On Constructing Confidence Region for Model Parameters in Stochastic Gradient Descent via Batch Means

Yi Zhu
Northwestern University
Jing Dong
Columbia Business School

Abstract

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

1 Introduction

Stochastic Gradient Descent and variants of it have been widely used in model-parameter estimation for either online learning or when data sizes are very large [13, 12, 8]. As the estimators we construct via stochastic gradient descent is random, it is desirable to be able to quantify the estimation errors incurred. While there is a rich literature studying convergence rate of the objective function or the parameter estimation errors based on stochastic gradient descent, e.g., [17, 1, 10], much less is known about the statistical inference for true model parameters [7, 16, 4, 2, 15]. Following the later line of work, in this paper, we propose a simple procedure to construct asymptotically valid confidence regions for model parameters based on a cancellation method known as the batch means. The confidence region we construct is tight as it takes the covariance structure of the parameters into account.

We consider the classic setting where the model parameters, x^*, can be characterized as the minimizer of a convex objective function, which is also known as the loss function. Specifically,

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

where h is a real-valued function, x is a d-dimensional parameter, 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 take $\bar{X}_t = t^{-1}\sum_{i=0}^{t-1} X_i$ as an estimator for x^*, then under certain regularity conditions, [12] establish that

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

where $\Rightarrow$ denotes convergence in distribution, $N(0,\Sigma)$ denotes a Gaussian random vector with mean 0 and covariance matrix Σ, and

$$\Sigma = \nabla^2 H(x^*)^{-1} U \nabla^2 H(x^*)^{-1},$$

where $\nabla^2 H(x^*)$ is the Hessian of H at x^*, and $U = E[\nabla_x h(x^*,\zeta)\nabla_x h(x^*,\zeta)^T]$. If we know the value of Σ, then a natural way to construct the 95% confidence region for x^* is

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

where $\chi^2_{d,0.05}$ is the 95%-quantile of the chi-squared distribution with d degrees of freedom. The confidence region is asymptotically valid in the sense that $\lim_{t\to\infty} \mathrm{pr}(x^* \in \hat{R}_t) = 0.95$. The main challenge here is that covariance matrix Σ is unknown and it is very costly to construct consistent estimators of Σ [2].

To address the challenge, we introduce a cancellation method, called the batch means, from the stochastic simulation literature [14, 5, 9]. The main idea is to construct the statistics in a special way to cancel out the unknown covariance matrix. The method was introduced to deal with steady-state estimation problems, where we use the time average of the stochastic process as an estimator of the steady-state mean. Despite the elegance of the method, existing results in the literature do not allow us to apply it directly in the stochastic gradient descent setting. This is because in steady-state estimation problems, the stochastic process is time-homogeneous, while in stochastic gradient descent, if we view $\{X_t : t \geq 0\}$ as a stochastic process, the transition kernel is time-dependent 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 stochastic gradient descent setting. This provides us with a simple way to construct asymptotically valid confidence regions for model parameters. The method utilizes the output of the stochastic gradient descent algorithm itself, and it does not require any modification to the underlying algorithm. We also extend the batch means method to allow more general batch size specifications and provide some guidance on how to select the batch sizes. Our analysis relies on the process-level convergence result for $\{\bar{X}_t : t \geq 0\}$, which is stronger than the large sample convergence result established in the literature.

2 Batch means method

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

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

where $E[\mathcal{G}(X_{t-1}, \zeta_t)|X_{t-1}] = \nabla H(X_{t-1})$ and $\gamma_t = at^{-r}$ for some $a > 0$ and $r \in (1/2, 1)$. The batch means method divides the stochastic gradient descent sample path $\{X_t : 0 \leq t \leq T\}$ into m non-overlapping batches, where the i-th batch is of size $b_i := \lceil Tw_i \rceil$. Here, $m \in \mathbb{Z}_+$ with $m > d$, and $w = (w_1, \ldots, w_m) \in \mathbb{R}_+^m$, where $\mathbb{Z}_+$ is the set of positive integers and $\mathbb{R}_+$ is the set of positive real numbers. m and w are the parameters for the batch means method. The method is asymptotically valid for a wide range of parameter specifications. As for the pre-limit performance, we will discuss how to "fine-tune" these parameters in Section 3. We define $\tau_i = \sum_{j=1}^{i} b_i$. Then the ith batch contains $\{X_{\tau_{i-1}+1}, \ldots, X_{\tau_i}\}$ and its batch mean is defined as $\Xi_i = b_i^{-1} \sum_{t=\tau_{i-1}+1}^{\tau_i} X_t$.

