Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

QUANTIFYING UNCERTAINTY IN SAMPLE AVERAGE APPROXIMATION

Henry Lam

Department of Industrial & Operations Engineering
University of Michigan
1205 Beal Ave.
Ann Arbor, MI 48109, USA

Enlu Zhou

School of Industrial & Systems Engineering
Georgia Institute of Technology
755 Ferst Drive NW
Atlanta, GA 30332, USA

ABSTRACT

We consider stochastic optimization problems in which the input probability distribution is not fully known, and can only be observed through data. Common procedures handle such problems by optimizing an empirical counterpart, namely via using an empirical distribution of the input. The optimal solutions obtained through such procedures are hence subject to uncertainty of the data. In this paper, we explore techniques to quantify this uncertainty that have potentially good finite-sample performance. We consider three approaches: the empirical likelihood method, nonparametric Bayesian approach, and the bootstrap approach. They are designed to approximate the confidence intervals or posterior distributions of the optimal values or the optimality gaps. We present computational procedures for each of the approaches and discuss their relative benefits. A numerical example on conditional value-at-risk is used to demonstrate these methods.

1 INTRODUCTION

We are interested in stochastic optimization problems in the form

$$\max_{\theta\in\Theta} E[h(X;\theta)], \tag{1}$$

where $\theta \in \Theta \in \mathbb{R}^p$ is the decision variable and $X \in \mathbb{R}^d$ is a random variable. In many applications, the underlying probability distribution that controls the expectation $E[\cdot]$ is not fully known and can only be accessed via limited data or Monte Carlo samples. Then it is customary to work on an empirical counterpart of the problem, namely by solving

$$\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} h(X_i;\theta), \tag{2}$$

where $X_1,\ldots,X_n$ are the data. This is well known as sample average approximation (SAA) (Shapiro, Dentcheva, and Ruszczyński 2014) in the stochastic programming literature. Obviously, (2) is subject to approximation errors with respect to (1), and this error is the main focus of this paper.

Our premise is that beyond the n observed samples, new samples are not easily accessible, either because of lack of data or because of limited computational capacity in running further Monte Carlo simulation. In this setup, previous results on quantifying the approximation errors have fallen in two directions. The first is central limit convergence of the optimality gap, i.e. difference between (1) and (2), typically to some Gaussian random variable under suitable scaling (Shapiro, Dentcheva, and Ruszczyński 2014; Kim, Pasupathy, and Henderson 2015). Such results can be used to construct confidence interval (CI) that has asymptotically correct coverage probability as sample size increases. The second approach consists of sampling complexity results, in the form of concentration inequalities on the optimality gap. For instance, Shapiro and Nemirovski (2005) present large deviations inequalities that depend on certain size measures of Θ and the Lipschitz constant of h. In classification-type learning problems, where h is often taken as an indicator function, deviation bounds have been obtained based on the entropy or the so-called Vapnik-Chervonenkis dimension of h (e.g. Mendelson 2003).

The above two approaches are both valuable in evaluating the reliability of an adopted solution or choosing a good sample size n in advance. However, they have antagonistic limitations: Central limit convergence typically only works well for the large-sample case and has no guarantee for smaller sample size, while sampling complexity results are typically loose because of their worst-case nature, even though they hold for any sample size. Driven by these limitations, we are interested in exploring techniques that can potentially yield improvement on both ends, by giving tight quantification of approximation errors even in situations of limited samples.

We consider three distinct approaches: the empirical likelihood (EL) method, the Dirichlet method, and nonparametric bootstrapping. These approaches differ in statistical philosophy and have distint benefits and disadvantages in terms of both the uncertainty assessment scope and computational load. To give some initial highlight, the EL method can be viewed as a nonparametric analog of classical likelihood inference, and can generate CI for the true optimal value or the optimality gap in the frequentist sense. On the other hand, the Dirichlet method is utilized from a Bayesian perspective to approximate a corresponding posterior distribution, whereas nonparametric bootstrapping can be viewed as a limit of the Dirichlet method by taking suitable asymptotic on the prior distribution.

To demonstrate our proposed approaches, we consider conditional value-at-risk (CVaR) as a specific example of (1). CVaR is widely used in portfolio risk management in finance and insurance, among other applications (Rockafellar and Uryasev 2000). The distributions of these portfolios are unknown and can only be represented via data in any real-life contexts. We shall provide some numerics on quantifying the uncertainty of the empirical optimization of CVaR.

Our paper is organized as follows. In Section 2, we first present the three approaches for quantifying uncertainty of the optimal value output by (2). We then adapt these approaches for optimality gap in Section 3. We discuss some comparisons among the approaches in Section 4, followed by some numerical results on CVaR in Section 5.

2 METHODS FOR QUANTIFYING UNCERTAINTY OF THE OPTIMAL VALUE

2.1 Empirical Likelihood Method

The EL method is a nonparametric analog of maximum likelihood estimation first proposed by Owen (1988). To introduce the method, let us first fix some notation. Given the set of data $X_1, X_2, \ldots, X_n$, we define a probability simplex over $\{X_1, \ldots, X_n\}$, denoted $\mathbf{w} = (w_1, \ldots, w_n)$ where $\sum_{i=1}^n w_i = 1$ and $w_i \geq 0$ for all i. We denote $\chi^2_{q,\beta}$ as the $1-\beta$ quantile of a χ^2 distribution with degree of freedom q. The EL method entails that the optimizations

$$\begin{array}{ll} \max/\min_{\mathbf{w}} & \max_{\theta\in\Theta}\sum_{i=1}^n w_i h(X_i;\theta) \\ \text{subject to} & -2\sum_{i=1}^n \log(nw_i) \leq \chi^2_{p+1,\beta} \\ & \sum_{i=1}^n w_i = 1 \\ & w_i \geq 0 \text{ for all } i \end{array} \tag{3}$$

where $\max/\min$ denote a pair of maximization and minimization, contain the true value of $\max_{\theta\in\Theta} E[h(X;\theta)]$ with probability at least $1-\beta$ asymptotically.

