Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

EFFICIENT UNCERTAINTY QUANTIFICATION OF BAGGING VIA THE CHEAP BOOTSTRAP

Arindam Roy Chowdhury[1] and Henry Lam[1]
[1]Dept. of Industrial Eng. and Operations Research, Columbia University, New York, NY, USA

ABSTRACT

Bagging has emerged as an effective tool for reducing variance and enhancing stability in model training, via repeated data resampling followed by a suitable aggregation. Recently, it has also been used to obtain performance bounds for data-driven solutions in stochastic optimization. However, quantifying statistical uncertainty for bagged estimators can be challenging, as standard bootstrap would require resampling at both the bagging and the bootstrap stages—-leading to multiplicative computation costs that can be prohibitively large. In this work, we propose a practical and theoretically justified approach using the *cheap bootstrap* methodology, which enables valid confidence interval construction for bagged estimators under a controllable number of model evaluation. We establish asymptotic validity of our approach and demonstrate its empirical performance through simulation experiments. Our results show that the proposed method achieves nominal coverages with significantly reduced computational burden than other benchmarks.

1 INTRODUCTION

Bagging, or bootstrap aggregation, is a popular approach to reduce variance and enhance stability in estimation or prediction tasks. Its main idea is to repeatedly resample data that are used to build new estimators, and aggregate these estimators via averaging to obtain a final estimate. This approach leads to some widely used machine learning predictors such as random forests (Breiman 2001), where the individual predictors built from the resampled data, often called the base learners, are decision trees. Recently, bagging has also been used in obtaining performance bounds, namely the optimality gaps of a data-driven solution or the optimal values, in stochastic optimization problems (Lam and Qian 2018a; Lam and Qian 2018b; Chen and Woodruff 2023). In this paper, we focus on the latter setting as a showcase of our idea, keeping in mind that the idea can potentially generalize broadly to many bagging-related problems.

To explain bagging more precisely, consider the problem of estimating an unknown quantity ψ using i.i.d. data $\xi_1,\dots,\xi_n \sim P$. Let $\hat{\psi}_n$ denote some estimator computed from the full dataset. To obtain a bagged estimator, we generate D resamples from the original full dataset, where each resample consists of n observations drawn randomly with or without replacement from the dataset, say $\xi_1^{*(d)},\dots,\xi_n^{*(d)}$, for $d \in \{1,\dots,D\}$. For each resample, we compute the corresponding estimate $\hat{\psi}_n^{(d)}$. Then, the bagged estimator is defined as

$$\hat{\psi}_n^{bag} := \frac{1}{D}\sum_{d=1}^{D} \hat{\psi}_n^{(d)}.$$

Intuitively, this estimator is more "smooth" than using the simple estimator $\hat{\psi}_n$ when it has certain "jump" behaviors, which is precisely the case in decision trees (Bühlmann and Yu 2002) or stochastic optimization with multiple nearly optimal solutions (Lam and Qian 2018b). This resulting smoothness consequently propels variance reduction in the estimation.

Our main focus in this paper is the quantification of statistical uncertainty associated with $\hat{\psi}_n^{bag}$, via the construction of statistically valid confidence intervals. To this end, a natural attempt would be to derive a central limit theorem (CLT) and utilize a normality interval. In the case of bagged estimator, the standard error inside such a normality interval would involve the so-called influence function (Efron 1992),

979-8-3315-8726-0/25/$31.00

which represents the gradient information of the estimator with respect to the data distribution, and appears naturally in view of the nonparametric delta method. This culminates in the *infinitesimal jackknife (IJ)* approach (Efron 2014; Wager et al. 2014), which directly approximates the standard error of the bagged estimator via a formula that can reuse the same resamples that construct the base estimator. While this idea is powerful and elegant, a rigorous analysis that pins down precisely the overall computation effort for sufficient-coverage interval construction appears to be still open.

As an alternative, we can consider using the bootstrap on bagged estimators. The bootstrap operates by resampling data and using the resampled estimators to construct intervals. This latter construction can be conducted by taking suitable quantiles of the resample estimators (leading to methods like the basic bootstrap or percentile bootstrap; Efron and Tibshirani (1994), Davison and Hinkley (1997)), or taking the variance of the resample estimators (leading to the standard error bootstrap; Efron (1981)), as well as other variants (such as the studentized bootstrap, e.g., Davison and Hinkley (1997) §2.4). Note that even though bagging includes a bootstrap step, it is used to construct a *point* estimator via averaging the bootstrapped estimators, and bears a conceptually different goal of increasing estimation stability rather than the quantification of uncertainty or confidence interval construction. More importantly, when implementing the bootstrap, we would need to resample the data and refit the estimator many times. Applying to the bagged estimator, this means that we need to first resample the data, and for each resample, we construct a bagged estimator via another layer of resampling. In other words, applying the bootstrap on bagged estimators typically requires a *multiplicative* computation effort due to nested layers of resampling, and hence is computationally expensive.

Our main contribution in this paper is to address the above computation challenge faced by the bootstrap when applying to bagged estimators, by utilizing a recent bootstrap variant called the *cheap bootstrap* (Lam 2022b; Lam 2022a; Lam and Liu 2023; Huang et al. 2023; Ohlendorff et al. 2025). The cheap bootstrap, when applying to standard estimation problems (without bagging), allows the construction of asymptotically exact confidence intervals with a fixed, small number of bootstrap resamples. When applying to bagged estimators, the cheap bootstrap would construct valid intervals with again a fixed, small number of bootstrap resamples in the outer layer of the nested procedure, thus leading to an overall computation effort within a constant order of the number of resamples in the bagging. Compared to the standard application of the bootstrap, this therefore alleviates the multiplicative computation effort from both the outer and inner resampling steps. Compared to IJ, our approach appears to be roughly on par in terms of computation, though we should note that both the analyses on our approach and the IJ appear not fully reflective in some aspects as we will explain in the sequel. In particular, in this work we justify our advantage by analyzing the asymptotic coverage of our approach for a simplified setup of letting the resample size in the bagging procedure to be fixed instead of growing in data size. We nonetheless test and show promising numerical performances when we use a large bagging resample size. Moreover, as discussed earlier, we will focus our discussion on the estimation of the optimality gap in stochastic optimization as an example of bagging estimators in this paper.