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 construct F type of statistics based on the m batch means. In particular, we consider the statistics

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

where $\bar{X}_T := T^{-1} \sum_{t=1}^{T} X_t$ and $S_m(T) := (m-1)^{-1} \sum_{i=1}^{m} (\Xi_i - \bar{X}_T)(\Xi_i - \bar{X}_T)^T$ How Γ_T works will be made precise in Theorem 1. The actual procedure to construct the confidence region is summarized in the following Algorithm.

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

```
Input: The SGD sample path $\{X_t : 0 \leq t \leq T\}$, the number of batches $m$, the relative batch length parameter $w$
Find the appropriate scaling parameter $\alpha_m(\delta, w)$.
Calculate the batch means $\Xi_i$ for $i = 1, 2, \ldots, m$
Calculate $\bar{X}_T$ and $S_m(T)$
Output: $R_T = \left\{ x \in \mathbb{R}^d : \frac{m(m-d)}{d(m-1)} (\bar{X}_T - x)^T S_m^{-1}(T)(\bar{X}_T - x) \leq \alpha_m(\delta, w) \right\}$.
```

The confidence regime constructed in Algorithm 1 is asymptotic valid in the sense that if the scaling parameter $\alpha_m(\delta, w)$ is properly chosen, then $\lim_{T\to\infty} \mathrm{pr}(x^* \in R_T) = 1 - \delta$. The key now is to calibrate the appropriate scaling parameter $\alpha_m(\delta, w)$. The value of $\alpha_m(\delta, w)$ is determined by the asymptotic behavior of Γ_T. Theorem 1 characterizes the limiting distribution of Γ_T as $T \to \infty$, and is the main result of this paper. Before we present the theorem, we first introduce a few assumptions, which are standard for the convergence analysis of Polyak-Ruppert averaging, e.g., [2, 12]. We define $\Delta_t := X_t - x^*$, and $\xi_t = (\xi_t(1), \ldots, \xi_t(d))$ as

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

Assumption 1. *$H(x)$ is continuously differentiable and strongly convex with parameter $C > 0$, i.e., for any x and y $H(y) \geq H(x) - \nabla H(x)^T(y-x) + \frac{C}{2}\|y-x\|_2^2$. $\nabla H(x)$ is Lipschitz continuous with parameter $L > 0$, i.e., for any x and y, $\|\nabla H(x) - \nabla H(y)\|_2 \leq L\|x-y\|_2$, and $\nabla^2 H(x^*)$ exists.*

Assumption 2. *$(\xi_t : t \geq 1)$ is a martingale-difference process with respect to the filtration $\mathcal{F} = \{\mathcal{F}\}_{t\geq 1}$ generated by $(\zeta_t : t \geq 1)$, and it satisfies the following:*
1. *The conditional covariance of ξ_t has an expansion around $x = x^*$: $E[\xi_t \xi_t^T | \mathcal{F}_{t-1}] = U + r(\Delta_{t-1})$, for some positive definite matrix U, and there exit constants $S_1 > 0$ and $S_2 > 0$, such that for any $\Delta \in \mathbb{R}^d$, $\|r(\Delta)\|_2 \leq S_1 \|\Delta\|_2 + S_2 \|\Delta\|_2^2$.*
2. *There exists $M \in (0, \infty)$, such that $\|\xi_t\| \leq M$ almost surely, $\forall t \geq 1$.*

Remark 1. *Assumption 1 ensures the convergence of $\bar{X}_t$ to a unique global optimal x^* [12]. Assumption 2 provides sufficient conditions to establish the functional Central Limit Theorem (FCLT) for partial sums of ξ_t's.*

We define the function $g_m : C^d[0,1] \times \mathbb{R}^m \to \mathbb{R}^{d\times d}$ as

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

where $c_0 = 0$ and $c_i = c_{i-1} + w_i$. We are now ready to introduce the main theorem.

