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

ASSESSING SOLUTION QUALITY IN STOCHASTIC OPTIMIZATION VIA BOOTSTRAP AGGREGATING

Henry Lam
Huajie Qian

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

ABSTRACT

We study a statistical method to estimate the optimality gap, as an assessment of the quality, of a given solution for a stochastic optimization using limited data. Our approach is based on bootstrap aggregating the resampled optimal values of sample average approximation (SAA), by connecting these SAA values with the classical notion of symmetric statistics. We discuss how this approach works on general stochastic optimization problems and is statistically more efficient than some previous methods. We substantiate our findings with several numerical experiments.

1 INTRODUCTION

Consider a stochastic optimization problem

$$Z^* = \min_{x\in\Theta}\{Z(x) = E_F[h(x,\xi)]\} \tag{1}$$

where $\xi \in \Xi$ is generated under some distribution F, and $E_F[\cdot]$ denotes its expectation. In many situations, F is not known, but instead a collection of i.i.d. data for ξ, say $\boldsymbol{\xi}_{1:n} = (\xi_1, \ldots, \xi_n)$, are available. Obtaining a good solution for (1) under this setting has been under active investigation both from the stochastic community and the optimization community. Common methods include the sample average approximation (SAA) (Shapiro et al. 2009), stochastic approximation (SA) or gradient descent (Kushner and Yin 2003; Nemirovski et al. 2009), and (distributionally) robust optimization (Delage and Ye 2010; Bertsimas et al. 2017). These methods all aim to find a solution that is nearly optimal, or in some way provide a safe approximation, for the problem (1).

In this paper, we take the view that a (good) solution, say $\hat{x}$, is given to us. Our goal is to assess the quality of $\hat{x}$ using the data we have. This quality is measured by the optimality gap

$$\mathscr{G}(\hat{x}) = Z(\hat{x}) - Z^* \tag{2}$$

in the sense that the smaller $\mathscr{G}(\hat{x})$ is, the closer is the solution $\hat{x}$ to the true optimum in terms of achieved objective value. More precisely, we are interested in drawing inference on $\mathscr{G}(\hat{x})$, by using potentially limited data.

This problem is motivated by the fact that many state-of-the-art guarantees of the aforementioned solution methods can only provide crude, worst-case performance bounds on the optimality guarantees. For instance, Shapiro and Nemirovski (2005) and Kleywegt et al. (2002) provide large deviations bounds on the optimality gap of SAA in terms of the diameter or cardinality of the decision space and the maximal variance of the function h. Nemirovski et al. (2009) and Ghadimi and Lan (2013) provide bounds on the expected value and deviation probabilities of optimality gaps of the SA iterates in terms of the strong

convexity parameters, space diameter and maximal variance. These bounds can be refined under additional structural information (e.g., Shapiro and Homem-de Mello 2000). However, while they are very useful in understanding the behaviors of the optimization procedures, using them as a precise assessment on the quality of an obtained solution may be conservative. Because of this, a stream of work Shapiro (2003), Mak et al. (1999), Bayraksan and Morton (2006), Lan et al. (2012), Stockbridge and Bayraksan (2013), Bayraksan and Morton (2011) study approaches to validate given solutions. Our investigation in this paper follows this latter line of work, focusing on situations when data are limited and hence the statistical efficiency of the validation scheme becomes important.

We discuss more precisely our target quantity. Given $\hat{x}$, we aim to construct an upper confidence bound for $\mathscr{G}(\hat{x})$. The reason why we focus on upper bound (instead of lower bound) is that it provides a way to screen out bad solutions, while the lower bound cannot do so (because a small lower bound does not translate to a belief of a small gap and a good solution quality). So from a risk perspective, the upper bound is the more important direction to investigate.

The crux of this problem is the estimation of the optimal value, Z^*, i.e., finding a lower confidence bound for Z^*, which arises in the second term in (2). Assuming $\boldsymbol{\xi}_{1:n}$ are independent of $\hat{x}$, then the first term, namely $Z(\hat{x})$, can be estimated as a standard population mean since $\hat{x}$ is assumed fixed (i.e., a $1-\alpha$ level confidence interval for $Z(\hat{x})$ is simply $(1/n)\bar{Z}(\hat{x}) \pm z_{1-\alpha/2}\sigma(\hat{x})/\sqrt{n}$ where $\bar{Z}(\hat{x}) = \sum_{i=1}^{n} h(\hat{x},\xi_i)$ and $\sigma^2(\hat{x}) = 1/(n-1)\sum_{i=1}^{n}(h(\hat{x},\xi_i)-\bar{Z}(\hat{x}))^2$ are the sample mean and variance of h evaluated at $\hat{x}$). On the other hand, Z^* involves operations much more complex than taking the mean. Thus the bulk of our discussion throughout this paper is on estimating Z^*.

