Gradient and Hessian of Joint Probability Function with Applications on Chance Constrained Programs

L. Jeff Hong

Department of Economics and Finance and Department of Management Sciences
City University of Hong Kong, Kowloon Tong, Hong Kong, China

Guangxin Jiang

Department of Economics and Finance
City University of Hong Kong, Kowloon Tong, Hong Kong, China

January 20, 2017

Abstract

Joint probability function refers to the probability function that requires multiple conditions to satisfy simultaneously. It appears naturally in chance constrained programs. In this paper we derive closed-form expressions of the gradient and Hessian of joint probability functions and develop Monte Carlo estimators of them. We then design a Monte Carlo algorithm, based on these estimators, to solve chance constrained programs. Our numerical study shows that the algorithm works well, especially only with the gradient estimators.

1 Introduction

Consider the following optimization problem:

$$\min_{x\in X} f(x) \quad \text{subject to} \quad h(x) = \Pr\{c_1(x,\xi) \geq 0, \ldots, c_m(x,\xi) \geq 0\} \geq 1-\alpha, \tag{1}$$

where $X \subset \Re^d$ is a convex set, ξ is a random vector and $\alpha \in (0,1)$. Problem (1) is called a chance constrained program (CCP). In particular, it is called a single CCP when $m = 1$ and a joint CCP when $m \geq 2$. CCPs arise naturally in many applications areas such as economics, finance and engineering. For instance, the cash matching problem of Dentcheva et al. (2004) maximizes the value of the portfolio at the end of the planning horizon while covering all scheduled payments with a probability at least 95%; the reservoir system design problem of Prékopa et al. (1978) minimizes the total building and penalty costs while satisfying demands for all sites and all periods with a

probability at least 80%. Chance-constrained optimization problems were introduced by Charnes et al. (1958) and Miller and Wagner (1965). Since then, they have been studied extensively in the stochastic programming literature. For a recent review of the topic, readers are referred to Prékopa (2003).

To solve Problem (1), there are three major difficulties. First, the feasible set defined by the chance constraint may not be convex even when $c_i, i = 1, \ldots, m$, are convex or linear in the decision variables x. Then, finding a global optimal solution to the problem becomes very difficult. Second, the mathematical tractability of an individual probability function $\Pr\{c_i(x, \xi) \geq 0\}$ is already poor since only the distribution of ξ may be known (even this may not be known in many practical situations). In problem (1), however, we need to handle a joint probability function. Third, designing efficient numerical methods to solve the problem are often difficult. In many cases, for instance, the joint probability function may only be evaluated through Monte-Carlo simulation.

In stochastic programming literature, different approaches have been proposed to address the three difficulties. For the first difficulty, Prékopa (2003) has shown that the joint probability function $h(x)$ is quasi-concave (which defines a convex feasible set) if $c_1(x, \xi), \ldots, c_m(x, \xi)$ are quasi-concave functions of (x, ξ) and ξ has a logconcave probability distribution, which includes uniform distribution, multivariate normal distribtion, and many others. However, this condition is very restrictive and it is not satisfied even when $c_i(x, \xi) = \xi' x$. Lagoa et al. (2005) show that an individual probability constraint in the form of $\Pr\{a^T x \leq b\} \geq 1 - \alpha$ defines a convex set provided that the vector $(a^T, b)^T$ has a symmetric logconcave density with $\alpha < 1/2$. When $h(x)$ is not quasi-concave (or at least not verifiable), many convex approximations of $h(x)$ have been proposed, e.g., the quadratic approximation of Ben-Tal and Nemirovski (2000), the conditional value-at-risk (CVaR) approximation of Rockafellar and Uryasev (2000) for an individual probability function, and the Bernstein approximation of Nemirovski and Shapiro (2006). These approximations typically find feasible but suboptimal solutions to Problem (1).

For the second difficulty, a common approach is to approximate the joint probability function by a set of individual probability functions. A popular choice is to use Boole inequality, which guarantees the satisfication of the joint probability constraint if $\Pr\{c_i(x, \xi) \geq 0\} \leq \alpha_i, i = 1, \ldots, m$, and $\alpha_1 + \cdots + \alpha_m = \alpha$ (e.g., Nemirovski and Shapiro 2006). There are also other sharper approximations, see, for instance, Chapter 6 of Prékopa (1995) and Chen et al. (2010).

However, they often require the joint probability function to have certain special structures. These approximations also find feasible but suboptimal solutions to Problem (1).

For the third difficulty, there are generally three approaches. When the chance constraint is in some special cases, the gradient of the joint probability function may be evaluated. Then the problem can be solved as a nonlinear optimization problem by standard gradient-based algorithms. For instance, $\nabla h(x)$ may be computed if $h(x) = \Pr\{Tx \geq \xi\}$ where T is a deterministic matrix (e.g., Prékopa 1995). Uryasev (1989) derived a general gradient formula for the joint probability function which invloves a surface integral. Prékopa (1995) pointed out that no numerical evaluation technique has been reported for the formula except when $h(x) = \Pr\{Tx \geq \xi\}$ and ξ has a multivariate normal distribution. Marti (2005) also reported several techniques for computing the gradient of $h(x)$, including integral transformation and orthogonal function series expansions which may be difficult to implement in practice. When the chance constraint is approximated by functions that are analytically tractable, e.g., the quadratic approximation of Ben-Tal and Nemirovski (2000), the problem can be solved using standard nonlinear optimization solvers.

Another approach that is often used to handle the third difficulty is the Monte Carlo method. Under this approach, a sample of ξ, denoted as $\{\xi_1, \ldots, \xi_n\}$, is first generated, and then optimization problem is approximated by the sample problem (often through, but not limited to, a sample-average approximation). The approximation problem is then solved using different methods. For instance, the CVaR approximation of Rockafellar and Uryasev (2000), the scenario approach of Calafiore and Campi (2005, 2006), the sequential convex approximation of Hong et al. (2011, 2014) are all solved using this approach, and Meng et al. (2010) and Sun et al. (2014) studied the asymptotic convergence. An advantage of this approach is that it does not require the density of ξ. Instead, it only requires a sample of ξ which may be sampled from a complicated simulation model.

Among different solution methods, convex approximations and scenario analysis are the two most popular ones, because both of them do not require too restrictive assumptions on the distribution of ξ. Under the assumption that $c_i(x, \xi), i = 1, \ldots, m$ are convex in x, both methods convert the original non-convex CCPs into convex programs that may be solved by many optimization packages. However, both of these methods only provide guarantees on the feasibility of the solution, but not on the optimality. To resolve the program, Hong et al. (2011) first propose

a sequential convex approximation algorithm, which solves a convex program in each iteration based on the information obtained from the previous iteration. They show that, under some technical conditions, the sequence of the solutions converge to a KKT point of the original Problem (1). Hu et al. (2013) and Hong et al. (2014) further improved the algorithm by removing some of the technical conditions and providing more insights.

In this paper we derive closed-form expressions of both the gradient $\nabla h(x)$ and the Hessian $\nabla^2 h(x)$ of the joint probability function $h(x)$ with very limited assumptions on the continuity and integrability of ξ and $c_i(x, \xi)$. Then, we show that $\nabla h(x)$ and $\nabla^2 h(x)$ can be computed (estimated) when a sample of ξ is given. Once $\nabla h(x)$ and $\nabla^2 h(x)$ are available, we may solve Problem (1) directly using nonlinear optimization algorithms. This may be viewed also as a Monte Carlo approach, and the local optimality of the solution is guaranteed by properties of the nonlinear optimization algorithms.

Our approach addresses the second and third difficulties in solving CCPs. First, we are able to handle joint probability constraint directly without approximating them by individual probability constraints. This removes the conservatism due to the approximation, and allows us to obtain much better solutions especially when m is large. Second, compared to some other Monte Carlo approaches, ours is more efficient especially when the sample size n is large. For instance, the number of constraints in the CVaR approximation of Rockafellar and Uryasev (2000) and the scenario analysis of Nemirovski and Shapiro (2006) increase linearly in the sample size n. Then, n cannot be too large; otherwise, the problems become difficult to solve numerically. In our approach, however, the samples are only used to evaluate $h(x)$, $\nabla h(x)$ and $\nabla^2 h(x)$, which is an $O(n)$ operation if both m and d are fixed. This allows us to use much larger sample sizes and obtain solutions with better precision.

Using Monte Carlo methods in gradient estimation is a classical topic in the simulation literature (see Fu 2006 for a thorough introduction). There are in general two approaches, pathwise method originated by Ho and Cao (1983) and later elaborated by Glasserman (1991) and Fu and Hu (1997), and likelihood ratio method proposed by Reiman and Weiss (1989) and Glynn (1990). Our method falls into the category of pathwise method. Traditionally, pathwise method cannot handle expectations of discontinuous function (for instance, the indicator function in joint probability function). However, in recent years, by combining kernel estimation, pathwise method has

also been used to estimate gradient of probability functions (Hong and Liu 2010), quantile functions (Hong 2009 and Fu et al. 2009) and other types of functions that may be expressed by the expectation of a discontinuous function (e.g., Hong and Liu 2010). The method proposed in this paper is specially designed to estimate gradients and Hessian of joint probability functions, which are new in the literature.

In summary, we make the following contributions in this paper:

1. We prove new closed-form expressions of the gradient and Hessian of a joint probability function where $c_i(x, \xi)$ may be linear or nonlinear in x and ξ.

2. We provide a Monte Carlo method to evaluate the gradient and Hessian of the joint probability function based only on samples of ξ. The method is an $O(n)$ operation, with our parameters fixed, where n is the sample size.

3. We propose a Monte Carlo method to solve CCPs using existing nonlinear optimization algorithms. The method guarantees the local optimality and works well for testing problems.

The rest of the paper is organized as follows: We derive the closed-form expressions of the gradient and Hessian of the joint probability function and introduce Monte Carlo methods to compute them in Sections 2 and 3, and give more details on how to use the gradient and Hessian to solve CCPs. The numerical illustrations are presented in Section 4, followed by the conclusions in Section 5.

