*IIE Transactions* (2013) **45**, 716–735

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/0740817X.2012.745205

# A smooth Monte Carlo approach to joint chance-constrained programs

ZHAOLIN HU[1], L. JEFF HONG[2,*] and LIWEI ZHANG[3]

[1]*School of Economics and Management, Tongji University, Shanghai 200092, China*
[2]*Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China*
*E-mail: hongl@ust.hk*
[3]*School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China*

Received June 2011 and accepted September 2012



This article studies Joint Chance-Constrained Programs (JCCPs). JCCPs are often non-convex and non-smooth and thus are generally challenging to solve. This article proposes a logarithm-sum-exponential smoothing technique to approximate a joint chance constraint by the difference of two smooth convex functions, and uses a sequential convex approximation algorithm, coupled with a Monte Carlo method, to solve the approximation. This approach is called *a smooth Monte Carlo approach* in this article. It is shown that the proposed approach is capable of handling both smooth and non-smooth JCCPs where the random variables can be either continuous, discrete, or mixed. The numerical experiments further confirm these findings.

**Keywords:** Joint chance-constrained program, Monte Carlo, stochastic optimization

## 1. Introduction

The use of optimization to support management decisions has become a common practice in the business world. In many industrial optimization problems, however, there are uncertainties associated with the set of constraints. The following are some examples.

1. To make production and inventory decisions, a manufacturer needs to maximize its profit while satisfying demands from customers. However, there are often uncertainties in the demands of the customers.
2. To make investment decisions, an agent needs to maximize her client's expected return while satisfying the client's cash requirements in the future. However, there are uncertainties in the returns of financial products.
3. To make resource allocation decisions, a planner needs to minimize the total allocation cost while meeting demands from different nodes. However, there are uncertainties in the demands of the nodes (Chen *et al.*, 2010).

To handle the uncertainties in the constraints, a natural approach is to require that all constraints be satisfied with a given high probability; e.g., 90%. The resulting optimization problem is called a Chance-Constrained Program (CCP). Its solution guarantees to satisfy the original constraints with the given probability. Note that the probability can also be viewed as the level of confidence. Therefore, this approach is also consistent with the typical approach of handling uncertainties used in statistical inferences.

In this article, we consider the CCP represented as follows:

$$\begin{aligned} \underset{\mathbf{x}\in X}{\text{minimize}} \quad & h(\mathbf{x}), \\ \text{subject to} \quad & \Pr\{c_1(\mathbf{x}, \boldsymbol{\xi}) \le 0, \ldots, c_m(\mathbf{x}, \boldsymbol{\xi}) \le 0\} \ge 1 - \alpha, \end{aligned} \tag{1}$$

where $\mathbf{x}$ is a $d$-dimensional vector of decision variables; $\boldsymbol{\xi}$ is a $k$-dimensional vector of uncertain parameters, and the support of $\boldsymbol{\xi}$, denoted as $\Xi$, is a closed subset of $\Re^k$; $X$ is a subset of $\Re^d$; $h : \Re^d \to \Re$ and $c_i : \Re^d \times \Xi \to \Re$, $i = 1, \ldots, m$, are real-valued functions. Moreover, throughout this article we assume that $X$ is a convex and compact set, which may be defined by some deterministic constraints, and the functions $h$ and $c_i$, $i = 1, \ldots, m$, are convex and continuously differentiable in $\mathbf{x}$ for every $\boldsymbol{\xi} \in \Xi$. Problem (1) is called a Single CCP (SCCP) if $m = 1$ and a Joint CCP (JCCP) if $m > 1$.

The literature on CCPs can be dated back to Charnes *et al.* (1958), who first considered an SCCP, and Miller and Wagner (1965), who first considered a JCCP. Since then, both theories and applications of CCPs have been studied extensively. For a comprehensive literature review on the

*Corresponding author

topic, readers are referred to Prékopa (2003), Nemirovski and Shapiro (2006), Hong *et al.* (2011), and references therein.

There are three major difficulties in solving a CCP. First, the chance constraint is typically difficult to evaluate. For instance, even when the original constraints are linear in $\mathbf{x}$ and the uncertain parameters follow a multivariate normal distribution, the joint chance constraint in general may not have a closed-form expression. Second, the chance constraint does not necessarily preserve the convexity of the original constraints. Even though the original constraints define a convex set for any given $\boldsymbol{\xi} \in \Xi$, the set defined by the chance constraint may not be convex. It is provably convex only under very restrictive conditions (see, e.g., Prékopa (2003) and Nemirovski and Shapiro (2006)). Third, the chance constraint does not necessarily preserve the smoothness of the original constraints. For instance, when $\boldsymbol{\xi}$ follows a discrete distribution or $m > 1$, the left-hand side of the chance constraint is often a non-smooth function of $\mathbf{x}$. Moreover, it is often difficult to know *a priori* whether a CCP is smooth.

A number of general approaches have been proposed in the literature to solve CCPs. In this article we introduce two of the most popular ones. The first one is the so-called convex conservative approximations, which include the quadratic approximation of Ben-Tal and Nemirovski (2000), the Bernstein approximation of Nemirovski and Shapiro (2006), the second-order conic program approximation of Chen *et al.* (2010), and the Conditional Value-at-Risk (CVaR) approximation of Rockafellar and Uryasev (2000). This approach seeks to find a convex subset of the (possibly non-convex) feasible set and finds the optimal solution in the subset. It handles the three difficulties at the same time by choosing a convex subset that is defined by a set of constraints that are easy to evaluate, convex, and smooth. These approximations are often significantly easier to solve than the CCPs and their solutions are guaranteed to be feasible for the CCPs. However, these solutions do not satisfy any optimality conditions of the CCPs and their qualities are hard to quantify. Among these convex conservative approximations, the CVaR approximation is known as the best because the feasible regions defined by all other approximations are known to be subsets of the feasible region defined by the CVaR approximation. Furthermore, these approximations are often designed to solve SCCPs. To handle JCCPs, probabilistic inequalities (e.g., Bonferroni's inequality) have to be used to break a joint chance constraint into multiple single chance constraints, which often makes the approximations even more conservative. The second approach is the so-called scenario approach (see, for instance, De Farias and Van Roy (2004) and Calafiore and Campi (2005, 2006)). It replaces the chance constraint in Problem (1) by a set of constraints $c_i(\mathbf{x}, \xi_\ell) \leq 0, i = 1, \ldots, m, \ell = 1, \ldots, n$, where $\{\xi_1, \ldots, \xi_n\}$ is a collection of independent scenarios of $\boldsymbol{\xi}$ and $n$ needs to be determined carefully to ensure the probability requirement. The new problem under the scenario approach is easier to solve because $c_i$ is convex, smooth, and easy to evaluate. However, Nemirovski and Shapiro (2006) and Hong *et al.* (2011) found that the solutions of the approach can be drastically different when different collections of scenarios are used and the solutions are in general very conservative.

Recently, Hong *et al.* (2011) proposed an $\varepsilon$-approximation approach, which reformulated a JCCP into a Difference of Convex (DC) program, and used an $\varepsilon$-approximation together with a Sequential Convex Approximation (SCA) algorithm to solve it. In each iteration of the SCA algorithm, the approach applies a gradient-based Monte Carlo method to solve a convex stochastic program. They showed that, under some technical conditions, the solutions found by their approach converge to the set of Karush–Kuhn–Tucker (KKT) points of the JCCP. To the best of our knowledge, the $\varepsilon$-approximation approach is the only approach in the literature that is capable of solving general JCCPs while guaranteeing certain optimality conditions.

Although the $\varepsilon$-approximation approach has appealing properties, it is designed to solve only smooth JCCPs. As we pointed out earlier in this article, the non-smoothness is a major difficulty of JCCPs. Therefore, by assuming that JCCPs are smooth, the $\varepsilon$-approximation approach avoids this difficulty but also significantly limits its applicability in practice.

In this article, based on the $\varepsilon$-approximation approach, we propose a Smooth Monte Carlo (SMC) approach that handles the third difficulty as well as the first two. In this approach, we no longer solve an $\varepsilon$-approximation. Instead, we smooth the $\varepsilon$-approximation using a logarithm-sum-exponential function. This results in a new smooth DC program even when the JCCP is not smooth. Furthermore, we show that once the $\varepsilon$-approximation is smoothed, the parameter $\varepsilon$ can be treated as a decision variable, often resulting in a better solution than the original $\varepsilon$-approximation. Similar to Hong *et al.* (2011), we propose to use an SCA algorithm to solve the smooth DC program and use a Monte Carlo method to solve the convex subproblem in each iteration of the SCA algorithm. We show that the SMC approach is a conservative approximation of the original JCCP and analyze its convergence under some technical conditions. Note that "conservative approximation" throughout this article means that the feasible region of the approximation problem is a subset of the original problem, which is consistent with the stochastic optimization literature.

Smooth approximations to non-smooth optimization problems have been studied extensively in the literature of non-linear optimization. Readers interested in this topic are referred to, e.g., Bertsekas (1975), Fukushima and Qi (1998), and Nesterov (2005) for comprehensive reviews. In the stochastic optimization context, Alexander *et al.* (2006) proposed a smoothing technique to handle CVaR optimization problems. They compared their approach with the well-known linear approach of Rockafellar and Uryasev

(2000) and found that their approach is computationally much more efficient than the linear approach. Inspired by Alexander *et al.* (2006), Xu and Zhang (2009) proposed a smooth scheme for the stochastic programs where the nonsmooth objective function is in an expectation form and proved the convergence of the smoothed sample-average approximation.

The rest of this article is organized as follows. In Section 2 we introduce the $\varepsilon$-approximation of Hong *et al.* (2011) and discuss its limitations. In Section 3 we propose a logarithm-sum-exponential smoothing technique to approximate a JCCP as a smooth DC program and show that this approach guarantees the desired convergence properties. Furthermore, we show that the smooth DC program can be strengthened by treating the parameter $\varepsilon$ as a decision variable. We implement Monte Carlo methods and propose an SCA algorithm to solve the smooth DC program in Section 4. The results of some numerical experiments are reported in Section 5. We conclude the article in Section 6 and include some lengthy proofs in the Appendix.

## 2. Background

Let us consider the joint chance constraint:

$$\Pr\{c_1(\mathbf{x}, \boldsymbol{\xi}) \le 0, \ldots, c_m(\mathbf{x}, \boldsymbol{\xi}) \le 0\} \ge 1 - \alpha.$$

Clearly, simultaneously handling multiple constraints $c_1(\mathbf{x}, \boldsymbol{\xi}) \le 0, \ldots, c_m(\mathbf{x}, \boldsymbol{\xi}) \le 0$ is difficult. One approach is to implement probabilistic inequalities (e.g., Bonferroni's inequality) to break the joint chance constraint into a number of single chance constraints (see, e.g., Nemirovski and Shapiro (2006)). However, this typically results in a conservative approximation of the original joint chance constraint since the probabilistic inequalities are generally not tight. Another approach is to use a maximum operator $c(\mathbf{x}, \boldsymbol{\xi}) = \max\{c_1(\mathbf{x}, \boldsymbol{\xi}), \ldots, c_m(\mathbf{x}, \boldsymbol{\xi})\}$. Then, it is easy to see that

$$\Pr\{c_1(\mathbf{x}, \boldsymbol{\xi}) \le 0, \ldots, c_m(\mathbf{x}, \boldsymbol{\xi}) \le 0\} = \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) \le 0\}.$$

Note that $c(\mathbf{x}, \boldsymbol{\xi})$ is convex in $\mathbf{x}$ for any $\boldsymbol{\xi} \in \Xi$. Therefore, we can convert a JCCP to an SCCP in this way without destroying the convexity structure embedded in the probability function. Then, Problem (1) can be rewritten as

$$\begin{aligned} \underset{\mathbf{x}\in X}{\text{minimize}} \quad & h(\mathbf{x}), \\ \text{subject to} \quad & \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\} \le \alpha. \end{aligned} \tag{2}$$

Although the approach of using a maximum operator converts a JCCP to an SCCP, it also introduces the additional difficulty that $c(\mathbf{x}, \boldsymbol{\xi})$ may no longer be smooth in $\mathbf{x}$ even though $c_i(\mathbf{x}, \boldsymbol{\xi})$, $i = 1, \ldots, m$ are smooth. In this article we mainly focus on this issue.

### 2.1. *$\varepsilon$-approximation*

We first introduce the $\varepsilon$-approximation of Hong *et al.* (2011) in this subsection and discuss its limitations in next subsection. Let

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

where $[z]^+ = \max\{z, 0\}$. Hong *et al.* (2011) showed that $\inf_{t>0} E[\pi(c(\mathbf{x}, \boldsymbol{\xi}), t)] = \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) \ge 0\}$ and suggested solving:

$$\begin{aligned} \underset{\mathbf{x}\in X}{\text{minimize}} \quad & h(\mathbf{x}), \\ \text{subject to} \quad & \inf_{t>0} E[\pi(c(\mathbf{x}, \boldsymbol{\xi}), t)] \le \alpha. \end{aligned} \tag{3}$$

Let $\Omega_0$ and $\Omega$ denote the feasible sets of Problem (2) and Problem (3), respectively. Note that $\Pr\{c(\mathbf{x}, \boldsymbol{\xi}) \ge 0\} = \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\} + \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) = 0\}$. Then, $\Omega \subset \Omega_0$. Therefore, Problem (3) in general is a conservative approximation of Problem (2).

Furthermore, Hong *et al.* (2011) make the following assumption.

**Assumption 1.** For any $\mathbf{x} \in X$, $\Pr\{c(\mathbf{x}, \boldsymbol{\xi}) = 0\} = 0$.

Then, when Assumption 1 is satisfied, Problems (2) and (3) are equivalent and one only needs to solve Problem (3) to find the optimal solutions of Problem (2).

Note that $E[\pi(c(\mathbf{x}, \boldsymbol{\xi}), t_1)] \le E[\pi(c(\mathbf{x}, \boldsymbol{\xi}), t_2)]$ when $0 < t_1 \le t_2$ and

$$\inf_{t>0} E[\pi(c(\mathbf{x}, \boldsymbol{\xi}), t)] = \lim_{t \searrow 0} E[\pi(c(\mathbf{x}, \boldsymbol{\xi}), t)],$$

where $t \searrow 0$ denotes that $t$ decreasingly goes to zero. Therefore, a natural approach to approximating Problem (3) is to fix $t = \varepsilon$ for some $\varepsilon > 0$ and solve

$$\begin{aligned} \underset{\mathbf{x}\in X}{\text{minimize}} \quad & h(\mathbf{x}), \\ \text{subject to} \quad & E[\pi(c(\mathbf{x}, \boldsymbol{\xi}), \varepsilon)] \le \alpha. \end{aligned} \tag{4}$$

Hong *et al.* (2011) call Problem (4) an $\varepsilon$-approximation. Let $\Omega_\varepsilon$ denote the feasible set of Problem (4). Note that

$$E[\pi(c(\mathbf{x}, \boldsymbol{\xi}), \varepsilon)] \ge \lim_{t \searrow 0} E[\pi(c(\mathbf{x}, \boldsymbol{\xi}), t)].$$

Then, $\Omega_\varepsilon \subset \Omega \subset \Omega_0$. Therefore, Problem (4) is also a conservative approximation to Problem (2) no matter whether or not Assumption 1 is satisfied.

To solve Problem (4), we let

$$\begin{aligned} g_1(\mathbf{x}, \varepsilon) &= E[[c(\mathbf{x}, \boldsymbol{\xi}) + \varepsilon]^+] - \alpha\varepsilon, \\ g_2(\mathbf{x}) &= E[[c(\mathbf{x}, \boldsymbol{\xi})]^+]. \end{aligned}$$

Then, Problem (4) can be written as the following optimization problem:

$$\begin{aligned} \underset{\mathbf{x}\in X}{\text{minimize}} \quad & h(\mathbf{x}), \\ \text{subject to} \quad & g(\mathbf{x}, \varepsilon) := g_1(\mathbf{x}, \varepsilon) - g_2(\mathbf{x}) \le 0. \end{aligned} \tag{5}$$

Note that both $g_1(\mathbf{x}, \varepsilon)$ and $g_2(\mathbf{x})$ are convex functions of $\mathbf{x}$. Therefore, $g(\mathbf{x}, \varepsilon)$ takes the form of the difference of two convex functions, which is known as a DC function, and Problem (5) is known as a DC program.

Hong *et al.* (2011) proposed to solve Problem (5) using an SCA algorithm. The algorithm starts with a feasible solution and, in each iteration (say iteration $k$), it solves

$$\begin{aligned} \underset{\mathbf{x}\in X}{\text{minimize}} \quad & h(\mathbf{x}), \\ \text{subject to} \quad & g_1(\mathbf{x}, \varepsilon) - [g_2(\mathbf{x}_{k-1}) + \nabla_x g_2(\mathbf{x}_{k-1})^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_{k-1})] \le 0, \end{aligned} \tag{6}$$

where $\mathbf{x}_{k-1}$ is the optimal solution of the problem in iteration $k-1$. Because it uses a first-order Taylor expansion at $\mathbf{x}_{k-1}$ to approximate the convex function $g_2(\mathbf{x})$ in Problem (5), Problem (6) is a convex optimization problem. Furthermore, Hong *et al.* (2011) showed that, under certain conditions, the sequence of the solutions generated by the SCA algorithm converges to the set of KKT points of Problem (5), which also converges to the set of KKT points of Problem (2) as $\varepsilon \to 0$. To actually solve Problem (6), Hong *et al.* (2011) proposed a gradient-based Monte Carlo method and also analyzed its convergence.

### 2.2. *Limitations of the ε-approximation*

Although the $\varepsilon$-approximation approach has the desired convergence property that is often missed by the algorithms in the JCCP literature, it also has some limitations. One of the major limitations is that the approach is designed to solve only smooth JCCPs. The theoretical framework and the SCA algorithm of this approach critically depend on the following two assumptions on the smoothness of $c(\mathbf{x}, \boldsymbol{\xi}) = \max\{c_1(\mathbf{x}, \boldsymbol{\xi}), \ldots, c_m(\mathbf{x}, \boldsymbol{\xi})\}$:

1. $c(\mathbf{x}, \boldsymbol{\xi})$ is a continuous random variable with a continuous density function for any $\mathbf{x} \in X$, and
2. $c(\mathbf{x}, \boldsymbol{\xi})$ is differentiable with respect to $\mathbf{x}$ with probability one at any $\mathbf{x} \in X$.

