Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

ON CONSTRUCTING CONFIDENCE REGION FOR MODEL PARAMETERS IN STOCHASTIC GRADIENT DESCENT VIA BATCH MEANS

Yi Zhu

WeRide Corp
2630 Orchard Pkwy
San Jose, CA 95134

Jing Dong

Graduate School of Business
Columbia University
New York, NY 10027, USA

ABSTRACT

We study an easy-to-implement algorithm to construct asymptotically valid confidence regions for model parameters in stochastic gradient descent. The main idea is to cancel out the covariance matrix which is hard/costly to estimate using the batch means method with a fixed number of batches. In developing the algorithm, we establish a process-level functional central limit theorem for Polyak-Ruppert averaging iterates. We also extend the batch means method to accommodate more general batch size specifications.

1 INTRODUCTION

Stochastic Gradient Descent (SGD) and variants of it have been widely used in model-parameter estimation in either online learning or when data sizes are very large (Robbins and Monro 1951; Polyak and Juditsky 1992). As the estimators we construct via SGD is not exact, it is desirable to quantify the estimation errors incurred. While there is a rich literature studying the convergence rate of SGD algorithms (see, e.g., Zhang 2004; Nemirovski, Juditsky, Lan, and Shapiro 2009; Agarwal, Bartlett, Ravikumar, and Wainwright 2012), much less is known about the uncertainty quantification for the true model parameters (see, however, Hsieh and Glynn 2002; Toulis and Airoldi 2017; Chen, Lee, Tong, and Zhang 2020; Su and Zhu 2018). In this paper, we propose a simple procedure to construct asymptotically valid confidence regions for model parameters based on a cancellation method called the batch means with a fixed number of batches.

We consider the setting where the model parameters, $x^* \in \mathbb{R}^d$, can be characterized as the minimizer of a convex objective function, which is also known as the loss function. Specifically,

$$x^* = \arg\min \left(H(x) := \mathbb{E}[h(x,\zeta)]\right), \tag{1}$$

where h is a real-valued function and ζ is a d'-dimensional random vector. Stochastic gradient descent is an iterative algorithm to solve (1). In its simplest form, the t-th iteration takes the form

$$X_t = X_{t-1} - \gamma_t \nabla_x h(X_{t-1}, \zeta_t),$$

where $\nabla_x h$ is the gradient of h with respect to x and γ_t is the step size. If we use $\bar{X}_t = t^{-1}\sum_{i=0}^{t-1} X_i$ to estimate x^*, then under certain regularity conditions, the paper (Polyak and Juditsky 1992) establishes that

$$t^{1/2}(\bar{X}_t - x^*) \Rightarrow N(0,\Sigma) \text{ as } t \to \infty,$$

where $\Rightarrow$ denotes convergence in distribution and $N(0,\Sigma)$ denotes a Gaussian random vector with mean 0 and covariance matrix Σ. Here, $\Sigma = \nabla^2 H(x^*)^{-1} U \nabla^2 H(x^*)^{-1}$, where $\nabla^2 H(x^*)$ is the Hessian of H at x^*, and $U = \mathbb{E}[\nabla_x h(x^*,\zeta)\nabla_x h(x^*,\zeta)^T]$. If we know the value of Σ, then a classic way to construct the 95% confidence region for x^* is

$$\hat{R}_t = \{x \in \mathbb{R}^d : t(\bar{X}_t - x)^T \Sigma^{-1} (\bar{X}_t - x) \le \chi^2_{d,0.05}\},$$

978-1-6654-3311-2/21/$31.00 ©2021 IEEE

where $\chi^2_{d,0.05}$ is the 95%-quantile of the chi-squared distribution with d degrees of freedom. The confidence region $\hat{R}_t$ is asymptotically valid in the sense that $\lim_{t\to\infty}\mathbb{P}(x^* \in \hat{R}_t) = 0.95$.

The key challenge here is that the covariance matrix Σ is often unknown in practice and can be very costly to estimate consistently. To address the challenge, we adopt the cancellation-based batch means method from the stochastic simulation literature (Schruben 1983; Glynn and Iglehart 1990). In particular, our work is closely related to the works that use the cancellation methods to conduct multivariate output analysis (Yang and Nelson 1992; Munoz and Glynn 2001). These methods were developed for steady-state estimation problems. The main idea is to construct the statistics in a special way that cancels out the unknown covariance matrix. Despite the elegance of the batch means idea, existing results in the literature do not allow us to apply it directly to stochastic gradient descent. This is because in steady-state estimation problems, we require the stochastic process to be time-homogeneous. However, the transition kernel of $\{X_t : t \geq 0\}$ in SGD is time-varying due to the decreasing step sizes. The main contribution of this paper is that we rigorously establish the validity of the batch means method in the SGD setting. This provides a simple way to construct asymptotically valid confidence regions for model parameters in SGD. The method utilizes the SGD iterates directly and does not require any modification to the underlying algorithm. We also extend the batch means method to allow more general batch size specifications. Our result relies on establishing appropriate process-level convergence of $(\bar{X}_t : t \geq 0)$, which is stronger than the large sample convergence results established in the literature.

As discussed above, the main advantage of cancellation-based batch means method is to avoid estimating Σ directly. Note that constructing consistent estimator of Σ can be quite challenging (Chen, Lee, Tong, and Zhang 2020). Recently, the paper (Chen, Lee, Tong, and Zhang 2020) develops a consistent estimator of Σ using the idea of batching. To achieve consistency, it requires the number of batches to go to infinity as number of SGD iterates increases. The cancellation method we considered here keeps the number of batches fixed, which is more desirable when the number of iterates is relatively small and/or the dimension of the parameters is relatively high. This is because the validity of the batching idea relies on having the batch means close enough to independent Gaussian random vectors. With a fixed number of iterates, when requiring a large number of batches, the size of each batch tends to be small, which could render the batch means being far from Gaussian.