2 Analysis of Gradient

Throughout this paper, we use c_i, $\nabla_x c_i$ and $\nabla_x^2 c_i$ denote $c_i(x, \xi)$, $\nabla_x c_i(x, \xi)$ and $\nabla_x^2 c_i(x, \xi)$ for all $i = 1, 2, \ldots, m$ when there is no ambiguity. Then

$$h(x) = \Pr\{c_1 \geq 0, \ldots, c_m \geq 0\} = \mathrm{E}\left[\prod_{i=1}^{m} 1_{\{c_i \geq 0\}}\right],$$

where $1_{\{\cdot\}}$ is an indicator function.

2.1 Background

In this paper, we make following assumptions on $c_i(x, \xi)$.

Assumption 1. *For all $i = 1, 2, \ldots, m$, $\mathrm{E}\left(|c_i(x,\xi)|^m\right) < \infty$, and $c_i(x,\xi)$ has a continuous density $f_{c_i}(t)$ in the neighborhood of $t = 0$, $c_i(x,\xi)$ is differentiable with respect to x, and there exists a function $K_i(x,\xi)$ with $\mathrm{E}\left([K_i(x,\xi)]^m\right) < \infty$ such that*

$$|c_i(x + \Delta x, \xi) - c_i(x,\xi)| \leq K_i(x,\xi) \cdot \|\Delta x\| \tag{2}$$

when $\|\Delta x\|$ is small enough, and $\|\cdot\|$ is the Euclidean norm.

Assumption 2. *For all $i = 1, 2, \ldots, m$, let*

$$q_i(t) = \mathrm{E}\left[\nabla_x c_i \cdot \prod_{j=1, j\neq i}^{m} 1_{\{c_j \geq 0\}} \middle| c_i = t\right].$$

Then $q_i(t)$ is continuous at $t = 0$.

Assumption 1 requires $c_i(x,\xi)$ to be a continuous random variable in a neighborhood of 0 and to have finite mth moment for all $i = 1, 2, \ldots, m$. It also requires $c_i(x,\xi)$ to satisfy a Lipschitz continuity for every x in its local neighborhood. This assumption is typically satisfied. For instance, when $c_i(x,\xi) = \xi^T x$, then we may set $K_i(x,\xi) = \|\xi\|$; when $c_i(x,\xi) = x^T \xi x$, then we may set $K_i(x,\xi) = (2\|x\| + 1) \cdot \|\xi\|$. Notice that the assumption also implies that $\mathrm{E}\left(\|\nabla_x c_i(x,\xi)\|^m\right) < \infty$ for all $i = 1, 2, \ldots, m$.

Assumption 2 is a more technical assumption. Since $c_i(x,\xi)$ is a continuous random variable in the neighborhood of 0, as assumed in Assumption 1, a small perturbation of t at $t = 0$ typically does not result in a sudden change to the conditional expectation. Therefore, this assumption is typically satisfied though it is difficult to verify in practice.

Notice that $\nabla h(x) = \nabla_x \mathrm{E}\left[\prod_{i=1}^{m} 1_{\{c_i(x,\xi)\geq 0\}}\right]$. Then a critical question is whether we can interchange the differentiation and the expectation. To answer this question and also to faciliate the analysis in the rest of this paper, we first prove the following lemma on the validity of interchanging differentiation and expectation.

Lemma 1. *Let $\theta(x,\xi)$ be a function of $x \in X \subset \Re^d$ and a random vector ξ. Suppose that $\theta(x,\xi)$ is differentiable with respect to x for every $x \in X$ for almost all ξ (almost surely with respect to ξ), and that there exists a function $K_\theta(x,\xi)$ with $\mathrm{E}\left[K_\theta(x,\xi)\right] < \infty$ such that*

$$|\theta(x + \Delta x, \xi) - \theta(x,\xi)| \leq K_\theta(x,\xi) \cdot \|\Delta x\|, \tag{3}$$

when $\|\Delta x\|$ is small enough. Then

$$\nabla \mathrm{E}\left[\theta(x,\xi)\right] = \mathrm{E}\left[\nabla_x \theta(x,\xi)\right].$$

Proof. Let $g(x) = \mathrm{E}\left[\theta(x,\xi)\right]$. Then $\nabla g(x) = \left(\partial g(x)/\partial x_1, \ldots, \partial g(x)/\partial x_d\right)^T$. Then it suffices to prove that $\partial g(x)/\partial x_i = \mathrm{E}\left[\partial\theta(x,\xi)/\partial x_i\right]$ for all $i = 1, 2, \ldots, d$. Let e_i denote the ith column of an identity matrix. Since

$$\left|\frac{\theta(x + \Delta x_i e_i, \xi) - \theta(x,\xi)}{\Delta x_i}\right| \leq K_\theta(x,\xi)$$

when $|\Delta x_i|$ is small and $\mathrm{E}\left[K_\theta(x,\xi)\right] < \infty$, then by the dominated convergence theorem (Durrett 1995),

$$\frac{\partial g(x)}{\partial x_i} = \lim_{\Delta x_i \to 0} \mathrm{E}\left[\frac{\theta(x + \Delta x_i e_i, \xi) - \theta(x,\xi)}{\Delta x_i}\right] = \mathrm{E}\left[\lim_{\Delta x_i \to 0} \frac{\theta(x + \Delta x_i e_i, \xi) - \theta(x,\xi)}{\Delta x_i}\right].$$

Since $\theta(x,\xi)$ is differentiable with respect to x almost surely for every $x \in X$, then $\partial g(x)/\partial x_i = \mathrm{E}\left[\partial\theta(x,\xi)/\partial x_i\right]$. This concludes the proof of the lemma. □

Lemma 1 is a result directly from the dominated convergence theorem. Note that indicator functions are not Lipschitz continuous. Then Equation (3) does not hold for $\theta(x,\xi) = \prod_{i=1}^m 1_{\{c_i \geq 0\}}$. Therefore, Lemma 1 is not applicable to $h(x) = \mathrm{E}\left[\prod_{i=1}^m 1_{\{c_i \geq 0\}}\right]$. To overcome this difficulty, we define

$$\Psi(x, y_1, \ldots, y_m) = \mathrm{E}\left[\prod_{i=1}^m (y_i - c_i) \cdot 1_{\{c_i \geq y_i\}}\right].$$

By Assumption 1, $\mathrm{E}(|c_i|^m) < \infty$. Then by Hölder's inequality (Durrett 1995)

$$\mathrm{E}\left[\left|\prod_{i=1}^m (y_i - c_i) \cdot 1_{\{c_i \geq y_i\}}\right|\right] \leq \prod_{i=1}^m \left[\mathrm{E}\left(|y_i - c_i|^m\right)\right]^{\frac{1}{m}} < \infty.$$

Therefore, $\Psi(x, y_1, \ldots, y_m)$ is well defined. Notice that $f(z) = z \cdot 1_{\{z \geq 0\}}$ is a Lipschitz continuous function with the Lipschitz constant being 1, it is differentiable everywhere except at $z = 0$, and $f'(z) = 1_{\{z \geq 0\}}$ when $z \neq 0$. Then we have the following lemma on the relationship between $\Psi(x, y_1, \ldots, y_m)$ and $h(x)$.

Lemma 2. *Suppose that Assumption 1 is satisfied. Then*

$$h(x) = \left.\frac{\partial^m \Psi(x, y_1, \ldots, y_m)}{\partial y_1 \cdots \partial y_m}\right|_{y_1 = \cdots = y_m = 0}.$$

Proof. Let $\Phi_1(y_1, x, \xi) = \prod_{i=1}^{m}(y_i - c_i) \cdot 1_{\{c_i \geq y_i\}}$. Then $\Psi(x, y_1, \ldots, y_m) = \mathrm{E}[\Phi_1(y_1, x, \xi)]$. We first prove that we can apply Lemma 1 on $\mathrm{E}[\Phi_1(y_1, x, \xi)]$.

Since $f(z) = z \cdot 1_{\{z \geq 0\}}$ is a Lipschitz continuous function with the Lipschitz constant being 1, then

$$|\Phi_1(y_1 + \Delta y_1, x, \xi) - \Phi_1(y_1, x, \xi)| \leq \left| \prod_{i=2}^{m}(y_i - c_i) \cdot 1_{\{c_i \geq y_i\}} \right| \cdot |\Delta y_1|.$$

By Assumption 1 and Hölder's inequality, $\mathrm{E}\left[\left|\prod_{i=2}^{m}(y_i - c_i) \cdot 1_{\{c_i \geq y_i\}}\right|\right] < \infty$. Since $f(z)$ is differentiable everywhere except at $z = 0$, then $\Phi_1(y_1, x, \xi)$ is differentiable with respect to y_1 except at $y_1 = c_1(x, \xi)$. By Assumption 1, $c_1(x, \xi)$ has a density in a neighborhood of 0, then $\Pr\{c_1(x, \xi) = y_1\} = 0$ when y_1 is in the neighborhood of 0. Therefore, by Lemma 1,

$$\left.\frac{\partial\, \Psi(x, y_1, \ldots, y_m)}{\partial y_1}\right|_{y_1=0} = \mathrm{E}\left[\frac{\partial \Phi_1(y_1, x, \xi)}{\partial y_1}\right]\Bigg|_{y_1=0} = \mathrm{E}\left[1_{\{c_1 \geq 0\}} \cdot \prod_{i=2}^{m}(y_i - c_i) \cdot 1_{\{c_i \geq y_i\}}\right].$$

We may use the same techniques to continue differentiating $\Psi(x, y_1, \ldots, y_m)$ with respect to $y_2, \ldots, y_m$. We have

$$\left.\frac{\partial^m\, \Psi(x, y_1, \ldots, y_m)}{\partial y_1 \cdots \partial y_m}\right|_{y_1=\cdots=y_m=0} = \mathrm{E}\left[\prod_{i=1}^{m} 1_{\{c_i \geq 0\}}\right] = h(x).$$

This concludes the proof of the lemma. □

By Lemma 2, to find $\nabla h(x)$ and $\nabla^2 h(x)$, we may interchange the order of differentiations [1] to obtain

$$\nabla h(x) = \nabla_x \left.\frac{\partial^m\, \Psi(x, y_1, \ldots, y_m)}{\partial y_1 \cdots \partial y_m}\right|_{y_1=\cdots=y_m=0} = \left.\frac{\partial^m\, \nabla_x \Psi(x, y_1, \ldots, y_m)}{\partial y_1 \cdots \partial y_m}\right|_{y_1=\cdots=y_m=0}, \quad (4)$$

$$\nabla^2 h(x) = \nabla_x^2 \left.\frac{\partial^m\, \Psi(x, y_1, \ldots, y_m)}{\partial y_1 \cdots \partial y_m}\right|_{y_1=\cdots=y_m=0} = \left.\frac{\partial^m\, \nabla_x^2 \Psi(x, y_1, \ldots, y_m)}{\partial y_1 \cdots \partial y_m}\right|_{y_1=\cdots=y_m=0}. \quad (5)$$

In the rest of this section and next section, we use these two equations to analyze $\nabla h(x)$ and $\nabla^2 h(x)$.