However, these two assumptions are often violated for practical problems. When $\boldsymbol{\xi}$ has a discrete distribution, $c(\mathbf{x}, \boldsymbol{\xi})$ is typically a discrete random variable, which violates the first assumption and often the second as well. When $\boldsymbol{\xi}$ has a continuous distribution, the second assumption may also be violated. Note that $c(\mathbf{x}, \boldsymbol{\xi}) = \max\{c_1(\mathbf{x}, \boldsymbol{\xi}), \ldots, c_m(\mathbf{x}, \boldsymbol{\xi})\}$ is typically not smooth for any fixed $\boldsymbol{\xi}$. The second assumption basically requires that the probability of being at the kink points is zero. For instance, when $c(\mathbf{x}, \boldsymbol{\xi}) = \max\{c_1(\mathbf{x}, \boldsymbol{\xi}), c_2(\mathbf{x}, \boldsymbol{\xi})\}$, $c(\mathbf{x}, \boldsymbol{\xi})$ is typically not differentiable at $\mathbf{x}$ where $c_1(\mathbf{x}, \boldsymbol{\xi}) = c_2(\mathbf{x}, \boldsymbol{\xi})$. Therefore, a condition often used to ensure the assumption is to have $\Pr\{c_1(\mathbf{x}, \boldsymbol{\xi}) = c_2(\mathbf{x}, \boldsymbol{\xi})\} = 0$. However, even for the simple case where $c_1(\mathbf{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}_1^{\mathrm{T}}\mathbf{x}$ and $c_2(\mathbf{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}_2^{\mathrm{T}}\mathbf{x}$, $\Pr\{c_1(\mathbf{x}, \boldsymbol{\xi}) = c_2(\mathbf{x}, \boldsymbol{\xi})\} = 1$ at $\mathbf{x} = 0$ and the assumption is violated. Furthermore, even when the assumptions are satisfied, it is often difficult to know *a priori*.

When the JCCP is not smooth, there exist a number of issues on the theory and implementation of the $\varepsilon$-approximation. First, KKT conditions and KKT points are only defined for smooth non-linear optimization problems. Therefore, when the JCCP is not smooth, the optimality conditions for smooth optimization are no longer applicable. Second, the SCA algorithm requires linearizing the convex function $g_2(\mathbf{x})$ using a first-order Taylor expansion. However, if $g_2(\mathbf{x})$ is not differentiable, the SCA algorithm can no longer be implemented. Third, the SCA algorithm requires solving a sequence of convex optimization problems that are typically non-smooth when either $g_1(\mathbf{x}, \varepsilon)$ or $g_2(\mathbf{x})$ are non-smooth. However, this may create significant difficulties because non-smooth optimization problems are in general much more difficult and slower to solve than their smooth counterparts and, especially, as we know, in the stochastic optimization context the use of non-smooth techniques is still relatively scarce.

In this article we handle the non-smoothness of JCCPs by a smooth approximation approach. This approach handles the three issues simultaneously. Furthermore, it provides an opportunity for us to further treat $\varepsilon$ as a decision variable that may significantly improve the performance of the SCA algorithm.

## 3. A smooth approximation to a JCCP

In this section, we propose a logarithm-sum-exponential smoothing approach to smooth the $\varepsilon$-approximation, study its convergence properties, and show how it may be improved by treating $\varepsilon$ as a decision variable.

### 3.1. *Logarithm-sum-exponential smoothing to maximum function*

Note that $g_1(\mathbf{x}, \varepsilon)$ and $g_2(\mathbf{x})$ may be written as

$$\begin{aligned} g_1(\mathbf{x}, \varepsilon) &= E[[c(\mathbf{x}, \boldsymbol{\xi}) + \varepsilon]^+] - \alpha\varepsilon \\ &= E[\max\{0, c_1(\mathbf{x}, \boldsymbol{\xi}) + \varepsilon, \ldots, c_m(\mathbf{x}, \boldsymbol{\xi}) + \varepsilon\}] - \alpha\varepsilon, \end{aligned} \tag{7}$$

$$\begin{aligned} g_2(\mathbf{x}) &= E[[c(\mathbf{x}, \boldsymbol{\xi})]^+] \\ &= E[\max\{0, c_1(\mathbf{x}, \boldsymbol{\xi}), \ldots, c_m(\mathbf{x}, \boldsymbol{\xi})\}], \end{aligned} \tag{8}$$

respectively. Therefore, it is clear that the non-smoothness of Problem (5) is caused by the maximum operator. In this subsection we use a logarithm-sum-exponential function to smooth the maximum operator.

Let $a = \max\{a_1, \ldots, a_m\}$. Then $[a]^+ = \max\{0, a_1, \ldots, a_m\}$. It is not difficult to show that for any $\mu > 0$ (see, e.g., Rockafellar and Wets (1998) and Boyd and Vandenberghe (2004)):

$$[a]^+ \le \mu \log\left[1 + \sum_{i=1}^{m} \exp(\mu^{-1} a_i)\right] \le [a]^+ + \mu \log(m+1), \tag{9}$$

where the second inequality is tight when $a_1 = \cdots = a_m = 0$. We call $\mu \log[1 + \sum_{i=1}^m \exp(\mu^{-1} a_i)]$ a *logarithm-sum-exponential approximation* of $[a]^+$.

To analyze the logarithm-sum-exponential approximation to $g_1(\mathbf{x}, \varepsilon)$ and $g_2(\mathbf{x})$, we consider a general vector-valued random function $\mathbf{F}(\mathbf{x}, \xi) = (f_1(\mathbf{x}, \xi), \ldots, f_m(\mathbf{x}, \xi)) : \Re^d \times \Xi \to \Re^m$ with each $f_i$ being a real-valued function. Let $f(\mathbf{x}, \xi) = \max\{f_1(\mathbf{x}, \xi), \ldots, f_m(\mathbf{x}, \xi)\}$. Then, we use the function

$$H(\mathbf{x}, \xi, \mu) := \mu \log\left[1 + \sum_{i=1}^m \exp\{\mu^{-1} f_i(\mathbf{x}, \xi)\}\right]$$

to smooth the function $[f(\mathbf{x}, \xi)]^+$ and consequently use

$$\Psi(\mathbf{x}, \mu) := E[H(\mathbf{x}, \xi, \mu)]$$

to smooth the expectation $E[[f(\mathbf{x}, \xi)]^+]$. One of the merits of the logarithm-sum-exponential smoothing approach is that it has very nice properties, which we summarize in the following proposition.

**Proposition 1.** *Suppose that $f_i(\mathbf{x}, \xi)$, $i = 1, \ldots, m$ are convex in $\mathbf{x}$ for every $\xi \in \Xi$. Then $H(\mathbf{x}, \xi, \mu)$ is jointly convex in $(\mathbf{x}, \mu)$ and non-decreasing in $\mu$ for every $\xi \in \Xi$, $\Psi(\mathbf{x}, \mu)$ is jointly convex in $(\mathbf{x}, \mu)$ and non-decreasing in $\mu$, and for any $\mu > 0$:*

$$E[[f(\mathbf{x}, \xi)]^+] \le \Psi(\mathbf{x}, \mu) \le E[[f(\mathbf{x}, \xi)]^+] + \mu \log(m + 1).$$

*Moreover,*

$$E[[f(\mathbf{x}, \xi)]^+] = \inf_{\mu > 0} \Psi(\mathbf{x}, \mu).$$

**Proof.** First, it is known that the logarithm-sum-exponential function $\log[1 + \sum_{i=1}^m \exp(a_i)]$ is convex in $(a_1, \ldots, a_m)^{\mathrm{T}}$. Since $\mu \log[1 + \sum_{i=1}^m \exp(\mu^{-1} a_i)]$ is the perspective function of the logarithm-sum-exponential function and the perspective operation preserves the convexity (for function $f : \Re^n \to \Re$, the perspective of $f$ is the function $g : \Re^{n+1} \to \Re$ defined by $g(\mathbf{x}, t) = tf(\mathbf{x}/t)$ with domain $\mathrm{dom} g = \{(\mathbf{x}, t) \mid \mathbf{x}/t \in \mathrm{dom} f, t > 0\}$; see, e.g., Section 3.2.6 of Boyd and Vandenberghe (2004)), we have $\mu \log[1 + \sum_{i=1}^m \exp(\mu^{-1} a_i)]$ is jointly convex in $(a_1, \ldots, a_m, \mu)^{\mathrm{T}}$. Moreover, $\mu \log[1 + \sum_{i=1}^m \exp(\mu^{-1} a_i)]$ is increasing in each component $a_i$, $i = 1, \ldots, m$. Note also that for every $\xi \in \Xi$, $f_i(\mathbf{x}, \xi)$, $i = 1, \ldots, m$ are convex in $\mathbf{x}$. Using the composition rules of convex functions (Boyd and Vandenberghe, 2004), we have that for every $\xi \in \Xi$, $\mu \log[1 + \sum_{i=1}^m \exp\{\mu^{-1} f_i(\mathbf{x}, \xi)\}]$ (i.e., $H(\mathbf{x}, \xi, \mu)$) is convex in $(\mathbf{x}, \mu)$. It follows that $\Psi(\mathbf{x}, \mu)$ is convex in $(\mathbf{x}, \mu)$.

Note that from Equation (9), we have for every $\xi \in \Xi$ and $\mu > 0$

$$[f(\mathbf{x}, \xi)]^+ \le H(\mathbf{x}, \xi, \mu) \le [f(\mathbf{x}, \xi)]^+ + \mu \log(m + 1).$$

This immediately implies $E[[f(\mathbf{x}, \xi)]^+] \le \Psi(\mathbf{x}, \mu) \le E[[f(\mathbf{x}, \xi)]^+] + \mu \log(m + 1)$.

To prove the monotonicity of $H(\mathbf{x}, \xi, \mu)$ and $\Psi(\mathbf{x}, \mu)$ in $\mu$, we only need to prove the function $\mu \log[1 + \sum_{i=1}^m \exp(\mu^{-1} a_i)]$ is non-decreasing in $\mu$ for every $(a_1, \ldots, a_m)^{\mathrm{T}}$. For any $(a_1, \ldots, a_m)^{\mathrm{T}}$ and $\mu > 0$, note that

$$\begin{aligned}
&\frac{\partial}{\partial \mu}\left\{\mu \log\left[1 + \sum_{i=1}^m \exp(\mu^{-1} a_i)\right]\right\} \\
&\quad = \log\left[1 + \sum_{i=1}^m \exp(\mu^{-1} a_i)\right] - \frac{\sum_{i=1}^m (\mu^{-1} a_i) \exp(\mu^{-1} a_i)}{1 + \sum_{i=1}^m \exp(\mu^{-1} a_i)} \\
&\quad \ge \log\left[1 + \sum_{i=1}^m \exp(\mu^{-1} a_i)\right] \\
&\qquad - \frac{[\max_{i=1,\ldots,m}\{\mu^{-1} a_i\}]^+ \sum_{i=1}^m \exp(\mu^{-1} a_i)}{1 + \sum_{i=1}^m \exp(\mu^{-1} a_i)} \\
&\quad \ge \log\left[1 + \sum_{i=1}^m \exp(\mu^{-1} a_i)\right] - \left[\max_{i=1,\ldots,m}\{\mu^{-1} a_i\}\right]^+ \\
&\quad \ge 0,
\end{aligned}$$

where the last inequality is due to Equation (9). Therefore, $\mu \log[1 + \sum_{i=1}^m \exp(\mu^{-1} a_i)]$ is non-decreasing in $\mu$. This implies that $H(\mathbf{x}, \xi, \mu)$ and $\Psi(\mathbf{x}, \mu)$ are non-decreasing in $\mu$. Finally, by the monotone convergence theorem (Durrett, 2005), we obtain

$$\begin{aligned}
\inf_{\mu > 0} \Psi(\mathbf{x}, \mu) &= E[\inf_{\mu > 0} H(\mathbf{x}, \xi, \mu)] \\
&= E[\max\{0, f_1(\mathbf{x}, \xi), \ldots, f_m(\mathbf{x}, \xi)\}] = E[[f(\mathbf{x}, \xi)]^+].
\end{aligned}$$

This concludes the proof of the proposition. ■

Proposition 1 provides a foundation for our smooth approximation. It shows that we may bound the (possibly) non-smooth functions $g_1(\mathbf{x}, \varepsilon)$ and $g_2(\mathbf{x})$ from both above and below using smooth functions. Furthermore, these approximations preserve the convexity of the original functions and the approximation errors can be bounded explicitly.

### 3.2. *Smooth approximation to the ε-approximation*

Now we return to Problems (1) and (5) and show how to use the logarithm-sum-exponential functions to approximate both $g_1(\mathbf{x}, \varepsilon)$ and $g_2(\mathbf{x})$ to obtain a smooth approximation to Problem (1). For any fixed $\varepsilon$, we let.

$$\begin{aligned}
H_1(\mathbf{x}, \varepsilon, \xi, \mu) &= \mu \log\left[1 + \sum_{i=1}^m \exp\{\mu^{-1}(c_i(\mathbf{x}, \xi) + \varepsilon)\}\right], \\
\Psi_1(\mathbf{x}, \varepsilon, \mu) &= E\left[H_1(\mathbf{x}, \varepsilon, \xi, \mu)\right], \\
H_2(\mathbf{x}, \xi, \mu) &= \mu \log\left[1 + \sum_{i=1}^m \exp\{\mu^{-1} c_i(\mathbf{x}, \xi)\}\right], \\
\Psi_2(\mathbf{x}, \mu) &= E\left[H_2(\mathbf{x}, \xi, \mu)\right].
\end{aligned}$$

Then, for any $\mu > 0$, both $H_1(\mathbf{x}, \varepsilon, \xi, \mu)$ and $H_2(\mathbf{x}, \xi, \mu)$ are continuously differentiable in $\mathbf{x}$ for every $\xi \in \Xi$. Let

$$\begin{aligned}
\bar{g}_1(\mathbf{x}, \varepsilon, \mu) &= \Psi_1(\mathbf{x}, \varepsilon, \mu) - \alpha\varepsilon \quad \text{and} \\
\bar{g}_2(\mathbf{x}, \mu) &= \Psi_2(\mathbf{x}, \mu) - \mu \log(m + 1).
\end{aligned}$$

By Equations (7) and (8) and Proposition 1, we have

$$g_1(\mathbf{x}, \varepsilon) \le \bar{g}_1(\mathbf{x}, \varepsilon, \mu) \le g_1(\mathbf{x}, \varepsilon) + \mu \log(m+1),$$
$$g_2(\mathbf{x}) - \mu \log(m+1) \le \bar{g}_2(\mathbf{x}, \mu) \le g_2(\mathbf{x}),$$

which implies that

$$\begin{aligned} g_1(\mathbf{x}, \varepsilon) - g_2(\mathbf{x}) &\le \bar{g}_1(\mathbf{x}, \varepsilon, \mu) - \bar{g}_2(\mathbf{x}, \mu) \\ &\le g_1(\mathbf{x}, \varepsilon) - g_2(\mathbf{x}) + 2\mu \log(m+1). \end{aligned} \tag{10}$$

Equation (10) serves as the basis for our smooth approximation. It suggests that we can use the following smooth optimization problem to approximate Problem (5).

$$\begin{aligned} (\mathrm{P}_\mu) \quad & \underset{\mathbf{x}\in X}{\text{minimize}} \quad h(\mathbf{x}), \\ & \text{subject to} \quad \bar{g}(\mathbf{x}, \varepsilon, \mu) := \bar{g}_1(\mathbf{x}, \varepsilon, \mu) - \bar{g}_2(\mathbf{x}, \mu) \le 0. \end{aligned}$$

From Proposition 1 we see that both $\bar{g}_1(\mathbf{x}, \varepsilon, \mu)$ and $\bar{g}_2(\mathbf{x}, \mu)$ are convex in $\mathbf{x}$. Therefore, Problem $(\mathrm{P}_\mu)$ is also a DC program. To analyze the smoothness of the problem, we make the following assumption, which is a standard assumption in stochastic optimization and is also a very weak assumption in practical situations (see, e.g., Broadie and Glasserman (1996), Hong and Liu (2009), and Shapiro *et al.* (2009)).

**Assumption 2.** There exist random functions $M_i(\xi) > 0$ and $K_i(\xi)$ with $E[M_i(\xi)] < \infty$ and $E[K_i(\xi)] < \infty$ such that $|c_i(\mathbf{x}, \xi)| \le M_i(\xi)$ and $\|\nabla_x c_i(\mathbf{x}, \xi)\| \le K_i(\xi)$ for all $\mathbf{x} \in X$, $\xi \in \Xi$, and $i = 1, \ldots, m$.

Let $M(\xi) = \sum_{i=1}^m M_i(\xi)$. By Assumption 2, we have $E[M(\xi)] < \infty$ and

$$|H_1(\mathbf{x}, \varepsilon, \xi, \mu)| \le M(\xi) + \varepsilon + \mu \log(m+1),$$
$$|H_2(\mathbf{x}, \xi, \mu)| \le M(\xi) + \mu \log(m+1)$$

for all $\mathbf{x} \in X$ and $\xi \in \Xi$. Similarly, let $K(\xi) = \sum_{i=1}^m K_i(\xi)$. We have $E[K(\xi)] < \infty$ and

$$\begin{aligned} &|H_1(\mathbf{x}_1, \varepsilon, \xi, \mu) - H_1(\mathbf{x}_2, \varepsilon, \xi, \mu)| \\ &\quad \le \sum_{i=1}^m \sup_{\mathbf{x}\in X} \|\nabla_x c_i(\mathbf{x}, \xi)\| \|\mathbf{x}_1 - \mathbf{x}_2\| \le K(\xi) \|\mathbf{x}_1 - \mathbf{x}_2\|, \\ &|H_2(\mathbf{x}_1, \xi, \mu) - H_2(\mathbf{x}_2, \xi, \mu)| \\ &\quad \le \sum_{i=1}^m \sup_{\mathbf{x}\in X} \|\nabla_x c_i(\mathbf{x}, \xi)\| \|\mathbf{x}_1 - \mathbf{x}_2\| \le K(\xi) \|\mathbf{x}_1 - \mathbf{x}_2\|, \end{aligned}$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in X$, and $\xi \in \Xi$. Then it follows from Theorem 7.52 of Shapiro *et al.* (2009) that both $\bar{g}_1(\mathbf{x}, \varepsilon, \mu)$ and $\bar{g}_2(\mathbf{x}, \mu)$ are continuously differentiable in $\mathbf{x}$ and

$$\begin{aligned} \nabla_x \bar{g}_1(\mathbf{x}, \varepsilon, \mu) &= E[\nabla_x H_1(\mathbf{x}, \varepsilon, \xi, \mu)] \quad \text{and} \\ \nabla_x \bar{g}_2(\mathbf{x}, \mu) &= E[\nabla_x H_2(\mathbf{x}, \xi, \mu)]. \end{aligned} \tag{11}$$

Therefore, Problem $(\mathrm{P}_\mu)$ is a smooth optimization problem, which we refer to as a smooth DC program or a smooth approximation to distinguish from the possibly non-smooth $\varepsilon$-approximation.