2 BATCH MEANS METHODS

Consider the case where $H(x)$ is strongly convex with a unique minimizer at x^*. We apply Polyak-Ruppert averaging. In particular, the iteration takes the form

$$X_t = X_{t-1} - \gamma_t \mathscr{G}(X_{t-1}, \zeta_t), \tag{2}$$

where $\mathbb{E}[\mathscr{G}(X_{t-1},\zeta_t)|X_{t-1}] = \nabla H(X_{t-1})$ and $\gamma_t = at^{-r}$ for some $a > 0$ and $r \in (1/2, 1)$. The batch means method divides the sample path $\{X_t : 0 \leq t \leq T\}$ into m non-overlapping batches, where the i-th batch is of size $b_i := \lceil Tw_i \rceil$. Let $\tau_i = \sum_{j=1}^{i} b_j$. Then, the ith batch contains iterates $\{X_{\tau_{i-1}+1}, \ldots, X_{\tau_i}\}$ and its batch mean is defined as

$$\Xi_i = \frac{1}{b_i} \sum_{t=\tau_{i-1}+1}^{\tau_i} X_t.$$

The basic idea of the batch means method is that for T large enough, Ξ_i's are approximately independent $N(x^*, (1/b_i)\Sigma)$. Then, we can consider an F type of statistic:

$$\Gamma_T = m(m-d)(d(m-1))^{-1}(\bar{X}_T - x^*)^T (S_m(T))^{-1}(\bar{X}_T - x^*) \tag{3}$$

where

$$\bar{X}_T := \frac{1}{T}\sum_{t=1}^{T} X_t \quad \text{and} \quad S_m(T) := \frac{1}{m-1}\sum_{i=1}^{m}(\Xi_i - \bar{X}_T)(\Xi_i - \bar{X}_T)^T.$$

Based on the form of Γ_T, the unknown Σ which appears in both $\bar{X}_T - x^*$ and $S_m(T)$ will cancel out in the limit (as $T \to \infty$). The actual procedure to construct confidence regions for x^* is summarized in the following Algorithm.

Input: The SGD sample path $\{X_t : 0 \le t \le T\}$, the number of batches $m \ge d$, the relative batch length parameter w

Find the appropriate scaling parameter $\alpha_m(\delta, w)$ according to Theorem 1.

Calculate the batch means Ξ_i for $i = 1, 2, \ldots, m$.

Calculate $\bar{X}_T$ and $S_m(T)$

Output: $R_T = \left\{ x \in \mathbb{R}^d : \frac{m(m-d)}{d(m-1)} (\bar{X}_T - x)^T S_m^{-1}(T) (\bar{X}_T - x) \le \alpha_m(\delta, w) \right\}$.

Algorithm 1: Construct a $100(1-\delta)\%$ confidence region for x^*

To apply Algorithm 1, we need to specify the parameters m and w. We require $m \in \mathbb{Z}_+$ with $m > d$ and $w = (w_1, \ldots, w_m) \in \mathbb{R}_+^m$ with $\sum_{i=1}^m w_i = 1$, where $\mathbb{Z}_+$ is the set of strictly positive integers and $\mathbb{R}_+$ is the set of strictly positive real numbers. The confidence region constructed in Algorithm 1 is asymptotically valid in the sense that if the scaling parameter $\alpha_m(\delta, w)$ is properly chosen, $\lim_{T\to\infty} \mathbb{P}(x^* \in R_T) = 1 - \delta$. Then, the key is to calibrate the scaling parameter $\alpha_m(\delta, w)$. The value of $\alpha_m(\delta, w)$ is determined by the asymptotic distribution of Γ_T, which is characterized in Theorem 1. Before we present the theorem, we first introduce some assumptions which are standard for the convergence analysis of Polyak-Ruppert averaging (see, e.g., (Chen, Lee, Tong, and Zhang 2020; Polyak and Juditsky 1992)). Define $\Delta_t := X_t - x^*$ and

$$\xi_t = (\xi_t(1), \ldots, \xi_t(d)) := \mathscr{G}(X_{t-1}, \zeta_t) - \nabla H(X_{t-1}). \tag{4}$$

Assumption 1 (1) $H(x)$ is differentiable and strongly convex with parameter $C > 0$, i.e., for any x and y $H(y) \ge H(x) + \nabla H(x)^T (y - x) + \frac{C}{2} \|y - x\|^2$. (2) $\nabla H(x)$ is Lipschitz continuous with parameter $L > 0$, i.e., for any x and y, $\|\nabla H(x) - \nabla H(y)\| \le L \|x - y\|$. (3) There exists $C_2 > 0$, such that $\|\nabla H(x) - \nabla^2 H(x^*)(x - x^*)\| \le C_2 \|x - x^*\|^2$. (4) $\nabla^2 H(x^*)$ exists.

Assumption 2 $(\xi_t : t \ge 1)$ are martingale differences with respect to the filtration $\mathscr{F} = \{\mathscr{F}\}_{t \ge 1}$ generated by $(\zeta_t : t \ge 1)$, and it satisfies the following two conditions:
(1) The conditional covariance of ξ_t has an expansion around x^*: $\mathbb{E}[\xi_t \xi_t^T | \mathscr{F}_{t-1}] = U + r(\Delta_{t-1})$, for some positive definite matrix U, and there exist constants $S_1, S_2 > 0$, such that for any $\Delta \in \mathbb{R}^d$, $\|r(\Delta)\| \le S_1 \|\Delta\| + S_2 \|\Delta\|^2$.
(2) There exists $M \in (0, \infty)$, such that $\|\xi_t\| \le M$ almost surely, $\forall t \ge 1$.

Assumption 1 ensures that $\bar{X}_t$ converges to a unique global optimal x^* (Polyak and Juditsky 1992). Assumption 2 provides sufficient conditions to establish the functional Central Limit Theorem (FCLT) for partial sums of ξ_t's.

Define $g_m : \mathscr{C}^d[0,1] \times \mathbb{R}^m \to \mathbb{R}^{d \times d}$ as

$$g_m(x, w) = \frac{1}{m-1} \sum_{i=1}^m \left(\frac{x(c_i) - x(c_{i-1})}{c_i - c_{i-1}} - x(1) \right) \left(\frac{x(c_i) - x(c_{i-1})}{c_i - c_{i-1}} - x(1) \right)^T,$$

with $c_0 = 0$ and $c_i = c_{i-1} + w_i$.

Theorem 1 Under Assumptions 1 and 2, for Γ_T defined in (3) with $m > d$ and $w \in \mathbb{R}_+^m$,

$$\Gamma_T \Rightarrow m(m-d)(d(m-1))^{-1} Z^T g_m(B, w)^{-1} Z \text{ as } T \to \infty,$$

where Z is a standard d-dimensional Gaussian random vector, B is a standard d-dimensional Brownian motion (BM), and Z is independent of $g_m(B, w)$. Furthermore, if we set $\alpha_m(\delta, w)$ as the $(1-\delta)$-quantile of $m(m-d)(d(m-1))^{-1} Z^T g_m(B, w)^{-1} Z$, then

$$\lim_{T\to\infty} \mathbb{P}(x^* \in R_T) = 1 - \delta.$$

The proof of Theorem 1 is delayed until Section 5. We note from Theorem 1 that the scaling parameter $\alpha_m(\delta, w)$ does not depend on the underline problem instances, $(X_t : 0 \le t \le T)$. It only depends on the batch means parameters m and w. In the special case of evenly-split batches, i.e., $w_i = 1/m$,

$$m(m-d)(d(m-1))^{-1} Z^T h_m(B,w)^{-1} Z \stackrel{d}{=} F_{d,m-d},$$

where $F_{d,m-d}$ denote the F distribution with d and $m-d$ degrees of freedom. We will discuss a different splitting scheme in Section 3.