To get some intuition, the quantity $-\sum_{i=1}^n \log(nw_i)$ can be interpreted as a statistical distance between two probability measures P and Q on the support $\{X_1, \ldots, X_n\}$, defined by the weights $\mathbf{w}$ and the uniform weights $(1/n, \ldots, 1/n)$ respectively. In fact, this distance belongs to the class of so-called ϕ-divergence, where the ϕ function can be identified as $\phi(x) = -\log x$ here (Pardo 2005). The optimizations in (3) can therefore be viewed from a distributionally robust optimization perspective: when the underlying probability distribution of a stochastic problem is not fully known, one can impose an uncertainty set in which the probability distribution is believed to lie, and calculate the worst-case scenario among all distributions within the set. In our case, (3) is computing the worst-case scenarios as the max or min for the quantity $\max_{\theta\in\Theta} E[h(X;\theta)]$, where the uncertainty set can be thought as the neighborhood around the uniform distribution on $\{X_1,\ldots,X_n\}$ measured by a ϕ-divergence with $\phi(x) = -\log x$.

2.1.1 Computational Load

We shall discuss how to compute (3), which consists of a min-max problem and a max-max problem. Note that the outer objective function $\max_{\theta\in\Theta}\sum_{i=1}^n w_i h(X_i;\theta)$ is a convex function in $\mathbf{w}$. Suppose that $h(X;\cdot)$ is concave in θ. Then evaluating $\max_{\theta\in\Theta}\sum_{i=1}^n w_i h(X_i;\theta)$ for fixed $\mathbf{w}$ boils down to a convex optimization, and hence both the outer and inner optimizations in the min-max problem are convex and efficiently solvable. Moreover, to speed up computation, by observing that the feasible region for $\mathbf{w}$ is compact one can consider an exchange of min-max to max-min via Sion's minimax theorem (Sion et al. 1958), which gives

$$\min_{\mathbf{w}\in\mathscr{A}}\max_{\theta\in\Theta}\sum_{i=1}^n w_i h(X_i;\theta) = \max_{\theta\in\Theta}\min_{\mathbf{w}\in\mathscr{A}}\sum_{i=1}^n w_i h(X_i;\theta) \tag{4}$$

where $\mathscr{A}$ denotes the feasible region in (3). Fixing θ, the inner minimization $\min_{\mathbf{w}\in\mathscr{A}}\sum_{i=1}^n w_i h(X_i;\theta)$ in the maximin formulation can be solved by Lagrangian relaxation

$$\begin{aligned} &\max_{\nu\geq 0,\lambda\in\mathbb{R}}\min_{\mathbf{w}\geq\mathbf{0}}\sum_{i=1}^n w_i h(X_i;\theta) + \nu\left(-2\sum_{i=1}^n \log(nw_i) - \chi^2_{p+1,\beta}\right) - \lambda\left(\sum_{i=1}^n w_i - 1\right) \\ = \quad &\max_{\nu\geq 0,\lambda\in\mathbb{R}}\sum_{i=1}^n \min_{w_i\geq 0}\{w_i(h(X_i;\theta)-\lambda) - 2\nu\log(nw_i)\} - 2\nu n\log n - \nu\chi^2_{p+1,\beta} + \lambda. \end{aligned} \tag{5}$$

Straightforward calculation reveals that

$$\min_{w_i\geq 0}\{w_i(h(X_i;\theta)-\lambda) - 2\nu\log(nw_i)\} = \begin{cases} 2\nu\left(1-\log\left(\frac{2\nu}{h(X_i;\theta)-\lambda}\right)\right) & \text{if } h(X_i;\theta)>\lambda \\ -\infty & \text{otherwise} \end{cases}$$

with the optimal solution of w_i being $2\nu/(h(X_i;\theta)-\lambda)$ or ∞ correspondingly. This implies that (5) is equal to

$$\max_{\nu\geq 0,\ \lambda<\min_i h(X_i;\theta)} -\sum_{i=1}^n 2\nu\log\left(\frac{2\nu}{h(X_i;\theta)-\lambda}\right) + 2\nu n(1-\log n) - \nu\chi^2_{p+1,\beta} + \lambda$$

where the optimal solution of w_i is $2\nu/(h(X_i;\theta)-\lambda)$. Hence the inner minimization of the right hand side of (4) becomes a convex maximization problem with two variables (instead of having n variables in the original formulation).

On the other hand, the max-max problem in (3) can be more challenging, because the outer optimization involves maximizing the convex function $\max_{\theta\in\Theta}\sum_{i=1}^n w_i h(X_i;\theta)$ over $\mathbf{w}$. A quick heuristic for obtaining a solution for the max-max problem is to do alternating maximization, namely repeatedly fixing $\mathbf{w}$ and maximizing over θ, and fixing θ and maximizing over $\mathbf{w}$, until no improvement is observed. Csiszar and Tusnady (1984) contains global convergence results for this sort of procedure under some improvement bound assumptions uniformly on the alternating steps. Similar as above, when θ is fixed, one can write $\max_{\mathbf{w}\in\mathscr{A}}\sum_{i=1}^n w_i h(X_i;\theta)$ via Lagrangian relaxation as

$$\min_{\nu\geq 0,\ \lambda>\max_i h(X_i;\theta)} 2\nu\sum_{i=1}^n \log\left(\frac{2\nu}{\lambda-h(X_i;\theta)}\right) + 2\nu n(\log n - 1) + \nu\chi^2_{p+1,\beta} + \lambda$$

with the optimal solution of w_i being $2\nu/(\lambda - h(X_i;\theta))$.

2.1.2 Statistical Guarantees

The statistical guarantee for (3) and the choice of the "neighborhood size" $\chi^2_{p+1,\beta}$ can be made rigorous through the theory of the EL method. The key of the method is an analog of the celebrated Wilks' Theorem (Cox and Hinkley 1979) in parametric likelihood inference, namely that the ratio between the maximum likelihood and the true likelihood, i.e. the so-called likelihood ratio, converges to a chi-square distribution in a suitable logarithmic scale.

