^1), (X^2, Y^2), \ldots, (X^N, Y^N)\}$. The sample average function is denoted as

$$g_N(z, \eta, v) := \frac{1}{N} \sum_{j=1}^{N} G(z, \eta, v, X^j, Y^j). \tag{2.2}$$

The sample average approximation of (RSD) is then stated as

$$\begin{aligned} &\min f(z) && \text{(SASD)} \\ &\text{s.t. } g_N(z, \eta, v) \leq 0, \quad (\eta, v) \in \mathcal{A} \times \widetilde{\mathcal{C}} && (2.3) \\ &\qquad z \in Z. \end{aligned}$$

In order to analyze the convergence of problem (SASD) to its true counterpart, it will be convenient to consider perturbed versions of (2.1) and (2.3). For a given $\epsilon \in \mathbb{R}$, we define the following ϵ-approximation of the feasible regions of (RSD) and (SASD):

$$S^\epsilon := \{z \in Z : g(z, \eta, v) \leq \epsilon, \ (\eta, v) \in \mathcal{A} \times \widetilde{\mathcal{C}}\}, \tag{2.4}$$

$$S_N^\epsilon := \{z \in Z : g_N(z, \eta, v) \leq \epsilon, \ (\eta, v) \in \mathcal{A} \times \widetilde{\mathcal{C}}\}. \tag{2.5}$$

Note that S^0 and S_N^0 are the feasible regions of (RSD) and (SASD) respectively. Let

$$\theta^\epsilon := \min_{z \in S^\epsilon} f(z), \tag{ϵ-RSD}$$

$$\theta_N^\epsilon := \min_{z \in S_N^\epsilon} f(z) \tag{ϵ-SASD}$$

be the optimal values of (ϵ-RSD) and (ϵ-SASD) respectively. Let Υ^ϵ and Υ_N^ϵ be the sets of optimal solutions of (ϵ-RSD) and (ϵ-SASD) respectively. In addition, define

$$\Phi(z, X, Y) := \|H(z, X)\| + \|Y\| + \iota, \tag{2.6}$$

and its expected value and sample average approximation as

$$\phi(z) := E[\Phi(z, X, Y)], \tag{2.7}$$

$$\phi_N(z) := \frac{1}{N}\sum_{j=1}^{N} \Phi(z, X^j, Y^j). \tag{2.8}$$

Note that $\|.\|$ defaults to the standard Euclidean norm in this paper unless otherwise specified. As we shall see later, the function $\Phi(\cdot)$ plays an important role in the analysis, since for all given $(z, X, Y) \in Z \times \Xi$, $\Phi(z, X, Y)$ dominates $G(z, \eta, v, X, Y)$ for any $(\eta, v) \in \mathcal{A} \times \widetilde{\mathcal{C}}$. Thus, we can identify some good behavior of $G(\cdot)$ by verifying the behavior of $\Phi(\cdot)$.

Define the diameter of a nonempty compact set K as

$$D(K) := \max_{y,y' \in K} \|y - y'\|, \tag{2.9}$$

and the distance between a point x and K as

$$d(x, K) := \begin{cases} \inf_{y \in K} \|x - y\| & \text{if } K \text{ is nonempty,} \\ \infty & \text{o.w.} \end{cases} \tag{2.10}$$

The following assumptions are made:

(A1) $Z \subset \mathbb{R}^n$ is a nonempty compact set.

(A2) $H(\cdot, X)$ is Lipschitz continuous on Z a.e. with respect to X, i.e., there exists a function $\Pi : \Xi_X \to \mathbb{R}_+$ such that $\|H(z_1, X) - H(z_2, X)\| \leq \Pi(X)\|z_1 - z_2\|$ a.e. for all $z_1, z_2 \in Z$. Assume that $\Pi(X)$ is an integrable random variable, and define $\pi := E[\Pi(X)]$.

(A3) For all $z \in Z$, the Moment Generation Function (MGF)[2] $M_\Phi^z(\cdot)$ of $\Phi(z, X, Y)$ is finite in a neighborhood of zero.

(A4) The MGF $M_\Pi(\cdot)$ of $\Pi(X)$ is finite in a neighborhood of zero.

Assumptions (A1) and (A2) impose some regularity conditions on the structure of the problem constraints. Assumptions (A3) and (A4) ensure that the random variables in the problem are reasonably well behaved, and hold in particular when Y has bounded support and the partial derivatives of $H(z, x)$ with respect to z have a uniform bound for all $(z, x) \in Z \times \Xi_X$. Assumptions of the form (A1)-(A4), which are required for the analysis, are common in the literature (see Shapiro et al. (2009)).

We now study some features of the integrand $G(\cdot)$ and its expected value function $g(\cdot)$, as well as the sample average function $g_N(\cdot)$. Under Assumptions (A1)-(A4), the analysis that follows shows boundedness, continuity and convergence of these functions. The following two basic propositions, which discuss these properties, provide the foundation for the remaining results in the paper.

Proposition 2.1

[2] The MGF of a random variable W is defined as $M(s) = E[e^{sW}]$.

(1) For all $(\eta, v) \in \mathcal{A} \times \widetilde{\mathcal{C}}$, $G(z,\eta,v,X,Y) \leq \Phi(z,X,Y)$ a.e.. If Assumptions (A1) and (A2) hold, there is an integrable random variable greater than or equal to $\Phi(z,X,Y)$ for all $z \in Z$ a.e., and hence $\phi(\cdot)$ is bounded on Z and $\phi_N(\cdot)$ is bounded on Z a.e..

(2) $G(z,\eta,v,X,Y)$ is Lipschitz continuous in $(\eta,v) \in \mathbb{R}^{m+1}$ a.e. for all $z \in Z$. If Assumptions (A1) and (A2) hold, $G(z,\eta,v,X,Y)$ is Lipschitz continuous in $Z \times \mathcal{A} \times \widetilde{\mathcal{C}}$ a.e..

(3) If Assumption (A3) holds, the MGF of $G(z,\eta,v,X,Y)$ is finite in a neighborhood of zero for all $(z,\eta,v) \in Z \times \mathcal{A} \times \widetilde{\mathcal{C}}$.

Proof:

(1) It is easy to see that

$$\begin{aligned}\sup_{(\eta,v)\in\mathcal{A}\times\widetilde{\mathcal{C}}} |G(z,\eta,v,X,Y)| \leq & \sup_{(\eta,v)\in\mathcal{A}\times\widetilde{\mathcal{C}}} |v^T H(z,X) - v^T Y - \iota| \\ \leq & \sup_{(\eta,v)\in\mathcal{A}\times\widetilde{\mathcal{C}}} \|v\|(\|H(z,X)\| + \|Y\|) + \iota \\ \leq & \Phi(z,X,Y).\end{aligned}$$

Here, the last inequality uses $\|v\| \leq 1$, which follows from the fact that $\|v\|_1 \leq 1$ for all $v \in \widetilde{\mathcal{C}}$.

Next, fix $z_0 \in Z$. Assumption (A2) implies that, for all $z \in Z$,

$$\|H(z,X)\| \leq \|H(z_0,X)\| + \Pi(X)\|z - z_0\| \leq \|H(z_0,X)\| + \Pi(X)D(Z).$$

Note that $D(Z)$ is finite by Assumption (A1). Thus,

$$\Phi(z,X,Y) \leq \|H(z_0,X)\| + \Pi(X)D(Z) + \|Y\| + \iota = \Phi(z_0,X,Y) + \Pi(X)D(Z). \tag{2.11}$$

The random variable $\Phi(z_0,X,Y)$ is integrable since we have assumed that $H(z,X)$ is integrable for all $z \in Z$, and also that Y and $\Pi(X)$ are integrable as well. It follows from (2.11) that, for any $z \in Z$,

$$\phi(z) = E[\Phi(z_0,X,Y)] + E[\Pi(X)]D(Z) < \infty$$

and similarly for $\phi_N(z)$.

(2) Given $z \in Z$, for any $s = (\eta, v),\ s' = (\eta', v') \in \mathbb{R}^{m+1}$, we have

$$\begin{aligned}&|G(z,s,X,Y) - G(z,s',X,Y)| \\ &\qquad \leq |G(z,\eta,v,X,Y) - G(z,\eta',v,X,Y)| + |G(z,\eta',v,X,Y) - G(z,\eta',v',X,Y)|.\end{aligned}$$

Further,

$$\begin{aligned}|G(z,\eta,v,X,Y) - G(z,\eta',v,X,Y)| \leq & |(\eta - v^T H(z,X))_+ - (\eta' - v^T H(z,X))_+| \\ & + |(\eta - v^T Y)_+ - (\eta' - v^T Y)_+| \\ \leq & 2|\eta - \eta'|,\end{aligned} \tag{2.12}$$

and

$$\begin{aligned}|G(z,\eta',v,X,Y)-G(z,\eta',v',X,Y)| &\le |(v-v')^TH(z,X)|+|(v-v')^TY| \\ &\le \|v-v'\|\Phi(z,X,Y). \end{aligned}\tag{2.13}$$

Thus, it follows that

$$|G(z,s,X,Y)-G(z,s',X,Y)| \ \le\ (\Phi(z,X,Y)+2)\|s-s'\|. \tag{2.14}$$

Moveover, for all $t=(z,\eta,v),\ t'=(z',\eta',v')\in Z\times\mathcal{A}\times\widetilde{\mathcal{C}}$, we have

$$\begin{aligned}|G(t,X,Y)-G(t',X,Y)| &\le |G(z,s,X,Y)-G(z,s',X,Y)|+|G(z,s',X,Y)-G(z',s',X,Y)| \\ &\le (\Phi(z,X,Y)+2)\|s-s'\|+|v^T(H(z,X)-H(z',X))| \\ &\le (\Phi(z,X,Y)+\Pi(X)+2)\|t-t'\|. \end{aligned}\tag{2.15}$$

As shown in part (1), under Assumptions (A1) and (A2) $\Phi(z,X,Y)$ is uniformly dominated by an integrable random variable for all $z\in Z$ a.e..

(3) It follows that

$$E\left[e^{sG(z,\eta,v,X,Y)}\right]\ \le\ E\left[e^{|sG(z,\eta,v,X,Y)|}\right]\ \le\ E\left[e^{|s|\Phi(z,X,Y)}\right]\ \le\ M_\Phi^z(|s|).$$

The proof is complete because $M_\Phi^z(\cdot)$ is finite in a neighborhood of zero for all $z\in Z$ by Assumption (A3). □

Proposition 2.2 *Suppose that Assumptions (A1) and (A2) hold. Then,*

(1) The function $g(\cdot)$ is bounded and $g_N(\cdot)$ is bounded a.e..

(2) The function $g(\cdot)$ is Lipschitz continuous on $Z\times A\times\widetilde{\mathcal{C}}$ and $g_N(\cdot)$ is Lipschitz continuous on $Z\times A\times\widetilde{\mathcal{C}}$ a.e..

(3) $g_N(\cdot)$ converges to $g(\cdot)$ a.e. uniformly on $Z\times\mathcal{A}\times\widetilde{\mathcal{C}}$.

Proof:
(1) Proposition 2.1 (1) implies that $|g(z,\eta,v)|\le\phi(z)$ for all $(z,\eta,v)\in Z\times\mathcal{A}\times\widetilde{\mathcal{C}}$. Moreover, the same proposition shows that $\phi(\cdot)$ is bounded. Similarly, we have that $|g_N(z,\eta,v)|\le\phi_N(z)$ and $\phi_N(\cdot)$ is bounded a.e..