Finally, we discuss whether our investigated setting applies when $\hat{x}$ is constructed using a data set that is the same or contains some of $\boldsymbol{\xi}_{1:n}$. In fact, estimating Z^* can utilize any of the data in constructing $\hat{x}$, whereas the data in estimating $Z(\hat{x})$ need to be independent from those for constructing $\hat{x}$ to get a valid upper bound. Then, through a Bonferroni correction, we can obtain a valid lower bound for Z^* and upper bound for $Z(\hat{x})$ simultaneously. From this view, in a typical problem setting, the amount of data available for inferring Z^* is at least as much as that for finding $\hat{x}$.

As our main contribution, we propose a resampling method based on bootstrap aggregating to estimate a lower confidence bound for Z^*. This approach works for a general class of cost function h and decision space Θ (e.g., it does not require Lipschitz continuity of h needed in invoking the classical asymptotic theory of SAA). In particular, it utilizes a general optimistic bound that exchanges the minimization and the expectation operators for SAA problems, which has been studied and used by, e.g., Mak et al. (1999) in designing estimation schemes based on data batching. By connecting an SAA problem with the classical notion of symmetric statistics, we show that simple recycling, i.e., resampling, of data can achieve greater statistical efficiency asymptotically. We demonstrate how this scheme can be viewed as a bootstrap aggregating (or so-called bagging, see Breiman 1996) procedure, and explain the mechanism and root of its efficiency in the estimation of Z^*.

2 USING ASYMPTOTICS OF SAMPLE AVERAGE APPROXIMATION

We first motivate the need for methods beyond the classical theory of SAA. When the cost function h is well-conditioned, then, in a sense, a central limit theorem (CLT) governs the behavior of the estimated optimal value in SAA, namely

Theorem 1 (Adopted from Theorem 5.7 in Shapiro et al. 2009) Suppose, a.s. in ξ, $h(\cdot,\xi)$ is Lipschitz continuous in the sense

$$\|h(x,\xi)-h(y,\xi)\| \le L(\xi)\|x-y\|$$

for any $x,y \in \Theta$ and $L(\xi)$ such that $E[L(\xi)^2] < \infty$. Suppose that $E[h(\tilde{x},\xi)^2] < \infty$ for some point $\tilde{x} \in \Theta$, Θ is compact, and there is a unique optimal solution x^* to problem (1). Given i.i.d. data $\boldsymbol{\xi}_{1:n} = (\xi_1,\ldots,\xi_n)$,

consider the SAA problem

$$\hat{Z}_n = \min_{x\in\Theta}\left\{\bar{Z}(x) = \frac{1}{n}\sum_{i=1}^{n} h(x,\xi_i)\right\}. \tag{3}$$

The SAA optimal value $\hat{Z}_n$ satisfies

$$\sqrt{n}(\hat{Z}_n - Z^*) \Rightarrow N(0,\sigma^2) \tag{4}$$

where $\sigma^2 = Var(h(x^*,\xi))$ and $\Rightarrow$ denotes convergence in distribution.

Theorem 1 stipulates that, under the depicted conditions, one can use

$$\hat{Z}_n - z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} \tag{5}$$

as a valid lower confidence bound for Z^*, where $\hat{\sigma}^2 = (1/(n-1))\sum_{i=1}^{n}(h(\hat{x}^*,\xi_i) - \hat{Z}_n)^2$, with $\hat{x}^*$ being an optimal solution of (3), is the empirical estimate of σ^2.

However, there are several reasons why one would need a method beyond the one offered by Theorem 1. When there are multiple optimal solutions, then the limit in (4) is the infimum of a Gaussian process instead of merely a Gaussian random variable, making inference on Z^* more challenging. Moreover, some common classes of optimization problems (e.g., integer programming) may not satisfy the continuity needed in the theorem. Ideally, we would like a method that is more general than in using Theorem 1.

3 EXISTING BATCHING PROCEDURES

To circumvent the limitations of Theorem 1, one approach is to use the optimistic bound (Mak et al. 1999; Shapiro 2003; Glasserman 2013)

$$E[\hat{Z}_n] \le Z^* \tag{6}$$

where $E[\cdot]$ in (6) is taken with respect to the data in constructing the SAA value $\hat{Z}_n$. The bound (6) holds for any $n \ge 1$, a direct consequence from exchanging the expectation and the minimization operator in the SAA.

The bound (6) offers a simple way to construct a lower bound for Z^* in much greater generality than Theorem 1. Note that the left hand side of (6) is a mean of SAA. Thus, if one can "sample" a collection of SAA values, then a confidence bound for Z^* can be constructed readily by using the standard sample-mean-based procedure. To sample SAA values, an approach suggested by Mak et al. (1999) is to batch the i.i.d. data set $\boldsymbol{\xi}_{1:n}$ into say m batches, each batch consisting of k observations, so that $mk = n$ (we ignore rounding issues). For each $j = 1,\ldots,m$, solve an SAA using the k observations in the j-th batch; call this $\hat{Z}_k^j$. Then we can use

$$\tilde{Z}_k - z_{1-\alpha/2}\frac{\tilde{\sigma}}{\sqrt{m}} \tag{7}$$

where $\tilde{Z}_k = (1/m)\sum_{j=1}^{m}\hat{Z}_k^j$ and $\tilde{\sigma}^2 = (1/(m-1))\sum_{j=1}^{m}(\hat{Z}_k^j - \tilde{Z}_k)^2$ is the sample mean and variance from $\hat{Z}_k^j, j = 1,\ldots,m$. This procedure is shown in Algorithm 1.