2.2 A Closed Form of Gradient

In this subsection we derive a closed form of $\nabla h(x)$ based on Equation (4). We first prove the following lemma on $\nabla_x \Psi(x, y_1, \ldots, y_m)$.

[1] There are technical conditions for interchanging the order of differentiations (see, for instance, Marsden and Hoffman 1993). The conditions are weak and typically satisfied by practical problems. To avoid too much technicality, we implicitly assume that the order can be interchanged throughout the paper.

Lemma 3. *Suppose that Assumption 1 is satisfied. Then*

$$\nabla_x \Psi(x, y_1, \ldots, y_m) = -\sum_{i=1}^{m} \mathrm{E}\left[\nabla_x c_i \cdot 1_{\{c_i \geq y_i\}} \cdot \prod_{j=1, j\neq i}^{m} (y_j - c_j) \cdot 1_{\{c_j \geq y_j\}}\right].$$

Proof. Let $\beta_i(x, \xi) = (y_i - c_i) \cdot 1_{\{c_i \geq y_i\}}$ for all $i = 1, 2, \ldots, m$. Notice that $\mathrm{E}\left[|\beta_i(x, \xi)|^m\right] < \infty$ by Assumption 1. Since $\beta_i(x, \xi) = -f(c_i - y_i)$ where $f(z) = z \cdot 1_{\{z \geq 0\}}$, then

$$\begin{aligned} & |\beta_i(x + \Delta x, \xi) - \beta_i(x, \xi)| \\ = \;& |f(c_i(x + \Delta x, \xi) - y_i) - f(c_i(x, \xi) - y_i)| \\ \leq \;& |c_i(x + \Delta x, \xi) - c_i(x, \xi)| \\ \leq \;& K_i(x, \xi) \cdot \|\Delta x\|, \end{aligned} \tag{6}$$

where the first inequality follows from the Lipschitz continuity of $f(z)$ and the second inequality follows from Assumption 1. Furthermore, by Assumption 1, we have

$$|\beta_i(x + \Delta x, \xi)| \leq |\beta_i(x, \xi)| + K_i(x, \xi) \cdot \|\Delta x\| \leq |\beta_i(x, \xi)| + K_i(x, \xi) \tag{7}$$

when $\|\Delta x\|$ is small enough.

Let $\beta(x, \xi) = \prod_{i=1}^{m} \beta_i(x, \xi)$. Then $\Psi(x, y_1, \ldots, y_m) = \mathrm{E}\left[\beta(x, \xi)\right]$. Notice that

$$\begin{aligned} & \beta(x + \Delta x, \xi) - \beta(x, \xi) \\ = \;& \prod_{i=1}^{m} \beta_i(x + \Delta x, \xi) - \prod_{i=1}^{m} \beta_i(x, \xi) \\ = \;& \sum_{i=1}^{m} \left\{\prod_{j=1}^{i-1} \beta_j(x, \xi) \cdot [\beta_i(x + \Delta x, \xi) - \beta_i(x, \xi)] \cdot \prod_{j=i+1}^{m} \beta_j(x + \Delta x, \xi)\right\}. \end{aligned}$$

Then by Equations (6) and (7),

$$|\beta(x + \Delta x, \xi) - \beta(x, \xi)| \leq \sum_{i=1}^{m} \left\{\prod_{j=1}^{i-1} |\beta_j(x, \xi)| \cdot K_i(x, \xi) \cdot \prod_{j=i+1}^{m} \left[|\beta_j(x, \xi)| + K_j(x, \xi)\right]\right\} \cdot \|\Delta x\|$$

when $\|\Delta x\|$ is small enough. Then

$$\begin{aligned}
&\mathrm{E}\left[\sum_{i=1}^{m}\left\{\prod_{j=1}^{i-1}|\beta_j(x,\xi)|\cdot K_i(x,\xi)\cdot\prod_{j=i+1}^{m}\left[|\beta_j(x,\xi)|+K_j(x,\xi)\right]\right\}\right]\\
&\leq \sum_{i=1}^{m}\left[\prod_{j=1}^{i-1}\mathrm{E}\left(|\beta_j(x,\xi)|^m\right)\cdot\mathrm{E}\left([K_i(x,\xi)]^m\right)\cdot\prod_{j=i+1}^{m}\mathrm{E}\left([|\beta_j(x,\xi)|+K_j(x,\xi)]^m\right)\right]^{\frac{1}{m}}\\
&\leq \sum_{i=1}^{m}\left[\prod_{j=1}^{i-1}\mathrm{E}\left(|\beta_j(x,\xi)|^m\right)\cdot\mathrm{E}\left([K_i(x,\xi)]^m\right)\cdot\prod_{j=i+1}^{m}2^{m-1}\left[\mathrm{E}\left(|\beta_j(x,\xi)|^m\right)+\mathrm{E}\left([K_j(x,\xi)]^m\right)\right]\right]^{\frac{1}{m}}\\
&< \infty,
\end{aligned}$$

where the first inequality follows from Hölder's inequality, the second inequality follows from Minkowvski's inequality (Durrett 1995), and the third inequality holds because $\mathrm{E}\left([K_j(x,\xi)]^m\right)<\infty$ and $\mathrm{E}\left(|\beta_j(x,\xi)|^m\right)<\infty$.

Furthermore, for all $x\in X$,

$$\nabla_x\beta(x,\xi)=-\sum_{i=1}^{m}\nabla_x c_i\cdot 1_{\{c_i\geq y_i\}}\cdot\prod_{j=1,j\neq i}^{m}(y_j-c_j)\cdot 1_{\{c_j\geq y_j\}}$$

for all ξ except when ξ satisfies $c_i(x,\xi)=y_i$ for some $i=1,2,\ldots,m$. Since $c_i(x,\xi)$ has a density in a neighborhood of 0 for all $i=1,2,\ldots,m$, then $\beta(x,\xi)$ is differentiable almost surely when y_i is close enough to 0 for all $i=1,2,\ldots,m$.

Then by Lemma 1,

$$\nabla_x\Psi(x,y_1,\ldots,y_m)=\mathrm{E}\left[\nabla_x\beta(x,\xi)\right]=-\sum_{i=1}^{m}\mathrm{E}\left[\nabla_x c_i\cdot 1_{\{c_i\geq y_i\}}\cdot\prod_{j=1,j\neq i}^{m}(y_j-c_j)\cdot 1_{\{c_j\geq y_j\}}\right].$$

This concludes the proof of the lemma. □

Now we can use Equation (4) to derive a closed form of $\nabla h(x)$. We summarize the result in the following theorem.

Theorem 1. *Suppose that Assumptions 1 and 2 are satisfied, and f_{c_i} is defined in Assumption 1. Then*

$$\nabla h(x)=\sum_{i=1}^{m}f_{c_i}(0)\cdot\mathrm{E}\left[\nabla_x c_i\cdot\prod_{j=1,j\neq i}^{m}1_{\{c_j\geq 0\}}\,\middle|\,c_i=0\right].$$

Proof. Let

$$\phi_i(x,y_1,\ldots,y_m)=-\mathrm{E}\left[\nabla_x c_i\cdot 1_{\{c_i\geq y_i\}}\cdot\prod_{j=1,j\neq i}^{m}(y_j-c_j)\cdot 1_{\{c_j\geq y_j\}}\right]$$

for all $i = 1, 2, \ldots, m$. By Lemma 3, $\nabla_x \Psi(x, y_1, \ldots, y_m) = \sum_{i=1}^m \phi_i(x, y_1, \ldots, y_m)$. Then by Equation (4), it suffices to prove that

$$\left.\frac{\partial \phi_i(x, y_1, \ldots, y_m)}{\partial y_1 \cdots \partial y_m}\right|_{y_1=\cdots=y_m=0} = f_{c_i}(0) \cdot \mathrm{E}\left[\nabla_x c_i \cdot \prod_{j=1, j\neq i}^m 1_{\{c_j \geq 0\}} \middle| c_i = 0\right]. \tag{8}$$

By the similar arguments used in the proof of Lemma 2, we can show that

$$\left.\frac{\partial \phi_i(x, y_1, \ldots, y_m)}{\partial y_1 \cdots \partial y_{i-1} \partial y_{i+1} \cdots \partial y_m}\right|_{y_1=\cdots=y_{i-1}=y_{i+1}=y_m=0} = -\mathrm{E}\left[\nabla_x c_i \cdot 1_{\{c_i \geq y_i\}} \cdot \prod_{j=1, j\neq i}^m 1_{\{c_j \geq 0\}}\right].$$

Notice that

$$\begin{aligned}
&\mathrm{E}\left[\nabla_x c_i \cdot 1_{\{c_i \geq y_i\}} \cdot \prod_{j=1, j\neq i}^m 1_{\{c_j \geq 0\}}\right] \\
&= \int_{y_i}^\infty \mathrm{E}\left[\nabla_x c_i \cdot \prod_{j=1, j\neq i}^m 1_{\{c_j \geq 0\}} \middle| c_i = t\right] f_{c_i}(t) dt \\
&= \int_{y_i}^\infty q_i(t) f_{c_i}(t) dt,
\end{aligned}$$

where $q_i(t)$ is defined in Assumption 2. Since both $f_{c_i}(t)$ and $q_i(t)$ are continuous in a neighborhood of $t = 0$ by Assumptions 1 and 2, then by fundamental theorem of calculus (Marsden and Hoffman 1993),

$$\left.\frac{\partial}{\partial y_i} \int_{y_i}^\infty q_i(t) f_{c_i}(t) dt\right|_{y_i=0} = -q_i(0) f_{c_i}(0).$$

Therefore,

$$\left.\frac{\partial \phi_i(x, y_1, \ldots, y_m)}{\partial y_1 \cdots \partial y_m}\right|_{y_1=\cdots=y_m=0} = f_{c_i}(0) q_i(0) = f_{c_i}(0) \cdot \mathrm{E}\left[\nabla_x c_i \cdot \prod_{j=1, j\neq i}^m 1_{\{c_j \geq 0\}} \middle| c_i = 0\right].$$

Then Equation (8) holds for all $i = 1, 2, \ldots, m$. This concludes the proof of the theorem. □

2.3 Monte Carlo Method for Computing Gradient

When solving chance-constrained optimization problems, we often know how to simulate ξ based on the model of ξ or we may have some historical observations of ξ. Therefore, in this subsection, we suppose that we have n independent and identically distributed (i.i.d.) observations of ξ, denoted as $\{\xi_1, \xi_2, \ldots, \xi_n\}$ and we discuss how to use these observations to compute $\nabla h(x)$ for any $x \in X$.