(2) Let

$$\Gamma := 2+\max_{z\in Z}\phi(z)\ <\ \infty. \tag{2.16}$$

From the proof of Proposition 2.1 (2), we know that given any $z\in Z$, we have $|g(z,s)-g(z,s')|\le\Gamma\|s-s'\|$ for any $s,\ s'\in\mathbb{R}^{m+1}$. Analogously, $\|g_N(z,s)-g_N(z,s')\|\le(2+\max_{z\in Z}\phi_N(z))\|s-s'\|$. Moreover, that proof also shows that $|g(t)-g(t')|\le(\Gamma+\pi)\|t-t'\|$ for all $t,\ t'\in\mathbb{R}^{m+n+1}$. Let

$$\pi_N := \frac{1}{N}\sum_{j=1}^{N}\Pi(X^j). \tag{2.17}$$

Then $|g_N(t) - g_N(t')| \leq (2 + \pi_N + \max_{z\in Z} \phi_N(z))\|t - t'\|$ for all $t,\ t' \in \mathbb{R}^{m+n+1}$.

(3) Proposition 2.1 shows that $G(z, \eta, v, X, Y)$ is dominated by an integrable function and is continuous on $Z \times \mathcal{A} \times \widetilde{\mathcal{C}}$ a.e.. Assumption (A1) ensures the compactness of $Z \times \mathcal{A} \times \widetilde{\mathcal{C}}$. Under these conditions, the uniform convergence a.e. immediately follows Proposition 7 in Shapiro (2003). □

3 Rate of Convergence Analysis of Sample Average Approximation

In this section we discuss the application of the SAA method to (RSD) and study conditions for convergence. As we shall see below, it is possible to show that (i) the optimal value and the set of optimal solutions of the SAA problem converge to the true values, and (ii) the feasible set of the SAA problem becomes arbitrarily close to the original feasible set with a probability that goes to one exponentially fast with the sample size.

Based on the distance function (2.10) of a point to a set, we denote the deviation of set K_1 from K_2 as

$$\mathbb{D}(K_1, K_2) := \begin{cases} \sup_{y\in K_1} d(y, K_2) & \text{if } K_1 \text{ is nonempty,} \\ 0 & \text{o.w.} \end{cases} \tag{3.1}$$

and the Hausdorff distance between these sets as

$$\mathbb{H}(K_1, K_2) := \max\{\mathbb{D}(K_1, K_2),\ \mathbb{D}(K_2, K_1)\}. \tag{3.2}$$

Let us first discuss the consistency of the estimator θ_N^ϵ of θ^ϵ, which are respectively the optimal value of (ϵ-SASD) and the optimal value of (ϵ-RSD) for a given $\epsilon \in \mathbb{R}$. Vogel (1994) studies the stability of the feasible set of the SAA problem and consistency of the SAA optimal value and solutions for stochastic programming problems for countably many constraints. We now show that the same conclusions apply to (ϵ-SASD) and (ϵ-RSD), which have uncountably many constraints, in Theorem 3.1 below. Note that the theorem does not make any assumptions about convexity. Let us first denote

$$\psi(z) := \sup_{(\eta,v)\in\mathcal{A}\times\widetilde{\mathcal{C}}} g(z, \eta, v), \tag{DCP}$$

$$\psi_N(z) := \sup_{(\eta,v)\in\mathcal{A}\times\widetilde{\mathcal{C}}} g_N(z, \eta, v). \tag{SDCP}$$

Thus, we have $S^\epsilon \equiv \{z \in Z : \psi(z) \leq \epsilon\}$ and $S_N^\epsilon \equiv \{z \in Z : \psi_N(z) \leq \epsilon\}$. Lemma 3.1 shows Lipschitz continuity and uniform convergence of $\psi_N(\cdot)$ and $\psi(\cdot)$.

Lemma 3.1

(1) If Assumption (A2) hold, $\psi(\cdot)$ is Lipschitz continuous on Z and $\psi_N(\cdot)$ is Lipschitz continuous on Z a.e..

(2) If Assumptions (A1) and (A2) hold, $\psi_N(\cdot)$ uniformly converges to $\psi(\cdot)$ on Z a.e..

Proof: If Assumption (A2) holds, for any $z,\ z' \in Z$, we have

$$\begin{aligned}|\psi(z)-\psi(z')| &\leq \sup_{(\eta,v)\in\mathcal{A}\times\widetilde{\mathcal{C}}} |g(z,\eta,v)-g(z',\eta,v)| \\ &\leq E\|H(z,X)-H(z',X)\| \\ &\leq \pi\|z-z'\|,\end{aligned}$$

so $\psi(\cdot)$ is Lipschitz continuous on Z. A similar argument shows the Lipschitz continuity of $\psi_N(\cdot)$. If Assumptions (A1) and (A2) hold, it follows from Proposition 2.2 (3) that $\psi_N(\cdot)$ converges uniformly to $\psi(\cdot)$ on Z a.e.. □

Theorem 3.1 *Suppose Assumptions (A1) and (A2) hold. Fix $\epsilon \in \mathbb{R}$ and suppose also that $\mathbb{D}(S^\epsilon, S^{\epsilon-\gamma}) \to 0$ as $\gamma \downarrow 0$. Then,*

(1) $\mathbb{H}(S_N^\epsilon, S^\epsilon) \to 0$ a.e as $N \to \infty$;

(2) If the objective function $f(\cdot)$ is continuous in a neighborhood of S^ϵ, then $\theta_N^\epsilon \to \theta^\epsilon$ and $\mathbb{D}(\Upsilon_N^\epsilon, \Upsilon^\epsilon) \to 0$ as $N \to \infty$ a.e..

Proof:
(1) Lemma 3.1 (1) shows the continuity of $\psi(\cdot)$ and $\psi_N(\cdot)$ a.e. on Z. By Lemma 3.1 (2), we know that $\psi_N(\cdot)$ converges uniformly to $\psi(\cdot)$ on Z a.e.. It follows from Proposition 7.15 in Rockafellar and Wets (1998) that $\psi_N(\cdot)$ both epi-converges and hypo-converges to $\psi_N(\cdot)$ on Z a.e.. By Theorem 3.1 in Vogel (1994), we have that S_N^ϵ is upper semiconvergent to S^ϵ a.e.. Obviously, $\mathbb{D}(S_N^\epsilon, S^\epsilon) \to 0$ as $N \to \infty$ a.e..

We now show that $\mathbb{D}(S^\epsilon, S_N^\epsilon) \to 0$ a.e.. Denote $(S^\epsilon)^o := \{z \in Z : \psi(z) < \epsilon\}$. It is trivial to prove that $D(S^\epsilon, S^{\epsilon-\gamma}) \to 0$ a.e. as $\gamma \downarrow 0$ if and only if $S^\epsilon \subseteq \mathrm{cl}((S^\epsilon)^o)$. It follows from Theorem 3.5 in Vogel (1994) that S_N^ϵ is lower semiconvergent to S^ϵ a.e.. Thus $\mathbb{D}(S^\epsilon, S_N^\epsilon,) \to 0$ as $N \to \infty$ a.e..

(2) Suppose that S^ϵ is a nonempty set. Υ^ϵ is also nonempty because of the continuity of $f(\cdot)$. Since $\|H(S_N^\epsilon, S^\epsilon)\| \to 0$. a.e., it follows that S_N^ϵ is semiconvergent to S^ϵ a.e. By Theorem 4.1 in Vogel (1994), we have that $\theta_N^\epsilon \to \theta^\epsilon$ a.e. and $D(\Upsilon_N^\epsilon, \Upsilon^\epsilon) \to 0$ a.e. as $N \to \infty$.

If S^ϵ is empty, S_N^ϵ must be empty for large enough N a.e.. Otherwise, $\mathbb{D}(S_N^\epsilon, S^\epsilon) = \infty$ for all $N \geq 1$ a.e. so that $\lim_{N\to\infty} \mathbb{D}(S_N^\epsilon, S^\epsilon) = \infty$ a.e., which contradicts the conclusion from part (1). Therefore, Υ^ϵ and Υ_N^ϵ are empty as well and thus both θ^ϵ and the limit of θ_N^ϵ are ∞. □

The condition $\mathbb{D}(S^\epsilon, S^{\epsilon-\gamma}) \to 0$ as $\gamma \downarrow 0$ ensures the stability of (ϵ-RSD). An arbitrary small perturbation of the function $g(\cdot)$ cannot result in a big change in the optimal solutions. Consider the case, for example, where $\epsilon = -\iota$, $Z = \mathbb{R}$, $H(z,X) = -z^2$ and Y is a Bernoulli random variable which takes values ± 1 with probability 1/2, so we can take $\mathcal{A} = [-1,1]$. Thus, $g(z,\eta) = (\eta + z^2)_+ - 1/2[(\eta-1)_+ + (\eta+1)_+] - \iota$. It is easily verified that $S^{-\iota} = \{z \in \mathbb{R} : z^2 \leq 0\} = \{0\}$. Then, $S^{-\iota-\gamma}$ is empty for all $\gamma > 0$ so that $\mathbb{D}(S^{-\iota}, S^{-\iota-\gamma}) \to \infty$. Suppose that we sample $2k+1$ ($k \in \mathbb{N}$) points consisting of k (-1)'s and $k+1$ (1)'s. Clearly, $S_N^{-\iota} = \{z \in \mathbb{R} : z^2 \leq -1/(2k+1)\}$ is empty under the perturbation.

Proposition 3.1 below shows that the condition $\mathbb{D}(S^\epsilon, S^{\epsilon-\gamma}) \to 0$ is automatically satisfied when (ϵ-RSD) is a convex problem and the Slater condition holds.

Proposition 3.1 *Suppose that (i) Z is convex, (ii) for all $(\eta, v) \in \mathcal{A}\times\widetilde{\mathcal{C}}$ the integrand $G(\cdot, \eta, v, X, Y)$ is convex a.e. with respect to (X, Y), and (iii) the Slater condition holds for (ϵ-RSD). Then $\mathbb{D}(S^\epsilon, S^{\epsilon-\gamma}) \to 0$ as $\gamma \downarrow 0$.*

Proof: Assumption (iii) ensures that $(S^\epsilon)^o$ is non-empty; the convexity property in assumptions (i)-(ii) guarantees that $S^\epsilon \subseteq \mathrm{cl}((S^\epsilon)^o)$. □

Theorem 3.1 shows that the feasible set of (ϵ-SASD) approaches that of (ϵ-RSD) as the sample size goes to infinity. Next, for $\epsilon > 0$, let us consider

$$P(S^{-\epsilon} \subseteq S_N^0 \subseteq S^\epsilon). \tag{3.3}$$

Wang and Ahmed (2008) study the same question as (3.3) for problems with finitely many expected value constraints. They show that this probability converges to 1 exponentially fast as the sample size N increases. We now extend their theorem to our setting of infinitely many constraints in Theorem 3.2 below. The proof follows similar steps to the proof in Wang and Ahmed (2008) and is given in Appendix A for completeness.