2 STOCHASTIC OPTIMIZATION AND BAGGING

2.1 Problem Formulation

Consider the following generic stochastic optimization problem:

$$Z^* = \min_{x \in \mathscr{X}} \left\{ Z(x) = \mathbb{E}_F[h(x,\xi)] \right\}, \tag{1}$$

where $\mathscr{X}$ denotes the decision space, and the random variable $\xi \in \Xi$ is drawn from a distribution F. The operator $\mathbb{E}_F[\cdot]$ represents the expectation under F. We focus on the setting in which the true distribution F is not explicitly known, but instead, we observe a finite sample of independent realizations of ξ, denoted by $\xi_{1:n} = (\xi_1, \ldots, \xi_n)$.

Our goal is to construct confidence intervals for the optimality gap of a data-driven solution of (1). Here, optimality gap of a solution, say $\hat{x}$, refers to $Z(\hat{x}) - Z^*$. It is quite straightforward to see that this problem can be largely reduced to finding an interval for the optimal value Z^*, since $Z(\hat{x})$ can be readily estimated as a sample mean once $\hat{x}$ is obtained. To construct intervals for Z^* and explain how this is related to bagging, we first note that a widely used technique for solving data-driven stochastic optimization problems like the above is *sample average approximation (SAA)* (Shapiro et al. 2009; Kleywegt et al. 2002; Higle and Sen 1991). Given i.i.d. samples $\xi_1, \ldots, \xi_n \sim F$, when running the SAA, the estimated objective takes the form:

$$\hat{Z}_n := \min_{x \in \mathscr{X}} \frac{1}{n} \sum_{i=1}^{n} h(x, \xi_i). \tag{2}$$

A direct application of Jensen's inequality reveals that the SAA objective $\hat{Z}_n$ is optimistically biased, in the sense

$$W_n := \mathbb{E}[\hat{Z}_n] \leq Z^*, \tag{3}$$

where the expectation is taken over the data used to compute $\hat{Z}_n$ defined in (2). This inequality holds under very general conditions, without requiring any regularity assumptions on the loss function $h(\cdot,\cdot)$, and is thus widely used to obtain lower bounds on the true optimal value Z^* (see, e.g., Mak et al. 1999; Shapiro 2003; Glasserman 2004). Since it is also known that under some regularity conditions, $\lim_{n\to\infty} W_n = Z^*$, a consistent lower bound on W_n will also be tight for Z^*. Ideally, this would require letting $n \to \infty$, but to simplify analysis we instead fix a sufficiently large k. Note that the optimistic bias in (3) holds for any k and so by constructing a valid lower confidence bound for W_k, we would still inherit a lower confidence bound for Z^*. Moreover, we should also keep in mind that k could be chosen large enough so that $W_k \approx Z^*$, which is how we set up our test cases in the later numerical experiments.

With its form shown in the left-hand side of (3), the natural empirical estimate of W_k is to repeatedly resample the SAA objective value and take their average. This is precisely *bagging* (Lam and Qian 2018b). In particular, D subsamples of size k are drawn (with or without replacement) from the dataset (where $k < n$ and hence subsamples). For each subsample $d \in \{1, \ldots, D\}$, the corresponding SAA value is computed, and their average defines the bagged estimator:

$$Z_{n,k}^{D} := \frac{1}{D} \sum_{d=1}^{D} Z_k^{(d)} = \frac{1}{D} \sum_{d=1}^{D} \left[\min_{x \in \mathscr{X}} \frac{1}{k} \sum_{i=1}^{k} h(x, \xi_i^{(d)}) \right].$$

where $(\xi_1^{(d)}, \ldots, \xi_k^{(d)})$ denote a single resample of size k from the data. This provides a point estimator for $W_k = \mathbb{E}[\hat{Z}_k]$, which as discussed is a lower bound on Z^* by (3).

2.2 Uncertainty Quantification and Related Existing Approaches

Our goal is to construct a statistically valid lower confidence bound for W_k based on the above bagged estimator. Specifically, we are interested in obtaining a lower confidence bound so that it also serves the same for the true optimal value Z^*. This also serves as an example on studying uncertainty quantification for bagged predictors more generally.

Before we detail our approach, we note that a natural first approach to consider is the classical bootstrap, which uses proper statistics (e.g., quantiles, variances) on the resample counterparts of the original estimator to construct intervals. Note that implementing the resampling in the classical bootstrap typically requires repeatedly running many model evaluations, say B, as a Monte Carlo approximation. In our setting, since the considered estimator is a bagged estimator which itself involves resampling the data, applying the classical bootstrap directly would require an additional layer of resampling on top of the existing one, effectively necessitating a $D \times B$ amount of model evaluations. This nested resampling imposes a significant computational burden, making the classical bootstrap potentially impractical for our purpose.

An elegant tractable alternative proposed in Efron (2014) is the IJ estimator. This method estimates variance by approximating the influence function directly via the nonparametric delta method. It does so by reusing the same bootstrap resamples used in constructing the bagged predictor, thereby avoiding an additional layer of resampling. In particular, they first consider the idealized IJ estimator of variance, denoted $\widehat{V}_{\text{IJ}}^{\infty}$, when $B = \infty$, and a practical de-biased estimator, denoted $\widehat{V}_{\text{IJ}}^{B-U}$, that achieves a mean squared error of order $O(1/n)$ in estimating $\widehat{V}_{\text{IJ}}^{\infty}$ with only $B = \Theta(n)$ bagging resamples. This linear dependence on n represents a significant computational advantage over the naive bootstrap procedure, which requires two nested layers of resampling. However, the translation from the estimation error of $\widehat{V}_{\text{IJ}}^{\infty}$ into the coverage of the confidence intervals appears open. In the special setting of stochastic optimization performance bounds, Lam and Qian (2018b) construct a lower confidence bound that they show is valid for covering W_k as long as $B/n \to \infty$.