Theorem 1. *Under Assumption 1 and 2, for Γ_T defined in (3), when $m > d$ and $w > 0$,*

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

where Z is a standard d-dimensional Gaussian random vector, B is standard d-dimensional Brownian motion, 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

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

We note from Theorem 1 that the scaling parameter $\alpha_m(\delta, w)$ does not depend on the underline problem instances. It only depends on the batch means parameters m and w. In the special case of evenly-split batch size, i.e., $w_i = 1/m$,

$$m(m-d)(d(m-1))^{-1} Z^T g_m(B, w)^{-1} Z$$

follows an F distribution with d and $m - d$ degrees of freedom. We will discuss a different splitting scheme in Section 3 and provide the corresponding scaling parameter table (Table 1).

3 Selection of the batch means parameters

The confidence region constructed using the batch means method is asymptotic valid regardless of our choice of m and w, as long as $m > d$ and $w > 0$. However, different m and w will affect the pre-limit performance of the procedure. In this section, we study how to choose the parameters for the batch means method. The analysis is divided into two parts. We first study for a fixed m, how to choose the batch sizes w. We then study how to choose m.

The pre-limit performance is essentially determined by how close the distribution of

$$\Big((b_1/\sqrt{T})(\Xi_1 - x^*), \ldots, (b_m/\sqrt{T})(\Xi_m - x^*)\Big)$$

is to $(G(B(c_1) - B(c_0)), \ldots, G(B(c_m) - B(c_{m-1})))$.

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 [2]. We first note that for t large, 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 $\Delta_t = X_t - x^*$, and the equality follows as $\nabla H(x^*) = 0$. Then 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, 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 $\lambda(A)$ denote 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, we have increasing batch sizes, i.e., w_i's are increasing in i. This is similar to the batch size splitting rule proposed in [2]. For what follows, we shall refer to this specification as the “increasing batch size" (IBS) allocation. The main difference between our method and the one in [2] is that the method in [2] requires sending m to infinity as T goes to infinity, while our method holds m fixed. We will conduct more comparisons of the two methods in Section 4.

Table 1 provides some of the commonly used scaling parameters for IBS with different values of d and m. As these quantiles are estimated using Monte Carlo simulation, we also provide the corresponding 95% confidence intervals.

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

d	1	2	3	4
$m = 10$	2.93 ± 0.01	2.92 ± 0.01	3.13 ± 0.01	3.50 ± 0.01
$m = 20$	2.18 ± 0.01	2.00 ± 0.01	1.95 ± 0.01	1.97 ± 0.01
$m = 30$	1.91 ± 0.01	1.71 ± 0.01	1.64 ± 0.01	1.62 ± 0.01
$m = 40$	1.76 ± 0.01	1.55 ± 0.01	1.47 ± 0.01	1.50 ± 0.01
$m = 60$	1.58 ± 0.01	1.38 ± 0.01	1.29 ± 0.01	1.26 ± 0.01
$m = 100$	1.42 ± 0.01	1.21 ± 0.01	1.13 ± 0.01	1.09 ± 0.01
$m = 120$	1.33 ± 0.01	1.15 ± 0.01	1.07 ± 0.01	1.03 ± 0.01
$m = 150$	1.28 ± 0.01	1.09 ± 0.01	1.01 ± 0.01	0.97 ± 0.01
d	10	50	80	100
$m = 10$	NA	NA	NA	NA
$m = 20$	2.46 ± 0.01	NA	NA	NA
$m = 30$	1.74 ± 0.01	NA	NA	NA
$m = 40$	1.48 ± 0.01	NA	NA	NA
$m = 60$	1.24 ± 0.01	2.49 ± 0.04	NA	NA
$m = 100$	1.02 ± 0.01	1.31 ± 0.01	1.81 ± 0.01	NA
$m = 120$	0.97 ± 0.01	1.18 ± 0.01	1.43 ± 0.01	1.81 ± 0.01
$m = 150$	0.91 ± 0.01	1.07 ± 0.01	1.22 ± 0.01	1.35 ± 0.01

We next show some numerical experiments about different choices of batch sizes. We compare three different specifications: i) IBS, ii) even splitting (ES), and iii) decreasing batch size (DBS) where we reverse the batch size specification of IBS. Table 2 summarizes results.