3 SELECTION OF THE BATCH MEANS PARAMETERS

The confidence region constructed using the batch means method is asymptotically valid regardless of our choice of m and w, as long as $m > d$ and $w \in \mathbb{R}_+^m$. However, different values of m and w will affect the pre-limit performance of the procedure. In this section, we study how to choose these parameters to achieve good pre-limit performance. The analysis is divided into two parts. We first study how to choose the batch sizes w with a fixed value of m. We then study how to choose m. The key intuition is that the pre-limit performance of the procedure is largely determined by how close the distribution of $\left((b_1/\sqrt{T})(\Xi_1 - x^*), \ldots, (b_m/\sqrt{T})(\Xi_m - x^*)\right)$ is to the distribution of $(G(B(c_1) - B(c_0)), \ldots, G(B(c_m) - B(c_{m-1})))$, where B is a d-dimensional BM.

3.1 Batch Size

Note that the pre-limit Ξ_i's are correlated while the limiting $(B(c_i) - B(c_{i-1}))$'s are uncorrelated. Thus, one important quantity we want to minimize is the correlation between Ξ_i and Ξ_{i+1}. To understand the correlation between Ξ_i and Ξ_{i+1}, we follow the arguments in (Chen, Lee, Tong, and Zhang 2020). First, note that for t large enough, X_t is close to x^*. Thus,

$$\nabla H(X_{t-1}) \approx \nabla H(x^*) + \nabla^2 H(x^*)(X_{t-1} - x^*) = A\Delta_{t-1},$$

where $A := \nabla^2 H(x^*)$ and recall that $\Delta_t = X_t - x^*$. The equality follows as $\nabla H(x^*) = 0$. Next, by the recursion formula (2), we have

$$\Delta_t \approx (I - \gamma_t A)\Delta_{t-1} + \gamma_t \xi_t,$$

where I is the identity matrix and ξ_t is defined in (4). This further indicates that for i and j large enough, the correlation between Δ_i and Δ_j is approximately $\prod_{t=i}^{j-1} ||I - \gamma_t A|| \approx \exp\left(-\lambda(A)\sum_{t=i}^{j-1}\gamma_t\right)$, where $\lambda(A)$ denotes the smallest eigenvalue of A. With the goal of balancing the correlation between Ξ_i and Ξ_{i+1}, we can choose w according to $\min_w \max_i \exp\left(-\lambda(A)\sum_{t=\tau_{i-1}}^{\tau_i}\gamma_t\right)$. It is easy to see that the minimum is achieved when $\sum_{t=\tau_{i-1}+1}^{\tau_i}\gamma_t$'s are equal. In this case, we can set

$$\tau_i = (i/m)^{1/(1-r)}\, T.$$

Note that for this specification of τ_i's, the batch sizes are gradually increasing, i.e., w_i is increasing in i. In what follows, we shall refer to this batch-size specification as the "increasing batch size" (InBS).

Table 1 provides some of the commonly used scaling parameters for InBS with different values of d and m. The quantiles are estimated using Monte Carlo simulation. We generate enough samples such that half-width of the corresponding 95% confidence interval is less than 0.01.

We next show some numerical experiments comparing the performance of three different batch size specifications: i) InBS, ii) even splitting (ES), and iii) decreasing batch size (DeBS) where we reverse the order of the batch size specification of InBS. Table 2 summarizes results.

For Table 2 and the subsequent numerical experiments, we consider two classes of examples: linear regression and logistic regression. For linear regression, $b_i = x^{*T} a_i + \varepsilon_i$ where a_i's and ε_i's are i.i.d. $N(0,1)$. In this case, $\zeta = (a,b)$ and $h(x,\zeta) = (b - x^T a)^2$. For logistic regression, $b_i \in \{-1,1\}$ with

Table 1: 95%-quantile of $\frac{m(m-d)}{d(m-1)} Z^T g_m(B,w)^{-1} Z$ with InBS allocation

d	1	2	3	4
$m=10$	2.93	2.92	3.13	3.50
$m=20$	2.18	2.00	1.95	1.97
$m=30$	1.91	1.71	1.64	1.62
$m=40$	1.76	1.55	1.47	1.50

d	10	50	80	100
$m=20$	2.46	NA	NA	NA
$m=60$	1.24	2.49	NA	NA
$m=100$	1.02	1.31	1.81	NA
$m=120$	0.97	1.18	1.43	1.81

$\mathbb{P}(b_i = 1|a_i) = (1+\exp(-x^{*T}a_i))^{-1}$. In this case $\zeta = (a,b)$ and $h(x,\zeta) = \log(1+\exp(-bx^T a))$. When not specified, the true parameters x^* is a d-dimensional vector linearly spaced between 0 and 1. We set the baseline number of iterations at $n := 10^5$. In all the experiments, our goal is to achieve a 95% coverage rate. The estimated coverage rate is based on 1000 independent replications of the procedure. We also report the corresponding 95% confidence interval for the coverage rate.

We observe from Table 2 that as the number of iterations increases, all three batch size specifications are approaching the target coverage rate, 0.95. For a small number of iterations, InBS and ES achieve a higher coverage rate than DeBS, and InBS performs slightly better than ES. In practice, when T is large, we suggest using ES for the ease of implementation. When T is small, we suggest using InBS.

Table 2: Coverage rate comparison for different batch size specifications

	n	$4n$	$7n$	$10n$
Linear regression with $d=2$				
InBS	0.975 ± 0.009	0.955 ± 0.013	0.970 ± 0.010	0.971 ± 0.009
ES	0.938 ± 0.015	0.947 ± 0.014	0.951 ± 0.013	0.950 ± 0.013
DeBS	0.787 ± 0.025	0.878 ± 0.020	0.909 ± 0.017	0.912 ± 0.019
Logistic regression with $d=2$				
InBS	0.934 ± 0.015	0.932 ± 0.015	0.946 ± 0.014	0.948 ± 0.013
ES	0.899 ± 0.018	0.917 ± 0.018	0.934 ± 0.015	0.933 ± 0.015
DeBS	0.842 ± 0.023	0.908 ± 0.017	0.932 ± 0.018	0.930 ± 0.015

