Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

SAMPLING UNCERTAIN CONSTRAINTS UNDER PARAMETRIC DISTRIBUTIONS

Henry Lam
Fengpei Li

Department of Industrial Engineering and Operations Research
Columbia University
500 W. 120th Street
New York, NY 10027, USA

ABSTRACT

We consider optimization problems with uncertain constraints that need to be satisfied probabilistically. When data are available, a common method to obtain feasible solutions for such problems is to impose sampled constraints, following the so-called scenario generation (SG) approach. However, when the data size is small, the sampled constraints may not support a guarantee on the feasibility of the obtained solution. This paper studies how to leverage parametric information and the power of Monte Carlo simulation to obtain feasible solutions even when the data are not sufficient to support the use of SG. Our approach makes use of a distributionally robust optimization formulation that informs the Monte Carlo sample size needed to achieve our guarantee.

1 INTRODUCTION

We consider optimization problems in the form

$$\begin{aligned} \min_{x\in\mathscr{X}\subseteq\mathbb{R}^d} \quad & c^T x, \\ \text{subject to} \quad & \mathbb{P}(x\in\mathscr{X}_\xi)\geq 1-\varepsilon, \end{aligned} \tag{1}$$

where $\mathbb{P}$ is a probability measure governing the random variable ξ, and $\mathscr{X}_\xi \subset \mathscr{X}$ is a set depending on ξ. Problem (1) enforces a solution x to satisfy $x\in\mathscr{X}_\xi$ with high probability, namely at least $1-\varepsilon$. This problem is often known as a chance-constrained or probabilistically constrained optimization (e.g., Prékopa 2003). It provides a natural framework for decision-making under stochastic resource capacity or risk tolerance, and has been applied in various domains such as production planning (Murr and Prékopa 2000), inventory management (Lejeune and Ruszczynski 2007), reservoir design (Prékopa and Szántai 1978; Prékopa et al. 1978), communications (Shi et al. 2015), and ranking and selection (Hong et al. 2015).

We focus on the situations where $\mathbb{P}$ is unknown, but some data, say $\xi_1,\ldots,\xi_n$, are available. One common approach to handle (1) in these situations is to use the so-called scenario generation (SG) or constraint sampling. This replaces the unknown constraint in (1) with $x\in\mathscr{X}_{\xi_i}, i=1,\ldots,n$, namely, by considering

$$\begin{aligned} \min_{x\in\mathscr{X}\subseteq\mathbb{R}^d} \quad & c^T x, \\ \text{subject to} \quad & x\in\mathscr{X}_{\xi_i},\ i=1,\ldots,n. \end{aligned} \tag{2}$$

Note that the chance-constrained problem (1) is generally difficult to solve, even when the set $\mathscr{X}_\xi$ is convex and tractable for any given ξ (Prékopa 2003). Thus, in the latter case, the sampled problem (2) offers an additional benefit in approximating the intractable problem with a more tractable one.

Our goal is to find a good feasible solution for (1) in the described data-driven context above. Note that, because of the statistical noise from the data, we can at best find a solution that is feasible with a high

confidence. Define $V(x,\mathbb{P}) = \mathbb{P}(x \notin \mathscr{X}_\xi)$ to be the violation probability of a solution x for the condition $x \in \mathscr{X}_\xi$ under probability measure $\mathbb{P}$. For any $\hat{x}$ that is obtained from finite data, we want to make sure that

$$\mathbb{P}_{data}(V(\hat{x},\mathbb{P}) \leq \varepsilon) \geq 1-\alpha, \tag{3}$$

for a given confidence level $1-\alpha$ (e.g., $\alpha = 5\%$), where $\mathbb{P}_{data}$ denotes the measure generating the data $\xi_i, i = 1,\ldots,n$ (and hence $\hat{x}$).

As $n \to \infty$, one would expect that the space of ξ is sufficiently populated by the sample and $V(\hat{x},\mathbb{P}) \to 0$. The question is then how many observations are enough to see this behavior. The seminal work of Campi and Garatti (2008) shows a tight bound on the number of observations, or sampled constraints, needed to guarantee (3), for given values of ε and α. In particular, define $\gamma(N,\varepsilon,d) = \sum_{i=0}^{d-1} \binom{N}{i} \varepsilon^i (1-\varepsilon)^{N-i}$ for positive integers $N \geq d \geq 1$ and $0 < \varepsilon < 1$. Under the convexity of $\mathscr{X}_\xi$ and some additional mild assumptions, Campi and Garatti (2008) proves that an optimal solution obtained from solving (2) satisfies

$$\mathbb{P}_{data}(V(\hat{x},\mathbb{P}) > \varepsilon) \leq \gamma(n,\varepsilon,d). \tag{4}$$

Thus, if we can find n such that $\gamma(n,\varepsilon,d) \leq \alpha$, then we achieve (3).

In this paper, we focus on the small-sample situation, so that our data size n is not enough to support $\gamma(n,\varepsilon,d) \leq \alpha$. Note that, in using this result from Campi and Garatti (2008), n can be seen to be linear in the decision dimension d, and so for high-dimensional problems this small-sample situation can happen frequently. This dimensional dependence also appears in other sample size bounds (e.g., De Farias and Van Roy 2004; Luedtke and Ahmed 2008). To overcome this challenge, several recent methods have been suggested, such as the use of support rank and solution-dependent support constraints (Schildbach et al. 2013; Campi and Garatti 2018), regularization (Campi and Carè 2013), and sequential approaches (Carè et al. 2014; Calafiore et al. 2011; Chamanbaz et al. 2016; Calafiore 2017). They aim to alleviate the dependence on d, and thus substantially extend the scope of applicability of SG.

Our main contribution in this paper is to study a different path in obtaining guarantee (3) in small-sample situation, in the settings where $\mathbb{P}$ is assumed a parametric structure. The unknown and estimable quantity is the set of parameters in $\mathbb{P}$. We also assume that one can simulate from the parametric model $\mathbb{P}$ using Monte Carlo (an assumption applied for all common parametric models). We will see how such a capability, which can be viewed as a way of generating additional synthetic data, can be combined with the relation (4) to obtain a scheme applicable when n does not support using (4) directly. Unlike some other techniques that reduce the data size needed for SG, our procedure resembles closely the standard SG (2). The differences are in the algorithmic parameters and the distribution we sample from in the Monte Carlo scheme.

