Sequential Convex Approximations to Joint Chance Constrained Programs: A Monte Carlo Approach

L. Jeff Hong
Department of Industrial Engineering and Logistics Management
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

Yi Yang
Department of Computer Science, University of California, Irvine, CA 92617, U.S.A.

Liwei Zhang
Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China

Abstract

When there is parameter uncertainty in the constraints of a convex optimization problem, it is natural to formulate the problem as a joint chance constrained program (JCCP) which requires that all constraints be satisfied simultaneously with a given large probability. In this paper, we propose to solve the JCCP by a sequence of convex approximations. We show that the solutions of the sequence of approximations converge to a Karush-Kuhn-Tuck (KKT) point of the JCCP under a certain asymptotic regime. Furthermore, we propose to use a gradient-based Monte Carlo method to solve the sequence of convex approximations.

1 Introduction

Consider the following optimization problem:

$$\begin{array}{ll} \text{minimize} & h(x) \\ \text{subject to} & c_1(x,\xi) \le 0, \ldots, c_m(x,\xi) \le 0, \\ & x \in X, \end{array} \tag{1}$$

where ξ is a k-dimensional parameter vector, X is a subset of $\Re^d$, $h : \Re^d \to \Re$ and $c_i : \Re^{d+k} \to \Re$, $i = 1, \ldots, m$, are real-valued functions. Furthermore, we assume that $h(x)$ and $c_i(x,\xi)$, $i = 1, \ldots, m$, are convex in x and X is a compact convex set. Then, Problem (1) is a standard constrained convex optimization problem. It has broad applications in communications and networks, product design, system control, statistics and finance, and it can be solved efficiently (e.g., see Boyd and Vandenberghe (2004) for a comprehensive introduction to convex optimization).

In many practical problems, however, the parameter vector ξ of Problem (1) may be uncertain. If this uncertainty is ignored (e.g., by using the expected values of ξ in the optimization), the optimal solution obtained by solving Problem (1) may actually be infeasible with a very high

probability. To illustrate this, we consider a very simple example. We let $\xi = (\xi_1, \ldots, \xi_m)$ where $\xi_1, \ldots, \xi_m$ are m independent observations of a standard normal distribution, let $X = \Re$, $h(x) = x$ and $c_i(x, \xi) = \xi_i - x$ for all $i = 1, \ldots, m$. If we ignore the parameter uncertainty by using $\mathrm{E}(\xi)$ to substitute ξ in Problem (1), the optimal solution is $x^* = 0$. However, the probability of $x^* = 0$ being a feasible solution equals

$$\Pr\{c_1(x^*, \xi) \leq 0, \ldots, c_m(x^*, \xi) \leq 0\} = \Pr\{x^* \geq \xi_1\} \cdots \Pr\{x^* \geq \xi_m\} = 0.5^m,$$

which is very small when m is large (for instance, it is already less than 0.001 when $m = 10$).

To consider this parameter uncertainty, we may formulate the problem as

$$\text{(P)} \qquad \begin{array}{ll} \text{minimize} & h(x) \\ \text{subject to} & \Pr\{c_1(x, \xi) \leq 0, \ldots, c_m(x, \xi) \leq 0\} \geq 1 - \alpha, \\ & x \in X. \end{array}$$

In Problem (P), we require all m uncertain constraints be satisfied simultaneously with a probability at least $1 - \alpha$, where $0 < \alpha < 1$ is often set as 0.01, 0.05 or 0.1. Therefore, the solution to Problem (P) is guaranteed to be a feasible solution to the original Problem (1) with a probability at least $1 - \alpha$. Problem (P) is called a joint chance constrained program (JCCP) and the probabilistic constraint is called a joint chance constraint. When $m = 1$, the constraint is called a single chance constraint because it requires only a single constraint to be satisfied with probability $1 - \alpha$. For simplicity of the notation, we let

$$p(x) = 1 - \Pr\{c_1(x, \xi) \leq 0, \ldots, c_m(x, \xi) \leq 0\}$$

and $p(x)$ is the probability that at least a constraint is violated. Then, the joint chance constraint of Problem (P) becomes $p(x) \leq \alpha$.

Many stochastic optimization problems can be formulated as a JCCP. For instance, the reservoir system design problem of Prékopa et al. (1978) minimizes the total building and penalty costs while satisfying demands for all sites and all periods with a probability at least 80%, and the cash matching problem of Dentcheva et al. (2004) maximizes the value of the portfolio at the end of the planning horizon while covering all scheduled payments with a probability at least 95%. JCCPs were first introduced and studied by Charnes et al. (1958), Miller and Wagner (1965) and Prékopa (1970). Since then, they have been studied extensively in the stochastic optimization literature. For a recent review of the topic, readers are referred to Prékopa (2003).

There are generally two major difficulties in solving a JCCP. First, $p(x)$ may not be a convex (or quasiconvex) function even though $c_1(x, \xi), \ldots, c_m(x, \xi)$ are all convex in x. Therefore, Problem

(P) may not be a convex optimization problem even though Problem (1) is. Then, it is difficult to find a global optimal solution. Second, $p(x)$ generally has no closed form and is typically difficult to evaluate.

Different approaches have been proposed in the stochastic optimization literature to address these difficulties. For the convexity of Problem (P), Prékopa (2003) showed that $p(x)$ is quasiconvex (which implies that Problem (P) is convex) if $c_1(x,\xi),\ldots,c_m(x,\xi)$ are quasiconvex functions of (x,ξ), and if ξ has a logconcave probability distribution, which includes uniform distribution, multivariate normal distribution, and many others. Lagoa et al. (2005) showed that an individual chance constraint in the form of $\Pr\{a^T x \le b\} \ge 1-\alpha$ defines a convex set provided that the vector $(a^T, b)^T$ has a symmetric logconcave density with $\alpha < 1/2$. Henrion (2007) showed that an individual chance constraint in the form of $\Pr\{\xi^T q(x) \le b\} \ge 1-\alpha$ defines a convex set provided that all components of $q(x)$ are nonnegative and convex, $\alpha < 1/2$, and the vector ξ has a elliptically symmetric distribution, whose parameters satisfy certain requirements. Henrion and Strugarek (2008) showed that a joint chance constraint in the form of $\Pr\{g_i(x) \ge \xi_i, i = 1,\ldots,m\} \ge 1-\alpha$ defines a convex set if $g_i(x)$ is $(-r_i)$-concave and $\xi_i, i = 1,\ldots,m$, are independent random variables with (r_i+1)-decreasing densities for some $r_i > 0$ for sufficiently small α values.

When $p(x)$ is not quasiconvex (or at least not verifiable), many convex conservative approximations of $p(x)$ have been proposed, e.g., the quadratic approximation of Ben-Tal and Nemirovski (2000), the conditional value-at-risk (CVaR) approximation of Rockafellar and Uryasev (2000), and the Bernstein approximation of Nemirovski and Shapiro (2006). These approximations typically find feasible but suboptimal solutions to Problem (P). Furthermore, most of these approximations only work on single chance constraints instead of a joint chance constraint. Therefore, one has to approximate the joint chance constraint by a set of individual chance constraints. A popular choice is to use Boole's inequality, which guarantees the satisfaction of the joint chance constraint if $\Pr\{c_i(x,\xi) \ge 0\} \ge 1-\alpha_i, i = 1,\ldots,m$, and $\alpha_1 + \cdots + \alpha_m = \alpha$ (e.g., Nemirovski and Shapiro 2006). However, it makes the solution even more conservative.

To evaluate $p(x)$, Monte Carlo simulations are often used when the closed form of $p(x)$ is not available. When the chance constraint is approximated by functions that are analytically tractable, e.g., the quadratic approximation or the Bernstein approximation, evaluations of these functions are easy. The resulted problems can be solved directly using standard nonlinear optimization algorithms. When the chance constraint is approximated by functions that are not analytically

tractable, e.g., the CVaR approximation, Monte Carlo simulations are also used to evaluate these functions (Rockafellar and Uryasev 2000).

Luedtke and Ahmed (2008) studied the sample-average approximation of the JCCPs. Their goal is to determine the sample size and appropriate probability requirement such that one can find a feasible solution of the original JCCP and also bound the optimality gap. Luedtke et al. (2007) considered linear programs with joint chance constraints and showed that the problems can be reformulated into mixed integer programs. They further demonstrated that the mixed integer programs can be solved efficiently when only the right-hand side vector is random.

Another approach to solving JCCPs is to use the scenario approach, which solves the following problem:

$$\begin{array}{ll} \text{minimize} & h(x) \\ \text{subject to} & c_i(x, \xi_\ell) \leq 0, \quad i = 1, \ldots, m, \ \ell = 1, \ldots, n, \\ & x \in X, \end{array} \tag{2}$$

where $\xi_1, \xi_2 \ldots, \xi_n$ are independent observations of ξ that are often generated from a Monte Carlo simulation. Problem (2) is a convex optimization problem and analytically tractable. The critical issue is how to determine the sample size n to ensure that the joint chance constraint is satisfied with a high probability. Calafiore and Campi (2005 and 2006) and De Farias and Van Roy (2004) studied this issue independently, and Erdgoğan and Iyengar (2006) further extended the earlier results to situations where the distribution of ξ is ambiguous. The scenario approach is simple to understand and easy to implement. However, its also has several drawbacks. First, Problem (2) is also a conservative approximation to the original JCCP; it finds feasible but suboptimal solutions. Second, the solutions found by the scenario approach are not stable. They can be drastically different when different sets of samples are used. Third, the performance of the approach cannot be improved by acquiring more samples of ξ, which is in contrast to many other Monte Carlo algorithms. Indeed, increasing sample size will make Problem (2) more conservative and may lead to worse solutions. Therefore, the performance of the approach cannot be improved when an ample amount of computational time is available.

In this paper, we propose a new approach to solving Problem (P). We first show that the function $p(x)$ can be represented as a limit of a DC function (i.e., difference of two convex functions). Then we use an ε-approximation to approximate Problem (P). We show that, as ε goes down to zero, the optimal solutions (either global optimal or KKT points) of the approximation converge to the optimal solutions of Problem (P), respectively. To solve the ε-approximation problem, we propose

to solve a sequence of convex optimization problems. We show that the sequence of solutions converge to a KKT point of the ε-approximation problem under moderate conditions. We also propose to use a Monte Carlo method to solve the sequence of convex optimization problems. We show that the solution of the Monte Carlo method converges with probability 1 as the sample size goes to infinity, and the sample problem can be solved efficiently using a gradient approach.

Compared to other approaches in the literature, our approach has several advantages. First, it converges to a KKT point of Problem (P), while nearly all other methods are conservative approximations whose solutions do not satisfy any optimality conditions of Problem (P). When the JCCP is convex (even though it may not be verifiable), our approach converges to the global optimal solution while other methods cannot. Second, our approach directly handles the joint chance constraint without breaking it into multiple single chance constraints. Therefore, it avoids the extra conservativeness introduced by using Boole's inequality.

Our approach also has several drawbacks compared to other approaches in the literature. First, it is computationally slow. Generally, Monte Carlo methods are slower than deterministic approximations because the sample problem is either of a large size or the functions need to estimated repeatedly. Our approach requires to solve a sequence of sample problems. Therefore, it can only solve problems with a small or moderate size, e.g., problems with less than 100 dimensions. Second, our approach requires the full joint distribution of the uncertain parameters in order to generate Monte Carlo samples. However, specifying a full joint distribution for a large number of parameters are often difficult in practice. When the distribution of ξ is ambiguous, our approach cannot be applied. To solve JCCPs with ambiguous distributions, many robust optimization approaches have been proposed in the literature, e.g., Ben-Tal and Nemirovski (2000), Bertsimas and Sim (2004) and Chen et al. (2009).

The rest of the paper is organized as follows. We provide a new formulation of the JCCP in Section 2, and show how the formulation can be solved by sequential convex approximations in Section 3. In Section 4, we propose an efficient Monte Carlo algorithm to solve the sequence of convex approximations. The numerical results are reported in Section 5, followed by the conclusions and future research in Section 6. Some lengthy proofs are included in the electronic companion to this paper.

2 A DC Formulation

Let $c(x,\xi) = \max\{c_1(x,\xi),\ldots,c_m(x,\xi)\}$. Note that $c(x,\xi)$ is a convex function of x since $c_i(x,\xi)$, $i = 1,\ldots,m$, are all convex in x. Then,

$$p(x) = 1 - \Pr\{c_1(x,\xi) \le 0,\ldots,c_m(x,\xi) \le 0\} = 1 - \Pr\{c(x,\xi) \le 0\} = \Pr\{c(x,\xi) > 0\}.$$

By doing so, we convert a joint chance constraint to a single chance constraint. We can do this in this paper because we do not need to exploit certain special structures of $c_i(x,\xi)$. For many papers that require $c_i(x,\xi)$ be a linear function of x or ξ and exploit this linearity, e.g., Ben-Tal and Nemirovski (2000) and Nemirovski and Shapiro (2006), $c(x,\xi)$ is no longer linear when $m > 1$ and, thus, handling a joint chance constraint is significantly more difficult than handling a single chance constraint. For us, however, handling a joint chance constraint is as difficult as handling a single chance constraint.

A major difficulty of solving Problem (P) is that $p(x)$ is generally not a convex function of x even though $c(x,\xi)$ is convex in x. In the literature, many algorithms have been proposed to approximate $p(x)$ by a conservative function $\tilde{p}(x)$, i.e., $\tilde{p}(x) \ge p(x)$ for all $x \in X$. Then, the solution to

$$\text{minimize } h(x), \quad \text{subject to } \tilde{p}(x) \le \alpha,\ x \in X$$

is a feasible solution of Problem (P). If $\tilde{p}(x)$ is close to $p(x)$, then the solution is a good approximation to the optimal solution of Problem (P). If $\tilde{p}(x)$ is a convex function, then the approximated problem is a convex program that may be easier to solve.

In this section, we first introduce the CVaR approximation of Rockafellar and Uryasev (2000), which is the "best" convex conservative approximation (Nemirovski and Shapiro 2006). Based on the CVaR approximation, we propose another conservative approximation to $p(x)$, called the DC approximation. We then study the properties of the DC approximation.

2.1 CVaR Approximation

Note that $p(x) = \Pr\{c(x,\xi) > 0\} = \mathrm{E}\left[1_{(0,+\infty)}(c(x,\xi))\right]$, where $1_A(z)$ denotes the indicator function of set A that equals to 1 if $z \in A$ and 0 if $z \notin A$. Since the indicator function $1_{(0,+\infty)}(z)$ is nonconvex (see the left panel of Figure 1), one way to approximate $p(x)$ is to find a convex approximation $\psi(z)$ of $1_{(0,+\infty)}(z)$ such that $\psi(z) \ge 1_{(0,+\infty)}(z)$ for any $z \in \Re$. Then, $\tilde{p}(x) = \mathrm{E}\left[\psi(c(x,\xi))\right]$ is a convex conservative approximation of $p(x)$. For instance, both the CVaR approximation of Rockafellar

and Uryasev (2000) and the Bernstein approximation of Nemirovski and Shapiro (2006) use this approach.

