Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

RESAMPLING STOCHASTIC GRADIENT DESCENT CHEAPLY

Henry Lam
Zitong Wang

Industrial Engineering and
Operations Research
Columbia University
500 West 120th Street
New York, NY 10027, USA

ABSTRACT

Stochastic gradient descent (SGD) or stochastic approximation has been widely used in model training and stochastic optimization. While there is a huge literature on analyzing its convergence, inference on the obtained solutions from SGD has only been recently studied, yet is important due to the growing need for uncertainty quantification. We investigate two easily implementable resampling-based methods to construct confidence intervals for SGD solutions. One uses multiple, but few, SGDs in parallel via resampling with replacement from the data, and another operates this in an online fashion. Our methods can be regarded as enhancements of established bootstrap schemes to substantially reduce the computation effort in terms of resampling requirements, while at the same time bypasses the intricate mixing conditions in existing batching methods. We achieve these via a recent cheap bootstrap idea and Berry-Esseen-type bound for SGD.

1 INTRODUCTION

Stochastic optimization commonly arises in many applications across machine learning, operations research, and scientific analysis. The problem can be formulated as:

$$\min_{x\in\mathbb{R}^d} H(x) \triangleq \mathbb{E}_{\zeta\sim P}[h(x,\zeta)] \tag{1}$$

where P is some underlying data distribution governing the randomness $\zeta \in \Omega$, and h is some known real-valued function. Stochastic gradient descent (SGD) or stochastic approximation is a popular numerical approach to solve (1). With an initial guess $x_0 \in \mathbb{R}^d$, SGD iteratively updates the solution using

$$x_{t+1} = x_t - \eta_t \nabla h(x_t, \zeta_{t+1}), \quad t = 0, \ldots, n-1 \tag{2}$$

where ζ_t is a sample drawn using a Monte Carlo model generator or real data. While the Robbins-Monro (Robbins and Monro 1951) procedure outputs x_n after a large number of iterations, one could also take the average $\bar{x}_n \triangleq \frac{1}{n}\sum_{t=1}^{n} x_t$ as the output, which is known as the Polyak-Ruppert-Juditsky averaging (Polyak and Juditsky 1992) or for convenience in this paper we call it averaged stochastic gradient descent (ASGD). Both approaches are common, with ASGD known to be more robust with respect to the step size η_t (Rakhlin et al. 2012).

We are interested in conducting inference, or quantifying the uncertainty, in SGD. That is, we aim to construct, using the iterates (2), a $1-\gamma$ confidence interval for (each component of) the true optimal solution x^* of problem (1). Despite the popularity of SGD, to our best knowledge, this problem has been

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

systematically studied only recently, driven by applications in exploration (Lattimore and Szepesvári 2020) and as stopping criteria (Su and Zhu 2018; Fang et al. 2018; Chen et al. 2020).

1.1 Existing Methods and Challenges

A main challenge in SGD inference is the serial dependence incurred by the sequence $\{x_t\}$, which makes the construction of a consistent standard error estimator intricate. We discuss the ideas and challenges of several recent works that address this issue. Chen et al. (2020) proposed two methods, one based on the delta method that directly approximates the asymptotic covariance of gradient $\nabla h(x,\zeta)$ and Hessian at optimal $\nabla^2 H(x^*)$. However, this requires the computation and storage of the Hessian matrix, which can be computationally demanding. For example, backpropagation can only provide first-order gradient information (Rumelhart et al. 1986), and arguably, a major advantage of SGD lies in its Hessian-free nature. In addition, storing a Hessian matrix requires an expensive $\mathscr{O}(d^2)$ space. These put aside the subtle regularity assumptions needed for consistency as noted by Chen et al. (2020) themselves.

Motivated by these, Chen et al. (2020)'s second method borrows the batch-means idea in stochastic simulation output analysis (Glynn and Iglehart 1990; Schmeiser 1982; Schruben 1983; Glynn and Lam 2018) and Markov Chain Monte Carlo (Geyer 1992; Flegal and Jones 2010; Jones et al. 2006). This approach divides the iterations of SGD into M batches of increasing sizes and aggregates the means of these batches to construct confidence intervals. However, the batch-mean method introduces hyperparameter M, the number of batches, to tune. Additionally, experiments show that this method is more sensitive to the quality of converges of SGD and could underperform other methods. Relatedly, Li et al. (2018) presented a batch-means method for inference in M-estimation by using SGD trajectory with a constant step size. Instead of using batches with increasing lengths, they use batches with a fixed length but separated by gaps to overcome the dependence between iterations of SGD. Zhu and Dong (2021) studied a batch-mean algorithm to construct a d-dimensional confidence region for the optimal solution to problem (1). Their method works by canceling out the asymptotic covariance matrix of the rescaled residue of SGD using an F-type statistic. Additionally, they explore the influence of varying the number and sizes of batches on the efficacy of their algorithm.

Another approach is to use the bootstrap which, advantageously, does not succumb to the computation load of variance estimation nor the tuning and sensitivity of batch sizes. Fang et al. (2018) developed an online bootstrap method that persistently maintains B perturbed version of SGD estimates, updated upon each data arrival. However, B is required to be large for their method to work properly. For linear regression problems of dimensions 10 or 20, they set $B = 200$, which means 200 times more computational cost compared to running the SGD itself or using batch means.

Yet another approach called HiGrad is proposed by Su and Zhu (2018) based on "splitting" an SGD trajectory. More specifically, it first runs SGD for several steps. Then, with the outcome of this thread as a starting point, perform multiple SGD threads using different data, and continue this branching process for each thread's outcome until data are used up. A confidence interval is constructed using all the obtained split outcomes. HiGrad requires a substantial modification to the original SGD runs; in fact, there is no more "original" run of SGD in HiGrad.

Finally, we briefly mention a line of work on quantifying algorithmic randomness, including Lunde et al. (2021) that applied the bootstrap on streaming principal component analysis (Oja 1982), and Chen and Lopes (2020) on randomized Newton methods. Moreover, Nesterov and Vial (2008) gave a complexity bound on the number of iterations of their method in relation to the confidence level on reaching the optimal value via SGD. However, all these works focus on assessing the uncertainty from algorithmic randomness and treat the data as fixed, and hence are less relevant to our focus in this paper.

1.2 Our Contributions

Our discussion above reveals that existing approaches in SGD inference encounter either intricate algorithmic tuning that relates to mixing conditions (batching), substantial modification on the SGD itself (HiGrad), or computation and storage challenges (delta method and online bootstrap). In this paper, we investigate a methodology that resolves these three challenges simultaneously. More precisely, we adopt the bootstrap approach, which does not require mixing-related tuning nor substantial modification to the original SGD. At the same time, we enhance the bootstrap to make it substantially lighter in terms of resampling cost. The latter is made possible by using a recent “cheap bootstrap" idea (Lam 2022b; Lam and Liu 2023; Lam 2022a).