Our derivation relies on casting the uncertain-parameter problem as a distributionally robust optimization (DRO) (Delage and Ye 2010; Wiesemann et al. 2014). This approach considers the worst-case situation among all parameters that lie in a so-called uncertainty set or ambiguity set. In the chance constraint framework, this entails replacing the chance constraint with unknown distribution with a worst-case chance constraint within the uncertainty set (Hanasusanto et al. 2015; Zymler et al. 2013; Hanasusanto et al. 2017; Li et al. ; Jiang and Guan 2016; Zhang et al. 2016). The DRO approach has been used in stochastic simulation, bearing names such as the robust Monte Carlo (Hu et al. 2012; Glasserman and Xu 2014; Lam 2016; Hu and Hong 2015; Lam 2018; Ghosh and Lam 2018). Our procedure will rely on a suitable DRO formulation using statistical distances (Petersen et al. 2000; Ben-Tal and Nemirovski 2000; Lim et al. 2006; Love and Bayraksan 2015), and the particular SG we use has a similar favor as the robust Monte Carlo considered in Hu et al. (2012) and Glasserman and Xu (2014), as we also utilize a change-of-measure argument in arriving at our scheme. Lastly, Erdoğan and Iyengar (2006) considers an SG for robust chance-constrained problems that is related to our proposal. Erdoğan and Iyengar (2006) considers uncertainty set based on the Prohorov distance and derives bounds for the required sample size. On the other hand, our approach uses the class of ϕ- or f-divergences, which is readily estimatable especially in the parametric setting. Moreover, our focus is on utilizing the convexity-based sample size estimate in Campi and Garatti (2008), in contrast to the Vapnik-Chervonenkis dimension used in Erdoğan and Iyengar

(2006). We will study the required Monte Carlo size in relation to the data size, and the choice of the Monte Carlo distribution, which are quite different from Erdoğan and Iyengar (2006).

2 OUTLINE OF THE METHOD

Recall that we are interested in finding a solution $\hat{x}$ such that (3) holds, in the case that $\mathbb{P}$ is observable only through data. Suppose that $\mathbb{P} \in \mathscr{P}$, the class of all possible probability distributions for ξ (which we shall exemplify later). Suppose that given our data, we can find an uncertainty set $\mathscr{U}_{data} \in \mathscr{P}$ such that

$$\mathbb{P}_{data}(\mathbb{P} \in \mathscr{U}_{data}) \geq 1-\alpha. \tag{5}$$

We then proceed to consider a distributionally robust chance-constrained problem

$$\begin{array}{ll} \min\limits_{x \in \mathscr{X} \subseteq \mathbb{R}^d} & c^T x, \\ \text{subject to} & \inf\limits_{\mathbb{Q} \in \mathscr{U}_{data}} \mathbb{Q}(x \in \mathscr{X}_\xi) \geq 1-\varepsilon, \end{array} \tag{6}$$

where $\mathbb{Q}$ is the probability measure for ξ. If we can find a solution $\hat{x}$ feasible for (6), then this $\hat{x}$ is also feasible for (1) with confidence at least $1-\alpha$. This is because if $\mathbb{P} \in \mathscr{U}_{data}$, then for any x feasible for (6), $\mathbb{P}(x \in \mathscr{X}_\xi) \geq \inf_{\mathbb{Q} \in \mathscr{U}_{data}} \mathbb{Q}(x \in \mathscr{X}_\xi) \geq 1-\varepsilon$, and therefore $\mathbb{P}_{data}(V(\hat{x}, \mathbb{P}) \leq \varepsilon) \geq \mathbb{P}_{data}(\mathbb{P} \in \mathscr{U}_{data}) \geq 1-\alpha$.

We now consider a Monte Carlo scheme from some "baseline" distribution $\mathbb{P}_0$ (which can depend on the data), to obtain a solution that is feasible for (6) with a confidence of, say, $1-\beta$. This is achievable by using a certain Monte Carlo size N. To obtain this number, we find a bound on $\sup\limits_{\mathbb{Q} \in \mathscr{U}_{data}} V(\hat{x}(\mathbb{P}_0), \mathbb{Q})$, where $\hat{x}(\mathbb{P}_0)$ is obtained from solving (2) using ξ_i's generated from $\mathbb{P}_0$. Call this bound $M(\mathbb{P}_0, \mathscr{U}_{data}, V(\hat{x}(\mathbb{P}_0), \mathbb{P}_0))$. In other words, it satisfies

$$\sup_{\mathbb{Q} \in \mathscr{U}_{data}} V(\hat{x}(\mathbb{P}_0), \mathbb{Q}) \leq M(\mathbb{P}_0, \mathscr{U}_{data}, V(\hat{x}(\mathbb{P}_0), \mathbb{P}_0)). \tag{7}$$

Moreover, suppose we also have that $M(\mathbb{P}_0, \mathscr{U}_{data}, v)$ is non-decreasing in $v > 0$. We then find a $\delta > 0$ such that

$$M(\mathbb{P}_0, \mathscr{U}_{data}, \delta) \leq \varepsilon. \tag{8}$$

Then, by using the result in Campi and Garatti (2008) discussed in the introduction, we know that $\mathbb{P}_{MC,0}(V(\hat{x}(\mathbb{P}_0), \mathbb{P}_0) > \delta) \leq \gamma(N, \delta, d)$ where $\mathbb{P}_{MC,0}$ is the measure generating N Monte Carlo samples from $\mathbb{P}_0$ to obtain $\hat{x}(\mathbb{P}_0)$, which holds independent of the choice of $\mathbb{P}_0$. We find N such that

$$\gamma(N, \delta, d) \leq \beta. \tag{9}$$

This choice of N then gives us, with confidence $1-\beta$, that $V(\hat{x}(\mathbb{P}_0), \mathbb{P}_0) \leq \delta$ and hence $\sup\limits_{\mathbb{Q} \in \mathscr{U}_{data}} V(\hat{x}(\mathbb{P}_0), \mathbb{Q}) \leq \varepsilon$ by (7), (8) and the monotonicity of M. This in turn leads to the conclusion that $\hat{x}(\mathbb{P}_0)$ is feasible for (6) with confidence at least $1-\beta$.

Thus, overall, if we choose $\mathscr{U}_{data}$ to satisfy (5) and N to satisfy (8) and (9), our obtained solution $\hat{x}(\mathbb{P}_0)$ is feasible for (1) with confidence at least $1-\alpha-\beta$. Note that our argument relies crucially on generating a bound M that translates the ambiguous violation probability $\sup\limits_{\mathbb{Q} \in \mathscr{U}_{data}} V(\hat{x}(\mathbb{P}_0), \mathbb{Q})$ into a quantity depending instead on the baseline violation probability $V(\hat{x}(\mathbb{P}_0), \mathbb{P}_0)$. Next section will address how to do so.