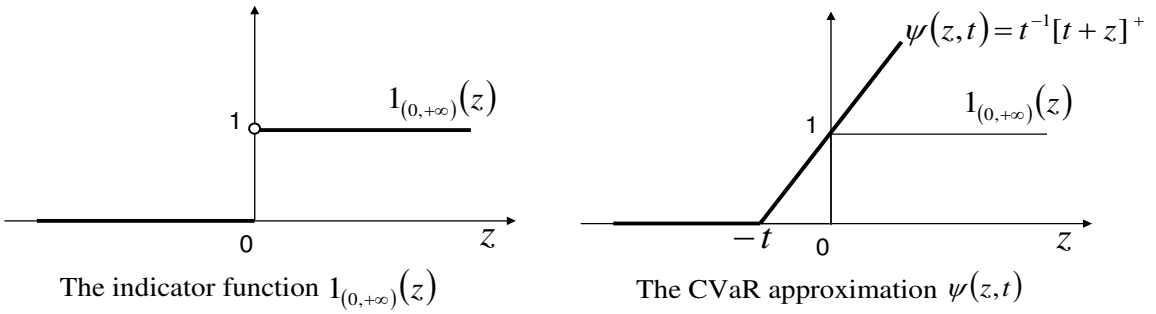

Figure 1: The CVaR approximation to the indicator function $1_{(0,+\infty)}(z)$.

Among all convex conservative approximations of this kind, the CVaR approximation is known to be the "best" (Nemirovski and Shapiro 2006). It uses

$$\psi(z,t)=\frac{1}{t}[t+z]^+$$

to approximate $1_{(0,+\infty)}(z)$, where $t>0$ and $a^+=\max\{a,0\}$ (see the right panel of Figure 1), and let

$$\tilde{p}(x)=\inf_{t>0}\ \mathrm{E}\left[\psi(c(x,\xi),t)\right]=\inf_{t>0}\ \frac{1}{t}\mathrm{E}\left[[t+c(x,\xi)]^+\right].$$

It can be shown that the new constraint $p'(x)\leq\alpha$ is equivalent to $\mathrm{CVaR}_{1-\alpha}(c(x,\xi))\leq 0$ (Nemirovski and Shapiro 2006), where

$$\mathrm{CVaR}_{1-\alpha}(z)=\inf_{\tau\in\Re}\left\{\tau+\frac{1}{\alpha}\mathrm{E}\left[[z-\tau]^+\right]\right\}.$$

This is why this approximation is called the CVaR approximation. The CVaR approximation problem can be solved using a Monte Carlo method. Rockafellar and Uryasev (2000) provided an approach that solves the problem with a single chance constraint and Nemirovski and Shapiro (2006) showed that the Boole's inequality can be used to extend it to JCCPs. Hong and Liu (2009) provided a gradient-based Monte Carlo algorithm that directly solves the CVaR approximations of JCCPs.

2.2 DC Approximation

Although the CVaR approximation may be the "best" convex conservative approximation of $p(x)$, it is clear that $\psi(z,t)$ is not a good approximation to the indicator function $1_{(0,+\infty)}(z)$ from Figure 1. The difference between the two functions grows unboundedly as $z\to+\infty$.

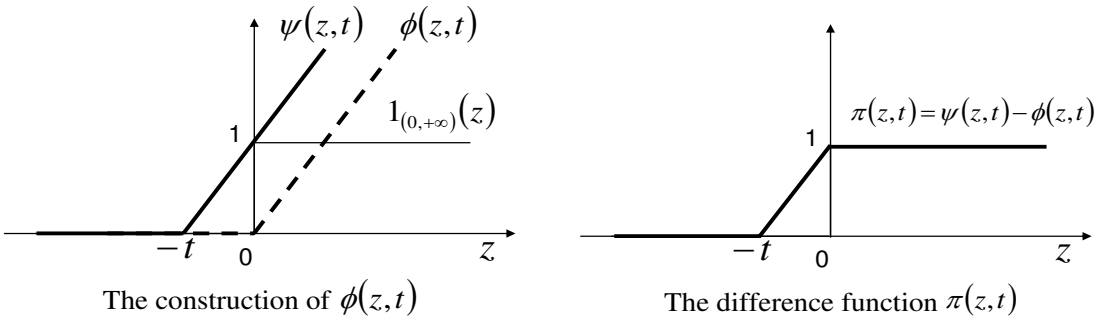

Figure 2: The DC approximation to the indicator function $1_{(0,+\infty)}(z)$.

To find a better approximation to $p(x)$, we first define

$$\phi(z,t) = \frac{1}{t}[z]^+$$

for any $t > 0$. Note that $\phi(z,t)$ can be obtained by shifting $\psi(z,t)$ to the right side by a distance of t (see the left panel of Figure 2). Then, $\pi(z,t) = \psi(z,t) - \phi(z,t)$ is a better approximation of $1_{(0,+\infty)}(z)$ than $\psi(z,t)$ is (see the right panel of Figure 2). Since both $\psi(z,t)$ and $\phi(z,t)$ are convex functions of z, $\pi(z,t)$ is a DC function of z. Furthermore, since $\pi(z,t) \geq 1_{(0,+\infty)}(z)$ for all $z \in \Re$ when $t > 0$, $\pi(z,t)$ is also a conservative approximation of $1_{(0,+\infty)}(z)$ when $t > 0$.

Let $g_1(x,t) = \mathrm{E}[t + c(x,\xi)]^+$ and $g_2(x) = g_1(x,0)$. Note that both functions are convex in x. Let

$$\tilde{p}(x,t) = \mathrm{E}\left[\pi(c(x,\xi),t)\right] = \frac{1}{t}\left[g_1(x,t) - g_2(x)\right]. \tag{3}$$

Then, $\tilde{p}(x,t)$ is a conservative DC approximation of $p(x)$ for any $t > 0$. Let

$$\tilde{p}(x) = \inf_{t>0} \ \tilde{p}(x,t). \tag{4}$$

Then, $\tilde{p}(x)$ is the best conservative approximation among all $\tilde{p}(x,t)$ when $t > 0$. In this paper, we suggest to solve

$$\text{(DC)} \qquad \text{minimize } h(x), \quad \text{subject to } \ \tilde{p}(x) \leq \alpha, \ x \in X.$$

We call Problem (DC) as the DC approximation of Problem (P).

2.3 Equivalence of the JCCP and DC Approximation

In this subsection, we prove that Problem (DC), which is a conservative approximation of Problem (P), is indeed equvialent to Problem (P). We make the following assumptions.

Assumption 1. *The set X is a compact and convex subset of $\Re^d$, and the support of ξ, denoted as Ξ, is a closed subset of $\Re^k$. For any $\xi \in \Xi$, $h(x)$ and $c_i(x,\xi)$, $i = 1,\ldots,m$, are continuously differentiable and convex in x for any $x \in \mathcal{O}$ where $\mathcal{O}$ is a bounded open set such that $X \subset \mathcal{O}$.*

Assumption 1 is used to clearly define Problem (P).

Assumption 2. *There exists a random variable K_i with $\mathrm{E}(K_i) < +\infty$ such that*

$$|c_i(x_1,\xi) - c_i(x_2,\xi)| \leq K_i \|x_1 - x_2\|$$

for any $x_1, x_2 \in \mathcal{O}$ and any $i = 1, 2, \ldots, m$.

Let $K = \sum_{i=1}^m K_i$. Assumption 2 implies that $|c(x_1,\xi) - c(x_2,\xi)| \leq K \|x_1 - x_2\|$ and $\mathrm{E}(K) < +\infty$. The Lipschitz continuity of $c(x,\xi)$ is critical in the analysis of differentiability of $\mathrm{E}[c(x,\xi)]$. It is a common assumption used to handle the differentiability of an expectation (e.g., Broadie and Glasserman (1996), Hong (2009) and Hong and Liu (2009)).

Assumption 3. *For any $x \in \mathcal{O}$, $c(x,\xi)$ is differentiable with respect to x w.p.1.*

To verify Assumption 3, we only consider the situation where ξ is a continuous random vector, because the following Assumption 4 is typically violated if ξ is discrete. By Assumption 1, $c_i(x,\xi)$ is continuously differentiable for all $i = 1,\ldots,m$. Therefore, if $\Pr\{c_i(x,\xi) = c_j(x,\xi)\} = 0$ for any $x \in \mathcal{O}$ and any $i, j = 1,\ldots,m$ with $i \neq j$, $c(x,\xi)$ is differentiable with respect to x w.p.1. In the electronic companion to this paper, we discuss how to verify and satisfy Assumption 3 if $\Pr\{c_i(x,\xi) = c_j(x,\xi)\} \neq 0$.

Let $F(t,x) = \Pr\{c(x,\xi) \leq t\}$ be the cumulative distribution function of $c(x,\xi)$. We make the following assumption on the continuity of $F(t,x)$.

Assumption 4. *There exists a certain $\delta > 0$ such that $F(t,x)$ is continuously differentiable when $(t,x) \in (-\delta,+\delta) \times \mathcal{O}$.*

Since $p(x) = 1 - F(0,x)$, Assumption 4 implies that $p(x)$ is continuously differentiable. Furthermore, note that $\partial_t F(t,x)$ is the density function of $c(x,\xi)$. Therefore, Assumption 4 implies that $c(x,\xi)$ has a continuous density in $(-\delta,+\delta)$ for any $x \in \mathcal{O}$.

Assumption 5. *Let $\Omega_0 = \{x \in X : p(x) \leq \alpha\}$ and $\Omega_0^I = \{x \in X : p(x) < \alpha\}$. Then $\Omega_0 = cl\, \Omega_0^I$.*

Note that Ω_0 is the set of feasible solutions to Problem (P). Assumption 5 is an assumption on the constraint qualification of Problem (P). When Ω_0 is a convex set, it is implied by the widely used Slater's condition (e.g., Boyd and Vandenberghe 2004). Assumption 5 is a commonly used condition in nonlinear programming, especially for numerical methods that approximate the optimal solutions by sequences of points in Ω^I, e.g., the barrier function method (see, for instance, Zangwill (1969) and Bazaraa et al. (1993)). We need this assumption because the method we propose in Section 3 is also such a numerical method. More discussions on Assumption 5 along with a counter-example are presented in the electronic companion to this paper.

Then, we have the following lemmas that are used repeatedly in the rest of the paper.

Lemma 1. *Suppose that Assumption 4 is satisfied. For any $x \in X$, $\tilde{p}(x,t)$ is nondecreasing in t when $t > 0$.*

Proof. For any $t > 0$ and any $z \in \Re$,

$$\pi(z,t) = \psi(z,t) - \phi(z,t) = \left[1 + \frac{1}{t}z\right] \cdot 1_{(-t,0]}(z) + 1_{(0,+\infty)}(z). \tag{5}$$

For any $t_1 > t_2 > 0$ and any $z \in \Re$,

$$\pi(z,t_1) - \pi(z,t_2) = \left[1 + \frac{1}{t_1}z\right] \cdot 1_{(-t_1,-t_2]}(z) + \left[\frac{1}{t_1} - \frac{1}{t_2}\right] z \cdot 1_{(-t_2,0]}(z) \geq 0.$$

Therefore, $\pi(z,t)$ is nondecreasing in t when $t > 0$. Since $\tilde{p}(x,t) = \mathrm{E}\left[\pi(c(x,\xi),t)\right]$, then $\tilde{p}(x,t)$ is also nondecreasing in t when $t > 0$. □

Lemma 2. *Suppose that Assumptions 1 to 4 are satisfied. Then, $g_1(x,t)$ is differentiable in $\mathcal{O} \times (-\delta,\delta)$, and*

$$\begin{aligned} \nabla_x g_1(x,t) &= \mathrm{E}\left[\nabla_x c_{i^*}(x,\xi) \cdot 1_{(-t,+\infty)}\left(c(x,\xi)\right)\right], \\ \frac{\partial}{\partial t} g_1(x,t) &= 1 - F(-t,x), \end{aligned}$$

where $i^* = argmax_{i=1,\ldots,m}\{c_i(x,\xi)\}$.

Proof. By Assumptions 1 to 3, $c(x,\xi) = \max_{i=1,\ldots,m} c_i(x,\xi)$ is differentiable with respect to x w.p.1 and $\nabla_x c(x,\xi) = \nabla_x c_{i^*}(x,\xi)$ w.p.1 when $x \in \mathcal{O}$. Note that $g_1(x,t) = \mathrm{E}\left\{[t + c(x,\xi)]^+\right\}$. Since $f(x) = (t+x)^+$ is differentiable except at $x = -t$ and $f'(x) = 1_{(-t,+\infty)}(x)$ when $x \neq -t$, and

$\Pr\{c(x,\xi) = t\} = 0$ when $t \in (-\delta, +\delta)$ by Assumption 4, then by Proposition 1 of Broadie and Glasserman (1996), $g_1(x,t)$ is differentiable in $\mathcal{O} \times (-\delta, +\delta)$ and

$$\begin{aligned} \nabla_x g_1(x,t) &= \mathrm{E}\left[\nabla_x c_{i^*}(x,\xi) \cdot 1_{(-t,+\infty)}\left(c(x,\xi)\right)\right], \\ \frac{\partial}{\partial t} g_1(x,t) &= \mathrm{E}\left[1_{(-t,+\infty)}\left(c(x,\xi)\right)\right] = \Pr\{c(x,\xi) > -t\} = 1 - F(-t,x). \end{aligned}$$

This concludes the proof of the lemma. □

Remark 1. *Since $g_2(x) = g_1(x,0)$, Lemma 2 also implies that $g_2(x)$ is differentiable and $\nabla g_2(x) = \mathrm{E}\left[\nabla_x c_{i^*}(x,\xi) \cdot 1_{(0,+\infty)}\left(c(x,\xi)\right)\right]$.*

Then, we have the following theorem on the equivalence of Problem (DC) and Problem (P).

Theorem 1. *Suppose that Assumptions 1 to 4 are satisfied. Then, Problem (DC) is equivalent to Problem (P).*

Proof. By Lemma 1, $\tilde{p}(x) = \inf_{t>0} \tilde{p}(x,t) = \lim_{t \searrow 0} \tilde{p}(x,t)$, where $t \searrow 0$ denotes that t decreasingly goes to 0. By Lemma 2,

$$\lim_{t \searrow 0} \tilde{p}(x,t) = \lim_{t \searrow 0} \frac{1}{t}\left[g_1(x,t) - g_1(x,0)\right] = \frac{\partial}{\partial t} g_1(x,0) = 1 - F(0,x) = p(x).$$

Then, $\tilde{p}(x) = p(x)$. Therefore, Problem (DC) is equivalent to Problem (P). This concludes the proof of the theorem. □

Theorem 1 is an important result of this paper. It shows that solving Problem (DC) is equivalent to solving Problem (P). In the rest of this paper, we study how to solve Problem (DC).