To apply the EL method on the data set $\{X_1,\ldots,X_n\}$, one would first define an empirical likelihood $\prod_{i=1}^n w_i$. Treating $\mathbf{w}$ as the "parameters", it is not difficult to see that the maximum value of $\prod_{i=1}^n w_i$, among all $\mathbf{w}$ in the probability simplex, is $\prod_{i=1}^n (1/n)$, which can be viewed as a nonparametric analog of maximum likelihood. The next step is to define a target parameter of interest, i.e. the quantity whose statistical uncertainty is to be assessed. In our case this is $\max_{\theta\in\Theta} E[h(X;\theta)]$. The EL method bridges to Wilks' Theorem via a quantity known as the profile likelihood, defined as the maximum ratio between the empirical likelihood and the nonparametric maximum likelihood, optimized over all probability weights with support over the data set and satisfying some given moment conditions incorporating the target parameter of interest. The crux is that profile likelihood satisfies similar asymptotic properties as the likelihood ratio in the parametric context.

We state the above mathematically. We first make the following assumptions:

Assumption 1

1. $E[h(X;\theta)]$ is differentiable in θ with $\nabla_\theta E[h(X;\theta)] = E[\nabla_\theta h(X;\theta)]$ for $\theta \in \Theta$.
2. $\nabla_\theta E[h(X;\theta_0)] = 0$ if and only if $\theta_0 \in \operatorname{argmax}_{\theta\in\Theta} E[h(X;\theta)]$, and $\sum_{i=1}^n w_i \nabla_\theta h(X_i;\theta_0) = 0$ if and only if $\theta_0 \in \operatorname{argmax}_{\theta\in\Theta} \sum_{i=1}^n w_i h(X_i;\theta)$ for any X_i and $\mathbf{w} = (w_1,\ldots,w_n)$ is a probability vector.
3. There exists a $\theta_0 \in \operatorname{argmax}_{\theta\in\Theta} E[h(X;\theta)]$ such that the covariance matrix of the random vector $(\nabla_\theta h(X;\theta_0), h(X;\theta_0))$ has rank $p+1>0$.

Note that there is no assumption on the uniqueness of θ_0. Condition 2 is a first order condition that is satisfied if, for instance, $h(X;\cdot)$ is a coersive concave function for any X a.s. and $\Theta = \mathbb{R}^p$. Condition 3 states that all partial derivatives of $h(X;\theta_0)$ and also $h(X;\theta_0)$ itself are linearly independent. Then we have:

Theorem 1 Let the maximum and minimum values of the programs (3) be $\overline{Z}$ and $\underline{Z}$ respectively. Then, under Assumption 1, we have

$$\liminf_{n\to\infty} P\left(\max_{\theta\in\Theta} E[h(X;\theta)] \in [\underline{Z},\overline{Z}]\right) \geq 1-\beta.$$

We shall prove Theorem 1. We denote θ_0 as an element in $\operatorname{argmax}_{\theta\in\Theta} E[h(X;\theta)]$ such that the covariance matrix of the random vector $(\nabla_\theta h(X;\theta_0), h(X;\theta_0))$ has rank $p+1>0$. The existence of such θ_0 is guaranteed in Condition 3 in Assumption 1. By Conditions 1 and 2 in Assumption 1, θ_0 satisfies $E[\nabla_\theta h(X;\theta_0)] = \mathbf{0}$, where $\mathbf{0}$ is a length p vector of zeros.

We define a profile likelihood as

$$\mathscr{R}(\theta,z) = \max\left\{\prod_{i=1}^n nw_i : \sum_{i=1}^n w_i \nabla_\theta h(X_i;\theta) = \mathbf{0},\ \sum_{i=1}^n w_i h(X_i;\theta) = z,\ \sum_{i=1}^n w_i = 1,\ w_i \geq 0 \text{ for all } i\right\}. \tag{6}$$

This is the maximum ratio between the empirical likelihood and the nonparametric maximum likelihood, among all probability weights that respect the condition $E[\nabla_\theta h(X;\theta_0)] = \mathbf{0}$ and $E[h(X;\theta_0)] = z$. The parameter z is our target. We let z_0 be the true optimal value of $\max_{\theta\in\Theta} E[h(X;\theta)]$, which is equal to $E[h(X;\theta_0)]$. The following result is an immediate application from Owen (2001):

Theorem 2 Under Assumption 1, the profile likelihood (6) satisfies $-2\log\mathscr{R}(\theta_0,z_0) \Rightarrow \chi^2_{p+1}$ as $n\to\infty$, where χ^2_q denotes the χ^2-distribution with degree of freedom q.

In other words, we know that $P(-2\log\mathscr{R}(\theta_0,z_0)\leq\chi^2_{q,\beta})\to 1-\beta$. So the set $\mathscr{C}=\{(\theta,z):-2\log\mathscr{R}(\theta,z)\leq\chi^2_{p+1,\beta}\}$ forms an asymptotically valid $1-\beta$ level confidence region for (θ_0,z_0), i.e. $P((\theta_0,z_0)\in\mathscr{C})\to 1-\beta$. This implies that with probability $1-\beta$ asymptotically, there exists a probability vector $\mathbf{w}$ such that $\sum_{i=1}^n w_i\nabla_\theta h(X_i;\theta_0)=\mathbf{0}$, $\sum_{i=1}^n w_i h(X_i;\theta_0)=z_0$ and $-2\sum_{i=1}^n\log(nw_i)\leq\chi^2_{p+1,\beta}$. This further implies that with probability $1-\beta$ asymptotically, z_0 is contained by the pair of optimizations

$$\begin{array}{ll}\max/\min_{\mathbf{w}} & \sum_{i=1}^n w_i h(X_i;\theta_0)\\ \text{subject to} & \sum_{i=1}^n w_i\nabla_\theta h(X_i;\theta_0)=\mathbf{0}\\ & -2\sum_{i=1}^n\log(nw_i)\leq\chi^2_{p+1,\beta}\\ & \sum_{i=1}^n w_i=1\\ & w_i\geq 0 \text{ for all } i\end{array}$$