Our methodology can be implemented in both offline and online fashions. The offline version, which we call the *Cheap Offline Bootstrap (COfB)*, reruns the SGD using resampling with replacement from the data B times and constructs confidence intervals from these resampled iterates via an approach similar to the standard error bootstrap. However, while this approach may appear to require heavy resampling effort, our key assertion is that the B in our implementation can be very small (such as 3). In this way, our approach is computationally less demanding than the delta method (Chen et al. 2020) and online bootstrap (Fang et al. 2018), does not require hyperparameter tuning in batch-means (Chen et al. 2020; Zhu and Dong 2021), and also does not substantially modify the SGD trajectory in HiGrad (Su and Zhu 2018).

A caveat of COfB is that we can only rerun SGD after all the data become available. Thus, it cannot be used in a single-pass streaming fashion. To address this, our online version, *Cheap Online Bootstrap (COnB)*, runs multiple $(B+1)$ SGDs in parallel on the fly as new data comes in. COnB borrows the idea of Fang et al. (2018) in perturbing the gradient estimate in the SGD iteration. However, like COfB, it is computationally much cheaper than Fang et al. (2018) as it only needs to maintain a very small number of SGD runs. In both our theory and experimentation, we illustrate that using $B=3$ already produces consistently better coverage than the existing approaches.

Our methodology synthesizes two recent ideas. One, as mentioned earlier, is the recent cheap bootstrap idea. Roughly speaking, instead of using the resemblance of the resample distribution to the sampling distribution as in classical bootstraps, the cheap bootstrap exploits the approximate independence between the resample and original estimates. Coupled with asymptotic normality, this allows to construct a pivotal statistic with an extremely small number (as low as 1) of resample runs B. While the cheap bootstrap gives rise to asymptotically exact-coverage intervals, it also comes with the cost of (arguably fair) larger interval lengths when B is small. Nonetheless, as discussed in Lam (2022b), Lam and Liu (2023), the interval length advantageously shrinks quickly as B increases away from 1. Our second methodological element, which constitutes our main technical development, is to show the asymptotic independence, more precisely a joint central limit theorem, for the original and the resampled SGD runs under resampling with replacement, and how to suitably aggregate the outputs guided by this theorem using the cheap bootstrap logic. To attain this independence, we generalize the recent non-asymptotic bounds for ASGD studied by Shao and Zhang (2022), Anastasiou et al. (2019) to hold uniformly for both the original and resampled runs, under SGD and ASGD settings.

2 METHODOLOGY

Denote the underlying data distribution by P. Let x_{out} be the output of (A)SGD, using step sizes $\eta_t = \eta t^{-\alpha}$ and *i.i.d.* data $\{\zeta_t\}_{t=1}^n$ drew from P. More precisely, in ASGD $x_{\text{out}} = \frac{1}{n}\sum_{t=1}^n x_t$, and in SGD $x_{\text{out}} = x_n$, where x_t is the solution obtained in the t-th iteration of (2). Let $\hat{P}_n$ denote the empirical distribution from data $\{\zeta_t\}_{t=1}^n$, i.e., $\hat{P}_n(\cdot) = \frac{1}{n}\sum_{t=1}^n I(\zeta_t \in \cdot)$, where $I(\cdot)$ denotes the indicator function. We also use $(\cdot)_i$ to denote the i-th entry of a vector and $(\cdot)_{i,j}$ to denote the (i,j)-th entry of a matrix.

Our first method, COfB, works as follows. After obtaining x_{out} with data $\{\zeta_t\}_{t=1}^n$, we repeatedly resample with replacement from the data (i.e., draw n observations from $\hat{P}_n$) and run (A)SGD on the resampled data for B times. Denote the resample outputs by x^{*b}_{COfB}, $b=1,\ldots,B$. As we will see in later discussions, the number of reruns B is not necessarily large as long as it is greater than 2. Then, the $1-\gamma$ confidence interval for the i-th entry of x^* is given by

$$\mathscr{I}_{i,n}^{\text{COfB}}=\left[(x_{\text{out}})_i-t_{B-1,1-\frac{\gamma}{2}}s_i^{\text{COfB}},(x_{\text{out}})_i+t_{B-1,1-\frac{\gamma}{2}}s_i^{\text{COfB}}\right] \tag{3}$$

where $s_i^{\text{COfB}}\triangleq\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left((x_{\text{COfB}}^{*b})_i-(\bar{x}_{\text{COfB}}^{*})_i\right)^2}$, $(\bar{x}_{\text{COfB}}^{*})_i\triangleq\frac{1}{B}\sum_{b=1}^{B}(x_{\text{COfB}}^{*b})_i$, and $t_{B-1,1-\frac{\gamma}{2}}$ denotes the $1-\frac{\gamma}{2}$ quantile of the student-t distribution with degree of freedom $B-1$. A pseudo-code for COfB can be found in Algorithm 1.

Algorithm 1 Cheap Offline Bootstrap (COfB)

```
Input: i.i.d. data $\{\zeta_t\}_{t=1}^n$, number of reruns $B\geq 2$, step size sequence $\{\eta_t\}$, initial guess $x_0$, nominal coverage level $1-\gamma$.
Output: $\mathscr{I}_{i,n}^{\text{COfB}}$, $i=1,\ldots,d$
Run (A)SGD (2) to obtain $x_{\text{out}}$.
for $b\leftarrow[1,2,\ldots,B]$ do
    Resample with replacement from $\{\zeta_t\}_{t=1}^n$ to obtain $\{\zeta_1^{*b},\ldots,\zeta_n^{*b}\}$.
    Run (A)SGD for $n$ steps on $\{\zeta_t^{*b}\}_{t=1}^n$ with initialization $x_0$ to obtain $x_{\text{COfB}}^{*b}$.
end for
for $i\leftarrow[1,2,\ldots,d]$ do
    $(\bar{x}_{\text{COfB}}^{*})_i\leftarrow\frac{1}{B}\sum_{b=1}^{B}(x_{\text{COfB}}^{*b})_i$
    $s_i^{\text{COfB}}\leftarrow\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left((x_{\text{COfB}}^{*b})_i-(\bar{x}_{\text{COfB}}^{*})_i\right)^2}$
    $\mathscr{I}_{i,n}^{\text{COfB}}\leftarrow\left[(x_{\text{out}})_i-t_{B-1,1-\frac{\gamma}{2}}s_i^{\text{COfB}},(x_{\text{out}})_i+t_{B-1,1-\frac{\gamma}{2}}s_i^{\text{COfB}}\right]$
end for
```

Note that COfB is an offline algorithm since resampling from $\{\zeta_i\}_{i=1}^n$ can only be accomplished when all these data points have been obtained. In contrast, our second method, COnB, works by maintaining $B+1$ parallel runs of ASGD starting from the same initialization. One of these trajectories is the original run following exactly (2). The other B trajectories update similarly, except that the gradient estimate $\nabla h(x_t,\zeta_{t+1})$ is perturbed by a factor $W_{t,b}$ following exponential distribution with rate 1. The confidence intervals $\mathscr{I}_{i,n}^{\text{COnB}}$ are constructed similarly as COfB with x_{out} and $\{x_{\text{COnB}}^{*b}\}_{b=1}^n$, except that the standard error term $s_i^{\text{COnB}}\triangleq\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left((x_{\text{COnB}}^{*b})_i-(x_{\text{out}})_i\right)^2}$ has x_{out} instead of $\bar{x}_{\text{COnB}}^{*}$ as the center of the squares. A pseudo-code of COnB is in Algorithm 2. Note that, when a new data ζ_t arrives, COnB uses only $B+1$ gradient calculations to update the original and resampled outputs.