In our work, we propose a new approach using the cheap bootstrap to address this issue. Rather than reusing the bagging resamples and approximating the influence function as in the IJ framework, we perform resampling directly on the bagged estimator. While this appears to revert back to the idea and hence limitations of the classical bootstrap, the cheap bootstrap distinctly requires only a small number of bootstrap replications. This idea follows from Lam (2022b), which demonstrates that a number of resamples as small as one suffices to construct valid confidence intervals under suitable regularity conditions. In our current context, we accordingly fix the number of outer resamples at a small value B, and based on each outer resample we compute the bagged predictor using D inner resamples. The resulting interval is constructed using the original bagged estimate and the bootstrapped estimates, as outlined in Algorithm 1. As we will elaborate further in the next section, our approach offers coverage-valid confidence intervals with computational complexity $\Theta(n^{1+\varepsilon})$ for any $\varepsilon > 0$. Roughly speaking, this order is mostly due to the requirement that the number of bagging resamples D must satisfy $D/n \to \infty$, while keeping B fixed at a small number. We note that this is at least on par with the asymptotic computation effort of the IJ-based approach in Lam and Qian (2018b) in the case of stochastic optimization performance bounds. Our numerical results, which we detail in Section 4, further suggest that our method is both practical and effective in capturing uncertainty with controllable computation overhead.

3 THEORETICAL GUARANTEES

To justify the validity of Algorithm 1, we begin by reviewing the cheap bootstrap, but with slight modifications tailored to our setting. We then expand this framework to the bagged SAA estimator described in Section 2.

3.1 The Cheap Bootstrap Confidence Interval

We discuss the cheap bootstrap in the general setting of estimating an unknown quantity ψ that depends on data distribution P. Let

$$\hat{\psi}_n = \hat{\psi}_n(\xi_1, \dots, \xi_n) \tag{4}$$

be an estimator based on i.i.d. data $\xi_1, \dots, \xi_n \sim P$. Denote by $\hat{P}_n$ the empirical distribution of the observed data, and let $\xi_1^{*b}, \dots, \xi_n^{*b} \sim \hat{P}_n$ for $b = 1, \dots, B$, drawn conditionally i.i.d., be the resamples. For each of the B resamples, we compute the corresponding bootstrap estimator:

$$\psi_n^{*b} = \hat{\psi}_n(\xi_1^{*b}, \dots, \xi_n^{*b}), \tag{5}$$

and define the empirical variance:

$$S_{n,B}^2 := \frac{1}{B} \sum_{b=1}^{B} (\psi_n^{*b} - \hat{\psi}_n)^2.$$

Using this variance estimate, the cheap bootstrap confidence interval is given by:

$$I_{n,B} := \left[\hat{\psi}_n \pm t_{B,1-\alpha/2} \cdot S_{n,B}\right], \tag{6}$$

Algorithm 1: Cheap Bootstrap Confidence Interval for W_k

Input: Data $\xi_{1:n}$;
Subsample size k;
Number of bagging resamples D;
Number of bootstrap replicates B;
Confidence level $1-\alpha$
Output: Confidence interval $I_{n,k,B,D}$ for W_k
Compute the bagged estimator $Z_{n,k}^{D}$
for $b=1$ **to** B **do**
  Draw a bootstrap resample $\xi_{1:n}^{*(b)}$ of size n from $\{\xi_1,\dots,\xi_n\}$ (with replacement);
  Compute $Z_{n,k}^{D,*(b)}$ for the resampled data $\xi_{1:n}^{*(b)}$;
Compute empirical variance:

$$S^2 := \frac{1}{B}\sum_{b=1}^{B}\left(Z_{n,k}^{D,*(b)} - Z_{n,k}^{D}\right)^2, \quad \text{where } S^2 = S^2(n,k,D,B) \text{ depends on } n,k,D,B.$$

Construct confidence interval:

$$I_{n,k,B,D} := \left[Z_{n,k}^{D} \pm t_{B,1-\alpha/2}\cdot S\right]$$

where $t_{B,1-\alpha/2}$ is the $1-\alpha/2$ quantile of the Student-t distribution with B degrees of freedom;
return $I_{n,k,B,D}$;
Note: The output interval is denoted by $I_{n,k,B,D}^{V}$ when bagging *with* replacement, and by $I_{n,k,B,D}^{U}$ when bagging *without* replacement is used. The bootstrap resamples are always drawn *with* replacement.

where $t_{B,1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the Student-t distribution with B degrees of freedom.

While (6) may resemble the so-called standard error bootstrap in the classical literature, the main insight is that B does not need to be large, and in fact to attain valid coverage B can be taken to be merely one. To justify such a claim, we begin with an assumption that resembles the one used to justify the classical bootstrap, which is a relaxed version of the formulation in Lam (2022b):

Assumption 1 As $n\to\infty$, assume there exists a normalizing sequence σ_n such that the plug-in estimator $\hat{\psi}_n$ from (4) satisfies a central limit theorem (CLT):

$$\frac{\hat{\psi}_n - \psi}{\sigma_n} \xrightarrow{d} \mathcal{N}(0,1). \tag{7}$$

Additionally, any bootstrap estimator ψ_n^* from (5) for any arbitrary b satisfies a conditional CLT given the data:

$$\frac{\psi_n^* - \hat{\psi}_n}{\sigma_n^*} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{in probability,} \tag{8}$$

where the scaling sequences satisfy $\sigma_n/\sigma_n^* \xrightarrow{p} 1$.

Note that the convergence in (8) is understood in terms of the conditional distribution:

$$P\left(\frac{\psi_n^* - \hat{\psi}_n}{\sigma_n^*} \le x \,\middle|\, \hat{P}_n\right) \xrightarrow{p} \Phi(x) \quad \text{for all } x\in\mathbb{R}.$$

where Φ denotes the standard normal distribution function. Assumption 1 mirrors the usual requirements for bootstrap validity, wherein both the estimator and its bootstrap version satisfy asymptotic normality with a common variance. Here, we allow for distinct normalizing sequences σ_n and σ_n^*, provided their ratio converges to one in probability. Note that σ_n and σ_n^* both can be random and data-dependent, though in our application σ_n would be deterministic. For notational clarity, the usual $\sqrt{n}$ scaling is absorbed into these sequences.