Note that the tightness of the obtained bound in this procedure depends on the choices of m and k. On one hand, m must be chosen big enough, e.g., > 30, so that one can use the CLT to justify the approximation (7). Moreover, the larger is m, the smaller magnitude is the standard error in the second term of (7). On the other hand, the larger is k, the closer $E[\hat{Z}_k^j]$ is to Z^* in (6). This is thanks to a monotonicity property in that $E[\hat{Z}_n]$ is non-decreasing in n (Mak et al. 1999), and thus one would expect $E[\hat{Z}_k^j]$, and hence $\tilde{Z}_k$, to be higher as k increases, leading to a tighter lower bound for Z^*. Therefore, there is a tradeoff between the statistical accuracy controlled by m (in terms of the validity of the CLT and the magnitude of the standard error term) and the tightness controlled by k (in terms of the position of $E[\hat{Z}_k^j]$ in (6)). In the batching or

Algorithm 1 The Batching Procedure from Mak et al. (1999)

Given n observations $\xi_1,\ldots,\xi_n$, choose m,k such that $mk=n$
for $j=1$ **to** m **do**
From $\{\xi_{(j-1)k+1},\ldots,\xi_{jk}\}$, solve

$$\hat{Z}_k^j = \min_{x\in\Theta} \frac{1}{k}\sum_{i=1}^{k} h(x,\xi_{(j-1)k+i})$$

end for
Compute $\tilde{Z}_k = \frac{1}{m}\sum_{j=1}^{m}\hat{Z}_k^j$ and $\tilde{\sigma}^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{Z}_k^j-\tilde{Z}_k)^2$
Output $\tilde{Z}_k - z_{1-\alpha/2}\frac{\tilde{\sigma}}{\sqrt{m}}$

the so-called multiple-replication approach of Mak et al. (1999), this tradeoff is confined to the relation $mk=n$.

Our proposed procedure that follows next breaks this tradeoff confinement. This procedure can be compared to the single or double-replication procedure in Bayraksan and Morton (2006), which also does so by generalizing the use of (5) to problems with multiple optimal solutions. Our procedure is more general than Bayraksan and Morton (2006) in the aspect of removing the continuity requirement needed in their scheme. However, to make our scheme fully efficient, we require, instead, smoothness conditions on h with respect to the sample ξ (instead of the decision x). Our scheme can thus be viewed as "transferring" the regularity requirement on the decision space to the sample space, and we do so by interpreting the optimistic bound (6) from a bootstrap aggregating (or so-called bagging) framework. We should also note that our procedure assumes adequate computing capacity to solve many SAA problems (i.e., the statistical efficiency, instead of the optimization efficiency, is our main concern).

4 BOOTSTRAP AGGREGATING FOR ESTIAMTING OPTIMAL VALUES

Instead of batching the data into k samples per batch, we resample the data. Specifically, we repeatedly solve many, say B, SAA problems, where each SAA problem consists of k resampled observations from $\boldsymbol{\xi}_{1:n}$. The resampling can be done with or without replacement (and we will discuss some differences in the arising statistical performances between the two cases). The procedure is summarized in Algorithm 2.

The output of the procedure using $\tilde{Z}_k^{bag}$ is an averaging of many bootstrap resampled SAA, which resembles the bagging scheme (Breiman 1996). The quantity $\widehat{Cov}_*(N_i^*,\hat{Z}_k^*)$ in (9) is the covariance between the count of a specific observation ξ_i in a bootstrap resample, denoted N_i^*, and the resulting resampled SAA value $\hat{Z}_k^*$. The quantity $\tilde{\sigma}_{IJ}^2 = \sum_{i=1}^{n}\widehat{Cov}_*(N_i^*,\hat{Z}_k^*)^2$ is an empirical version of the so-called infinitesimal jackknife (IJ) estimator (Efron 2014), which has been used in the uncertainty quantification of bagging schemes, including random forests or tree ensembles in machine learning (Wager et al. 2014). The additional constant factor $(n/(n-k))^2$ in the second line of (8) is a finite sample correction specific to resampling without replacement, which has appeared in the context of random forests (Wager and Athey 2018). Although the IJ variance estimator is not affected by this factor in asymptotic sense, we find it significantly improves finite sample performance of our method.