Note that Equation (10) shows that the difference between the constraint functions $\bar{g}(\mathbf{x}, \varepsilon, \mu)$ and $g(\mathbf{x}, \varepsilon)$ can be bounded by $2\mu \log(m+1)$, which decreases to zero as the smoothing parameter $\mu$ decreasingly goes to zero. This motivates us to consider the asymptotic behaviors of the smooth DC program; i.e., Problem $(\mathrm{P}_\mu)$, as $\mu$ goes to zero. To cope with the analysis, we make the following assumption on the constraint qualification of the $\varepsilon$-approximation. It is worthwhile noting that this type of assumption is commonly used in non-linear programming (e.g., Zangwill (1969) and Bazaraa *et al.* (1993)). For a detailed justification of Assumption 3, readers are referred to the discussion that follows Assumption 5 of Hong *et al.* (2011) as well as the Electronic Companion of Hong *et al.* (2011).

**Assumption 3.** Let $\Omega_\varepsilon^I = \{\mathbf{x} \in X : g_1(\mathbf{x}, \varepsilon) - g_2(\mathbf{x}) < 0\}$. Then $\Omega_\varepsilon = \mathrm{cl}(\Omega_\varepsilon^I)$ ($\mathrm{cl}(\cdot)$ denotes the closure of a set).

For sets $A, B \subset \Re^d$, let $\mathrm{dist}(\mathbf{x}, A) = \inf_{\mathbf{x}'\in A} \|\mathbf{x} - \mathbf{x}'\|$ denote the distance from $\mathbf{x} \in \Re^d$ to $A$ and $\mathbb{D}(A, B) = \sup_{\mathbf{x}\in A} \mathrm{dist}(\mathbf{x}, B)$ denote the deviation of the set $A$ from the set $B$ (Shapiro *et al.*, 2009). Note that it is a measure of the distance between two sets. Let $\Phi(\varepsilon, \mu) = \{\mathbf{x} \in X : \bar{g}(\mathbf{x}, \varepsilon, \mu) \le 0\}$. Recall that it is the feasible set of Problem $(\mathrm{P}_\mu)$. Let $S(\varepsilon, \mu)$ and $S(\varepsilon)$ denote the sets of optimal solutions of Problem $(\mathrm{P}_\mu)$ and Problem (5) respectively, and $\nu(\varepsilon, \mu)$ and $\nu(\varepsilon)$ denote the corresponding optimal values, respectively. We have the following theorem, whose proof is provided in the Appendix.

**Theorem 1.**

(a) $\Phi(\varepsilon, \mu) \subset \Omega_\varepsilon$ *for any* $\mu > 0$, *and* $\Phi(\varepsilon, \mu_2) \subset \Phi(\varepsilon, \mu_1)$ *for any* $0 < \mu_1 \le \mu_2$.

(b) *Suppose that Assumption 3 is satisfied. Then,*

$$\lim_{\mu \searrow 0} \Phi(\varepsilon, \mu) = \Omega_\varepsilon, \ \lim_{\mu \searrow 0} \nu(\varepsilon, \mu) = \nu(\varepsilon) \quad \textit{and}$$
$$\lim_{\mu \searrow 0} \mathbb{D}(S(\varepsilon, \mu), S(\varepsilon)) = 0.$$

Theorem 1 summarizes the properties of the smooth approximation. The first property of Theorem 1 shows that the smooth approximation is a conservative approximation of the $\varepsilon$-approximation and is therefore also a conservative approximation of the original JCCP. It also indicates that the feasible region of the smooth approximation enlarges as the smoothing parameter $\mu$ decreases. The second property shows that the feasible region of the smooth approximation converges to the feasible region of Problem (5)—i.e., the $\varepsilon$-approximation—as $\mu$ decreasingly goes to zero. Furthermore, it shows that both the optimal value and the set of optimal solutions of the smooth approximation converge to those of the $\varepsilon$-approximation, respectively.

Theorem 1 builds the convergence of optimal solutions for the smooth approximation. However, the optimal solutions of the smooth approximation may not be guaranteed by typical numerical algorithms due to the non-convexity. Thus, it is natural to study the convergence of the stationary points for the smooth approximation. Because Problem $(\mathrm{P}_\mu)$ is a smooth non-linear optimization problem, we can

use the conventional KKT conditions as its optimality conditions. Let $\Lambda(\varepsilon, \mu)$ denote the set of KKT pairs of Problem $(\mathrm{P}_\mu)$. Then,

$$\Lambda(\varepsilon, \mu) = \left\{ (\mathbf{x}, \lambda) \in \Phi(\varepsilon, \mu) \times \Re_+ : \begin{array}{c} 0 \in \nabla_x h(\mathbf{x}) + \lambda[\nabla_x \bar{g}_1(\mathbf{x}, \varepsilon, \mu) \\ -\nabla_x \bar{g}_2(\mathbf{x}, \mu)] + N_X(\mathbf{x}) \\ \lambda\,[\bar{g}_1(\mathbf{x}, \varepsilon, \mu) - \bar{g}_2(\mathbf{x}, \mu)] \\ = 0 \end{array} \right\},$$

where $N_X(\mathbf{x})$ denotes the normal cone to $X$ at $\mathbf{x}$. The major issue arising is that the $\varepsilon$-approximation—i.e., Problem (5)—may be non-smooth and non-convex, which makes its optimality conditions difficult to define. However, we note that $g(\mathbf{x}, \varepsilon)$ is a DC function, and the subdifferentials for both its first convex part $g_1(\mathbf{x}, \varepsilon)$ and its second convex part $g_2(\mathbf{x})$ are well defined (Shapiro *et al.*, 2009). This allows us to borrow the definitions of the optimality conditions for non-smooth convex optimization problems to define the optimality conditions for Problem (5). Specifically, we define the set of stationary pairs of Problem (5) as

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

where $\partial_x g_1(\mathbf{x}, \varepsilon)$ and $\partial_x g_2(\mathbf{x})$ denote the sets of subdifferentials with respect to $\mathbf{x}$ for $g_1(\mathbf{x}, \varepsilon)$ and $g_2(\mathbf{x})$ (Shapiro *et al.*, 2009). When $g_1(\mathbf{x}, \varepsilon)$ and $g_2(\mathbf{x})$ are differentiable, we have $\partial_x g_1(\mathbf{x}, \varepsilon) = \nabla_x g_1(\mathbf{x}, \varepsilon)$ and $\partial_x g_2(\mathbf{x}) = \nabla_x g_2(\mathbf{x})$. Consequently, the newly defined optimality conditions reduce to the KKT conditions and $\Lambda(\varepsilon)$ becomes the set of KKT pairs of Problem (5). Problem (5) is essentially a quasi-differentiable optimization problem in the sense of Demyanov and Rubinov (1980), and we can also use the methodology in Shapiro (1984) to investigate its optimality conditions. However, the optimality conditions introduced here are suitable for analyzing convergence properties of numerical algorithms.

As the optimality conditions for both the approximation problem and original problem have been defined, we can now analyze the convergence of the stationary points for the smooth approximation. The following theorem summarizes the convergence property. The proof of the theorem is provided in the Appendix.

**Theorem 2.** *Suppose that Assumptions 2 and 3 are satisfied. Then* $\limsup_{\mu \searrow 0} \Lambda(\varepsilon, \mu) \subset \Lambda(\varepsilon)$.

Theorems 1 and 2 ensure that Problem $(\mathrm{P}_\mu)$ can approximate Problem (5) very well. Therefore, we can solve Problem $(\mathrm{P}_\mu)$ instead of Problem (5) provided that the smoothing parameter $\mu$ is sufficiently close to zero. It is worthwhile noting that Problem $(\mathrm{P}_\mu)$ is always a conservative approximation of Problem (5), and hence of the original JCCP, no matter whether the related convergence properties can be guaranteed. Thus, by solving Problem $(\mathrm{P}_\mu)$, even when optimality cannot be reached (e.g., the related assumptions are not satisfied), we can still obtain a good feasible solution for the JCCP.

### 3.3. *Optimizing over* $\varepsilon$

An important question in the $\varepsilon$-approximation is how to set the parameter $\varepsilon$. Based on the convergence analysis, one should select a very small $\varepsilon$ to reduce the approximation error. However, as reported by Hong *et al.* (2011, p. 630), "extremely small $\varepsilon$ may cause numerical problems and may require longer time to solve the subproblem in each iteration." This motivates us to imbed the selection of $\varepsilon$ in the optimization process. In this subsection, we show how we may treat $\varepsilon$ as a decision variable in the smooth approximation and improve the performance of the approximation.

By treating $\varepsilon$ as a decision variable (hereafter $\varepsilon$ is denoted as $t$ to avoid confusion), Problem $(\mathrm{P}_\mu)$ may be strengthened as follows:

$$(\mathrm{P}^o_\mu) \quad \begin{array}{ll} \underset{\mathbf{x} \in X,\ t \geq 0}{\text{minimize}} & h(\mathbf{x}), \\ \text{subject to} & \bar{g}(\mathbf{x}, t, \mu) := \bar{g}_1(\mathbf{x}, t, \mu) - \bar{g}_2(\mathbf{x}, \mu) \leq 0. \end{array}$$

We can see that the functions $c_i(\mathbf{x}, \xi) + t, i = 1, \ldots, m$ are jointly convex in $(\mathbf{x}, t)$ for every $\xi \in \Xi$. It follows from Proposition 1 that $\Psi_1(\mathbf{x}, t, \mu)$ is jointly convex in $(\mathbf{x}, t)$. Consequently, the function $\bar{g}_1(\mathbf{x}, t, \mu)$ is jointly convex in $(\mathbf{x}, t)$. Note that $\bar{g}_2(\mathbf{x}, \mu)$ does not include $t$. Therefore, Problem $(\mathrm{P}^o_\mu)$ is also a DC program for $(\mathbf{x}, t)$. In what follows we study the properties of this new DC program and show that it could be a superior approximation to the JCCP, compared with Problem $(\mathrm{P}_\mu)$. Analogous to Assumption 3, to analyze the asymptotic behaviors of the new DC program, we need the following assumption on the constraint qualification of the original JCCP.

**Assumption 4.** Let $\Omega^I = \{\mathbf{x} \in X : \Pr\{c(\mathbf{x}, \xi) \geq 0\} < \alpha\}$. Then $\Omega = \mathrm{cl}(\Omega^I)$.

Let $Z^o(\mu)$ be the feasible set of Problem $(\mathrm{P}^o_\mu)$ and $\Phi^o(\mu)$ be the projection of $Z^o(\mu)$ on $X$; i.e., $\Phi^o(\mu) = \{\mathbf{x} \in X : \exists\ t \geq 0,\ \text{such that } (\mathbf{x}, t) \in Z^o(\mu)\}$. Let $\nu^o(\mu)$ be the optimal value of Problem $(\mathrm{P}^o_\mu)$ and $S^o(\mu)$ be the projection of the set of optimal solutions of Problem $(\mathrm{P}^o_\mu)$ on $X$; i.e., $S^o(\mu) = \{\mathbf{x} \in X : \exists\ t \geq 0,\ \text{such that } (\mathbf{x}, t) \text{ is an optimal solution of Problem } (\mathrm{P}^o_\mu)\}$. The following theorem summarizes the properties of Problem $(\mathrm{P}^o_\mu)$. The proof of the theorem is deferred to the Appendix.

**Theorem 3.**

(a) *For any* $\mu > 0$ *and* $\varepsilon > 0$, $\Phi(\varepsilon, \mu) \subset \Phi^o(\mu) \subset \Omega$ *and* $\nu \leq \nu^o(\mu) \leq \nu(\varepsilon, \mu)$.

(b) $\Phi^o(\mu_2) \subset \Phi^o(\mu_1)$ *for any* $0 < \mu_1 \leq \mu_2$.
(c) *Suppose that Assumption 4 is satisfied. Then,*

$$\lim_{\mu \searrow 0} \Phi^o(\mu) = \Omega, \ \lim_{\mu \searrow 0} v^o(\mu) = v \text{ and}$$
$$\lim_{\mu \searrow 0} \mathbb{D}(S^o(\mu), S) = 0.$$

Comparing Theorem 3 with Theorem 1, we see that Problem ($\mathrm{P}^o_\mu$) is also a smooth conservative DC approximation of the original JCCP and is a better approximation than the smooth approximation ($\mathrm{P}_\mu$). Therefore, to solve the JCCP, we can solve either Problem ($\mathrm{P}_\mu$) or Problem ($\mathrm{P}^o_\mu$). In Section 4 and the numerical experiments reported in Section 5 we approximate and solve JCCPs using Problem ($\mathrm{P}^o_\mu$).

Similarly, as Problem ($\mathrm{P}^o_\mu$) may be non-convex, its optimal solutions may not be guaranteed by the optimization procedures such as the SCA algorithm that will be introduced in Section 4. Therefore, what remains is the convergence of the "possible" stationary points for the strengthened smooth approximation Problem ($\mathrm{P}^o_\mu$). We observe that Problem ($\mathrm{P}^o_\mu$) directly approaches to Problem (3) (or the original JCCP) as $\mu \searrow 0$. However, the probability function $\Pr\{c(\mathbf{x}, \xi) > 0\}$ in the JCCP is in general non-smooth, non-convex, and also may not be locally Lipschitz continuous. Consequently, none of its gradient, subdifferential in convex context, and Clarke's generalized gradient (Clarke, 1983) are available, which makes the conventional KKT conditions for smooth optimization, the subdifferential conditions for non-smooth convex optimization, and the generalized gradient conditions for locally Lipschitz continuous optimization not applicable to the JCCP. In this article, we try to give a depiction of the possible optimality conditions for the JCCP and show the convergence of the stationary points for Problem ($\mathrm{P}^o_\mu$) in the proposed new context. As the analysis is quite involved, we include it in the Appendix (Section A4). In the rest of the article we shall focus on the computational and implementation issues of the proposed smooth approach.

## 4. Solving smooth approximations using a Monte Carlo approach

In this section, we discuss how to solve Problem ($\mathrm{P}^o_\mu$). As has been discussed, Problem ($\mathrm{P}^o_\mu$) is a stochastic DC program. Thus, solving it typically requires combining Monte Carlo methods with algorithms that solve deterministic DC programs. DC programs have been studied extensively in recent years. They can typically be solved using an SCA algorithm (Hong *et al.*, 2011). In the following subsection, we show how to fit the SCA algorithm to Problem ($\mathrm{P}^o_\mu$) and build the convergence of the algorithm accordingly.

### 4.1. *SCA*

To facilitate the analysis, we use $\mathbf{z} \in \Re^{d+1}$ to denote $(\mathbf{x}, t) \in \Re^{d+1}$. Since $\bar{g}(\mathbf{x}, t, \mu) \to +\infty$ uniformly on $X$ as $t \to \infty$, in Problem ($\mathrm{P}^o_\mu$) we can further assume that $t$ is constrained in a compact set, say, $t \in [0, T]$ for some large $T$. Due to this, we define $Z = X \times [0, T]$. If a function $f(\mathbf{x})$ only depends on $\mathbf{x}$, we simply define $f(\mathbf{z}) = f(\mathbf{x})$. The basic idea of using the SCA algorithm to solve a DC program is to convexify the DC constraint via a first-order Taylor approximation. Specifically, for Problem ($\mathrm{P}^o_\mu$), we can use the first-order Taylor expansion $\bar{g}_2(y, \mu) + \nabla_z \bar{g}_2(y, \mu)^{\mathrm{T}}(\mathbf{z} - y)$ at any point $y \in Z^o(\mu)$ to approximate $\bar{g}_2(\mathbf{z}, \mu)$. Because $\bar{g}_2(\mathbf{z}, \mu)$ is convex in $\mathbf{z}$, we have

$$\bar{g}_2(\mathbf{z}, \mu) \geq \bar{g}_2(y, \mu) + \nabla_z \bar{g}_2(y, \mu)^{\mathrm{T}}(\mathbf{z} - y), \ \forall\, \mathbf{z} \in Z,$$

which implies that

$$\bar{g}_1(\mathbf{z}, \mu) - \bar{g}_2(\mathbf{z}, \mu) \leq \bar{g}_1(\mathbf{z}, \mu) - [\bar{g}_2(y, \mu) + \nabla_z \bar{g}_2(y, \mu)^{\mathrm{T}}(\mathbf{z} - y)]. \quad (12)$$

Let

$$Z(\mu, y) = \{\mathbf{z} \in Z : \bar{g}_1(\mathbf{z}, \mu) - [\bar{g}_2(y, \mu) + \nabla_z \bar{g}_2(y, \mu)^{\mathrm{T}}(\mathbf{z} - y)] \leq 0\}.$$

It follows from Equation (12) that for any $y \in Z^o(\mu)$, the inclusion $Z(\mu, y) \subset Z^o(\mu)$ holds. Moreover, since the right-hand side of Equation (12) is a smooth convex function of $\mathbf{z}$, we have $Z(\mu, y)$ is a convex subset of $Z^o(\mu)$. Let Problem $\mathrm{CP}(\mu, y)$ denote the following optimization problem:

$$(\mathrm{CP}(\mu, y)) \qquad \text{minimize } \{h(\mathbf{z}) : \mathbf{z} \in Z(\mu, y)\}.$$

Then, for any $y \in Z^o(\mu)$, $\mathrm{CP}(\mu, y)$ is a convex conservative approximation of Problem ($\mathrm{P}^o_\mu$). Because $y$ is a feasible solution of Problem $\mathrm{CP}(\mu, y)$, the optimal solution of Problem $\mathrm{CP}(\mu, y)$ is at least as good as $y$. Then, we can repeat the above process at the newly found solution. This leads us to the following algorithm.