2.4 ε-Approximation

Note that $\tilde{p}(x) = \lim_{t \searrow 0} \tilde{p}(x,t)$ by Lemma 1. However, $\tilde{p}(x,t)$ is not well defined at $t = 0$. Therefore, we approximate $\tilde{p}(x)$ by $\tilde{p}(x,\varepsilon) = \frac{1}{\varepsilon}\left[g_1(x,\varepsilon) - g_2(x)\right]$ for a small $\varepsilon \in (0,\delta)$ where δ is defined in Assumption 4, and approximate Problem (DC) by

$$(\mathrm{P}_\varepsilon) \qquad \begin{array}{ll} \text{minimize} & h(x) \\ \text{subject to} & g_1(x,\varepsilon) - g_2(x) \le \varepsilon\alpha, \\ & x \in X. \end{array}$$

By Theorem 1, Problem (DC) is equivalent to Problem (P). Therefore, Problem (P_ε) is also an approximation to Problem (P). In the rest of this subsection, we show that Problem (P_ε) is a good approximation to Problem (P).

Let $\Omega(\varepsilon) = \{x \in X : g_1(x,\varepsilon) - g_2(x) \leq \varepsilon\alpha\}$ denote the feasible set of Problem (P_ε). Then, we have the following lemma on the relationship between $\Omega(\varepsilon)$ and Ω_0, which is the feasible set of Problem (P).

Lemma 3. *Suppose that Assumptions 1 to 5 are satisfied. Then,* $\lim_{\varepsilon \searrow 0} \Omega(\varepsilon) = \Omega_0$.

Proof. By Lemma 1, $\tilde{p}(x,t)$ is nondecreasing with respect to t. Then, for any $\varepsilon_2 \geq \varepsilon_1 > 0$,

$$\frac{1}{\varepsilon_1}\left[g_1(x,\varepsilon_1) - g_2(x)\right] \leq \frac{1}{\varepsilon_2}\left[g_1(x,\varepsilon_2) - g_2(x)\right],$$

which in turn implies that $\Omega(\varepsilon_2) \subset \Omega(\varepsilon_1)$. Therefore, it follows from Exercise 4.3 of Rockafellar and Wets (1998) that $\lim_{\varepsilon \searrow 0} \Omega(\varepsilon)$ exists.

We first prove that $\lim_{\varepsilon \searrow 0} \Omega(\varepsilon) \subset \Omega_0$. For any $x \in \lim_{\varepsilon \searrow 0} \Omega(\varepsilon)$, there exists $\varepsilon_k \searrow 0$ and $x^k \in \Omega(\varepsilon_k)$ such that $x^k \to x$. Since $x^k \in \Omega(\varepsilon_k)$, then $x^k \in X$ and

$$\frac{1}{\varepsilon_k}[g_1(x^k,\varepsilon_k) - g_2(x^k)] \leq \alpha. \tag{6}$$

By Taylor expansion and Lemma 2, we have

$$g_1(x^k,\varepsilon_k) = g_1(x^k,0) + \frac{\partial}{\partial t} g_1(x^k,\tilde{\varepsilon}_k)\varepsilon_k = g_1(x^k,0) + \left[1 - F(-\tilde{\varepsilon}_k, x)\right]\varepsilon_k \tag{7}$$

for some $\tilde{\varepsilon}_k \in (0,\varepsilon_k)$. Combining Equations (6) and (7) and taking $k \to +\infty$, we have $x \in X$ and $1 - F(0,x) \leq \alpha$, which is equivalent to $p(x) \leq \alpha$. Therefore, $x \in \Omega_0$, which implies that $\lim_{\varepsilon \searrow 0} \Omega(\varepsilon) \subset \Omega_0$.

We then prove that $\lim_{\varepsilon \searrow 0} \Omega(\varepsilon) \supset \Omega_0$. For any $x \in \Omega_0^I$, since $p(x) = \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon}\left[g_1(x,\varepsilon) - g_2(x)\right]$ and $p(x) < \alpha$, then $\frac{1}{\varepsilon}[g_1(x,\varepsilon) - g_2(x)] < \alpha$ for some $\varepsilon > 0$ small enough. Therefore, $x \in \Omega(\varepsilon)$. So we obtain that $\liminf_{\varepsilon \searrow 0} \Omega(\varepsilon) \supset \Omega_0^I$. Since $\Omega(\varepsilon)$ is a closed set for any $\varepsilon > 0$, then $\lim_{\varepsilon \searrow 0} \Omega(\varepsilon)$ is also a closed set. Then, by Assumption 5, $\lim_{\varepsilon \searrow 0} \Omega(\varepsilon) \supset \Omega_0$.

Therefore, $\lim_{\varepsilon \searrow 0} \Omega(\varepsilon) = \Omega_0$. This concludes the proof of the lemma. □

For sets $A, B \subset \Re^d$, let $\mathrm{dist}(x,A) = \inf_{x' \in A} \|x - x'\|$ denote the distance from $x \in \Re^d$ to A, and

$$\mathbb{D}(A,B) = \sup_{x \in A} \mathrm{dist}(x,B)$$

denote the deviation of the set A from the set B (Shapiro and Ruszczyński 2008). Let $S(\varepsilon)$ and $\nu(\varepsilon)$ be the set of optimal solutions and the optimal value of Problem (P_ε), S_0 and ν_0 be the set of optimal solutions and the optimal value of Problem (P). Then, we have the following theorem.

Theorem 2. *Suppose that Assumptions 1 to 5 are satisfied. Then,* $\lim_{\varepsilon \searrow 0} \mathbb{D}(S(\varepsilon), S_0) = 0$ *and* $\lim_{\varepsilon \searrow 0} \nu(\varepsilon) = \nu_0$.

Proof. Let $\overline{h}(x) = h(x) + I_{\Omega_0}(x)$ and $\overline{h}_\varepsilon(x) = h(x) + I_{\Omega(\varepsilon)}(x)$, where $I_A(x) = 0$ if $x \in A$ and $I_A(x) = +\infty$ if $x \notin A$. By Lemma 3, $\lim_{\varepsilon \searrow 0} \Omega(\varepsilon) = \Omega_0$. Then, by Proposition 7.4(f) of Rockafellar and Wets (1998), we have that $I_{\Omega(\varepsilon)}(\cdot)$ epi-converges to $I_{\Omega_0}(\cdot)$ as $\varepsilon \searrow 0$. Since $h(\cdot)$ is continuous, we have that $\overline{h}_\varepsilon(\cdot)$ epi-converges to $\overline{h}(\cdot)$ as $\varepsilon \searrow 0$. As $\Omega(\varepsilon)$ and Ω_0 are compact, we have that $\overline{h}_\varepsilon(\cdot)$ and $\overline{h}(\cdot)$ are lower semi-continuous and proper[1]. Then, by Theorem 7.33 of Rockafellar and Wets (1998), we have $\nu(\varepsilon) \to \nu_0$ and

$$\limsup_{\varepsilon \searrow 0} S(\varepsilon) \subset S_0. \tag{8}$$

Since $S(\varepsilon)$ and S_0 are subsets of the compact set X, they are uniformly compact. By the discussions in Example 4.13 of Rockafellar and Wets (1998), we have that Equation (8) implies $\lim_{\varepsilon \searrow 0} \mathbb{D}(S(\varepsilon), S_0) = 0$. This concludes the proof of the theorem. □

Theorem 2 shows that the optimal solutions of Problem (P_ε) provide good approximations to the optimal solutions of Problem (P) when ε is close enough to 0.

However, Problem (P_ε) is generally a non-convex problem. Therefore, finding an optimal solution to the problem may be difficult. We often only find KKT points of Problem (P_ε) (as the method introduced in Section 3 does). In the rest of this subsection, we analyze the convergence of the KKT points of Problem (P_ε) to the KKT points of Problem (P) as $\varepsilon \searrow 0$.

Let Λ_0 and $\Lambda(\varepsilon)$ denote the sets of KKT pairs of Problems (P) and (P_ε), respectively, namely

$$\Lambda_0 = \left\{ (x, \lambda) \in \Omega_0 \times \Re_+ : \begin{array}{c} 0 \in \nabla h(x) + \lambda \, \nabla p(x) + N_X(x), \\ \lambda \, [p(x) - \alpha] = 0 \end{array} \right\},$$

and

$$\Lambda(\varepsilon) = \left\{ (x, \lambda) \in \Omega(\varepsilon) \times \Re_+ : \begin{array}{c} 0 \in \nabla h(x) + \lambda \left[\dfrac{\nabla_x g_1(x, \varepsilon) - \nabla g_2(x)}{\varepsilon} \right] + N_X(x), \\ \lambda \left[\dfrac{g_1(x, \varepsilon) - g_2(x)}{\varepsilon} - \alpha \right] = 0 \end{array} \right\},$$

where $N_X(x)$ denotes the normal cone to X at x (Bonnans and Shapiro 2000), and the differentiability of $p(x)$, $g_1(x, \varepsilon)$ and $g_2(x)$ is ensured by Assumption 4 and Lemma 2. Then, we have the following theorem that shows the relation between Λ_0 and $\Lambda(\varepsilon)$.

[1] A function $f : \Re^d \to \Re \cup \{\pm\infty\}$ is lower semi-continuous at $x_0 \in \Re^d$ if $f(x_0) \leq \liminf_{x \to x_0} f(x)$. It is lower semi-continuous if it is lower semi-continuous at every $x \in \Re^d$. It is a proper function if $f(x) > -\infty$ for every $x \in \Re^d$ and there is at least one point $x \in \Re^d$ such that $f(x) < +\infty$.

Theorem 3. *Suppose that Assumptions 1 to 5 are satisfied. Then,* $\limsup_{\varepsilon \searrow 0} \Lambda(\varepsilon) \subset \Lambda_0$.

Proof. For any $(x, \lambda) \in \limsup_{\varepsilon \searrow 0} \Lambda(\varepsilon)$, there exists $(x_k, \lambda_k) \in \Lambda(\varepsilon_k)$ such that $(x_k, \lambda_k) \to (x, \lambda)$. The inclusion $(x_k, \lambda_k) \in \Lambda(\varepsilon_k)$ means

$$0 \in \nabla h(x_k) + \lambda_k \left[\frac{\nabla_x g_1(x_k, \varepsilon_k) - \nabla g_2(x_k)}{\varepsilon_k} \right] + N_X(x_k), \tag{9}$$

$$\lambda_k \left[\frac{g_1(x_k, \varepsilon_k) - g_2(x_k)}{\varepsilon_k} - \alpha \right] = 0, \ \lambda_k \geq 0. \tag{10}$$

By Lemma 2, $\frac{\partial}{\partial t} g_1(x, t) = 1 - F(-t, x)$ when $x \in X$ and $t \in (-\delta, \delta)$. Since $F(t, x)$ is continuously differentiable by Assumption 4, then $\nabla_x \frac{\partial}{\partial t} g_1(x, t) = -\nabla_x F(-t, x)$ is continuous in t and x. Then, by Exercise 6.24 of Marsden and Hoffman (1993),

$$\frac{\partial}{\partial t} \nabla_x g_1(x, t) = \nabla_x \frac{\partial}{\partial t} g_1(x, t) = -\nabla_x F(-t, x).$$

Since $g_2(x) = g_1(x, 0)$, by the mean-value theorem, we have

$$\frac{\nabla_x g_1(x_k, \varepsilon_k) - \nabla g_2(x_k)}{\varepsilon_k} = \frac{\partial}{\partial t} \nabla_x g_1(x, \tilde{\varepsilon}_k) = -\nabla_x F(-\tilde{\varepsilon}_k, x)$$

for some $\tilde{\varepsilon}_k \in (0, \varepsilon_k)$, $k = 1, 2, \cdots$. Since $F(t, x)$ is continuously differentiable by Assumption 4, then we have that

$$\lim_{k \to +\infty} \frac{\nabla_x g_1(x_k, \varepsilon_k) - \nabla g_2(x_k)}{\varepsilon_k} = -\lim_{k \to +\infty} \nabla_x F(-\tilde{\varepsilon}_k, x) = -\nabla_x F(0, x) = \nabla p(x). \tag{11}$$

Furthermore, by Lemma 2,

$$\lim_{k \to +\infty} \frac{g_1(x_k, \varepsilon_k) - g_2(x_k)}{\varepsilon_k} = 1 - F(0, x) = p(x). \tag{12}$$

By Lemma 3, we know that $\Omega(\varepsilon)$ increases as $\varepsilon \searrow 0$ and $\lim_{\varepsilon \searrow 0} \Omega(\varepsilon) = \Omega_0$. Therefore, $\Omega(\varepsilon) \subset \Omega_0$ for $\varepsilon > 0$. It follows from $x_k \in \Omega(\varepsilon_k)$ and $x_k \to x$ that $x \in \Omega_0$. By Proposition 6.6 of Rockafellar and Wets (1998),

$$\limsup_{x_k \to x} N_X(x_k) = N_X(x) \tag{13}$$

when $x, x_k \in X$. Then, by taking $k \to +\infty$ in Equations (9) and (10), we obtain that $(x, \lambda) \in \Lambda_0$. This concludes the proof of the theorem. □

To obtain a stronger convergence result of the KKT pairs, we make the following assumption.

Assumption 6. *Suppose that the following regularity condition holds for every feasible point $x \in \Omega_0$:*

$$\left.\begin{array}{l} 0 \in \lambda \nabla p(x) + N_X(x) \\ \lambda \geq 0, \lambda[p(x) - \alpha] = 0 \end{array}\right\} \Longrightarrow \lambda = 0. \tag{14}$$

Assumption 6 is a constraint qualification of the constraints $x \in X$ and $p(x) - \alpha \leq 0$, which ensures the existence of KKT pairs. It is a frequently used condition for a set involving the abstract constraint $x \in X$, see, for instance, the basic constraint qualification defined in Theorem 6.14 of Rockafellar and Wets (1998). Let Σ be the set of all stationary points of Problem (P). Then, $\Sigma \subset X$ is a compact set. For any $x \in \Sigma$, it follows from Exercise 6.39 and Example 6.40 of Rockafellar and Wets (1998) that Condition (14) is Robinson constraint qualification, and it is Mangasarian-Fromovitz (MF) constraint qualification when X is a compact polyhedron.

Then, we have the following theorem that is stronger than Theorem 3. The proof of the theorem is provided in the electronic companion to this paper.

Theorem 4. *Suppose that Assumptions 1 to 6 are satisfied and $X \subset \Re^d$ is a compact polyhedron. Then,* $\lim_{\varepsilon \searrow 0} \mathbb{D}(\Lambda(\varepsilon), \Lambda_0) = 0$.

Theorems 3 and 4 show that the cluster points of the sequence of KKT pairs of Problem (P_ε) are KKT pairs of Problem (P). Therefore, the KKT points of Problem (P_ε) are good approximations to the KKT points of Problem (P) when ε is small. In the rest of this paper, we consider how to find a KKT point of Problem (P_ε).