We explain how Algorithm 2 arises and its connection to bagging. In a nutshell, the $\tilde{Z}_k^{bag}$ in Algorithm acts as a point estimator for $E[\hat{Z}_k]$ in (6), whereas $\tilde{\sigma}_{IJ}^2$ captures the standard error in using this point estimator. Let us introduce a functional viewpoint and write

$$W_k(F) = E_{F^k}[H_k(\xi_1,\ldots,\xi_k)] \tag{10}$$

Algorithm 2 Bagging-based Procedure

Given n observations $\boldsymbol{\xi}_{1:n} = \{\xi_1, \ldots, \xi_n\}$, select a positive integer k
for $b = 1$ **to** B **do**
Randomly sample $\boldsymbol{\xi}_k^b = (\xi_1^b, \ldots, \xi_k^b)$ uniformly from $\boldsymbol{\xi}_{1:n}$ (with or without replacement), and solve

$$\hat{Z}_k^b = \min_{x \in \Theta} \frac{1}{k} \sum_{i=1}^{k} h(x, \xi_i^b)$$

end for
Compute $\tilde{Z}_k^{bag} = \frac{1}{B} \sum_{b=1}^{B} \hat{Z}_k^b$ and

$$\tilde{\sigma}_{IJ}^2 = \begin{cases} \sum_{i=1}^{n} \widehat{Cov}_*(N_i^*, \hat{Z}_k^*)^2, & \text{if resampling is with replacement} \\ \left(\frac{n}{n-k}\right)^2 \sum_{i=1}^{n} \widehat{Cov}_*(N_i^*, \hat{Z}_k^*)^2, & \text{if resampling is without replacement} \end{cases} \tag{8}$$

where

$$\widehat{Cov}_*(N_i^*, \hat{Z}_k^*) = \frac{1}{B} \sum_{b=1}^{B} (N_i^b - \frac{k}{n})(\hat{Z}_k^b - \tilde{Z}_k^{bag}) \tag{9}$$

and N_i^b is the number of ξ_i that shows up in the b-th resample
Output $\tilde{Z}_k^{bag} - z_{1-\alpha/2}\tilde{\sigma}_{IJ}$

where

$$H_k(\xi_1, \ldots, \xi_k) = \min_{x \in \Theta} \frac{1}{k} \sum_{i=1}^{k} h(x, \xi_i)$$

is the SAA value, expressed more explicitly in terms of the underlying data used. Here, the expectation $E_{F^k}[\cdot]$ is generated with respect to i.i.d. variables $(\xi_1, \ldots, \xi_k)$, i.e., F^k denotes the product measure of k F's. For convenience, we denote $E[\cdot]$ as the expectation with respect to F or the product measure of F's when no confusion arises. Also, we denote $W_k = W_k(F)$.

Under these notations, the optimistic bound (6) can be expressed as

$$W_k(F) \leq Z^*$$

with the best bound being $W_\infty = \lim_{k \to \infty} W_k \leq Z^*$ thanks to the monotonicity property discussed before.

Suppose that we have used sampling with replacement in our bootstrap scheme in Algorithm 2. Also say we use infinitely many bootstrap replications, i.e., $B = \infty$. Then, the estimator $\tilde{Z}_k^{bag}$ in Algorithm 2 becomes precisely

$$\tilde{Z}_k^{bag} = W_k(\hat{F})$$

where $\hat{F}$ is the empirical distribution formed by $\boldsymbol{\xi}_{1:n}$, i.e., $\hat{F}(\cdot) = (1/n)\sum_{i=1}^{n} \delta_{\xi_i}(\cdot)$ where $\delta_{\xi_i}(\cdot)$ is the delta measure at ξ_i. If $W_k(\cdot)$ is "smooth" in some sense, then one would expect $W_k(\hat{F})$ to be close to $W_k(F)$. Indeed, $W_k(F)$, expressed as the k-fold expectation under F in (10), is Gateaux differentiable from the theory of von Mises statistical functionals (Serfling 2009). This gives in particular a CLT in approximating $W_k(F)$ using $W_k(\hat{F})$.

Note that $W_k(\hat{F})$ consists exactly of the average of $H_k(\xi_{i_1}, \ldots, \xi_{i_k})$ over all possible combinations of $\{\xi_{i_1}, \ldots, \xi_{i_k}\}$ drawn with replacement from $\boldsymbol{\xi}_{1:n}$. This is equivalent to

$$V_{n,k} = \frac{1}{n^k} \sum_{i_j \in \{1,\ldots,n\}, j=1,\ldots,k} H_k(\xi_{i_1}, \ldots, \xi_{i_k}) \tag{11}$$

which is the so-called V-statistic. If we have used sampling without replacement in Algorithm 2, we arrive at the estimator (assuming $B = \infty$)

$$U_{n,k} = \frac{1}{\binom{n}{k}} \sum_{(i_1,\ldots,i_k)\in\mathscr{C}_k} H_k(\xi_{i_1},\ldots,\xi_{i_k}) \tag{12}$$

where $\mathscr{C}_k$ denotes the collection of all subsets of size k in $\{1,\ldots,n\}$. The quantity (12) is known as the U-statistic. The estimators (11) and (12) belong to the class of symmetric statistics (Serfling 2009; Van der Vaart 2000), since the estimator is unchanged by shuffling the ordering of the data $\boldsymbol{\xi}_{1:n}$. Symmetric statistics generalize the sample mean, the latter corresponding to the case when $k = 1$.