For Table 2 and subsequent numerical experiments, we focus on two classes of examples: linear regression and logistic regression. For linear regression, we write $b_i = x^{*T} a_i + \epsilon_i$ where ϵ_i's are iid $N(0,1)$. In this case, $\zeta = (a,b)$ and $h(x,\zeta) = (b - x^T a)^2$. For logistic regression, we consider $b_i \in \{-1, 1\}$ with $\mathbb{P}(b_i = 1|a_i) = (1 + \exp(-x^{*T} a_i))^{-1}$. 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 examples, our goal is to achieve 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 iteration increases, all three batch size specifications are approaching the correct coverage rate, 0.95. For a relatively small number of iterations, IBS and ES achieve a higher coverage rate than DBS.

3.2 Number of batches

We next look into different choices of m for $m \geq d + 1$. We divide the analysis into two parts. We first analyze the limiting volume of the confidence region for different choices

Table 2: Coverage rate comparison for different batch size specifications

	n	$4n$	$7n$	$10n$
Linear regression with $d=2$				
IBS	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
DBS	0.787 ± 0.025	0.878 ± 0.020	0.909 ± 0.017	0.912 ± 0.019
Logistic regression with $d=2$				
IBS	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
DBS	0.842 ± 0.023	0.908 ± 0.017	0.932 ± 0.018	0.930 ± 0.015

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

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

Thus, in Figure 1, we compare

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

for different values of m. We observe that as m increases, the volume of the confidence region decreases. However, there is a diminishing effect of increasing m on decreasing the volume. Moreover, for pre-limit, the larger m is, the smaller the size of each batch would be, which implies that the batch means are further from their corresponding asymptotic distributions. These suggest that m should not be too large. This is especially important when T is relatively small.

In Table 3, we compare the pre-limit performance for different values of m. We use IBS for the batch size specification. We focus on a relatively small number of iterations in these examples and we observe that when the numbers of iterations are small, large values of m can lead to substantial under-coverage.

4 Comparison to other methods

In this section, we compare our batch means method to two recently developed methods to draw statistical inference for model parameters in SGD. Specifically, the methods are developed in [2] and [15], which we refer to as batch means with an increasing number

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

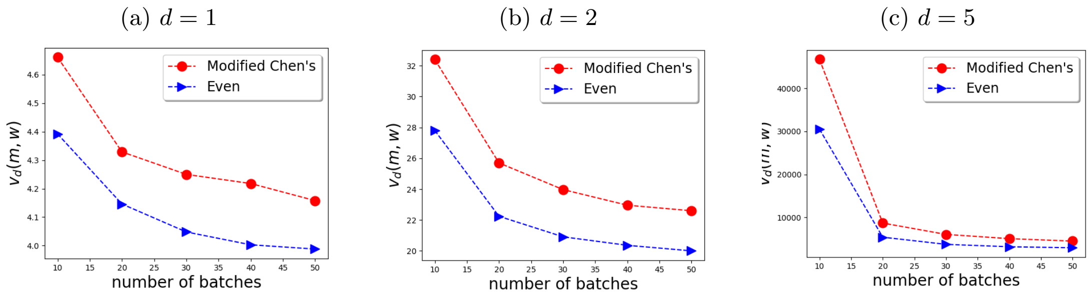

(a) $d = 1$ (b) $d = 2$ (c) $d = 5$

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

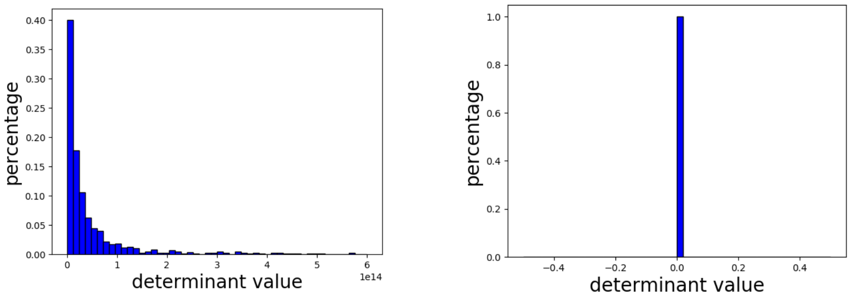
of batches (BMI) and hierarchical incremental gradient descent (HiGrad), respectively. We also introduce a fourth method, which is known as the sectioning method [7]. This method is similar to the batch means method, but instead of dividing a single sample path into m batches, we generate m independent sample path of equal length. This method can also be viewed as a special case of HiGrad where the number of levels is 1.