3.2 Number of Batches

We next look into different choices of m for $m > d$. We divide the analysis into two parts. We first analyze the limiting volume of the confidence region for different values of m. We then analyze the pre-limit performance. The volume of the confidence region, which is a d-dimensional ellipsoid, takes the form

$$V_d(m,w) := \left(\frac{d(m-1)}{m(m-d)}\right)^{d/2} \det(S_m(T)^{1/2}) \alpha_m(\delta,w)^{d/2} q_d,$$

where $q_d = \pi^{d/2}/\Gamma(d/2+1)$ is the volume of a d-dimensional unit sphere and Γ denotes the Gamma function. From Theorem 2, we have

$$\det((TS_m(T))^{1/2}) \Rightarrow \det(G)^2 \det(h_m(B,w)^{1/2}) \text{ as } T \to \infty.$$

Then, we can compare the limiting volume of the confidence region by comparing

$$v_d(m,w) := \left(\frac{d(m-1)}{m(m-d)}\right)^{d/2} \mathbb{E}[\det(h_m(B,w)^{1/2})] \alpha_m(\delta,w)^{d/2}$$

for different values of m. We show one such comparison in Figure 1. We observe in Figure 1 that as m increases, $v_d(m,w)$ decreases. However, there is a diminishing decreasing effect. On the other hand, for a

finite number of iterations, the larger m is, the smaller the sizes of the batches are. Then, the batch means are further from their asymptotic Gaussian distributions. In Table 3, we compare the pre-limit performance for different values of m. We use InBS for the batch size specification and focus on a relatively small number of iterations. We observe that when the numbers of iterations are small, large values of m can lead to substantial under-coverage. In practice, we suggest setting m between 15 and 30 when $d < 10$, and m between $d+5$ and $d+10$ when $d > 10$.

Figure 1: Compare $v_d(m,w)$ for different values of m and d.

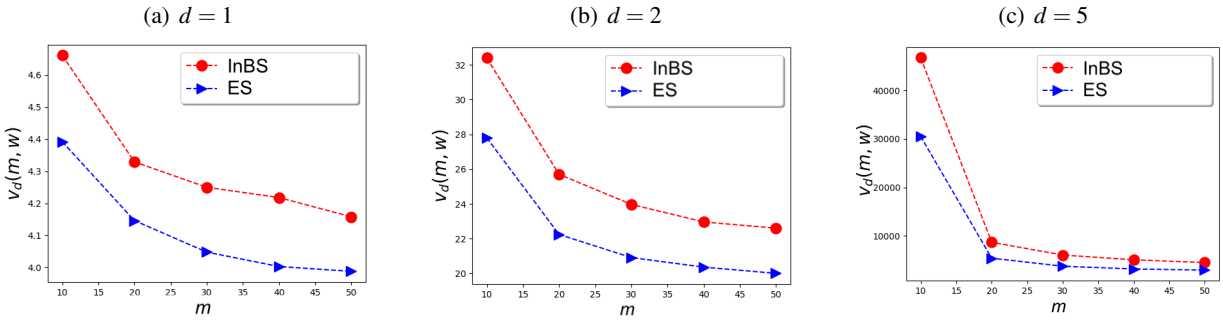

Table 3: Coverage comparison for different values of m, logistic regression with $d = 3$

	$0.1n$	$0.4n$	$0.7n$	n
$m = 10$	0.913 ± 0.017	0.933 ± 0.015	0.947 ± 0.013	0.933 ± 0.015
$m = 20$	0.814 ± 0.024	0.897 ± 0.018	0.919 ± 0.017	0.927 ± 0.016
$m = 30$	0.730 ± 0.027	0.876 ± 0.020	0.909 ± 0.017	0.906 ± 0.018
$m = 40$	0.615 ± 0.030	0.817 ± 0.024	0.845 ± 0.022	0.883 ± 0.019

4 COMPARISON TO OTHER METHODS

In this section, we compare our batch means method to several existing state-of-the-art methods. In particular, we consider two recently developed statistical inference methods for model parameters in SGD: the batch means method with an increasing number of batches (BMI) developed in (Chen, Lee, Tong, and Zhang 2020) and the hierarchical incremental gradient descent method (HiGrad) developed in (Su and Zhu 2018). These methods have been demonstrated to achieve superior performance (in terms of the accuracy and the computational complexity) over other existing methods. We also consider the sectioning method (Hsieh and Glynn 2002), which is similar to the batch means method, but instead of dividing a single sample path into m batches, it generates m independent paths of equal length. This method can also be viewed as a special case of HiGrad with only 1 level.

BMI is mainly designed to draw marginal inference, i.e., it constructs confidence intervals for each parameter (dimension) separately. Thus, it does not require $m \geq d+1$. However, Lemma 3 in Section 5 indicates that when $m \leq d$, the estimated covariance matrix $S_m(T)$ is likely to be degenerate. HiGrad has versions for both marginal inference and joint inference. Comparing to our method, HiGrad has a lot more parameters to be specified (e.g., the tree structure and the partition of the data set) and requires modification to the original SGD procedure. The sectioning method has the advantage that estimators constructed based on different sections are independent. Thus, the asymptotic independence requirement is automatically satisfied. However, with a fixed budget, the length of each path (section) in the sectioning method is only $1/m$ the length of the path in the batch means. If we have a limited amount of computational budget,

focusing on a single long run (as in the batch means) instead of multiple shorter runs (as in the sectioning method) may get us closer to x^*.

In Tables 4 and 5 we compare the finite sample coverage rate of our batch means method (BM) and other benchmark methods for logistic regression examples. For BM, we set $m = 30$ and use InBS for batch sizes. When conducting joint inference using BMI, we set the marginal confidence level to be $1 - 0.05/d$ based on the Bonferroni correction. For HiGrad, we use a two-layer tree structure with 5 and 6 nodes for the two layers respectively. When doing marginal inference using BM, we can construct the batch means confidence interval for each parameter (dimension) separately. In Table 4, we show the coverage rates of confidence regions (joint inference) constructed using different methods and with different sample sizes (number of iterations). In Table 5, we show results for confidence intervals (marginal inference). The reported coverage rate in Table 5 is the average coverage rate over the d parameters. We observe that BM achieves superior coverage rate in all cases. We also note that the coverage rate deteriorates as the dimension of the problem, d, increases. Lastly, since all the methods are asymptotically valid, we expect all of them to achieve the target coverage rate when the number of iterates is large enough.