**Algorithm Smooth-SCA**

*Step 0.* Give an initial point $\mathbf{z}_0 \in Z^o(\mu)$ and set $k = 0$.
*Step 1.* Stop if $\mathbf{z}_k$ is a KKT point of Problem ($\mathrm{P}^o_\mu$).
*Step 2.* Solve $\mathrm{CP}(\mu, \mathbf{z}_k)$ to obtain its optimal solution $\mathbf{z}_{k+1}$.
*Step 3.* Set $k = k + 1$ and go to Step 1.

Algorithm smooth-SCA is easy to implement, since we only need to solve the convex optimization problem $\mathrm{CP}(\mu, \mathbf{z}_k)$ in each iteration. It also has some desired properties, which we summarize in the following theorem. Note that we say that Slater's condition holds at $y \in Z$ if $\operatorname{int} Z(\mu, y) \neq \phi$, where $\operatorname{int} A$ denotes the interior of a set $A$. Slater's condition is one of the most commonly used constraint qualifications for convex optimization (Boyd and Vandenberghe, 2004). The proof of Theorem 4 follows directly from the results of Hong *et al.* (2011) and it is provided in the Appendix for completeness.

**Theorem 4.** *Suppose* $\{\mathbf{z}_k\}$ *is a sequence of solutions generated by Algorithm Smooth-SCA for Problem* ($\mathrm{P}^o_\mu$) *starting from* $\mathbf{z}_0 \in Z^o(\mu)$. *Then,*

1. $\{\mathbf{z}_k\} \subset Z^o(\mu)$ *and* $\{h(\mathbf{z}_k)\}$ *is a convergent non-increasing sequence.*
2. *if* $\mathbf{z}_{k+1} = \mathbf{z}_k$ *at which Slater's condition holds, then* $\mathbf{z}_k$ *is a KKT point of Problem* $(\mathrm{P}^o_\mu)$.
3. *Suppose* $\bar{\mathbf{z}}$ *is a cluster point of* $\{\mathbf{z}_k\}$ *satisfying Slater's condition. Then* $\bar{\mathbf{z}}$ *is a KKT point of Problem* $(\mathrm{P}^o_\mu)$.

The first property in Theorem 4 shows that we always search better solutions in the feasible region, which is similar to the framework of interior-point methods. It also shows that we make improvement at each iteration and the sequence of objective values converges to a certain value. The second property shows that, if our algorithm terminates after a finite number of iterations, we actually reach a KKT point. The third property ensures that all limit points of the sequence of solutions generated are KKT points. Together with the second property, it demonstrates that our algorithm has the desired convergence property. If Problem $(\mathrm{P}^o_\mu)$ only has a single KKT point or has only a KKT point that is better than the initial solution, or Problem $(\mathrm{P}^o_\mu)$ is convex, our algorithm guarantees to converge to a global optimal solution.

Note that Algorithm Smooth-SCA can start from any feasible solution of the JCCP. However, there is a natural initial solution for our problem. We can drop the second term $\bar{g}_2(\mathbf{z}, \mu)$ from Problem $(\mathrm{P}^o_\mu)$ to obtain the following optimization problem:

$$\begin{aligned} \underset{\mathbf{z}\in Z}{\text{minimize}} \quad & h(\mathbf{z}), \\ \text{subject to} \quad & \bar{g}_1(\mathbf{z}, \mu) \le 0. \end{aligned} \tag{13}$$

Note that $\bar{g}_1(\mathbf{z}, \mu) \le 0$ is the smooth conservative approximation of $E[[c(\mathbf{x}, \boldsymbol{\xi}) + t]^+] - \alpha t \le 0$, while the latter constraint is equivalent to the CVaR constraint in the CVaR approximation mentioned in Section 1 (Nemirovski and Shapiro, 2006). Therefore, Problem (13) is also a convex conservative approximation of the original JCCP and thus its optimal solution can be used as the initial solution to Algorithm Smooth-SCA. In this article we call Problem (13) a smooth CVaR approximation and call its optimal solution a smoothed CVaR solution. In the numerical studies reported in Section 5 we use the smoothed CVaR solutions as the initial solutions.

### 4.2. *Sample-average approximation*

To implement Algorithm Smooth-SCA, we need to solve Problem CP$(\mu, y)$ efficiently. Note that Problem CP$(\mu, y)$ is a standard smooth convex stochastic program. It has been studied extensively in the literature (Shapiro *et al.*, 2009). When $\boldsymbol{\xi}$ is a discrete random vector and takes values $\xi_1, \xi_2, \ldots, \xi_n$ with probabilities $p_1, p_2, \ldots, p_n$ ($\sum_{j=1}^n p_j = 1$), the functions

$$\Psi_1(\mathbf{z}, \mu) = \sum_{j=1}^n p_j H_1(\mathbf{z}, \xi_j, \mu) \quad \text{and}$$
$$\Psi_2(\mathbf{z}, \mu) = \sum_{j=1}^n p_j H_2(\mathbf{z}, \xi_j, \mu)$$

are deterministic smooth convex functions. In this case, Problem $(\mathrm{P}^o_\mu)$ degenerates to a deterministic DC program. Consequently, Problem CP$(\mu, y)$ becomes a deterministic convex optimization problem and thus it can be solved efficiently. When $\boldsymbol{\xi}$ is a continuous random vector, the closed form of the right-hand side of Equation (12) is typically unavailable. In this case, we can solve Problem CP$(\mu, y)$ using a Sample-Average Approximation (SAA) approach. In what follows we implement the SAA approach to solve CP$(\mu, y)$. For a comprehensive review about the SAA, readers are referred to Shapiro *et al.* (2009).

Suppose that we have an independent and identically distributed (i.i.d.) sample $\{\xi_1, \xi_2, \ldots, \xi_n\}$ from the random vector $\boldsymbol{\xi}$, which may be generated from a simulation study or extracted from historical data. Using the sample, the functions $\bar{g}_1(\mathbf{z}, \mu)$ and $\bar{g}_2(y, \mu)$ may be estimated by

$$\bar{g}_{1,n}(\mathbf{z}, \mu) := \frac{1}{n}\sum_{j=1}^n H_1(\mathbf{z}, \xi_j, \mu) - \alpha t \quad \text{and}$$
$$\bar{g}_{2,n}(y, \mu) := \frac{1}{n}\sum_{j=1}^n H_2(y, \xi_j, \mu) - \mu \log(m+1),$$

respectively. It follows from Equation (11) that the gradient $\nabla_z \bar{g}_2(y, \mu)$ can be estimated by $(1/n)\sum_{j=1}^n \nabla_z H_2(y, \xi_j, \mu)$. Therefore, we can use the following sample problem to approximate Problem CP$(\mu, y)$:

$$\begin{aligned} \underset{\mathbf{z}\in Z}{\text{minimize}} \quad & h(\mathbf{z}), \\ \text{subject to} \quad & \bar{g}_{1,n}(\mathbf{z}, \mu) \\ & - \left[\bar{g}_{2,n}(y, \mu) + \frac{1}{n}\sum_{j=1}^n \nabla_z H_2(y, \xi_j, \mu)^{\mathrm{T}}(\mathbf{z} - y)\right] \le 0. \end{aligned} \tag{14}$$

Let $v_n(\mu, y)$ and $S_n(\mu, y)$ denote the optimal value and the set of optimal solutions of Problem (14), respectively, and $v(\mu, y)$ and $S(\mu, y)$ denote the optimal value and the set of optimal solutions of Problem CP$(\mu, y)$, respectively. Intuitively, when the sample size $n$ is sufficiently large, $v_n(\mu, y)$ and $S_n(\mu, y)$ should provide good approximations to $v(\mu, y)$ and $S(\mu, y)$, respectively. This intuition is confirmed by the following theorem. The proof of the theorem is provided in the Appendix.

**Theorem 5.** *Suppose that Assumption 2 is satisfied, and Slater's condition holds at y. Then with probability one (w.p.1.):*

$$\lim_{n\to\infty} v_n(\mu, y) = v(\mu, y) \quad \textit{and}$$
$$\lim_{n\to\infty} \mathbb{D}(S_n(\mu, y), S(\mu, y)) = 0.$$

Note that Problem (14) is a deterministic convex optimization problem, in which the gradients of the objective function and constraint function can be evaluated. We can then use a gradient-based method to directly solve Problem (14). In the numerical results reported in Section 5, we combine the SAP with Algorithm Smooth-SCA to solve Problem ($P^o_\mu$). Note that this approach is called the SMC approach in this article, as mentioned in Section 1.

### 4.3. *Numerical stability*

Similar to the $\varepsilon$-approximation, a very first step of implementing the SMC approach is to set the smoothing parameter $\mu$. To reduce the approximation error of the smooth approximation, as has been demonstrated, we often prefer very small $\mu$ values. However, this may cause certain numerical stability issue, as the computation involves dangerous operations such as $\exp(\mu^{-1}(\cdot))$; see, for instance, Nesterov (2005). In this article, we adopt the following approach that was suggested by Nesterov (2005) to address this issue. For any vector $(a_0, a_1, \ldots, a_m)$, let $\bar{a} = \max\{a_0, a_1, \ldots, a_m\}$ and $b_i = a_i - \bar{a}$, $i = 0, 1, \ldots, m$. Then

$$\mu \log\left[\sum_{i=0}^{m} \exp(\mu^{-1} a_i)\right] = \bar{a} + \mu \log\left[\sum_{i=0}^{m} \exp\left(\mu^{-1} b_i\right)\right].$$

It is easily seen that all $b_i$, $i = 0, 1, \ldots, m$ are non-positive and at least one of them has a level of zero. Therefore, the function value of the logarithm-sum-exponential function can be computed with small numerical errors. Furthermore, it can be verified that

$$\nabla_a\, \mu \log\left[\sum_{i=0}^{m} \exp(\mu^{-1} a_i)\right] = \nabla_b\, \mu \log\left[\sum_{i=0}^{m} \exp\left(\mu^{-1} b_i\right)\right].$$

Therefore, we can also use the technique to compute the gradient of the logarithm-sum-exponential function. Nesterov (2005) showed that this approach is very effective even for a quite small $\mu$ (e.g., $\mu = 10^{-3}, 10^{-4}$) and has quite high accuracy.

## 5. Numerical experiments

We study the performances of our method through two classes of examples. The first class is non-smooth JCCPs where the random parameters follow discrete distributions, whereas the second one is smooth JCCPs where the random distributions are continuous. We use the examples to test the efficacy of our SMC approach and compare the SMC approach with the $\varepsilon$-approximation of Hong *et al.* (2011). We implement Algorithm Smooth-SCA in MATLAB and use MATLAB function fmincon to solve the convex optimization problem in each iteration. All of the programs are run on a laptop with Intel Core 2 Duo CPU (2.26 GHz, 2.27 GHz) and 4 GB of RAM.

### 5.1. *JCCPs with discrete random variables*

In this subsection we assume that the random vector $\boldsymbol{\xi}$ in Problem (1) takes only $n$ scenarios $\xi_1, \ldots, \xi_n$ with equal probabilities. It can be checked that the smoothness of this JCCP is violated. We use such examples to show that the SMC approach can handle non-smooth JCCPs with discrete distributions.

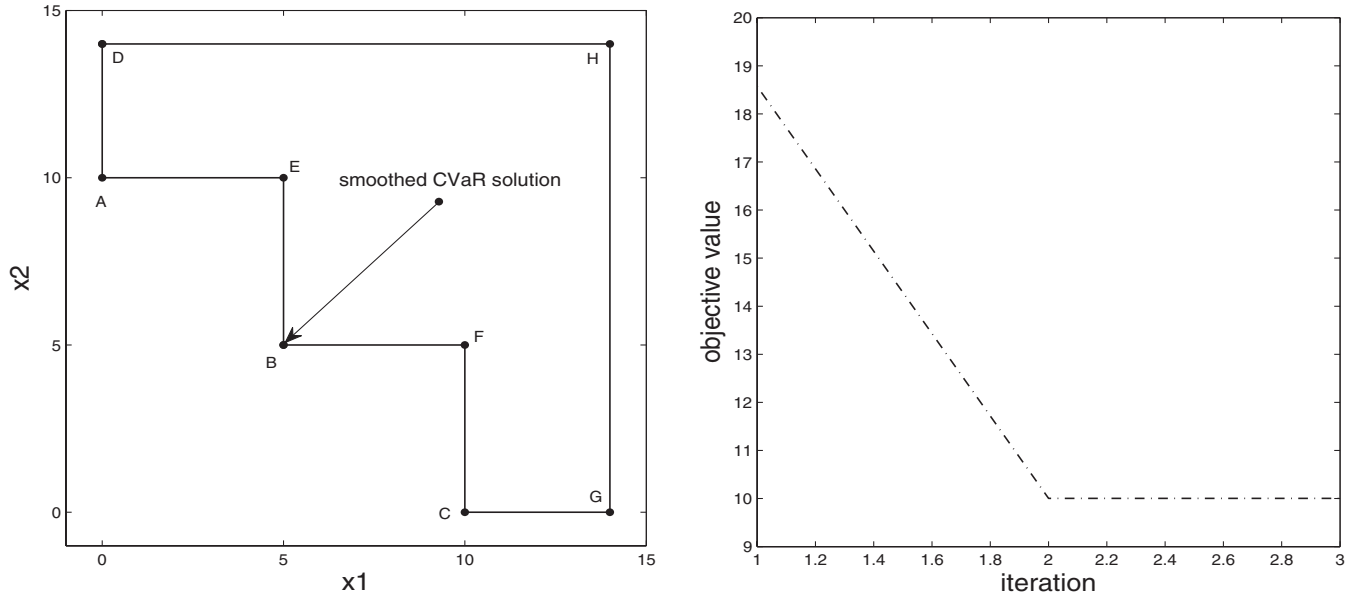
#### 5.1.1. *Linear JCCPs with discrete random variables*

We first consider a very simple instance that takes the form of Formulation (1.3) of Prékopa (2003):

$$\begin{aligned} \underset{\mathbf{x}\in X}{\text{minimize}} \quad & \mathbf{e}^{\mathrm{T}}\mathbf{x} \\ \text{subject to} \quad & \Pr\{\boldsymbol{\xi} - \mathbf{B}\mathbf{x} \le 0\} \ge 1 - \alpha, \end{aligned} \tag{15}$$

where $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$ is a $d$-dimensional vector in $\Re^d$, $\boldsymbol{\xi} = (\xi^1 \ldots, \xi^m)^{\mathrm{T}}$ is an $m$-dimensional random vector, $\mathbf{e}$ is a $d \times 1$ deterministic vector, and $\mathbf{B}$ is an $m \times d$ deterministic matrix. Specifically, we set $d = 2$, $k = 2$, $\alpha = 0.42$, $\mathbf{e} = (1, 1)^{\mathrm{T}}$, $\mathbf{B} = \mathbf{I}$, where $\mathbf{I}$ is a $2 \times 2$ identity matrix, and $\mathbf{X} = \{\mathbf{x} : -14 \le \mathbf{x} \le 14\}$. We assume $\xi^j$, $j = 1, 2$, independently take values in $\{-10, -5, 0, 5, 10\}$ with equal probabilities. Thus, the random vector $\boldsymbol{\xi}$ takes 25 scenarios with equal probabilities. Under the setting, the feasible region of Problem (15) is the irregular polygon AEBFCGHDA shown in Fig. 1(a), and the optimal solutions are A(0, 10), B(5, 5), and C(10, 0) with optimal objective value being 10. They are used as benchmarks to evaluate the performances of our method. It can be verified that for this example the two sets $\Omega^I$ and $\Omega$ are the same and they are just the polygon minus the segments DA, AE, EB, BF, FC, and CG. This means that Assumption 4 is not satisfied for the example (but we do have cl $\Omega^I = \Omega_0$). However, we do not care about the violation of the constraint qualification. We just use the SMC approach to solve the problem and see what will happen. In the implementation of the SMC approach, we use the smoothed CVaR solution as the initial solution for our algorithm and stop the algorithm if the difference of objective function values between two consecutive iterations is less than or equal to $10^{-4}$. We first set the smoothing parameter $\mu$ as $10^{-4}$. Because the instance problem is a small-scale deterministic optimization problem, our algorithm always converges to the same solution in only three iterations with a negligible amount of computational time (less than 1 second).

The convergence of the solutions is shown in Fig. 1(a). From the plot we see that our algorithm succeeds in converging to the optimal solution B starting from the smoothed CVaR solution. The convergence of the objective




(a) (b)

**Fig. 1.** Performances of SMC approach for discrete random variables: (a) the convergence of the solutions and (b) the convergence of the objective values.

values is also shown in Fig. 1(b). The optimal value of the smooth CVaR approximation is about 18.6, and our algorithm improves on this value and converges to the optimal value 10, which is significantly lower than the objective function value guaranteed by the smooth CVaR approximation. It seems that in this example the kink points E and F keep the smooth CVaR approximation from finding better solutions. Once we further make convex approximation around the smoothed CVaR solution, the effects of points E and F vanish, which then enables the SMC approach to find the optimal point B. This example suggests that the SMC approach could be very effective in solving the JCCPs even when the JCCPs admit irregular characteristics and even when they are not covered by the convergence theory built in this article (Assumption 4 is violated). We also used a number of different feasible points as the initial solution for our algorithm. The experiments show that our algorithm has similar performances.