or equivalently

$$\begin{array}{ll}\max/\min_{\mathbf{w}} & \max_{\theta\in\Theta}\sum_{i=1}^n w_i h(X_i;\theta)\\ \text{subject to} & \theta_0 \text{ is an optimal solution for } \max_{\theta\in\Theta}\sum_{i=1}^n w_i h(X_i;\theta)\\ & -2\sum_{i=1}^n\log(nw_i)\leq\chi^2_{p+1,\beta}\\ & \sum_{i=1}^n w_i=1\\ & w_i\geq 0 \text{ for all } i\end{array}\tag{7}$$

by using Condition 2 in Assumption 1. By relaxing the first constraint in (7), we conclude that with probability at least $1-\beta$ asymptotically, z_0 is contained in (3), which proves Theorem 1.

Theorem 2 is an asymptotic statement. For small samples, accuracy (measured in terms of the empirical coverage probability being correctly larger than $1-\beta$) can be enhanced by other calibration methods than using $\chi^2_{p+1,\beta}$ in the optimization (3). Chapters 2 and 13 in Owen (2001) provide some discussion.

2.2 Dirichlet Methods

The empirical likelihood method is a frequentist approach based on asymptotics. In contrast, the Bayesian approach takes a different perspective by characterizing the likelihood of possible alternatives based on the available data and the chosen prior. It encodes all the information in the posterior distribution and does not rely on asymptotics. In this section, we develop a Bayesian, non-parametric method to quantify the uncertainty in the optimal value of (1). To have a tractable posterior of distribution over a general support, we will use the Dirichlet process, which was first introduced by Ferguson (1973) and is currently one of the most popular Bayesian nonparametric models. To ease the explanation, we will first illustrate the approach on the simple case when the input distribution has a finite support.

2.2.1 Input Distribution with Finite Support

Suppose the true input distribution of X lives on a finite support $\{x_1,\ldots,x_n\}$. The data we have about the input distribution are N_i's — the number of times that we observe each x_i. Please note that the finite support is not merely a simplification; such scenario does happen, for example, when we want to estimate the distribution of the customer demand for a certain product based on observations of sales (assuming no lost sales), where the sales quantity is an integer ranging from 0 to the stock-out level. We denote by $F=[p_1,\ldots,p_n]$ an empirical distribution on the support $\{x_1,\ldots,x_n\}$. The Bayesian approach views F as a random variable, and yields a belief (posterior distribution) of F based on a chosen prior and the likelihood of observing the data $\psi\triangleq[N_1,\ldots,N_n]$:

$$l(\psi|F)=\Pi_{i=1}^n p_i^{N_i}.$$

If we have no prior knowledge of the true distribution, we can assume a uniform prior, which is equivalent to a Dirichlet distribution with the parameter being a unit vector, denoted as $Dir(e)$ where e is the unit vector. We can also use an informative prior by choosing an appropriate parameter value in the Dirichlet prior. Under the noninformative prior, the posterior distribution is the Dirichlet distribution with parameter $e+\psi$, i.e., $Dir(e+\psi)$ with the probability density function

$$f(p;e+\psi)=\frac{1}{B(e+\psi)}\Pi_{i=1}^{n}p_i^{N_i},$$

where the normalization constant $B(\cdot)$ is the multinomial Beta function. Hence, the Dirichlet distribution $Dir(e+\psi)$ is essentially a distribution of the distribution F given the data ψ.

The uncertainty in the optimal value of (1) can be characterized by the posterior distribution on the optimal value given the data, i.e., $P(\max_{\theta\in\Theta}E[h(X;\theta)]|\psi)$. Hence, we need to propagate the posterior on the distribution of X to the optimal function value, and that can be done by the following two steps:

1. Sampling: draw $F^j=[p_1^j,\ldots,p_n^j], j=1,\ldots,M$ from $Dir(e+\psi)$.
2. Optimization: for each $j=1,\ldots,M$, solve the optimization problem

$$\max_{\theta\in\Theta}\left\{E_{F^j}[h(X;\theta)]=\sum_{i=1}^{n}p_i^j h(x_i;\theta)\right\},$$

and denote the optimal value by h^{*j}.

The procedure above will output M sample optimal values $\{h^{*1},\ldots,h^{*M}\}$, which form an empirical distribution that approximates the posterior distribution of the optimal value h^* of (1). From this empirical distribution further information can be extracted. Specifically, we can sort the sample optimal values from the smallest to the largest such that $h^{*(1)}\leq\ldots\leq h^{*(M)}$, and take $[h^{*(\lceil M\frac{\beta}{2}\rceil)},h^{*(\lceil M(1-\frac{\beta}{2})\rceil)}]$ as an estimate of the $(1-\beta)$ Bayesian confidence interval (also called "credible interval") for the optimal value h^*. The procedure above can be made more efficient by first evaluating and storing the function values on the support points, i.e., $h(x_1;\theta),\ldots,h(x_n;\theta)$, and then simply taking a linear combination of these stored values to get the objective function $E_{F^j}[h(X;\theta)]$ according to the probabilities of each generated F^j.

The following theorem shows that the estimated credible interval $[h^{*(\lceil M\frac{\beta}{2}\rceil)},h^{*(\lceil M(1-\frac{\beta}{2})\rceil)}]$ is asymptotically the true credible interval as the sample size M goes to infinity. First note that the posterior distribution of the optimal value h^* given data ψ is

$$P_{h^*}(t|\psi)\triangleq Pr\{\max_{\theta\in\Theta}E[h(X;\theta)]\leq t|\psi\}.$$

Hence, the true γ-quantile of this posterior distribution is $h_\gamma^*(\psi)=\inf\{h:P_{h*}(t|\psi)\geq\gamma\}$. Then we have the following theorem.