3 MAIN THEORETICAL GUARANTEES

Our main theoretical guarantees on COnB and COfB is on the asymptotic coverage exactness, for B as low as either one or two. To explain and state this result more precisely, Let $H_n(\cdot)=\frac{1}{n}\sum_{i=1}^n h(x,\zeta_i)$ denote the sample average approximation (SAA) of (1) and $\hat{x}_n$ the minimizer of $H_n(\cdot)$. $\|x\|_p$ denotes $(\mathbb{E}[\|x\|^p])^{\frac{1}{p}}$ for a random variable x and $\|\cdot\|$ denotes the standard Euclidean 2-norm for vectors. Let $\mathscr{X}_1$ be a bounded subset of $\mathbb{R}^d$, containing x^* in its interior, and let $\mathscr{X}=\{x|\sup_{y\in\mathscr{X}_1}\|x-y\|\leq\varepsilon_1\}$ for some $\varepsilon_1>0$. For each i,j, define the function classes $\mathscr{F}_{i,j}=\{\partial_{i,j}^2 h(x,\zeta)|x\in\mathscr{X}\}$ and $\tilde{\mathscr{F}}_i=\{(\partial_i h(x_1,\zeta)-\sum_j\partial_{i,j}^2 h(x_2,\zeta)(x_1-x_2))/\|x_1-x_2\||x_1\in\mathscr{X},x_2\in\mathscr{X}_1,x_1\neq x_2\}$. These function classes represent the scopes of the higher-order terms of the Taylor expansion of H at x^*, which are crucial in developing the required asymptotic properties. Let $G(x)=\nabla^2 H(x)$ and $S(x)=\mathbb{E}[\nabla h(x,\zeta)(\nabla h(x,\zeta))^t op]$ be the Hessian of H and covariance matrix of $\nabla h(x,\zeta)$ respectively. Given n data points, define $G_n(x)=\frac{1}{n}\sum_{i=1}^n\nabla^2 h(x,\zeta_i)$ and $S_n(x)=\frac{1}{n}\sum_{i=1}^n\nabla h(x,\zeta_i)(\nabla h(x,\zeta_i))^t op$.

Algorithm 2 Cheap Online Bootstrap (COnB)

```
Input: $i.i.d.$ data $\{\zeta_t\}_{t=1}^n$, number of bootstrap runs $B \geq 1$, step size sequence $\{\eta_t\}$, initial guess $x_0$, nominal coverage level $1-\gamma$.
Output: $\mathscr{I}_{i,n}^{\text{COnB}}$, $i = 1,\dots,d$
for $t \leftarrow [1,2,\dots,n]$ do
    $x_t \leftarrow x_{t-1} - \eta_t \nabla h(x_{t-1},\zeta_t)$
    for $b \leftarrow [1,2,\dots,B]$ do
        Randomly generate $W_{b,t}$ from exponential distribution with rate 1.
        $x_t^{*b} \leftarrow x_{t-1}^{*b} - \eta_t W_{b,t} \nabla h(x_{t-1}^{*b},\zeta_t)$
    end for
end for
$x_{\text{out}} \leftarrow \frac{1}{n}\sum_{t=1}^n x_t$
for $b \leftarrow [1,2,\dots,B]$ do
    $x_{\text{COnB}}^{*b} \leftarrow \frac{1}{n}\sum_{t=1}^n x_t^{*b}$
end for
for $i \leftarrow [1,2,\dots,d]$ do
    $s_i^{\text{COnB}} \leftarrow \sqrt{\frac{1}{B}\sum_{b=1}^B \left((x_{\text{COnB}}^{*b})_i - (x_{\text{out}})_i\right)^2}$
    $\mathscr{I}_{i,n}^{\text{COnB}} \leftarrow \left[(x_{\text{out}})_i - t_{B,1-\frac{\gamma}{2}} s_i^{\text{COnB}}, (x_{\text{out}})_i - t_{B,1-\frac{\gamma}{2}} s_i^{\text{COnB}}\right]$
end for
```

Assumption 1 h and H are twice continuously differentiable in x. And the eigenvalues of $\nabla^2 h(x,\zeta)$ lies in $[l,L]$ for some positive real numbers $0 < l < L$ for all x,ζ.

Assumption 2 The noise of estimated gradient $\{\nabla h(x_{t-1},\zeta_t) - \nabla H(x_{t-1})\}_{t=1}^n$ is $i.i.d.$ with mean 0.

Assumption 1 specifies that the objective function h exhibits strong convexity along with a bounded Hessian, which implies the same property holds for H, in particular its strong convexity. Thus, it guarantees the existence and uniqueness of x^* that satisfies the first-order optimality condition $\nabla H(x^*) = 0$. Assumption 2 stipulates that the evaluation noise in the first-order gradient oracle is unbiased, which is a standard assumption to ensure the convergence of (A)SGD. To establish asymptotic normality, an additional assumption on the variability of $\nabla h(x,\zeta)$ is required:

Assumption 3 There are $\tau_0, \tau > 0$ such that $\|x_0 - x^*\| \leq \tau_0$ and $\|\nabla h(x^*,\zeta)\|_4 \leq \tau$. The eigenvalues of $S(x^*) = \mathbb{E}[\nabla h(x^*,\zeta)(\nabla h(x^*,\zeta))^t op]$ lie in the interval $[\lambda_1,\lambda_2]$ for some positive constants $\lambda_1 < \lambda_2$.

We also need the SAA solution, namely, $\hat{x}_n$, to be consistent in the sense that the difference between $H_n(\hat{x}_n)$ and $H_n(x^*)$ converges to 0 in probability. The following assumption is sufficient for this requirement.

Assumption 4 $\sup_{x\in\mathbb{R}^d} |H_n(x) - H(x)| \xrightarrow{P} 0$.