There are two major difficulties in computing $\nabla h(x)$ using Theorem 1. First, the densities of $c_i, i = 1, 2, \ldots, m$ are typically unknown. Second, the conditional expectations condition on $\{c_i(x, \xi) = 0\}$, which is a probability zero event for every i, is typically satisfied by none of the n observations of ξ.

To overcome these difficulties, notice that

$$f_B(b) \cdot \mathrm{E}[A|B = b] = \partial_b \mathrm{E}\left[A \cdot 1_{\{B \geq b\}}\right] = \lim_{\delta \to 0} \frac{1}{\delta} \mathrm{E}\left[A \cdot 1_{\left\{-\frac{\delta}{2} \leq B - b \leq \frac{\delta}{2}\right\}}\right], \tag{9}$$

where the first equality refers to the proof of Theorem 1 of Hong (2009), and the second equality holds by the derivative definition. Then we may set δ small and compute $1/\delta \cdot \mathrm{E}\left[A \cdot 1_{\{-\delta/2 \leq B - b \leq \delta/2\}}\right]$ by using $\{\xi_1, \ldots, \xi_n\}$. Notice that $f_B(b)$ is not required and $\{-\delta/2 \leq B - b \leq \delta/2\}$ is no longer a probability zero event. Therefore, we may estimate the expectation by a standard sample mean.

However, this method does not utilize the observations that do not satisfy $\{-\delta/2 \leq B - b \leq \delta/2\}$, even though they may include useful information. In this paper, we suggest to use the kernel method to compute $f_B(b) \cdot \mathrm{E}[A|B = b]$. A one-dimensional kernel function K is a symmetric density such that $uK(u) \to 0$ as $|u| \to \infty$ and $\int_{-\infty}^{\infty} u^2 K(u)\, du < \infty$ (Bosq 1998). For instance, the standard normal density $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$ is a widely used one-dimensional kernel function. Similar to Equation (9), we have (see Liu and Hong (2009) for detail)

$$f_B(b) \cdot \mathrm{E}[A|B = b] = \lim_{\delta \to 0} \frac{1}{\delta} \mathrm{E}\left[A \cdot K\left(\frac{B - b}{\delta}\right)\right]$$

by Bochner's lemma (Parzen 1962). Let $\{(A_1, B_1), (A_2, B_2), \ldots, (A_n, B_n)\}$ be an i.i.d. sample of (A, B). Then $f_B(b) \cdot \mathrm{E}[A|B = b]$ can be estimated by the following kernel estimator

$$\widehat{H}(b) = \frac{1}{n\delta_n} \sum_{\ell=1}^{n} A_\ell \cdot K\left(\frac{B_\ell - b}{\delta_n}\right). \tag{10}$$

It can be shown that $\widehat{H}(b)$ is a consistent estimator of $f_B(b) \cdot \mathrm{E}[A|B = b]$ as $n \to \infty$ if $\delta_n \to 0$ and $n\delta_n \to \infty$. Furthermore, the rate of convergence is $(n\delta_n)^{-1/2}$ under some technical conditions (Bosq 1998). The asymptotic properties of $\widehat{H}(b)$ hold even when the sequence $\{(A_1, B_1), \ldots, (A_n, B_n)\}$ is not an i.i.d. sequence. For instance, the properties also hold when the sequence is ϕ-mixing, which means it is stationarily dependent but does not have long-range dependence.

Therefore, we may estimate $\nabla h(x)$ by the following estimator

$$\widehat{\nabla} h(x) = \frac{1}{n\delta_n} \sum_{\ell=1}^{n} \sum_{i=1}^{m} \nabla_x c_i(x, \xi_\ell) \cdot \prod_{j=1, j \neq i}^{m} 1_{\{c_j(x, \xi_\ell) \geq 0\}} \cdot K\left[\frac{c_i(x, \xi_\ell)}{\delta_n}\right]. \tag{11}$$

Notice that computing $\widehat{\nabla} h(x)$ is an $O(n)$ operation with a fixed m and the dimension d. It can be done efficiently when the sample size n is large. If we also consider m and d, then computing $\widehat{\nabla} h(x)$ is an $O(nmd)$ operation, and it also scales well in both m and d.

3 Analysis of Hessian

In this section we apply Equation (5) to analyze $\nabla^2 h(x)$ and discuss how to compute it through a Monte Carlo method.

3.1 Background

We make the following additional assumptions on $c_i(x), i = 1, 2, \ldots, m$.

Assumption 3. *For all $i = 1, 2, \ldots, m$, $f_{c_i}(t)$ is differentiable at $t = 0$, $\mathrm{E}\left(|c_i(x,\xi)|^{m+1}\right) < \infty$ and $\mathrm{E}\left([K_i(x,\xi)]^{m+1}\right) < \infty$ where $K_i(x,\xi)$ is defined in Assumption 1.*

Assumption 4. *For all $i = 1, 2, \ldots, m$, $c_i(x,\xi)$ is twice differentiable with respect to x, and there exists a function $G_i(x,\xi)$ with $\mathrm{E}\left([G_i(x,\xi)]^{m+1}\right) < \infty$ such that*

$$\|\nabla_x c_i(x + \Delta x, \xi) - \nabla_x c_i(x,\xi)\| \le G_i(x,\xi) \cdot \|\Delta x\|$$

when $\|\Delta x\|$ is small enough. Furthermore, for all $i, j = 1, 2, \ldots, m$ and $i \neq j$, (c_i, c_j) has a continuous joint density $f_{c_i,c_j}(t,s)$ in a neighborhood of $(0,0)$.

Assumption 5. *For all $i, j = 1, 2, \ldots, m$ and $i \neq j$, let*

$$\begin{aligned}
u_i(t) &= \mathrm{E}\left[\nabla_x c_i \nabla_x c_i^T \cdot \prod_{j=1, j\neq i}^{m} 1_{\{c_j \ge 0\}} \,\middle|\, c_i = t\right], \\
v_i(t) &= \mathrm{E}\left[\nabla_x^2 c_i \cdot \prod_{j=1, j\neq i}^{m} 1_{\{c_j \ge 0\}} \,\middle|\, c_i = t\right], \\
w_{i,j}(t,s) &= \mathrm{E}\left[\nabla_x c_i \nabla_x c_j^T \cdot \prod_{k=1, k\neq i,j}^{m} 1_{\{c_k \ge 0\}} \,\middle|\, c_i = t, c_j = s\right].
\end{aligned}$$

Then $u_i(t)$ is differentiable at $t = 0$, $v_i(t)$ is continuous at $t = 0$, and $w_{i,j}(t,s)$ is continuous at $(t,s) = (0,0)$.

Assumption 3 adds on Assumption 1. It requires $f_{c_i}(t)$ to be differentiable at $t = 0$, whereas Assumption 1 only requires it to be continuous. It requires $c_i(x, \xi)$ and $K_i(x, \xi)$ to have finite $(m+1)$st moments, whereas Assumption 1 only requires them to have finite mth moments. Notice that Assumption 3 also implies that $\mathrm{E}\left(\|\nabla_x c_i(x, \xi)\|^{m+1}\right) < \infty$.

Assumption 4 extends the local Lipschitz continuity to $\nabla_x c_i(x, \xi)$. It is typically satisfied. For instance, when $c_i(x, \xi) = \xi^T x$, we may set $G_i(x, \xi) = 0$; when $c_i(x, \xi) = x^T \xi x$, we may set $G_i(x, \xi) = \|\xi\|$. Notice that the assumption also implies that $\mathrm{E}\left(\|\nabla_x^2 c_i(x, \xi)\|^{m+1}\right) < \infty$ for all $i = 1, 2, \ldots, m$.

Assumption 5 is an extension of Assumption 2. Since (c_i, c_j) is a continuous random vector in a neighborhood of $(0, 0)$, as assumed in Assumption 4, a small perturbation of t or s or both t and s at $t = 0$ and $s = 0$ typically does not result in a sudden change to the conditional expectations. Therefore, this assumption is typically satisfied though it is difficult to verify in practice.

3.2 A Closed Form of Hessian

In this subsection we derive a closed form of $\nabla^2 h(x)$. By Lemma 3,

$$\nabla_x^2 \Psi(x, y_1, \ldots, y_m) = \sum_{i=1}^{m} \nabla_x \phi_i(x, y_1, \ldots, y_m), \tag{12}$$

where

$$\phi_i(x, y_1, \ldots, y_m) = -\mathrm{E}\left[\nabla_x c_i \cdot 1_{\{c_i \geq y_i\}} \cdot \prod_{j=1, j\neq i}^{m} (y_j - c_j) \cdot 1_{\{c_j \geq y_j\}}\right].$$

Since $\nabla_x c_i \cdot 1_{\{c_i \geq y_i\}}$ is discontinuous in x, then the expression inside of the expectation is also discontinuous in x. Therefore, we cannot apply Lemma 1 directly on $\phi(x, y_1, \ldots, y_m)$ to obtain $\nabla_x \phi_i(x, y_1, \ldots, y_m)$.