2.1 Bounding Ambiguous Violation Probability

We study the bound $M(\mathbb{P}_0, \mathscr{U}_{data}, V(\hat{x}(\mathbb{P}_0), \mathbb{P}_0))$. In fact, we will consider a more general result. Consider any measurable set $\mathscr{A} \subset \mathscr{Y}$ and set $\mathscr{U} \subset \mathscr{P}$. We will find an M such that

$$\sup_{\mathbb{Q} \in \mathscr{U}} \mathbb{Q}(\xi \in \mathscr{A}) \leq M(\mathbb{P}_0, \mathscr{U}, \mathbb{P}_0(\xi \in \mathscr{A})).$$

For two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ that are both dominated by a common measure ν on $\mathscr{Y}$, with Radon-Nikodym derivatives $\frac{d\mathbb{P}_1}{d\nu}$ and $\frac{d\mathbb{P}_2}{d\nu}$ respectively, we define the χ^2-distance between $\mathbb{P}_1$ and $\mathbb{P}_2$ as

$$\chi^2(\mathbb{P}_1,\mathbb{P}_2)=\int_{\mathscr{Y}}\frac{(\frac{d\mathbb{P}_2}{d\nu}-\frac{d\mathbb{P}_1}{d\nu})^2}{\frac{d\mathbb{P}_1}{d\nu}}\nu(dy)=\int_{\mathscr{Y}}(\frac{d\mathbb{P}_2}{d\nu}/\frac{d\mathbb{P}_1}{d\nu}-1)^2\frac{d\mathbb{P}_1}{d\nu}\nu(dy).$$

When $\mathbb{P}_2$ is absolutely continuous with respect to $\mathbb{P}_1$, we have in particular

$$\chi^2(\mathbb{P}_1,\mathbb{P}_2)=\int_{\mathscr{Y}}(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}-1)^2\mathbb{P}_1(dy)=\int_{\mathscr{Y}}\frac{d\mathbb{P}_2}{d\mathbb{P}_1}\mathbb{P}_2(dy)-1.$$

Suppose that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}_0$. We have

$$\begin{aligned}
\sup_{\mathbb{Q}\in\mathscr{U}}\mathbb{Q}(\xi\in\mathscr{A}) &= \mathbb{P}_0(\xi\in\mathscr{A})+\Big(\sup_{\mathbb{Q}\in\mathscr{U}}\mathbb{Q}(\xi\in\mathscr{A})-\mathbb{P}_0(\xi\in\mathscr{A})\Big)\\
&= \mathbb{P}_0(\xi\in\mathscr{A})+\sup_{\mathbb{Q}\in\mathscr{U}}\int\mathbf{1}\{y\in\mathscr{A}\}\mathbb{Q}(dy)-\int\mathbf{1}\{y\in\mathscr{A}\}\mathbb{P}_0(dy)\\
&= \mathbb{P}_0(\xi\in\mathscr{A})+\sup_{\mathbb{Q}\in\mathscr{U}}\int\mathbf{1}\{y\in\mathscr{A}\}\left(\frac{d\mathbb{Q}}{d\mathbb{P}_0}-1\right)\mathbb{P}_0(dy)\\
&\le \mathbb{P}_0(\xi\in\mathscr{A})+\sup_{\mathbb{Q}\in\mathscr{U}}\left(\int_{\mathscr{Y}}\mathbf{1}\{y\in\mathscr{A}\}\mathbb{P}_0(dy)\right)^{1/2}\left(\int_{\mathscr{Y}}\left(\frac{d\mathbb{Q}}{d\mathbb{P}_0}-1\right)^2\mathbb{P}_0(dy)\right)^{1/2}\\
&= \mathbb{P}_0(\xi\in\mathscr{A})+\mathbb{P}_0(\xi\in\mathscr{A})^{1/2}\cdot(\sup_{\mathbb{Q}\in\mathscr{U}}\chi^2(\mathbb{P}_0,\mathbb{Q}))^{1/2},
\end{aligned}\tag{10}$$

where the inequality follows from the Cauchy-Schwarz inequality. Note that the bound (10) holds if $\mathscr{A}$ and $\mathscr{U}$ are random sets, potentially dependent on each others, but independent from ξ generated from $\mathbb{Q}$ or $\mathbb{P}_0$ above. Thus, by plugging in $\hat{x}(\mathbb{P}_0)\notin\mathscr{X}_\xi$ as $\xi\in\mathscr{A}$ and $\mathscr{U}_{data}$ as $\mathscr{U}$, we have

$$\sup_{\mathbb{Q}\in\mathscr{U}_{data}}V(\hat{x}(\mathbb{P}_0),\mathbb{Q})\le V(\hat{x}(\mathbb{P}_0),\mathbb{P}_0)+V(\hat{x}(\mathbb{P}_0),\mathbb{P}_0)^{1/2}\cdot(\sup_{\mathbb{Q}\in\mathscr{U}_{data}}\chi^2(\mathbb{P}_0,\mathbb{Q}))^{1/2}.$$

Thus, we have identified

$$M(\mathbb{P}_0,\mathscr{U}_{data},v)=v+v^{1/2}\cdot(\sup_{\mathbb{Q}\in\mathscr{U}_{data}}\chi^2(\mathbb{P}_0,\mathbb{Q}))^{1/2},\tag{11}$$

which is non-decreasing in v.

2.2 Choices of Baseline and Uncertainty Set

Our next step as outlined in the beginning of Section 2 is to find δ such that $M(\mathbb{P}_0,\mathscr{U}_{data},\delta)\le\varepsilon$, and moreover a $\mathscr{U}_{data}$ that satisfies (5). For these, we now make an assumption that $\mathbb{P}$, the true distribution of ξ, is known to lie in a parametric family $\mathscr{P}=\{\mathbb{P}_\theta\}_{\theta\in\Theta\subset\mathbb{R}^p}$ indexed by θ, which has dimension p.

We make a convenient choice for $\mathbb{P}_0$ and $\mathscr{U}_{data}$. Namely, we choose $\mathbb{P}_0$ to be $\mathbb{P}_{\hat{\theta}}$, where $\hat{\theta}$ is the maximum likelihood estimator for θ from the data $\xi_i, i=1,\ldots,n$. Then we set

$$\mathscr{U}_{data}=\left\{\mathbb{Q}\in\mathscr{P}:\chi^2(\mathbb{P}_{\hat{\theta}},\mathbb{Q})\le\frac{\chi^2_{1-\alpha,p}}{n}\right\},\tag{12}$$

where $\chi^2_{1-\alpha,p}$ is the $1-\alpha$-quantile of the χ^2-distribution with degree of freedom p. By divergence-based inference (e.g., Pardo 2005), we have that (12) satisfies

$$\lim_{n\to\infty}\mathbb{P}_{data}(\mathbb{P}\in\mathscr{U}_{data})=1-\alpha.$$