When $B < \infty$, then $V_{n,k}$ and $U_{n,k}$ above are approximated by a random sampling of the summands on the right hand side of (11) and (12). These are known as incomplete V and U statistics (Serfling 2009; Van der Vaart 2000). As B is chosen large enough, $\tilde{Z}_k^{bag}$ will well approximate $V_{n,k}$ and $U_{n,k}$.

To discuss further, we make the following assumption:

Assumption 2 We have

$$E \sup_{x\in\Theta} |h(x,\xi)|^2 < \infty.$$

Assumption 2 implies that $EH_k(\xi_1,\ldots,\xi_k)^2 < \infty$ for any k, since

$$EH_k(\xi_1,\ldots,\xi_k)^2 \le \frac{1}{k^2} E \sup_{x\in\Theta} \left(\sum_{i=1}^{k} h(x,\xi_i) \right)^2 \le E \sup_{x\in\Theta} |h(x,\xi)|^2 < \infty$$

by the Minkowski inequality.

To state our next assumption, denote $g_k(\xi) = E[H_k(\xi_1,\ldots,\xi_k)|\xi_1 = \xi]$. Denote $Var(\cdot) = Var_F(\cdot)$ as the variance under F. We make the following assumption on g_k:

Assumption 3 We have $0 < Var(g_k(\xi)) < \infty$.

Then we have the following result regarding the asymptotic of $U_{n,k}$ and $V_{n,k}$:

Theorem 4 Suppose $k \ge 1$ is fixed, and Assumptions 2 and 3 hold. Then

$$\sqrt{n}(U_{n,k} - W_k) \Rightarrow N(0, k^2 Var(g_k(\xi)))$$

and

$$\sqrt{n}(V_{n,k} - W_k) \Rightarrow N(0, k^2 Var(g_k(\xi)))$$

as $n \to \infty$, where $N(0, k^2 Var(g_k(\xi)))$ is a normal distribution with mean 0 and variance $k^2 Var(g_k(\xi))$.

Theorem 4 is a consequence of the classical CLT for symmetric statistics (e.g., Van der Vaart 2000). The expression $kg_k(\xi)$, as a function defined on the space Θ, is the so-called influence function of $W_k(F)$, which can be viewed as its functional derivative (Hampel 1974). Alternately, for a U-statistic $U_{n,k}$, this is the so-called Hajek projection (Van der Vaart 2000), which is the projection of the statistic onto the subspace generated by the linear combinations of $f_i(\xi_i), i = 1,\ldots,n$ and any measurable function f_i. It turns out that these two views coincide, and the U- and V-statistics obey the same CLT as depicted in Theorem 4. Finally, the quantity $\tilde{\sigma}_{IJ}^2$ in Algorithm 2, an empirical IJ estimator, can be shown to approximate the true variance $k^2 Var(g_k(\xi))$ as $n, B \to \infty$, by borrowing recent results in bagging (Efron 2014; Wager and Athey 2018).

The discussion above holds for a fixed k, the sample size we use in obtaining the resampled SAA values. It shows that, at least asymptotically, using with or without replacement does not matter. However, with a bigger n, intuitively we can afford to use a bigger k without distorting the asymptotic behavior, and this in turn allows us to get a tighter $W_k = E[\hat{Z}_k]$ in the optimistic bound (6). The next section shows that we could indeed do so to a certain extent.

5 ASYMPTOTIC BEHAVIOR OF BAGGING-BASED PROCEDURES WITH GROWING RESAMPLE SIZE

We first make the following strengthened version of Assumption 2:

Assumption 5 ($L_{2+\delta}$-bounded modulus of continuity) We have

$$E \sup_{x\in\Theta} |h(x,\xi) - h(x,\xi')|^{2+\delta} < \infty$$

where ξ, ξ' are i.i.d. generated from F.

We also make the following assumption:

Assumption 6 (Non-degeneracy) We have

$$P\left(\min_{x\in\Theta}\{h(x,\xi) - Z(x)\} > 0\right) + P\left(E\left[\min_{x\in\Theta}\{h(x,\xi) - h(x,\xi')\}\Big|\xi'\right] > 0\right) > 0 \tag{13}$$

where $\xi, \xi' \overset{i.i.d.}{\sim} F$.

Assumption 5 can be a bit obscure. There are several sufficient conditions for Assumption 5, such as:

Assumption 7 (Uniform boundedness) $h(\cdot,\cdot)$ is uniformly bounded over $\Theta \times \Xi$.