Theorem 3 Assume $P_{h^*}(t|\psi)$ is a continuous distribution. Given the data ψ,
(i) as $M\to\infty$, the empirical distribution formed by $\{h^{*1},\ldots,h^{*M}\}$ provides a uniformly consistent estimator of the posterior distribution of h^*, i.e., $P_{h^*}(t|\psi)$;
(ii) for any $\gamma\in(0,1),\lim_{M\to\infty}h^{*(\lceil M\gamma\rceil)}=h_\gamma^*(\psi)$ almost surely.

The proof is similar to that of Theorem 1 in Xie, Nelson, and Barton (2015). Note that $\{h^{*1},\ldots,h^{*M}\}$ is an i.i.d. sample from the posterior distribution $P_{h^*}(t|\psi)$. By the Glivenko-Cantelli theorem, the empirical distribution formed by $\{h^{*1},\ldots,h^{*M}\}$ converges uniformly to $P_{h^*}(t|\psi)$ almost surely. Since $P_{h^*}(t|\psi)$ is continuous, according to Lemma 21.2 in Vaart (1998), as $M\to\infty$ the quantile estimate $h^{*(\lceil M\gamma\rceil)}$ converges to the true quantile $h_\gamma^*(\psi)$ almost surely. By setting γ to be $\beta/2$ and $(1-\beta)/2$, we conclude that $[h^{*(\lceil M\frac{\beta}{2}\rceil)},h^{*(\lceil M(1-\frac{\beta}{2})\rceil)}]$ is a consistent estimator of the credible interval on h^*.

2.2.2 General Input Distributions

The approach above can be generalized to the case when we have a general input distribution. It follows essentially the same idea but uses the Dirichlet process instead of the Dirichlet distribution to address the general support. Similar to a Dirichlet distribution, a Dirichlet process can be viewed as a distribution over distributions, i.e., each sample from a Dirichlet process is a distribution. To facilitate understanding, we will first give a brief overview of the Dirichlet process (more introductions on Dirichlet process can be found in, for example, Teh 2010, Ghosal 2010). Specifically, let F_0 be a distribution over the probability space Θ and a be a positive number. Then G is Dirichlet process distributed with base distribution F_0 and concentration parameter a, denoted as $G \sim DP(a, F_0)$, if

$$(G(A_1),\dots,G(A_r)) \sim Dir(aF_0(A_1),\dots,aF_0(A_r)),$$

for every finite measurable partition $\{A_1,\dots,A_r\}$ of Θ. Simply put, a Dirichlet process is a stochastic process that has Dirichlet distributed finite-dimensional marginal distributions, just as a Gaussian process has Gaussian distributed finite-dimensional marginal distributions.

Just like the Dirichlet distribution is a conjugate prior, the Dirichlet process is also conjugate for estimating an unknown distribution based on i.i.d. observations. Let $X_1,\dots,X_n$ be i.i.d. observations from the unknown input distribution F^c. Denote the empirical distribution by $\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le t\}$. If we choose the prior over F as $DP(a, F_0)$, then posterior of F given the data $\psi = \{X_1,\dots,X_n\}$ is

$$F|\psi \sim DP(a+n, \frac{a}{a+n}F_0 + \frac{n}{a+n}\hat{F}).$$

Note that the base measure of the posterior is a weighted average of the prior base F_0 and the empirical distribution $\hat{F}$, and the weight associated with the prior base is proportional to a. Hence, it is easy to see the role of the parameters F_0 and a in the Bayesian sense: the base measure F_0 is the mean of the prior, and the concentration parameter a is the strength of the prior. As $a \to 0$, the posterior becomes the empirical distribution and it boils down to a non-informative prior. On the other hand, as n, the number of data points, increases, the posterior is dominated by the empirical distribution, which in turn becomes a close approximation of the true distribution. The Dirichlet process is weakly consistent at the true distribution F^c with a convergence rate $1/\sqrt{n}$; or in other words, as the number of observations/data, n, goes to infinity, the Dirichlet process converges to the true distribution F^c in probability (Lo 1983, Ghosal 2010). This justifies the use of Dirichlet process as the posterior distribution to estimate the unknown true distribution.

We are now ready to present our approach to characterizing the posterior distribution of the optimal value of (1) given the data. Conceptually that can be done by the following two steps:

1. Sampling: first, draw sample distributions $F^1,\dots,F^M$ from $DP(a+n, \frac{a}{a+n}F_0 + \frac{n}{a+n}\hat{F})$; then, draw i.i.d. samples $Y_i^j,\dots,Y_N^j$ from each F^j.
2. Optimization: for each $j = 1,\dots,M$, solve the optimization problem

$$\max_{\theta\in\Theta}\left\{E_{F^j}[h(X;\theta)] \approx \frac{1}{N}\sum_{i=1}^{N} h(Y_i^j;\theta)\right\},$$

   and denote the optimal value as h^{*j}.

The procedure above outputs the sample optimal values $\{h^{*j}\}$, which form an empirical distribution that approximates the posterior distribution of h^* given the data. In particular, $[h^{*(\lceil M\frac{\beta}{2}\rceil)}, h^{*(\lceil M(1-\frac{\beta}{2})\rceil)}]$, where $h^{*(1)} \le \dots \le h^{*(M)}$, is an estimate of the $(1-\beta)$ credible interval for the optimal value. Similar as Theorem 3, under the assumption that the posterior distribution of h^* given data is a continuous distribution, we can show that $[h^{*(\lceil M\frac{\beta}{2}\rceil)}, h^{*(\lceil M(1-\frac{\beta}{2})\rceil)}]$ converges to the true credible interval almost surely as $M \to \infty$ and $N \to \infty$.