A further sufficient condition for Assumption 4 is that the function class $\{h(x,\zeta)|x \in \mathbb{R}^d\}$ is Glivenko-Cantelli, which can be implied by Assumption 1 if the space of x is a bounded subset of $\mathbb{R}^d$ (Van der Vaart 2000), though we do not assume the latter here. Essentially, a function class $\mathscr{F}$ is Glivenko-Cantelli if the law of large numbers holds uniformly in functions over $\mathscr{F}$.

The following two assumptions are specialized for ASGD and SGD considered in this work respectively. The specific choice of step size guarantees the convergence of (A)SGD in distribution. The Glivenko-Cantelli assumptions help us analyze the vanishing property of some terms in our analysis of the residual $x_{\text{out}} - x^*$.

Assumption 5 The step size satisfies $\eta_t = \eta t^{-\alpha}$ for some $\alpha \in (\frac{1}{2}, 1]$. For each i,j, function class $\mathscr{F}_{i,j}$ is P-Glivenko-Cantelli.

Assumption 6 The step size is $\eta_t = \eta t^{-1}$, and the initial step size η satisfies $\eta l > \frac{1}{2}$. For each i, j, function classes $\mathscr{F}_{i,j}$ and $\tilde{\mathscr{F}}_i$ are P-Glivenko-Cantelli and $\mathbb{P}(x_t = x^*) = 0,\ \forall t$.

With the above assumptions, we have the following theorem:

Theorem 7 Under Assumptions 1, 2, 3, 4 and 5 for COfB running ASGD, or Assumptions 1, 2, 3, 4 and 6 for COfB running SGD, we have, for any fixed $B \geq 2$, $i = 1, \ldots, d$, the COfB $1-\gamma$ confidence interval for i-th entry is asymptotically exact in the sense

$$\lim_{n\to\infty} \mathbb{P}(x_i^* \in \mathscr{I}_{i,n}^{\text{COfB}}) = 1-\gamma \tag{4}$$

Moreover, under Assumptions 1, 2, 3, 4 and 5 for COnB, we have, for any fixed $B \geq 1$, $i = 1, \ldots, d$, the COnB $1-\gamma$ confidence interval for i-th entry is asymptotically exact in the sense

$$\lim_{n\to\infty} \mathbb{P}(x_i^* \in \mathscr{I}_{i,n}^{\text{COnB}}) = 1-\gamma \tag{5}$$

Theorem 7 states that COfB and COnB attain asymptotically exact coverage as the sample size $n \to \infty$, *regardless of any fixed choice of $B \geq 2$ for COfB and $B \geq 1$ for COnB*. This light computation hinges on our interval construction step at the end that differs from standard bootstraps. Note the subtlety that COfB requires $B \geq 2$, but COnB is valid even for B as small as 1. This discrepancy comes from the slight difference in the joint asymptotic limits among the original and resample (A)SGD runs of COfB and COnB respectively, which will be discussed in Theorem 8 in the following section. One may also notice that COnB works only for ASGD. Whether it will work for SGD is still open to us, as the asymptotic behavior for SGD is actually more delicate.

4 IDEA BEHIND THE MAIN GUARANTEES

We present the development for Theorem 7 in three layers. First is the cheap bootstrap idea that relies on asymptotic independence among original and resample (A)SGD runs. Second is the conversion from conditional convergence of resample estimates given the data, a condition that is widely utilized in classical bootstraps, into asymptotic independence. The third and most challenging step is the development of uniform non-asymptotic bounds to argue this conditional convergence.

4.1 Bootstrap via Asymptotic Independence

We start with the following result on asymptotic independence among original and resample runs.

Theorem 8 Under the same assumptions as Theorem 7, we have

$$\sqrt{n}\begin{pmatrix} x_{\text{out}} - x^* \\ x_{\text{COfB}}^{*1} - \hat{x}_n \\ \vdots \\ x_{\text{COfB}}^{*B} - \hat{x}_n \end{pmatrix} \xrightarrow[n\to\infty]{d} \begin{pmatrix} Z_0 \\ Z_1 \\ \vdots \\ Z_B \end{pmatrix} \tag{6}$$

and

$$\sqrt{n}\begin{pmatrix} x_{\text{out}} - x^* \\ x_{\text{COnB}}^{*1} - x_{\text{out}} \\ \vdots \\ x_{\text{COnB}}^{*B} - x_{\text{out}} \end{pmatrix} \xrightarrow[n\to\infty]{d} \begin{pmatrix} Z_0 \\ Z_1 \\ \vdots \\ Z_B \end{pmatrix} \tag{7}$$

where $Z_b,\ b = 0, \ldots, B$ are *i.i.d.* d-dimensional Gaussian random variables with mean 0. When x_{out} stands for the ASGD output $\frac{1}{n}\sum_{t=1}^{n} x_t$, the covariance matrix of Z_b is $G(x^*)^{-1}S(x^*)G(x^*)^{-1}$, $b = 0, \ldots, B$, where $G(x^*) = \nabla^2 H(x^*)$ and $S(x^*) = \mathbb{E}[\nabla h(x^*, \zeta)(h(x^*, \zeta))^\top]$. When x_{out} stands for the SGD output x_n, consider the singular value decomposition $G(x^*) = QDQ^\top$ with $D = \mathrm{diag}(d_1, \ldots, d_d)$, where $d_1, \ldots, d_d$ are eigenvalues of $G(x^*)$ in decreasing order and Q the matrix consisting of eigenvectors. For $i, j = 1, \ldots, d$ and $b = 0, \ldots, B$, the covariance between $(Z_b)_i$ and $(Z_b)_j$ is given by $\sigma_{i,j}^2 = \eta^2(\eta d_i + \eta d_j - 1)^{-1}(Q^\top S(x^*)Q)_{i,j}$.

Theorem 8 involves two aspects. First, the asymptotic distribution of the error of the resample run compared with the SAA solution in COfB, namely, $x_{\text{COfB}}^{*b} - \hat{x}_n$ (or compared with the original ASGD run in COnB, $x_{\text{COnB}}^{*b} - x_{\text{out}}$) is the same as that of the original run compared with the true minimizer, $x_{\text{out}} - x^*$. Second, more importantly, is the asymptotic independence among all these errors. Derivation of Theorem 8 will be discussed in the later sections. Taking Theorem 8 for granted, the following is a proof of Theorem 7.