Theorem 3.2 *Suppose (A1)-(A4) hold. Define*

$$\sigma^2 := \max\left\{ \mathit{Var}[\Pi(X)],\ \max_{z\in Z} \mathit{Var}[\Phi(z, X, Y)],\ \max_{t\in Z\times\mathcal{A}\times\widetilde{\mathcal{C}}} \mathit{Var}[G(t, X, Y)] \right\}.$$

Then, given $\epsilon > 0$, there exists $\tau \in (0, 2\epsilon)$ such that the following holds for all $N \geq 1$:

$$P(S^{-\epsilon} \subseteq S_N^0 \subseteq S^\epsilon) \ \geq\ 1 - \left(3 + \frac{(\tau + 4\pi)^n D(Z)^n}{\tau^n} + \frac{2D(Z\times\mathcal{A}\times\widetilde{\mathcal{C}})^{m+n+1}}{\gamma^{m+n+1}} \right) e^{\left(-\frac{N\tau^2}{12\sigma^2}\right)},$$

where $\gamma := \frac{2\epsilon-\tau}{4\Gamma+4\pi+5\tau}$. In particular, given $\beta \in [0, 1]$, if

$$N \ \geq\ \frac{12\sigma^2}{\tau^2} \log\left[\frac{1}{\beta} \left(3 + \frac{(\tau + 4\pi)^n D(Z)^n}{\tau^n} + \frac{2D(Z\times\mathcal{A}\times\widetilde{\mathcal{C}})^{m+n+1}}{\gamma^{m+n+1}} \right) \right],$$

then $P(S^{-\epsilon} \subseteq S_N^0 \subseteq S^\epsilon) \geq 1 - \beta$.

Theorem 3.2 shows the exponential convergence (in probability) of feasible regions of (SASD) to that of the true problem. This provides a theoretical foundation to control the probability of constraint violation by properly choosing N. Alternatively, N can be determined by fixing a probability β.

4 Reformulation of Sample Average Approximation

Problem (SASD) has infinitely many constraints (2.3), defined on the uncountable set $\mathcal{A}\times\widetilde{\mathcal{C}}$. Similarly to the proof of Theorem 3.2, we can shrink $\mathcal{A}\times\widetilde{\mathcal{C}}$ to a finite subset K with $\mathbb{D}(\mathcal{A}\times\widetilde{\mathcal{C}}, K) \leq \gamma$. For a given $\epsilon > 0$, we could in principle use the set $\{z \in Z : g_N(z, \eta, v) \leq \epsilon\ (\eta, v) \in K\}$ to approximately represent S_N^ϵ. However, this is impractical since the subset K is hard to build. Hence it is necessary to work out an efficient way to find a particular finite support of $\mathcal{A}\times\widetilde{\mathcal{C}}$ to reformulate the constraints. In what follows we describe such an approach. Throughout this section and the next, the samples (X^j, Y^j), $j = 1, \ldots, N$ are fixed, so the results refer to the corresponding sample path. We also assume that Assumption (A0) holds and that the set $\mathcal{A}$ is defined as in Proposition 1.2.

Proposition 4.1 *For all $\epsilon \in \mathbb{R}$, we can rewrite the set S_N^ϵ as*

$$S_N^\epsilon \;=\; \{z \in Z : g_N(z, v^T Y^i, v) \leq \epsilon,\; i = 1, \ldots, N,\; v \in \widetilde{\mathcal{C}}\}. \tag{4.1}$$

Proof: In fact, given any $v \in \widetilde{\mathcal{C}}$, we can regard the samples, $\{v^T Y^j : j = 1, \ldots, N\}$, as the equally likely outcomes of a discrete random variable. Then the proof immediately follows Proposition 3.2 in Dentcheva and Ruszczyński (2003). □

Consider now the problems

$$\min_{v \in \widetilde{\mathcal{C}}} -g_N(\hat{z}, v^T Y^i, v), \tag{SDCP$_i$}$$

for $i = 1, \ldots, N$. By Proposition 4.1, we can rewrite (SDCP) as a combination of (SDCP$_i$):

$$\psi_N(z) = -\min_{(i,v) \in \{1,\ldots,N\} \times \widetilde{\mathcal{C}}} -g_N(\hat{z}, v^T Y^i, v). \tag{4.2}$$

Let us first consider a special case, (SASD) with the $\mathcal{P}$-dominance constraints, i.e., $\widetilde{\mathcal{C}}$ is a polytope. Homem-de-Mello and Mehrotra (2009) provide an equivalent representation of (4.1) with a finite number of v's. They prove that all the needed v's are components of the vertices of certain polyhedra. For reference, we state that result below, adapted to our context.

Theorem 4.1 *Suppose $\widetilde{\mathcal{C}}$ is a polytope. Let*

$$\mathcal{P}_i := \{(v, y) \in \mathbb{R}^{m+N} : y_j \geq v^T (Y^i - Y^j),\; y_j \geq 0,\; v \in \widetilde{\mathcal{C}},\; j = 1, \ldots, N\},\; i = 1, \ldots, N. \tag{4.3}$$

Then the set S_N^ϵ in (4.1) satisfies

$$S_N^\epsilon = \{z \in Z : g_N(z, {v^{ik}}^T Y^i, v^{ik}) \leq \epsilon,\; i = 1, \ldots, N,\; k = 1, \ldots, \nu_i\}, \tag{4.4}$$

where v^{ik}, $k = 1, \ldots, \nu_i$ are the v-components of the vertex solutions of $\mathcal{P}_i$.

Theorem 4.1 writes S_N^ϵ as a set consisting of a finite number of constraints. Recall that S_N^0 is the feasible region of (SASD). Then, by replacing (2.3) with (4.4) (for $\epsilon = 0$) we obtain the problem

$$\begin{aligned} &\min f(z) && \text{(FSASD)} \\ &\text{s.t. } g_N(z, {v^{ik}}^T Y^i, v^{ik}) \leq 0,\; i = 1, \ldots, N,\; k = 1, \ldots, \nu_i && (4.5) \\ &\qquad z \in Z. \end{aligned}$$

(FSASD) can be reformulated by introducing the intermediate variable r^{ijk} as in Homem-de-Mello and Mehrotra (2009).

$$
\begin{aligned}
&\min\ f(z) && \text{(FullNLp)}\\
&\text{s.t.}\ \sum_{j=1}^{N} r^{ijk} \le \sum_{j=1}^{N} (v^{ik^T} Y^i - v^{ik^T} Y^j)_+ + \iota, \quad i = 1,\ldots,N,\ k = 1,\ldots,\nu_i \\
&\quad r^{ijk} \ge (v^{ik^T} Y^i - v^{ik^T} H(z, X^j)), \quad i,j = 1,\ldots,N,\ k = 1,\ldots,\nu_i && (4.6)\\
&\quad r^{ijk} \ge 0, \quad i,j = 1,\ldots,N,\ k = 1,\ldots,\nu_i.
\end{aligned}
$$

When $H(\cdot, X)$ is a linear function a.e. with respect to X, (FullNLp) is a linearly constrained program.

(FullNLp) is still impractical when the underlying random vectors are high dimensional since the number of vertices in $\mathcal{P}_i$ grows exponentially fast with that dimension. Furthermore, it is not clear, in principle, whether (2.3) can be reduced to finitely many constraints when $\widetilde{\mathcal{C}}$ is not polyhedral. Homem-de-Mello and Mehrotra (2009) suggest a cut-generation approach which solves a sequence of relaxations of (FSASD), over a subset of constraints (4.5). Here, we extend that algorithm to the broader class of $\mathcal{C}$-dominance constrained problems. This is described in the next section.

5 A Cut-Generation Algorithm with Sample Average Approximation

Recall from Proposition 4.1 that the feasible region of (SASD), S_N^0 written in (4.1), contains infinitely many constraints because of $\widetilde{\mathcal{C}}$. In the cut-generation approach, we consider (HSASD$_k$), a sequence of relaxed (SASD) over some finite subsets of $\widetilde{\mathcal{C}}$. (HSASD$_k$) are solved by using the formulations given in (FullNLp). At a solution of $\hat{z}$ of a relaxed problem we study subproblems (SDCP$_i$) for $i = 1, \ldots, N$. Given $\epsilon > 0$, we choose a δ such that $0 < \delta < \epsilon$. Then, let $\sigma = \epsilon - \delta$. If the σ-optimal values[3] of all the (SDCP$_i$), are bigger than or equal to $-\delta$, we stop and declare $\hat{z} \in S_N^\epsilon$, i.e., $\hat{z}$ is a feasible solution of (ϵ-SASD) at which the objective value is in $[\theta_N^\epsilon, \theta_N^0]$. Otherwise, there exists a σ-optimal solution v^σ of (SDCP$_i$) with an objective value less than $-\delta$. Using this solution, we generate a valid cut $g_N(z, v^{\sigma^T} Y^i, v^\sigma) \le 0$ for $\hat{z}$. Algorithm 1 summarizes the procedure.

Note that step 2 of Algorithm 1 involves solving (SDCP$_i$), which is a DC programming problem — i.e., it minimizes difference of two convex polyhedral functions over a closed convex set. As we shall see soon, a σ-optimal solution of (SDCP$_i$) can be found in a finite number of steps.

We discuss the convergence of Algorithm 1 in Theorem 5.1 below. Let (HSASD) be the last (HSASD$_k$) after Algorithm 1 terminates.

Theorem 5.1 *Suppose that Assumptions (A0) and (A1) hold. Suppose also that there exist $z_0 \in Z$ and a constant $M > 0$ such that $\|H(z_0, X)\| \le M$ a.e.. Then, Algorithm 1 converges after generating a finite number of cuts. Let $\widetilde{\theta}_N$ be the optimal value of the main problem (HSASD), upon termination of Algorithm 1. Then $\theta_N^\epsilon \le \widetilde{\theta}_N \le \theta_N^0$.*

[3] A σ-optimal solution is a feasible solution whose objective function value — called a σ-optimal value — is within σ of the true optimal value.

Algorithm 1 A Cut-Generation Algorithm for (SASD)

```
0. Given $\epsilon > 0$, choose $\delta \in (0, \epsilon)$. Let $\sigma = \epsilon - \delta$.
   Let $k = 0$ and choose an arbitrary finite set $\mathcal{V}^0 \subset \mathcal{A} \times \widetilde{\mathcal{C}}$.

1. Find an optimal solution $\hat{z}$ of

       $\min f(z)$                                              (HSASD$_k$)
       s.t. $g_N(z, \eta, v) \leq 0, \ (\eta, v) \in \mathcal{V}^k,$
            $z \in Z,$

   which can be done by solving (FullNLp).

2. Let $\mathcal{V}^{k+1} = \mathcal{V}^k$.
   For $i = 1, \ldots, N$,
       solve the problems (SDCP$_i$), let $v_i^\sigma$ and $\psi_i^\sigma$ be a $\sigma$-optimal solution and objective value;
       if $\psi_i^\sigma < -\delta$, $\mathcal{V}^{k+1} = \mathcal{V}^{k+1} \cup \{({v_i^\sigma}^T Y^i, v_i^\sigma)\}$.

3. If $\mathcal{V}^{k+1} \neq \mathcal{V}^k$, let $k = k + 1$, go to Step 1; otherwise, exit.
```

Proof: By Proposition 2.2 (2), we know that $g_N(z, v^T Y^i, v)$ is uniformly Lipschitz continuous with respect to v for all $z \in Z$ and $i = 1, \ldots, N$ a.e., so that for any $v, \ v' \in R^m$ we have

$$\begin{aligned} & |g_N(z, v^T Y^i, v) - g_N(z, (v')^T Y^i, v')| \\ & \qquad \leq \ 2|v^T Y^i - (v')^T Y^i| + \|v - v'\| \phi_N(z, X, Y) \qquad \text{(from (2.12) and (2.13))} \\ & \qquad \leq \ C\|v - v'\|, \end{aligned} \tag{5.1}$$

where

$$C := \max_{i \in \{1, \ldots, N\}} 2\|Y^i\| + \max_{z \in Z} \phi_N(z, X, Y). \tag{5.2}$$

Assumption (A0) ensures that Y is bounded. Moreover, by Proposition 2.1, $\Phi(z, X, Y)$ is uniformly dominated by the random variable $\Phi(z_0, X, Y)$ for all $z \in Z$. Thus, (A0) and the assumption that $\|H(z_0, X)\|$ is bounded together imply that $\Phi(z_0, X, Y)$ is bounded and hence so is C, i.e., there exists a constant $c > 0$ such that $C \leq c$ a.e..