To solve this problem, we use the same techniques used in Section 2. We define

$$\Gamma_i(x, y_1, \ldots, y_m) = -\mathrm{E}\left[\nabla_x c_i \cdot \prod_{j=1}^{m} (y_j - c_j) \cdot 1_{\{c_j \geq y_j\}}\right]$$

for all $i = 1, 2, \ldots, m$. By Assumption 3 and Hölder's inequality,

$$\mathrm{E}\left[\left\|\nabla_x c_i \cdot \prod_{j=1}^{m} (y_j - c_j) \cdot 1_{\{c_j \geq y_j\}}\right\|\right] \leq \left[\mathrm{E}\left(\|\nabla_x c_i\|^{m+1}\right) \cdot \prod_{i=1}^{m} \mathrm{E}\left(|y_i - c_i|^{m+1}\right)\right]^{\frac{1}{m+1}} < \infty.$$

Then $\Gamma_i(x, y_1, \ldots, y_m) < \infty$ is well define. Similar to Lemma 2, we can prove that

$$\phi_i(x, y_1, \ldots, y_m) = \frac{\partial}{\partial y_i}\Gamma_i(x, y_1, \ldots, y_m).$$

Then by interchanging the order of differentiations, we have

$$\nabla_x \phi_i(x, y_1, \ldots, y_m) = \nabla_x \left[\frac{\partial}{\partial y_i}\Gamma_i(x, y_1, \ldots, y_m)\right] = \frac{\partial}{\partial y_i}\left[\nabla_x \Gamma_i(x, y_1, \ldots, y_m)\right]. \tag{13}$$

Similar to Lemma 3, we can prove that

$$\begin{aligned}
&\nabla_x \Gamma_i(x, y_1, \ldots, y_m) \\
&= -\mathrm{E}\left[\nabla_x \left\{\nabla_x c_i \cdot \prod_{j=1}^{m}(y_j - c_j)\cdot 1_{\{c_j \geq y_j\}}\right\}\right] \\
&= -\mathrm{E}\left[\nabla_x^2 c_i \cdot \prod_{j=1}^{m}(y_j - c_j)\cdot 1_{\{c_j \geq y_j\}}\right] \\
&\quad + \sum_{j=1}^{m}\mathrm{E}\left[\nabla_x c_i \nabla_x c_j^T \cdot 1_{\{c_j \geq y_j\}} \cdot \prod_{k=1, k\neq j}^{m}(y_k - c_k)\cdot 1_{\{c_k \geq y_k\}}\right].
\end{aligned} \tag{14}$$

By Equations (5), (12), (13) and (14),

$$\begin{aligned}
&\nabla^2 h(x) \\
&= \sum_{i=1}^{m} \frac{\partial^{m+1}}{\partial y_1 \cdots \partial y_m \partial y_i}\left[\nabla_x \Gamma_i(x, y_1, \ldots, y_m)\right]\Big|_{y_1 = \cdots = y_m = 0} \\
&= -\sum_{i=1}^{m} \frac{\partial^{m+1}}{\partial y_1 \cdots \partial y_m \partial y_i}\mathrm{E}\left[\nabla_x^2 c_i \cdot \prod_{j=1}^{m}(y_j - c_j)\cdot 1_{\{c_j \geq y_j\}}\right]\Big|_{y_1 = \cdots = y_m = 0} \\
&\quad + \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\partial^{m+1}}{\partial y_1 \cdots \partial y_m \partial y_i}\mathrm{E}\left[\nabla_x c_i \nabla_x c_j^T \cdot 1_{\{c_j \geq y_j\}} \cdot \prod_{k=1, k\neq j}^{m}(y_k - c_k)\cdot 1_{\{c_k \geq y_k\}}\right]\Big|_{y_1 = \cdots = y_m = 0}
\end{aligned} \tag{15}$$

Now we analyze the three terms on the right-hand side of Equation (15). By the techniques

used in the proof of Lemma 2, for all $i = 1, 2, \ldots, m$,

$$\begin{aligned}
&\frac{\partial^{m+1}}{\partial y_1 \cdots \partial y_m \partial y_i} \mathrm{E}\left[\nabla_x^2 c_i \cdot \prod_{j=1}^{m} (y_j - c_j) \cdot 1_{\{c_j \geq y_j\}}\right]\Bigg|_{y_1=\cdots=y_m=0} \\
&= \frac{\partial}{\partial y_i} \mathrm{E}\left[\nabla_x^2 c_i \cdot 1_{\{c_i \geq y_i\}} \cdot \prod_{j=1, j\neq i}^{m} 1_{\{c_j \geq 0\}}\right]\Bigg|_{y_i=0} \\
&= \frac{\partial}{\partial y_i}\left\{\int_{y_i}^{\infty} \mathrm{E}\left[\nabla_x^2 c_i \cdot \prod_{j=1, j\neq i}^{m} 1_{\{c_j \geq 0\}} \,\middle|\, c_i = t\right] \cdot f_{c_i}(t)\, dt\right\}\Bigg|_{y_i=0} \\
&= -f_{c_i}(0) \cdot \mathrm{E}\left[\nabla_x^2 c_i \cdot \prod_{j=1, j\neq i}^{m} 1_{\{c_j \geq 0\}} \,\middle|\, c_i = 0\right], \qquad (16)
\end{aligned}$$

where the last equation follows from Assumptions 1 and 5 and fundamental theorem of calculus.

Similar, for all $i, j = 1, 2, \ldots, m$ and $i \neq j$,

$$\begin{aligned}
&\frac{\partial^{m+1}}{\partial y_1 \cdots \partial y_m \partial y_i} \mathrm{E}\left[\nabla_x c_i \nabla_x c_j^T \cdot 1_{\{c_j \geq y_j\}} \cdot \prod_{k=1, k\neq j}^{m} (y_k - c_k) \cdot 1_{\{c_k \geq y_k\}}\right]\Bigg|_{y_1=\cdots=y_m=0} \\
= &\frac{\partial^2}{\partial y_i \partial y_j}\left\{\int_{y_i}^{\infty}\int_{y_j}^{\infty} \mathrm{E}\left[\nabla_x c_i \nabla_x c_j^T \cdot \prod_{k=1, k\neq i,j}^{m} 1_{\{c_k \geq 0\}} \,\middle|\, c_i = t, c_j = s\right] \cdot f_{c_i,c_j}(t,s)\, ds dt\right\}\Bigg|_{y_i=y_j=0} \\
= &f_{c_i,c_j}(0,0) \cdot \mathrm{E}\left[\nabla_x c_i \nabla_x c_j^T \cdot \prod_{k=1, k\neq i,j}^{m} 1_{\{c_k \geq 0\}} \,\middle|\, c_i = 0, c_j = 0\right], \qquad (17)
\end{aligned}$$

where the last equation follows from Assumptions 4 and 5 and fundamental theorem of calculus.

Similar, for all $i = 1, 2, \ldots, m$,

$$\begin{aligned}
&\frac{\partial^{m+1}}{\partial y_1 \cdots \partial y_m \partial y_i} \mathrm{E}\left[\nabla_x c_i \nabla_x c_i^T \cdot 1_{\{c_i \geq y_i\}} \cdot \prod_{j=1, j\neq i}^{m} (y_j - c_j) \cdot 1_{\{c_j \geq y_j\}}\right]\Bigg|_{y_1=\cdots=y_m=0} \\
&= \frac{\partial^2}{\partial y_i^2} \mathrm{E}\left[\nabla_x c_i \nabla_x c_i^T \cdot 1_{\{c_i \geq y_i\}} \cdot \prod_{j=1, j\neq i}^{m} 1_{\{c_j \geq 0\}}\right]\Bigg|_{y_i=0} \\
&= \frac{\partial^2}{\partial y_i^2}\left\{\int_{y_i}^{\infty} \mathrm{E}\left[\nabla_x c_i \nabla_x c_i^T \cdot \prod_{j=1, j\neq i}^{m} 1_{\{c_j \geq 0\}} \,\middle|\, c_i = t\right] \cdot f_{c_i}(t)\, dt\right\}\Bigg|_{y_i=0} \\
&= -\frac{\partial}{\partial y_i}\left\{f_{c_i}(y_i) \cdot \mathrm{E}\left[\nabla_x c_i \nabla_x c_i^T \cdot \prod_{j=1, j\neq i}^{m} 1_{\{c_j \geq 0\}} \,\middle|\, c_i = y_i\right]\right\}\Bigg|_{y_i=0}, \qquad (18)
\end{aligned}$$

where the second last equation follows from Assumptions 1 and 5 and fundamental theorem of calculus, and the differentiability of the last equation is given by Assumptions 3 and 5.

Combining Equations (15) to (18), we have the following theorem on $\nabla^2 h(x)$.

Theorem 2. *Suppose that Assumptions 1 to 5 are satisfied. Then*

$$\begin{aligned}\nabla^2 h(x) &= \sum_{i=1}^{m} f_{c_i}(0)\cdot \mathrm{E}\left[\nabla_x^2 c_i \cdot \prod_{j=1,j\neq i}^{m} 1_{\{c_j\geq 0\}} \,\middle|\, c_i = 0\right] \\ &\quad + \sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m} f_{c_i,c_j}(0,0)\cdot \mathrm{E}\left[\nabla_x c_i \nabla_x c_j^T \cdot \prod_{k=1,k\neq i,j}^{m} 1_{\{c_k\geq 0\}} \,\middle|\, c_i = 0, c_j = 0\right] \\ &\quad - \sum_{i=1}^{m} \frac{\partial}{\partial y_i}\left\{ f_{c_i}(y_i)\cdot \mathrm{E}\left[\nabla_x c_i \nabla_x c_i^T \cdot \prod_{j=1,j\neq i}^{m} 1_{\{c_j\geq 0\}} \,\middle|\, c_i = y_i\right]\right\}\Bigg|_{y_i=0}. \end{aligned} \tag{19}$$

Though there is a differentiation sign in the third term of Equation (19), we show in next subsection that it can be computed easily by the kernel method.