BMI is mainly designed to draw marginal inference, i.e., it constructs confidence intervals for each parameter (dimension) separately. Thus, it does not impose $m \geq d + 1$. However, we note from Lemma 3 that when $m \leq d$, the estimated covariance matrix $S_m(T)$ is likely to be degenerate. Indeed, Figure 2 plots the histogram of the determinant of $S_m(T)$ for a logistic regression problem with n iterations. Note that in this case, BMI suggests setting $m = \lceil n^{0.25} \rceil = 18$. We compare two cases, one has $d = 10 < m$, the other has $d = 20 > m$. We observe that when $d > m$, the determinant of $S_m(T)$ is concentrated around zero.

HiGrad has versions for both marginal inference and joint inference. However, we note that there are a lot more parameters to be specified (e.g., the tree structure and partition of data set) for successful implementation of this method. HiGrad also requires modification to the original SGD procedure. The sectioning method, a special case of HiGrad, has the advantage that estimators constructed for different sections are independent. Thus, the asymptotic independence requirement is automatically satisfied. However, if we have limited amount of computational budget, focusing on a single long run instead of multiple shorter runs may get us closer to x^* and the normality requirement.

In Table 4 and 5 we compare the finite sample coverage rate of our batch means method and other benchmark methods for logistic regression examples. For the batch means method (BM), we set $m = 30$ and use IBS for batch size specification. When doing

Figure 2: Histogram for the determinant of $S_m(T)$ with $m = 18$



(a) $d = 10$

(b) $d = 20$

joint inference, we set the marginal confidence level at $1 - 0.05/d$ for BMI based on the Bonferroni correction.

For HiGrad, we use a two-layer tree structure with 5 and 6 nodes for the respective layers. Note that in this case, we have 30 branches in total. When doing marginal inference, for BM, we can construct the batch means confidence interval for each parameter (dimension) separately. Algorithm 2 summarizes our marginal inference procedure.

Algorithm 2 Construct the marginal $100(1 - \delta)\%$ confidence interval for each dimension of the model parameter x^*

```
Input: The SGD sample path of $\{X_t : 0 \le t \le T\}$, the number of batches $m$, the relative batch length parameter $w$
Find the appropriate scaling parameter $\alpha_m(\delta, w)$ with $d = 1$.
Calculate the batch means $\Xi_i$ for $i = 1, 2, \ldots, m$
For $k = 1, \ldots, d$, calculate
    $$\bar{X}_T(k) := \frac{1}{T}\sum_{t=1}^{T} X_t(k), \quad \sigma_{m,T}(k) := \sqrt{\frac{1}{m-1}\sum_{k=1}^{m}(\Xi_i(k) - \bar{X}_T(k))^2}.$$
Output:
    $$R_T(k) = \left[\bar{X}_T(k) - \sqrt{\frac{\alpha_m(\delta, w)}{m}}\sigma_{m,T}(k), \bar{X}_T(k) + \sqrt{\frac{\alpha_m(\delta, w)}{m}}\sigma_{m,T}(k)\right],$$
    for $k = 1, \ldots d$.
```

In Table 4, we show results for confidence regions (joint inference). 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 comparing to the benchmark methods in all cases. As all the methods we compare are asymptotically valid, we expect all these methods to achieve good coverage rate when the number of iterations (samples) is large enough. We also note that the coverage rates deteriorate as the dimension of the problem, d, increases.