In the worst case, each iteration of Algorithm 1 generates N cuts by (SDCP$_i$) for all $i = 1, \ldots, N$. It suffices to prove that each (SDCP$_i$) generates a finite number of cuts. Without loss of generality, we assume that a new (η, v), where $\eta = v^T Y^i$, is added into $\mathcal{V}^k$ by (SDCP$_i$) in each iteration. Let $\mathcal{V}^{k-1} = \{(\eta^0, v^0), \ldots, (\eta^{k-1}, v^{k-1})\}$. For each v^j, we denote its neighborhood as $\mathcal{N}(v^j) := \left\{ v \in \widetilde{\mathcal{C}} : \|v - v^j\| \leq \delta / c \right\}$. If Algorithm 1 fails to stop at iteration k, we get a σ-optimal solution v^k of (SDCP$_i$) at which the objective function value is less than $-\delta$. We claim that $v^k \notin \bigcup_{j=0,\ldots,k-1} \mathcal{N}(v^j)$. Suppose by contradiction that $v^k \in \bigcup_{j=0,\ldots,k-1} \mathcal{N}(v^j)$. Then, there exists some v^j, $j \in \{0, \ldots, k-1\}$, such that $\|v^k - v^j\| \leq \delta / c$. It follows that $|g_N(\hat{z}, {v^k}^T Y^i, v^k) - g_N(\hat{z}, {v^j}^T Y^i, v^j)| \leq \delta$. Since $g_N(\hat{z}, {v^j}^T Y^i, v^j) \leq 0$, it follows that $g_N(\hat{z}, {v^k}^T Y^i, v^k) \leq \delta$, which contradicts the fact that the objective function value of (SDCP$_i$) at v^k is less than $-\delta$. Considering each $i = 1, \ldots, N$, Algorithm 1 will generate at most $N \left\lceil \frac{cD(\widetilde{\mathcal{C}})}{\delta} \right\rceil^m$ cuts.

Obviously, $\widetilde{\theta}_N \leq \theta^0_N$ since the feasible set of (HSASD) contains that of (SASD). Now, we show that $\theta^\epsilon_N \leq \widetilde{\theta}_N$. Let ψ^σ_i be the σ-optimal value of (SDCP$_i$). When Algorithm 1 terminates, we have that $\psi^0_i \geq \psi^\sigma_i - \sigma \geq -\delta - \sigma = -\epsilon$ for all $i = 1, \ldots, N$. By Proposition 4.1, the optimal solution $\hat{z}$ of (HSASD) belongs to S^ϵ_N. □

Homem-de-Mello and Mehrotra (2009) propose a branch-and-cut algorithm for minimizing (SDCP$_i$). Although the method is designed for a polyhedral feasible region, we can adapt it to a general convex set. Horst et al. (1995) present an outer approximation method for a compact convex set. Using their idea, we construct a polytope $\widetilde{\mathcal{P}}$ containing $\widetilde{\mathcal{C}}$ and consider the minimization of (SDCP$_i$) over $\widetilde{\mathcal{P}}$ to obtain a optimal solution $\hat{v}$. Consider a neighborhood of $\widetilde{\mathcal{C}}$ defined as $\mathcal{N}_{\widetilde{\mathcal{C}}} := \{v \in \mathbb{R}^m : d(v, \widetilde{\mathcal{C}}) \leq \sigma/c\}$. If $\hat{v} \notin \mathcal{N}_{\widetilde{\mathcal{C}}}$, we cut that point away from $\widetilde{\mathcal{P}}$ and repeat the procedure. As $\widetilde{\mathcal{C}}$ is compact, there exists a unique $\hat{u} \in \widetilde{\mathcal{C}}$ such that

$$\hat{u} = \arg\min_{u \in \widetilde{\mathcal{C}}} \|\hat{v} - u\|. \tag{5.3}$$

Thus

$$(\hat{v} - \hat{u})^T (v - \hat{u}) \leq 0 \tag{5.4}$$

is a valid cut for $\hat{v}$. In case $\hat{v} \in \mathcal{N}_{\widetilde{\mathcal{C}}}$, (5.1) assures that there exists a feasible solution of (SDCP$_i$) whose objective value is within σ of the value of (SDCP$_i$) at $\hat{v}$.

In what follows we use an outer approximation method to extend the branch-and-cut algorithm in Homem-de-Mello and Mehrotra (2009) in order to minimize polyhedral concave functions over general convex sets. We first construct an initial outer simplex containing $\widetilde{\mathcal{C}}$, $\widetilde{\mathcal{P}} = \{v \in \mathbb{R}^m_+ : \|v\|_1 \leq 1\}$.

The branch-and-cut algorithm yields a global optimal solution of (SDCP$_i$) over $\widetilde{\mathcal{P}}$. If it is not a feasible solution of (SDCP$_i$) over $\mathcal{N}_{\widetilde{\mathcal{C}}}$, the cut (5.4) is generated. The procedure is repeated until a feasible global optimal solution $\hat{v}^\sigma_i$ of (SDCP$_i$) over $\mathcal{N}_{\widetilde{\mathcal{C}}}$ is found. If this solution is infeasible for $\widetilde{\mathcal{C}}$, i.e., $\hat{v}^\sigma_i \in \mathcal{N}_{\widetilde{\mathcal{C}}} \setminus \widetilde{\mathcal{C}}$, we project $\hat{v}^\sigma_i$ onto $\widetilde{\mathcal{C}}$, obtaining

$$v^\sigma_i := \arg\min_{u \in \widetilde{\mathcal{C}}} \|\hat{v}^\sigma_i - u\|. \tag{5.5}$$

Let $\hat{\psi}^\sigma_i$ be the σ-optimal value of (SDCP$_i$) at $\hat{v}^\sigma_i$ and ψ^σ_i be the objective value at v^σ_i. By construction, we have $\|v^\sigma_i - \hat{v}^\sigma_i\| \leq \sigma/c$. It follows from (5.1) that

$$0 \;\leq\; \psi^\sigma_i - \hat{\psi}^\sigma_i = g_N(\hat{z}, {v^\sigma_i}^T Y^i, v^\sigma_i) - g_N(\hat{z}, {\hat{v}^\sigma_i}{}^T Y^i, \hat{v}^\sigma_i) \;\leq\; c\|v^\sigma_i - \hat{v}^\sigma_i\| \;\leq\; \sigma.$$

Then, ψ^σ_i is also a σ-optimal value. Note that $\hat{\psi}^\sigma_i \in [\psi^0_i - \sigma, \psi^0_i]$ since it is obtained over a feasible region that contains $\widetilde{\mathcal{C}}$, whereas $\psi^\sigma_i \in [\psi^0_i, \psi^0_i + \sigma]$ since it corresponds to a feasible point in $\widetilde{\mathcal{C}}$. The algorithm, called Algorithm 2, is summarized below.

Recall the proof of Theorem 5.1 depends on the finite convergence of Algorithm 2. We show now that the algorithm stops after finitely many iterations.

Theorem 5.2 *Algorithm 2 terminates after a finite number of iterations.*

Proof: Homem-de-Mello and Mehrotra (2009) show that the branch-and-cut algorithm solving (SDCP$_i$) over $\widetilde{\mathcal{P}}$ stops after a finite number of iterations. Furthermore, step 2 makes no effect on

Algorithm 2 A Branch-and-Cut Algorithm for (SDCP$_i$)

```
0. Construct an initial simplex $\widetilde{\mathcal{P}} \supseteq \widetilde{\mathcal{C}}$.

1. Run the branch-and-cut algorithm in Homem-de-Mello and Mehrotra (2009) and obtain a
   global optimal solution $\hat{v}$ of

       $\min_{v \in \widetilde{\mathcal{P}}} -g_N(\hat{z}, v^T Y^i, v).$

2. Compute $\hat{u}$ in (5.3). If $\|\hat{v} - \hat{u}\| > \sigma/c$, let

       $\widetilde{\mathcal{P}} = \widetilde{\mathcal{P}} \cap \{v \in \mathbb{R}^m : (\hat{v} - \hat{u})^T (v - \hat{u}) \leq 0\},$

   and go to Step 1; otherwise, exit with a $\sigma$-optimal solution $v_i^\sigma = \hat{u}$.
```

the convergence of the branch-and-cut algorithm. Hence, in order to show that the extension keeps a finite convergence, we only need to discuss step 2.

Now suppose that Algorithm 2 fails to exit from step 2 in finitely many iterations. It follows that we repeat the iteration from step 1 to 2 and obtain an infinite sequence $\{v^k : k = 1, 2, \dots\} \subset \widetilde{\mathcal{P}}$, such that $v^k \notin \mathcal{N}_{\widetilde{\mathcal{C}}}$. Note that $\widetilde{\mathcal{P}}$ is compact in the initial construction. From the above discussion on the cuts we see that for any v^i, v^j in the sequence with $j > i$ we have $v^j \in \widetilde{\mathcal{P}} \cap \{v \in \mathbb{R}^m : (v^i - u^i)^T (v - u^i) \leq 0\}$, where u^i is the closest point in $\widetilde{\mathcal{C}}$ to v^i. It follows that $\|v^i - v^j\| \geq \|v^i - u^i\| \geq \sigma/c$. As a result, it is impossible to find a Cauchy subsequence in $\{v^k\}$, which contradicts the compactness of $\widetilde{\mathcal{P}}$. Therefore, Algorithm 2 must exit from step 2 after a finite number of iterations. □

6 Lower and Upper Bounds

The analysis in the previous sections shows that the optimal value and solution of (SASD) are good approximations of their true counterparts of (RSD) with sufficiently large N. In this section, we discuss how to build statistical lower and upper bounds for the true optimal value. First, we use the optimal values of the relaxed and stringent sample problems ($\pm\epsilon$-SASD) for the upper and lower bounds. However, this approach requires calculations of difficult quantities. Next, we consider a practical lower bound by constructing a Lagrangian function for the problem. We solve a technical difficulty that the Lagrangian multiplier of (RSD) is a function on the uncountable set $\mathcal{A} \times \widetilde{\mathcal{C}}$ as (RSD) is a semi-infinite problem. Finally, we propose a practical upper bound. Given a feasible solution of ($-\epsilon$-SASD), we statistically test its feasibility to (RSD). If the solution is satisfied, the corresponding objective value is a reasonable upper bound. Also, to get a tighter bound, we develop a bisection algorithm.