Table 4: Joint coverage rate comparison for different methods: logistic regression

	n	$4n$	$7n$	$10n$
$d = 2$				
BM	0.919 ± 0.017	0.942 ± 0.013	0.936 ± 0.015	0.945 ± 0.014
BMI	0.890 ± 0.019	0.919 ± 0.017	0.897 ± 0.018	0.899 ± 0.018
HiGrad	0.833 ± 0.023	0.879 ± 0.020	0.901 ± 0.018	0.913 ± 0.017
Sectioning	0.659 ± 0.029	0.807 ± 0.024	0.842 ± 0.023	0.859 ± 0.021
$d = 20$				
BM	0.638 ± 0.029	0.847 ± 0.020	0.878 ± 0.020	0.900 ± 0.018
BMI	0.537 ± 0.030	0.642 ± 0.031	0.680 ± 0.029	0.698 ± 0.028
HiGrad	0.090 ± 0.017	0.427 ± 0.030	0.510 ± 0.029	0.570 ± 0.028
Sectioning	0.024 ± 0.009	0.226 ± 0.026	0.311 ± 0.028	0.384 ± 0.030

Table 5: Marginal coverage rate comparison: logistic regression

	n	$4n$	$7n$	$10n$
$d = 2$				
BM	0.938 ± 0.015	0.949 ± 0.014	0.945 ± 0.014	0.953 ± 0.013
BMI	0.905 ± 0.018	0.920 ± 0.017	0.927 ± 0.016	0.932 ± 0.015
HiGrad	0.860 ± 0.020	0.903 ± 0.018	0.913 ± 0.017	0.915 ± 0.017
Sectioning	0.757 ± 0.026	0.851 ± 0.020	0.872 ± 0.020	0.880 ± 0.020
$d = 20$				
BM	0.901 ± 0.019	0.937 ± 0.015	0.945 ± 0.014	0.953 ± 0.013
BMI	0.835 ± 0.023	0.861 ± 0.021	0.860 ± 0.029	0.866 ± 0.021
HiGrad	0.457 ± 0.030	0.610 ± 0.029	0.631 ± 0.031	0.650 ± 0.029
Sectioning	0.367 ± 0.030	0.535 ± 0.031	0.564 ± 0.031	0.580 ± 0.030

5 PROOF OF THE MAIN RESULT

The proof of Theorem 1 involves two main steps. The first step establishes the process level convergence of $\bar{X}_t$ (Theorem 2). The second step shows that $g_m(B,w)$ is positive definite almost surely (Lemma 3).

5.1 Process level convergence of $\bar{X}_t$

For the first step, we start by presenting two auxiliary lemmas. The first lemma extends the Azuma-Hoeffding inequality to the multidimensional setting. Its proof follows similar lines of arguments as Theorem 1.8 in (Hayes 2005) and is thus omitted here.

Lemma 1 Let $\mathscr{M}$ be a martingale in $\mathbb{R}^d$ with $\mathscr{M}_0 = 0$, and for every n, the martingale difference $\mathscr{M}_n - \mathscr{M}_{n-1}$ satisfies $\|\mathscr{M}_n - \mathscr{M}_{n-1}\| \le \sigma_n \le 1/2$. Then for any $a > 1$,

$$\mathbb{P}(\|\mathscr{M}_n\| \ge a) \le 2\exp\left(1 - (a-1)^2/(\sum_{i=1}^{n} 2\sigma_i^2)\right).$$

The second lemma characterizes the convergence rate of an important term in the SGD iterrates. It tightens the bound established in (Polyak and Juditsky 1992). Let

$$\bar{\beta}_s^t := \gamma_s \sum_{i=s}^{t-1} \prod_{k=s+1}^{i} \left(I - \gamma_k \nabla^2 H(x^*)\right) \text{ and } \phi_s^t = \bar{\beta}_s^t - \left(\nabla^2 H(x^*)\right)^{-1},$$

where $\prod_{k=s+1}^{s} \left(I - \gamma_k \nabla^2 H(x^*)\right) \equiv I$.

Lemma 2 For $\gamma_t = at^{-r}$ with some $a > 0$ and $1/2 < r < 1$, $\sum_{s=0}^{t-1} \|\phi_s^t\| = O(t^r)$.

Proof. We first summarize some useful results from (Polyak and Juditsky 1992). Let $\beta_s^s = I$ and $\beta_s^{t+1} = \beta_s^t(I - \gamma_t \nabla^2 H(x^*))$ for $t \ge s$. There exist $\lambda, K > 0$, such that for any $s \ge 0$ and $t \ge s$,

$$\|\beta_s^t\| \le K\exp\left(-\lambda \sum_{i=s}^{t-1} \gamma_i\right), \text{ where } \textstyle\sum_{i=s}^{s-1} \gamma_i \equiv 0. \tag{5}$$

Let $S_s^t = \sum_{i=s}^{t-1} (\gamma_s - \gamma_i)\beta_s^i$. It can be shown that $\phi_s^t = S_s^t - \left(\nabla^2 H(x^*)\right)^{-1} \beta_s^t$. Let $m_s^i = \sum_{k=s}^{i} \gamma_k$. Then,

$$\sum_{i=s}^{t-1} m_s^i \exp(-\lambda m_s^i) = O(1/\gamma_s) \quad \text{and} \quad \|\bar{\beta}_s^t\| \le K.$$

Second, note that $\|\phi_s^t\| \le \|S_s^t\| + \|(\nabla^2 H(x^*))^{-1}\|\|\beta_s^t\|$. We next establish bounds for $\|S_s^t\|$ and $\|\beta_s^t\|$ respectively. First,

$$\begin{aligned}
\|S_j^t\| = \left\| \sum_{i=j+1}^{t-1} \left(\sum_{k=j+1}^{i} (\gamma_{k-1} - \gamma_k) \right) \beta_j^i \right\| &= \left\| \sum_{i=j+1}^{t-1} \left(\sum_{k=j+1}^{i} (\gamma_{k-1} - \gamma_k)\gamma_{k-1}(\gamma_{k-1})^{-1} \right) \beta_j^i \right\| \\
&\le K(\gamma_j - \gamma_{j+1})(\gamma_j)^{-1} \sum_{i=j}^{t-1} m_j^i \exp(-\lambda m_j^i) \\
&\le K(\gamma_j - \gamma_{j+1})/\gamma_j^2 \quad \text{for } j \text{ large enough.}
\end{aligned}$$

By L'Hospital's rule, $(\gamma_j - \gamma_{j+1})/\gamma_j^2 = O(j^{-(1-r)})$. Thus, for t large enough,

$$\sum_{j=0}^{t-1} \|S_j^t\| \le K \sum_{j=0}^{t-1} j^{-(1-r)} \le K \int_0^t x^{-(1-r)} dx = O(t^r). \tag{6}$$

Lastly, note that

$$\sum_{j=0}^{t-1} \|\beta_j^t\| \le K \sum_{j=0}^{t-1} \exp(-\lambda(t-j)\gamma_t) \le \frac{K}{1 - \exp(-\lambda\gamma_t)} = O(\gamma_t^{-1}) = O(t^r). \tag{7}$$

Combining (6) and (7), we have $\sum_{j=0}^{t-1} \|\phi_j^t\| \le \sum_{j=0}^{t} \|S_j^t\| + \left\|\nabla^2 H(x^*)\right\|^{-1} \left(\sum_{j=0}^{t} \|\beta_j^t\|\right) = O(t^r)$. □

Theorem 2 Under Assumptions 1 and 2, there exists a matrix G, such that

$$n^{1/2}t(\bar{X}_{nt} - x^*) \Rightarrow GB(t) \text{ in } D(0,\infty) \text{ as } n \to \infty,$$

where $D(0,\infty)$ denotes the space of right continuous functions with left limit endowed with Skorokhod J_1 topology.