The difficulty with the conceptual procedure above is in the first step, i.e., sampling from the Dirichlet process. It is impossible to draw a distribution F^j directly from the Dirichlet process, since a full distribution requires storing an infinite amount of information. Therefore, we resort to sampling strategies that circumvents drawing distributions from the Dirichlet process but rather draws samples that approximates the distribution F^j. Here we describe two such sampling methods. The first, Polya urn scheme (Teh 2010), treats each F^j as a latent distribution and directly draws i.i.d. samples $\{Y_i^j\}$ from F^j. It is described below (the superscript j is dropped for convenience):

- Draw Y_1 from $\frac{a}{a+n}F_0 + \frac{n}{a+n}\hat{F}$.
- For $i = 2,\ldots,N$, with probability $\frac{a+n}{a+n+i-1}$ draw Y_i from $\frac{a}{a+n}F_0 + \frac{n}{a+n}\hat{F}$; with probability $\frac{n_y}{a+n+i-1}$ set $Y_i = y$, where n_y is the number of previous samples that take value y.

As N goes to infinity, the Polya urn scheme ensures that $\{Y_i\}$ becomes an exact draw from the Dirichlet process. An alternative and simpler sampling method is to draw $\{Y_1,\ldots,Y_N\}$ directly from the distribution $\frac{a}{a+n}F_0 + \frac{n}{a+n}\hat{F}$, which is equal to $E[F(t)|\psi]$ and thus a natural Bayesian estimate for $F(t)$. This is essentially bootstrap sampling from the weighted average of the prior and the empirical distribution. Although this sampling strategy does not give an exact draw from the Dirichlet process, it has the same asymptotic performance as the exact sampling method (see, e.g. Hjort 1985, Lo 1987).

2.3 Bootstrap Approaches

The Dirichlet method above has a close relation with the nonparametric bootstrap approaches proposed by Barton and Schruben (1993) and Barton and Schruben (2001) for quantifying the uncertainty in stochastic simulation. Specifically, they proposed the direct bootstrap approach based on Efron (1982) and the Bayesian bootstrap approach based on Rubin (1981), both resampling empirical distributions from the data set and then using direct simulation to construct confidence intervals on the simulation output. We will now adapt their direct bootstrap and Bayesian bootstrap approaches to our setting in quantifying the uncertainty in stochastic optimization.

1. Sampling: for each $j = 1,\ldots,M$,
   - If direct bootstrap: draw i.i.d. samples $Y_1^j,\ldots,Y_N^j$ from $\hat{F}$. Then $F^j(t) = \frac{1}{N}\sum_{i=1}^N I\{Y_i^j \le t\}$ is an empirical distribution.
   - If Bayesian bootstrap (also called "uniformly randomized resampling"): draw i.i.d. samples $Y_1^j,\ldots,Y_N^j$ from $\hat{F}$, order the samples as $Y_{(1)}^j \le \ldots \le Y_{(N)}^j$; generate ordered samples $u_{(1)}^j \le \ldots \le u_{(N)}^j$ from a Uniform(0,1) distribution, and let $u_{(0)}^j = 0$. Then $F^j(t) = \sum_{i=1}^N (u_{(i)}^j - u_{(i-1)}^j) I\{Y_{(i)}^j \le t\}$ is a smoothed empirical distribution.
2. Optimization: for each $j = 1,\ldots,M$, solve the optimization problem

$$\max_{\theta\in\Theta} E_{F^j}[h(X;\theta)],$$

and denote the optimal value as h^{*j}.

Direct bootstrap resamples with replacement from the data and assigns equal probabilities to the samples; whereas Bayesian bootstrap also resamples with replacement but reweights the samples using uniform spacings, because the joint distribution of the probabilities $F(Y_{(1)}^j),\ldots,F(Y_{(N)}^j)$ corresponds to that of the N order statistics from a uniform distribution. The Bayesian bootstrap leads to a more smoothed empirical distribution than the direct bootstrap.

The two bootstrap resampling schemes are closely related with the Dirichlet methods. When the concentration parameter a in the Dirichlet process goes to 0 (which is the non-informative case), the Dirichlet process collapses to a finite-dimensional Dirichlet distribution, and a distribution from the Dirichlet process can only have probability mass on the observed data points. Hence, drawing a sample distribution from the Dirichlet process becomes bootstrap resampling from the data set. More precisely, when $a \to 0$, sampling from the Dirichlet process boils down to Bayesian bootstrap, which explains the name of "Bayesian" bootstrap (Lo 1987); or in other words, samples from the Bayesian bootstrap correspond to discrete distributions supported at the observed data points with Dirichlet distributed weights. On the other hand, the approximate sampling scheme for Dirichlet process in Section 2.2.2 is equivalent to direct bootstrap as $a \to 0$. Therefore, both direct and Bayesian bootstrap approaches can be interpreted as Dirichlet methods with non-informative priors.

3 QUANTIFYING UNCERTAINTY OF THE OPTIMALITY GAP

The three approaches in Section 2 can all be adapted readily to quantify the uncertainty for the optimality gap between (1) and (2). In this section, we suppose that $\hat{\theta}$ has been picked as our selected solution, and we are interested in the uncertainty of the difference $\Delta = \max_{\theta\in\Theta} E[h(X;\theta)] - E[h(X;\hat{\theta})]$. We present the adaptation for each of the methods:

Empirical likelihood: Using the notation in Section 2.1, we can obtain a confidence interval for Δ as follows:

Theorem 4 Under Assumption 1, with $(\nabla_\theta h(X;\theta_0), h(X;\theta_0))$ in Condition 3 replaced by $(\nabla_\theta h(X;\theta_0), h(X;\theta_0) - h(X;\hat{\theta}))$, the programs

$$\begin{array}{ll} \max / \min_{\mathbf{w}} & \max_{\theta\in\Theta} \sum_{i=1}^n w_i(h(X_i;\theta) - h(X_i;\hat{\theta})) \\ \text{subject to} & -2\sum_{i=1}^n \log(nw_i) \le \chi^2_{q,\beta} \\ & \sum_{i=1}^n w_i = 1 \\ & w_i \ge 0 \text{ for all } i \end{array} \tag{8}$$

where max / min denote a pair of maximization and minimization, contain the true value of Δ with probability at least $1-\beta$ asymptotically as $n \to \infty$.