6.1 Theoretical Lower and Upper Bounds

We generate M independent sample groups, $(X_j^1, Y_j^1), \dots, (X_j^N, Y_j^N)$, $j = 1, \dots, M$, each of which consists of N independent and identically distributed (i.i.d.) sample pairs. Let $S_N^\epsilon(j)$ be set (2.5) composed by the jth group of sample pairs. Correspondingly, $\theta_N^\epsilon(j)$ is the optimal value of the

jth (ϵ-SASD), obtained from sample pairs of size N. Recall that θ^0 is the optimal value of the true problem (RSD). Given $\epsilon > 0$, we have that $p^l := P(\theta_N^\epsilon(j) \leq \theta^0) \geq P(S^0 \subseteq S_N^\epsilon(j))$. Under Assumptions (A1)-(A4), Theorem 3.2 yields

$$P(S^0 \subseteq S_N^\epsilon(j)) \ \geq \ p_N := 1 - \left(3 + \frac{(\tau + 4\pi)^n D(Z)^n}{\tau^n} + \frac{D(Z \times \mathcal{A} \times \widetilde{\mathcal{C}})^{m+n+1}}{\gamma^{m+n+1}}\right) e^{\left(-\frac{N\tau^2}{12\sigma^2}\right)}. \quad (6.1)$$

Next, we rearrange $\theta_N^\epsilon(j)$ in nondecreasing order, so $\theta_N^\epsilon(1) \leq \ldots \leq \theta_N^\epsilon(M)$. For a positive integer $L \leq M$, we use the quantity $\theta_N^\epsilon(L)$ as a statistical lower bound for θ^0, as done by Nemirovski and Shapiro (2006) for chance constrained problems. We have

$$\begin{aligned} P(\theta_N^\epsilon(L) > \theta^0) \ &= \sum_{j=0}^{L-1} \begin{pmatrix} M \\ j \end{pmatrix} (p^l)^j (1-p^l)^{M-j} \\ &=: b(L-1; p^l, M) \\ &\leq b(L-1; p_N, M). \end{aligned} \quad (6.2)$$

The last inequality follows from the fact that $P(\theta_N^\epsilon(L) > \theta^0)$ is decreasing in p^l. Similarly, we have that $p^h := P(\theta_N^{-\epsilon}(j) \geq \theta^0) \geq P(S_N^{-\epsilon}(j) \subseteq S^0)$. For $H \leq M$, it follows that

$$P(\theta_N^{-\epsilon}(H) < \theta^0) \ = \ b(M-H; p^h, M) \ \leq \ b(M-H; p_N, M). \quad (6.3)$$

Hence, we use $\theta_N^{-\epsilon}(H)$ as a statistical upper bound. The results are summarized in the following theorem.

Theorem 6.1 *Suppose (A1)-(A4) hold. Given $\epsilon > 0$, $\beta \in (0,1)$ and $N \geq 1$, we can choose positive integers M, H, and L in such a way that*

$$b(\max\{L-1, M-H\}; p_N, M) \leq \beta, \quad (6.4)$$

where $p_N = 1 - \left(3 + \frac{(\tau+4\pi)^n D(Z)^n}{\tau^n} + \frac{D(Z\times\mathcal{A}\times\widetilde{\mathcal{C}})^{m+n+1}}{\gamma^{m+n+1}}\right) e^{\left(-\frac{N\tau^2}{12\sigma^2}\right)}$ by Theorem 3.2. Then with probability at least $1-\beta$, the random quantities $\theta_N^\epsilon(L)$ and $\theta_N^{-\epsilon}(H)$ respectively give a lower and upper bound for the optimal value θ^0.

Proof: We fix $L = 1$ and $H = M$. It is easy to see that $b(L-1; p_N, M) = b(M-H; p_N, M) = (1-p_N)^M \to 0$ as $M \to \infty$. Thus, we can always find M, H, and L such that (6.4) holds. □

Note that the complexity of the sample average approximation problem may grow fast with N. For this reason, we fix N first and then allow the values L, H, and M to change. To get tighter bounds, a larger L and smaller H should be chosen for a small M. Obviously, the answer is the largest L and smallest H satisfying (6.4). On the other hand, if none of L and H satisfying (6.4) exists, we increase M, which makes the left hand side of (6.4) go to 0 by growing to infinity.

6.2 Practical Lower Bound

The quantity p_N defined in (6.1) is very difficult to compute. Thus, in general Theorem 6.1 only gives theoretical bounds. In actual use, we need more efficient bounds, easily obtained but well approximating the true optimal value. Shapiro et al. (2009) propose a method to use the Lagrangian of the expected value constrained problem to obtain the lower bound. Here, we extend the idea to (RSD). As shown in Dentcheva and Ruszczyński (2009), the Lagrangian of (RSD) is

$$\mathcal{L}(z,\mu) := f(z) + \int_{\mathcal{A}\times\widetilde{\mathcal{C}}} g(z,\eta,v)d\mu, \tag{6.5}$$

where μ belongs to the space $\mathrm{rca}(\mathcal{A}\times\widetilde{\mathcal{C}})$ of a regular countably additive measures on $\mathcal{A}\times\widetilde{\mathcal{C}}$ having finite variation. Given any $\widetilde{\mu}\in\mathrm{rca}(\mathcal{A}\times\widetilde{\mathcal{C}})$, we have that

$$\theta^0 \geq \inf_{z\in Z} \mathcal{L}(z,\widetilde{\mu}). \tag{6.6}$$

It follows that the sample average of (6.6) can be used to construct a statistical lower bound. The key in such an approach is to determine $\widetilde{\mu}$ for a tighter lower bound. In principle, we could use the SAA method with an independent sample group to solve

$$\widetilde{\mu} = \arg \max_{\mu\in\mathrm{rca}(\mathcal{A}\times\widetilde{\mathcal{C}})} \min_{z\in Z} \mathcal{L}_N(z,\mu), \tag{6.7}$$

where $\mathcal{L}_N(z,\mu) := f(z) + \int_{\mathcal{A}\times\widetilde{\mathcal{C}}} g_N(z,\eta,v)d\mu$ is the sample average of $\mathcal{L}(z,\mu)$. Clearly, (6.7) is a difficult problem to solve. Here, we will discuss a particular but practical way to choose $\widetilde{\mu}$ and then show the quality of this approach.

By running Algorithm 1 with an initial i.i.d sample group, $(X_0^1,Y_0^1),\ldots,(X_0^{N_l},Y_0^{N_l})$, we obtain a finite set, $\mathcal{V}_{N_l} := \{(\eta^k,v^k)\}\in\mathcal{A}\times\widetilde{\mathcal{C}}$, to generate the constraints of the main problem (HSASD). The corresponding optimal Lagrangian multipliers of that problem are $\widetilde{\lambda}_{N_l}(\eta^k,v^k)$. We can view $\widetilde{\lambda}_{N_l}(\cdot)$ as a measure on $\mathcal{A}\times\widetilde{\mathcal{C}}$ with mass function on $\mathcal{V}_{N_l}$ since $\widetilde{\lambda}_{N_l}(\eta,v)\equiv 0$ for all $(\eta,v)\in(\mathcal{A}\times\widetilde{\mathcal{C}})\setminus\mathcal{V}_{N_l}$. Clearly, $\widetilde{\lambda}_{N_l}\in\mathrm{rca}(\mathcal{A}\times\widetilde{\mathcal{C}})$ a.e.. We construct

$$\varphi_{N_l}^0 := \inf_{z\in Z} \mathcal{L}_{N_l}^0(z,\widetilde{\lambda}_{N_l}), \tag{6.8}$$

where $\mathcal{L}_{N_l}^0(\cdot)$ is the sample average function of $\mathcal{L}(\cdot)$ with the initial group of N_l samples. Let

$$\varphi_N^j := \inf_{z\in Z} \mathcal{L}_N^j(z,\widetilde{\lambda}_{N_l}), \quad j=1,\ldots,M, \tag{6.9}$$

be formulated by using the measure $\widetilde{\lambda}_{N_l}$ with M independently generated groups of samples of size N each.

Proposition 6.1 *Let $\mathcal{F}_0$ denote the σ-algebra generated by the initial sample $(X_0^1,Y_0^1),\ldots,(X_0^{N_l},Y_0^{N_l})$. Then, $E[\varphi_N^j|\mathcal{F}_0]\leq\theta^0$ a.e. for $j=1,\ldots,M$.*

Proof: Since $\widetilde{\lambda}_{N_l}\in\mathrm{rca}(\mathcal{A}\times\widetilde{\mathcal{C}})$ a.e., the proof directly follows (6.6) and the fact that for any $\widetilde{u}\in\mathrm{rca}(\mathcal{A}\times\widetilde{\mathcal{C}})$, $\inf_{z\in Z}\mathcal{L}(z,\widetilde{u})\geq E[\inf_{z\in Z}\mathcal{L}_N(z,\widetilde{u})]$. Therefore, $\theta^0\geq\inf_{z\in Z}\mathcal{L}(z,\widetilde{\lambda}_{N_l})\geq E[\varphi_N^j|\mathcal{F}_0]$ a.e.. □

Using Proposition 6.1, we now build a statistical lower bound by substituting the sample average of φ_N^j for the conditional expectation $E[\varphi_N^j|\mathcal{F}_0]$. Using a similar method to that described in Mak et al. (1999), we compute

$$\overline{\theta}_{N_l,N,M} := \frac{1}{M}\sum_{j=1}^{M}\varphi_N^j \tag{6.10}$$

to estimate the lower bound for θ^0. The variance of $\overline{\theta}_{N_l,N,M}$ is estimated by

$$\overline{\sigma}^2_{N_l,N,M} := \frac{1}{M}\left[\frac{1}{M-1}\sum_{j=1}^{M}(\varphi_N^j - \overline{\theta}_{N_l,N,M})^2\right]. \tag{6.11}$$

In general, the random variables φ_N^j are not normally distributed. However, since the φ_N^j, $j = 1, \ldots, M$ are i.i.d., by taking M sufficiently large we can apply the Central Limit Theorem. This fact supports that

$$L_{N_l,N,M} = \overline{\theta}_{N_l,N,M} - \nu_\alpha\,\overline{\sigma}_{N_l,N,M} \tag{6.12}$$

be used as an approximate $100(1-\alpha)\%$ confidence lower bound for the conditional expectation of $\varphi_N^j(\widetilde{\lambda}_{N_l})$. In (6.12), ν_α denotes the α-critical value of the normal distribution.

We now show the convergence of $L_{N_l,N,M}$ as $N_l, N \to \infty$. One difficulty is that the cardinality of $\mathcal{V}_{N_l}$ may go to infinity. In each iteration, step 2 of Algorithm 1 processes N_l separation problems (SDCP$_i$) in parallel. In the worst case, with a same $v \in \widetilde{\mathcal{C}}$, N_l new constraints are added in the main problem (HSASD$_k$). As a result, the number of nonzero $\widetilde{\lambda}_{N_l}$'s may go to infinity, which complicates the analysis. We make a modification to prevent this case from happening. As illustrated in Figure 1, the set $\Xi_Y \times \widetilde{\mathcal{C}}$ is divided into cubes of diameter $\gamma \le \frac{\delta}{c+2}$, where c is a constant bounding the (random) Lipschitz constant C in (5.2) — we assume here that the conditions of Theorem 5.1 hold. Now in step 2 of Algorithm 1, only one point from each γ-cube is allowed to construct a new constraint. Given two arbitrary points, (Y^1, v^1) and (Y^2, v^2), in a γ-cube, it follows that

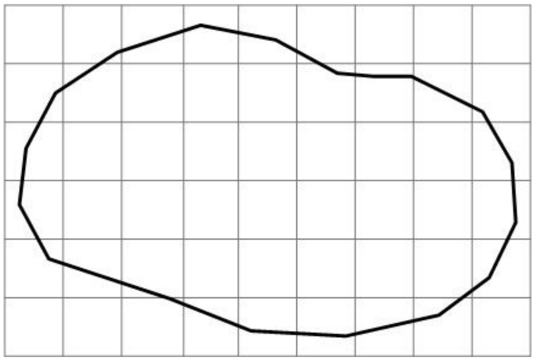

Figure 1: Set $\Xi_Y \times \widetilde{\mathcal{C}}$ Covered by γ-cubes

$$\begin{aligned}|g_N(z, {v^1}^T Y^1, v^1) - g_N(z, {v^2}^T Y^2, v^2)| &\le c\|v^1 - v^2\| + 2\|Y^1 - Y^2\| \\ &\le (c+2)\|(Y^1, v^1) - (Y^2, v^2)\| \\ &\le \delta.\end{aligned}$$

Then the condition $g_N(z, {v^1}^T Y^1, v^1) \le 0$, implies $g_N(z, {v^2}^T Y^2, v^2) \le \delta$, so v^2 is not cut by step 2 of Algorithm 1. By the new policy, $|\mathcal{V}_{N_l}| \le \left\lceil \frac{D(\Xi_Y \times \widetilde{\mathcal{C}})}{\gamma} \right\rceil^{2m}$ for all $N_l > 0$. Furthermore, it is easy to verify that $\theta^{\epsilon}_{N_l} \le \widetilde{\theta}_{N_l} \le \theta^0_{N_l}$ in Theorem 5.1 still holds.