Proof. We start by summarizing some useful results from (Polyak and Juditsky 1992). We first note that $\bar{X}_t$ has the following decomposition:

$$\bar{X}_t - x^* = J^{(0)}(t) + J^{(1)}(t) + J^{(2)}(t) + J^{(3)}(t), \quad \text{where } J^{(0)}(t) = -\frac{1}{t}\sum_{s=0}^{t-1} \beta_0^s \Delta_0,$$

$$J^{(1)}(t) = -\frac{1}{t}\sum_{s=0}^{t-1} \bar{\beta}_s^t \left(\nabla H(X_s) - \nabla^2 H(x^*)\Delta_s\right), \; J^{(2)}(t) = \frac{1}{t}\sum_{s=0}^{t-1} \left(\nabla^2 H(x^*)\right)^{-1} \xi_s, \; J^{(3)}(t) = \frac{1}{t}\sum_{s=0}^{t-1} \phi_s^t \xi_s.$$

Recall that $\xi_t = \mathscr{G}(X_{t-1}, \zeta_t) - \nabla H(X_{t-1})$ and $\Delta_t = X_t - x^*$. We have the following properties for terms in the decomposition: P1) $t^{-1/2}\sum_{i=1}^{t-1} \|\Delta_i\|^2 \to 0$ almost surely (a.s.) as $t \to \infty$. P2) $\|\phi_s^t\| \le K$ for some $K \in (0,\infty)$. P3) $\sum_{s=1}^{t} \|\phi_s^t\| = O(t^r)$. We comment that P3 is not provided in (Polyak and Juditsky 1992). We establish it in Lemma 2.

We are now ready to establish the functional level convergence results for each part in the decomposition. For $J^{(0)}$, from (5), we have $tn^{1/2}J^{(0)}(nt) \to 0$ in $D(0,\infty)$ as $n \to \infty$. For $J^{(1)}$, we have

$$\begin{aligned}
\sup_{0\le t\le T} \|tn^{1/2}J^{(1)}(nt)\| &\le \sup_{0\le t\le T} n^{-1/2} \sum_{s=1}^{nt-1} \|((\nabla^2 H(x^*))^{-1} + \phi_s^t)(\nabla H(X_s) - \nabla^2 H(x^*)\Delta_s)\| \\
&\le (\|\nabla^2 H(x^*)\|^{-1} + K) C_2 \sup_{0\le t\le T} n^{-1/2} \sum_{i=1}^{nt-1} \|\Delta_i\|^2 \text{ by P2 and Assumption 1} \\
&\le C_2(\|\nabla^2 H(x^*)\|^{-1} + K) T^{1/2} ((nT)^{1/2})^{-1} \sum_{i=1}^{nT-1} \|\Delta_i\|^2 \to 0 \text{ a.s. as } n \to \infty \text{ by P1.}
\end{aligned}$$

Thus, $tn^{1/2}J^{(1)}(nt) \Rightarrow 0$ in $D(0,\infty)$ as $n \to \infty$.

For $J^{(2)}$, let $\mathscr{M}_n(t) := n^{-1/2}\sum_{s=1}^{nt} \xi_s$. We next establish FCLT for $\mathscr{M}_n$: there exists a matrix U such that

$$\mathscr{M}_n(t) \Rightarrow UB(t) \text{ in } D(0,\infty) \text{ as } n \to \infty. \tag{8}$$

Under Assumption 2, ξ_t's are Martingale differences. From Theorem 8.1 in (Pang, Talreja, and Whitt 2007), to establish (8), we only need to verify the following two conditions:

C1) For each $t > 0$, $\lim_{n\to\infty} E[\mathscr{J}(\mathscr{M}_n, t)] = 0$, where $\mathscr{J}$ is the maximum jump function, i.e. $\mathscr{J}(x,t) := \sup\{\|x(s) - x(s-)\| : 0 < s \le t\}$.

C2) For each (i,j), $1 \le i,j \le d$, there exists a constant U_{ij}, such that $[\mathscr{M}_{n,i}, \mathscr{M}_{n,j}](t) \Rightarrow U_{ij}t$ as $n \to \infty$, where $M_{n,i}$ denotes i-th entry of $\mathscr{M}_n$, and $[\mathscr{M}_{n,i}, \mathscr{M}_{n,j}]$ is the square-bracket process.

For C1), under the boundedness condition of the Martingale differences (Assumption 2),

$$\mathbb{E}[\mathscr{J}(\mathscr{M}_n, T)] = \mathbb{E}\left[n^{-1/2} \sup_{0<s\le Tn} \|\xi_s\|\right] \le M/n^{1/2} \to 0 \text{ as } n \to \infty.$$

For C2), we have

$$[\mathscr{M}_{ni}, \mathscr{M}_{nj}](t) = \frac{t}{nt}\sum_{s=1}^{nt} \xi_{si}\xi_{sj} = \underbrace{\frac{t}{nt}\sum_{s=1}^{nt} (\xi_{si}\xi_{sj} - \mathbb{E}[\xi_{si}\xi_{sj}|\mathscr{F}_{s-1}])}_{(a)} + \underbrace{\frac{t}{nt}\sum_{s=1}^{tn} \mathbb{E}[\xi_{si}\xi_{sj}|\mathscr{F}_{s-1}]}_{(b)}.$$

For (a), under Assumption 2, $\sum_{s=1}^{nt}(\xi_{si}\xi_{sj}-\mathbb{E}[\xi_{si}\xi_{sj}|\mathscr{F}_{s-1}])$ is again a martingale. We can thus apply martingale law of large numbers (Csörgő 1968):

$$\frac{1}{nt}\sum_{s=1}^{nt}(\xi_{si}\xi_{sj}-\mathbb{E}[\xi_{si}\xi_{sj}|\mathscr{F}_{s-1}])\Rightarrow 0 \text{ as } n\to\infty.$$

For (b), under Assumption 2, we have

$$\frac{t}{nt}\sum_{s=1}^{tn}\mathbb{E}[\xi_{si}\xi_{sj}|\mathscr{F}_{s-1}]\Rightarrow U_{ij}t \text{ as } n\to\infty.$$

Then, setting $G=\nabla^2 H(x^*)^{-1}U$, we have $tn^{1/2}J^{(2)}(nt)\Rightarrow GB(t)$ in $D(0,\infty)$ as $n\to\infty$, i.e., (8).