With the above assumption, we have the following guarantee:

Theorem 1 (Asymptotic Validity of Cheap Bootstrap Confidence Intervals) Under Assumption 1, as $n \to \infty$ and for any fixed $B \geq 1$,

$$T_{n,B} := \frac{\hat{\psi}_n - \psi}{S_{n,B}} \xrightarrow{d} t_B$$

where t_B denotes the Student-t distribution with B degree of freedom. Consequently, the cheap bootstrap confidence interval satisfies:

$$\lim_{n\to\infty} \mathbb{P}(\psi \in I_{n,B}) = 1 - \alpha.$$

In summary, the interval $I_{n,B}$ achieves asymptotically exact $(1-\alpha)$-level coverage even when the number of resamples B is kept fixed at a small number. This contrasts with standard bootstrap procedures, which require $B \to \infty$ for validity.

Sketch of Proof. We use the conditional Slutsky theorem (see Section A) to conclude that Assumption 1 also implies:

$$\frac{\psi_n^* - \hat{\psi}_n}{\sigma_n} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{in probability, as } n \to \infty.$$

We then establish the joint weak convergence:

$$\left(\frac{\hat{\psi}_n - \psi}{\sigma_n}, \frac{\psi_n^{*1} - \hat{\psi}_n}{\sigma_n}, \ldots, \frac{\psi_n^{*B} - \hat{\psi}_n}{\sigma_n}\right) \xrightarrow{d} (Z_0, Z_1, \ldots, Z_B) \text{ as } n \to \infty,$$

where $Z_0, Z_1, \ldots, Z_B$ are i.i.d. $\mathcal{N}(0,1)$ random variables. Applying the continuous mapping theorem yields:

$$T_{n,B} = \frac{\frac{\hat{\psi}_n - \psi}{\sigma_n}}{\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\frac{\psi_n^{*b} - \hat{\psi}_n}{\sigma_n}\right)^2}} \xrightarrow{d} \frac{Z_0}{\sqrt{\frac{1}{B}\sum_{b=1}^{B} Z_b^2}} \sim t_B. \quad [\text{ as } n \to \infty \text{ and for any fixed } B \geq 1]$$

□

3.2 Expanding Cheap Bootstrap to Bagged SAA Estimator

We begin by analyzing the setting where the number of bagging resamples $D \to \infty$. In this case, the bagged estimator converges to a fully aggregated form that averages over all possible subsamples of size k. Since the dataset is finite, there are finitely many such subsamples, either with or without replacement. Specifically, there are n^k ordered tuples when sampling with replacement, and $\binom{n}{k}$ unordered subsets when sampling without replacement. These give rise to the well-known V- and U-statistic estimators, respectively. Let us introduce the notation:

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

We denote the limiting forms of the bagged SAA estimator $Z_{n,k}^D$ as follows:

- **With replacement (V-statistic):**

  $$V_{n,k} := \frac{1}{n^k} \sum_{\beta \in \mathscr{B}_k} H_k(\xi_{\beta_1}, \ldots, \xi_{\beta_k}),$$

  where $\mathscr{B}_k$ is the set of all ordered k-tuples (with replacement) from $\{1, \ldots, n\}$.

- **Without replacement (U-statistic):**

$$U_{n,k} := \frac{1}{\binom{n}{k}} \sum_{\beta \in \mathscr{C}_k} H_k(\xi_{\beta_1}, \ldots, \xi_{\beta_k}),$$

where $\mathscr{C}_k$ is the set of all unordered k-subsets (without replacement) from $\{1, \ldots, n\}$.

The fully bagged estimators $U_{n,k}$ and $V_{n,k}$ serve as idealized targets for estimating the quantity W_k. We now define cheap bootstrap confidence intervals using these estimators. For each of the B bootstrap resamples (with B fixed, possibly as small as 1), we compute the fully bagged statistics $U_{n,k}^{*(b)}$ and $V_{n,k}^{*(b)}$ by averaging over all subsets of size k, drawn without and with replacement, respectively, from the resampled data. Let the empirical variances of the resample estimates be:

$$S_U^2 := \frac{1}{B} \sum_{b=1}^{B} \left(U_{n,k}^{*(b)} - U_{n,k} \right)^2, \qquad S_V^2 := \frac{1}{B} \sum_{b=1}^{B} \left(V_{n,k}^{*(b)} - V_{n,k} \right)^2. \tag{9}$$

The corresponding cheap bootstrap confidence intervals for W_k are:

$$I_{n,k,B,\infty}^{U} := \left[U_{n,k} \pm t_{B,1-\alpha/2} \cdot S_U \right], \qquad I_{n,k,B,\infty}^{V} := \left[V_{n,k} \pm t_{B,1-\alpha/2} \cdot S_V \right], \tag{10}$$

where $t_{B,1-\alpha/2}$ is the $1-\alpha/2$ quantile of the Student-t distribution with B degrees of freedom.

To establish the asymptotic validity of these intervals, we next introduce some regularity assumptions:

Assumption 2 We assume a uniform moment bound:

$$\mathbb{E}\left[\sup_{x \in \mathscr{X}} |h(x, \xi)|^3 \right] < \infty.$$

This uniform bound plays a key role in controlling higher-order terms in the central limit approximation.

Assumption 3 Define the function

$$g_k(\xi) := \mathbb{E}[H_k(\xi_1, \ldots, \xi_k) \mid \xi_1 = \xi],$$

i.e., the conditional expectation of the SAA value when the first sample is fixed at ξ. We assume:

$$0 < \mathrm{Var}(g_k(\xi)) < \infty.$$

This assumption ensures a non-degenerate real-valued limiting distribution in our limit analysis.