3.3 Monte Carlo Method for Computing Hessian

Suppose that we have n i.i.d. observations of ξ, denoted as $\{\xi_1, \xi_2, \ldots, \xi_n\}$. Then for any $x \in X$, all three terms of Equation (19) can be computed through the kernel method. The first term is in the form of $f_B(b)\cdot \mathrm{E}[A|B=b]$, which can be estimated by $\widehat{H}(b)$ of Equation (10).

Now we consider the second term of Equation (19). Let $G(u,v)$ be a symmetric bivariate density such that $\|(u,v)\|^2 \cdot G(u,v) \to 0$ as $\|(u,v)\| \to \infty$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \|(u,v)\|^2 \cdot G(u,v)\, dudv < \infty$. Then $G(u,v)$ is a two-dimensional kernel function (Bosq 1998). For instance, the bivariate normal density $G(u,v) = \frac{1}{2\pi} e^{-(u^2+v^2)/2}$ is a two-dimensional kernel function. Then $f_{B,C}(b,c)\cdot \mathrm{E}[A|B=b,C=c]$ can be estimated by

$$\widehat{J}(b,c) = \frac{1}{n\delta_n^2}\sum_{\ell=1}^{n} A_\ell \cdot G\left(\frac{B_\ell - b}{\delta_n}, \frac{C_\ell - c}{\delta_n}\right),$$

where $\{(A_1,B_1,C_1),\ldots,(A_n,B_n,C_n)\}$ is a sequence of i.i.d. observations of (A,B,C). It can be shown that $\widehat{J}(b,c)$ is a consistent estimator of $f_{B,C}(b,c)\cdot \mathrm{E}[A|B=b,C=c]$ as $n\to\infty$ if $\delta_n \to 0$ and $n\delta_n^2 \to \infty$. Furthermore, the rate of convergence is $(n\delta_n^2)^{-1/2}$ under some technical conditions (Bosq 1998). The asymptotic properties also hold when the observations are not i.i.d., e.g., they are ϕ-mixing. Since the second term of Equation (19) is in the form of $f_{B,C}(b,c)\cdot \mathrm{E}[A|B=b,C=c]$. We may use $\widehat{J}(b,c)$ to estimate it.

Notice that the third term of Equation (19) is in the form of $\frac{d}{db}\{f_B(b)\cdot \mathrm{E}[A|B=b]\}$. Since $f_B(b)\cdot \mathrm{E}[A|B=b]$ can be estimated by $\widehat{H}(b)$ of Equation (10), then it is natural to estiamte

$\frac{d}{db}\{f_B(b) \cdot \mathrm{E}[A|B=b]\}$ by

$$\widehat{L}(b) = \frac{d}{db}\widehat{H}(b) = -\frac{1}{n\delta_n^2}\sum_{\ell=1}^{n} A_\ell \cdot K'\left(\frac{B_\ell - b}{\delta_n}\right),$$

where $K'(u)$ is the derivative of $K(u)$. For instance, $K'(u) = -\frac{u}{\sqrt{2\pi}}e^{-u^2/2}$ when $K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$.

Therefore, we can estimate $\nabla^2 h(x)$ by the following estimator

$$\begin{aligned}
&\widehat{\nabla^2} h(x) \\
&= \frac{1}{n\gamma_{1,n}}\sum_{\ell=1}^{n}\sum_{i=1}^{m} \nabla_x^2 c_i(x,\xi_\ell) \cdot \prod_{j=1,j\neq i}^{m} 1_{\{c_j(x,\xi_\ell)\geq 0\}} \cdot K\left[\frac{c_i(x,\xi_\ell)}{\gamma_{1,n}}\right] \\
&+ \frac{1}{n\gamma_{2,n}^2}\sum_{\ell=1}^{n}\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m} \nabla_x c_i(x,\xi_\ell)\cdot\nabla_x c_j(x,\xi_\ell)^T \cdot \prod_{k=1,k\neq i,j}^{m} 1_{\{c_k(x,\xi_\ell)\geq 0\}} \cdot G\left[\frac{c_i(x,\xi_\ell)}{\gamma_{2,n}}, \frac{c_j(x,\xi_\ell)}{\gamma_{2,n}}\right] \\
&+ \frac{1}{n\gamma_{3,n}^2}\sum_{\ell=1}^{n}\sum_{i=1}^{m} \nabla_x c_i(x,\xi_\ell)\cdot\nabla_x c_i(x,\xi_\ell)^T \cdot \prod_{j=1,j\neq i}^{m} 1_{\{c_j(x,\xi_\ell)\geq 0\}} \cdot K'\left[\frac{c_i(x,\xi_\ell)}{\gamma_{3,n}}\right]. \qquad (20)
\end{aligned}$$

We use $\gamma_{1,n}$, $\gamma_{2,n}$ and $\gamma_{3,n}$ instead of δ_n to denote the bandwidth parameter in $\widehat{\nabla^2} h(x)$, because they are not necessarily the same as the δ_n used in $\widehat{\nabla} h(x)$. Notice that computing $\widehat{\nabla^2} h(x)$ is also an $O(n)$ operation with a fixed m and d. It can also be done efficiently when the sample size n is large. If we also consider the effects of m and d, the computational complexity of computing $\widehat{\nabla^2} h(x)$ becomes $O(nm^2d^2)$. Therefore, when m and d are large, computing $\widehat{\nabla^2} h(x)$ is significantly slower than computing $\widehat{\nabla} h(x)$.

3.4 Chance-Constrained Programs

In this section we outline a simulation-based approach to solve Problem (1). We first generate n i.i.d. observations of ξ through Monte Carlo simulation. We denote them as $\xi_1, \ldots, \xi_n$. We may estimate $h(x)$ by

$$\hat{h}(x) = \frac{1}{n}\sum_{\ell=1}^{n}\prod_{i=1}^{m} 1_{\{c_i(x,\xi_\ell))\geq 0\}},$$

$\nabla h(x)$ by $\widehat{\nabla} h(x)$ of Equation (11) and $\nabla^2 h(x)$ by $\widehat{\nabla^2} h(x)$ of Equation (20) for any $x \in X$. Then we may feed the estimated values of $h(x)$, $\nabla h(x)$ and $\nabla^2 h(x)$ together with $f(x)$, $\nabla f(x)$ and $\nabla^2 f(x)$ into a nonlinear optimization solver to solve the problem. Notice that solving a nonlinear

program is often equivalent to finding solutions to the KKT conditions. Therefore, the convergence of the proposed simulation-based algorithm can be guaranteed by the convergence results on stochastic generalized equations as the sample size $n \to \infty$ (see, for instance, Section 5.2 of Shapiro et al. 2009).

Since Problem (1) may not be a convex program, there may exist multiple local optimal solutions. Our approach may not be able to find the global optimal solution of the problem when the problem is non-convex. Although we cannot guarantee global optimality, we may try to find a good local optimal solution. To achieve this, we suggest to first solve either the CVaR approximation of Rockafellar and Uryasev (2000) or the ϵ-approximation of Hong et al. (2011), then use the solution as the starting solution to our approach. In this way, we ensure that the solution found by our approach is a local optimal solution that is at least better than the CVaR approximation or the ϵ-approximation. Moreover, if the CVaR approximation is used to find a starting solution, we do not recommend using the algorithm proposed by Rockafellar and Uryasev (2000), because it requires solving a convex optimization program with over $n \times m$ constraints and it will be computationally very expensive if the sample size n is large. Instead, we suggest using the algorithm proposed by Hong and Liu (2009) because, similar to our algorithm, it only uses the sample to estimate the value and the gradient of the CVaR function and it can handle cases with a very large sample size.

4 Numerical Example

We consider the norm optimization example, which is used by Hong et al. (2011). Let $\mathbf{x} = (x_1, \ldots, x_d)^T$ denote a d-dimensional vector on $\Re^d$, and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m)$ with $\boldsymbol{\xi}_i = (\xi_{i,1}, \ldots, \xi_{i,d})^T$ be a $d \times m$ matrix of mutually independent and identically distributed (i.i.d.) standard normal random variables. Let $\boldsymbol{\xi}_i \circ \mathbf{x} = (\xi_{i,1}x_1, \ldots, \xi_{i,d}x_d)^T$ be the Hadamard product. Suppose that we are interested in solving the following optimization problem:

$$\begin{aligned} \max \quad & \|\mathbf{x}\|_1 = \sum_{j=1}^{d} |x_i|, \\ s.t. \quad & \Pr\{\|\boldsymbol{\xi}_i \circ \mathbf{x}\| \le M, i = 1, 2, \ldots, m\} \ge 1 - \alpha, \\ & x_j \ge 0, j = 1, \ldots, d. \end{aligned}$$

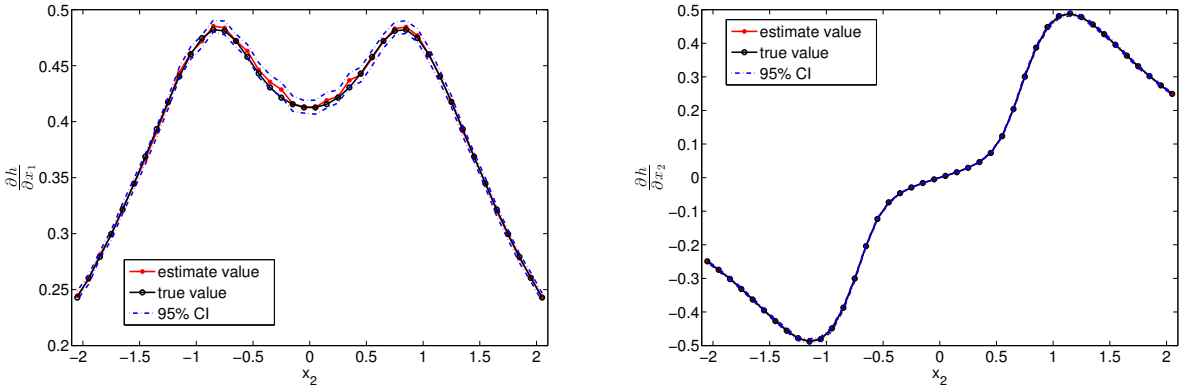
Let $c_i(\mathbf{x}, \boldsymbol{\xi_i}) = M^2 - \sum_{j=1}^{d} x_j^2 \xi_{j,i}^2$. Then by Equation (1), the problem can be reformulated as

$$\begin{aligned} \min \quad & -\sum_{j=1}^{d} x_i, \\ s.t. \quad & h(x) \geq 1 - \alpha, \\ & x_j \geq 0, j = 1, \ldots, d. \end{aligned}$$