Assumption 8 (Uniform Lipschitz condition) $h(x,\xi)$ is Lipschitz continuous with respect to ξ, where the Lipschitz constant is uniformly bounded in $x \in \Theta$, i.e.,

$$|h(x,\xi) - h(x,\xi')| \le L\|\xi - \xi'\|$$

where $\|\cdot\|$ is some norm in Ξ. Moreover, $E\|\xi\|^{2+\delta} < \infty$.

Assumption 9 (Majorization)

$$|h(x,\xi) - h(x,\xi')| \le f(\xi) + f(\xi')$$

where $Ef(\xi)^{2+\delta} < \infty$.

That Assumption 7 implies Assumption 5 is straightforward. To see how Assumption 8 implies Assumption 5, note that, if the former is satisfied, we have

$$E \sup_{x\in\Theta} |h(x,\xi) - h(x,\xi')|^{2+\delta} \le L^{2+\delta} E\|\xi - \xi'\|^{2+\delta} < \infty.$$

Similarly, Assumption 9 implies Assumption 5 because the former leads to

$$E \sup_{x\in\Theta} |h(x,\xi) - h(x,\xi')|^{2+\delta} \le E(f(\xi) + f(\xi'))^{2+\delta} < \infty.$$

On the other hand, Assumption 6 roughly means that ξ is sufficiently mixed so that the optimal value of a data-driven optimization problem with only one (or two) data point can deviate away from its mean. This assumption holds, e.g., when Θ lies in a positive region in the real space that is bounded away from the origin. The assumption can be somewhat stringent, but it can be relaxed in practical problems. For example, one can replace Θ in (13) by a smaller region that can possibly contain any candidates of optimal solutions.

An important consequence of this assumption is that it implies $k^2 Var(g_k(\xi))$ is bounded away from 0 even as k grows, thus leading to a behavior similar to Assumption 3 for the finite k case.

Lemma 10 Suppose Assumption 2 holds. Assumption 6 implies that $k^2 Var(g_k(\xi)) > \varepsilon > 0$ for some constant ε, when k is sufficiently large.

We have the following result:

Theorem 11 Suppose Assumptions 2, 5 and 6 hold. Then, for $k = O(n^\gamma)$ where $\gamma < 1/2$, we have

$$\frac{\sqrt{n}(U_{n,k} - W_k)}{k\sqrt{Var(g_k(\xi))}} \Rightarrow N(0,1)$$

and

$$\frac{\sqrt{n}(V_{n,k} - W_k)}{k\sqrt{Var(g_k(\xi))}} \Rightarrow N(0,1)$$

where $N(0,1)$ is the standard normal variable.

Theorem 11 is an analog of Theorem 4 when $k \to \infty$. We discuss its implications. First, there is a limit in how large k we can take relative to n, which is thresholded at order $\sqrt{n}$ (The U-statistic case can be generalized slight to allow $k = o(\sqrt{n})$ instead of $O(n^\gamma)$ for $\gamma < 1/2$). We note that a symmetric statistic as $k \to \infty$ is known as an infinite-order symmetric statistic (Frees 1989), and Theorem 11 complements the existing theory in this context, specializing to the case where the kernel, i.e., $H_k(\cdot)$, is precisely the SAA value.

Next, we have the following result regarding the bias of $U_{n,k}$ and $V_{n,k}$:

Theorem 12 Under the same assumptions as Theorem 11, the bias of $U_{n,k}$ in estimating W_k is 0, whereas the bias of $V_{n,k}$ in estimating W_k is $O((k^2/n)^l + k/n)$ where l is any fixed positive integer.

The zero-bias property of $U_{n,k}$ is trivial: Each summand in its definition is an SAA value with distinct i.i.d. data, and thus has mean exactly W_k. On the other hand, the summands in $V_{n,k}$ are SAA values constructed from potentially repeated observations, which induces bias relative to W_k. In this sense, $U_{n,k}$ is again more advantageous compared to $V_{n,k}$ as k grows.

6 NUMERICAL EXPERIMENTS

In this section we provide numerical tests to demonstrate the validity of our resampling method and compare its performance to the batching procedure (Mak et al. 1999) in the case of limited data. The stochastic optimization problem we solve is the $(1-\alpha)$-level conditional value at risk (CVaR) of a standard normal variable ξ

$$\min_x \{x + \frac{1}{\alpha} E[(\xi - x)_+]\}$$

where $(\cdot)_+ := \max\{\cdot, 0\}$ denotes the positive part. We set $\alpha = 0.1$, namely, we are solving for the 0.9-level CVaR of the standard normal, whose true value can be calculated to be 1.755.