To investigate the effects of the smoothing parameter $\mu$ and the behavior of the $t$-component of the decision vector, we now consider different values of $\mu$. The results obtained are summarized in Table 1. The first row of Table 1 is the number of iterations that are taken for the algorithm to converge. It shows that for all of the $\mu$ values considered, the algorithm can quickly converge to the optimal solution from the smoothed CVaR solution. The second row shows the optimal value for the smooth approximation obtained by the algorithm, from which we find that the approximation error becomes smaller and smaller as $\mu$ gradually decreases, and the optimal value of the smooth approximation is already very close to the optimal value of the original JCCP when $\mu$ is as small as $10^{-2}$. The last row displays the behavior of the $t$-component of the optimal solutions produced by the algorithm. From this row we see that the $t$-component gets closer and closer to zero as $\mu$ gradually reduces to zero.

### 5.1.2. *Non-linear JCCPs with discrete random variables*

We have studied some properties of the SMC approach via the above simple example. Now we implement our approach to a slightly more complicated model that takes the following form:

$$\begin{aligned} \underset{0\le \mathbf{x}\le 100}{\text{minimize}} \quad & \mathbf{x}^{\mathrm{T}}\Sigma_0\mathbf{x} + \mathbf{a}^{\mathrm{T}}\mathbf{x} \\ \text{subject to} \quad & \Pr\left\{\boldsymbol{\xi}_i^{\mathrm{T}}\Sigma_i\mathbf{x} + b_i \le 0, i = 1, \ldots, m\right\} \ge 1-\alpha, \end{aligned} \tag{16}$$

where $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$ is a $d$-dimensional vector in $\Re^d$, $\boldsymbol{\xi}_i = (\xi_i^1 \ldots, \xi_i^d)^{\mathrm{T}}$, $i = 1, \ldots, m$ are $d$-dimensional random

**Table 1.** Performances of SMC approach for different $\mu$

| | $\mu$ | | | | | |
|---|---|---|---|---|---|---|
| | $10^{-1}$ | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | $10^{-10}$ |
| Number of iterations | 4 | 3 | 3 | 3 | 3 | 3 |
| Optimal value | 14.1718 | 10.4172 | 10.0417 | 10.0042 | 10.0004 | 10.0000 |
| Optimal $t$ | 1.9444 | 0.1944 | 0.0196 | 0.0021 | $1.9238\ 10^{-4}$ | $1.6731\ 10^{-6}$ |

vectors, $\Sigma_i$, $i = 0, 1, \ldots, m$ are $d \times d$ deterministic positive semi-definite matrices, $\mathbf{a}$ is a $d \times 1$ deterministic vector, and $b_i$, $i = 1, \ldots, m$ are deterministic parameters. Problem (16) is a non-smooth JCCP with a quadratic objective function. It is often referred to as a *chance-constrained quadratic program* (Zheng *et al.*, 2012).

To conduct the experiments, we set $m = 10$ and $b_i = -200$, $i = 1, \ldots, m$. In each replication of the experiments, we randomly generate parameters $\Sigma_i$, $i = 0, 1, \ldots, m$ from the matrix $UU^{\mathrm{T}}$ where $U$ is a $d$-dimensional random vector whose components are independent and follow [0, 1] uniform distribution and generate $\mathbf{a}$ from a random vector $U$ whose components are independent and follow $[-100, 0]$ uniform distribution. Also in each replication, we construct each $\boldsymbol{\xi}_i$ via the following procedure (Steps (i) and (ii)): (i) independently generate 500 scenarios from a random vector $U$ whose components are independent and follow $[-10, 10]$ uniform distribution; (ii) assume that $\boldsymbol{\xi}_i$ takes the 500 scenarios with equal probabilities. Once the input parameters and the scenarios are given, Problem (16) becomes a deterministic problem. However, because the scale of the problem is very large, it is difficult to solve the problem to obtain its optimal values. We mainly compare the SMC approach with the smooth CVaR approximation. We conduct experiments for $d = 10, 50, 100$ and $\alpha = 0.1, 0.2, 0.3, 0.4$. For each combination of $d$ and $\alpha$, we make five replications (i.e., randomly generate the input parameters and the discrete random vector and conduct the experiments five times). In the experiments, we set the smoothing parameter $\mu = 10^{-4}$. The stopping rule is the same as that for Problem (15). Table 2 shows the typical computational results.

As can be seen in Table 2, there are five columns for each combination of $d$ and $\alpha$. Column 1 (SCVaR) shows the optimal objective value of the smooth CVaR approximation, whereas Column 2 (SMC) shows the optimal objective value obtained by the SMC approach. The percentage changes of Column 2 from Column 1 are reported

**Table 2.** Performances of SMC approach for chance-constrained quadratic program

| | $\alpha = 0.1$ | | | | | $\alpha = 0.2$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *d* | *SCVaR* | *SMC* | *Change (%)* | *Ite* | *CPUT* | *SCVaR* | *SMC* | *Change (%)* | *Ite* | *CPUT* |
| 10 | −1099 | −1257 | 14.4 | 10 | 54 | −1024 | −1313 | 28.2 | 6 | 34 |
| 10 | −1649 | −2251 | 36.5 | 11 | 50 | −1164 | −1479 | 27.1 | 10 | 76 |
| 10 | −926 | −1114 | 20.3 | 5 | 21 | −1013 | −1277 | 26.1 | 13 | 107 |
| 10 | −1010 | −1260 | 24.8 | 7 | 34 | −924 | −1173 | 27.0 | 21 | 120 |
| 10 | −1109 | −1454 | 31.1 | 7 | 48 | −2018 | −2970 | 47.2 | 7 | 35 |
| 50 | −581 | −680 | 17.0 | 6 | 63 | −750 | −916 | 22.1 | 10 | 217 |
| 50 | −674 | −807 | 19.7 | 4 | 34 | −721 | −867 | 20.3 | 11 | 216 |
| 50 | −622 | −714 | 14.8 | 12 | 149 | −800 | −1134 | 41.8 | 8 | 153 |
| 50 | −618 | −739 | 19.6 | 8 | 163 | −810 | −990 | 22.2 | 9 | 192 |
| 50 | −653 | −765 | 17.2 | 7 | 156 | −767 | −977 | 27.4 | 7 | 134 |
| 100 | −429 | −498 | 16.1 | 5 | 139 | −547 | −685 | 25.2 | 8 | 281 |
| 100 | −487 | −588 | 20.7 | 13 | 600 | −603 | −776 | 28.7 | 10 | 655 |
| 100 | −605 | −716 | 18.4 | 6 | 407 | −566 | −689 | 21.7 | 8 | 546 |
| 100 | −476 | −536 | 12.6 | 7 | 350 | −594 | −737 | 24.1 | 18 | 883 |
| 100 | −617 | −753 | 22.0 | 12 | 571 | −543 | −658 | 21.2 | 7 | 373 |
| | $\alpha = 0.3$ | | | | | $\alpha = 0.4$ | | | | |
| *d* | *SCVaR* | *SMC* | *Change (%)* | *Ite* | *CPUT* | *SCVaR* | *SMC* | *Change (%)* | *Ite* | *CPUT* |
| 10 | −1185 | −1771 | 49.5 | 8 | 44 | −1068 | −1577 | 47.7 | 8 | 56 |
| 10 | −1317 | −1730 | 31.4 | 7 | 48 | −1000 | −1271 | 27.1 | 10 | 56 |
| 10 | −910 | −1304 | 43.3 | 6 | 20 | −1274 | −2103 | 65.1 | 10 | 59 |
| 10 | −1265 | −1595 | 26.1 | 8 | 60 | −1245 | −1756 | 41.0 | 11 | 88 |
| 10 | −1216 | −1561 | 28.4 | 9 | 48 | −1176 | −1755 | 49.2 | 7 | 30 |
| 50 | −677 | −851 | 25.7 | 17 | 482 | −850 | −1179 | 38.7 | 7 | 77 |
| 50 | −865 | −1094 | 26.5 | 13 | 313 | −847 | −1100 | 29.9 | 9 | 245 |
| 50 | −691 | −855 | 23.7 | 12 | 329 | −798 | −1100 | 37.8 | 10 | 214 |
| 50 | −768 | −971 | 26.4 | 11 | 200 | −738 | −1020 | 38.2 | 9 | 189 |
| 50 | −825 | −1063 | 28.9 | 9 | 203 | −757 | −986 | 30.1 | 14 | 312 |
| 100 | −543 | −696 | 28.2 | 9 | 561 | −642 | −832 | 29.6 | 10 | 577 |
| 100 | −589 | −726 | 23.3 | 11 | 589 | −627 | −841 | 34.1 | 9 | 528 |
| 100 | −657 | −861 | 31.1 | 10 | 390 | −546 | −742 | 35.9 | 11 | 608 |
| 100 | −567 | −745 | 31.4 | 13 | 679 | −567 | −754 | 33.0 | 12 | 707 |
| 100 | −669 | −907 | 35.6 | 10 | 599 | −574 | −776 | 35.2 | 6 | 193 |

in Column 3 (Change (%)). The results show that the improvement achieved by the SMC approach is quite significant. The last two columns show the number of iterations (Ite) and the CPU time (CPUT, seconds, computed by MATLAB function cputime) for the algorithm to terminate. Note that in our experiments the CPU time is significantly longer (1.5–2 times) than the Elapsed time computed using MATLAB function tic/toc, whereas the Elapsed time is the real running time of the algorithm. The results suggest that the SMC approach can often stop in a number of iterations, spending a reasonable computational time. We have also observed the behaviors of the $t$-component of the solutions and found that for all of the replications the $t$-component approaches to zero. Table 2 empirically shows that the SMC approach could be very effective in solving realistic-scale non-smooth JCCPs.

### 5.2. *JCCPs with continuous random variables*

In this subsection we solve the norm optimization problem of Hong *et al.* (2011) where the random variables take continuous distributions and the problem is indeed smooth. We use this example to compare our SMC approach with the $\varepsilon$-approximation approach. After some transformation the problem can be written as the following JCCP:

$$\begin{aligned} \underset{\mathbf{x}\geq 0}{\text{minimize}} \quad & -\sum_{j=1}^{d} x_j \\ \text{subject to} \quad & \Pr\left\{\sum_{j=1}^{d} \xi_{ij}^2 x_j^2 \leq 100,\ i=1,\ldots,m\right\} \geq 1-\alpha, \end{aligned} \tag{17}$$

where $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$ is a $d$-dimensional vector in $\Re^d$, and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1 \ldots, \boldsymbol{\xi}_d)$, with $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{id})^{\mathrm{T}}$ for $i = 1, \ldots, m$ is a $d \times m$ matrix of random variables.

Let $c_i(\mathbf{x}, \boldsymbol{\xi}) = \sum_{j=1}^{d} \xi_{ij}^2 x_j^2 - 100$, $i = 1, \ldots, m$, and $c(\mathbf{x}, \boldsymbol{\xi}) = \max_{i=1,\ldots,m}\{c_i(\mathbf{x}, \boldsymbol{\xi})\}$. Then, Problem (17) takes the form of Problem (2). According to Hong *et al.* (2011), this JCCP is smooth ($c(\mathbf{x}, \boldsymbol{\xi})$ satisfies the smooth conditions) and thus the $\varepsilon$-approximation can be used. Our major aim of using this example is to demonstrate the benefits of treating $\varepsilon$ as a decision variable. It is worth noting that implementing the SMC approach does not require checking the smoothness. To conduct the experiments, we set $d = 10$, $m = 10$, and $\alpha = 0.1$. We assume the random variables $\xi_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, d$ are i.i.d. standard normal random variables. Under the assumption, the optimal solution and optimal value of Problem (17) can be derived analytically, which are $(2.082, 2.082, \ldots, 2.082)^{\mathrm{T}}$ and $-20.82$, respectively (Hong *et al.*, 2011). We set the sample size as 10 000 and still set the smoothing parameter $\mu$ as $10^{-4}$. The algorithm is stopped if the difference of objective values between two consecutive iterations is less than or equal to $10^{-2}$. We ran Algorithm Smooth-SCA multiple times and these replications always showed similar performances. We also ran the $\varepsilon$-approximation where the parameter $\varepsilon$ was set as 0.05, as used in Hong *et al.* (2011). We report a typical simulation run for the two approaches in Fig. 2.

We plot the objective values in Fig. 2(a). From the plot we see that the SMC approach starts from the objective value of the smooth CVaR approximation and converges to the optimal value in less than 10 iterations, while the $\varepsilon$-approximation starts from the objective value of the CVaR approximation and converges to the optimal value in about

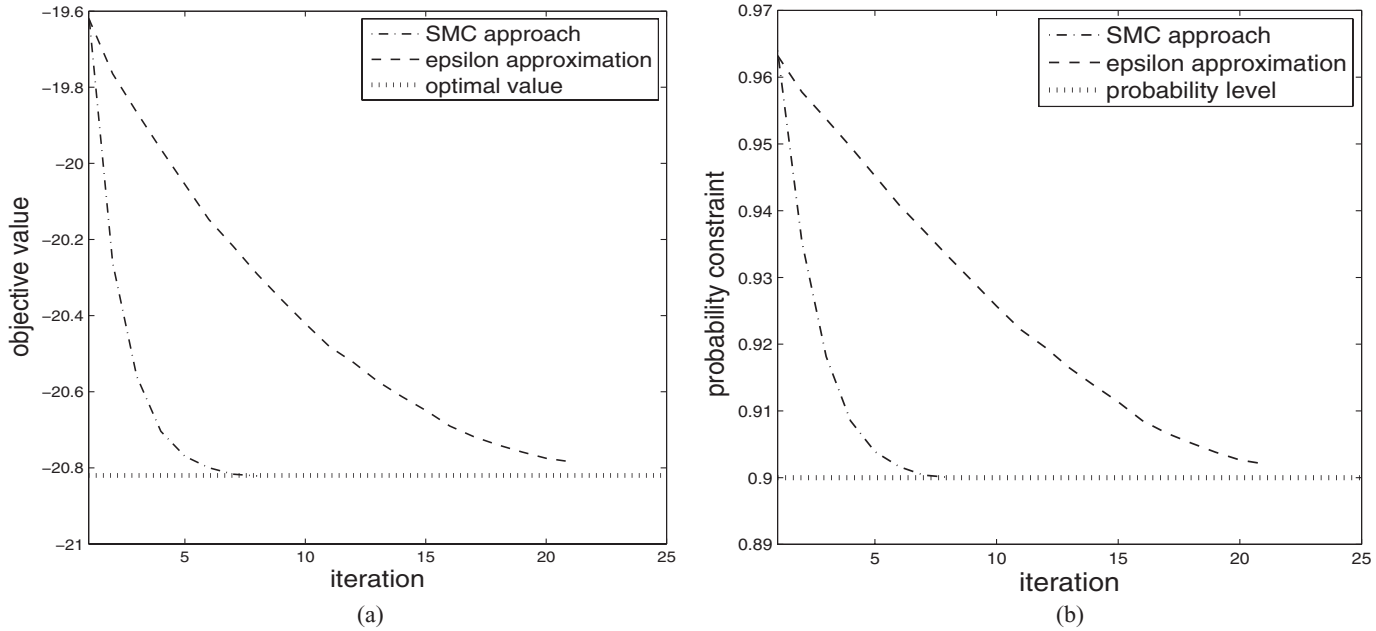



**Fig. 2.** Performances of SMC and $\varepsilon$-approximation approaches: (a) convergence of the objective values and (b) the values of the probability constraint.

**Table 3.** Performances for different $\varepsilon$

| | $\varepsilon$ | | | |
|---|---|---|---|---|
| | *0.1* | *0.05* | *0.02* | |
| Number of iterations | 17 | 23 | 32 | 8 |
| Elapsed time | 285 | 405 | 688 | 171 |

20 iterations. Figure 2(b) shows the values of the left-hand side of the joint chance constraint, estimated at solutions generated by the two approaches. From the plot we see that both approaches keep on relaxing the joint chance constraint until the probability level decreases to the prespecified level $1-\alpha$. In the two approaches, the Elapsed time per iteration is similar. However, the SMC approach reduces both the number of iterations and the total computational time by more than half compared with the $\varepsilon$-approximation. This shows the superiority of the SMC approach. To further highlight this phenomenon, we varied the parameter $\varepsilon$ in the $\varepsilon$-approximation and ran the $\varepsilon$-approximation to see the effects of $\varepsilon$. Especially, we considered $\varepsilon = 0.05, 0.02, 0.1$ and randomly made five replications for each value. We report the average number of iterations and average Elapsed (running) time in Table 3. The experiments show that while the optimal values are somewhat insensitive to $\varepsilon$, the running time of the algorithm varies significantly over the three values. We also randomly ran Algorithm Smooth-SCA five times and report the average results in the last column of Table 3. The experiments suggest that the SMC approach enjoys a faster convergence than the $\varepsilon$-approximation for reasonably selected $\varepsilon$. One possible explanation for this is that we treat $\varepsilon$ as a decision variable $t$ in the SMC approach and $t$ can automatically adapt itself to the Taylor expansion in each iteration and thus can help find the global optimal solution very quickly.

Similar to the first example, we varied the parameter $\mu$ to see the effects of $\mu$ on the performances of Algorithm Smooth-SCA. The results obtained show a pattern similar to those reported in Table 1, except that the current problem admits a stochastic nature and there exists some minor difference across different simulation runs. Especially, we again observed that the $t$-component of the solutions provided by Algorithm Smooth-SCA gets closer and closer to zero as $\mu$ gradually decreases to zero.

We also studied two additional cases. In the first case, we assumed that the elements of the random vector $\boldsymbol{\xi}$ were dependent and followed normal distributions. In the second case, we assumed that the elements of $\boldsymbol{\xi}$ were independent and followed a lognormal distribution. For the two cases, both the SMC approach and the $\varepsilon$-approximation show similar patterns as for the independent normal case reported above.

## 6. Conclusions

In this article, we propose a logarithm-sum-exponential smoothing approach to approximate possibly non-smooth JCCPs as smooth stochastic DC programs. We then combine Monte Carlo methods with an SCA algorithm to solve the smooth stochastic DC programs. The SMC approach developed in this article can handle both smooth and non-smooth JCCPs where the random variables can be either continuous or discrete. We also show that this approach guarantees desired convergence properties under certain conditions. Even when the conditions are violated, the proposed smooth approximation is still implementable and is a conservative approximation of the original JCCP and thus the solutions obtained by the SMC approach still guarantee feasibility for the original JCCP. The numerical results show that the SMC approach works well on both smooth and non-smooth JCCPs.