We discuss next the efficiency of the statistical lower bound $L_{N_l,N,M}$ as the sample sizes N_l and N both increase. The following theorem states the limit behavior.

Theorem 6.2 *Suppose that (i) Assumptions (A1) and (A2) hold, (ii) $f(\cdot)$ is finite and convex in a neighborhood of Z, (iii) $G(\cdot, \eta, v, X, Y)$ is convex a.e. (with respect to X, Y) for all $(\eta, v) \in \mathcal{A} \times \widetilde{\mathcal{C}}$, (iv) the Slater condition holds for problem (RSD), and (v) the assumptions of Theorem 5.1 hold. Then*

$$\theta^{\epsilon} \ \le\ \liminf_{N_l \to \infty} \lim_{N \to \infty} L_{N_l,N,M} \ \le\ \limsup_{N_l \to \infty} \lim_{N \to \infty} L_{N_l,N,M} \ \le\ \theta^0 \ \textit{a.e.}.$$

Proof: As $N \to \infty$, we have

$$\varphi^j_N \to \inf_{z \in Z} \mathcal{L}(z, \widetilde{\lambda}_{N_l}) \quad \textit{a.e.}$$

for $j = 0, \ldots, M$ by Proposition 2.2 (3). Note that the assumption that $f(\cdot)$ is convex and $G(\cdot, \eta, v, X, Y)$ is convex for all $(\eta, v, X, Y) \in \mathcal{A} \times \widetilde{\mathcal{C}} \times \Xi$ implies that problems (RSD) and (HSASD) are convex. Moreover, if the Slater condition holds for (RSD) then it must hold for (HSASD) for sufficiently large N_l a.e.. By the strong duality of (HSASD), the optimal value of the dual problem is that of (HSASD), $\widetilde{\theta}_{N_l}$, when Algorithm 1 terminates. By Proposition 3.1, Theorem 3.1, and 5.1, it follows that $\theta^{\epsilon} \le \liminf_{N_l \to \infty} \widetilde{\theta}_{N_l} \le \limsup_{N_l \to \infty} \widetilde{\theta}_{N_l} \le \theta^0$ a.e..

Let $(\widetilde{z}_{N_l}, \widetilde{\lambda}_{N_l})$ be a saddle point of the Lagrangian of (HSASD) (for brevity, we use $\widetilde{\lambda}_{N_l}$ to denote the vector whose components are $\{\widetilde{\lambda}_{N_l}(\eta, v) : (\eta, v) \in \mathcal{V}_{N_l}\}$). We now show that $\widetilde{\lambda}_{N_l}$ is uniformly bounded for large enough N_l a.e.. For a given (η, v), let $g'(z; d, \eta, v)$ be the directional derivative of $g(\cdot)$ along vector d at point z. By Proposition VII.2.2.4 in Hiriart-Urruty and Lemaréchal (1993), the Slater condition implies that, for all $z \in Z$, there exists $d \in \mathbb{R}^n$ such that $g'(z; d, \eta, v) \le -\alpha_1$ for some $\alpha_1 > 0$ and all $(\eta, v) \in \mathcal{A} \times \widetilde{\mathcal{C}}$. Furthermore, convexity of $G(\cdot, \eta, v, X, Y)$ a.e. implies that $\lim_{N_l \to \infty} g'_{N_l}(z; d, \eta, v) = g'(z; d, \eta, v)$ a.e. for all $(z, \eta, v) \in Z \times \mathcal{A} \times \widetilde{\mathcal{C}}$. It follows that there exists $\widetilde{N}_l$ (possibly dependent on the sample path) such that, for all $N_l \ge \widetilde{N}_l$, $g'_{N_l}(z; d, \eta, v) \le -\alpha_2$ for some $\alpha_2 \in (0, \alpha_1]$ a.e. Since $(\widetilde{z}_{N_l}, \widetilde{\lambda}_{N_l})$ is a saddle point, we have

$$f'(\widetilde{z}_{N_l}; d) + \sum_{(\eta,v) \in \mathcal{V}_{N_l}} \widetilde{\lambda}_{N_l}(\eta, v) g'_{N_l}(\widetilde{z}_{N_l}; d, \eta, v) \ \ge\ 0.$$

As each $\widetilde{\lambda}_{N_l}(\eta, v)$ is nonnegative, we obtain

$$\sum_{(\eta,v) \in \mathcal{V}_{N_l}} |\widetilde{\lambda}_{N_l}(\eta, v)| = \sum_{(\eta,v) \in \mathcal{V}_{N_l}} \widetilde{\lambda}_{N_l}(\eta, v) \le \frac{f'(\widetilde{z}_{N_l}; d)}{\alpha_2}.$$

Without loss of generality, assume that $\|d\| = 1$. By assumption (ii), it follows from Theorem IV.3.1.2 in Hiriart-Urruty and Lemaréchal (1993) that there exists $M \ge 0$ such that $|f(z + td) - f(z)| \le Mt$ for all $z \in Z$ when t is sufficiently small. We now let $K = M/\alpha_2$ so that $\widetilde{\lambda}_{N_l}(\eta, v) \le K$

for all $(\eta, v) \in \mathcal{V}_{N_l}$. Thus, we have

$$
\begin{aligned}
& \lim_{N_l \to \infty} \lim_{N \to \infty} \left| \varphi^0_{N_l} - \varphi^j_N \right| \\
= & \lim_{N_l \to \infty} \left| \inf_{z \in Z} \left\{ f(z) + \sum_{(\eta,v) \in \mathcal{V}_{N_l}} \widetilde{\lambda}_{N_l}(\eta, v) g_{N_l}(z, \eta, v) \right\} - \inf_{z \in Z} \left\{ f(z) + \sum_{(\eta,v) \in \mathcal{V}_{N_l}} \widetilde{\lambda}_{N_l}(\eta, v) g(z, \eta, v) \right\} \right| \\
\leq & \lim_{N_l \to \infty} \sup_{z \in Z} \left| \sum_{(\eta,v) \in \mathcal{V}_{N_l}} \widetilde{\lambda}_{N_l}(\eta, v) [g_{N_l}(z, \eta, v) - g(z, \eta, v)] \right| \\
\leq & \lim_{N_l \to \infty} K \left\lceil \frac{D(\Xi_Y \times \widetilde{\mathcal{C}})}{\gamma} \right\rceil^{2m} \sup_{(z,\eta,v) \in Z \times \mathcal{A} \times \widetilde{\mathcal{C}}} |g_{N_l}(z, \eta, v) - g(z, \eta, v)| \\
= & \ 0.
\end{aligned}
$$

It follows that

$$
\theta^\epsilon \ \leq \ \liminf_{N_l \to \infty} \lim_{N \to \infty} \overline{\theta}_{N_l,N,M} \ \leq \ \limsup_{N_l \to \infty} \lim_{N \to \infty} \overline{\theta}_{N_l,N,M} \ \leq \ \theta^0 \text{ a.e.}
$$

and $\overline{\sigma}_{N_l,N,M} \to 0$ as $N \to \infty$ a.e.. □

6.3 Practical Upper Bound

Consider again the perturbed SAA approximation (ϵ-SASD) defined in Section 2. For a given $\tau \leq 0$, we can statistically test the feasibility of a solution of (τ-SASD) to (RSD). If the solution is satisfied, the corresponding objective value of (τ-SASD) is an upper bound to the true optimal value. We now choose a statistical test method.

Recall that $\psi(.)$ and $\psi_N(.)$ are the optimal values of (DCP) and (SDCP) respectively. By using M independent sample groups, the method in Mak et al. (1999) gives an approximate $100(1-\alpha)\%$ confidence upper bound for $\psi(\hat{z})$ where $\hat{z}$ is a solution to (τ-SASD):

$$
U_{N,M}(\hat{z}) := \overline{\psi}_{N,M}(\hat{z}) + \nu_\alpha \, \widetilde{\sigma}_{N,M}(\hat{z}), \tag{6.13}
$$

where

$$
\overline{\psi}_{N,M}(\hat{z}) := \frac{1}{M} \sum_{j=1}^{M} \psi^j_N(\hat{z}),
$$

$$
\widetilde{\sigma}^2_{N,M}(\hat{z}) := \frac{1}{M} \left[\frac{1}{M-1} \sum_{j=1}^{M} (\psi^j_N(\hat{z}) - \overline{\psi}_{N,M}(\hat{z}))^2 \right],
$$

and ν_α denotes the α-critical value of the normal distribution. Note that $\psi^j_N(\cdot)$ corresponds to the optimal value of (SDCP) with the jth sample group ($j = 1, \ldots, M$). Again, the Central Limit Theorem can be applied if M is sufficiently large, since the random variables $\psi^j_N(\hat{z})$ are i.i.d.. If $U_{N,M}(\hat{z}) \leq 0$, we claim that $f(\hat{z})$ is a $100(1-\alpha)\%$ confidence upper bound for (RSD). However, if $U_{N,M}(\hat{z}) > 0$ we cannot make any claims. Therefore, we would like to have a feasible solution of (SASD) which is also most likely feasible to (RSD). We compute a solution of (τ-SASD), called

$\hat{z}_N$, when τ is the smallest (negative) number that keeps (τ-SASD) feasible. Obviously, $\hat{z}_N$ is a minimizer of the function $\psi_N(\cdot)$ in (SDCP) over Z, defined with the same sample for (τ-SASD). We first show that $\hat{z}_N$ increasingly approaches a minimizer of $\psi(\cdot)$ in (DCP) as the sample size increases. Clearly, the minimizers of $\psi(\cdot)$ are the "most likely" feasible solutions of (RSD). We say "most" since (RSD) is an infeasible problem if the minimizers of $\psi(\cdot)$ are not feasible to (RSD). Theorem 6.3 states the result.

Theorem 6.3 *Suppose Assumptions (A1) and (A2) hold. Let $\widetilde{\zeta}$ and $\widetilde{Z}$ be, respectively, the optimal value and the set of the optimal solutions of*

$$\min_{z \in Z} \psi(z). \tag{minDCP}$$

Correspondingly, denote $\widetilde{\zeta}_N$ and $\widetilde{Z}_N$ be the optimal value and the set of the optimal solutions of

$$\min_{z \in Z} \psi_N(z). \tag{minSDCP}$$

If $\widetilde{Z}$ is nonempty and $\widetilde{Z}_N$ is nonempty for N large enough a.e., then $\widetilde{\zeta}_N \to \widetilde{\zeta}$ and $\mathbb{D}(\widetilde{Z}_N, \widetilde{Z}) \to 0$ a.e. as $N \to \infty$.

Proof: By Lemma 3.1, we know that $\psi(\cdot)$ is continuous on Z and $\psi_N(\cdot)$ uniformly converges to $\psi(\cdot)$ on Z a.e. Then we can complete the proof by Theorem 5.3 in Shapiro et al. (2009). □

It is clear that problem (minSDCP) has a very a similar structure to problem (SASD). Thus, we can adapt Algorithm 1 to this case. Algorithm 3 below describes the procedure.