3 Sequential Convex Approximations

Problem (P_ε) is a DC program, because the left-hand side of the constraint $g_1(x, \varepsilon) - g_2(x) - \varepsilon\alpha \leq 0$ is a DC function of x. To solve the problem, we propose to use a sequence of convex approximations. In Section 3.1, we first introduce an algorithm to solve this type of DC programs. Then, in Section 3.2, we show how to apply the algorithm to solve Problem (P_ε). The algorithm starts with an initial feasible solution of Problem (P_ε), we show how to find a good initial solution in Section 3.3.

3.1 Algorithm SCA

Consider the nonlinear optimization problem of the form

$$\text{(DCP)} \qquad \begin{array}{ll} \text{minimize} & h(x) \\ \text{subject to} & g_1(x) - g_2(x) \leq 0, \\ & x \in X, \end{array}$$

where $X \subset \Re^d$ is a nonempty convex compact set, $h : \Re^d \to \Re$, $g_i : \Re^d \to \Re$, $i = 1, 2$, are real-valued continuously differentiable and convex functions in a bounded open set $\mathcal{O} \supset X$. Note that Problem (P_ε) is an example of Problem (DCP).

Let $\Omega = \{x \in X : g_1(x) - g_2(x) \leq 0\}$ and

$$\Omega_y = \{x \in X : g_1(x) - [g_2(y) + \nabla g_2(y)^T(x - y)] \leq 0\}$$

for any $y \in X$. Note that $g_2(y) + \nabla g_2(y)^T(x - y)$ defines a tangent plane of $g_2(x)$ at $x = y$. Since $g_2(x)$ is convex, we have for any $y \in X$,

$$g_2(x) \geq g_2(y) + \nabla g_2(y)^T(x - y), \forall x \in X,$$

which implies

$$g_1(x) - g_2(x) \leq g_1(x) - [g_2(y) + \nabla g_2(y)^T(x - y)]. \tag{15}$$

Then, $\Omega_y \subset \Omega$ for any $y \in X$. Furthermore, since $g_1(x) - [g_2(y) + \nabla g_2(y)^T(x - y)]$ is a convex function of x, then Ω_y is a convex subset of Ω for any $y \in X$. Define Problem (CP(y)) as

$$(\mathrm{CP}(y)) \qquad \text{minimize}\, \{h(x) : x \in \Omega_y\}.$$

Then, CP(y) is a convex conservative approximation of Problem (DCP) for any $y \in X$. We suggest using the following algorithm to solve Problem (DCP).

Algorithm SCA

Step 0 Give $x_0 \in \Omega$ and set $k = 0$.

Step 1 Stop if x_k satisfies the KKT condition of Problem (DCP).

Step 2 Solve CP(x_k) to obtain its optimal solution x_{k+1}.

Step 3 Set $k = k + 1$ and go to Step 1.

A similar algorithm was proposed by Smola et al. (2005) as an approach to solving DC programs in the form of Problem (DCP). However, they did not provide rigorous analysis on the convergence of the algorithm, although they claim that the algorithm can find a KKT point of Problem (DCP). In the rest of this subsection, we analyze the properties of the algorithm and prove the claim of Smola et al. (2005) in a rigorous way.

Property 1. *If $\{x_k\}$ is generated by Algorithm SCA for Problem (DCP) starting from $x_0 \in \Omega$, then $\{x_k\} \subset \Omega$.*

Proof. Note that $x_1 \in \Omega_{x_0}$, and $\Omega_y \subset \Omega$ for any $y \in X$ by Equation (15). Then $x_1 \in \Omega$. Therefore, by the principle of induction, we have $\{x_k\} \subset \Omega$. □

Property 2. *If $\{x_k\}$ is generated by Algorithm SCA for Problem (DCP) starting from $x_0 \in \Omega$, then $\{h(x_k)\}$ is a convergent non-increasing sequence.*

Proof. Since Ω_{x_k} is a convex compact set and h is convex, CP(x_k) has an nonempty compact solution set and $x_{k+1} \in \operatorname{argmin}\{h(x) : x \in \Omega_{x_k}\}$. Noting that $x_k \in \Omega_{x_k}$ for every $k \geq 1$, we have that $h(x_{k+1}) \leq h(x_k)$. As $h(\cdot)$ is continuous and X is compact, we obtain that $\inf_k\{h(x_k)\}$ is finite, which is greater than or equal to $\inf_{x\in X} h(x)$, and $\lim_{k\to+\infty} h(x_k) = \inf_k\{h(x_k)\}$. □

The next property states that the cluster points of $\{x_k\}$ are all KKT points of Problem (DCP). To prove it, we need a constraint qualification. We say that Slater's condition holds at $y \in \Omega$ if $\operatorname{int}\Omega_y \neq \emptyset$. Note that Slater's condition is the most commonly used constraint qualification in convex optimization (Boyd and Vandenberghe 2004). The proof of the property is quite lengthy, so we include it in the online companion to this paper.

Property 3. *Let $\{x_k\}$ be the sequence of solutions generated by Algorithm SCA for Problem (DCP) starting from $x_0 \in \Omega$. Suppose that $\bar{x}$ is a cluster point of $\{x_k\}$ satisfying Slater's condition. Then, $\bar{x}$ is a KKT point of Problem (DCP). Moreover, if h is strictly convex in $\mathcal{O}$, then $\{x_k\}$ converges to a KKT point of Problem (DCP).*

By Properties 1 to 3, we see that Algorithm SCA has many desired properties. In the next subsection, we show how to apply it to solve Problem (P_ε).

3.2 Algorithm SCA for Problem (P_ε)

Note that Problem (P_ε) is exactly in the form of Problem (DCP), where we only need to define $g_1(x) = g_1(x, \varepsilon) - \varepsilon\alpha$. Then we can apply Algorithm SCA to solve Problem (P_ε) directly.

By Assumption 1 and the conclusions of Lemma 2, we can verify that Problem (P_ε) satisfies the definition of Problem (DCP) when ε is small enough. Let $\{x_k\}$ be the sequence of solutions generated by Algorithm SCA for Problem (P_ε) starting from $x_0 \in \Omega(\varepsilon)$. By Properties 1 and 2,

we have that $\{x_k\} \subset \Omega(\varepsilon)$ and $\{h(x_k)\}$ is a convergent non-increasing sequence when ε is small enough.

To apply Property 3 to Problem (P_ε), however, we need to prove that all cluster points of $\{x_k\}$ satisfy Slater's condition. Let $\bar{X}$ denote a cluster point of $\{x_k\}$. Since $g_1(x,\varepsilon)$ and $g_2(x)$ are both continuous in x, then $\Omega(\varepsilon)$ is a closed set. Since $\{x_k\} \subset \Omega(\varepsilon)$, we have $\bar{x} \in \Omega(\varepsilon)$. Then, we only need to prove that $\operatorname{int}\Omega(\varepsilon)_{\bar{x}} \neq \emptyset$, which is implied by the conclusion of the following lemma.

Lemma 4. *Suppose that Assumptions 1 to 6 are satisfied. Then, int* $\Omega(\varepsilon)_y \neq \emptyset$ *for any* $y \in \Omega(\varepsilon)$ *when* $\varepsilon > 0$ *is small enough.*

Proof. By contradiction. Suppose that there exist $\varepsilon_i \searrow 0$ and $y_i \in \Omega(\varepsilon_i)$ such that $\operatorname{int}\Omega(\varepsilon_i)_{y_i} = \emptyset$. Note that

$$\Omega(\varepsilon)_{y_i} = \{x \in X : g_1(x,\varepsilon_i) - [g_2(y_i) + \nabla g_2(y_i)(x - y_i)] - \varepsilon_i\alpha \leq 0\}.$$

Then, the equality $\operatorname{int}\Omega(\varepsilon_i)_{y_i} = \emptyset$ implies $g_1(y_i,\varepsilon_i) - g_2(y_i) - \varepsilon_i\alpha = 0$ and

$$y_i \in \operatorname{argmin}_{x\in X} \left\{g_1(x,\varepsilon_i) - [g_2(y_i) + \nabla g_2(y_i)(x - y_i)] - \varepsilon_i\alpha\right\}. \tag{16}$$

As X is a convex compact set, the optimization problem of Equation (16) is a convex problem. Then, we have from the necessary optimality condition that

$$-[\nabla_x g_1(y_i,\varepsilon_i) - \nabla g_2(y_i)] \in N_X(y_i), \quad g_1(y_i,\varepsilon_i) - g_2(y_i) - \varepsilon_i\alpha = 0,$$

or equivalently

$$-\frac{\nabla_x g_1(y_i,\varepsilon_i) - \nabla_x g_1(y_i,0)}{\varepsilon_i} \in N_X(y_i), \quad \frac{g_1(y_i,\varepsilon_i) - g_1(y_i,0)}{\varepsilon_i} - \alpha = 0. \tag{17}$$

Since $\{y_i\} \subset X$ and X is compact, $\{y_i\}$ has a cluster point, say $\bar{y} \in \Omega_0$. Assume that there is a subsequence $\{y_{k_j}\}$ such that $y_{k_j} \to \bar{y}$. Letting $j \to +\infty$ and by Equations (11) to (13), we have

$$-\nabla p(\bar{y}) \in N_X(\bar{y}), \quad p(\bar{y}) - \alpha = 0, \quad \bar{y} \in \Omega_0,$$

which implies that $\lambda = 1$ is a solution of

$$0 \in \lambda\nabla p(x) + N_X(x), \quad \lambda \geq 0, \quad \lambda[p(x) - \alpha] = 0$$

when $x = \bar{y} \in \Omega_0$. Then, it contradicts Assumption 6. This concludes the proof of the lemma. □

Note that Lemma 4 shows that int $\Omega(\varepsilon)_y \neq \emptyset$ for any $y \in \Omega(\varepsilon)$. Then, it also holds for $y = \bar{x}$. Therefore, the conclusions of Property 3 also hold when Algorithm SCA is applied to solve Problem (P_ε) when ε is small enough.

For completeness, we summarize the three properties for Problem (P_ε) in the following theorem.

Theorem 5. *Let $\{x_k\}$ be the sequence of solutions generated by Algorithm SCA for Problem (P_ε) starting from $x_0 \in \Omega(\epsilon)$. Suppose that Assumptions 1 to 6 are satisfied. Then, for any $\varepsilon > 0$ small enough, $\{x_k\} \subset \Omega(\epsilon)$, $\{h(x_k)\}$ is a convergent non-increasing sequence, all cluster points of $\{x_k\}$ are KKT points of Problem (P_ε). Furthermore, $\{x_k\}$ converges to a KKT point of Problem (P_ε) if h is strictly convex.*

3.3 Initial Solutions for Problem (P_ε)

To apply Algorithm SCA, we need an initial solution $x_0 \in \Omega(\varepsilon)$. In this subsection, we provide two natural choices.

In the first choice, we let $\Omega_0(\varepsilon) = \{x \in X : g_1(x, \varepsilon) \leq \varepsilon\alpha\}$. Note that $g_2(x) = \mathrm{E}\left[c(x, \xi)\right]^+ \geq 0$ for all $x \in X$. Then, $\Omega_0(\varepsilon) \subset \Omega(\varepsilon)$. Furthermore, $\Omega_0(\varepsilon)$ is a convex set since $g_1(x, \varepsilon)$ is a convex function of x when $x \in \mathcal{O}$. Then,

$$\text{minimize } h(x), \quad \text{subject to } x \in \Omega_0(\varepsilon)$$

is a convex optimization problem. Let $x_\varepsilon \in \operatorname{argmin} \{h(x) : x \in \Omega_0(\varepsilon)\}$, we have $x_\varepsilon \in \Omega(\varepsilon)$.

In the second choice, we let $\Omega_{\mathrm{CVaR}} = \{x \in X : \mathrm{CVaR}_{1-\alpha}(c(x, \xi)) \leq 0\}$. By the discussions in Section 2.1,

$$\Omega_{\mathrm{CVaR}} = \left\{ x \in X : \inf_{t>0} \frac{1}{t} g_1(x, t) \leq \alpha,\ x \in X \right\}.$$

Let $x_{\mathrm{CVaR}} \in \operatorname{argmin} \{h(x) : x \in \Omega_{\mathrm{CVaR}}\}$, which is the optimal solution of the CVaR approximation of Rockafellar and Uryasev (2000). Let $\varepsilon^* = q_{1-\alpha}(c(x_{\mathrm{CVaR}}, \xi))$, which is the $1 - \alpha$ quantile of $c(x_{\mathrm{CVaR}}, \xi)$. By Pflug (2000), $\varepsilon^* > 0$ and

$$\inf_{t>0} \frac{1}{t} g_1(x_{\mathrm{CVaR}}, t) = \frac{1}{\varepsilon^*} g_1(x_{\mathrm{CVaR}}, \varepsilon^*).$$

Then, $x_{\mathrm{CVaR}} \in \{x \in X : g_1(x, \varepsilon^*) \leq \varepsilon^*\alpha\}$. Since $g_2(x) \geq 0$, then

$$x_{\mathrm{CVaR}} \in \{x \in X : g_1(x, \varepsilon^*) - g_2(x) \leq \varepsilon^*\alpha\} = \Omega(\varepsilon^*).$$

By Lemma 1, $\Omega(\varepsilon^*) \subset \Omega(\varepsilon)$ for any $0 < \varepsilon \leq \varepsilon^*$. Then, $x_{\text{CVaR}} \in \Omega(\varepsilon)$ for any $0 < \varepsilon \leq \varepsilon^*$. Therefore, one may first solve the CVaR approximation and find x_{CVaR} and then select $\varepsilon \in (0, \varepsilon^*]$. Then, x_{CVaR} may be used as an initial solution to Algorithm SCA.

When we let $x_0 = x_{\text{CVaR}}$, the sequence of solutions $\{x_k\}$ generated by Algorithm SCA are improving and at least as good as the CVaR approximation, which is the "best" convex conservative approximation. Furthermore, $\{x_k\}$ converges to the set of KKT points of Problem (P_ε) by Theorem 5, which converges to the set of KKT points of Problem (P) as $\varepsilon \searrow 0$ by Theorem 4. Compared to other approximation algorithms, e.g., CVaR approximation, quadratic approximation, Bernstein approximation and scenario analysis, which only find a (good) feasible solution to Problem (P), our algorithm has more desirable properties. If Problem (P) is convex (even though it may not be verifiable), our algorithm converges to its global optimal solution, while others do not.

4 A Gradient-based Monte Carlo Method

To implement Algorithm SCA, we need to repeatedly solve

$$\begin{array}{ll} \text{minimize} & h(x) \\ \text{subject to} & g_1(x,\varepsilon) - \left[g_2(y) + \nabla g_2(y)^T(x-y)\right] \leq \varepsilon\alpha, \\ & x \in X \end{array} \tag{18}$$

for different y values. Though Problem (18) is a convex optimization problem, it is difficult to solve because we generally do not have the closed form expressions of $g_1(x,\varepsilon)$, $g_2(x)$ and $\nabla g_2(y)$. To overcome this difficulty, we propose to use a Monte Carlo method.