## Acknowledgements

The authors would like to thank the Associate Editor and two referees for their insightful comments. The research of the first author was partially supported by the National Natural Science Foundation of China under grants 71090404/71090400 and 71201117, the research of the second author was partially supported by the Hong Kong Research Grants Council under projects GRF613011 and N_HKUST626/10, and the research of the third author was partially supported by the National Natural Science Foundation of China under projects 11071029 and 91130007.

## References

Alexander, S., Coleman, T.F. and Li, Y. (2006) Minimizing CVaR and VaR for a portfolio of derivatives. *Journal of Bank and Finance*, **30**, 583–605.

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

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

Bertsekas, D. (1975) Nondifferentiable optimization via approximation, in *Mathematical Programming Study 3, Nondifferentiable Optimization*, Balinski, M.L. and Wolfe, P. (eds), North-Holland, Amsterdam, pp. 1–25.

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

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

Calafiore, G. and Campi, M.C. (2005) Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, **102**, 25–46.

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

Charnes, A., Cooper, W.W. and Symonds, G.H. (1958) Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil. *Management Science*, **4**, 235–263.
Chen, W., Sim, M., Sun, J. and Teo, C.-P. (2010) From CVaR to uncertainty set: implications in joint chance constrained optimization. *Operations Research*, **58**, 470–485.
Clarke, F.H. (1983) *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, NY.
De Farias, D.P. and Van Roy, B. (2004) On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, **29**, 462–478.
Demyanov, V.F. and Rubinov, A.M. (1980) On quasidifferentiable functionals. *Soviet Mathematics-Doklady*, **21**, 14–17.
Durrett, R. (2005) *Probability: Theory and Examples*, third edition, Duxbury Press, Belmont, MA.
Fukushima, M. and Qi, L. (1998) *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*, Kluwer, Dordrecht, The Netherlands.
Hong, L.J. and Liu, G. (2009) Simulating sensitivities of conditional value-at-risk. *Management Science*, **55**, 281–293.
Hong, L.J., Yang, Y. and Zhang, L. (2011) Sequential convex approximations to joint chance constrained programs: a Monte Carlo approach. *Operations Research*, **59**, 617–630.
Miller, L.B. and Wagner, H. (1965) Chance-constrained programming with joint constraints. *Operations Research*, **13**, 930–945.
Molchanov, I. (2005) *Theory of Random Sets*, Springer-Verlag, London.
Nemirovski, A. and Shapiro, A. (2006) Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, **17**, 969–996.
Nesterov, Y. (2005) Smooth minimization of nonsmooth functions. *Mathematical Programming*, **103**, 127–152.
Prékopa, A. (2003) Probabilistic programming, in *Stochastic Programming, Handbooks in OR&MS*, Vol. 10, Ruszczynski, A. and Shapiro, A. (eds), Elsevier, Amsterdam, pp. 267–351.
Rockafellar, R.T. and Uryasev, S. (2000) Optimization of conditional value-at-risk. *The Journal of Risk*, **2**, 21–41.
Rockafellar, R.T. and Uryasev, S. (2002) Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, **26**, 1443–1471.
Rockafellar, R.T. and Wets, R.J.-B. (1998) *Variational Analysis*, Springer-Verlag, New York, NY.
Shapiro, A. (1984) On optimality conditions in quasidifferentiable optimization. *SIAM Journal on Control and Optimization*, **22**, 610–617.
Shapiro, A., Dentcheva, D. and Ruszczyński, A. (2009) *Lectures on Stochastic Programming: Modeling and Theory*, SIAM, Philadelphia, PA.
Xu, H. and Zhang, D. (2009) Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications. *Mathematical Programming*, **119**, 371–401.
Zangwill, W.I. (1969) *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ.
Zheng, X.J., Sun, X.L., Li, D. and Cui, X.T. (2012) Lagrangian decomposition and mixed-integer quadratic programming reformulations for probabilistically constrained quadratic programs. *European Journal of Operational Research*, **221**, 38–48.



## Appendix

### A1: Proof of Theorem 1

We first prove (a). For any $\mu > 0$ and any $\mathbf{x} \in \Phi(\varepsilon, \mu)$, we have from Equation (10) that

$$g_1(\mathbf{x}, \varepsilon) - g_2(\mathbf{x}) \le \bar{g}_1(\mathbf{x}, \varepsilon, \mu) - \bar{g}_2(\mathbf{x}, \mu) \le 0,$$

which implies $\mathbf{x} \in \Omega_\varepsilon$. Therefore, the inclusion $\Phi(\varepsilon, \mu) \subset \Omega_\varepsilon$ holds.

Let

$$\bar{\lambda}_j(\mathbf{x}, \varepsilon, \xi, \mu) = \frac{\exp(\mu^{-1}(c_j(\mathbf{x}, \xi) + \varepsilon))}{\sum_{i=0}^m \exp(\mu^{-1}(c_i(\mathbf{x}, \xi) + \varepsilon))}, \quad j = 0, 1, \ldots, m,$$
$$\widetilde{\lambda}_j(\mathbf{x}, \varepsilon, \xi, \mu) = \frac{\exp(\mu^{-1} c_j(\mathbf{x}, \xi))}{\sum_{i=0}^m \exp(\mu^{-1} c_i(\mathbf{x}, \xi))}, \quad j = 0, 1, \ldots, m, \tag{A1}$$

where we define both $c_0(\mathbf{x}, \xi) \equiv 0$ and $c_0(\mathbf{x}, \xi) + \varepsilon \equiv 0$ for consistency of notation. Then simple calculation shows

$$H_2(\mathbf{x}, \xi, \mu) = \sum_{j=0}^m \widetilde{\lambda}_j(\mathbf{x}, \varepsilon, \xi, \mu) c_j(\mathbf{x}, \xi) - \mu \sum_{j=0}^m \widetilde{\lambda}_j(\mathbf{x}, \varepsilon, \xi, \mu) \log[\widetilde{\lambda}_j(\mathbf{x}, \varepsilon, \xi, \mu)].$$

By calculating the derivative of $\bar{g}(\mathbf{x}, \varepsilon, \mu)$ with respect to $\mu$, we have

$$\begin{aligned}
&\frac{\partial}{\partial \mu} \bar{g}(\mathbf{x}, \varepsilon, \mu) \\
&= \mu^{-1} E\left[ H_1(\mathbf{x}, \varepsilon, \xi, \mu) - \sum_{j=0}^m \bar{\lambda}_j(\mathbf{x}, \varepsilon, \xi, \mu) \left(c_j(\mathbf{x}, \xi) + \varepsilon\right) \right] \\
&\quad - \mu^{-1} E\left[ H_2(\mathbf{x}, \xi, \mu) - \sum_{j=0}^m \widetilde{\lambda}_j(\mathbf{x}, \varepsilon, \xi, \mu) c_j(\mathbf{x}, \xi) \right] \\
&\quad + \log(m+1) \\
&\ge \mu^{-1} E\left[ [c(\mathbf{x}, \xi) + \varepsilon]^+ - \sum_{j=0}^m \bar{\lambda}_j(\mathbf{x}, \varepsilon, \xi, \mu) \left(c_j(\mathbf{x}, \xi) + \varepsilon\right) \right] \\
&\quad - \mu^{-1} E\left[ H_2(\mathbf{x}, \xi, \mu) - \sum_{j=0}^m \widetilde{\lambda}_j(\mathbf{x}, \varepsilon, \xi, \mu) c_j(\mathbf{x}, \xi) \right] \\
&\quad + \log(m+1) \\
&\ge -\mu^{-1} E\left[ H_2(\mathbf{x}, \xi, \mu) - \sum_{j=0}^m \widetilde{\lambda}_j(\mathbf{x}, \varepsilon, \xi, \mu) c_j(\mathbf{x}, \xi) \right] \\
&\quad + \log(m+1) \\
&= E\left[ \sum_{j=0}^m \widetilde{\lambda}_j(\mathbf{x}, \varepsilon, \xi, \mu) \log[\widetilde{\lambda}_j(\mathbf{x}, \varepsilon, \xi, \mu)] \right] \\
&\quad + \log(m+1) \text{ (here } 0 \log 0 \equiv 0) \\
&\ge \min\left\{ \sum_{j=0}^m \lambda_j \log \lambda_j : \sum_{j=0}^m \lambda_j = 1, \lambda_j \ge 0, j = 0, \ldots, m \right\} \\
&\quad + \log(m+1) \\
&= 0.
\end{aligned}$$

Therefore, the function $\bar{g}(\mathbf{x}, \varepsilon, \mu)$ is non-decreasing in $\mu$. Thus, we have $\bar{g}(\mathbf{x}, \varepsilon, \mu_1) \le \bar{g}(\mathbf{x}, \varepsilon, \mu_2)$ for $0 < \mu_1 \le \mu_2$ and $\Phi(\varepsilon, \mu_2) \subset \Phi(\varepsilon, \mu_1)$.

Now we prove (b). As $\Phi(\varepsilon, \mu)$ is monotone in $\mu$, it follows from Exercise 4.3 of Rockafellar and Wets (1998) that

$\lim_{\mu\searrow 0}\Phi(\varepsilon,\mu)$ exists. From (a) and the fact that $\Omega_\varepsilon$ is a closed set (ensured by Assumption 3), we can easily get the inclusion $\lim_{\mu\searrow 0}\Phi(\varepsilon,\mu)\subset\Omega_\varepsilon$. We only need to prove the opposite inclusion. For any $x\in\Omega_\varepsilon^I$, let $\eta=g_2(\mathbf{x})-g_1(\mathbf{x},\varepsilon)$. Then, $\eta>0$. Let $\bar{\mu}=\eta/[4\log(m+1)]$. Then from Equation (10) we have

$$\begin{aligned}\bar{g}_1(\mathbf{x},\varepsilon,\mu)-\bar{g}_2(\mathbf{x},\mu)&\le g_1(\mathbf{x},\varepsilon)-g_2(\mathbf{x})\\&\quad+2\mu\log(m+1)\le-\eta+\eta/2=-\eta/2<0\end{aligned}$$

for any $\mu\in(0,\bar{\mu})$, which implies that $\mathbf{x}\in\Phi(\varepsilon,\mu)$ for any $\mu\in(0,\bar{\mu})$. Therefore, we have $\mathbf{x}\in\lim_{\mu\searrow 0}\Phi(\varepsilon,\mu)$. Thus, we obtain that $\lim_{\mu\searrow 0}\Phi(\varepsilon,\mu)\supset\Omega_\varepsilon^I$. Since $\lim_{\mu\searrow 0}\Phi(\varepsilon,\mu)$ is a closed set, we have by Assumption 3, $\lim_{\mu\searrow 0}\Phi(\varepsilon,\mu)\supset\Omega_\varepsilon$. Therefore, $\lim_{\mu\searrow 0}\Phi(\varepsilon,\mu)=\Omega_\varepsilon$.

As $\lim_{\mu\searrow 0}\Phi(\varepsilon,\mu)=\Omega_\varepsilon$ holds and $\bar{g}(\mathbf{x},\varepsilon,\mu)$ is continuous in $\mathbf{x}$, the rest of (b) can be proven using the same argument for Theorem 2 of Hong *et al.* (2011). ■

### A2: Proof of Theorem 2

For consistency of notation, we define $\bar{g}_2(\mathbf{x},\varepsilon,\mu)=\bar{g}_2(\mathbf{x},\mu)$ and $g_2(\mathbf{x},\varepsilon)=g_2(\mathbf{x})$ throughout the proof of Theorem 2. To prove Theorem 2, we need the following lemma.

**Lemma A1.** *Let* $\mathbf{x}\in X$. *Then:*

(a) *for* $i=1,2$, $\lim_{\mu\searrow 0}\nabla_x\bar{g}_i(\mathbf{x},\varepsilon,\mu)$ *exists and belongs to* $\partial_x g_i(\mathbf{x},\varepsilon)$;
(b) *for* $i=1,2$.

$$\limsup_{\mathbf{x}'\to\mathbf{x},\mu\searrow 0}\{\nabla_x\bar{g}_i(\mathbf{x}',\varepsilon,\mu)\}\subset\partial_x g_i(\mathbf{x},\varepsilon).$$

**Proof.** Without loss of generality, we only prove the conclusion for $i=1$.

(a) Since:

$$\nabla_x H_1(\mathbf{x},\varepsilon,\boldsymbol{\xi},\mu)=\sum_{j=0}^m\bar{\lambda}_j(\mathbf{x},\varepsilon,\boldsymbol{\xi},\mu)\nabla_x(c_j(\mathbf{x},\boldsymbol{\xi})+\varepsilon),$$

where $\bar{\lambda}_j(\mathbf{x},\varepsilon,\boldsymbol{\xi},\mu)$, $j=0,1,\ldots,m$ are defined by Equation (A1), we have

$$\begin{aligned}\nabla_x\Psi_1(\mathbf{x},\varepsilon,\mu)&=E[\nabla_x H_1(\mathbf{x},\varepsilon,\boldsymbol{\xi},\mu)]\\&=E\left[\sum_{j=0}^m\bar{\lambda}_j(\mathbf{x},\varepsilon,\boldsymbol{\xi},\mu)\nabla_x(c_j(\mathbf{x},\boldsymbol{\xi})+\varepsilon)\right].\end{aligned}$$

Note that we can rewrite $\bar{\lambda}_j(\mathbf{x},\varepsilon,\boldsymbol{\xi},\mu)$, $j=0,1,\ldots,m$ as

$$\begin{aligned}&\bar{\lambda}_j(\mathbf{x},\varepsilon,\boldsymbol{\xi},\mu)\\&\quad=\frac{\exp(\mu^{-1}(c_j(\mathbf{x},\boldsymbol{\xi})+\varepsilon-[c(\mathbf{x},\boldsymbol{\xi})+\varepsilon]^+))}{\sum_{i=0}^m\exp(\mu^{-1}(c_i(\mathbf{x},\boldsymbol{\xi})+\varepsilon-[c(\mathbf{x},\boldsymbol{\xi})+\varepsilon]^+))},\\&\quad j=0,1,\ldots,m.\end{aligned}$$

Then it can be verified that

$$\lim_{\mu\searrow 0}\bar{\lambda}_j(\mathbf{x},\varepsilon,\boldsymbol{\xi},\mu)=\begin{cases}\dfrac{1}{|I(\mathbf{x},\boldsymbol{\xi},\varepsilon)|} & j\in I(\mathbf{x},\boldsymbol{\xi},\varepsilon),\\ 0 & j\notin I(\mathbf{x},\boldsymbol{\xi},\varepsilon),\end{cases}$$

where $I(\mathbf{x},\boldsymbol{\xi},\varepsilon)=\{j: c_j(\mathbf{x},\boldsymbol{\xi})+\varepsilon=[c(\mathbf{x},\boldsymbol{\xi})+\varepsilon]^+, 0\le j\le m\}$. Therefore, we have from Lebesgue Dominated Convergence Theorem that

$$\begin{aligned}&\lim_{\mu\searrow 0}\nabla_x\Psi_1(\mathbf{x},\varepsilon,\mu)\\&\quad=E\left[\lim_{\mu\searrow 0}\sum_{j=0}^m\bar{\lambda}_j(\mathbf{x},\varepsilon,\boldsymbol{\xi},\mu)\nabla_x(c_j(\mathbf{x},\boldsymbol{\xi})+\varepsilon)\right]\\&\quad=E\left[\sum_{j\in I(\mathbf{x},\boldsymbol{\xi},\varepsilon)}\frac{1}{|I(\mathbf{x},\boldsymbol{\xi},\varepsilon)|}\nabla_x(c_j(\mathbf{x},\boldsymbol{\xi})+\varepsilon)\right]\\&\quad\in E[\partial_x[c(\mathbf{x},\boldsymbol{\xi})+\varepsilon]^+]=\partial_x g_1(\mathbf{x},\varepsilon).\end{aligned}$$

From the definition of $\bar{g}_1(\mathbf{x},\varepsilon,\mu)$, we have that (a) holds.