For $J^{(3)}$, by Assumption 2, we have for any $\delta>0$ and n large enough,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{1\le t\le nT}\left\|n^{-1/2}\sum_{i=1}^{t}\phi_i^t\xi_i\right\|\ge\delta\right)=\mathbb{P}\left(\sup_{1\le t\le nT}\left\|(2MK)^{-1}\sum_{i=1}^{t}\phi_i^t\xi_i\right\|\ge n^{1/2}\delta(2MK)^{-1}\right)\\
\le\ &\sum_{t=1}^{nT}2\exp\left(1-(n^{1/2}\delta(2MK)^{-1}-1)^2/\left(2\sum_{s=1}^{t}M^2(2MK)^{-2}\|\phi_s^t\|^2\right)\right) \text{ by Lemma 1}\\
\le\ &\sum_{t=1}^{nT}2\exp\left(1-2K^2(\delta(2MK)^{-1}-n^{-1/2})^2/\left(n^{-1}\sum_{s=1}^{t}\|\phi_s^t\|\right)\right)\\
\le\ &\sum_{t=1}^{nT}2\exp\left(1-2K^2(\delta(2MK)^{-1}-n^{-1/2})^2/\left(n^{-1}C't^r\right)\right) \text{ for some } C'>0 \text{ by P3}\\
\le\ &2nT\exp\left(1-2K^2(\delta(2MK)^{-1}-n^{-1/2})^2(C'T^r)^{-1}n^{1-r}\right)\to 0 \text{ as } n\to\infty.
\end{aligned}$$

Thus, $tn^{1/2}J^{(3)}(nt)\Rightarrow 0$ in $D(0,\infty)$ as $n\to\infty$. □

We note from Theorem 2 that if we fix $t=1$, then $n^{1/2}(\bar{X}_n-x^*)\Rightarrow N(0,G)$ as $n\to\infty$. This implies that the FCLT we established is stronger than the large sample central limit theorem. We also comment that FCLT is required for a more general class of cancellation methods known as the standardized time series (Glynn and Iglehart 1990).

5.2 Non-degeneracy of $g_m(B,w)$

For the batch means method to be valid, we require that the number of batches $m\ge d+1$. This is because when $m\le d$, the estimated covariance matrix, $S_m(T)$, is likely to be degenerate. Specifically, from Theorem 2, we have that for any $m\in\mathbb{Z}^+$, $TS_m(T)\Rightarrow Gg_m(B,w)G^T$ as $T\to\infty$. The following lemma characterizes the behavior of $g_m(B,w)$ for different values of m, including $m\le d$.

Lemma 3 For $w\in\mathbb{R}_+^m$, when $m\ge d+1$, $g_m(B,w)$ is positive definite with probability 1; when $m\le d$, $g_m(B,w)$ is degenerate with probability 1.

Proof. Recall that $w_i=c_i-c_{i-1}$. Let $N_i=w_i^{-1/2}(B(c_i)-B(c_{i-1}))$ and

$$r_i=[-w_1^{1/2},\ldots,-w_{i-1}^{1/2},\ w_i^{-1/2}-w_i^{1/2},\ -w_{i+1}^{1/2},\ldots,-w_m^{1/2}]^T,$$

for $i=1,2,\ldots,m$. Then,

$$g_m(B,w)=(m-1)^{-1}N\left(\sum_{i=1}^{m}r_ir_i^T\right)N^T=(m-1)^{-1}NVN^T,$$

where $N = [N_1, \ldots, N_m]$ is a $d \times m$ matrix whose columns are independent and identically distributed d-dimensional standard Gaussian random vectors, and V is an $m \times m$ matrix with $V_{ii} = 1/w_i - 2 + mw_i$ and $V_{ij} = -(w_i/w_j)^{1/2} - (w_j/w_i)^{1/2} + m(w_i w_j)^{1/2}$ for $i \neq j$. Let $V_i = [V_{i1}, \ldots, V_{im}]$. In what follows, we shall prove that V has rank $m-1$.

We first note that because $\sum_{i=1}^m w_i^{1/2} V_i = 0$, $\text{rank}(V) \leq m-1$. We next look at the 'upper-left corner' $(m-1) \times (m-1)$ sub-matrix of V, which we denoted as $\tilde{V}$. We can decomposition $\tilde{V}$ as $\tilde{V} = \tilde{V}^1 + \Delta$, where $\tilde{V}^1_{ij} = \tilde{V}_{ij}$ for $i \neq j$, and $\tilde{V}^1_{ii} = (m-1)/(mw_i) - 2 + mw_i$; $\Delta_{ij} = 0$ for $i \neq j$, and $\Delta_{ii} = (mw_i)^{-1} > 0$. Let

$$\tilde{w}_i = \frac{m}{m-1} w_i, \quad \tilde{r}_i = \left[-\tilde{w}_1^{1/2}, \ldots, (\tilde{w}_i)^{-1/2} - (\tilde{w}_i)^{1/2}, \ldots, -(\tilde{w}_{m-1})^{1/2} \right]^T,$$

for $i = 1, \ldots, m-1$. Then we have $\tilde{V}^1 = \sum_{i=1}^{m-1} \tilde{r}_i \tilde{r}_i^T$. This suggests $\tilde{V}^1$ is positive semi-definite. As Δ is strictly positive definite, $\tilde{V}$ is positive definite. This indicates that $\text{rank}(V) \geq m-1$. Thus, $\text{rank}(V) = m-1$.

For $m \leq d$, $\text{rank}(g_m(B,w)) \leq \text{rank}(V) \leq m-1 < d$, Thus, $g_m(B,w)$ is degenerate.