Proof of Theorem 7. Consider COfB first. Observe that

$$\frac{(x_{\text{out}})_i - x_i^*}{s_i^{\text{COfB}}} = \frac{\sqrt{n}\left((x_{\text{out}})_i - x_i^*\right)}{\sqrt{n \times (s_i^{\text{COfB}})^2}} = \frac{\sqrt{n}\left((x_{\text{out}})_i - x_i^*\right)}{\sqrt{n \times \frac{\sum_{b=1}^B \left((x_{\text{COfB}}^{*b})_i - (\bar{x}_{\text{COfB}}^*)_i\right)^2}{B-1}}} = \frac{\sqrt{n}\left((x_{\text{out}})_i - x_i^*\right)}{\sqrt{\frac{\sum_{b=1}^B \left(\sqrt{n}\left((x_{\text{COfB}}^{*b})_i - (\hat{x}_n)_i\right) - \sqrt{n}\left((\bar{x}_{\text{COfB}}^*)_i - (\hat{x}_n)_i\right)\right)^2}{B-1}}}$$

As n goes to infinity, we have

$$\frac{\sqrt{n}\left((x_{\text{out}})_i - x_i^*\right)}{\sqrt{\frac{\sum_{b=1}^B \left(\sqrt{n}\left((x_{\text{COfB}}^{*b})_i - (\hat{x}_n)_i\right) - \sqrt{n}\left((\bar{x}_{\text{COfB}}^*)_i - (\hat{x}_n)_i\right)\right)^2}{B-1}}} \xrightarrow{d} \frac{(Z_0)_i}{\sqrt{\frac{\sum_{b=1}^B \left((Z_b)_i - (\bar{Z})_i\right)^2}{B-1}}} \overset{d}{=} \frac{N}{\sqrt{\frac{\chi_{B-1}^2}{B-1}}} \overset{d}{=} t_{B-1}$$

where $\bar{Z} = (1/B)\sum_{b=1}^B Z_b$, and N stands for a standard normal variable, χ_{B-1}^2 a χ^2-variable with $B-1$ degree of freedom, t_{B-1} a student t-variable with $B-1$ degree of freedom, and "$\overset{d}{=}$" equality in distribution. The convergence in distribution above comes from the continuous mapping theorem. The first equality in distribution comes from the *i.i.d.* normality limit in Theorem 8 and the elementary relation between χ^2 and normal. The second equality in distribution comes from the elementary construction of a t variable. Thus, by a pivotal argument, we obtain the confidence interval generated from COfB.

A similar argument works for COnB, except that we use x_{out} directly in place of $\bar{x}_{\text{COfB}}^*$ in the pivotal construction and correspondingly, it would result in student t-distribution with degree of freedom B. □

4.2 From Conditional Convergence to Asymptotic Independence

Our next step is to prove Theorem 8. As a sub-step, the conclusions in Theorem 8 can be implied by a conditional convergence:

Theorem 9 Suppose

$$\sqrt{n}(x_{\text{out}} - x^*) \xrightarrow[n\to\infty]{d} Z_0 \tag{8}$$

In addition, if

$$\sqrt{n}(x_{\text{COfB}}^{*b} - \hat{x}_n^*) \xrightarrow[n\to\infty]{d} Z_0 \ \text{ conditional on } \zeta_1, \zeta_2, \ldots \tag{9}$$

then (6) holds, and if

$$\sqrt{n}(x_{\text{COnB}}^{*b} - x_{\text{out}}) \xrightarrow[n\to\infty]{d} Z_0 \ \text{ conditional on } \zeta_1, \zeta_2, \ldots \tag{10}$$

then (7) holds, where Z_0 denotes a d-dimensional Gaussian random variable with mean 0 and covariance matrix as described in Theorem 8.

Therefore, if we can show (8), (9) and (10), then we obtain the conclusions in Theorem 8 and subsequently the guarantees of COfB and COnB in Theorem 7. Note that (8) is the classical asymptotic normality of (A)SGD guaranteed by our assumptions (Chung 1954; Sacks 1958). On the other hand, the type of conditional convergence in (9) and (10) is the main driver of classical bootstrap methods that allow the approximation of a sampling distribution using the resample counterpart. Here, converting the latter into the conclusion in Theorem 8 presents a new bootstrap methodological route that substantially reduces B. The proof for Theorem 9 generalizes the proof of Proposition 1 in Lam (2022b) and we refer readers to the details therein.

4.3 From Uniform Non-Asymptotic Bound to Conditional Convergence

For COnB, the desired conditional convergence result (10) is well established in the proof of Theorem 1 in Fang et al. (2018). We focus on proving (9) for COfB under both SGD and ASGD settings, which constitutes our main technical development.

Let us abstractize our discussion to denote $\psi(P)$ as the minimizer for (1) with data distribution P, where ψ is viewed as a mapping from the data distribution to $\mathbb{R}^d$. Correspondingly, define ψ_n as the mapping from the data distribution to the outcome of (A)SGD. Then $\psi_n(P) \in \mathbb{R}^d$ is the (random) outcome of (A)SGD after n iterations, as a function of data distribution P with h and $\{\eta_t\}_{t=1}^n$ implicitly chosen.

Classical results (Polyak and Juditsky 1992; Chung 1954) state that the weak limit of $\sqrt{n}(\psi_n(P) - \hat{\psi}(P))$ exists and equals Z_0. This Z_0 is the Gaussian variable described in Theorems 8 and 9 whose variance depends on P. Let $\hat{Z}_m$ denote the normal variable that replaces P in its variance with $\hat{P}_m$, conditional on the collected data. With these new notations, (9) holds if for any Borel measurable set $D \subset \mathbb{R}^d$, we have the following

$$\lim_{n\to\infty} |\mathbb{P}^*(\sqrt{n}\left(\psi_n(\hat{P}_n) - \hat{\psi}(\hat{P}_n)\right) \in D) - \mathbb{P}(Z_0 \in D)| = 0 \quad \text{w.p.1} \tag{11}$$

where $\mathbb{P}^*$ denotes the probability conditional on the data (we will also use $\mathbb{E}^*$ to denote the corresponding conditional expectation). By the triangle inequality, one can obtain

$$\begin{aligned}&|\mathbb{P}^*(\sqrt{n}\left(\psi_n(\hat{P}_n) - \hat{\psi}(\hat{P}_n)\right) \in D) - \mathbb{P}(Z_0 \in D)|\\ \le&|\mathbb{P}^*(\sqrt{n}\left(\psi_n(\hat{P}_n) - \hat{\psi}(\hat{P}_n)\right) \in D) - \mathbb{P}^*(\hat{Z}_n \in D)| + |\mathbb{P}^*(\hat{Z}_n \in D) - \mathbb{P}(Z_0 \in D)|\end{aligned}$$

It can be proved that the second term above vanishes with probability 1. On the other hand, we have the following theorem for the first term:

Theorem 10 Under the same assumptions as in theorem 7 and focusing on COfB, for any measurable set D, we have

$$\lim_{n\to\infty} |\mathbb{P}^*\left(\sqrt{n}(\psi_n(\hat{P}_n) - \hat{\psi}(\hat{P}_n)) \in D\right) - \mathbb{P}^*(\hat{Z}_n \in D)| = 0 \quad \text{w.p.1} \tag{12}$$

The proof invokes an expansive analysis on the behavior of the (A)SGD output. From the iterative scheme (2), one obtains the following

$$x_{n+1} = B_{0n}x_1 - \sum_{m=1}^{n} \eta_m B_{mn}\delta_m - \sum_{m=1}^{n} \eta_m B_{mn}E_m \tag{13}$$

where $\delta_k \triangleq \delta(x_k) = \nabla H(x_k) - G(x^*)(x_k - x^*)$ is the second-order residual of the Taylor expansion of ∇H at x^*, $E_{k-1} = \nabla h(x_{k-1}, \zeta_k) - \nabla H(x_{k-1})$, and $B_{mn} = \prod_{j=m+1}^{n}(I - \eta_j G) \in \mathbb{R}^{d\times d}$.