Algorithm 3 A Cut-Generation Algorithm for (minSDCP)

```
0. Given $\epsilon > 0$, choose $\delta \in (0, \epsilon)$. Let $\sigma = \epsilon - \delta$.
   Let $k = 0$ and choose an arbitrary finite set $\mathcal{V}^0 \subset \{(v, j) : v \in \widehat{\mathcal{C}}_j,\ j \in \{1, \ldots, N\}\}$.

1. Find an optimal solution $\widetilde{z}_k$ and optimal value $t_k$ of

       $\min\ t$
       $\text{s.t. } t \geq g_N(z, v^T Y^i, v),\ (v, i) \in \mathcal{V}^k,$
       $z \in Z,$

   which can be done by solving (FullNLp).

2. Let $\mathcal{V}^{k+1} = \mathcal{V}^k$.
   For $i = 1, \ldots, N$,
       solve the problems (SDCP$_i$), let $v^\sigma_{ik}$ and $\psi^\sigma_{ik}$ be respectively a $\sigma$-optimal solution and
       $\sigma$-optimal value;
       if $\psi^\sigma_{ik} < t_k - \delta$, $\mathcal{V}^{k+1} = \mathcal{V}^{k+1} \cup \{(v^\sigma_{ik}, i)\}$.

3. If $\mathcal{V}^{k+1} \neq \mathcal{V}^k$, let $k = k + 1$, go to Step 1; otherwise, exit.
```

Note that if S^0 has an interior point, we have $\widetilde{\zeta} < 0$. Thus, when N is sufficiently large, at least a feasible solution of (RSD) can be found with high probability from Theorem 6.3. This idea suggests an algorithm to build a tighter upper bound at $100(1-\alpha)\%$ confidence level. Starting from a small sample size N_u, we solve (minSDCP), using Algorithm 3. The optimal solution $\widetilde{z}_{N_u}$ is tested in (6.13). If $U_{N,M}(\widetilde{z}_{N_u}) \leq 0$, $f(\widetilde{z}_{N_u})$ is the desired upper bound. Otherwise, we increase N_u by a constant Δ and repeat the procedure until $\widetilde{z}_{N_u}$ is verified to be a feasible solution of (RSD) in probability or N_u reaches a set bound. Afterwards, we may want a tighter bound if $\widetilde{z}_{N_u}$ is statistically feasible to (RSD). An observation is that $\widetilde{z}_{N_u}$ is an optimal solution of ($\widetilde{\zeta}_{N_u}$-SASD) too. We can relax ($\widetilde{\zeta}_{N_u}$-SASD) by solving (τ-SASD) with the same sample for $\widetilde{\zeta}_{N_u} < \tau \leq 0$. Let $\widetilde{z}_\tau$ be an optimal solution of (τ-SASD). Then we test the feasibility of $\widetilde{z}_\tau$ to (RSD). This idea suggests a bisection search method for $\tau \in [\zeta_{N_u}, 0]$; for each τ, we solve (τ-SASD); if the statistical test $U_{N,M}(\widetilde{z}_\tau) \leq 0$, then we increase τ; otherwise, we decrease τ.

Algorithm 4 summarizes the procedure. Note that if Z is convex and $G(\cdot, \eta, v, X, Y)$ is convex a.e. for all $(\eta, v) \in \mathcal{A} \times \widetilde{\mathcal{C}}$, the feasible region of (RSD) is also convex. Consequently, the bisection search method can be simplified without solving (τ_k-SASD) in step 4. Recall that $(\widetilde{z}_{N_l}, \widetilde{\lambda}_{N_l})$ is a saddle point of Lagrangian of (HSASD) which is used to construct the lower bound for the true optimal value. We can improve Algorithm 4 by searching along the line connecting $\widetilde{z}_{N_u}$ and $\widetilde{z}_{N_l}$. This change is summarized in Algorithm 5, which updates step 3 and 4 of Algorithm 4 while keeping the same for the other steps.

Algorithm 4 A Line Search Algorithm for Upper Bound

```
0. Set starting point $N_u$, bound $B > N_u$, and step size $\Delta$.

1. Solve (minSDCP) to obtain the optimal solution $\widetilde{z}_{N_u}$ and optimal value $\widetilde{\zeta}_{N_u}$.
   Evaluate $U_{N,M}(\widetilde{z}_{N_u})$.
   Stop or go to step 3 for a tighter upper bound if $U_{N,M}(\widetilde{z}_{N_u}) \leq 0$.

2. $N_u := N_u + \Delta$.
   Stop if $N_u > B$; otherwise go to step 1.

3. Let $k := 1,\ a_1 := \widetilde{\zeta}_{N_u},\ b_1 := 0$.
   Choose $\epsilon > 0$ and let $n$ be the smallest positive integer such that $(1/2)^n \leq \epsilon/|\widetilde{\zeta}_{N_u}|$.

4. Let $\tau_k := (a_k + b_k)/2$ and compute an optimal solution $\hat{z}_k$ of ($\tau_k$-SASD), defined with the
   same sample used in (minSDCP).
   Evaluate $U_{N,M}(\hat{z}_k)$.
   Go to step 6 if $U_{N,M}(\hat{z}_k) > 0$; otherwise, $f(\hat{z}_k)$ is a wanted upper bound.
   Stop or go to step 5 for a tighter upper bound.

5. Let $a_{k+1} := \tau_k$ and $b_{k+1} := b_k$; go to step 7.

6. Let $a_{k+1} := a_k$ and $b_{k+1} := \tau_k$; go to step 7.

7. If $k = n$, stop; otherwise, replace $k$ by $k + 1$ and go back to step 3.
```

Algorithm 5 A Line Search Algorithm for Upper Bound as (RSD) has a Convex Feasible Region

```
3. Let $k = 1,\ a_1 = \widetilde{z}_{N_u},\ b_1 = \widetilde{z}_{N_l}$.
   Choose $\epsilon > 0$ and let $n$ be the smallest positive integer such that $(1/2)^n \leq \epsilon/\|\widetilde{z}_{N_u} - \widetilde{z}_{N_l}\|$.

4. Let $\tau_k = (a_k + b_k)/2$ and evaluate $U_{N,M}(\tau_k)$.
   Go to step 6 if $U_{N,M}(\tau_k) > 0$; otherwise, $f(\tau_k)$ is a wanted upper bound.
   Stop or go to step 5 for a tighter upper bound.
```

7 Conclusions

We have studied optimization problems with multivariate stochastic dominance constraints. The multivariate aspect of the problem is dealt with by using weighted combinations of the random vectors over a convex set of weights. This concept of stochastic dominance requires that all weighted combinations of multiple random gains depending our decisions be preferable to those of random benchmarks. We can also apply this concept to compare random losses with corresponding benchmarks. If X and Y represent random losses, X is preferred to Y if and only if $-v^T X$ dominates $-v^T Y$ in the second order for all v in a convex subset of $\mathbb{R}^m_+$. The difficulty with the resulting problem is that not only does it have uncountably many expected-value constraints but also calculating the expectations exactly is typically impossible in case of very large or infinite number of scenarios. We have addressed these issues by using the Sample Average Approximation (SAA) method. In the analyses of the approximation, we have discussed four crucial issues: 1) convergence of the approach as the sample size goes to infinity; 2) quality of solutions of the sample problems obtained with finite many samples; 3) derivation of an algorithm to solve the problem; and 4) construction of lower and upper bounds for the true optimal values. Our results provide a practical way to solve the problem that has solid mathematical foundation.

Acknowledgements

We thank two anonymous referees for their constructive comments. This research has been supported by NSF Grant CMMI-0727532 and ONR N00014-09-10518.

References

G. N. Aretoulis, G. P. Kalfakakou, and F. Z. Striagka. Construction material supplier selection under multiple criteria. *Operational Research*, 2009. inprint.

J. Atlason, M. Epelman, and G. Henderson. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research*, 127:333–358, 2004.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Springer-Verlag, New York, NY, 2nd. edition, 1998.

D. Dentcheva and A. Ruszczyński. Optimization with stochastic dominance constraints. *SIAM J. Optim.*, 14(2):548–566, 2003.

D. Dentcheva and A. Ruszczyński. Optimality and duality theory for stochastic optimization problems with nonlinear dominance constraints. *Math. Programming*, 99:329–350, 2004.

D. Dentcheva and A. Ruszczyński. Portfolio optimization with stochastic dominance constraints. *Journal of Banking Finance*, 30:433–451, 2006.

D. Dentcheva and A. Ruszczyński. Optimization with multivariate stochastic dominance constraints. *Math. Programming*, 117:111–127, 2009.

D. Dentcheva, R. Henrion, and A. Ruszczyński. Stability and sensitivity of optimization problems with first order stochastic dominance constraints. *SIAM J. Optim.*, 18(1):322–337, 2007.

D. Drapkin and R. Schultz. An algorithm for stochastic programs with first-order dominance constraints induced by linear recourse. Manuscript, Department of Mathematics, University of Duisburg-Essen, Duisburg, Germany, 2007.

R. Gollmer, U. Gotzes, F. Neise, and R. Schultz. Risk modeling via stochastic dominance in power systems with dispersed generation. Manuscript, Department of Mathematics, University of Duisburg-Essen, Duisburg, Germany, 2007.

J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I.* Springer-Verlag, New York, 1993.

T. Homem-de-Mello and S. Mehrotra. A cutting surface method for linear programs with polyhedral stochastic dominance constraints. *SIAM Journal on Optimization*, 20(3):1250–1273, 2009.

R. Horst, P. M. Pardalos, and N. V. Thoai. *Introduction to Global Optimization.* Kluwer Academic Publishers, Boston, 1995.

J. Hu, T. Homem-de-Mello, and S. Mehrotra. Multi-criterion robust and stochastic dominance-constrained models with application to budget allocation in homeland security. Manuscript, available at `http://www.optimization-online.org/DB_HTML/2010/04/2605.html`, 2010.

N. E. Karoui and A. Meziou. Constrained optimization with respect to stochastic dominance: Application to portfolio insurance. *Mathematical Finance*, 16(1):103–117, 2006.

A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello. The sample average approximation method for stochastic discrete optimiztion. *SIAM Journal on Optimization*, 12:479–502, 2001.

J. Luedtke. New formulations for optimization under stochastic dominance constraints. *SIAM Journal on Optimization*, 19(3):1433–1450, 2008.

W. K. Mak, D. P. Morton, and R. K. Wood. Monte carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24:47–56, 1999.

A. Müller and D. Stoyan. *Comparison Methods for Stochastic Models and Risks.* John Wiley & Sons, Chichester, 2002.

A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM J. Optim.*, 17:969–996, 2006.

Y. Nie, X. Wu, and T. Homem-de-Mello. Optimal path problems with second-order stochastic dominance constraints. Manuscript, Northwestern University, 2009.

M. O'Brien. Techniques for incorporating expected value constraints into stochastic programs. PHD Dissertation, Stanford University, 2000.

T. Prato and G. Herath. Multiple-criteria decision analysis for integrated catchment management. *Ecological Economics*, 63(2-3):627–632, 2007.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis.* Springer, Berlin, 1998.

D. Roman, K. Darby-Dowman, and G. Mitra. Portfolio construction based on stochastic dominance and target return distributions. *Math. Programming*, 108:541–569, 2006.

M. Shaked and J. G. Shanthikumar. *Stochastic Orders and their Applications.* Academic Press, Boston, 1994.

A. Shapiro. Monte Carlo sampling methods. In A. Ruszczynski and A. Shapiro., editors, *Handbook of Stochastic Optimization.* Elsevier Science Publishers B.V., Amsterdam, Netherlands, 2003.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming : modeling and theory.* SIAM, 2009.