Moreover, (11) becomes

$$v + v^{1/2}\left(\frac{\chi^2_{1-\alpha,p}}{n}\right)^{1/2}.$$

Thus, we choose δ such that $\delta + \delta^{1/2}\left(\frac{\chi^2_{1-\alpha,p}}{n}\right)^{1/2} \leq \varepsilon$, or equivalently

$$\delta = \varepsilon + \frac{\chi^2_{1-\alpha,p}}{2n} - \sqrt{\varepsilon \cdot \frac{\chi^2_{1-\alpha,p}}{n} + \frac{\left(\chi^2_{1-\alpha,p}\right)^2}{4n^2}}, \tag{13}$$

and the Monte Carlo size N such that $\gamma(N,\delta,d) \leq \beta$. Using this N number of samples generated from $\mathbb{P}_{\hat{\theta}}$ to construct (2) then guarantees that the obtained solution $\hat{x}(\mathbb{P}_{\hat{\theta}})$ is feasible for (1) with a confidence $1-\alpha-\beta+o(1)$ as $n \to \infty$. We note that formula (13) appears in a related context in Proposition 2 in Tseng et al. (2016), but here we are motivated by the use of SG in very limited data situations and investigate the joint construction of the uncertainty set and SG with overall statistical guarantees, which is different from the work of Tseng et al. (2016).

Note that the χ^2-distance we used above is one of many distances between probability distributions that can be categorized under the framework of ϕ- or f-divergences (Vajda 1972). There are reasons to believe that χ^2 is close to the best under our discussed framework. First is that absolute continuity between the true distribution $\mathbb{P}$ and the baseline distribution $\mathbb{P}_0$ to generate Monte Carlo samples is critical for our approach. Suppose they are not absolutely continuous. One could attempt to form an uncertainty set $\mathscr{U}_{data}$ that contains both $\mathbb{P}$ and $\mathbb{P}_0$ with high probability (this can be done using, e.g., Wasserstein distance; Esfahani and Kuhn 2015; Blanchet and Kang 2016; Gao and Kleywegt 2016). However, the difference $\sup_{\mathbb{Q}\in\mathscr{U}_{data}} \mathbb{Q}(\xi \in \mathscr{A}) - \mathbb{P}_0(\xi \in \mathscr{A})$ becomes more difficult to control as it can depend intricately on the set $\mathscr{A}$. On the other hand, one can instead use $\mathscr{U}_{data}$ that contains distributions absolutely continuous with respect to $\mathbb{P}$ or vice versa. But if $\mathbb{P}_0$ lies outside this set, the difference $\sup_{\mathbb{Q}\in\mathscr{U}_{data}} \mathbb{Q}(\xi \in \mathscr{A}) - \mathbb{P}_0(\xi \in \mathscr{A})$ is again hard to control. Therefore, to apply our framework, choosing a set $\mathscr{U}_{data}$ that contains distributions absolutely continuous with respect to $\mathbb{P}_0$ and also contains $\mathbb{P}$ facilitates the error control tremendously.

Next, many other ϕ-divergences can work under our framework, since one can find $\mathscr{U}_{data}$ in much the same way using general divergence-based inference tools. This comes from the fact that these divergences between two distributions indexed by θ_1 and θ_2 have the same expansion up to the second order (Nielsen and Nock 2014). Then, as long as we can bound the difference $\sup_{\mathbb{Q}\in\mathscr{U}_{data}} \mathbb{Q}(\xi \in \mathscr{A}) - \mathbb{P}_0(\xi \in \mathscr{A})$, the same argument to look for a suitable N will apply. However, it appears that χ^2-distance can perform better than other common candidates, including the Kullback-Leibler (KL) divergence, precisely because of this difference bound. In the KL case, Pinkser's inequality gives rise to an analog of (10) as

$$\sup_{\mathbb{Q}\in\mathscr{U}_{data}} \mathbb{Q}(\xi \in \mathscr{A}) \leq \mathbb{P}_0(\xi \in \mathscr{A}) + 0.5^{1/2}\left(\sup_{\mathbb{Q}\in\mathscr{U}_{data}} KL(\mathbb{P}_0,\mathbb{Q})\right)^{1/2}.$$

where $KL(\mathbb{P}_0,\mathbb{Q})$ denotes the KL divergence between $\mathbb{P}_0$ and $\mathbb{Q}$, which, using our machinery described above, is not as tight as using (10) when the ultimate value of δ is less than 0.5.

Lastly, note that we have constructed $\mathscr{U}_{data}$ that satisfies (5) asymptotically. To provide finite-sample bound, we can use concentration inequalities regarding the MLE estimators (see Korostelev and Korosteleva 2011). Though we do not go into the details here, we note that our previous discussion is exact for Gaussian distributions with unknown means.

3 GENERALIZATIONS AND MIXTURE BASELINE DISTRIBUTIONS

Note that an important element to determine the sample size requirement of our scheme is the quantity $\sup_{\mathbb{Q}\in\mathscr{U}_{data}} \chi^2(\mathbb{P}_0,\mathbb{Q})$ in (11). We have chosen a convenient choice in the last section for this quantity, which involves choosing $\mathbb{P}_0$ and $\mathscr{U}_{data}$. Within the class of ϕ- or f-divergences, the choice $\mathscr{U}_{data}$ is asymptotically the same. To express this more precisely, we re-parametrize the uncertainty set $\tilde{\mathscr{U}}_{data}$ to be over the space of the parameter $\theta\in\Theta$, and work with $\sup_{\theta\in\tilde{\mathscr{U}}_{data}} \chi^2(\mathbb{P}_0,\mathbb{P}_\theta)$ instead. We let

$$\tilde{\mathscr{U}}_{data} \triangleq \left\{\theta\in\Theta : (\theta-\hat{\theta}_n)^T I(\hat{\theta}_n)(\theta-\hat{\theta}_n) \le \frac{\chi^2_{1-\alpha,p}}{n}\right\}, \tag{14}$$

where $I(\hat{\theta}_n)$ is the estimated Fisher information, and $\hat{\theta}_n$ is the MLE in which we highlight the dependence on n. All divergence-based uncertainty sets $\mathscr{U}_{data}$ has an equivalent asymptotic form as (14).