The argument for (4) is similar to the proof of Theorem 1, but now with the profile likelihood defined as

$$\mathscr{R}(\theta,\Delta) = \max\left\{\prod_{i=1}^n nw_i : \sum_{i=1}^n w_i \nabla_\theta h(X_i;\theta) = \mathbf{0},\ \sum_{i=1}^n w_i(h(X_i;\theta) - h(X_i;\hat{\theta})) = \Delta,\ \sum_{i=1}^n w_i = 1,\ w_i \ge 0 \text{ for all } i\right\}$$

since Δ is now the parameter of interest. $\mathscr{R}(\theta,\Delta)$ satisfies the same behavior as Theorem 1 under the modified version of Assumption 1 in Theorem 4, and the argument follows suit.

In terms of computation, (8) has essentially the same complexity as (3).

Dirichlet method and Bootstrap approaches: In the optimization step, for each j, instead of outputting $h^{*j} = \max_{\theta\in\Theta} E_{F^j}[h(X;\theta)]$, one should output

$$h^{*j} - E_{F^j}[h(X;\hat{\theta})],$$

where $E_{F^j}[h(X;\hat{\theta})]$ is approximated by $\frac{1}{N}\sum_{i=1}^N h(Y_i^j;\hat{\theta})$ in the Dirichlet method, or computed according to the empirical distribution F^j in the bootstrap approaches.

4 COMPARISON OF THE PROPOSED METHODS

We have presented three methods for quantifying uncertainty in the optimal value and the optimality gap of a stochastic optimization problem. The EL method is a frequentist approach and based on asymptotics of large samples; it does not require sampling, but needs to solve a two-layer deterministic optimization problem. The Dirichlet method is fully Bayesian; it requires sampling and needs to solve a significant number (usually at least 1000 recommended) of single-layer deterministic optimization problems. The two bootstrap approaches can be viewed as the Dirichlet method with non-informative prior under the exact and approximate sampling schemes respectively, and they also require to solve the same number of single-layer optimization problems as the Dirichlet method. The computational efficiency of these methods is largely determined by the time of solving the two-layer optimization problem or solving a large number of single-layer problems. For the numerical example that we are presenting next, we find that solving the two-layer optimization in the EL method appears more efficient than the latter. However, as discussed before, the max-max problem is non-convex and there is no guarantee of reaching a global optimum. Moreover, there is also a silver lining in spending more computational effort in the Dirichlet and the bootstrap approaches: these methods output much richer information by providing an estimate of the entire posterior distribution of the optimal value. This approximate posterior distribution can be used for estimating many quantities of interest, such as statistics of the optimal value (mean, variance, etc.), and Bayesian confidence intervals under different confidence levels. In contrast, the EL method is not able to provide estimates of such statistics, and will have to solve a different two-layer optimization problem each time when we change the confidence level in order to get the corresponding confidence interval.

5 NUMERICAL STUDY

We demonstrate the presented methods numerically on a simple example of estimating $\mathrm{CVaR}_{\alpha,F^c}(X)$, the α-level Conditional-Value-at-Risk (CVaR) of a random variable X, which we assume follows an unknown distribution F^c. This can be rewritten as a stochastic optimization problem:

$$\min_{\theta}\left\{\theta+\frac{1}{1-\alpha}E[(X-\theta)^+]\right\}, \tag{9}$$

where $(\cdot)^+$ is short for $\max(\cdot,0)$.

We assume that F^c is a standard normal distribution, and set $\alpha=0.9$. Assuming we are given n data from the normal distribution, we implement the EL method (EL), Dirichlet method using Polya urn process (Dir), approximate Dirichlet method (AD), direct bootstrap (DB), and Bayesian bootstrap (BB), to obtain the inferred 95% confidence upper and lower bounds for the optimal value of (9). Moreover, we also compare with the confidence intervals obtained from central limit theorem and the delta method (Shapiro, Dentcheva, and Ruszczyński 2014, Theorem 5.7), given by

$$\left[\hat{Z}^*\pm z_{1-\beta/2}\frac{\hat{\sigma}(\hat{\theta}^*)}{\sqrt{n}}\right]$$

where $z_{1-\beta/2}$ is the critical value of the standard normal distribution with confidence $1-\beta$, $\hat{\theta}^*$ is the empirical optimal solution, $\hat{Z}^*$ is the empirical optimal value given by $(1/n)\sum_{i=1}^n h(X_i;\hat{\theta}^*)$, and $\hat{\sigma}^2(\hat{\theta}^*)$ is the empirical standard deviation of the objective value at the optimal solution given by $\sqrt{(1/(n-1))\sum_{i=1}^n(h(X_i;\hat{\theta}^*)-\hat{Z}^*)^2}$.

Regarding the parameter specifications, we pick the base distribution $F_0\sim N(0,1)$ and concentration parameter $a=0.1$ for the Dirichlet process prior (note that F_0 is exactly the unknown distribution F^c and so we would expect a good performance of the Dirichlet method, which is confirmed below). For both Dirichlet and bootstrap methods, we pick the number of empirical distributions $M=2000$ to be drawn in the outer layer and the number of samples $N=n$ in the inner layer.

We consider three settings $n=10$, 50 and 100. For each setting, we repeat the experiment 100 times, and note down the empirical coverage probability, mean upper and lower bounds, and the mean and standard deviation of the interval width for each method. The results are summarized in Table 1. Note that the true optimal value can be accurately calculated in our setting, and is given by 1.7550.