For simplicity of the notation, we let

$$g(x) = g_1(x,\varepsilon) - \left[g_2(y) + \nabla g_2(y)^T(x-y)\right].$$

Then, by the definitions of g_1, g_2 and Lemma 2, we have

$$g(x) = \text{E}\left\{[c(x,\xi)+\varepsilon]^+\right\} - \left[\text{E}\left\{[c(y,\xi)]^+\right\} + \text{E}\left\{\left[\nabla_x c_{i^*}(y,\xi)\cdot 1_{(0,+\infty)}\left(c(y,\xi)\right)\right]\right\}^T (x-y)\right], \tag{19}$$

where $c(x,\xi) = \max\{c_i(x,\xi)\}$ and $i^* = \text{argmax}_{i=1,\ldots,m}\{c_i(y,\xi)\}$. Let $\xi_1,\ldots,\xi_n$ denote an independent and identically distributed (i.i.d.) sample of ξ. Let $\bar{g}_2(y) = \frac{1}{n}\sum_{\ell=1}^n c(y,\xi_\ell)$ and $\bar{\nabla} g_2(y) = \frac{1}{n}\sum_{\ell=1}^n \nabla_x c_{i^*}(y,\xi_\ell)\cdot 1_{(-t,+\infty)}\left(c(y,\xi_\ell)\right)$. Then, a natural estimator of $g(x)$ is

$$\bar{g}(x) = \frac{1}{n}\sum_{\ell=1}^n [c(x,\xi_\ell)+\varepsilon]^+ - \left[\bar{g}_2(y) + \bar{\nabla} g_2(y)^T(x-y)\right]. \tag{20}$$

We suggest solving

(MC) $$\text{minimize} \;\; h(x), \quad \text{subject to} \;\; \bar{g}(x) \leq \varepsilon\alpha, \; x \in X,$$

and using its optimal solution to approximate the optimal solution of Problem (18). Let S and ν^* denote the set of optimal solutions and the optimal objective value of Problem (18), and $\hat{S}_n$ and $\hat{\nu}_n^*$ denote the set of optimal solutions and the optimal objective value of Problem (MC). Then there are two critical issues when we use this approximation. First, do $\hat{S}_n$ and $\hat{\nu}_n^*$ converge to S and ν^*? Second, how do we solve Problem (MC) efficiently?

To answer the first question, we have the following theorem.

Theorem 6. *Suppose that Assumptions 1 to 6 are satisfied. When $\varepsilon > 0$ is small enough, $\mathbb{D}(\hat{S}_n, S) \to 0$ w.p.1 and $\hat{\nu}_n^* \to \nu^*$ w.p.1 as $n \to \infty$.*

Proof. Note that we may write $g(x)$ and $\bar{g}(x)$ as

$$g(x) = \mathrm{E}\left\{[c(x,\xi)+\varepsilon]^+ - [c(y,\xi)]^+ - \left[\nabla_x c_{i^*}(y,\xi)\cdot 1_{(0,+\infty)}\left(c(y,\xi)\right)\right]^T (x-y)\right\}, \tag{21}$$

and

$$\bar{g}(x) = \frac{1}{n}\sum_{\ell=1}^{n}\left\{[c(x,\xi_\ell)+\varepsilon]^+ - [c(y,\xi_\ell)]^+ - \left[\nabla_x c_{i^*}(y,\xi_\ell)\cdot 1_{(0,+\infty)}\left(c(y,\xi_\ell)\right)\right]^T (x-y)\right\}.$$

Then, $\bar{g}(x)$ is the sample average approximation of $g(x)$. By the strong law of large numbers (Durrett 2005), $\bar{g}(x) \to g(x)$ w.p.1 for any fixed $x \in X$. Furthermore, since the integrand of Equation (21) is a convex function of x for any $\xi \in \Xi$ when $x \in \mathcal{O}$, by Theorem 6.38 of Shapiro and Ruszczyński (2008), $\bar{g}(x)$ converges to $g(x)$ uniformly on X w.p.1 as $n \to \infty$, i.e.,

$$\sup_{x\in X} |\bar{g}(x) - g(x)| \to 0 \quad \text{w.p.1 as } n \to \infty.$$

Also, by Lemma 4, Slater's condition holds for Problem (18) when Assumptions 1 to 6 are satisfied and $\varepsilon > 0$ is small enough. Then, by Theorem 4.5 of Shapiro and Ruszczyński (2008) and the discussions followed the theorem, the conclusions of our theorem hold. □

To answer the second question, we propose two methods to solve Problem (MC). In the first method, based on Equation (20) and the definition of $c(x,\xi)$, we reformulate Problem (MC) as

$$\begin{array}{ll} \text{minimize} & h(x) \\ \text{subject to} & c_i(x,\xi_\ell) \leq z_\ell, \quad i = 1,\ldots,m, \; \ell = 1,\ldots,n, \\ & \frac{1}{n}\sum_{\ell=1}^{n} z_\ell - \left[\bar{g}_2(y) + \bar{\nabla} g_2(y)^T (x-y)\right] \leq \varepsilon\alpha, \\ & z_\ell \geq 0, \quad \ell = 1,\ldots,n, \\ & x \in X. \end{array} \tag{22}$$

Note that Problem (22) is similar to the formulation of the sample CVaR problem of Rockafellar and Uryasev (2000). It is a convex optimization problem. Furthermore, it is a linear program if $c_i(x,\xi)$ are linear functions of x for all $i = 1,\ldots,m$. However, Problem (22) is often slow to solve because of the large numbers of decision variables and constraints, especially when the sample size n is large.

To efficiently solve Problem (MC) when n is large, we propose a second method. By Equation (19) and Lemma 2,

$$\nabla g(x) = \mathrm{E}\left[\nabla_x c_{j^*}(x,\xi)\cdot 1_{(-\varepsilon,+\infty)}(c(x,\xi))\right] - \mathrm{E}\left[\nabla_x c_{i^*}(y,\xi)\cdot 1_{(0,+\infty)}(c(y,\xi))\right],$$

where $j^* = \mathrm{argmax}_{j=1,\ldots,m}\{c_j(x,\xi)\}$ and $i^* = \mathrm{argmax}_{i=1,\ldots,m}\{c_i(y,\xi)\}$. Then, $\nabla g(x)$ can be estimated by

$$\bar{\nabla} g(x) = \frac{1}{n}\sum_{\ell=1}^{n}\left[\nabla_x c_{j^*}(x,\xi_\ell)\cdot 1_{(-\varepsilon,+\infty)}\left(c(x,\xi_\ell)\right) - \nabla_x c_{i^*}(y,\xi_\ell)\cdot 1_{(0,+\infty)}\left(c(y,\xi_\ell)\right)\right].$$

Although $g(x)$ is differentiable in $\mathcal{O}$, $\bar{g}(x)$ is not when $\xi_1,\ldots,\xi_n$ are given. It is only piecewise differentiable. At the points where $\bar{g}(x)$ is differentiable, $\nabla\bar{g}(x) = \bar{\nabla} g(x)$; at the points where $\bar{g}(x)$ is not differentiable, $\bar{\nabla} g(x)$ is a subgradient of $\bar{g}(x)$. Therefore, one may use a subgradient-based algorithm (see, for instance, Freund (2004)) to solve Problem (MC). In this paper, however, we suggest using an approximation method. Since $\bar{g}(x)$ converges to $g(x)$ which is continuously differentiable as $n\to\infty$, we may approximate $\bar{\nabla} g(x)$ as a smooth function when n is large, and use $\bar{\nabla} g(x)$ as its gradient. Then, we can use gradient-based algorithms to solve Problem (MC) directly. Note that this method can also be viewed as directly solving Problem (18) with estimated $g(x)$ and $\nabla g(x)$.

When we use the gradient-based method to solve Problem (MC), the samples are only used to compute $\bar{g}(x)$ and $\bar{\nabla} g(x)$, which is an $O(n)$ operation. The method is generally much faster than the first method that solves Problem (22). Hong and Liu (2009) compared the two methods for the CVaR approximation problem through numerical examples. They reported that the two methods find solutions with similar quality, but the gradient-based method is at least an order-of-magnitude faster when n is of moderate or large size, e.g., $n \geq 2000$. In our numerical experiments, we observe good performances of the gradient-based method as well.

5 Numerical Illustration

In this section, we consider two JCCP problems, a norm optimization problem and a network optimization problem. We use them to illustrate the performances of our method, and compare our method to the CVaR approximation and the scenario approach.

5.1 A Norm Optimization Problem

Let $x = (x_1, \ldots, x_d)^T$ denote a d-dimensional vector in $\Re^d$, and let $\xi = (\xi_1, \ldots, \xi_m)$, with $\xi_i = (\xi_{i1}, \ldots, \xi_{id})^T$ for any $i = 1, \ldots, m$, be a $d \times m$ matrix of random variables. Let $\|x\|_1$ and $\|x\|$ denote the 1-norm and 2-norm of x respectively, i.e., $\|x\|_1 = \sum_{j=1}^d |x_j|$ and $\|x\| = \left(\sum_{j=1}^d x_j^2\right)^{-1/2}$, and let $\xi_i \circ x = (\xi_{i1}x_1, \ldots, \xi_{id}x_d)^T$ denote the Hadamard product (or entrywise product) of ξ_i and x. We are interested in solving the following problem:

$$\begin{array}{ll} \text{maximize} & \|x\|_1 \\ \text{subject to} & \Pr\{\|\xi_i \circ x\| \le 10, \quad i = 1, \ldots, m\} \ge 1 - \alpha, \\ & x_j \ge 0, \quad j = 1, \ldots, d. \end{array} \tag{23}$$

We may reformulate Problem (23) as

$$\begin{array}{ll} \text{minimize} & -\sum_{j=1}^d x_j \\ \text{subject to} & \Pr\left\{\sum_{j=1}^d \xi_{ij}^2 x_j^2 \le 100, \quad i = 1, \ldots, m\right\} \ge 1 - \alpha, \\ & x_j \ge 0, \quad j = 1, \ldots, d. \end{array} \tag{24}$$

Note that Problem (24) is a JCCP as defined in Problem (P).

Let $c_i(x, \xi) = \sum_{j=1}^d \xi_{ij}^2 x_j^2 - 100$ for all $i = 1, \ldots, m$. For any $x \neq 0$, $c_i(x, \xi)$ is a continuous random variable and $c_i(x, \xi) = c_j(x, \xi)$ with probability 0. Therefore, Assumption 3 can be satisfied easily. When $x = 0$, $c_i(x, \xi) = -100$ for all $i = 1, \ldots, m$. By the definition of differentiability, for any $j = 1, \ldots, d$,

$$\frac{\partial}{\partial x_j} c(0, \xi) = \lim_{\delta \to 0} \frac{1}{\delta} \left[c(e_j\delta, \xi) - c(0, \xi)\right] = \lim_{\delta \to 0} \frac{1}{\delta} \max_{i=1,\ldots,m} \left(\xi_{ij}^2 \delta^2\right) = \lim_{\delta \to 0} \max_{i=1,\ldots,m} \xi_{ij}^2 \cdot \delta = 0,$$

where e_j denotes the ith column of a $d \times d$ identity matrix. Therefore, $c(x, \xi)$ is differentiable at $x = 0$ for any ξ and Assumption 3 is satisfied.

In the rest of this subsection, we apply Algorithm SCA with the gradient-based Monte Carlo method to solve this problem.

5.1.1 Independent Case

We consider the case where ξ_{ij}, $i = 1, \ldots, m$ and $j = 1, \ldots, d$, are independent and identically distributed standard normal random variables. We will call this case *independent case* in the rest

of this section. Note that convexity of Problem (23) in this situation is not clear. However, by symmetry, the optimal solution satisfies $x_1 = \cdots = x_d$. Then,

$$\Pr\left\{\sum_{j=1}^{d} \xi_{ij}^2 x_j^2 \leq 100, \quad i = 1, \ldots, m\right\} = \left[\Pr\left\{x_1^2 \sum_{j=1}^{d} \xi_{1j}^2 \leq 100\right\}\right]^m.$$

Note that $\sum_{j=1}^{d} \xi_{1j}^2$ follows a chi-square distribution with d degrees of freedom. Let $F_{\chi_d^2}(\cdot)$ denote its distribution function. Then, the joint chance constraint of Problem (24) is equivalent to

$$F_{\chi_d^2}\left(\frac{100}{x_1^2}\right) \geq (1-\alpha)^{1/m}.$$

Let $1 - \beta = (1-\alpha)^{1/m}$ and let $F_{\chi_d^2}^{-1}(\cdot)$ denote the inverse distribution function of a chi-square distribution with d degrees of freedom. Then, it is clear that the optimal solution x^* of Problem (24) is

$$x_1^* = \cdots x_d^* = \frac{10}{F_{\chi_d^2}^{-1}(1-\beta)}.$$

We apply our gradient-based Monte Carlo method directly to solve Problem (24) without exploring its special structure (e.g., the independence of ξ), and we use x^* as a benchmark to evaluate the performances of our method.

We set $d = 10$, $m = 10$ and $\alpha = 0.1$. Then, the optimal solution of Problem (24) is $x_1^* = \cdots = x_d^* = 2.08$ and the optimal objective value is $f^* = -20.82$. We set $\varepsilon = 0.05$[2] and use a sample size $n = 10000$. We use both x_{CVaR} and x_ε as initial solutions to compare their performances, and stop the algorithm if the difference between the objective values of two consecutive iterations is less than or equal to 0.01. We implement the algorithm in Matlab and use Matlab's own nonlinear optimization solver `fmincon` to solve the optimization problem in each iteration with the estimated constraint values and estimated gradients. The programs were run on a desktop computer with Intel Duo Core CPU (3.16GHz, 3.16GHz) and 4 GB of RAM.

Although our algorithm uses Monte Carlo samples, we find the performances of the algorithm are very stable. We run the algorithm 100 replications, it always converges to similar solutions. We report the typical performances of the algorithm in Figure 3. In the left panel of Figure 3, we plot the objective values of all iterations. From the plot, we can see that the algorithm converges to the optimal objective value from both x_{CVaR} and x_ε, even though the convexity of the problem is not

[2]Based on the convergence analysis, we may want to set ε small to reduce the bias. However, extremely small ε may cause numerical problems and may require longer time to solve the subproblem in each iteration. Given the error in Monte Carlo estimation, we do not suggest setting ε extremely small. Finding the optimal setting of ε is an important problem for future research.