In the ASGD case, from (13) and using results from Shao and Zhang (2022), we show that there is a 4-tuple $(\hat{\tau}_0, \hat{\tau}, \hat{C}, N)$ such that

$$\sup_{n>N} |\mathbb{P}^*(\sqrt{n}\left(\psi_n(\hat{P}_n) - \hat{\psi}(\hat{P}_n)\right) \in D) - \mathbb{P}^*(\hat{Z}_n \in D)| \le \hat{C}(d^{3/2} + \hat{\tau}^3 + \hat{\tau}_0^3)(d^{1/2}n^{-1/2} + n^{-\alpha+1/2}) \tag{14}$$

for any measurable D. Shao and Zhang (2022) gave an inequality similar to (14) but with a fixed distribution instead of a varying distribution $\hat{P}_n$ depending on n. Additional arguments are required to show that the

Table 1: Average time for different methods for logistic regression.

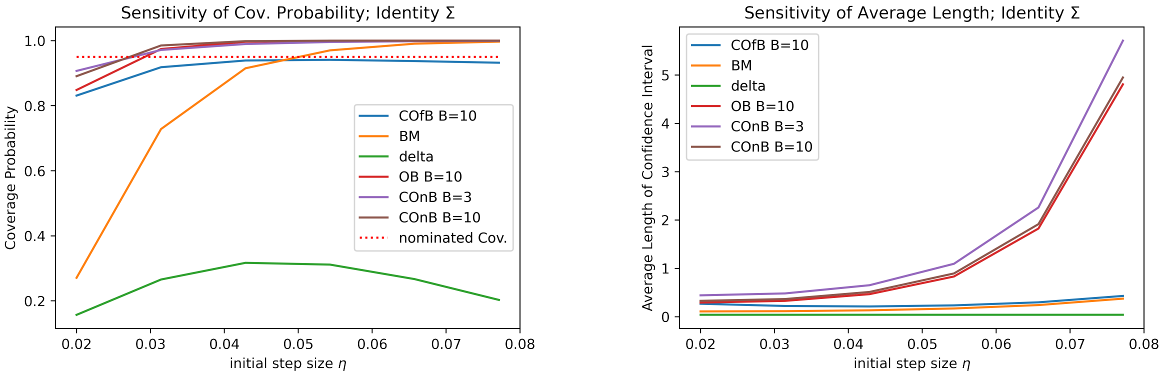
Method	delta	BM	OB ($B = 100$)	HiGrid$_{(2,2)}$	COfB ASGD	COnB
Average Time (s)	10.53	1.10	104.49	0.95	3.15	4.09

established rate is uniform across all data distributions including the empirical distribution. Detailed proof for (14) is omitted here due to space limit.

For the SGD case, the first two terms in (13) correspond to the interaction of the error of the initial solution and second-order residual in the Taylor expansion of ∇H. One can show the following vanishing property

$$\lim_{n\to\infty} \mathbb{E}[\|B_{0n}x_1 - \sum_{m=1}^{n} \eta_m B_{mn}\delta_m\|] = 0$$

On the other hand, the last term $\sum_{m=1}^{n} \eta_m B_{mn} E_m$ consists of the difference between sample gradient ∇h and true gradient ∇H, which converges to a normal distribution. We use the Berry-Esseen-type result from Lemma 4 in Shao and Zhang (2022) to give a non-asymptotic convergence result for this term and thus establish a bound similar to (14) that is uniform across all data distributions including the empirical distribution.



(a)

(b)

Figure 1: Performance of methods with respect to different choices of initial step size η, with sample size $n = 10^4$. We compare the delta method, batch-mean method with $M = n^{0.25}$, COfB method with $B = 10$, COnB with $B = 3, 10$, and the online bootstrap method with $B = 10$, when $n = 10^4$. The left figure shows the sensitivity of the average coverage probability against η; The right figure shows the sensitivity of the average length of confidence interval against η. We report the result for identity Σ.

5 EXPERIMENTS

In this section, we illustrate the numerical performances of our approaches and compare with the other methods. We consider logistic regression, with loss function $h(x,\zeta) = \log(1 + e^{-b\times a^\top x})$. The data $\zeta = (a,b)$ coming from distribution P consists of the independent variable $a \in \mathbb{R}^d$ and dependent variable $b \in \{-1,1\}$. where a follows a multivariate normal distribution with mean 0 and covariance matrix Σ, and $b = 1$ with probability $\frac{1}{1+e^{-a^\top x^*}}$. In this case, we have $\nabla h(x,\zeta) = \frac{-b\times a}{1+e^{b\times a^\top x}}$ and $\nabla^2 h(x,\zeta) = \frac{aa^\top}{(1+e^{a^\top x})(1+e^{-a^\top x})}$. The Hessian information $\nabla^2 h(x,\zeta)$ above will only be used in the delta method.

Baselines. The batch-mean method (BM) splits $\{x_i\}_{i=0}^n$ into M batches, with e_k and s_k denoting the ending index and starting index of the k-th batch respectively. n_k denotes the number of iterates in k-th batch, and the estimator is defined to be $\frac{1}{M}\sum_{k=1}^{M} n_k(\bar{x}_{n_k} - \bar{x}_M)(\bar{x}_{n_k} - \bar{x}_M)^\top$, where $\bar{x}_{n_k} = \frac{1}{n_k}\sum_{i=s_k}^{e_k} x_i$ and