$n = 10$	Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
EL	0.39	0.80	1.65	0.85	0.56
Dir	0.53	0.73	1.91	1.18	0.38
AD	0.25	1.40	1.76	0.35	0.16
DB	0.20	1.38	1.65	0.27	0.20
BB	0.20	1.37	1.65	0.27	0.20
CLT	0.63	0.94	2.35	1.41	1.02

$n = 50$	Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
EL	0.90	1.21	2.31	1.10	0.40
Dir	0.88	1.21	2.20	0.99	0.31
AD	0.72	1.46	1.97	0.50	0.17
DB	0.67	1.46	1.95	0.49	0.19
BB	0.67	1.46	1.95	0.49	0.19
CLT	0.86	1.22	2.23	1.01	0.42

$n = 100$	Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
EL	0.98	1.34	2.28	0.94	0.27
Dir	0.95	1.31	2.13	0.82	0.20
AD	0.78	1.46	1.96	0.50	0.12
DB	0.71	1.45	1.95	0.50	0.14
BB	0.72	1.45	1.95	0.50	0.14
CLT	0.89	1.36	2.07	0.71	0.21

Table 1: Comparison of performance among different methods, EL, Dir: Dirichet method using Polya urn, AD: approximate Dirichlet, DB: direct bootstrap, BB: Bayesian bootstrap, CLT: classical method

We see that the methods are relatively comparable. For all three cases, EL, Dir and CLT have the highest coverage probabilities, and among the three EL and Dir show more accurate coverage probabilities than CLT as n becomes larger. The other methods, namely AD, DB, and BB, produce much narrower intervals, which is a favorable characteristic that compensates their much lower coverage probabilities. However, their mean lower and upper bounds do not cover the true optimal value 1.7750 when $n = 10$, suggesting these methods are not reliable when data size is small. Overall EL and Dir strike a good balance between coverage probability and interval width when there is a reasonable amount of data. Note that when n is small Dir is sensitive to choice of the prior (i.e., the base distribution F_0) and the concentration parameter a, but the effect of the prior fades away when n becomes larger or when a goes to zero. In the case when we do not have much prior information, it is better to use a noninformative prior such as a very flat normal distribution and a small a; and when we have a good prior, then we can use this prior distribution with a relatively large a.

ACKNOWLEDGMENTS

The first author is grateful to the support by the National Science Foundation under Grants CMMI-1400391 and CMMI-1436247. The second author is grateful to the support by the National Science Foundation under Grant CMMI-1413790 and Grant CAREER CMMI-1453934, and Air Force Office of Scientific Research under Grant YIP FA-9550-14-1-0059.

REFERENCES

Barton, R., and L. Schruben. 1993. "Uniform and Bootstrap Resampling of Input Distributions.". In *Proceedings of the 1993 Winter Simulation Conference*, 503–508.

Barton, R., and L. Schruben. 2001. “Resampling Methods for Input Modeling”. In *Proceedings of the 2001 Winter Simulation Conference*, edited by D. J. M. B. A. Peters, J. S. Smith and e. M. W. Rohrer, 372–378.

Cox, D. R., and D. V. Hinkley. 1979. *Theoretical Statistics*. CRC Press.

Csiszar, I., and G. Tusnady. 1984. “Information Geometry and Alternating Minimization Procedures”. *Statistics & Decisions* Supplement Issue (1): 205–237.

Efron, B. 1982. “The Jackknife, the Bootstrap and Other Resampling Plans”. In *Volume 38 of CBMS-NSF Regional Conference Series in Applied Mathematics*.

Ferguson, T. 1973. “A Bayesian Analysis of Some Nonparametric Problems”. *The Annals of Statistics* 1 (2): 209–230.

Ghosal, S. 2010. *Bayesian Nonparametrics*, Chapter 2: The Dirichlet process, related priors and posterior asymptotics, 35–79. Cambridge Series in Statistical and Probabilistic Mathematics (No. 28). Cambridge University Press.

Hjort, N. 1985. “Bayesian Nonparametric Bootstrap Confidence Intervals”. Technical Report 240, Department of Statistics, Stanford University.

Kim, S., R. Pasupathy, and S. G. Henderson. 2015. “A Guide to Sample Average Approximation”. In *Handbook of Simulation Optimization*, 207–243. Springer.

Lo, A. 1983. “Weak Convergence for Dirichlet Processes”. *Sankhya: The Indian Journal of Statistics, Series A* 45 (1): 105–111.

Lo, A. 1987. “A Large Sample Study of the Bayesian Bootstrap”. *The Annals of Statistics* 15 (1): 360–375.

Mendelson, S. 2003. “A Few Notes on Statistical Learning Theory”. In *Advanced lectures on machine learning*, 1–40. Springer.

Owen, A. B. 1988. “Empirical Likelihood Ratio Confidence Intervals for a Single Functional”. *Biometrika* 75 (2): 237–249.

Owen, A. B. 2001. *Empirical Likelihood*. CRC press.

Pardo, L. 2005. *Statistical inference based on divergence measures*. CRC Press.

Rockafellar, R. T., and S. Uryasev. 2000. “Optimization of Conditional Value-at-Risk”. *Journal of risk* 2:21–42.

Rubin, D. 1981. “The Bayesian Bootstrap”. *Annals of Statistics* 9:130–134.

Shapiro, A., D. Dentcheva, and A. Ruszczyński. 2014. *Lectures on Stochastic Programming: Modeling and Theory*, Volume 16. SIAM.

Shapiro, A., and A. Nemirovski. 2005. “On Complexity of Stochastic Programming Problems”. In *Continuous optimization*, 111–146. Springer.

Sion, M. et al. 1958. “On General Minimax Theorems”. *Pacific J. Math* 8 (1): 171–176.

Teh, Y. 2010. “Dirichlet Process”. *Encyclopedia of Machine Learning*.

Vaart, A. V. D. 1998. *Asymptotic Statistics*. Cambridge UK: Cambridge University Press.

Xie, W., B. Nelson, and R. Barton. 2015. “A Bayesian Framework for Quantifying Uncertainty in Stochastic Simulation”. *Operations Research*. Accepted.

AUTHOR BIOGRAPHIES

HENRY LAM is an Assistant Professor in the Department of Industrial and Operations Engineering at the University of Michigan, Ann Arbor. His research focuses on stochastic simulation, risk analysis, and simulation optimization. His email address is khlam@umich.edu.

ENLU ZHOU is an Assistant Professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Institute of Technology. Her research interests include stochastic control and simulation optimization, with applications towards financial engineering. Her email address is enlu.zhou@isye.gatech.edu and her web page is http://enluzhou.gatech.edu/.