(b) We first prove the following inclusion:

$$\limsup_{\mathbf{x}'\to\mathbf{x},\ \mu\searrow 0}\{\nabla_x H_1(\mathbf{x}',\varepsilon,\boldsymbol{\xi},\mu)\}\subset\partial_x[c(\mathbf{x},\boldsymbol{\xi})+\varepsilon]^+. \quad \text{(A2)}$$

For any $v\in\limsup_{\mathbf{x}'\to\mathbf{x},\ \mu\searrow 0}\{\nabla_x H_1(\mathbf{x}',\varepsilon,\boldsymbol{\xi},\mu)\}$, there exists a sequence $\{(\mathbf{x}_k,\mu_k)\}$ satisfying $\mathbf{x}_k\to\mathbf{x}$ and $\mu_k\searrow 0$ such that $v=\lim_{k\to\infty}\nabla_x H_1(\mathbf{x}_k,\varepsilon,\boldsymbol{\xi},\mu_k)$. From the definition of $I(\mathbf{x},\boldsymbol{\xi},\varepsilon)$, there exists some $\delta>0$ such that

$$c_j(\mathbf{x},\boldsymbol{\xi})+\varepsilon-[c(\mathbf{x},\boldsymbol{\xi})+\varepsilon]^+\le-\delta \text{ for } j\notin I(\mathbf{x},\boldsymbol{\xi},\varepsilon).$$

Since $\mathbf{x}'\to c_j(\mathbf{x}',\boldsymbol{\xi})+\varepsilon-[c(\mathbf{x}',\boldsymbol{\xi})+\varepsilon]^+$ is continuous, for large enough $k>0$ we have

$$c_j(\mathbf{x}_k,\xi)+\varepsilon-[c(\mathbf{x}_k,\xi)+\varepsilon]^+\le-\delta/2 \text{ for } j\notin I(\mathbf{x},\boldsymbol{\xi},\varepsilon).$$

It follows that

$$\bar{\lambda}_j(\mathbf{x}_k,\varepsilon,\boldsymbol{\xi},\mu_k)\le\exp\left(-\mu_k^{-1}\delta/2\right)\to 0 \text{ for } j\notin I(\mathbf{x},\boldsymbol{\xi},\varepsilon).$$

On the other hand, there exist non-negative scalars $\bar{\mu}_j$ for $j\in I(\mathbf{x},\boldsymbol{\xi},\varepsilon)$, such that $\sum_{j\in I(\mathbf{x},\boldsymbol{\xi},\varepsilon)}\bar{\mu}_j=1$ and there exists a subsequence $\{k_i\}$ satisfying

$$\bar{\lambda}_j(\mathbf{x}_{k_i},\boldsymbol{\xi},\varepsilon,\mu_{k_i})\to\bar{\mu}_j \text{ for } j\in I(\mathbf{x},\boldsymbol{\xi},\varepsilon).$$

Therefore,

$$\begin{aligned}v&=\lim_{k\to\infty}\nabla_x H_1(\mathbf{x}_k,\varepsilon,\boldsymbol{\xi},\mu_k)\\&=\lim_{i\to\infty}\nabla_x H_1(\mathbf{x}_{k_i},\varepsilon,\boldsymbol{\xi},\mu_{k_i})\\&=\sum_{j\in I(\mathbf{x},\boldsymbol{\xi},\varepsilon)}\bar{\mu}_j\nabla_x(c_j(\mathbf{x},\boldsymbol{\xi})+\varepsilon)\in\partial_x[c(\mathbf{x},\boldsymbol{\xi})+\varepsilon]^+,\end{aligned}$$

and the inclusion (A2) is proved. Noting (A2), we have from Dominated Convergence for Selection Expectation (Theorem 1.38 in Molchanov (2005)) that

$$\begin{aligned}&\limsup_{\mathbf{x}'\to\mathbf{x},\mu\searrow 0} \nabla_x \Psi_1(\mathbf{x}', \varepsilon, \mu)\\ &\quad= \limsup_{\mathbf{x}'\to\mathbf{x},\mu\searrow 0} E[\nabla_x H_1(\mathbf{x}', \varepsilon, \boldsymbol{\xi}, \mu)]\\ &\quad\subset E\left[\limsup_{\mathbf{x}'\to\mathbf{x},\mu\searrow 0} \nabla_x H_1(\mathbf{x}', \varepsilon, \boldsymbol{\xi}, \mu)\right]\\ &\quad\subset E[\partial_x[c(\mathbf{x}, \boldsymbol{\xi}) + \varepsilon]^+] = \partial_x g_1(\mathbf{x}, \varepsilon).\end{aligned}$$

From the definition of $\bar{g}_1(\mathbf{x}, \varepsilon, \mu)$, we have (b) holds.

■

Now we can prove Theorem 2.

**Proof.** For any $(\mathbf{x}, \lambda) \in \limsup_{\mu\searrow 0} \Lambda(\varepsilon, \mu)$, there exist $\{(\mathbf{x}_k, \lambda_k)\}$ and $\{\mu_k\}$, such that $(\mathbf{x}_k, \lambda_k) \in \Lambda(\varepsilon, \mu_k)$, $\mu_k \searrow 0$ and $(\mathbf{x}_k, \lambda_k) \to (\mathbf{x}, \lambda)$. The inclusion $(\mathbf{x}_k, \lambda_k) \in \Lambda(\varepsilon, \mu_k)$ means:

$$\begin{aligned}&-[\nabla_x h(\mathbf{x}_k) + \lambda_k(\nabla_x \bar{g}_1(\mathbf{x}_k, \varepsilon, \mu_k)\\ &\quad -\nabla_x \bar{g}_2(\mathbf{x}_k, \varepsilon, \mu_k))] \in N_X(\mathbf{x}_k), \qquad \text{(A3)}\\ &\lambda_k[\bar{g}_1(\mathbf{x}_k, \varepsilon, \mu_k) - \bar{g}_2(\mathbf{x}_k, \varepsilon, \mu_k)] = 0,\ \lambda_k \geq 0. \qquad \text{(A4)}\end{aligned}$$

It follows from Lemma A1 that

$$\limsup_{k\to\infty} \{\nabla_x \bar{g}_i(\mathbf{x}_k, \varepsilon, \mu_k)\} \subset \partial_x g_i(\mathbf{x}, \varepsilon),\ i = 1, 2,$$

from which there exist an subsequence $\{k_j\}$, two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ such that

$$\lim_{j\to\infty} \nabla_x \bar{g}_i(\mathbf{x}_{k_j}, \varepsilon, \mu_{k_j}) = \mathbf{v}_i \in \partial_x g_i(\mathbf{x}, \varepsilon),\ i = 1, 2.$$

Noting the outer semi-continuity of $N_X$, letting $j \to \infty$, we have from Equations (A3) and (A4) that

$$\begin{aligned}&-[\nabla_x h(\mathbf{x}) + \lambda(\mathbf{v}_1 - \mathbf{v}_2)] \in N_X(\mathbf{x}),\\ &\quad \mathbf{v}_1 \in \partial_x g_1(\mathbf{x}, \varepsilon),\ \mathbf{v}_2 \in \partial_x g_2(\mathbf{x}, \varepsilon),\\ &\quad \lambda[g_1(\mathbf{x}, \varepsilon) - g_2(\mathbf{x}, \varepsilon)] = 0,\ \lambda \geq 0,\end{aligned}$$

which implies $(\mathbf{x}, \lambda) \in \Lambda(\varepsilon)$. The proof is completed. ■

### A3: Proof of Theorem 3

(a) Note that in Problem $(\mathrm{P}_\mu)$, $\varepsilon$ is fixed. Then, for any $\mathbf{x} \in \Phi(\varepsilon, \mu)$, we have $(\mathbf{x}, \varepsilon) \in Z^o(\mu)$ for the fixed $\varepsilon$. This implies $\mathbf{x} \in \Phi^o(\mu)$. Therefore, $\Phi(\varepsilon, \mu) \subset \Phi^o(\mu)$.
Now we prove the second inclusion. For any $\mathbf{x} \in \Phi^o(\mu)$, by the definition of $\Phi^o(\mu)$ we can find some $t \geq 0$ such that $(\mathbf{x}, t) \in Z^o(\mu)$. We argue that $t \neq 0$. If $t = 0$, we have

$$0 \geq \bar{g}_1(\mathbf{x}, t, \mu) - \bar{g}_2(\mathbf{x}, \mu) = \mu \log(m + 1) > 0.$$

This is a contradiction. Therefore, we must have $t > 0$. From Equation (10), we have

$$g_1(\mathbf{x}, t) - g_2(\mathbf{x}) \leq \bar{g}_1(\mathbf{x}, t, \mu) - \bar{g}_2(\mathbf{x}, \mu) \leq 0.$$

It follows that $(1/t)\{E[[c(\mathbf{x}, \boldsymbol{\xi}) + t]^+] - E[[c(\mathbf{x}, \boldsymbol{\xi})]^+]\} \leq \alpha$, which further implies that $\mathbf{x} \in \Omega$. Therefore, $\Phi^o(\mu) \subset \Omega$. Because the inclusions $\Phi(\varepsilon, \mu) \subset \Phi^o(\mu) \subset \Omega$ hold, we have $\nu \leq \nu^o(\mu) \leq \nu(\varepsilon, \mu)$.

(b) Given any $0 < \mu_1 \leq \mu_2$, consider any $\mathbf{x} \in \Phi^o(\mu_2)$. From the definition of $\Phi^o(\mu_2)$, we can find $\varepsilon \geq 0$ such that $\bar{g}(\mathbf{x}, \varepsilon, \mu_2) \leq 0$. Similarly, as in (a), we can further show that $\varepsilon > 0$. From the proof of (a) of Theorem 1 we have that $\bar{g}(\mathbf{x}, \varepsilon, \mu)$ is non-decreasing in $\mu$. Therefore, $\bar{g}(\mathbf{x}, \varepsilon, \mu_1) \leq \bar{g}(\mathbf{x}, \varepsilon, \mu_2)$. It follows that $\bar{g}(\mathbf{x}, \varepsilon, \mu_1) \leq 0$, which implies $\mathbf{x} \in \Phi^o(\mu_1)$. Therefore, $\Phi^o(\mu_2) \subset \Phi^o(\mu_1)$.

(c) From (b) we have $\lim_{\mu\searrow 0} \Phi^o(\mu)$ exists. From (a) and the fact that $\Omega$ is a closed set (ensured by Assumption 4), we have $\lim_{\mu\searrow 0} \Phi^o(\mu) \subset \Omega$. Now we prove the opposite inclusion. For any $\mathbf{x} \in \Omega^I$, because $\inf_{t>0} E[\pi(c(\mathbf{x}, \boldsymbol{\xi}), t)] = \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) \geq 0\} < \alpha$, we can find $\varepsilon > 0$ such that $E[\pi(c(\mathbf{x}, \boldsymbol{\xi}), \varepsilon)] < \alpha$. This implies $g_1(\mathbf{x}, \varepsilon) - g_2(\mathbf{x}) < 0$. Since $\bar{g}(\mathbf{x}, \varepsilon, \mu)$ converges to $g_1(\mathbf{x}, \varepsilon) - g_2(\mathbf{x})$ as $\mu \searrow 0$, we can find $\bar{\mu} > 0$ such that $\bar{g}(\mathbf{x}, \varepsilon, \bar{\mu}) < 0$. This means $\mathbf{x} \in \Phi^o(\bar{\mu}) \subset \lim_{\mu\searrow 0} \Phi^o(\mu)$. Therefore, $\Omega^I \subset \lim_{\mu\searrow 0} \Phi^o(\mu)$. Since $\lim_{\mu\searrow 0} \Phi^o(\mu)$ is a closed set, we have by Assumption 4, that $\Omega \subset \lim_{\mu\searrow 0} \Phi^o(\mu)$. Therefore, $\lim_{\mu\searrow 0} \Phi^o(\mu) = \Omega$. As $\lim_{\mu\searrow 0} \Phi^o(\mu) = \Omega$ holds and $\bar{g}(\mathbf{x}, t, \mu)$ is continuous in $(\mathbf{x}, t)$, using the same argument for Theorem 2 of Hong *et al.* (2011), we have $\lim_{\mu\searrow 0} \nu^o(\mu) = \nu$ and $\lim_{\mu\searrow 0} \mathbb{D}(S^o(\mu), S) = 0$. This concludes the proof of the theorem. ■

### A4: Asymptotic behaviors of strengthened smooth approximation

Let con[$A$] denote the convex hull of a set $A$ (Rockafellar and Wets, 1997). We define

$$\begin{aligned}&\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}\\ &\quad= \mathrm{con}\left[\limsup_{\mathbf{x}'\to\mathbf{x},t\searrow 0,\mu\searrow 0} t^{-1}\{\nabla_x \Psi_1(\mathbf{x}, t, \mu) - \nabla_x \Psi_2(\mathbf{x}, \mu)\}\right],\end{aligned}$$

and

$$\Lambda = \left\{(\mathbf{x}, \lambda) \in \Omega \times \Re_+ : \begin{array}{c} 0 \in \nabla_x h(\mathbf{x}) + \lambda \partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\} \\ \quad + N_X(\mathbf{x}) \\ \lambda[\Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\} - \alpha] = 0 \end{array}\right\}.$$

The set such as $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ is often used in nonsmooth optimization, especially when we use a sequence of functions to approximate a function; see, e.g., Rockafellar and Wets (1997). In this article we still call the set $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ the subdifferential of the function $\Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ and call $\Lambda$ the set of stationary pairs of the JCCP. In the case where $\Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ is locally Lipschitz continuous, following Clarke (1983), the generalized

directional derivative of $\Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$, at point $\mathbf{x}$ in the direction $\mathbf{v}$, is defined as

$$\overset{\circ}{\Pr}\{c(\mathbf{x}, \boldsymbol{\xi}) > 0; \mathbf{v}\} := \limsup_{y \to \mathbf{x}, \lambda \downarrow 0} \frac{\Pr\{c(y + \lambda \mathbf{v}, \boldsymbol{\xi}) > 0\} - \Pr\{c(y, \boldsymbol{\xi}) > 0\}}{\lambda},$$

and Clarke's generalized gradient of $\Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$, at point $\mathbf{x}$, is defined as

$$\bar{\partial}_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\} := \left\{\boldsymbol{\zeta} \in \Re^d : \overset{\circ}{\Pr}\{c(\mathbf{x}, \boldsymbol{\xi}) > 0; \mathbf{v}\} \geq \langle \mathbf{v}, \boldsymbol{\zeta} \rangle \text{ for all } \mathbf{v} \text{ in } \Re^d\right\}.$$

Clarke (1983) showed that $\bar{\partial}_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ is well defined and is a compact set. More important, because under Assumption 1

$$\frac{1}{t}\{\Psi_1(\mathbf{x}, t, \mu) - \Psi_2(\mathbf{x}, \mu)\} \to \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}, \quad \text{as } \mu \to 0, t \to 0,$$

it follows from Corollary 8.47 and Theorem 9.61 of Rockafellar and Wets (1997) that $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ is also a compact set, and $\bar{\partial}_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\} \subset \partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$. When the function $\Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ is not locally Lipschitz continuous, Clarke's generalized gradient is not defined, but the set $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ above is still available. Note especially that, in the smooth case, we have the following proposition.

**Proposition A1.** *Suppose that Assumptions 1 to 5 of Hong et al.* (2011) *are satisfied. Then* $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\} = \nabla_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$.

**Proof.** Consider any $\mathbf{x}_k \to \mathbf{x}$, $t_k \searrow 0$, and $\mu_k \searrow 0$. Because Assumptions 1 to 5 of Hong *et al.* (2011) are satisfied, we have $\Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ is continuously differentiable and $\Psi_1(\mathbf{x}, t, \mu)$ and $\Psi_2(\mathbf{x}, \mu)$ are continuously differentiable in $\mathbf{x}$ for any $t > 0$ and $\mu > 0$. This implies that we can change the order of the operators $\lim_{k\to\infty}$ and $\nabla_x$. Furthermore, by the Lebesgue Dominated Convergence Theorem, we can change the order of the operators $\lim_{k\to\infty}$ and $E[\cdot]$. Therefore, we have

$$\begin{aligned}
&\lim_{k\to\infty} \frac{1}{t_k}\left[\nabla_x \Psi_1(\mathbf{x}_k, t_k, \mu_k) - \nabla_x \Psi_2(\mathbf{x}_k, \mu_k)\right] \\
&= \nabla_x E\left[\lim_{k\to\infty} \frac{1}{t_k}\left[\mu_k \log\left[1 + \sum_{i=1}^m \exp\left\{\frac{1}{\mu_k}(c_i(\mathbf{x}_k, \boldsymbol{\xi}) + t_k)\right\}\right]\right.\right. \\
&\quad \left.\left. - \mu_k \log\left[1 + \sum_{i=1}^m \exp\left\{\frac{1}{\mu_k} c_i(\mathbf{x}_k, \boldsymbol{\xi})\right\}\right]\right]\right] \\
&= \nabla_x E\left[\lim_{k\to\infty} \frac{\sum_{i=1}^m \exp\left\{(1/\mu_k)(c_i(\mathbf{x}_k, \xi) + \tilde{t}_k)\right\}}{1 + \sum_{i=1}^m \exp\left\{(1/\mu_k)(c_i(\mathbf{x}_k, \xi) + \tilde{t}_k)\right\}}\right] \\
&= \nabla_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\},
\end{aligned}$$

where the second equality follows from the mean-value theorem and $\tilde{t}_k \in [0, t_k]$ is some value that depends on $\boldsymbol{\xi}, \mathbf{x}_k, t_k, \mu_k$. To justify the last equality, consider any $\boldsymbol{\xi} \in \Xi$. Suppose $c(\mathbf{x}, \boldsymbol{\xi}) < 0$; i.e., $c_i(\mathbf{x}, \boldsymbol{\xi}) < 0$ for all $i = 1, \ldots, m$. Since $t_k \searrow 0$ and $c_i(y, \boldsymbol{\xi})$ is continuous in $y$, we have for $k$ large enough $c_i(\mathbf{x}_k, \boldsymbol{\xi}) + \tilde{t}_k < 0$ and $c_i(\mathbf{x}_k, \boldsymbol{\xi}) + \tilde{t}_k \to c_i(\mathbf{x}, \boldsymbol{\xi}) < 0$. Therefore $a := \sum_{i=1}^m \exp\left\{\mu_k^{-1}(c_i(\mathbf{x}_k, \xi) + \tilde{t}_k)\right\} \to 0$, which implies $a/(1 + a) \to 0$. Suppose $c(\mathbf{x}, \boldsymbol{\xi}) > 0$; i.e., there exists $j$ such that $c_j(\mathbf{x}, \boldsymbol{\xi}) > 0$. Then, $\exp\left\{\mu_k^{-1}(c_j(\mathbf{x}_k, \boldsymbol{\xi}) + \tilde{t}_k)\right\} \to +\infty$. It follows that $a := \sum_{i=1}^m \exp\left\{\mu_k^{-1}(c_i(\mathbf{x}_k, \boldsymbol{\xi}) + \tilde{t}_k)\right\} \to +\infty$, which implies $a/(1 + a) \to 1$. Note that $c(\mathbf{x}, \boldsymbol{\xi})$ is a continuous random variable. We have the last equality holds. Recall the definition of the outer limit "lim sup," we obtain $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\} = \nabla_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$. ■

Proposition A1 suggests that in the smooth case, the set $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ defined in this article coincides with the conventional gradient, and consequently $\Lambda$ defined in this article is consistent with the set of stationary pairs $\Lambda_0$ of Hong *et al.* (2011). The above facts suggest that it may be appropriate to depict the optimality conditions for the JCCP using $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ and $\Lambda$, although we believe that other better alternatives may exist. To facilitate the discussion, we make the following assumption about the original JCCP.

**Assumption A1.** For every $\mathbf{x} \in X$ and $\mathbf{v} \in \partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$, the following regularity condition holds:

$$\left.\begin{array}{c} 0 \in \lambda \mathbf{v} + N_X(\mathbf{x}) \\ \lambda \geq 0; \lambda\left[\Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\} - \alpha\right] = 0 \end{array}\right\} \Longrightarrow \lambda = 0.$$

Assumption A1 is a non-smooth analogy to Assumption 6 of Hong *et al.* (2011), whereas Assumption 6 of Hong *et al.* (2011) is a very commonly used regularity condition in numerical optimization. Readers may refer to the discussion that follows Assumption 6 in Hong *et al.* (2011).