The natural question to ask is which baseline measure $\mathbb{P}_0$ we should choose to generate the Monte Carlo sample. For convenience, we denote $\mathscr{D}_{data}(\mathbb{P}_0) = \sup_{\theta\in\tilde{\mathscr{U}}_{data}} \chi^2(\mathbb{P}_0,\mathbb{P}_\theta)$. Ideally, we would like to choose $\mathbb{P}_0$ so that $\mathscr{D}_{data}(\mathbb{P}_0)$ is computable and also minimized, which then leads to a computable and minimized required number of sample N. To proceed, suppose further that $\mathbb{P}_\theta$ has a density $p(y;\theta)$ on $\mathscr{Y}$. Furthermore, we focus on $\mathbb{P}_0$ that has a density $p_0(y)$ in the mixture form

$$p_0(y) = \int_{\tilde{\mathscr{U}}_{data}} p(y;\theta)\mu(d\theta),$$

where μ is a distribution on $\theta\in\tilde{\mathscr{U}}_{data}$, and for convenience we call this associated class of distributions $\mathscr{P}(\tilde{\mathscr{U}}_{data})$. This choice of density is handy to simulate from because we can simply sample $\theta\sim\mu(d\theta)$ and then $\xi\sim\mathbb{P}_\theta$.

To minimize $\mathscr{D}_{data}(\mathbb{P}_0)$, we write

$$\begin{aligned}
\mathscr{D}_{data}(\mathbb{P}_0) &= \sup_{\theta\in\tilde{\mathscr{U}}_{data}} \int_{\mathscr{Y}} \Big(\frac{p(y;\theta)}{p_0(y)}-1\Big)^2 p_0(y)dy \\
&= \sup_{\theta\in\tilde{\mathscr{U}}_{data}} \int_{\mathscr{Y}} \frac{(p(y;\theta))^2}{p_0(y)}dy - 1 \\
&= \sup_{\theta\in\tilde{\mathscr{U}}_{data}} \int_{\mathscr{Y}} \frac{(p(y;\theta))^2}{\int_{\tilde{\mathscr{U}}_{data}} p(y;\theta)\mu(d\theta)}dy - 1.
\end{aligned} \tag{15}$$

We define

$$L(\mu,\theta) \triangleq \int_{\mathscr{Y}} \frac{(p(y;\theta))^2}{\int_{\tilde{\mathscr{U}}_{data}} p(y;\theta)\mu(d\theta)}dy \quad\text{and}\quad \phi(\mu) \triangleq \sup_{\theta\in\tilde{\mathscr{U}}_{data}} L(\mu,\theta).$$

It follows from (15) that minimizing $\mathscr{D}_{data}(\mathbb{P}_0)$ is equivalent to solving:

$$\min_{\mu\in\mathscr{P}(\tilde{\mathscr{U}}_{data})} \max_{\theta\in\tilde{\mathscr{U}}_{data}} L(\mu,\theta) = \min_{\mu\in\mathscr{P}(\tilde{\mathscr{U}}_{data})} \phi(\mu). \tag{16}$$

Note that the right hand side of (16) is a convex minimization problem, thanks to the fact that $1/x$ is a convex function for $x>0$.

However, even though the minimization in (16) is convex, solving it exactly is difficult due to the inner maximization. Thus we will aim for some heuristic methods to improve the choice of the baseline distribution. To gain some insights, we first consider the value of $\phi(\mu)$ in the case where we pick

$\mu = \delta_{\hat{\theta}_n} \in \mathscr{P}(\tilde{\mathscr{U}}_{data})$, the point mass at the MLE estimator at $\delta_{\hat{\theta}_n}$, so that $\mathbb{P}_0 = \mathbb{P}_{\hat{\theta}_n}$. Now we want to see if we can find some other $\mu \in \mathscr{P}(\tilde{\mathscr{U}}_{data})$ with a smaller value of $\phi(\mu)$. In this way, even if we cannot exactly solve (16), we can still decrease the value of $\mathscr{D}_{data}(\mathbb{P}_0)$. Here, we propose a mixture $\mu_{prop}(d\theta) \in \mathscr{P}(\tilde{\mathscr{U}}_{data})$ on the boundary of $\tilde{\mathscr{U}}_{data}$ that is easy to simulate and shows promising practical performance. Specifically, we define the support of measure $\mu_{prop}(d\theta)$ to be

$$\Theta(\hat{\theta}_n) = \{\theta \in \tilde{\mathscr{U}}_{data} : (\theta - \hat{\theta}_n)^T I(\hat{\theta}_n)(\theta - \hat{\theta}_n) = \frac{\chi^2_{1-\alpha,p}}{n}\}.$$

Now, to sample $\theta \sim \mu_{prop}(d\theta)$, we first sample a random vector $\eta \in \mathbb{R}^p$ uniformly on the surface of the p dimension ball with radius $\sqrt{\frac{\chi^2_{1-\alpha,p}}{n}}$. In particular, this can be achieved by sampling from p number of independent standard normal variables and scale them to have $\sqrt{\frac{\chi^2_{1-\alpha,p}}{n}}$ unit of length (see Muller (1959)). Then, we set $\theta = \hat{\theta}_n + (I(\hat{\theta}_n))^{-1/2}\eta$. In other words, we have chosen μ_{prop} such that if $\theta \sim \mu_{prop}(d\theta)$, then $(I(\hat{\theta}_n))^{1/2}(\theta - \hat{\theta}_n)$ is uniformly distributed on the surface of the p-dimensional ball with radius $\sqrt{\frac{\chi^2_{1-\alpha,p}}{n}}$. We want to simplify problem (16) by finding $0 < t < 1$ that minimizes

$$\phi(t) = \phi((1-t)\delta_{\hat{\theta}_n} + t\mu_{prop}),$$

using line search or the bisection method. In the case where we have checked $\phi(1) < \phi(0)$, we are certain to make improvement over using μ as the point mass at $\hat{\theta}_n$.

In general, if we cannot directly compute $p_0(y) = \int_{\tilde{\mathscr{U}}_{data}} p(y;\theta)\mu_{prop}(d\theta)$ and solve for $\phi(\mu_{prop})$, an alternate is to use Monte Carlo samples of $\{\theta_i\}_{i \le S}$ and approximate $\hat{p_0}_S(y) = \sum_{i=1}^S p(y,\theta_i)/N = \int_{\tilde{\mathscr{U}}_{data}} p(y;\theta)\mu_{emprical}(d\theta)$, where $\mu_{empirical}$ is the empirical distribution of these θ_i's. Then we try to calculate $\phi((1-t)\delta_{\hat{\theta}_n} + t\mu_{empricial})$ for different $0 < t < 1$ to see if there is improvement.

4 NUMERICAL EXPERIMENTS

We present some numerical examples to illustrate the performance of our method. We first focus on the computational costs for different level of accuracy ε and dimension d. Next, we compare with the result from standard SG. Finally, we demonstrate how choosing an effective baseline distribution $\mathbb{P}_0$ can reduce the required sample size and decrease the computational cost.

We will perform experiments on multivariate Gaussian random variables and exponential random variables on single linear chance constrained programs (CCP) and joint linear CCP.