$\bar{x}_M = \frac{1}{e_M - e0}\sum_{i=s_1}^{e_M} x_i$. Let $N = \frac{n^{1-\alpha}}{M+1}$, and e_k to be the closest integer to $((k+1)N)^{\frac{1}{1-\alpha}}$ for each $k = 0, \dots, M$ as suggested in Chen et al. (2020). The confidence interval for each entry of x^* is constructed using diagonal entries of the batch-mean estimator and a normal quantile. The delta method (Chen et al. 2020) generate confidence intervals using normal quantile and $\tilde{\Sigma}_n^2 = \tilde{G}_n^{-1}\tilde{S}_n\tilde{G}_n^{-1}$, where $\tilde{G}_n = \frac{1}{n}\sum_{i=1}^{n}\nabla^2 h(x_{i-1}, \zeta_i)$, and $\tilde{S}_n = \frac{1}{n}\sum_{i=1}^{n}\nabla h(x_{i-1}, \zeta_i)(\nabla h(x_{i-1}, \zeta_i))^\top$ are computed on the fly. The online bootstrap method (OB) (Fang et al. 2018) works by generating B sequences of $i.i.d.$ exponential random variables, $\{W_i^{(b)}\}_{i=1}^{n}$, $b = 1, \dots, B$. And for each b, run ASGD using $\{\zeta_i\}_{i=1}^{n}$ and step sizes $\{W_i^{(b)}\eta_i\}_{i=1}^{n}$. The confidence interval is then constructed using the outcomes and normal quantile. The HiGrad method (Su and Zhu 2018) takes two tuples $(B_1, B_2, \dots, B_K)$ and $(n_0, n_1, \dots, n_K)$ as hyperparameters describing when to break the SGD thread into how many branches. B_k represents the number of branches a single branch divides into at the k-th break, and n_k represents the number of data each thread uses between the k-th and $(k+1)$-th breaks. After all the breaks, there will be $T = \prod_{k=1}^{K} B_k$ threads and one obtain $\{x^{(j)}\}_{j=1}^{t} op$ by averaging each thread. The confidence interval for the i-th entry of x^* is calculated by aggregating $\{(x^{(j)})_i\}_{j=1}^{t} op$.

Hyperparameters. The nominal coverage probability is set to be 95%. We report results for three choices of dimensions, $d = 5, 20, 200$. Two choices of Σ are tested, namely identity ($\Sigma = I_d$) and Toeplitz ($\Sigma_{i,j} = 0.5^{|i-j|}$). For all experiments, the decay rate of step size α is set to be 0.501, so $\eta_t = \eta t^{-0.501}$. We set $x^* = [0, \frac{1}{d-1}, \frac{2}{d-1}, \dots, 1]$ and $x_0 = [0, \dots, 0] \in \mathbb{R}^d$.

For each set of hyperparameters, we run 500 independent trials and report the mean and standard deviation of the coverage probabilities and the average length of the intervals across d dimensions. We tune the initial step size η within the range $[0.2, 0.7]$ and report the result with the most accurate average coverage probability. For the batch-mean method, M is selected to be the nearest integer to $n^{0.25}$ as suggested in Chen et al. (2020). For HiGrad, the architecture we experiment on is $((2,2), (n/7, n/7, n/7))$. We report the performance of COfB and COnB with $B = 3$, and set $B = 200$ for the online bootstrap method unless otherwise specified, as suggested in Fang et al. (2018).

Results. Results for this logistic regression experiment can be found in Table 2. We use **bold** numbers to denote good results in the sense that the coverage probability is between 92% and 98%, and use *italic* numbers to denote bad results with coverage probability less than 80%.

Generally speaking, for any d and Σ discussed in this experiment, the delta method and batch-mean method are outperformed by other methods. For the HiGrad method, we observe a significant drop in performance when $d = 200$, which might be caused by its shorter SGD trajectory compared with other methods. When $d = 200$, the number of data/iterations is inadequate for SGD to converge properly. The performance of the delta method drops significantly as d increases since it requires a Monte-Carlo estimation of a full Hessian matrix, the accuracy of which suffers when the dimension increases, while the benefit of reusing data in bootstrap-type methods is more apparent. For the online bootstrap method, although it gives a comparable coverage probability, our methods are significantly faster. As a trade-off, the average length of the confidence interval of our method is larger compared with the batch-mean method and the delta method, which is due to the t-quantile with degree of freedom $B-1$ (B) in our COfB (COnB). Recall that entries of x^* are linearly spaced in $[0,1]$, and the confidence intervals are of magnitude 10^{-2}. Thus the benefits in computational efficiency appear to significantly outweigh the increase in interval length.

Table 1 presents the computation time for different methods. The experiment was performed on a single processor at 3.4GHz. HiGrad and the batch-mean method are the fastest as they require no extra gradient steps or computing Hessian. Roughly speaking, the computation time is proportional to the number of gradient steps except for the delta method.

Sensitivity Analysis. In Figure 1, we compare the performance of our methods, the delta method, the batch-mean method, and the online bootstrap method with number of samples $n = 10^4$ for the linear regression problem. Observe that the coverage probability of COfB methods remains stable around 95% regardless of changes in the initial step size. On the other hand, the batch-mean method requires a careful choice of the initial step size to give a comparable coverage rate, and the optimal choice is not the same

Table 2: Coverage probabilities and average interval lengths in the logistic regression experiment.

	$d=5$		$d=20$		$d=200$	
	Cov (%)	Len ($\times 10^{-2}$)	Cov (%)	Len ($\times 10^{-2}$)	Cov (%)	Len ($\times 10^{-2}$)
			Identity Σ, $n=10^5$			
delta	**95.00** (0.07)	3.10 (0.00)	**94.12** (0.07)	3.68 (0.00)	*61.92* (0.15)	5.85 (0.00)
BM	89.33 (0.10)	2.64 (0.00)	87.29 (0.11)	3.11 (0.01)	*57.47* (0.16)	5.56 (0.02)
OB	**94.83** (0.07)	3.18 (0.00)	**96.96** (0.05)	4.41 (0.01)	99.91 (0.01)	50.71 (0.23)
HiGrad$_{(2,2)}$	**94.33** (0.07)	5.77 (0.02)	**95.46** (0.07)	7.10 (0.02)	80.61 (0.13)	10.26 (0.04)
COfB$_{ASGD}$	**94.12** (0.07)	4.21 (0.00)	**95.03** (0.07)	7.32 (0.01)	**92.34** (0.08)	19.00 (0.02)
COfB$_{SGD}$	**95.40** (0.07)	9.38 (0.01)	**94.99** (0.07)	9.42 (0.01)	**94.72** (0.07)	25.61 (0.03)
COnB	**94.33** (0.07)	4.71 (0.02)	**95.62** (0.06)	6.56 (0.03)	99.42 (0.02)	75.25 (0.43)
			Toeplitz Σ, $n=10^5$			
delta	**94.83** (0.07)	4.05 (0.00)	**93.29** (0.08)	5.59 (0.00)	*53.69* (0.16)	9.56 (0.00)
BM	84.00 (0.12)	3.16 (0.01)	*75.25* (0.14)	3.75 (0.01)	*34.93* (0.15)	7.30 (0.03)
OB	**95.00** (0.07)	4.24 (0.00)	**94.67** (0.07)	6.70 (0.01)	99.78 (0.01)	69.28 (0.26)
HiGrad$_{(2,2)}$	**95.33** (0.07)	7.18 (0.03)	**93.38** (0.08)	8.92 (0.03)	*57.02* (0.16)	10.27 (0.04)
COfB$_{ASGD}$	**94.12** (0.07)	5.70 (0.00)	**94.77** (0.07)	11.49 (0.01)	**93.79** (0.08)	42.27 (0.05)
COfB$_{SGD}$	**95.36** (0.07)	9.48 (0.01)	**94.80** (0.07)	9.65 (0.01)	**94.41** (0.07)	30.53 (0.03)
COnB	**94.00** (0.08)	6.22 (0.02)	**94.71** (0.07)	10.27 (0.05)	**97.82** (0.05)	99.61 (0.60)

across different problems. The delta method suffers from a huge under-coverage and fails to give a valid confidence interval. The delta method and the batch-mean method have smaller average lengths, which can be associated with their under-coverages. It can be observed that the coverage probability and the average length of the batch-mean estimator both increase as η increases. Our COnB method has a similar sensitivity as the online bootstrap method. Nonetheless, as mentioned earlier, COnB is substantially faster. Additionally, the average length of our COnB becomes almost the same as that of the online bootstrap method when increasing B to 10.