Our next goal is to verify Assumption 1 for the estimators $U_{n,k}$ and $V_{n,k}$. In particular, the CLT required in (7) has already been established in prior work. We restate the result below, adapted from Theorem 2 of Lam and Qian (2018b):

Theorem 2 Suppose $k \geq 1$ is fixed, and Assumptions 2 and 3 hold. Then, as $n \to \infty$,

$$\sqrt{n}(U_{n,k} - W_k) \xrightarrow{d} \mathscr{N}(0, k^2 \,\mathrm{Var}(g_k(\xi))) \tag{11}$$

$$\sqrt{n}(V_{n,k} - W_k) \xrightarrow{d} \mathscr{N}(0, k^2 \,\mathrm{Var}(g_k(\xi))) \tag{12}$$

where $W_k = \mathbb{E}_{F^k}[H_k(\xi_1, \ldots, \xi_k)]$, and $g_k(\xi) = \mathbb{E}[H_k(\xi_1, \ldots, \xi_k) \mid \xi_1 = \xi]$.

Note that $\frac{k^2 Var(g_k(\xi))}{n}$ serves our purpose of σ_n^2. The original result in Lam and Qian (2018b) assumes a slightly weaker moment condition, namely $\mathbb{E}[\sup_{x \in \mathscr{X}} |h(x, \xi)|^2] < \infty$, to establish the CLT. However, for verifying the full bootstrap validity in Assumption 1, we require a uniform third-moment bound. Therefore, we adopt the stronger condition in Assumption 2 for our analysis.

We next move on to establish the second part of Assumption 1, namely the conditional bootstrap CLT in (8):

Theorem 3 (Bootstrap Normality of Bagged Estimators) Under Assumptions 2 and 3, as $n \to \infty$ and for any fixed k, there exists a normalizing sequence $\sigma^*_{n,k}$ such that

$$\left| \mathbb{P}\left(\frac{U^*_{n,k} - V_{n,k}}{\sigma^*_{n,k}} \leq x \,\middle|\, P_n \right) - \Phi(x) \right| \xrightarrow{p} 0,$$

and the scaling satisfies

$$\frac{n\sigma^{*2}_{n,k}}{k^2 \operatorname{Var}(g_k(\xi))} \xrightarrow{p} 1.$$

Proof. (Sketch of proof) Suppose $\xi_1, \ldots, \xi_n$ comprise a sample obtained from P, and let P_n denote the corresponding empirical distribution function. Let ξ^*_i be i.i.d resamples from the measure P_n and E^* denotes the expectation with respect to this measure. To prove the result, we use the flow of Theorem 12.3 from van der Vaart (1998) using the Hajek Projection of $U^*_{n,k}$, denoted by

$$\mathring{U}^*_{n,k} := V_{n,k} + \frac{k}{n} \sum_{i=1}^{n} [g^*_k(\xi^*_i) - V_{n,k}]$$

where $g^*_k(\xi) := \mathbb{E}^*[H_k(\xi^*_1, \ldots, \xi^*_k) | \xi^*_1 = \xi]$. $\sigma^*_{n,k}$ is chosen to be the variance of $\mathring{U}^*_{n,k}$ conditioned on the data. Observe that $\mathbb{E}^*[\mathring{U}^*_{n,k}] = V_{n,k}$ (and not $U_{n,k}$). This is because the bootstrap resampling is done with replacement. By using property of Hajek projection that $\mathbb{E}^*[(U^*_{n,k} - \mathring{U}^*_{n,k})(\mathring{U}^*_{n,k} - V^*_{n,k})] = 0$, it can be shown that as $n \to \infty$:

$$P\left(\frac{U^*_{n,k} - \mathring{U}^*_{n,k}}{\sigma^*_{n,k}} > \delta \mid P_n \right) \xrightarrow{p} 0 \tag{13}$$

Finally, we use the Berry-Esseen Theorem to conclude:

$$\left| P\left(\frac{\mathring{U}^*_{n,k} - V_{n,k}}{\sigma^*_{n,k}} \leq x \mid P_n \right) - \Phi(x) \right| \leq \frac{C}{\sqrt{n}} f(P_n, k) \xrightarrow{p} 0 \textit{ as } n \to \infty \tag{14}$$

where $f(P_n, k)$ is the ratio of the second and third moments of $g^*_k(\xi^*_i)$, which is essentially a V-statistic of the original sample, and can be shown to be bounded in probability under Assumptions 3 and 2. The conclusion follows by again using the conditional Slutsky on (13) and (14). □

Under mild conditions on the objective function, such as those in Assumptions 2 and 3, it can be shown that as $n \to \infty$, both $\sqrt{n}(U_{n,k} - V_{n,k}) \xrightarrow{p} 0$ and $\sqrt{n}(U^*_{n,k} - V^*_{n,k}) \xrightarrow{p} 0$. Thus, as a consequence of Theorem 3, we obtain the following:

Corollary 4 (Bootstrap Normality Around Population and Empirical Means) Under the conditions of Theorem 3, as $n \to \infty$, we also have:

$$\left| \mathbb{P}\left(\frac{U^*_{n,k} - U_{n,k}}{\sigma^*_{n,k}} \leq x \,\middle|\, P_n \right) - \Phi(x) \right| \xrightarrow{p} 0, \quad \text{and} \quad \left| \mathbb{P}\left(\frac{V^*_{n,k} - V_{n,k}}{\sigma^*_{n,k}} \leq x \,\middle|\, P_n \right) - \Phi(x) \right| \xrightarrow{p} 0.$$

Thus, along with Theorem 2, for any fixed $k \geq 1$, the confidence intervals defined in (10) satisfy:

$$\mathbb{P}(W_k \in I^U_{n,k,B,\infty}) \to 1 - \alpha, \quad \text{and} \quad \mathbb{P}(W_k \in I^V_{n,k,B,\infty}) \to 1 - \alpha, \quad \text{as } n \to \infty.$$

We now turn to the practical case where bagging is performed with a finite number of resamples D. Under certain regularity conditions, the following result has been established in Lam and Qian (2018b):

Lemma 1 (Consistency of Finite-D Bagged Estimators) Let $Z_{n,k}^{D,U}$ and $Z_{n,k}^{D,V}$ denote the finite-D bagged estimators using sampling without replacement and with replacement, respectively. Suppose $\lim_{n\to\infty} D_n/n = \infty$. Then, as $n \to \infty$ we have:

$$Z_{n,k}^{D_n,U} - U_{n,k} = o_p\left(\frac{1}{\sqrt{n}}\right), \qquad Z_{n,k}^{D_n,V} - V_{n,k} = o_p\left(\frac{1}{\sqrt{n}}\right).$$

The following result establishes that the confidence intervals $I_{n,k,B,D}^{U}$ and $I_{n,k,B,D}^{V}$, computed using Algorithm 1, achieve asymptotically valid coverage for W_k.