For $m > d$, define $\mathscr{P}(N) = \det(NN^T)$, which is a polynomial function over entries of N. Because all entries of N are independent and identically distributed standard Normal random variables, $\{X \in \mathbb{R}^{d \times m} : \mathscr{P}(X) = 0\}$ has Lebesgue measure 0, i.e., $\mathbb{P}(\mathscr{P}(N) = 0) = 0$. This indicates that N has rank d a.s.. Since V has rank $m-1$ and is positive semi-definite, we can decompose it as $V = P\Lambda P^T$, where Λ is a diagonal matrix with $\Lambda_{ii} > 0$ for $i = 1, 2, \ldots, m-1$ and $\Lambda_{mm} = 0$, P is an orthogonal matrix. Next, note that

$$NP\Lambda^{1/2} \stackrel{d}{=} N\Lambda^{1/2} = [\tilde{N}, 0] \text{ where } \tilde{N} = [\sqrt{\Lambda_1} N_1, \ldots, \sqrt{\Lambda_{m-1}} N_{m-1}].$$

The elements of $\tilde{N}$ are independent and identically distributed standard Normal random variables. Thus, $\text{rank}(\tilde{N}) = d$ a.s. as well. Lastly, because

$$g_m(B,w) = (m-1)^{-1} NVN^T \stackrel{d}{=} (m-1)^{-1} \tilde{N}\tilde{N}^T \text{ and } \text{rank}(\tilde{N}) = d \text{ a.s.},$$

$g_m(B,w)$ is positive definite a.s.. □

5.3 Proof of Theorem 1

Proof of Theorem 1. Recall that B denotes a d-dimensional BM. The proof builds on verifying the following conditions for $g_m(x,c)$ in Theorem 1 of (Munoz and Glynn 2001):

a. $g_m(Gx,w) = Gg_m(x,w)G^T$ for any non-singular $d \times d$ matrix G.
b. $g_m(x - \beta\eta, w) = g_m(x,w)$ for $x \in C[0,1]^d$ and $\beta \in R^d$, where $\eta(t) := t$, $0 \leq t \leq 1$.
c. $g_m(B,w)$ is positive definite and symmetric almost surely.
d. $\mathbb{P}(B \in D(g_m(\cdot,w))) = 0$ where $D(g_m(\cdot,w))$ is the set of discontinuities of $g_m(\cdot,c)$.

For (a), we note that

$$\begin{aligned} g_m(Gx,w) &= \frac{1}{m-1} \sum_{i=1}^m \left(\frac{Gx(c_i) - Gx(c_{i-1})}{c_i - c_{i-1}} - Gx(1) \right) \left(\frac{Gx(c_i) - Gx(c_{i-1})}{c_i - c_{i-1}} - Gx(1) \right)^T \\ &= \frac{G}{m-1} \sum_{i=1}^m \left(\frac{x(c_i) - x(c_{i-1})}{c_i - c_{i-1}} - x(1) \right) \left(\frac{x(c_i) - x(c_{i-1})}{c_i - c_{i-1}} - x(1) \right)^T G^T = Gg_m(x,w)G^T. \end{aligned}$$

For (b), we have

$$g_m(x - \beta J, w) = \frac{1}{m-1} \sum_{i=1}^m \left(\frac{x(c_i) - x(c_{i-1})}{c_i - c_{i-1}} - x(1) \right) \left(\frac{x(c_i) - x(c_{i-1})}{c_i - c_{i-1}} - x(1) \right)^T = g_m(x,w).$$

(c) follows from Lemma 3. Because $g_m(\cdot, w)$ is continuous on $C[0,1]^d$, (d) is also satisfied.

Let $\bar{Y}_T(u) = T^{-1}\sum_{i=1}^{uT} X_i$, $0 \le u \le 1$. Note that $S_m(T) = g_m(\bar{Y}_T, w)$. From Theorem 2, $\bar{Y}_T(t) \Rightarrow GB(t)$ in $D[0,1]$ as $T \to \infty$. Then, from Theorem 1 in (Munoz and Glynn 2001), we have

$$\Gamma_T = m(m-d)/(d(m-1))(\bar{X}_T - x^*)^T S_m^{-1}(T)(\bar{X}_T - x^*) \Rightarrow m(m-d)/(d(m-1))B^T(1)g_m(B,w)^{-1}B(1)$$

as $T \to \infty$. Moreover, we note that

$$\frac{B(c_i) - B(c_{i-1})}{c_i - c_{i-1}} - B(1) = \frac{1}{c_i - c_{i-1}}\left(B(c_i) - c_i B(1) - (B(c_{i-1}) - c_{i-1}B(1))\right).$$

Because $B(u) - uB(1)$, $0 \le u \le 1$, is independent of $B(1)$, $g_m(B,w)$ independent of $B(1)$. □

REFERENCES

Agarwal, A., P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. 2012. "Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization". *IEEE Transactions on Information Theory* 58(5):3235–3249.

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.

Csörgő, M. 1968. "On the strong law of large numbers and the central limit theorem for martingales". *Transactions of the American Mathematical Society* 131(1):259–275.

Glynn, P., and D. Iglehart. 1990. "Simulation output analysis using standardized time series". *Mathematics of Operations Research* 15(1):1–16.

Hayes, T. P. 2005. "A large-deviation inequality for vector-valued martingales". Technical Report, available at: https://www.cs.unm.edu/ hayes/papers/VectorAzuma/, accessed 19 May, 2021.

Hsieh, M.-h., and P. W. Glynn. 2002. "Confidence regions for stochastic approximation algorithms". In *Proceedings of the Winter Simulation Conference*, edited by E. Y. amd C.-H. Chen, J. Snowdon, and J. Charnes, 370–376. Piscataway, New Jersey: IEEE.

Munoz, D., and P. Glynn. 2001. "Multivaraite standardized time series for steady-state simulation output analysis". *Operations Research* 49(3):413–422.

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.

Pang, G., R. Talreja, and W. Whitt. 2007. "Martingale proofs of many-server heavy-traffic limits for Markovian queues". *Probability Surveys* 4:193–267.

Polyak, B. T., and A. B. Juditsky. 1992. "Acceleration of stochastic approximation by averaging". *SIAM Journal on Control and Optimization* 30(4):838–855.

Robbins, H., and S. Monro. 1951. "A stochastic approximation method". *The Annals of Mathematical Statistics* 22(3):400–407.

Schruben, L. 1983. "Confidence interval estimation using standardized time series". *Operations Research* 31(6):1090–1108.

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.

Toulis, P., and E. M. Airoldi. 2017. "Asymptotic and finite-sample properties of estimators based on stochastic gradients". *The Annals of Statistics* 45(4):1694–1727.

Yang, W.-N., and B. L. Nelson. 1992. "Multivariate batch means and control variates". *Management science* 38(10):1415–1431.

Zhang, T. 2004. "Solving large scale linear prediction problems using stochastic gradient descent algorithms". In *Proceedings of the 21st International Conference on Machine Learning*.

AUTHOR BIOGRAPHIES

Yi Zhu is a researcher in WeRide Corp. He received his PhD in Industrial Engineering and Management Sciences from Northwestern University. His research interests are in uncertainty quantification, machine learning, and autonomous driving. His email address is yizhu2020@u.northwestern.edu.

Jing Dong is an associate professor in the Decision, Risk, and Operations Division at Columbia Business School. She received her PhD in Operations Research from Columbia University. Her research interests are at the interface of applied probability and service operations management. She also develops efficient simulation algorithms to facilitate better decision-making in complex stochastic systems. Her email address is jing.dong@gsb.columbia.edu.