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 Concluding remarks

In this paper, we adapt the batch means method to construct asymptotically valid confidence regions for model parameters in SGD. Our construct is simple and does not require any modification to the underline SGD algorithm. We extend the class of batch means method to allow unequal batch sizes. We also extend the asymptotic analysis of Polyak-Ruppert averaging by establishing a process level functional central limit theorem.

Our construction requires that the number of batches $m > d$. However, we do not want m to be too large, especially when the sample size T is small. Following extensive numerical experiments, we suggest setting m between 20 and 40 when $d < 10$, and m

between $d+5$ and $d+10$ when $d \geq 10$. In terms of the batch size, both ES and IBS work well. Lastly, if we do not have a good knowledge of the starting value for SGD, we would also recommend discard the first few iterations when constructing batches to eliminate the initial transient bias.

A Appendix

The proof of Theorem 1 involves two main steps. The first step is to establish the process level convergence of $\bar{X}_t$. The second step is to establish some important properties of the function g_m. In particular, we need to show that $g_m(B, w)$ is positive definite with probability 1.

For the first step, we start by presenting two auxiliary lemmas. The first lemma extends the Azuma-Hoeffding inequality to the multidimensional case.

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

$$\mathrm{pr}(\|\mathcal{M}_n\| \geq a) \leq 2\exp\left(1 - (a-1)^2/(\sum_{i=1}^{n} 2\sigma_i^2)\right).$$

Proof of Lemma 1. The lemma follows similar lines of arguments but extends the results of Theorem 1.8 in [6].

As $\mathcal{M}$ is a martingale, we have the following decomposition for $\mathcal{M}_i$: $\mathcal{M}_i = \alpha_i \mathcal{M}_{i-1} + P_i$, where $\alpha_i = (\|\mathcal{M}_{i-1}\|^2)^{-1}\langle \mathcal{M}_i, \mathcal{M}_{i-1}\rangle$ and P_i is orthogonal to $\mathcal{M}_{i-1}$. Define $A_i = (\alpha_i - 1)\|\mathcal{M}_{i-1}\|$. Then $\|\mathcal{M}_i - \mathcal{M}_{i-1}\|^2 = A_i^2 + \|P_i\|^2 \leq \sigma_i^2$.

We next define $D = (D_i : 1 \leq i \leq n)$, $Z = (Z_i : 0 \leq i \leq n)$ and $Y = (Y_i : 0 \leq i \leq n)$ by the following recursion: $Y_0 = 1 + \sum_{i=1}^n \sigma_i^2/(a-1)$ and $Z_0 = 0$. For $i \geq 1$,

$$D_i = \mathrm{sign}(Z_{i-1})\left((Y_{i-1}^2 + 2Y_{i-1}A_i + \sigma_i^2)^{1/2} - Y_{i-1}\right); Z_i = Z_{i-1} + D_i; \quad Y_i = Y_0 + |Z_i|;$$

where $\mathrm{sign}(z) = 1 - 2\mathbb{1}\,(z < 0)$.

We first prove that

$$\|M_i\| \leq Y_i. \tag{5}$$

The proof is divided into two steps. We first establish a bound for $|Z_i|$. We then prove (5) by induction. Note that as

$$\begin{aligned} & \mathrm{sign}(Z_{i-1})Z_i = \mathrm{sign}(Z_{i-1})Z_{i-1} + \mathrm{sign}(Z_{i-1})D_i \\ = \;& |Z_{i-1}| + (Y_{i-1}^2 + 2Y_{i-1}A_i + \sigma_i^2)^{1/2} - Y_{i-1} = (Y_{i-1}^2 + 2Y_{i-1}A_i + \sigma_i^2)^{1/2} - Y_0, \end{aligned}$$

we have $|Z_i| = \left|(Y_{i-1}^2 + 2Y_{i-1}A_i + \sigma_i^2)^{1/2} - Y_0\right|$.

Now by definition $Y_0 > 0 = \|\mathcal{M}_0\|$. Suppose (5) holds for $i-1$. Then

$$Y_i \;=\; Y_0 + |Z_i| = Y_0 + \left|(Y_{i-1}^2 + 2Y_{i-1}A_i + (\sigma_i)^2)^{1/2} - Y_0\right| \geq (Y_{i-1}^2 + 2Y_{i-1}A_i + \sigma_i^2)^{1/2},$$

which implies $Y_i^2 \geq Y_{i-1}^2 + 2Y_{i-1}A_i + \sigma_i^2$.