S. Vogel. A stochastic approach to stability in stochastic programming. *Journal of Computational and Applied Mathematics*, 56:65–96, 1994.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, NJ, 2nd. edition, 1947.

W. Wang and S. Ahmed. Sample average approximation of expected value constrained stochastic programs. *Operations Research Letters*, 36:515–519, 2008.

K. Yang and J. Trewn. *Multivariate statistical methods in quality management.* McGraw-Hill, 2004.

A The Proof of Theorem 3.2

Before proving the theorem, we state some auxiliary lemmas that will be used in the proof.

Lemma A.1 *Let W and V be two random variables, and let $b \in \mathbb{R}$. Then,*

$$P(W > b - V) \;\le\; P(W > b - a) + P(V > a) \quad \textit{for all } a \in \mathbb{R}.$$

Proof:

$$\begin{aligned} P(W > b - V) &= P(W > b - V,\ V > a) + P(W > b - V,\ V \le a) \\ &\le P(W > b - V,\ V > a) + P(W > b - a) \qquad \text{(since } a + W \ge V + W > b \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{when } W > b - V \text{ and } V \le a) \\ &\le P(V > a) + P(W > b - a). \end{aligned}$$

□

Lemma A.2 *Let W be a random variable such that the MGF of W (denoted $M_W(\cdot)$) is finite in a neighborhood of zero. Let $W_1, W_2, \ldots$ be i.i.d. samples of W, and define $\overline{W}_N := (1/N)\sum_{i=1}^N W_i$. Then, for any $N \ge 1$,*

$$P(\overline{W}_N - E[W] > \delta) \;\le\; e^{-NI_W(E[W]+\delta)}$$

and

$$P(E[W] - \overline{W}_N > \delta) \;\le\; e^{-NI_W(E[W]-\delta)},$$

where $I_W(\cdot)$ is the rate function of W, defined as $I_W(z) = \sup_{\lambda\in\mathbb{R}}\{\lambda z - \log M_W(\lambda)\}$.

Moreover, when δ is sufficiently small we have

$$I_W(E[W] \pm \delta) \;\ge\; \frac{\delta^2}{3\,Var[W]}.$$

Proof: The first assertion is a well-known result, called Chernoff bound; for a proof see, e.g., Dembo and Zeitouni (1998). The second assertion follows from the Taylor expansion of the function $I_W(\cdot)$, see Kleywegt et al. (2001). □

We now prove Theorem 3.2. Let $\epsilon > 0$ be given. For arbitrary $\tau \in (0, \epsilon)$, define the following three quantities:

$$\begin{aligned} P_1^\tau :=& P\left(S^{-\epsilon} \nsubseteq S_N^0,\ \sup_{z \in Z} \phi_N(z) \leq \Gamma - 2 + \tau\right), \\ P_2^\tau :=& P\left(S_N^0 \nsubseteq S^\epsilon,\ \sup_{z \in Z} \phi_N(z) \leq \Gamma - 2 + \tau\right), \\ P_3^\tau :=& P\left(\exists z \in Z,\ \phi_N(z) > \Gamma - 2 + \tau\right), \end{aligned}$$

where $\phi_N(\cdot)$ and Γ are defined in (2.8) and (2.16) respectively. Note that

$$\begin{aligned} & P(S^{-\epsilon} \subseteq S_N^0 \subseteq S^\epsilon) \\ & \quad = 1 - P\left(S^{-\epsilon} \nsubseteq S_N^0 \text{ or } S_N^0 \nsubseteq S^\epsilon\right) \\ & \quad \geq 1 - P\left(S^{-\epsilon} \nsubseteq S_N^0 \text{ or } S_N^0 \nsubseteq S^\epsilon,\ \sup_{z \in Z} \phi_N(z) \leq \Gamma - 2 + \tau\right) - P\left(\sup_{z \in Z} \phi_N(z) > \Gamma - 2 + \tau\right) \\ & \quad \geq 1 - P_1^\tau - P_2^\tau - P_3^\tau. \qquad \text{(A.1)} \end{aligned}$$

We first work with the probability P_1^τ. Compactness of $Z \times \mathcal{A} \times \widetilde{\mathcal{C}}$ implies that, given any $\gamma > 0$, there exists a finite set $K \subseteq Z \times \mathcal{A} \times \widetilde{\mathcal{C}}$ with $|K| \leq D(Z \times \mathcal{A} \times \widetilde{\mathcal{C}})^{m+n+1}/\gamma^{m+n+1}$ such that, for all $t = (z, \eta, v) \in Z \times \mathcal{A} \times \widetilde{\mathcal{C}}$, we have $t' \in K$ satisfying $\|t - t'\| \leq \gamma$. Let t and t' be two such points. From Proposition 2.2 (2), we conclude that

$$|g(t) - g(t')| \ \leq \ (\Gamma + \pi + \tau)\gamma$$

and

$$|g_N(t) - g_N(t')| \ \leq \ (2 + \sup_{z \in Z} \phi_N(z) + \pi_N)\gamma.$$

Moreover, on the event $\{g_N(t) - g(t) > \epsilon,\ \sup_{z \in Z} \phi_N(z) \leq \Gamma - 2 + \tau\}$ we have

$$\begin{aligned} g_N(t') - g(t') \ &= \ g_N(t') - g_N(t) + g_N(t) - g(t) + g(t) - g(t') \\ &> \ \epsilon - (\Gamma + \pi + \tau)\gamma - (2 + \sup_{z \in Z} \phi_N(z) + \pi_N)\gamma \\ &\geq \ \epsilon - (\Gamma + \pi + \tau)\gamma - (\Gamma + \tau + \pi_N)\gamma \\ &= \ \epsilon - (2\Gamma + 2\tau + \pi + \pi_N)\gamma. \qquad \text{(A.2)} \end{aligned}$$

It follows that

$$\begin{aligned}
P_1^\tau &= P(\exists t \in Z \times \mathcal{A} \times \widetilde{\mathcal{C}} \text{ s.t. } g(t) \leq -\epsilon \text{ and } g_N(t) > 0, \sup_{z \in Z} \phi_N(z) \leq \Gamma - 2 + \tau) \\
&\leq P(\exists t \in Z \times \mathcal{A} \times \widetilde{\mathcal{C}} \text{ s.t. } g_N(t) - g(t) > \epsilon, \sup_{z \in Z} \phi_N(z) \leq \Gamma - 2 + \tau) \\
&\leq P(\exists t' \in K \text{ s.t. } g_N(t') - g(t') > \epsilon - (2\Gamma + 2\tau + \pi + \pi_N)\gamma) \quad \text{(from (A.2))}.
\end{aligned}$$

By applying Lemma A.1 to the above expression (with $V = (2\Gamma + 2\tau + \pi + \pi_N)\gamma$ and $a = \epsilon - \tau/2$), this is

$$\begin{aligned}
&\leq P((2\Gamma + 2\tau + 2\pi + \pi_N - \pi)\gamma > \epsilon - \tau/2) + P(\exists t' \in K \text{ s.t. } g_N(t') - g(t') > \tau/2) \\
&= P(\pi_N - \pi > \frac{2\epsilon - \tau}{2\gamma} - 2\Gamma - 2\tau - 2\pi) + P(\exists t' \in K \text{ s.t. } g_N(t') - g(t') > \tau/2) \\
&\leq P(\pi_N - \pi > \tau/2) + P(\exists t' \in K \text{ s.t. } g_N(t') - g(t') > \tau/2),
\end{aligned}$$

where the latter inequality holds since $\gamma = \frac{2\epsilon - \tau}{4\Gamma + 4\pi + 5\tau}$, which implies that $\frac{2\epsilon - \tau}{2\gamma} - 2\Gamma - 2\tau - 2\pi = \tau/2$. Now, using Lemma A.2, this is

$$\leq e^{-NI_\Pi(\pi + \tau/2)} + \sum_{t \in K} e^{-NI_t(g(t) + \tau/2)}, \tag{A.3}$$

where $I_\Pi(\cdot)$ is the rate function of $\Pi(X)$ and $I_t(\cdot)$ is that of $G(t, X, Y)$ at a given $t \in K$. Lemma A.2 also implies, via Assumption (A3), that

$$I_t(g(t) + \tau/2) \geq \frac{\tau^2}{12\text{Var}[G(t, X, Y)]} \geq \frac{\tau^2}{12\sigma^2}$$

for all $t \in K$ if τ is sufficiently small. Also, by Assumption (A4) and the same lemma, we have $I_\Pi(\pi + \tau/2) \geq \tau^2/(12\sigma^2)$. As the result, we obtain

$$P_1^\tau \leq \left(1 + \frac{D(Z \times \mathcal{A} \times \widetilde{\mathcal{C}})^{m+n+1}}{\gamma^{m+n+1}}\right) e^{\left(-\frac{N\tau^2}{12\sigma^2}\right)}. \tag{A.4}$$

A similar calculation yields

$$P_2^\tau \leq \left(1 + \frac{D(Z \times \mathcal{A} \times \widetilde{\mathcal{C}})^{m+n+1}}{\gamma^{m+n+1}}\right) e^{\left(-\frac{N\tau^2}{12\sigma^2}\right)}. \tag{A.5}$$

We now use a similar method to build an upper bound for P_3^τ. Construct a finite set $U \subseteq Z$ with $|U| \leq [(\tau + 4\pi)D(Z)]^n/\tau^n$ such that, for all $z \in Z$, there exists $z' \in U$ satisfying $\|z - z'\| \leq \tau/(\tau + 4\pi)$.

Let z and z' be two such points. Then we have, by Assumption (A2),

$$\left|\phi(z)-\phi(z')\right| \;\le\; E[\|H(z,X)-H(z',X)\|] \;\le\; \frac{\tau\pi}{\tau+4\pi}$$

and

$$\left|\phi_N(z)-\phi_N(z')\right| \;\le\; \frac{\tau\pi_N}{\tau+4\pi},$$

which in turn imply that

$$\begin{aligned}
P_3^\tau \;&\le P(\exists z\in Z \text{ s.t. } \phi_N(z)-\phi(z)\ge\tau)\\
&\le\; P\left(\exists z'\in U \text{ s.t. } \phi_N(z')-\phi(z')\ge\tau-\frac{\tau(\pi+\pi_N)}{\tau+4\pi}\right)\\
&\le\; P\left(\frac{\tau(\pi+\pi_N)}{\tau+4\pi}\ge\tau/2\right)+P\left(\exists z'\in U \text{ s.t. } \phi_N(z')-\phi(z')\ge\tau/2\right) \quad \text{(from Lemma A.1)}\\
&\le\; P(\pi_N-\pi\ge\tau/2)+\sum_{z'\in U}P\left(\phi_N(z')-\phi(z')\ge\tau/2\right)\\
&\le\; \left(1+\frac{(\tau+4\pi)^nD(Z)^n}{\tau^n}\right)e^{\left(-\frac{N\tau^2}{12\sigma^2}\right)}, \qquad\qquad (A.6)
\end{aligned}$$

where the last inequality follows from applying Lemma A.2. Combining (A.1), (A.4), (A.5), and (A.6), we complete the proof for the first part. Also, the second part follows by imposing that $\left(3+\frac{(\tau+4\pi)^nD(Z)^n}{\tau^n}+\frac{2D(Z\times\mathcal{A}\times\widetilde{\mathcal{C}})^{m+n+1}}{\gamma^{m+n+1}}\right)e^{\left(-\frac{N\tau^2}{12\sigma^2}\right)}\le\beta$.