Theorem 5 (Asymptotic Validity of Finite-D Cheap Bootstrap Intervals) Let $I_{n,k,B,D}^{U}$ and $I_{n,k,B,D}^{V}$ denote the confidence intervals constructed via Algorithm 1, using D_n bagging resamples with $\frac{D_n}{n} \to \infty$ and B outer bootstrap resamples kept fixed. Under Assumptions 2 and 3, and for any fixed $k \geq 1$, we have:

$$\mathbb{P}(W_k \in I_{n,k,B,D_n}^{U}) \to 1-\alpha, \quad \text{and} \quad \mathbb{P}(W_k \in I_{n,k,B,D_n}^{V}) \to 1-\alpha, \quad \text{as } n \to \infty.$$

As a consequence of the inequality $Z^* \geq W_k$, the lower endpoints of the confidence intervals constructed for W_k also serve as valid lower bounds for the true optimal Z^*. Specifically, we have:

$$\mathbb{P}\left(\left(I_{n,k,B,D_n}^{U}\right)_L \leq W_k \leq Z^*\right) \to 1-\alpha, \quad \text{and} \quad \mathbb{P}\left(\left(I_{n,k,B,D_n}^{V}\right)_L \leq W_k \leq Z^*\right) \to 1-\alpha, \quad \text{as } n \to \infty,$$

where $(I)_L$ denotes the lower endpoint of the interval I.

The proof of Theorem 3 relies on the fact that the conditional moments of the Hájek projection are essentially V-statistics of the original sample, thereby utilizing their convergence to their respective expectations. However, evaluating the full V-statistic is not strictly necessary for such results. For example, DiCiccio and Romano (2022) establishes a bootstrap CLT for incomplete U-statistics under the condition that $D_n/n \to \infty$. This insight can be leveraged in our setting, where bagging is performed only over D_n resamples, to argue the validity of Theorem 5 without evaluating the fully bagged estimators.

4 NUMERICAL EXPERIMENTS

In this section, we illustrate the validity of our proposed procedure by constructing confidence intervals on a simple stochastic linear program:

$$\begin{aligned} \min_{x} \quad & \mathbb{E}[-0.05x + (3-2x)\xi] \\ \text{s.t.} \quad & -1 \leq x \leq 1 \end{aligned} \tag{15}$$

where $\xi \sim \mathcal{N}(0,1)$ and $x \in \mathbb{R}$. The objective is to minimize the expected value of a linear function involving ξ, subject to simple box constraints. It is easy to verify that the optimal solution is $x^* = 1$, since the expectation reduces to minimizing a linear function in x. This problem highlights the case where the solution sharply depends on the data. Small fluctuations in the data can lead to abrupt jumps in the estimated optimizer inflating variance, motivating the utility of a bagging approach to stabilize the objective.

As discussed earlier, a key limitation of fixing k is that the target quantity W_k may not perfectly represent the true optimal value Z^*, particularly when k is small. This is illustrated in Figure 1, which displays the behavior of W_k as a function of k. Nonetheless, our simulation results indicate that for any fixed k, the proposed confidence interval achieves approximately 95% coverage for W_k, provided the sample size n is sufficiently large.

For all experiments, we use $D = \lfloor n^{1.1} \rfloor$ as the number of bagging resamples and evaluate performance across different values of the bootstrap size B. We evaluate both the $I_{n,k,B,D}^{V}$ and $I_{n,k,B,D}^{U}$ intervals multiple times and report their empirical coverage of W_k, and their average length as a function of n. We present

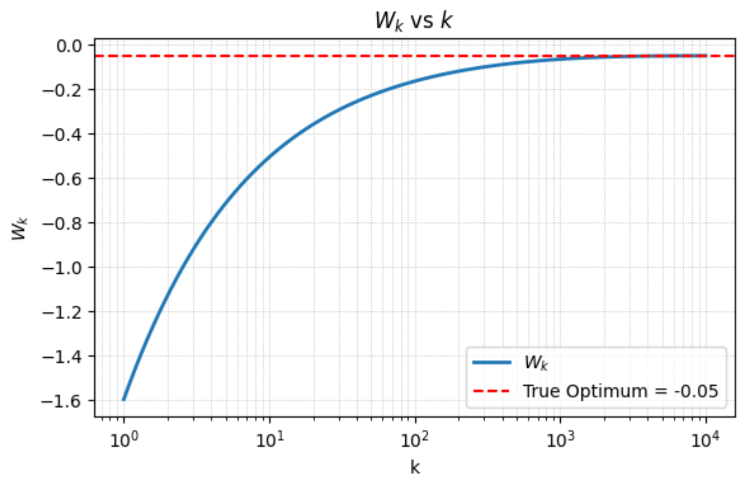

Figure 1: W_k vs. k for the problem (15)