4.1 Estimation of Gradient and Hessian

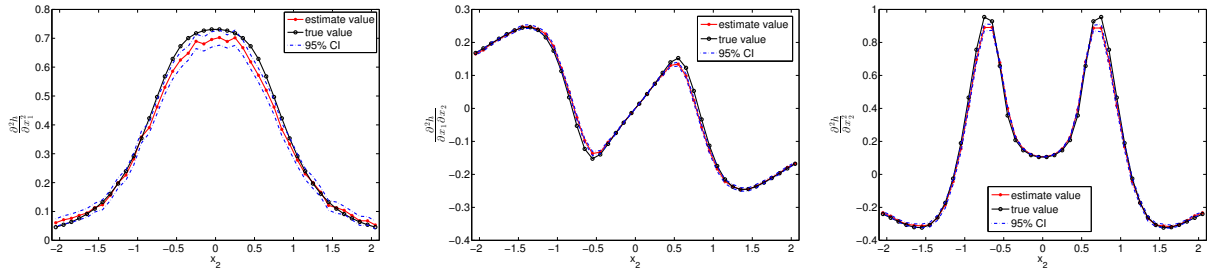
In this subsection, we consider the performance of the gradient and Hessian estimators. During the implementation, the kernel function is chosen to be the standard normal density function. Let $d = 2$, $m = 2$, and $M = 2$. Because $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ are i.i.d., $\Pr\{c_i(\mathbf{x}, \boldsymbol{\xi_i}) \geq 0, i = 1, 2\} = \big(\Pr\{c_1(\mathbf{x}, \boldsymbol{\xi_1}) \geq 0\}\big)^2$. Moreover, $\xi_{1,1}$ and $\xi_{1,2}$ are i.i.d. standard normal random variables, then by the convolution of the density functions, the true value of the gradient and hessian matrix of $h(x)$ can be calculated analytically. We fix $x_1 = 1$, and change x_2 from -2 to 2. Let the sample size $n = 10000$, and the bandwidth $\delta_n = n^{-1/5}$ for the gradient estimator and $\gamma_n = 0.2n^{-1/4}$ for Hessian matrix estimator. Replicate 100 times of the estimators, and the estimated values, true values, and 95% confident intervals (CI) of each elements in gradient and Hessian matrix are reported in Figures 1 and 2. These figures illustrate that our estimators can estimate the gradient and Hessian matrix accurately.

Figure 1: Estimating each component of the gradient



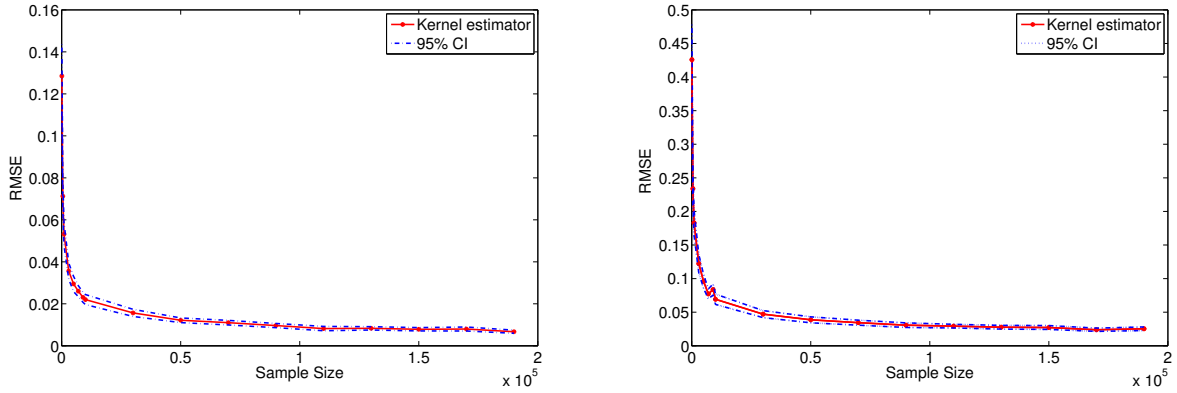
Next, we consider the performance of the estimators against the sample sizes, and obtain Figure 3, which indicates that the estimation accuracy will benefit from the increasing of the sample size.

Figure 2: Estimating each component of the Hessian matrix



However, accuracies of kernel estimators also depend on the choice of the bandwidth. Therefore, in order to improve the estimation accuracy further, bandwidth should be carefully chosen.

Figure 3: RMSE for gradient (left) and Hessian matrix (right) w.r.t different sample sizes



4.2 Optimization results

In this subsection, we study the optimization results by applying the kernel estimators of the gradient and Hessian in optimization solvers. Similar as in Hong et al. (2011), let $d = m = 10$, $M = 10$, $\alpha = 0.1$ sample size $n = 10000$, and setting stopping criteria $X_{tol} = 10^{-3}$. We solve the optimization problem in MATLAB and use the nonlinear optimization solver **FMINCON**. First, we use the CVaR approximation and the ε-approximation solutions as the initial solutions, and then compare the performance of the methods with gradient estimator (denote by G-CVaR and G-eps, respectively), and with both gradient and Hessian matrix estimators (denote by H-CVaR and H-eps, respectively) in the following box plots.

Note that, by Figure 4 and Table 1, Hessian matrix appears not beneficial for the optimization

Figure 4: Box plots of optimal objective value and iteration for G-CVaR, G-eps, H-CVaR, and H-eps

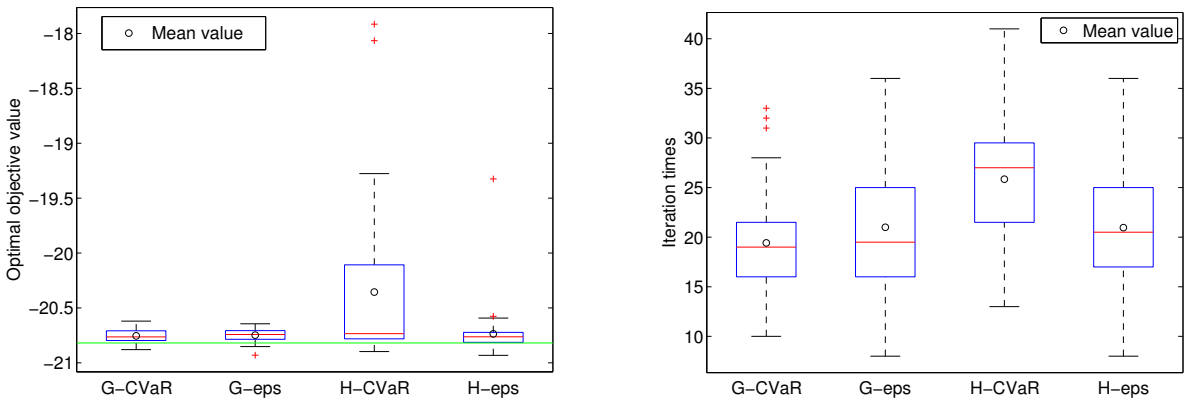


Table 1: CPU time for G-CVaR, G-eps, H-CVaR, and H-eps

	G-CVaR	G-eps	H-CVaR	H-eps
mean	$1.727s$	$1.690s$	$175.1s$	$142.6s$
maximum	$3.011s$	$5.184s$	$269.0s$	$242.2s$
minimun	$0.7504s$	$0.7193s$	$92.20s$	$54.78s$

solver **FMINCON**. The optimal objective values and numbers of iterations of the H-CVaR and the H-eps do not improve comparing with the G-CVaR and the G-eps, respectively. On the contrary, due to the computation of Hessian matrix, the CPU time increases significantly. We also test $d = 2$ and $m = 2$ case, which has an accurate estimation of Hessian matrix according to Figure 2. The results also indicate that adding Hessian matrix in the optimization solver costs more time, and the optimization effects are not improved. Therefore, we conclude that, even though Hessian may be estimated accurately, its use in the solver **FMINCON** does not help solve the joint CCP. However, there might be other solvers that may work well with the information of Hessian and this is a direction of future research.