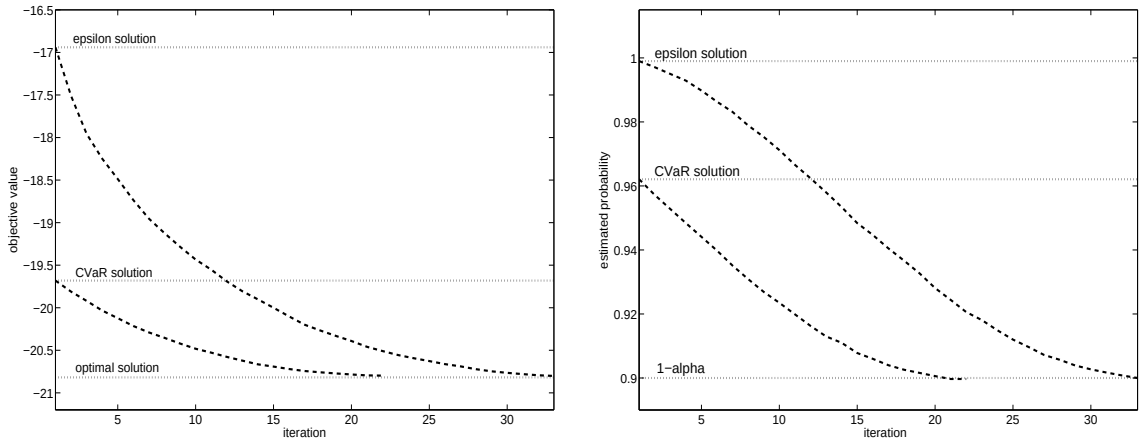

Figure 3: Performance of Algorithm SCA for Independent Case

clear. Furthermore, we can see that x_{CVaR} is a better initial solution, since the algorithm converges faster from it than from x_ϵ. In the right panel of Figure 3, we plot the values of left-hand side of the joint chance constraint, estimated using the sample, for all iterations. From the plot, we see that the joint chance constraint becomes tight when our algorithm stops, while it is not tight at x_ϵ and x_{CVaR}. The algorithm typically requires about 20 iterations to converge when x_{CVaR} is used as the initial solution and about 35 iterations when x_ε is used. The CPU time of each iteration is on average 6.8 seconds with a range of 4.4 to 12.2 seconds.

5.1.2 Dependent Case

We also considered the case where ξ_{ij}, $i = 1, \ldots, m$ and $j = 1, \ldots, d$, are dependent. We let ξ_{ij} be a normal random variable with mean j/d and variance 1, $\text{Cov}(\xi_{ij}, \xi_{i'j}) = 0.5$ when $i \neq i'$ and $\text{Cov}(\xi_{ij}, \xi_{i'j'}) = 0$ when $j \neq j'$. We will call this case *dependent case* in the rest of this subsection. Then, the joint chance constraint of Problem (24) can no longer be converted to a single chance constraints and the optimal solution is no longer known. However, we can still apply our method to solve the problem. With the same setting of the parameters as in the independent case, we report the performances of the algorithm in Figure 4. From the plots, we can see that changing the dependence structure of ξ does not alter the performances of our algorithm. Both initial solutions lead the algorithm to converge to the same objective value and the joint chance constraint becomes tight when the algorithm stops. The number of iterations and the CPU time in this case are also similar to the ones in the independent case.

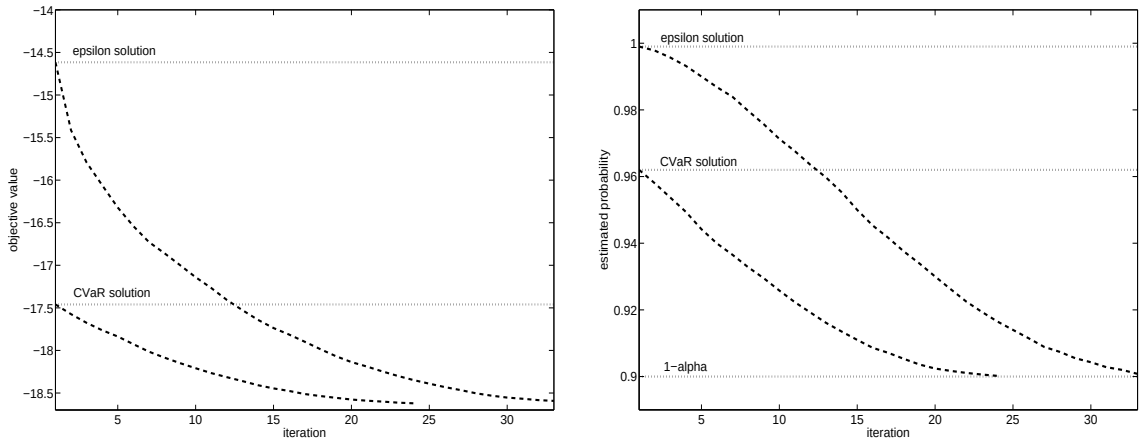

Figure 4: Performance of Algorithm SCA for Dependent Case

5.1.3 Scenario Approach

We also implement the scenario approach for this problem with $d = m = 10$ and $\alpha = 0.1$. By Calafiore and Campi (2006), we let the sample size equal to 1038. Then, by the formulation of Problem (2), we need to solve an optimization problem with 10380 constraints. We also use Matlab's nonlinear optimization solver `fmincon` to solve the problem. We run the algorithm 100 replications. To our surprise, the problem can be solved very efficiently (0.74 seconds on average) even though there are more than 10000 constraints. This is because the number of active constraints is only about 10 for all the replications that we have tried. However, the solutions are not stable, i.e., they are drastically different from replication to replication due to the randomness in the sample. For the independent case, the average objective value is -17.6 with the best being -18.6 and the estimated left-hand side value of the joint chance constraint is in the range of 0.982 to 0.999. For the dependent case, the average objective value is -15.7 with the best being -16.7 and the estimated left-hand side value of the joint chance constraint being in the range of 0.984 to 0.999. For both cases, the solutions found by the scenario approach are too conservative, and they are significantly worse than the solutions found by the CVaR approximation and our method, even though it is the fastest among all three.

5.2 A Network Optimization Problem

A generalized network flow problem (GNFP) is an extension of the classical network flow problem. In a GNFP, the flow on an arc is subject to change. If we use x_{ij} to denote the flow on arc e_{ij}

when the flow leaves node i, then the flow will become $\eta_{ij}x_{ij}$ when it arrives at node j, where $0 \leq \eta_{ij} \leq 1$ is a parameter indicating the changing rate. The GNFP has many useful applications. For example, in telecommunication, it models the packet loss on an unreliability transmission link; in production planning, it models the yield rate when one material is converted to another material. In our computational experiments, we consider a GNFP for electricity distribution where η_{ij} is used to model the power loss incurred on a transmission line (Jensen and Bard 2003).

In the problem, there are three electricity generating stations that serve ten different areas. Each station has a different generating cost and a capacity, and each area has a demand. The electricity transmission from a station to an area or to another station is subject to a random percentage loss, which is denoted by η_{ij}. We model η_{ij} by a beta distribution, whose mean and variance depend on the distance between i and j, and $0 \leq \eta_{ij} \leq 1$. The problem is to use the minimum cost to serve all areas where the probability of under-supply for any area or generating station is below a certain bound.

Let c_i, ℓ_i and u_i denote the marginal generation cost, the minimum generation amount, and the capacity at station i, respectively, let d_k denote the demand at area k, and let x_i, y_{ik} and z_{ij} denote the electricity generated at station i, the electricity transmitted from station i to area k, and the electricity transmitted from station i to station j, respectively, for all $i, j = 1, 2, 3$, $i \neq j$, and $k = 1, \ldots, 10$. Note that x_i, y_{ik} and z_{ij} are decision variables. Then, the problem can be formulated as a JCCP as follows:

$$
\begin{aligned}
\text{minimize} \quad & \textstyle\sum_{i=1}^{3} c_i x_i \\
\text{subject to} \quad & \Pr\left\{ \begin{array}{rl} x_i + \sum_{j \neq i} \eta_{ji} z_{ji} \geq \sum_{k=1}^{10} y_{ik} + \sum_{j \neq i} z_{ij}, & i = 1, 2, 3, \\ \sum_i \eta_{ik} y_{ik} \geq d_k, & k = 1, \ldots, 10, \end{array} \right\} \geq 1 - \alpha \\
& \ell_i \leq x_i \leq u_i, y_{ik} \geq 0, z_{ij} \geq 0, \;\; i, j = 1, 2, 3, \; i \neq j, \; k = 1, \ldots, 10.
\end{aligned}
$$

For the test problem we considered, we set $\alpha = 0.1$ and use a sample size of 10000. We compare the CVaR approximation and our method, and run both algorithms 100 times. The (estimated) optimal cost from the CVaR approximation is 2.60×10^6; while the (estimated) optimal cost from our method is 2.07×10^6, which is roughly a 20% reduction of from the CVaR solution.

6 Conclusion and Future Research

In this paper, we propose a sequential convex approximation algorithm that directly solves the stochastic optimization problems with a joint chance constraint. In each iteration of the algorithm, it solves a convex optimization problem using a Monte-Carlo method. We have shown that the algorithm finds a KKT point in the limit if the sample size of the Monte Carlo method goes to infinity and the parameter ε goes down to zero. Among the algorithms that have been proposed to solve this type of problems, to the best of our knowledge, our algorithm is the first one that has provable convergence to the set of KKT points. It is conceptually more attractive than conservative convex approximation algorithms and the scenario approach which find (good) feasible solutions that, nevertheless, satisfy no optimality conditions.

To apply this algorithm for practical problems, especially large-scale practical problems, there are several impediments. First, the algorithm is generally slower than other convex approximation algorithms, because it needs to solve a sequence of convex optimization problems, and because it uses a Monte Carlo method in each iteration. How to speed up the algorithm is a very important topic for future research. Second, our method requires the full joint distribution of the uncertain parameters in order to generate Monte Carlo samples. However, specifying a full joint distribution for a large number of parameters are often very challenging in practice. Different robust optimization algorithms have been proposed to address this issue. A future research topic is how to apply our algorithm under the robust optimization framework to find better solutions.

Acknowledgments

The authors would like to thank the Area Editor, Shane Henderson, the Associate Editor and three anonymous referees for their helpful comments and suggestions. This research was partially supported by Hong Kong Research Grants Council grant CERG 613907.

References

Bazaraa, M. S., H. D. Sherali, and C. M. Shetty. 1993. *Nonlinear Programming: Theory and Algorithms*, 2nd edition, John Wiley & Sons, New York.

Ben-Tal, A. and A. Nemirovski. 2000. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, **88** 411-424.

Bertsimas, D. and M. Sim. 2004. Price of robustness. *Operations Research*, **52** 35-53.

Bonnans, J. F. and A. Shapiro. 2000. *Perturbation Analysis of Optimization Problems*, Springer, New York.

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

Broadie, M. and P. Glasserman. 1996. Estimating security price derivatives using simulation. *Management Science*, **42** 269-285.

Calafiore, G. and M. C. Campi. 2005. Uncertain convex programs: Randomized solutions and confidence levels. *Mathematical Programming*, **102** 25-46.

Calafiore, G. and M. C. Campi. 2006. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, **51** 742-753.

Charnes, A., W. W. Cooper, and G. H. Symonds. 1958. Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil. *Management Science*, **4** 235-263.

Chen, W., M. Sim, J. Sun and C-P Teo. 2009. From CVaR to uncertainty set: Implications in joint chance constrained optimization. *Operations Research*, to appear.

de Farias, D. P. and B. Van Roy. 2004. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, **29** 462-478.

Dentcheva, D., B. Lai, and A. Ruszczyński. 2004. Dual methods for probabilistic optimization problems. *Mathematical Methods of Operations Research*, **60** 331-346.

Durrett, R. 2005. *Probability: Theory and Examples*, Third Edition. Duxbury Press, Belmont.

Erdoğan, E. and G. Iyengar. 2006. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, **107** 37-61.

Freund, R. M. 2004. Subgradient optimization, generalized programming, and nonconvex duality. Technical report, MIT.

Henrion, R. 2007. Structural properties of linear probabilistic constraints. *Optimization*, **56** 425-440.

Henrion, R. and C. Strugarek. 2008. Convexity of chance constraints with independent random variables. *Computational Optimization and Applications*, **41** 263-276.

Hong, L. J. 2009. Estimating quantile sensitivities. *Operations Research*, **57** 118-130.

Hong, L. J. and G. Liu. 2009. Simulating sensitivities of conditional value-at-risk. *Management Science*, **55** 281-293.

Jensen, P. A. and J. F. Bard. 2003. *Operations Research: Models and Methods.* John Wiley & Sons, New York.

Lagoa, C. M., X. Li, and M. Sznaier. 2005. Probabilistically constrained linear programs and risk-adjusted controller design. *SIAM Journal on Optimization*, **15** 938-951.

Luedtke, J. and S. Ahmed. 2008. A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, **19** 674-699.

Luedtke, J., S. Ahmed, and G. Nemhauser. 2007. An integer programming approach for linear programs with probabilistic constraints. *Lecture Notes on Computer Science*, **4513** 410-423.

Marsden, J. and M. Hoffman. 1993. *Elementary Classical Analysis*, Second Edition. Freeman, New York.

Miller, L. B. and H. Wagner. 1965. Chance-constrained programming with joint constraints. *Operations Research*, **13** 930-945.

Nemirovski, A. and A. Shapiro. 2006. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, **17** 969-996.

Pflug, G. 2000. Some remarks on the value-at-risk and the conditional value-at-risk. S. Uryasev, ed. *Probabilistic Constrained Optimization: Methodology and Applications.* Kluwer, Dordrecht.

Prékopa, A. 1970. On probabilistic constrained programming. *Proceedings of the Princeton Symposium on Mathematical Programming*, 113-138.

Prékopa, A. 2003. Probabilistic programming. In *Stochastic Programming, Handbooks in OR&MS.* Vol. 10, A. Ruszczynski and A. Shapiro, eds., Elsevier.

Prékopa, A., T. Rapcsák, and I. Zsuffa. 1978. Serially linked reservoir system design using stochastic programming. *Water Resources Research*, **14** 672-678.

Rockafellar, R. T. and S. Uryasev. 2000. Optimization of conditional value-at-risk. *The Journal of Risk*, **2** 21-41.

Rockafellar, R. T. and R. J.-B. Wets. 1998. *Variational Analysis.* Springer-Verlag, New York.

Shapiro, A. and A. Ruszczyński. 2008. *Lectures on Stochastic Programming*, available on `http://www2.isye.gatech.edu/ ashapiro/publications.html`.

Smola, A.J., S. V. N. Vishwanathan, and T. Hofmann. 2005. Kernel methods for missing variables. *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 325-332.

Zangwill, W. I. 1969. *Nonlinear Programming: A Unified Approach*, Prentice Hall, Englewood Cliffs, New Jersey.

Electronic Companion

A Discussions on Assumption 3

Without loss of generality, we consider $m = 2$ and $c(x, \xi) = \max\{c_1(x, \xi), c_2(x, \xi)\}$. When $m > 2$, we can verify the assumption by iteratively considering two functions at a time. For instance, when $m = 3$, we can first consider $\max\{c_1(x, \xi), c_2(x, \xi)\}$ and then consider $\max\{\max\{c_1(x, \xi), c_2(x, \xi)\}, c_3(x, \xi)\}$.