By the definition of A_i, we have $\|\mathcal{M}_i\|^2 = (\|\mathcal{M}_{i-1}\| + A_i)^2 + \|P_i\|^2$. Then

$$Y_i^2 - \|\mathcal{M}_i\|^2 \geq (Y_{i-1} - \|\mathcal{M}_{i-1}\|)(Y_{i-1} + \|\mathcal{M}_{i-1}\| + 2A_i) + (\sigma_i^2 - A_i^2 - \|P_i\|^2) > 0,$$

where the last inequality is due the facts that i) $Y_{i-1} > \|X_{i-1}\|$, ii) $Y_{i-1} \geq Y_0 \geq 1 \geq 2\sigma_i \geq 2|A_i|$, and iii) $\sigma_i^2 \geq A_i^2 + \|P_i\|^2$.

We next prove that for $\lambda = (a-1)/(\sum_{k=1}^n \sigma_k^2)$,

$$E[\exp(\lambda Z_n)] \leq \prod_{i=1}^n \cosh(\lambda\sigma_i) \leq \exp\left(\sum_{i=1}^n \lambda^2\sigma_i^2/2\right). \tag{6}$$

The second inequality follows straightforwardly. We shall thus focus on establish the first inequality in (6). To do so, we first establish a bound for $\exp(\lambda D_i)$. In particular, we shall first prove that

$$\exp(\lambda D_i) \leq \cosh(\lambda\sigma_i) + A_i/\sigma_i \text{sign}(Z_{i-1})\sinh(\lambda\sigma_i). \tag{7}$$

Fix Z_{i-1}, D_i can be view as a function of A_i. We can thus define $f(x) := \exp(\lambda D_i(x))$. Note that $D_i(\sigma_i) = \text{sign}(Z_{i-1})\sigma_i$ and $D_i(-\sigma_i) = -\text{sign}(Z_{i-1})\sigma_i$. Then the line linking $(-\sigma_i, f(-\sigma_i))$ and $(\sigma_i, f(\sigma_i))$ takes the form $y(x) = \cosh(\lambda\sigma_i) + x/\sigma_i\text{sign}(Z_{i-1})\sinh(\lambda\sigma_i)$. Then, to prove (7), it suffices to show $\partial^2 f(x)/\partial x^2 > 0$ on the interval $[-\sigma_i, \sigma_i]$.

$$\begin{aligned}\partial^2 f(x)/\partial x^2 &= \left((\lambda\partial D_i(x)/\partial x)^2 + \lambda\partial^2 D_i(x)/\partial a^2\right) f(x) \\ &= \left(\lambda(Y_{i-1}^2 + 2Y_{i-1}x + \sigma_i^2)^{1/2} - \text{sign}(Z_{i-1})\right)\frac{\lambda Y_{i-1}^2 f(x)}{(Y_{i-1}^2 + 2Y_{i-1}x + 1)^{3/2}}.\end{aligned}$$

Now, for $\lambda = (a-1)/\sum_{k=1}^n \sigma_k^2$ and $x \in [-\sigma_i, \sigma_i]$, we have

$$\lambda(Y_{i-1}^2 + 2Y_{i-1}A_i + \sigma_i^2)^{1/2} \geq \lambda(Y_0 - \sigma_i) = 1 + \lambda(1 - \sigma_i) \geq 1 \geq \text{sign}(Z_{i-1}).$$

Thus, $\partial^2 f(x)/\partial x^2 > 0$ and we have proved (7). Next, we note that $E[\exp(\lambda D_n)|Z_{n-1}] = E[E[\exp(\lambda D_n)|\mathcal{M}_0, \ldots, \mathcal{M}_{n-1}]|Z_{n-1}]$. As by (7),

$$\begin{aligned}&E[\exp(\lambda D_n)|\mathcal{M}_0, \ldots, \mathcal{M}_{n-1}] \\ \leq\ &\cosh(\lambda\sigma_n) + \text{sign}(Z_{n-1})\sinh(\lambda\sigma_n)\mathbb{E}[A_n|\mathcal{M}_0, \ldots, \mathcal{M}_{n-1}] = \cosh(\lambda\sigma_n),\end{aligned}$$