Next, we compare the G-CVaR and the G-eps with the sequential convex approximation (SCA) algorithm of Hong et al. (2011), with either CVaR or ϵ-approximation as the starting point (denoted as DC-CVaR or DC-eps), and obtain figure 5 and Table 2 with more strict stopping criteria $X_{tol} = 10^{-6}$. For the SCA approximation method, we set the same stopping criteria, i.e., $X_{tol} = 10^{-6}$, in

Figure 5: Box plots of optimal objective value for the G-CVaR, G-eps, DC-CVaR, and DC-eps

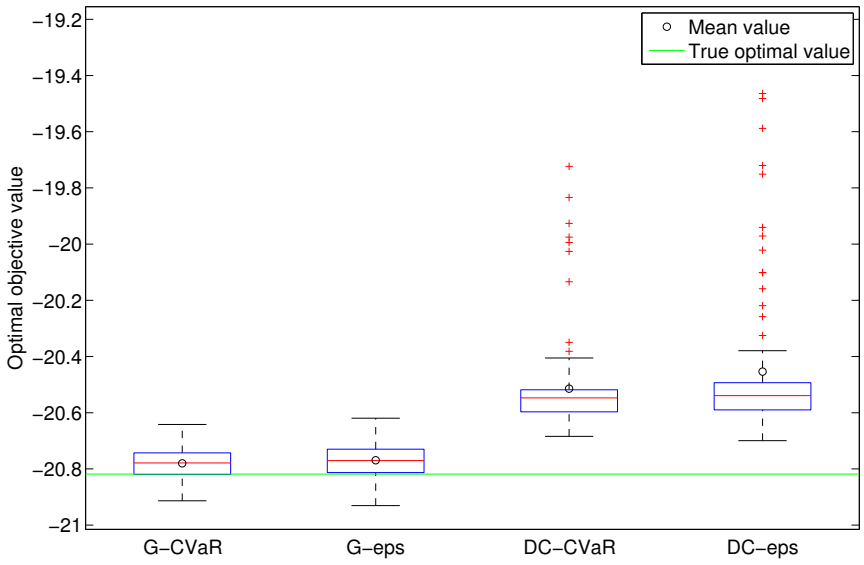
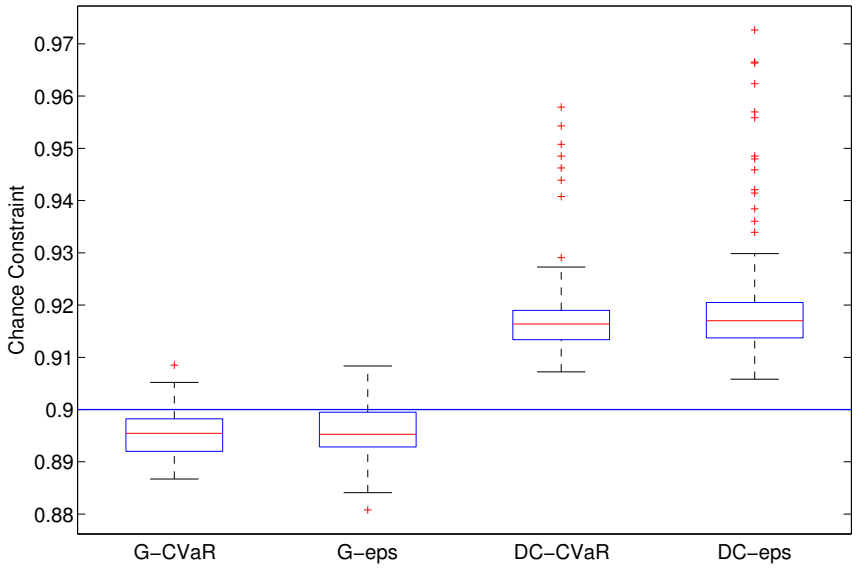


Figure 6: Box plots of the left-hand side of the chance constraint for the solutions found by G-CVaR, G-eps, DC-CVaR, and DC-eps



each iterative optimization, and the stopping criteria for the macro iteration is 10^{-3}.

In Figure 5, the performances of the G-CVaR and the G-eps are better than those of the DC-CVaR and the DC-eps, respectively, i.e. the mean values and the mediums of the G-CVaR and

the G-eps are closer to the optimum. Moreover, from Table 2, the CPU times of the G-CVaR and the G-eps are also much smaller than those of the DC-CVaR and the DC-eps. In Figure 5, we also notice that some optimal values found by G-CVaR and G-eps are better than the true optimal values, indicating that the constraints may be violated. Notice that this is a common issue in sample-based methods where solutions may violate the constraints due to the randomness in the sample. To understand how serious the constraint violation is, we calculate the true values of the left-hand side of the chance constraint for all solutions found by all four algorithms and plot them in Figure 6. From this figure, we see that the solutions found by DC-CVaR and DC-eps always satisfy the chance constraint, while the solutions found by G-CVaR and G-eps typically violate the chance constraint, but the amount of violations are typically quite small. If such constraint violation is a concern in practical applications, one may add a buffer to α (say $\alpha + \beta$). Recently, Lam (2016) provided an approach based on the empirical divergence to determine the appropriate buffer size.

Table 2: CPU time for G-CVaR, G-eps, DC-CVaR, and DC-eps

	G-CVaR	G-eps	DC-CVaR	DC-eps
mean	$17.15s$	$22.13s$	$471.3s$	$671.6s$
maximum	$27.72s$	$31.74s$	$883.4s$	$940.9s$
minimun	$7.557s$	$13.03s$	$33.14s$	$207.0s$

5 Conclusions

In this paper we derive closed-form expressions of the gradient and Hessian of joint probability functions and develop Monte Carlo estimators of them. We then design a Monte Carlo algorithm, based on these estimators, to solve chance constrained programs. Our numerical study shows that the algorithm works well, especially only with the gradient estimators.

There are a few directions for future research in this field. First, we are interested in developing new or utilizing existing algorithms (or solvers) that may take advantage of the information on Hessian matrix. Second, quantifying the convergence and the rate of convergence of the proposed algorithm is certainly an interesting problem that is worth studying. Third, it would be interesting to develop a general guidelines on how to choose the sample size n or a buffer size on constraint

threshold α to reach a pre-determined precision level and satisfy the constraint with a given probability.

Acknowledgement

This research was supported by the Hong Kong Research Grants Council [GRF 613213].

References

A. Ben-Tal and A. Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Math. Programming*, 88(3):411–424, 2000.

D. Bosq. *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Springer, New York, 1998.

G. Calafiore and M. C. Campi. Uncertain convex programs: Randomized solutions and confidence levels. *Math. Programming*, 102(1):25–46, 2005.

G. Calafiore and M. C. Campi. The scenario approach to robust control design. *IEEE Trans. Automat. Control*, 51(5):742–753, 2006.

A. Charnes, W. W. Cooper, and G. H. Symonds. Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil. *Management Sci.*, 4(3):235–263, 1958.

W. Chen, M. Sim, J. Sun, and C.-P. Teo. From CVaR to uncertainty set: Implications in joint chance-constrained optimization. *Oper. Res.*, 58(2):470–485, 2010.

D. Dentcheva, B. Lai, and A. Ruszczyński. Dual methods for probabilistic optimization problems. *Math. Methods Oper. Res.*, 60(2):331–346, 2004.

R. Durrett. *Probability: Theory and Examples*. Duxury Press, Belmont, MA, 2nd edition, 1995.

M. C. Fu. Gradient estimation. In S. G. Henderson and B. L. Nelson, editors, *Handbooks in Operations Research and Management Science*, volume 13, chapter 19, pages 575–616. Elsevier, Amsterdam, 2006.

M .C. Fu and J. Q. Hu. *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Kluwer Academic Publishers, Norwell, MA, 1997.

M. C. Fu, L. J. Hong, and J. Q. Hu. Conditional monte carlo estimation of quantile sensitivities. *Oper. Res.*, 55(12):2019–2027, 2009.

P. Glasserman. *Gradient Estimation via Perturbation Analysis*. Kluwer Academic Publishers, Norwell, MA, 1991.

P. W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Comm. ACM*, 33(10): 75–84, 1990.

Y. C. Ho and X.-R. Cao. Perturbation analysis and optimization of queueing networks. *J. Optim. Theory Appl.*, 40(4):559–582, 1983.

L. J. Hong. Estimating quantile sensitivities. *Oper. Res.*, 57(1):118–130, 2009.

L. J. Hong and G. Liu. Simulating sensitivities of conditional value-at-risk. *Management Sci.*, 55 (2):281–293, 2009.

L. J. Hong and G. Liu. Pathwise estimation of probability sensitivities through terminating and steady-state simulations. *Oper. Res.*, 58(2):357–370, 2010.

L. J. Hong, Y. Yang, and L. Zhang. Sequential convex approximations to joint chance constrained programs: A monte carlo approach. *Oper. Res.*, 59(3):617–630, 2011.

L. J. Hong, Z. Hu, and L. Zhang. Conditional value-at-risk approximation to value-at-risk constrained programs: A remedy via monte carlo. *INFORMS J. Comput.*, 26(2):385–400, 2014.

Z. Hu, L. J. Hong, and L. Zhang. A smooth monte carlo approach to joint chance-constrained programs. *IIE Trans.*, 45(7):716–735, 2013.

C. M. Lagoa, X. Li, and M. Sznaier. Probabilistically constrained linear programs and risk-adjusted controller design. *SIAM J. Optim.*, 15(3):938–951, 2005.

H. Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. Working paper, Department of Industrial and Operations Engineering, University of Michigan, 2016.

G. Liu and L. J. Hong. Kernel estimation of quantile sensitivities. *Nav. Res. Log.*, 56:511–525, 2009.

J. E. Marsden and M. J. Hoffman. *Elementary Classical Analysis*. W. H. Freeman and Co., New York, 2nd edition, 1993.

K. Marti. *Stochastic Optimization Methods*. Springer, Berlin, 2005.

F. W. Meng, J. Sun, and M. Goh. Stochastic optimization problems with CVaR risk measure and their sample average approximation. *J. Optim. Theory Appl.*, 146(2):399–418, 2010.

L. B. Miller and H. Wagner. Chance-constrained programming with joint constraints. *Oper. Res.*, 13(6):930–945, 1965.

A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM J. Optim.*, 17(4):969–996, 2006.

E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Stat.*, 33(3): 1065–1076, 1962.

A. Prékopa. *Stochastic Programming*. Springer, Netherlands, 1995.

A. Prékopa. Probabilistic programming. In A. Ruszczynski and A. Shapiro, editors, *Handbooks in Operations Research and Management Science*, volume 10, chapter 5, pages 267–351. Elsevier, Amsterdam, 2003.

A. Prékopa, T. Rapcsák, and I. Zsuffa. Serially linked reservoir system design using stochastic programming. *Water Resources Res.*, 12(4):672–678, 1978.

M. I. Reiman and A. Weiss. Sensitivity analysis for simulations via likelihood ratios. *Oper. Res.*, 37(5):830–844, 1989.

R. T. Rockafellar and S. P. Uryasev. Optimization of conditional value-at-risk. *J. Risk*, 2(3):21–41, 2000.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelphia, 2009.

H. Sun, H. Xu, and Y. Wang. Asymptotic analysis of sample average approximation for stochastic optimization problems with joint chance constraints via conditional value at risk and difference of convex functions. *J. Optim. Theory Appl.*, 161(1):257–284, 2014.

S. P. Uryasev. A differentiation formula for integrals over sets given by inclusion. *Numer. Funct. Anal. Optim.*, 10(7):827–841, 1989.