- For every experiment in each problem, we obtain an optimal solution $\hat{x}$ by solving SG, and evaluate the violation probability $V(\hat{x},\mathbb{P})$ under the true probability measure $\mathbb{P}_{\theta_0}$ (where θ_0 denotes the true parameter value) either through exact calculation by CDF or numerical expeiments of 10000 times. Then we calculate the value $\hat{\varepsilon}$ as the average of violation probability $V(\hat{x},\mathbb{P})$ across 1000 cases. In addition, we compute the 95th quantile of the violation probability and also the avarage of objective value as "Ave.Obj.Val" in these 1000 cases.
- In all examples we consider $\alpha = 0.05$ and $\beta = 0.05$ for convenience and alternate between different values of ε and d.
- For each ε and d, we let N_{ext} to be the sample size we need if we can sample from $\mathbb{P}$ directly, and N_{amb} be the Monte Carlo size using our method given the data size n. This number varies under different choices of baseline measure $\mathbb{P}_0$. Note that we must have $n < N_{ext} < N_{amb}$.

4.1 Single Linear Chance Constraints

We consider a single linear CCP

$$\begin{aligned} \min_{x\in\mathscr{X}\subseteq\mathbb{R}^d} \quad & c^T x, \\ \text{subject to} \quad & \mathbb{P}((a+\xi)^T x \le b) \ge 1-\varepsilon, \end{aligned}$$

where $x \in \mathbb{R}^d$ is the decision variable, $a, c \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are constants. The random vector $\xi \in \mathbb{R}^d$ follows $\mathscr{N}(\vartheta, I_d)$ for some unknown ϑ. We suitably choose the parameters so that our problem is always feasible. The true underlying measure has $\vartheta = 0$. We choose a baseline measure $\mathbb{P}_0$ using $\hat{\vartheta}_n$, the MLE estimator. The χ^2-distance in $\mathscr{U}_{data}$ can be explicitly computed in this case. For different levels of ε and d, we compute N_{ext} and set $n < N$. Then we compare the values of N_{ext} and N_{amb} to gain insight on the difference of the computational cost in our method with that of the standard SG. Finally, we report "Ave.Obj.Val" and $\hat{\varepsilon}$ and its 95% quantiles Q_{95} from 1000 experiments. Table 1 shows our results.

Table 1: A Single Linear CCP for Gaussian with unknown mean.

	Different levels of ε and d					
	$\varepsilon = 0.1$ $d = 5$	$\varepsilon = 0.1$ $d = 10$	$\varepsilon = 0.1$ $d = 20$	$\varepsilon = 0.05$ $d = 5$	$\varepsilon = 0.05$ $d = 10$	$\varepsilon = 0.05$ $d = 20$
n	60	100	180	100	200	300
N_{ext}	89	154	275	181	311	554
N_{amb}	342	585	1010	748	1144	2209
$Ave.Obj.Val$	0.8086	0.8686	0.9054	0.8037	0.8566	0.8944
$\hat{\varepsilon}$	0.0094	0.0191	0.0221	0.0071	0.0091	0.0097
Q_{95}	0.0203	0.0341	0.0333	0.0137	0.0149	0.0145

Note that the data size is not enough to support a standard SG in all our considered settings. For comparison, we show the Monte Carlo N_{amb} needed in our method, which are quite big compared to N_{ext}, though one should keep in mind that the Monte Carlo samples are easy to generate. As we can see, the $\hat{\varepsilon}$ from our method are all below the tolerance level, exhibiting the validity of our method.

Next, we explore the case where $\xi \sim exp(\lambda)$ with some unknown λ. We set the true underlying measure $\mathbb{P}$ to have $\lambda = 1$ and we choose the baseline measure $\mathbb{P}_0$ to use $\hat{\lambda}_n$, the MLE estimator. We consider the case where $\varepsilon = 0.01$, $d = 1$ and summarize the results in Table 2. Again, the $\hat{\varepsilon}$ we obtain is well below the tolerance level.

Table 2: A Single Linear CCP for Exponential with unknown mean.

	n	N_{exact}	N_{amb}	$Ave.Obj.Val$	$\hat{\varepsilon}$	Q_{95}
$Value$	100	299	1761	0.3644	0.00036	0.0014

Finally, we consider the case where $\xi \in \mathscr{N}(\vartheta, \Sigma)$ with both parameters unknown. In such cases, we use the Wilks' theorem and construct a joint confidence region for both ϑ and Σ based on the χ^2-distribution (Greene 2008; Arnold and Shavelle 1998). Specifically, We show the results for $d = 1$, $\varepsilon = 0.01$ for demonstration. We set the true underlying measure $\mathbb{P}$ to have $\vartheta = 0$, $\sigma^2 = 1$ and we choose the baseline measure $\mathbb{P}_0$ to use $(\hat{\vartheta}_n, \hat{\sigma}_n)$, the MLE estimator. We summarize the results in Table 3, which show similar patterns as the previous cases.

Table 3: A Single Linear CCP for Gaussian with unknown mean and unknown variance.

	n	N_{exact}	N_{amb}	$Ave.Obj.Val$	$\hat{\varepsilon}$	Q_{95}
Value	200	299	1426	0.5207	0.0011	0.0044

4.2 Joint Linear Chance Constraints

We consider a joint linear CCP

$$\begin{aligned} \min_{x\in\mathscr{X}\subseteq\mathbb{R}^d} \quad & c^T x, \\ \text{subject to} \quad & \mathbb{P}((A+\xi)^T x \le b) \ge 1-\varepsilon, \end{aligned}$$

where $x \in \mathbb{R}^d$ is the decision variable and $c \in \mathbb{R}^d$,$b \in \mathbb{R}^l$ and $A \in \mathbb{R}^{d\times l}$ are constants. The random vector $\xi \in \mathbb{R}^{d\times l}$ satisfies $vec(\xi) \sim \mathscr{N}(0,\Sigma)$ where $\Sigma \in \mathbb{R}^{dl\times dl}$ is some non-identity positive definite matrix. Again, we suitably choose the parameters so that our problem is always feasible. Also, we choose different values of d,l and ε to demonstrate the results. We summarize the results in Table 4.

Table 4: A Joint Linear CCP.

	Different levels of d,l and ε			
	$\varepsilon = 0.1$ $d = 5$ $l = 10$	$\varepsilon = 0.1$ $d = 10$ $d = 15$	$\varepsilon = 0.05$ $d = 5$ $l = 10$	$\varepsilon = 0.05$ $d = 10$ $l = 15$
n	60	100	100	200
N_{ext}	89	154	181	311
N_{amb}	342	585	748	1144
$Ave.Obj.Val$	0.6388	0.6563	0.6387	0.6669
$\hat{\varepsilon}$	0.0017	0.0024	0.0017	0.0044
Q_{95}	0.0041	0.0052	0.0037	0.0091

In this joint CCP case, the number of samples N_{amb} can grow quickly when ε decreases. Our $\hat{\varepsilon}$ are still well below the tolerance level, thus showing the validity of our approach. However, to reduce the computational costs, we should consider choosing a more efficient $\mathbb{P}_0$. We will address this question in the following subsection.