Let $(\mathbf{x}, t, \lambda)$ be any KKT pair of Problem $(\mathrm{P}_\mu^o)$, and $\Lambda^o(\mu)$ be the set of KKT pairs of Problem $(\mathrm{P}_\mu^o)$. Let $\Lambda^\Pi$ be the projection of $\Lambda$ on $X$. Then we have the following theorem.

**Theorem A1.** *Suppose that* $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ *is bounded, and Assumptions* 1, 2, 4, *and* A1 *are satisfied. For any* $\{\mu_k\}$ *with* $\mu_k \to 0$, *Suppose* $(\mathbf{x}_k, t_k, \lambda_k) \in \Lambda^o(\mu_k)$, $\mathbf{x}_k \to \bar{\mathbf{x}}$ *and* $t_k \to 0$. *Then* $\bar{\mathbf{x}} \in \Lambda^\Pi$.

**Proof.** Following the definition of KKT pairs, $(\mathbf{x}_k, t_k, \lambda_k) \in \Lambda^o(\mu_k)$ reads:

$$\begin{aligned}
&(\mathbf{x}_k, t_k, \lambda_k) \in Z^o(\mu_k) \times \Re_+ \\
&0 \in \nabla_x h(\mathbf{x}_k) + \lambda_k\left[\nabla_x \bar{g}_1(\mathbf{x}_k, t_k, \mu_k) - \nabla_x \bar{g}_2(\mathbf{x}_k, \mu_k)\right] \\
&\quad + N_X(\mathbf{x}_k) \qquad \text{(A5)} \\
&0 \in \nabla_t h(\mathbf{x}_k) + \lambda_k\left[\nabla_t \bar{g}_1(\mathbf{x}_k, t_k, \mu_k) - \nabla_t \bar{g}_2(\mathbf{x}_k, \mu_k)\right] \\
&\quad + N_{[0,+\infty)}(t_k) \qquad \text{(A6)} \\
&\lambda_k\left[\bar{g}_1(\mathbf{x}_k, t_k, \mu_k) - \bar{g}_2(\mathbf{x}_k, \mu_k)\right] = 0.
\end{aligned}$$

It follows from Theorem 3 that $\bar{\mathbf{x}} \in \Omega$. Suppose there exists $\{k_j\}$ such that $\lambda_{k_j} = 0$, $j = 1, 2, \ldots$. Then from Equation

(A5) we have $0 \in \nabla_x h(\mathbf{x}_{k_j}) + N_X(\mathbf{x}_{k_j})$. Letting $j \to \infty$, and noting the fact that $h$ is continuously differentiable, we obtain $0 \in \nabla_x h(\bar{\mathbf{x}}) + N_X(\bar{\mathbf{x}})$, which indicates that $\bar{\mathbf{x}}$ is a global optimal solution of the JCCP. For this case it can be verified that $(\bar{\mathbf{x}}, 0) \in \Lambda$ and thus $\bar{\mathbf{x}} \in \Lambda^{\Pi}$. Therefore, in the rest of the proof we can assume without loss of generality that starting from a sufficiently large $k$, $\lambda_k \neq 0$.

Because $(\mathbf{x}_k, t_k) \in Z^o(\mu_k)$, as in the proof of Theorem 3, we can verify that $t_k > 0$, which implies $N_{[0,+\infty)}(t_k) = \{0\}$. From Equation (A6) we immediately obtain $\nabla_t \bar{g}_1(\mathbf{x}_k, t_k, \mu_k) = 0$, which is equivalent to the following equality by changing the order of $\nabla_t$ and $E[\cdot]$:

$$E\left[\frac{\sum_{i=1}^m \exp\left\{\mu_k^{-1}(c_i(\mathbf{x}_k, \boldsymbol{\xi}) + t_k)\right\}}{1 + \sum_{i=1}^m \exp\left\{\mu_k^{-1}(c_i(\mathbf{x}_k, \boldsymbol{\xi}) + t_k)\right\}}\right] - \alpha = 0. \quad \text{(A7)}$$

Note that $t_k \to 0$ and Assumption 1 holds. Letting $k \to \infty$ on both sides of Equation (A7) and using the argument in the proof of Proposition A1 again we obtain $\Pr\{c(\bar{\mathbf{x}}, \boldsymbol{\xi}) > 0\} - \alpha = 0$.

Because $t_k \neq 0$, Equations (A5) and (A6) can be rewritten as

$$0 \in \nabla_x h(\mathbf{x}_k) + \lambda_k t_k \left[\frac{\nabla_x \Psi_1(\mathbf{x}_k, t_k, \mu_k) - \nabla_x \Psi_2(\mathbf{x}_k, \mu_k)}{t_k}\right] + N_X(\mathbf{x}_k) \quad \text{(A8)}$$

$$\lambda_k t_k [\nabla_t \bar{g}_1(\mathbf{x}_k, t_k, \mu_k)] = 0. \quad \text{(A9)}$$

Now we argue $\lambda_k t_k \not\to +\infty$. Suppose not. Note that $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ is always closed (Rockafellar and Wets, 1998). The assumption that $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ is bounded implies the compactness of $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$. Therefore, by passing to a subsequence if necessary, we can assume $t_k^{-1}\{\nabla_x \Psi_1(\mathbf{x}_k, t_k, \mu_k) - \nabla_x \Psi_2(\mathbf{x}_k, \mu_k)\}$ converges to a vector $\bar{\mathbf{v}} \in \partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$. Dividing both sides of Equation (A8) by $\lambda_k t_k$ and letting $k \to \infty$, we obtain $0 \in \bar{\mathbf{v}} + N_X(\bar{\mathbf{x}})$. Dividing both sides of Equation (A9) by $\lambda_k t_k$ and letting $k \to \infty$, by the above analysis we obtain $\Pr\{c(\bar{\mathbf{x}}, \boldsymbol{\xi}) > 0\} - \alpha = 0$. This contradicts Assumption A1. The contradiction shows that $\lambda_k t_k \not\to +\infty$.

Because $\lambda_k t_k \not\to +\infty$, $\{\lambda_k t_k\}$ has a convergent subsequence. By passing to a subsequence if necessary, we can assume $\{\lambda_k t_k\}$ converges; i.e., $\lambda_k t_k \to \bar{\lambda}$ for some $\bar{\lambda} \geq 0$. Again, because $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ is compact, we can assume that $t_k^{-1}\{\nabla_x \Psi_1(\mathbf{x}_k, t_k, \mu_k) - \nabla_x \Psi_2(\mathbf{x}_k, \mu_k)\}$ converges to a vector $\bar{\mathbf{v}} \in \partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$. Letting $k \to \infty$ on both sides of Equation (A8) yields

$$0 \in \nabla_x h(\bar{\mathbf{x}}) + \bar{\lambda}\bar{\mathbf{v}} + N_X(\bar{\mathbf{x}}).$$

Letting $k \to \infty$ on both sides of Equation (A9) we obtain

$$\bar{\lambda}[\Pr\{c(\bar{\mathbf{x}}, \boldsymbol{\xi}) > 0\} - \alpha] = 0.$$

This shows $(\bar{\mathbf{x}}, \bar{\lambda}) \in \Lambda$, which means $\bar{\mathbf{x}} \in \Lambda^{\Pi}$. The proof is finished. ■

Note that when $\Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ is smooth, $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ degenerates to a singleton and thus is bounded. When $\Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ is locally Lipschitz continuous, $\partial_x \Pr\{c(\mathbf{x}, \boldsymbol{\xi}) > 0\}$ is compact and thus is still bounded. Theorem A1 essentially shows that the $\mathbf{x}$-component of the cluster point (as $\mu \to 0$) of the sequence of KKT points of Problem $(\mathrm{P}^o_\mu)$ is a stationary point of the original JCCP. Therefore, the KKT points of Problem $(\mathrm{P}^o_\mu)$ also serve as a good approximation to the possible stationary points of Problem (3) or of the original JCCP.

As we relax $\varepsilon$ to a variable $t$ in Problem $(\mathrm{P}^o_\mu)$, a natural question is then whether the $t$-component of the KKT point of Problem $(\mathrm{P}^o_\mu)$ will finally converge to zero as $\mu \searrow 0$. The answer is yes in most real situations, but there exist exceptions. We briefly discuss it in the following. Suppose now $(\bar{\mathbf{x}}(\mu), \bar{t}(\mu))$ is a KKT point of Problem $(\mathrm{P}^o_\mu)$. If the JCCP can attain its optimal value at some interior points of its feasible region, then from Theorem 3 Problem $(\mathrm{P}^o_\mu)$ can also attain its optimal value at some interior points of its feasible region for any sufficiently small $\mu$. In this case, $\bar{g}(\mathbf{x}, t, \mu) \leq 0$ does not need to be tight at $(\bar{\mathbf{x}}(\mu), \bar{t}(\mu))$, and $\bar{\mathbf{x}}(\mu)$ may finally converge to some optimal solution $\bar{\mathbf{x}}(0)$ of the JCCP as $\mu \searrow 0$ while $\bar{t}(\mu)$ has freedom to take multiple values and does not need to converge to zero. Note that in this case we still have $\bar{\mathbf{x}}(0) \in \Lambda^{\Pi}$ because $(\bar{\mathbf{x}}(0), 0) \in \Lambda$.

Consider now the more interesting case where all of the optimal solutions of the JCCP make the joint chance constraint tight. By passing to a subsequence if necessary, we assume that $(\bar{\mathbf{x}}(0), \bar{t}(0))$ is the limit of $\{(\bar{\mathbf{x}}(\mu), \bar{t}(\mu))\}$ as $\mu \searrow 0$. Since the JCCP cannot attain its optimal value at any feasible points at which the chance constraint is not tight, we have $\bar{g}(\mathbf{x}, t, \mu) = 0$ at $(\bar{\mathbf{x}}(\mu), \bar{t}(\mu))$. (It can be verified that if $\bar{g}(\mathbf{x}, t, \mu) < 0$ at $(\bar{\mathbf{x}}(\mu), \bar{t}(\mu))$, then $\bar{\mathbf{x}}(\mu)$ is an optimal solution of the JCCP and at the same time $\Pr\{c(\bar{\mathbf{x}}(\mu), \boldsymbol{\xi}) > 0\} < \alpha$.) It follows that

$$\frac{\Psi_1(\bar{\mathbf{x}}(\mu), \bar{t}(\mu), \mu) - \bar{g}_2(\bar{\mathbf{x}}(\mu), \mu)}{\bar{t}(\mu)} = \alpha. \quad \text{(A10)}$$

Because $\Psi_1(\mathbf{x}, t, \mu)$ converges uniformly to $E[[c(\mathbf{x}, \boldsymbol{\xi}) + t]^+]$ and $\bar{g}_2(\mathbf{x}, \mu)$ converges uniformly to $E\left[[c(\mathbf{x}, \boldsymbol{\xi})]^+\right]$ as $\mu \searrow 0$, letting $\mu \searrow 0$ on both sides of Equation (A10) we have

$$\frac{E\left[[c(\bar{\mathbf{x}}(0), \boldsymbol{\xi}) + \bar{t}(0)]^+\right] - E\left[[c(\bar{\mathbf{x}}(0), \boldsymbol{\xi})]^+\right]}{\bar{t}(0)} = \alpha, \quad \text{(A11)}$$

provided that $\bar{t}(0) \neq 0$. Note that it may happen that the random variable $c(\bar{\mathbf{x}}(0), \boldsymbol{\xi})$ has no mass at some neighborhood of zero. Therefore, in this subcase, $\bar{t}(0)$ still does not need to be zero while the left-hand side of Equation (A11) can still be equal to $\inf_{t>0} E[\pi(c(\bar{\mathbf{x}}(0), \boldsymbol{\xi}), t)]$. Finally, we show $\bar{t}(0) \neq 0$ only when $c(\bar{\mathbf{x}}(0), \boldsymbol{\xi})$ has no mass between $(-\bar{t}(0), 0]$. From the proof of

Theorem A1 we have $\bar{t}(\mu)$ minimizes $\bar{g}_1(\bar{\mathbf{x}}(\mu), t, \mu)$. Observe that $\bar{g}_1(\bar{\mathbf{x}}(\mu), t, \mu) \to E[[c(\bar{\mathbf{x}}(0), \boldsymbol{\xi}) + t]^+] - \alpha t$ uniformly on $\Re^+$. Then $t = \bar{t}(0)$ minimizes $E[[c(\bar{\mathbf{x}}(0), \boldsymbol{\xi}) + t]^+] - \alpha t$, which implies $\Pr\{c(\bar{\mathbf{x}}(0), \xi) > -\bar{t}(0)\} \leq \alpha$ (Rockafellar and Uryasev, 2002). On the other hand, from Equation (A11) we have

$$E\left[[c(\bar{\mathbf{x}}(0), \boldsymbol{\xi}) + \bar{t}(0)]^+\right] - \alpha\bar{t}(0) = E[[c(\bar{\mathbf{x}}(0), \boldsymbol{\xi})]^+].$$

This means that $t = 0$ also minimizes $E[[c(\bar{\mathbf{x}}(0), \boldsymbol{\xi}) + t]^+] - \alpha t$. Noting that under Assumption 1, $E[[c(\bar{\mathbf{x}}(0), \boldsymbol{\xi}) + t]^+] - \alpha t$ is differentiable at $t = 0$, we have $\Pr\{c(\bar{\mathbf{x}}(0), \boldsymbol{\xi}) > 0\} = \alpha$. Therefore,

$$\alpha = \Pr\{c(\bar{\mathbf{x}}(0), \boldsymbol{\xi}) > 0\} \leq \Pr\{c(\bar{\mathbf{x}}(0), \boldsymbol{\xi}) > -t(0)\} \leq \alpha.$$

This means $c(\bar{\mathbf{x}}(0), \boldsymbol{\xi})$ has no mass between $(-\bar{t}(0), 0]$.

The above analysis also confirms that if for any $\delta > 0$, $c(\bar{\mathbf{x}}(0), \boldsymbol{\xi})$ has mass in $[-\delta, 0]$ (this includes the case where $c(\bar{\mathbf{x}}(0), \boldsymbol{\xi})$ has a continuous density at a neighborhood of zero), then we must have $\bar{t}(0) = 0$. In the numerical experiments we observed that for both classes of our test problems the $t$-component of the solutions converges to zero. It is worthwhile noting that the above multiple scenarios further strengthen the advantage of treating $\varepsilon$ as a decision variable: it can automatically identify these different scenarios.

### A5: Proof of Theorem 4

1. For any $k \geq 0$, $\mathbf{z}_{k+1}$ is an optimal solution of Problem CP$(\mu, \mathbf{z}_k)$. Thus, $\mathbf{z}_{k+1} \in Z(\mu, \mathbf{z}_k)$. Since $Z(\mu, y) \subset Z^o(\mu)$ for any $y \in Z$, we have $\mathbf{z}_{k+1} \in Z^o(\mu)$. Therefore, $\{\mathbf{z}_k\} \subset Z^o(\mu)$. Note that $\mathbf{z}_{k+1}$ is an optimal solution of Problem CP$(\mu, \mathbf{z}_k)$ and $\mathbf{z}_k$ is a feasible solution of Problem CP$(\mu, \mathbf{z}_k)$. We have $h(\mathbf{z}_{k+1}) \leq h(\mathbf{z}_k)$. This shows that $\{h(\mathbf{z}_k)\}$ is non-increasing. Because $X$ is compact and $h(\cdot)$ is continuous, we have that $\{h(\mathbf{z}_k)\}$ is bounded. It follows that $\{h(\mathbf{z}_k)\}$ is convergent.
2. It follows from Lemma 6 of Hong *et al.* (2011) that the result holds.
3. It follows from Property 3 of Hong *et al.* (2011) and corresponding proof that the result holds. ■

### A6: Proof of Theorem 5

Note that Assumption 2 holds and $\xi_1, \xi_2, \ldots, \xi_n$ are i.i.d. For any $\mathbf{z} \in Z$, by the strong law of large numbers (Durrett, 2005), we have that

$$\bar{g}_{1,n}(\mathbf{z}, \mu) - \left[\bar{g}_{2,n}(y, \mu) + \frac{1}{n}\sum_{j=1}^{n} \nabla_z H_2(y, \xi_j, \mu)^{\mathrm{T}} (\mathbf{z} - y)\right] \tag{A12}$$

converges to

$$\bar{g}_1(\mathbf{z}, \mu) - \left[\bar{g}_2(y, \mu) + \nabla_z \bar{g}_2(y, \mu)^{\mathrm{T}} (\mathbf{z} - y)\right] \tag{A13}$$

w.p.1. as $n \to \infty$. Because both Equations (A12) and (A13) are convex functions of $\mathbf{z}$, we have from Theorem 7.50 of Shapiro *et al.* (2009) that Equation (A12) converges to Equation (A13) w.p.1. uniformly on $\mathbf{z}$ as $n \to \infty$. Given Slater's condition, it follows from Theorem 5.5 of Shapiro *et al.* (2009) and the discussion that follows that the results of this theorem hold. ■

## Biographies

Zhaolin Hu is an Assistant Professor in the School of Economics and Management at Tongji University in Shanghai, China. He obtained his Ph.D. degree in Industrial Engineering and Logistics Management from the Hong Kong University of Science and Technology and his B.Sc. degree in Mathematics and Applied Mathematics from Zhejiang University. His current research interests include stochastic optimization, robust optimization, Monte Carlo methods, risk management, and environmental issues.

L. Jeff Hong is a Professor in the Department of Industrial Engineering and Logistics Management and the Director of the Financial Engineering Laboratory at the Hong Kong University of Science and Technology. He received his Ph.D. degree from Northwestern University. His research interests include Monte Carlo methods, stochastic modeling, stochastic optimization, and financial risk management.

Liwei Zhang received his Ph.D. degree from the Department of Applied Mathematics at Dalian University of Technology in 1998. He is now a Professor in the School of Mathematical Sciences at Dalian University of Technology, China. His research interests include conic optimization, stochastic programming, and mathematical programs with equilibrium constraints.