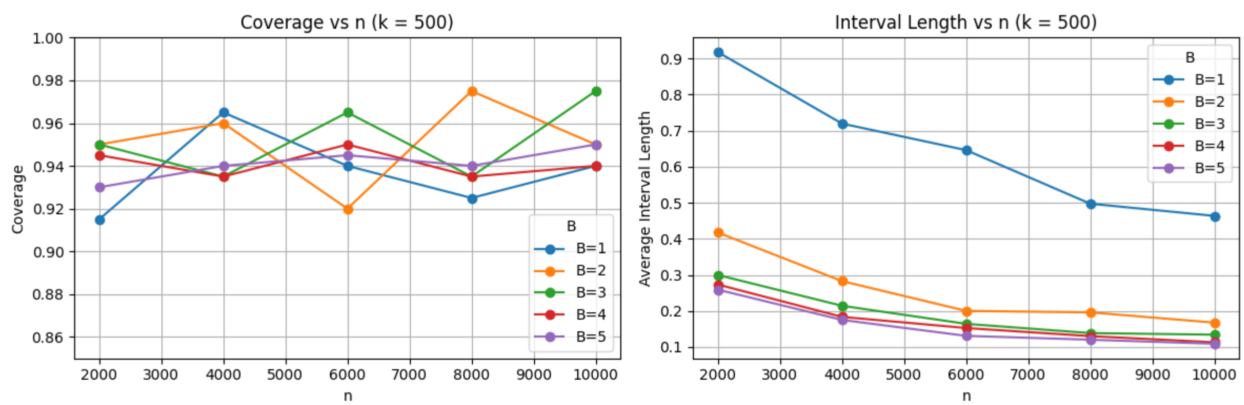

Figure 2: Coverage and average length of $I^V_{n,k,B,D}$ as a function of sample size (n), with $k = 500$.

results for two representative values of k: namely, $k = 500$, for which $W_{500} \approx -0.082$, and $k = 5000$, for which $W_{5000} \approx 0.0509$, in contrast to the true optimal value $Z^* = -0.05$.

Figures 2 and 3 show the behaviors of the interval $I^V_{n,k,B,D}$. We see that all values of B, including as low as $B = 1$, achieve empirical coverage rates close to the nominal 95% level. However, the corresponding interval length is large for $B = 1$, reflecting the increased variability in variance estimation under extremely limited resampling. In contrast, even moderately larger values of B maintain nominal coverage while producing significantly tighter confidence intervals, underscoring the practical effectiveness of our approach. Figures 4 and 5 show similar performances for $I^U_{n,k,B,D}$. These results confirm that the desired coverage can be achieved without requiring $B \to \infty$, validating the efficiency of the cheap bootstrap procedure in producing valid confidence intervals for bagged estimators.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government, Laboratory for AI-Powered Financial Technologies, and the Columbia Innovation Hub Award.

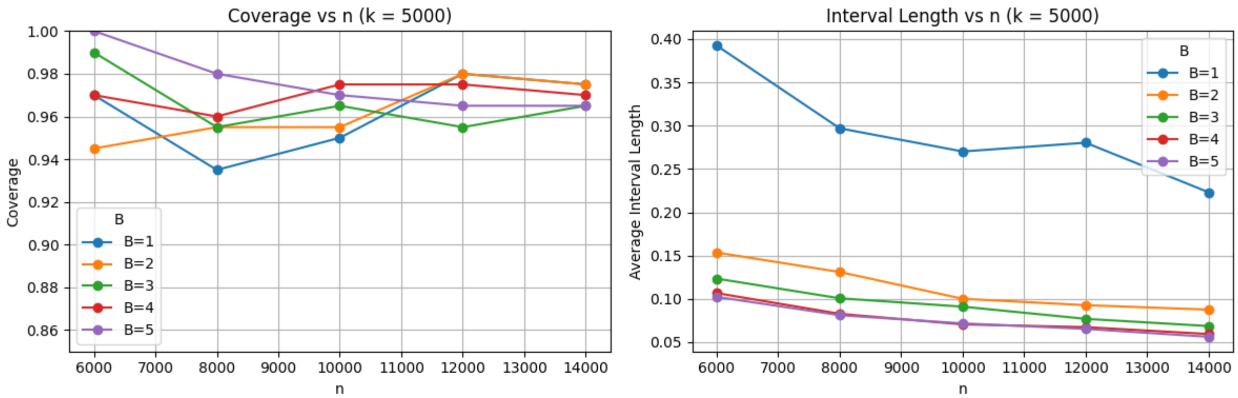

Figure 3: Coverage and average length of $I^V_{n,k,B,D}$ as a function of sample size (n), with $k = 5000$.

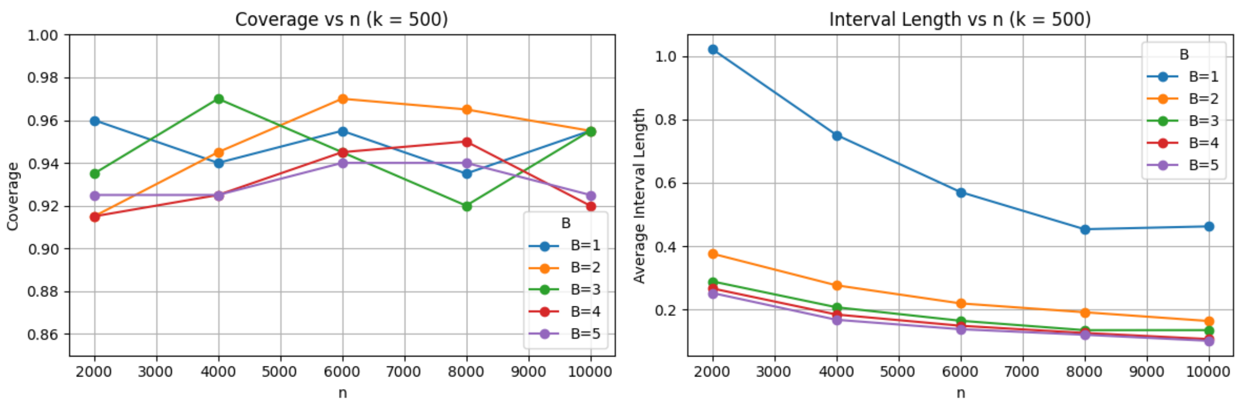

Figure 4: Coverage and average length of $I^U_{n,k,B,D}$ as a function of sample size (n), with $k = 500$.

A APPENDIX

Theorem 6 (Conditional Slutsky) Let X_n, Y_n be real-valued random variables, and let Z_n be an auxiliary sequence of random variables. Suppose $\mathbb{P}(X_n \leq x \mid Z_n) \xrightarrow{p} F(x)$ for all continuity points x of the c.d.f. F, and for any $\delta > 0$, $\mathbb{P}(|Y_n| > \delta \mid Z_n) \xrightarrow{p} 0$. Then, $\mathbb{P}(X_n + Y_n \leq x \mid Z_n) \xrightarrow{p} F(x)$ for all continuity points x of F and for any $\delta > 0$, $\mathbb{P}(|X_n Y_n| > \delta \mid Z_n) \xrightarrow{p} 0$.