In the following three situations, Assumption 3 may be satisfied. In the first situation, $c_1(x, \xi) - c_2(x, \xi)$ is a continuous random variable. For instance, when $c_1(x, \xi)$ and $c_2(x, \xi)$ are independent continuous random variables, $c_1(x, \xi) - c_2(x, \xi)$ is a continuous random variable. In this situation, $\Pr\{c_1(x, \xi) = c_2(x, \xi)\} = 0$. Because $c_1(x, \xi)$ and $c_2(x, \xi)$ are differentiable with respect to x for all $\xi \in \Xi$ by Assumption 1, $c(x, \xi) = \max\{c_1(x, \xi), c_2(x, \xi)\}$ is differentiable w.p.1 in this situation and Assumption 3 is satisfied.

In the second situation, $c_1(x, \xi) - c_2(x, \xi) = \beta$ for some constant $\beta \neq 0$. For instance, $c_1(x, \xi) - c_2(x, \xi) = b_1 - b_2$ at $x = 0$ if $c_1(x, \xi) = \xi_1^T x + b_1$ and $c_2(x, \xi) = \xi_2^T x + b_2$, and $c_1(x, \xi) - c_2(x, \xi) = (a_1 - a_2)x + (b_1 - b_2)$ if $c_1(x, \xi) = a_1 x + b_1 + \xi$ and $c_2(x, \xi) = a_2 x + b_2 + \xi$. In this situation, because $\beta \neq 0$, $\Pr\{c_1(x, \xi) = c_2(x, \xi)\} = 0$ and Assumption 3 is satisfied based on the same analysis in the first situation.

In the third situation, $c_1(x, \xi) - c_2(x, \xi) = 0$. For instance, $c_1(x, \xi) - c_2(x, \xi) = 0$ at $x = 0$ if $c_1(x, \xi) = \xi_1^T x$ and $c_2(x, \xi) = \xi_2^T x$. By Assumption 4, $\Pr\{c_1(x, \xi) = 0\} = 0$ and $\Pr\{c_2(x, \xi) = 0\} = 0$. Then, in this situation, $\Pr\{c_1(x, \xi) = \gamma c_2(x, \xi)\} = 0$ for any $\gamma > 0$ and $\gamma \neq 1$. Note that $c_2(x, \xi) \leq 0$ is equivalent to $\gamma c_2(x, \xi) \leq 0$. We may pre-process all the constraints that correspond to this situation by multiplying different constants on their left-hand sides. Then, for all the newly defined constraints, we have $\Pr\{c_i(x, \xi) = c_j(x, \xi)\} = 0$ for all $i \neq j$ and, hence, Assumption 3 is satisfied based on the same analysis in the first situation.[3]

Although we have considered the three situations where $c_i(x, \xi) - c_j(x, \xi)$ is either a continuous random variable or a constant, one may construct examples where $c_i(x, \xi) = c_j(x, \xi)$ with a certain probability between 0 or 1 at certain x values. In such a situation, one can approximate $c(x, \xi)$ by $\delta \log\left[\sum_{i=1}^m e^{c_i(x,\xi)/\delta}\right]$, which is differentiable with respect to x, with a small $\delta > 0$. By Rockafellar

[3] We want to point out that, in this situation, it may still be possible to prove the assumption by using the definition of differentiability in some cases without using pre-processing (see, for instance, the example in Section 5 at $x = 0$).

(1970), when $\delta > 0$,

$$c(x,\xi) \leq \delta \log \left[\sum_{i=1}^{m} e^{\frac{1}{\delta} c_i(x,\xi)} \right] \leq c(x,\xi) + \delta \log m. \tag{25}$$

Therefore, this approximation has a small error if δ is a small positive value. Once the algorithm moves to other points, Assumption 3 may be satisfied.

We may also use the approximation in Equation (25) on all $x \in X$. Then, Assumption 3 is no longer necessary. However, there will be an approximation error. How to quantify the error is beyond the scope of this paper. Nevertheless, it is an interesting problem and should be studied in the future.

B Discussions on Assumption 5

To better understand Assumption 5, we consider a general constraint set of the form $\Omega = \{x \in X : g(x) \leq 0\}$, where $g : \Re^n \to \Re$ is a real-valued function, Ω is a nonconvex set when g is not convex. Further, let $\Omega^I = \{x \in X : g(x) < 0\}$. Many numerical methods, e.g., the barrier function method, solve an optimization problem by approximating the optimal solutions by a sequence of points in Ω^I. Note that such numerical methods can only find solutions in cl Ω^I. If cl $\Omega^I \neq \Omega$, the optimal solution in Ω may not be in cl Ω^I and such numerical methods may not be able to find the optimal solution. This explains why similar assumptions are widely used in designing nonlinear optimization algorithms (see, for instance, Page 264 of Zangwill (1969) and Theorem 9.4.3 of Bazaraa et al. (1993)).

To gain more insight on this type of assumptions, we consider an example. Let $X = [-1,1] \times \Re_+$ and $g(x) = x_2 - (x_1^+)^2$. Then, $\Omega = \{x : x_1 \in [0,1], x_2 - x_1^2 \leq 0\} \cup \{x : x_1 \in [-1,0], x_2 = 0\}$, as in Figure 5. However, note that $\Omega^I = \{x : x_1 \in (0,1], x_2 - x_1^2 < 0\}$ (the shaded area in Figure 5). Then, $cl\left(\Omega^I\right) = \{x : x_1 \in [0,1], x_2 - x_1^2 \leq 0\} \neq \Omega$. Suppose that the optimal solution is $x^* = (-1,0)$ as in Figure 5. Then, there does not exist an sequence of points from Ω^I that can approach x^*. Therefore, any numerical methods that approximate the optimal solutions by sequences of points in Ω^I, e.g., the barrier function method, cannot find the optimal solution of this problem.

Back to our setting, we consider another example. Let $X = [-2,2]$,

$$p(x) = \Pr\left\{[1 + (x^+)^2]\xi - 1 \leq 0\right\}$$

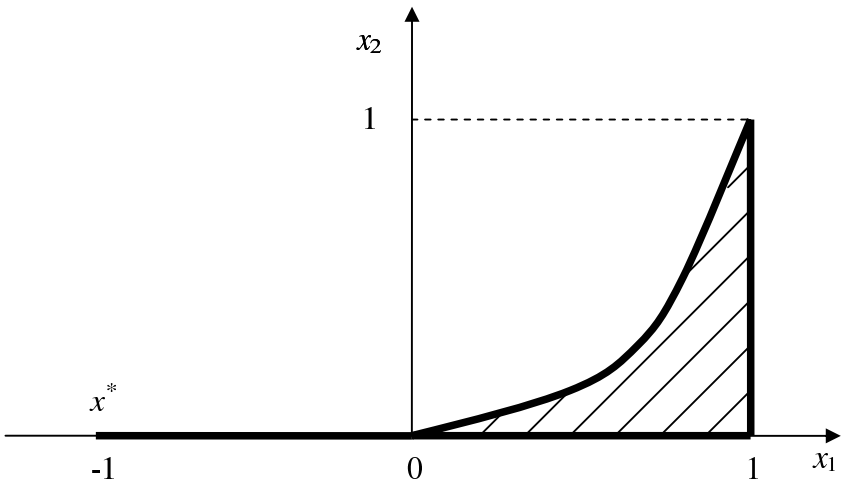

Figure 5: An example where $cl\left(\Omega^I\right) \neq \Omega$.

and $\alpha := \Pr\{\xi \leq 1\}$. One has that

$$p(x) = \begin{cases} \Pr\{\xi \leq 1\} = \alpha & \text{if } x \leq 0, \\ \Pr\left\{\xi \leq \dfrac{1}{1+x^2}\right\} < \alpha & \text{if } x > 0, \end{cases}$$

so that $\Omega = [-2, 2]$ and $\Omega_0^I = (0, 2]$. Therefore cl $\Omega_0^I = [0, 2] \neq \Omega$ and Assumption 5 does not hold for this simple example. If we take $x^* = -2$, we cannot find a sequence in Ω_0^I to approach it.

Our approach tries to find a sequence $\{x_\varepsilon : x_\varepsilon \in \Omega(\varepsilon)\}$, where $\Omega(\varepsilon)$ is defined in Subsection 2.4 satisfying $\Omega(\varepsilon) \subset \Omega_0$. It is possible that there is a sequence $\varepsilon_i \searrow 0$ such that $x_{\varepsilon_i} \in \Omega_0^I$. We wish to prove that any cluster of $\{x_{\varepsilon_i}\}$, say $\widetilde{x}$, is an optimal solution to Problem (P). Since $\widetilde{x} \in$ cl Ω_0^I, Assumption 5 is a natural condition to assume.

C Proof of Theorem 4

Proof. As X is a compact polyhedron, it can be expressed as

$$X = \{x \in \Re^n : Ax - b \in \Re_-^{m_1} \times \{0_{m_2}\}\},$$

where $A \in \Re^{(m_1+m_2)\times n}, b \in \Re^{m_1+m_2}$ for some integers $m_1 \geq 0$ and $m_2 \geq 0$. The Lagrange functions of Problems (P) and (P_ε) are define by

$$L(x, \lambda, \mu) = h(x) + \lambda(p(x) - \alpha) + \langle \mu, A - b \rangle$$

and

$$L_\varepsilon(x, \lambda, \mu) = h(x) + \lambda(p_\varepsilon(x) - \alpha) + \langle \mu, A - b \rangle,$$

respectively, where

$$p_\varepsilon(x) = \frac{g_1(x,\varepsilon) - g_2(x)}{\varepsilon}.$$

Let Σ and Σ_ε be the sets of all stationary points of Problems (P) and (P_ε), respectively. Then, $\Sigma \subset X$ and $\Sigma_\varepsilon \subset X$ are both compact sets. For any $x \in \Sigma$, let $\Gamma(x)$ be the set of Lagrange multipliers of Problem (P) at x, i.e.,

$$\Gamma(x) = \{(\lambda,\mu) \in \Re^{1+m_1}_+ \times \Re^{m_2} : \nabla L(x,\lambda,\mu) = 0, \lambda(p(x)-\alpha) = 0, \langle \mu, Ax-b\rangle = 0\}.$$

For any $x \in \Sigma_\varepsilon$, let $\Gamma_\varepsilon(x)$ be the set of Lagrange multipliers of Problem (P_ε) at x, i.e.,

$$\Gamma_\varepsilon(x) = \{(\lambda,\mu) \in \Re^{1+m_1}_+ \times \Re^{m_2} : \nabla L_\varepsilon(x,\lambda,\mu) = 0, \lambda(p_\varepsilon(x)-\alpha) = 0, \langle \mu, Ax-b\rangle = 0\}.$$

Then, we have

$$\Lambda_0 = \{(x,\lambda) : x \in \Sigma, \exists \mu, (\lambda,\mu) \in \Gamma(x)\}, \quad \Lambda(\varepsilon) = \{(x,\lambda) : x \in \Sigma_\varepsilon, \exists \mu, (\lambda,\mu) \in \Gamma_\varepsilon(x)\}. \tag{26}$$

For any $x \in \Sigma$, it follows from Exercise 6.39 and Example 6.40 of Rockafellar and Wets (1998) that Condition (14) is Robinson constraint qualification, and it is Mangasarian-Fromovitz (MF) constraint qualification when X is a compact polyhedron. We have from Gauvin (1977) that the set of Lagrange multipliers of Problem (P) at x, $\Gamma(x)$ is bounded. As any element $(\lambda,\mu) \in \Gamma(x)$ is a solution of the following system

$$N(x)\begin{bmatrix} \lambda \\ \mu \end{bmatrix} + q(x) = 0,\ \lambda \geq 0,\ \mu_j \geq 0,\ j = 1,\ldots,m_1, \tag{27}$$

where

$$N(x) = \begin{bmatrix} \nabla p(x)^T & A^T \\ p(x)-\alpha & (Ax-b)^T \end{bmatrix},\ q(x) = \begin{bmatrix} \nabla f(x) \\ 0 \end{bmatrix}.$$

Since $N(x)$ and $q(x)$ are continuous in x, we have from the classical Hoffman lemma in Hoffman (1952) that for each point x' in an open neighborhood of x, $\Gamma(x')$, if nonempty, is bounded. It follows from the compactness of Σ, the set $\bigcup_{x\in\Sigma} \Gamma(x)$ is bounded. Therefore, it follows from (26) that Λ_0 is a compact set.

It follows from Remark 2.88 of Bonnans and Shapiro (2000) that for $\delta > 0$ small enough, MF constraint qualification of Problem (P) holds at every point $x \in \Omega_0 + \delta\mathbf{B}$. Having the same reason as the boundedness of $\bigcup_{x\in\Sigma} \Gamma(x)$, we have that the set $\bigcup_{x\in\Omega_0+\delta\mathbf{B}} \Gamma(x)$ is bounded.

From Lemma 3, we have that $\Omega(\varepsilon) \subset \Omega_0 + \delta\mathbf{B}$ for small $\varepsilon > 0$. If δ is small enough, we have from Remark 2.88 of Bonnans and Shapiro (2000) that MF constraint qualification of Problem (P_ε) holds at every point $x \in \Omega_0 + \delta\mathbf{B}$. Then, for any $\Omega_0 + \delta\mathbf{B}$, any point $(\lambda, \mu) \in \Gamma_\varepsilon(x)$ satisfies

$$N_\varepsilon(x) \left[\begin{array}{c} \lambda \\ \mu \end{array} \right] + q(x) = 0,\ \lambda \geq 0,\ \mu_j \geq 0,\ j = 1, \ldots, m_1, \tag{28}$$

where

$$N_\varepsilon(x) = \left[\begin{array}{cc} \nabla p_\varepsilon(x)^T & A^T \\ p_\varepsilon(x) - \alpha & (Ax - b)^T \end{array} \right].$$

It follows from Lemma 2 that $p_\varepsilon(x)$ and $\nabla p_\varepsilon(x)$ are small perturbations of $p(x)$ and $\nabla p(x)$, respectively, if $\varepsilon > 0$ small enough. Therefore, the system (28) is a linear system in (λ, μ) perturbated from the system (27) and we have from the classical Hoffman lemma that

$$\Gamma_\varepsilon(x) \subset \Gamma(x) + \gamma\overline{\mathbf{B}},$$

for some $\gamma > 0$ when $0 < \varepsilon \leq \varepsilon_0$, where $\varepsilon_0 > 0$ is some positive scalar and $\overline{\mathbf{B}} \subset \Re^{1+m_1+m_2}$ is the unit ball. Therefore, for $\varepsilon > 0$ small enough,

$$\bigcup_{x \in \Sigma_\varepsilon} \Gamma_\varepsilon(x) \subset \bigcup_{x \in \Omega_0 + \delta\mathbf{B}} \Gamma_\varepsilon(x) \subset \bigcup_{x \in \Omega_0 + \delta\mathbf{B}} [\Gamma(x) + \gamma\overline{\mathbf{B}}],$$

which, from (26) and the boundedness of $\bigcup_{x \in \Omega_0 + \delta\mathbf{B}} \Gamma(x)$, implies that $\Lambda(\varepsilon)$ is uniformly bounded for $\varepsilon > 0$ less than a small positive number ε_0. We have from Theorem 3 that $\limsup_{\varepsilon \searrow 0} \Lambda(\varepsilon) \subset \Lambda_0$, which implies that $\lim_{\varepsilon \searrow 0} \mathbb{D}(\Lambda(\varepsilon), \Lambda_0) = 0$ from the discussions in Example 4.13 of Rockafellar and Wets (1998). This concludes the proof of the theorem. □

D Proof of Property 3

We first prove the following two lemmas.