4.3 Different Choices of the Baseline Distribution

We investigate different choices of $\mathbb{P}_0$, motivated from results in the last subsection that using $\mathbb{P}_0$ at the MLE estimator $\hat{\theta}_n$ may lead to a large N_{amb}. For demonstration, we consider the case where the parametric family is Gaussian with unknown mean and unit variance. We propose different choices of the mixture distribution $\mu(d\theta)$, compute $p_0(y) = \int_{\tilde{\mathscr{U}}_{data}} p(y;\theta)\mu d\theta$ and solve directly for $\mathscr{D}_{data}(\mathbb{P}_0)$ to compute δ and N_{amb}. Note that, in the case $\xi \sim \mathscr{N}(\vartheta,1)$ and $\alpha = 0.05$, the uncertainty set $\tilde{\mathscr{U}}_{data}$ is simply the interval $[\hat{\vartheta}_n - \frac{1.96}{\sqrt{n}}, \hat{\vartheta}_n + \frac{1.96}{\sqrt{n}}]$.

We compare 4 types of mixture on $\tilde{\mathscr{U}}_{data}$. We denote μ_1 to be the point mass $\delta_{\hat{\theta}_n}$, μ_2 to be the mixture of two equal point mass at the two points that split the interval $\tilde{\mathscr{U}}_{data}$ into three equal-distance pieces. Then, we set μ_3 to be the uniform distribution on the entire interval $\tilde{\mathscr{U}}_{data}$. Finally, μ_{prop} is the uniform distribution on the boundary of $\tilde{\mathscr{U}}_{data}$, which is just two equal point masses at two end points. We set

$\varepsilon = 0.05$. We summarize the results in Table 5. As, we can see, in this example, μ_{prop} has the best performance as it gives the lowest number of required Monte Carlo samples.

Table 5: A Comparison among four choices of $\mathbb{P}_0$ for single CCP.

	$\mathscr{D}_{data}(\mathbb{P}_0)$	δ	N_{amb}
$\delta_{\hat{\theta}_n}$	0.46837	0.0153	195
μ_2	0.42611	0.0164	182
μ_3	0.36702	0.0182	164
μ_{prop}	0.28765	0.0215	138

We also demonstrate the case of multivariate Gaussian with unknown mean. In this case, we can still explicitly compute the $p_0(y)$ under μ_{prop} using the hypergeometric function defined in Nath (1951). We show the comparisons in Table 6. We again set $\varepsilon = 0.05$, $d = 2$ and $n = 80$. In this case $N_{ext} = 93$. Our choice of μ_{prop} gives a much smaller N_{amb} than using μ set as the point mass at $\hat{\theta}_n$. Moreover, the small value of $\hat{\varepsilon}$ shows that our solution statistically satisfies feasibility for the CCP, which demonstrates that this choice of μ_{prop} is valid for our method.

Table 6: A Comparison among two choices of $\mathbb{P}_0$ for joint CCP.

	$\mathscr{D}_{data}(\mathbb{P}_0)$	δ	N_{amb}	Ave.Obj.Val	$\hat{\varepsilon}$	Q_{95}
$\delta_{\hat{\theta}_n}$	0.0155	0.0153	305	0.9126	0.0016	0.0080
μ_{prop}	0.00543	0.0360	130	0.9211	0.0153	0.0402

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CMMI-1542020 and CAREER CMMI-1653339/1834710.

REFERENCES

Arnold, B. C., and R. M. Shavelle. 1998. "Joint Confidence Sets for the Mean and Variance of a Normal Distribution". *The American Statistician* 52(2):133–140.

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

Blanchet, J., and Y. Kang. 2016. "Sample Out-Of-Sample Inference based on Wasserstein Distance". *arXiv preprint arXiv:1605.01340*.

Calafiore, G. C. 2017. " Repetitive Scenario Design". *IEEE Transactions on Automatic Control* 62(3):1125–1137.

Calafiore, G. C., F. Dabbene, and R. Tempo. 2011. " Research on Probabilistic Methods for Control System Design". *Automatica* 47(7):1279–1293.

Campi, M. C., and A. Carè. 2013. "Random Convex Programs with L_1-Regularization: Sparsity and Generalization". *SIAM Journal on Control and Optimization* 51(5):3532–3557.

Campi, M. C., and S. Garatti. 2008. "The Exact Feasibility of Randomized Solutions of Uncertain Convex Programs". *SIAM Journal on Optimization* 19(3):1211–1230.

Campi, M. C., and S. Garatti. 2018. "Wait-and-Judge Scenario Optimization". *Mathematical Programming* 167(1):155–189.

Carè, A., S. Garatti, and M. C. Campi. 2014. "FAST–Fast Algorithm for the Scenario Technique". *Operations Research* 62(3):662–671.

Chamanbaz, M., F. Dabbene, R. Tempo, V. Venkataramanan, and Q.-G. Wang. 2016. "Sequential Randomized Algorithms for Convex Optimization in the Presence of Uncertainty". *IEEE Transactions on Automatic Control* 61(9):2565–2571.

De Farias, D. P., and B. Van Roy. 2004. "On Constraint Sampling in the Linear Programming Approach to Approximate Dynamic Programming". *Mathematics of Operations Research* 29(3):462–478.

Delage, E., and Y. Ye. 2010. "Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems". *Operations Research* 58(3):595–612.

Erdoğan, E., and G. Iyengar. 2006. "Ambiguous Chance Constrained Problems and Robust Optimization". *Mathematical Programming* 107(1-2):37–61.

Esfahani, P. M., and D. Kuhn. 2015. "Data-Driven Distributionally Robust Optimization using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations". *Mathematical Programming*:1–52.

Gao, R., and A. J. Kleywegt. 2016. "Distributionally Robust Stochastic Optimization with Wasserstein Distance". *arXiv preprint arXiv:1604.02199*.

Ghosh, S., and H. Lam. 2018. "Robust Analysis in Stochastic Simulation: Computation and Performance Guarantees". *To appear in Operations Research, available as arxiv preprint at arXiv:1507.05609*.

Glasserman, P., and X. Xu. 2014. "Robust Risk Measurement and Model Risk". *Quantitative Finance* 14(1):29–58.

Greene, W. 2008. *Econometric Analysis*. 6th ed. Upper Saddle River, New Jersey: Pearson/Prentice Hall.