REFERENCES

Breiman, L. 2001. "Random Forests". *Machine Learning* 45(1):5–32 https://doi.org/10.1023/a:1010933404324.

Bühlmann, P., and B. Yu. 2002. "Analyzing Bagging". *The Annals of Statistics* 30(4):927–961 https://doi.org/10.1214/aos/1031689014.

Chen, X., and D. L. Woodruff. 2023. "Software for Data-Based Stochastic Programming Using Bootstrap Estimation". *INFORMS Journal on Computing* 35(6):1218–1224.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Number 1 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press https://doi.org/10.1017/CBO9780511802843.

DiCiccio, C., and J. Romano. 2022. "CLT For U-statistics With Growing Dimension". *Statistica Sinica* https://doi.org/10.5705/ss.202020.0048.

Efron, B. 1981. "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods". *Biometrika* 68(3):589–599.

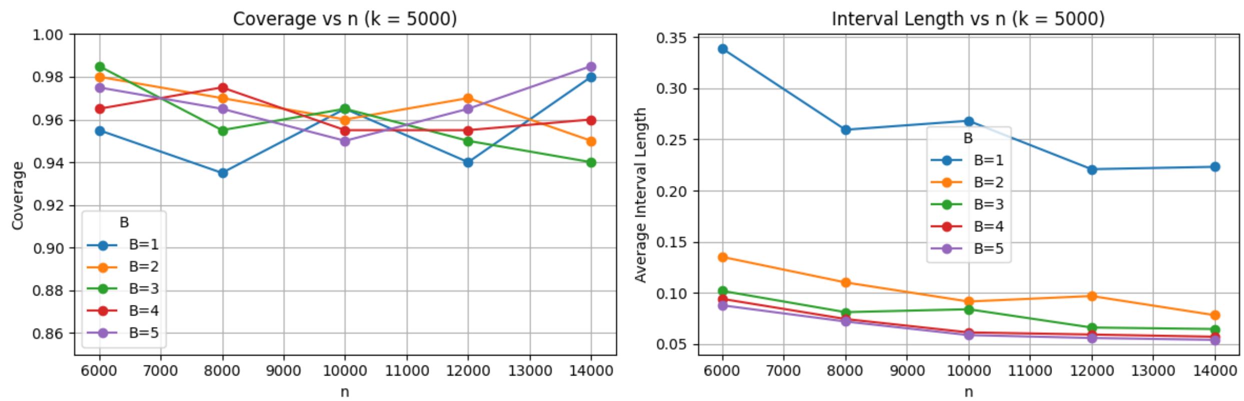

Figure 5: Coverage and average length of $I^U_{n,k,B,D}$ as a function of sample size (n), with $k = 5000$.

Efron, B. 1992, September. "Jackknife-After-Bootstrap Standard Errors and Influence Functions". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 54(1):83–111 https://doi.org/10.1111/j.2517-6161.1992.tb01866.x.

Efron, B. 2014, July. "Estimation and Accuracy After Model Selection". *Journal of the American Statistical Association* 109(507):991–1007 https://doi.org/10.1080/01621459.2013.823775.

Efron, B., and R. Tibshirani. 1994, May. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*, Volume 53. Springer.

Higle, J. L., and S. Sen. 1991, August. "Stochastic Decomposition: An Algorithm for Two-Stage Linear Programs with Recourse". *Mathematics of Operations Research* 16(3):650–669 https://doi.org/10.1287/moor.16.3.650.

Huang, Z., H. Lam, and H. Zhang. 2023. "Efficient Uncertainty Quantification and Reduction for Over-Parameterized Neural Networks". *Advances in Neural Information Processing Systems* 36:64428–64467.

Kleywegt, A. J., A. Shapiro, and T. Homem-de Mello. 2002, January. "The Sample Average Approximation Method for Stochastic Discrete Optimization". *SIAM Journal on Optimization* 12(2):479–502 https://doi.org/10.1137/s1052623499363220.

Lam, H. 2022a. "Cheap Bootstrap for Input Uncertainty Quantification". In *2022 Winter Simulation Conference (WSC)*, 2318–2329. IEEE.

Lam, H. 2022b. "A Cheap Bootstrap Method for Fast Inference". *arXiv preprint arXiv:2202.00090*.

Lam, H., and Z. Liu. 2023. "Bootstrap in high dimension with low computation". In *International Conference on Machine Learning*, 18419–18453. PMLR.

Lam, H., and H. Qian. 2018a. "Assessing Solution Quality in Stochastic Optimization via Bootstrap Aggregating". In *2018 Winter Simulation Conference (WSC)*, 2061–2071. IEEE.

Lam, H., and H. Qian. 2018b. "Bounding Optimality Gap in Stochastic Optimization via Bagging: Statistical Efficiency and Stability". *arXiv preprint arXiv:1810.02905*.

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.

Ohlendorff, J. S., A. Munch, K. K. Sørensen, and T. A. Gerds. 2025. "Cheap Subsampling Bootstrap Confidence Intervals for Fast and Robust Inference in Biostatistics". *arXiv preprint arXiv:2501.10289*.

Shapiro, A. 2003. "Monte Carlo Sampling Methods". *Handbooks in operations research and management science* 10:353–425.

Shapiro, A., D. Dentcheva, and A. Ruszczyński. 2009, January. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics https://doi.org/10.1137/1.9780898718751.

van der Vaart, A. W. 1998, October. *Asymptotic Statistics*. Cambridge University Press https://doi.org/10.1017/cbo9780511802256.

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

AUTHOR BIOGRAPHIES

ARINDAM ROY CHOWDHURY is a PhD student in the Department of Industrial Engineering and Operations Research at Columbia University. His email address is ar4445@columbia.edu.

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. His email address is henry.lam@columbia.edu and his website is http://www.columbia.edu/~khl2114/.