Lemma 5. *For Problem (DCP), let $\{y_k\} \subset X$ be a sequence convergent to $\bar{y} \in \Omega$ at which Slater's condition holds. Then,* $\lim_{k \to +\infty} \Omega_{y_k} = \Omega_{\bar{y}}$.[4]

[4]The terminologies of outer limit, inner limit and limit of a sequence of sets are given by Definition 4.1 of Rockafellar and Wets (1998). Here $\limsup_{k \to +\infty} \Omega_{y_k}$ is the outer limit of Ω_{y_k} which is defined by $\limsup_{k \to +\infty} \Omega_{y_k} := \{z : \exists k_j \to +\infty, \exists z_j \in \Omega_{y_{k_j}} \text{ such that } z_j \to z\}$, $\liminf_{k \to +\infty} \Omega_{y_k}$ is the inner limit of Ω_{y_k} which is defined by $\liminf_{k \to +\infty} \Omega_{y_k} := \{z : \exists z_k \in \Omega_{y_k} \text{ for all } k \text{ beyond some } \bar{k} \in \mathbf{N}, z_k \to z\}$; and if the outer limit and the inner limit are equal, we say $\lim_{k \to +\infty} \Omega_{y_k}$ exists and $\lim_{k \to +\infty} \Omega_{y_k} = \limsup_{k \to +\infty} \Omega_{y_k} = \liminf_{k \to +\infty} \Omega_{y_k}$.

Proof. Note that $\liminf_{k\to+\infty}\Omega_{y_k}\subset\limsup_{k\to+\infty}\Omega_{y_k}$. Then it suffices to prove that

$$\limsup_{k\to+\infty}\Omega_{y_k}\subset\Omega_{\bar{y}}\subset\liminf_{k\to+\infty}\Omega_{y_k}.$$

We first prove that $\limsup_{k\to+\infty}\Omega_{y_k}\subset\Omega_{\bar{y}}$. Let $z\in\limsup_{k\to+\infty}\Omega_{y_k}$, then there is a sequence $\{k_j\}\subset\mathbf{N}$ such that $z=\lim_{j\to\infty}z_j$ for some $z_j\in\Omega_{y_{k_j}}$. We have $z_j\in X$ and $g_1(z_j)-[g_2(y_{k_j})+\nabla g_2(y_{k_j})^T(z_j-y_{k_j})]\leq 0$. Letting $j\to+\infty$, we have $z\in X$ and $g_1(z)-[g_2(\bar{y})+\nabla g_2(\bar{y})^T(z-\bar{y})]\leq 0$, which implies $z\in\Omega_{\bar{y}}$. Therefore, $\limsup_{k\to+\infty}\Omega_{y_k}\subset\Omega_{\bar{y}}$.

We then prove that $\Omega_{\bar{y}}\subset\liminf_{k\to+\infty}\Omega_{y_k}$. Let $z\in\Omega_{\bar{y}}$, we only need to prove that there exists $\{z_j\}$ such that $z_j\in\Omega_{y_j}$ for j large enough and $z_j\to z$. Let $l_2(y,y')=g_2(y)+\nabla g_2(y)^T(y'-y)$,

$$\overline{G}(y')=\begin{bmatrix} y' \\ g_1(y')-l_2(\bar{y},y') \end{bmatrix},\quad G_j(y')=\begin{bmatrix} y' \\ g_1(y')-l_2(y_j,y') \end{bmatrix}$$

and

$$\mathcal{F}_{\overline{G}}(y')=\overline{G}(y')-X\times\Re_-,\quad \mathcal{F}_{G_j}(y')=G_j(y')-X\times\Re_-.$$

We choose $z\in\Omega_{\bar{y}}$. As $\bar{y}\in\Omega$ satisfies $\operatorname{int}\Omega_{\bar{y}}\neq\emptyset$, we have that Robinson constraint qualification of $\Omega_{\bar{y}}$ at z is satisfied, i.e.,

$$0\in\operatorname{int}\{\overline{G}(z)+\mathcal{J}\overline{G}(z)\Re^n-X\times\Re_-\}.$$

It follows from Theorem 2.89 of Bonnans and Shapiro (2000) that the mapping $\mathcal{F}_{\overline{G}}$ is metrically regular at $(z,0)$ at a rate $c>0$. Let $D(y')=G_j(y')-\overline{G}(y')$, namely

$$D(y')=\begin{bmatrix} 0 \\ l_2(y_j,y')-l_2(\bar{y},y') \end{bmatrix},$$

and for $y'\in X, y''\in X$

$$D(y')-D(y'')=\begin{bmatrix} 0 \\ (y'-y'')^T(\nabla g_2(y_j)-\nabla g_2(\bar{y})) \end{bmatrix}.$$

Therefore, we have

$$\|D(y')-D(y'')\|=|(y'-y'')^T(\nabla g_2(y_j)-\nabla g_2(\bar{y}))|\leq\|\nabla g_2(y_j)-\nabla g_2(\bar{y})\|\|y'-y''\|.$$

As $\nabla g_2(\cdot)$ is continuous at $\bar{y}$, $\|\nabla g_2(y_j)-\nabla g_2(\bar{y})\|$ can be arbitrarily small as j is large enough. Assume that j is large enough such that

$$\kappa_j:=\|\nabla g_2(y_j)-\nabla g_2(\bar{y})\|\leq\frac{1}{2c}.$$

We have from Theorem 2.84 of Bonnans and Shapiro (2000) that, when j is large enough, $\mathcal{F}_{G_j}$ is metrically regular at $(z, G_j(z) - \overline{G}(z))$. Namely for (z', η) close to $(z, G_j(z) - \overline{G}(z))$,

$$\operatorname{dist}(z', \mathcal{F}_{G_j}^{-1}(\eta)) \leq c_j \operatorname{dist}(G_j(z') - \eta, X \times \Re_-)$$

for the constant $c_j = 2c(1 - c\kappa_j)^{-1}$. As $c_j \in (0, 2c]$, for $\eta = 0$, when z' is close to z,

$$\operatorname{dist}(z', \Omega_{y_j}) \leq 2c \operatorname{dist}(G_j(z'), X \times \Re_-). \tag{29}$$

For any $y' \in X$, we have

$$\begin{aligned}
\left\|G_j(y') - \overline{G}(y')\right\| &= \left|l_2(y_j, y') - l_2(\bar{y}, y')\right| \\
&= \left|g_2(y_j) + \nabla g_2(y_j)^T(y' - y_j) - [g_2(\bar{y}) + \nabla g_2(\bar{y})^T(y' - \bar{y})]\right| \\
&= \Big|(g_2(y_j) - g_2(\bar{y})) + y'^T(\nabla g_2(y_j) - \nabla g_2(\bar{y})) \\
&\qquad -\nabla g_2(y_j)^T(y_j - \bar{y}) - \bar{y}^T\nabla g_2(y_j) - \nabla g_2(\bar{y}))\Big| \\
&\leq |g_2(y_j) - g_2(\bar{y})| + \left\|y'\right\| \|\nabla g_2(y_j) - \nabla g_2(\bar{y})\| \\
&\qquad + \|\nabla g_2(y_j)\| \|y_j - \bar{y}\| + \|\bar{y}\| \|\nabla g_2(y_j) - \nabla g_2(\bar{y})\| \\
&\leq |g_2(y_j) - g_2(\bar{y})| + 2M_0 \|\nabla g_2(y_j) - \nabla g_2(\bar{y})\| + M_1 \|y_j - \bar{y}\|, \qquad (30)
\end{aligned}$$

where $M_0 = \max\{\|y\| : y \in X\}$ and $M_1 = \max\{\|\nabla g_2(y)\| : y \in X\}$. By setting $z' = z$ in Equation (29) and by Equation (30), we obtain

$$\begin{aligned}
\operatorname{dist}(z, \Omega_{y_j}) &\leq 2c \operatorname{dist}(G_j(z), X \times \Re_-) \\
&\leq 2c \left\|G_j(z) - \overline{G}(z)\right\| \\
&\leq 2c \left[|g_2(y_j) - g_2(\bar{y})| + 2M_0\|\nabla g_2(y_j) - \nabla g_2(\bar{y})\| + M_1\|y_j - \bar{y}\|\right].
\end{aligned}$$

Therefore, for large j, there exists $z_j \in \Omega_{y_j}$ such that

$$\|z_j - z\| \leq 2c[|g_2(y_j) - g_2(\bar{y})| + 2M_0\|\nabla g_2(y_j) - \nabla g_2(\bar{y})\| + M_1\|y_j - \bar{y}\|],$$

which implies $z_j \to z$. Hence $\Omega_{\bar{y}} \subset \liminf_{j \to +\infty} \Omega_{y_j}$. This concludes the proof of the lemma. □

Lemma 6. *Let $\{x_k\}$ denote the sequence of solutions generated by Algorithm SCA for Problem (DCP) starting from $x_0 \in \Omega$. If $x_{k+1} = x_k$ at which Slater's condition holds, then x_k is a KKT point of Problem (DCP).*

Proof. As x_{k+1} is an optimal solution to CP(x_k), we have that x_k is the optimal solution of CP(x_k). Since CP(x_k) is a convex optimization problem and Slater's condition holds at x_k, we obtain

$$0 \in \nabla h(x_k) + [\nabla g_1(x_k) - \nabla g_2(x_k)] N_{\Re_-}(g_1(x_k) - g_2(x_k)) + N_X(x_k),$$

which implies that x_k is a KKT point of Problem (DCP). □

Now we can prove Property 3.

Proof. It follows from Property 1 that $\{x_k\} \subset \Omega$, and thus $\{x_k\}$ has cluster points and any cluster point $\bar{x} \in \Omega$. Since, from Property 2, $\{h(x_k)\}$ is non-increasing and convergent to $\inf_k\{h(x_k)\}$, we have that $h(\bar{x}) = \inf_k\{h(x_k)\}$.

Now we prove that $\bar{x}$ is an optimal solution of CP($\bar{x}$). Let $\{k_j\}$ satisfy that $x_{k_j} \to \bar{x}$. Let

$$\overline{h}(x) = h(x) + I_{\Omega_{\bar{x}}}(x), \quad \overline{h}_{k_j}(x) = h(x) + I_{\Omega_{x_{k_j}}}(x),$$

where

$$I_A(x) = \begin{cases} 0 & \text{if } x \in A \\ +\infty & \text{if } x \notin A \end{cases}.$$

It follows from Lemma 5 that $\Omega_{x_{k_j}} \to \Omega_{\bar{x}}$, and from Proposition 7.4(f) of Rockafellar and Wets (1998) that $I_{\Omega_{x_{k_j}}}(\cdot)$ epi-converges to $I_{\Omega_{\bar{x}}}$. Since $h(\cdot)$ is continuous, we have that $\overline{h}_{k_j}(\cdot)$ epi-converges to $\overline{h}(\cdot)$. It follows from Theorem 7.33 of Rockafellar and Wets (1998) that

$$\inf_x \overline{h}_{k_j}(x) \to \inf_x \overline{h}(x)$$

and

$$\limsup_{j \to +\infty} \operatorname{argmin} \overline{h}_{k_j} \subset \operatorname{argmin} \overline{h}. \tag{31}$$

Noting that argmin $\overline{h}_{k_j}$ coincides with the solution set of CP(x_{k_j}), which contains x_{k_j+1}, and argmin $\overline{h}$ coincides with the solution set of CP($\bar{x}$), which is nonempty and compact. From Equation (31) we have that x_{k_j+1} has a cluster point $\tilde{x} \in \Omega$, which is a point in argmin $\overline{h}$. Since $h(\tilde{x}) \le h(\bar{x})$ and $h(\bar{x}) = \inf_k h(x_k)$, we have that $h(\tilde{x}) = h(\bar{x})$. Noting again that $\bar{x} \in \Omega_{\bar{x}}$, we have that $\bar{x}$ is an optimal solution of CP($\bar{x}$). From Lemma 6, we have that $\bar{x}$ is a KKT point of Problem (DCP).

If h is strictly convex over $\mathcal{O}$, then argmin $\overline{h}_{k_j}$ is the singleton $\{x_{k_j+1}\}$ and argmin $\overline{h}$ is also singleton, say $\{\tilde{x}\}$. From Equation (31) we have that $x_{k_j+1} \to \tilde{x} \in \Omega$. Since $h(\tilde{x}) \le h(\bar{x})$ and $h(\bar{x}) = \inf_k h(x_k)$, we have that $h(\tilde{x}) = h(\bar{x})$. Once again, noting that the solution set of CP($\bar{x}$) is singleton, we obtain $\tilde{x} = \bar{x}$. Therefore, $x_{k_j+1} \to \bar{x}$. In a similar way, we can show that $x_{k_j+2} \to \bar{x}$,

$x_{k_j+3} \to \bar{x} \cdots$. Or equivalently, $x_{k_j+s} \to \bar{x}$ for all $s \in \mathbf{N}$. Therefore, $x_k \to \bar{x}$ and $\bar{x}$ is the solution of CP($\bar{x}$), and also a KKT point of Problem (DCP). □

References

Bazaraa, M. S., H. D. Sherali, and C. M. Shetty. 1993. *Nonlinear Programming: Theory and Algorithms*, 2nd edition, John Wiley & Sons, Inc, New York.

Bonnans, J. F. and A. Shapiro. 2000. *Perturbation Analysis of Optimization Problems*, Springer, New York.

Gauvin, J. 1977. A necessary and sufficient regular condition to bounded multipliers in nonconvex programming, *Mathematical Programming*, **12** 136-138.

Hoffman, A. 1952. On approximate solutions of systems of linear inequalities, *Journal of Research of the National Bureau of Standards, Section B, Mathematical Sciences*, **49** 263-265.

Rockafellar, R. T. 1970. *Convex Analysis.* Princeton University Press, Princeton, New Jersey.

Rockafellar, R. T. and R. J.-B. Wets. 1998. *Variational Analysis.* Springer-Verlag, New York.

Zangwill, W. I. 1969. *Nonlinear Programming: A Unified Approach*, Prentice Hall, Englewood Cliffs, New Jersey.