ACKNOWLEDGEMENTS

We gratefully acknowledge support from the InnoHK initiative, the Government of the HKSAR, and Laboratory for AI-Powered Financial Technologies.

REFERENCES

Anastasiou, A., K. Balasubramanian, and M. A. Erdogdu. 2019. “Normal Approximation for Stochastic Gradient Descent via Non-asymptotic Rates of Martingale CLT”. In *Conference on Learning Theory*, 115–137. PMLR.

Chen, J. X., and M. Lopes. 2020, 13–18 Jul. “Estimating the Error of Randomized Newton Methods: A Bootstrap Approach”. In *Proceedings of the 37th International Conference on Machine Learning*, edited by H. D. III and A. Singh, Volume 119 of *Proceedings of Machine Learning Research*, 1649–1659: PMLR.

Chen, X., J. D. Lee, X. T. Tong, and Y. Zhang. 2020. “Statistical Inference for Model Parameters in Stochastic Gradient Descent”. *The Annals of Statistics* 48(1):251–273.

Chung, K. L. 1954. “On a Stochastic Approximation Method”. *The Annals of Mathematical Statistics*:463–483.

Fang, Y., J. Xu, and L. Yang. 2018. “Online Bootstrap Confidence Intervals for the Stochastic Gradient Descent Estimator”. *Journal of Machine Learning Research* 19(78):1–21.

Flegal, J. M., and G. L. Jones. 2010. “Batch Means and Spectral Variance Estimators in Markov Chain Monte Carlo”. *The Annals of Statistics* 38(2):1034–1070.

Geyer, C. J. 1992. “Practical Markov Chain Monte Carlo”. *Statistical Science* 7(4):473–483.

Glynn, P. W., and D. L. Iglehart. 1990. “Simulation Output Analysis using Standardized Time Series”. *Mathematics of Operations Research* 15(1):1–16.

Glynn, P. W., and H. Lam. 2018. “Constructing Simulation Output Intervals under Input Uncertainty via Data Sectioning”. In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1551–1562: Institute of Electrical and Electronics Engineers, Inc.

Jones, G. L., M. Haran, B. S. Caffo, and R. Neath. 2006. “Fixed-width Output Analysis for Markov Chain Monte Carlo”. *Journal of the American Statistical Association* 101(476):1537–1547.

Lam, H. 2022a. “Cheap Bootstrap for Input Uncertainty Quantification”. In *Proceedings of the 2022 Winter Simulation Conference*, edited by B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. G. Corlu, L. H. Lee, E. P. Chew, T. Roeder, and P. Lendermann, 2318–2329: Institute of Electrical and Electronics Engineers, Inc.

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”. *International Conference on Machine Learning (ICML)*.

Lattimore, T., and C. Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press.

Li, T., L. Liu, A. Kyrillidis, and C. Caramanis. 2018, Apr.. “Statistical Inference Using SGD”. *Proceedings of the AAAI Conference on Artificial Intelligence* 32(1).

Lunde, R., P. Sarkar, and R. Ward. 2021. “Bootstrapping the Error of Oja’s Algorithm”. *Advances in Neural Information Processing Systems* 34:6240–6252.

Nesterov, Y., and J. P. Vial. 2008, jun. “Confidence Level Solutions for Stochastic Programming”. *Automatica* 44(6):1559–1568.

Oja, E. 1982. “Simplified Neuron Model as a Principal Component Analyzer”. *Journal of Mathematical Biology* 15(3):267–273.

Polyak, B. T., and A. B. Juditsky. 1992. “Acceleration of Stochastic Approximation by Averaging”. *SIAM Journal on Control and Optimization* 30(4):838–855.

Rakhlin, A., O. Shamir, and K. Sridharan. 2012. “Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization”. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, 1571–1578.

Robbins, H., and S. Monro. 1951. “A Stochastic Approximation Method”. *The Annals of Mathematical Statistics* 22(3):400–407.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. “Learning Representations by Back-propagating Errors”. *Nature* 323(6088):533–536.

Sacks, J. 1958. “Asymptotic Distribution of Stochastic Approximation Procedures”. *The Annals of Mathematical Statistics* 29(2):373–405.

Schmeiser, B. 1982. “Batch Size Effects in the Analysis of Simulation Output”. *Operations Research* 30(3):556–568.

Schruben, L. 1983, dec. “Confidence Interval Estimation Using Standardized Time Series”. *Operations Research* 31(6):1090–1108.

Shao, Q.-M., and Z.-S. Zhang. 2022. “Berry–Esseen Bounds for Multivariate Nonlinear Statistics with Applications to M-estimators and Stochastic Gradient Descent Algorithms”. *Bernoulli* 28(3):1548–1576.

Su, W. J., and Y. Zhu. 2018. “Uncertainty Quantification for Online Learning and Stochastic Approximation via Hierarchical Incremental Gradient Descent”. *arXiv preprint arXiv:1802.04876*.

Van der Vaart, A. W. 2000. *Asymptotic Statistics*, Volume 3. Cambridge university press.

Zhu, Y., and J. Dong. 2021. “On Constructing Confidence Region for Model Parameters in Stochastic Gradient Descent Via Batch Means”. In *Proceedings of the 2021 Winter Simulation Conference*, edited by S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, 1–12: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. His research interests include Monte Carlo methods, uncertainty quantification, risk analysis and data-driven optimization. He serves as the area editor for Stochastic Models and Data Science in Operations Research Letters, and on the editorial boards of Operations Research, INFORMS Journal on Computing, Applied Probability Journals, Stochastic Models, Manufacturing and Service Operations Management, and Queueing Systems. His email address is henry.lam@columbia.edu and his website is http://www.columbia.edu/~khl2114/.

ZITONG WANG is a Ph.D. student of Industrial Engineering and Operations Research at Columbia University. His primary research interests are stochastic optimization and uncertainty quantification. His email address is zw2690@columbia.edu.