We use Algorithm 1, Algorithm 2 without replacement (U-statistic) and Algorithm 2 with replacement (V-statistic) to compute a lower bound of the true optimal value with 95% confidence level. Specifically, we first simulate an i.i.d. standard normal data set $\xi_1, \ldots, \xi_n$, of size n, and then compute a lower bound of CVaR using each of the three methods under different choices of resample size k. Note that for Algorithm 1, the parameter k refers to the batch size, and the number of batches m is automatically set to n/k once given n and k. Also, since our main interest is the statistical efficiency, we set B, the number of resamples, in Algorithm 2 to be sufficiently large (we use $B = 5000$) to offset the effect of Monte Carlo error in estimating the covariances, which is in accordance with the assumption of adequate capacity in solving the optimization problems.

The three methods are tested and compared under data size $n = 50$ (Table 1) and $n = 300$ (Table 2). We draw 1000 realizations of the data, compute a lower confidence bound based on each of them, and then average the results to estimate coverage probability, mean of the lower bound and standard deviation of the lower bound. For a more complete comparison, two sets of results are provided for the batching procedure, one using the critical value of standard normal, and the other (in parentheses) using t-distribution with $m-1$ degrees of freedom to enhance finite-sample performances as suggested in Mak et al. (1999).

The two tables show that for a wide range of resample sizes, from 10 to more than half of the data size, our bagging-based procedure generates statistically valid lower bounds in the sense that the coverage probabilities are equal to or above the nominal value 95%. The results also verify the relation between the resample size k and tightness of the optimistic bound (6). To be specific, in both tables, as the resample size k grows, the mean of the lower bound gets closer to the true optimal value 1.755. In particular, in the case of $n = 300$ both the U-statistic and V-statistic approaches can provide a lower bound as good as 1.55 with coverage probability 97%-98% by using $k = 100, 150, 250$. Therefore, with the resampling procedure, one can obtain a relatively tight bound for the optimal value and in the meantime retain good statistical accuracy, by using a resample size k that is roughly half the data size. The good performance for such a large k also hints that the choice of k in Theorem 11 can be pushed larger in this example, likely because the considered performance function is quite smooth in the decision variable.

On the other hand, such a balance may not be easy to establish for the batching procedure because of the trade-off discussed in Section 3, especially when the data size is small. When the standard normal critical value is used, batching procedure either loses statistical accuracy when k is chosen large (coverage drops to 92% when $n = 50$, and 86% when $n = 300$), or generating relatively loose bounds (1.12 versus 1.26 when $n = 50$) when k is chosen small. As a result, a good choice of k may not be even possible to have when n is small (e.g., 50); when n is larger, good choices of k can exist, e.g. the case $k = 50$ in Table 2, but they are in general not known beforehand. The statistical accuracy seems to be improved when the t critical value is used, as all choices of k now lead to valid coverages under both sample sizes. However, the bounds get loose when k is chosen to be too small (because of the looseness of the optimistic bound (6)), and too large as well in this case (because of a small degree of freedom in the considered t-distribution that results in a large standard error).

Lastly, note that when $n = 50$ our resampling approaches with $k = 25, 40$ outperform all the cases of the batching approach, in terms of either statistical accuracy or tightness of the bound. This reveals that the issue of tightness and statistical accuracy trade-off becomes severer when there are limited data, and highlights the benefits of our bagging procedures in these situations.

Table 1: $n = 50$. "mean" and "std." represent the mean and standard deviation of the confidence bound.

	Batching with z-value (t-value)			U-statistic			V-statistic		
k	coverage(%)	mean	std.	coverage(%)	mean	std.	coverage(%)	mean	std.
10	99.0(99.4)	1.12(1.00)	0.27(0.29)	99.4	1.16	0.22	99.7	1.16	0.20
25	92.1(97.1)	1.29(0.36)	0.34(0.96)	98.9	1.23	0.22	99.6	1.23	0.21
40	NA	NA	NA	98.6	1.26	0.23	98.5	1.26	0.23

Table 2: $n = 300$. "mean" and "std." represent the mean and standard deviation of the confidence bound.

	Batching with z-value (t-value)			U-statistic			V-statistic		
k	coverage(%)	mean	std.	coverage(%)	mean	std.	coverage(%)	mean	std.
10	100(100)	1.37(1.36)	0.10(0.10)	100	1.38	0.09	100	1.38	0.08
30	98.4(99.1)	1.50(1.48)	0.11(0.11)	99.8	1.50	0.09	99.5	1.51	0.09
50	97.3(97.8)	1.53(1.50)	0.11(0.12)	98.7	1.53	0.10	98.9	1.52	0.10
100	92.6(96.9)	1.58(1.44)	0.13(0.19)	97.9	1.55	0.10	98.0	1.55	0.09
150	86.2(96.1)	1.59(1.20)	0.15(0.42)	98.0	1.55	0.10	97.6	1.55	0.10
250	NA	NA	NA	96.9	1.55	0.10	98.1	1.56	0.10