Hanasusanto, G. A., V. Roitch, D. Kuhn, and W. Wiesemann. 2015. "A Distributionally Robust Perspective on Uncertainty Quantification and Chance Constrained Programming". *Mathematical Programming* 151(1):35–62.

Hanasusanto, G. A., V. Roitch, D. Kuhn, and W. Wiesemann. 2017. "Ambiguous Joint Chance Constraints under Mean and Dispersion Information". *Operations Research* 65(3):751–767.

Hong, L. J., J. Luo, and B. L. Nelson. 2015. "Chance Constrained Selection of the Best". *INFORMS Journal on Computing* 27(2):317–334.

Hu, Z., J. Cao, and L. J. Hong. 2012. "Robust Simulation of Global Warming Policies using the DICE Model". *Management Science* 58(12):2190–2206.

Hu, Z., and L. J. Hong. 2015. "Robust Simulation of Stochastic Systems with Input Uncertainties Modeled by Statistical Divergences". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz et al., 643–654. Piscataway, New Jersey: IEEE.

Jiang, R., and Y. Guan. 2016. "Data-Driven Chance Constrained Stochastic Program". *Mathematical Programming* 158(1-2):291–327.

Korostelev, A. P., and O. Korosteleva. 2011. *Mathematical Statistics: Asymptotic Minimax Theory*. Providence, Rhode Island: American Mathematical Society.

Lam, H. 2016. "Robust Sensitivity Analysis for Stochastic Systems". *Mathematics of Operations Research* 41(4):1248–1275.

Lam, H. 2018. "Sensitivity to Serial Dependency of Input Processes: A Robust Approach". *Management Science* 64(3):1311–1327.

Lejeune, M. A., and A. Ruszczynski. 2007. "An Efficient Trajectory Method for Probabilistic Production-Inventory-Distribution Problems". *Operations Research* 55(2):378–394.

Li, B., R. Jiang, and J. L. Mathieu. "Ambiguous Risk Constraints with Moment and Unimodality Information". *Mathematical Programming*:1–42.

Lim, A. E., J. G. Shanthikumar, and Z. M. Shen. 2006. "Model Uncertainty, Robust Optimization, and Learning". In *Tutorials in Operations Research: Models, Methods, and Applications for Innovative Decision Making*, edited by M. P. Johnson et al., 66–94. Catonsville, Maryland: INFORMS.

Love, D., and G. Bayraksan. 2015. "Phi-Divergence Constrained Ambiguous Stochastic Programs for Data-Driven Optimization". Technical report, Department of Integrated Systems Engineering, The Ohio State University, Columbus, Ohio.

Luedtke, J., and S. Ahmed. 2008. "A Sample Approximation Approach for Optimization with Probabilistic Constraints". *SIAM Journal on Optimization* 19(2):674–699.

Muller, M. E. 1959. "A Note on a Method for Generating Points Uniformly on N-Dimensional Spheres". *Communications of the ACM* 2(4):19–20.

Murr, M. R., and A. Prékopa. 2000. "Solution of a Product Substitution Problem Using Stochastic Programming". In *Probabilistic Constrained Optimization*, edited by U. Stanislav, 252–271. Manhattan, New York: Springer.

Nath, P. 1951. "Confluent Hypergeometric Function". *Sankhyā: The Indian Journal of Statistics*:153–166.

Nielsen, F., and R. Nock. 2014. "On the Chi Square and Higher-Order Chi Distances for Approximating f-Divergences". *IEEE Signal Processing Letters* 21(1):10–13.

Pardo, L. 2005. *Statistical Inference Based on Divergence Measures*. New York: Chapman and Hall/CRC.

Petersen, I. R., M. R. James, and P. Dupuis. 2000. "Minimax Optimal Control of Stochastic Uncertain Systems with Relative Entropy Constraints". *IEEE Transactions on Automatic Control* 45(3):398–412.

Prékopa, A. 2003. "Probabilistic Programming". In *Handbooks in Operations Research and Management Science, Volume 10: Stochastic Programming*, edited by A. Ruszczynski and A. Shapiro. Amsterdam, Netherlands: Elsevier.

Prékopa, A., T. Rapcsák, and I. Zsuffa. 1978. "Serially Linked Reservoir System Design Using Stochastic Programing". *Water Resources Research* 14(4):672–678.

Prékopa, A., and T. Szántai. 1978. "Flood Control Reservoir System Design Using Stochastic Programming". In *Mathematical Programming in Use*, edited by M. Balinski and C. Lemarechal, 138–151. Manhattan, New York: Springer.

Schildbach, G., L. Fagiano, and M. Morari. 2013. "Randomized Solutions to Convex Programs with Multiple Chance Constraints". *SIAM Journal on Optimization* 23(4):2479–2501.

Shi, Y., J. Zhang, and K. B. Letaief. 2015. "Optimal Stochastic Coordinated Beamforming for Wireless Cooperative Networks with CSI Uncertainty". *IEEE Transactions on Signal Processing* 63(4):960–973.

Tseng, S.-H., E. Bitar, and A. Tang. 2016. "Random Convex Approximations of Ambiguous Chance Constrained Programs". In *2016 IEEE 55th Conference on Decision and Control*, edited by A. Giua et al., 6210–6215. Piscataway, New Jersey: IEEE.

Vajda, I. 1972. "On the f-Divergence and Singularity of Probability Measures". *Periodica Mathematica Hungarica* 2(1-4):223–234.

Wiesemann, W., D. Kuhn, and M. Sim. 2014. "Distributionally Robust Convex Optimization". *Operations Research* 62(6):1358–1376.

Zhang, Y., R. Jiang, and S. Shen. 2016. "Ambiguous Chance-Constrained Bin Packing under Mean-Covariance Information". *arXiv preprint arXiv:1610.00035*.

Zymler, S., D. Kuhn, and B. Rustem. 2013. "Distributionally Robust Joint Chance Constraints with Second-Order Moment Information". *Mathematical Programming*:1–32.

AUTHOR BIOGRAPHIES

HENRY LAM is an Associate Professor in the Department of IEOR at Columbia University. He received his Ph.D. degree in statistics from Harvard University in 2011, and was on the faculty of Boston University and University of Michigan before joining Columbia in 2017. His research focuses on Monte Carlo simulation, risk and uncertainty quantification, and stochastic optimization. His email address is khl2114@columbia.edu.

FENGPEI LI is a Ph.D. candidate in the Department of IEOR at Columbia University. He receives his B.S. degree in mathematics from UCSD. His research focuses on Monte Carlo simulation, stochastic modeling, stochastic optimization and robust optimization. His email address is fl2412@columbia.edu.