7 CONCLUSION

We have studied a bagging approach to estimate bounds for the optimal value, and consequently also the optimality gap, for stochastic optimization problems under minimal assumptions. Our motivation is that for many of these problems classical CLT convergences may fail to hold. Our approach is based on an established universal optimistic bound for SAA. Previous approaches that use batching to utilize this bound can run into a trade-off between tightness and statistical accuracy that is difficult to control when the sample size is small. We proposed a resampling-based method that views the optimistic bound as a target quantity estimated by symmetric statistics, which can be cast as bagging procedures and lead to less sample size requirements than batching. We explained how our method succumbs to a CLT without any smoothness conditions on the objective function. We also investigated convergences when the batch size grows relative to the sample size, which requires smoothness assumptions with respect to the sample space, a property orthogonal to existing approaches that impose smoothness to the decision. Lastly, we numerically demonstrated the benefits of our approach in improving the overall tightness-accuracy balance, especially in small-sample situations. Future work includes investigations taking into account computational considerations and further comparisons with other potential techniques.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CMMI-1542020, CMMI-1523453 and CAREER CMMI-1653339/1834710.

REFERENCES

Bayraksan, G., and D. P. Morton. 2006. "Assessing Solution Quality in Stochastic Programs". *Mathematical Programming* 108(2-3):495–514.

Bayraksan, G., and D. P. Morton. 2011. "A Sequential Sampling Procedure for Stochastic Programming". *Operations Research* 59(4):898–913.

Bertsimas, D., V. Gupta, and N. Kallus. 2017. "Robust Sample Average Approximation". *Mathematical Programming* doi:10.1007/s10107-017-1174-z.

Breiman, L. 1996. "Bagging Predictors". *Machine Learning* 24(2):123–140.

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

Efron, B. 2014. "Estimation and Accuracy After Model Selection". *Journal of the American Statistical Association* 109(507):991–1007.

Frees, E. W. 1989. "Infinite Order U-Statistics". *Scandinavian Journal of Statistics* 16(1):29–45.

Ghadimi, S., and G. Lan. 2013. "Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming". *SIAM Journal on Optimization* 23(4):2341–2368.

Glasserman, P. 2013. *Monte Carlo Methods in Financial Engineering*, Volume 53. Springer Science & Business Media.

Hampel, F. R. 1974. "The Influence Curve and Its Role in Robust Estimation". *Journal of the American Statistical Association* 69(346):383–393.

Kleywegt, A. J., A. Shapiro, and T. Homem-de Mello. 2002. "The Sample Average Approximation Method for Stochastic Discrete Optimization". *SIAM Journal on Optimization* 12(2):479–502.

Kushner, H., and G. G. Yin. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*, Volume 35. Springer Science & Business Media.

Lan, G., A. Nemirovski, and A. Shapiro. 2012. "Validation Analysis of Mirror Descent Stochastic Approximation Method". *Mathematical Programming* 134(2):425–458.

Mak, W.-K., D. P. Morton, and R. K. Wood. 1999. "Monte Carlo Bounding Techniques for Determining Solution Quality in Stochastic Programs". *Operations Research Letters* 24(1-2):47–56.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro. 2009. "Robust Stochastic Approximation Approach to Stochastic Programming". *SIAM Journal on Optimization* 19(4):1574–1609.

Serfling, R. J. 2009. *Approximation Theorems of Mathematical Statistics*, Volume 162. John Wiley & Sons.

Shapiro, A. 2003. "Monte Carlo Sampling Methods". In *Stochastic Programming*, edited by A. Ruszczynski and A. Shapiro, Volume 10 of *Handbooks in Operations Research and Management Science*, 353–425. Elsevier.

Shapiro, A., and T. Homem-de-Mello. 2000. "On the Rate of Convergence of Optimal Solutions of Monte Carlo Approximations of Stochastic Programs". *SIAM Journal on Optimization* 11(1):70–86.

Shapiro, A., and A. Nemirovski. 2005. "On Complexity of Stochastic Programming Problems". In *Continuous Optimization: Current Trends and Modern Applications*, edited by V. Jeyakumar and A. Rubinov, 111–146. Boston, MA: Springer US.

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

Stockbridge, R., and G. Bayraksan. 2013. "A Probability Metrics Approach for Reducing the Bias of Optimality Gap Estimators in Two-Stage Stochastic Linear Programming". *Mathematical Programming* 142(1-2):107–131.

Van der Vaart, A. W. 2000. *Asymptotic Statistics*, Volume 3. Cambridge University Press.

Wager, S., and S. Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". *Journal of the American Statistical Association* doi:10.1080/01621459.2017.1319839.

Wager, S., T. Hastie, and B. Efron. 2014. "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife". *Journal of Machine Learning Research* 15(1):1625–1651.

AUTHOR BIOGRAPHIES

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. His research focuses on Monte Carlo simulation, risk and uncertainty quantification, and stochastic optimization. His email address is khl2114@columbia.edu.

HUAJIE QIAN is a Ph.D. student in the Department of Industrial Engineering and Operations Research at Columbia University. His research interest lies in simulation uncertainty quantification, data-driven simulation analysis, stochastic optimization and control. His email address is h.qian@columbia.edu.