we have

$$\begin{aligned}&E[\exp(\lambda Z_n)] = E[\exp(\lambda Z_{n-1})\exp(\lambda D_n)] = E[\exp(\lambda Z_{n-1})E[\exp(\lambda D_n)|Z_{n-1})] \\ \leq\ &E[\exp(\lambda Z_{n-1})]\cosh(\lambda\sigma_n) \leq \prod_{i=1}^n \cosh(\lambda\sigma_i) \text{ by recursion.}\end{aligned}$$

Putting (5) and (7) together, we have

$$\begin{aligned}E[\exp(\lambda\|\mathcal{M}_n\|)] &\leq E[\exp\lambda Y_n] = \mathbb{E}[\exp(\lambda(Y_0 + |Z_n|))] \\ &\leq e^{\lambda Y_0}\mathbb{E}[\exp(\lambda Z_n) + \exp(-\lambda Z_n)] \leq 2e^{\lambda Y_0}\exp\left(\sum_{i=1}^n \lambda^2\sigma_i^2/2\right).\end{aligned}$$

Lastly,

$$\begin{aligned}\mathrm{pr}(\|\mathcal{M}_n\| \geq a) &\leq E[\exp(\lambda\|\mathcal{M}_n\| - \lambda a)] \leq 2\exp\left(\lambda Y_0 - \lambda a + \sum_{i=1}^{n} \sigma_i^2 \lambda^2/2\right) \\ &= 2\exp\left(1 - (a-1)^2/(2\sum_{k=1}^{n}\sigma_k^2)\right) \text{ by plugging in the value of } \lambda.\end{aligned}$$

□

The second lemma characterizes the convergence rate of an important term in stochastic gradient descent iterations. It tightens the bound established in [12].

Let

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

where we define $\prod_{k=s+1}^{s} \left(I - \gamma_k \nabla^2 H(x^*)\right) = I$. We also define $\phi_s^t = \bar{\beta}_s^t - \nabla^2 H(x^*)^{-1}$.

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 of Lemma 2. We start by summarizing some useful results from [12]. Let $\beta_s^s = I$ and $\beta_s^{t+1} = \beta_s^t(I - \gamma_t A)$ for $t \geq s$. There exists $\lambda, K \in (0, \infty)$ such that for any $s \geq 0$ and $t \geq s$, $\|\beta_s^t\| \leq K \exp\left(-\lambda \sum_{i=s}^{t-1} \gamma_i\right)$, where we define $\sum_{i=s}^{s-1} \gamma_i = 0$. Now let $S_s^t = \sum_{i=s}^{t-1} (\gamma_s - \gamma_i)\beta_s^i$. Then it can be shown that $\phi_s^t = S_s^t - \nabla^2 H(x^*)^{-1}\beta_s^t$. Let $m_s^i = \sum_{k=s}^{i} \gamma_k$. Then

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

We are now ready to prove the lemma. Note that $\|\phi_s^t\| \leq \|S_s^t\| + \|\nabla^2 H(x^*)^{-1}\| \|\beta_s^t\|$.

In what follows, we shall establish bounds for $\|S_s^t\|$ and $\|\beta_s^t\|$ respectively. We first note that

$$\begin{aligned}\|S_j^t\| &= \left\| \sum_{i=j+1}^{t-1} \left(\sum_{k=j+1}^{i} (\gamma_{k-1} - \gamma_k) \right) \beta_j^i \right\| \\ &= \left\| \sum_{i=j+1}^{t-1} \left(\sum_{k=j+1}^{i} (\gamma_{k-1} - \gamma_k)\gamma_{k-1}(\gamma_{k-1})^{-1} \right) \beta_j^i \right\| \\ &\leq K(\gamma_j - \gamma_{j+1})(\gamma_j)^{-1} \sum_{i=j}^{t-1} m_j^i \exp(-\lambda m_j^i).\end{aligned}$$

Thus, for j large enough, $\|S_j^t\| \leq K(\gamma_j - \gamma_{j+1})/\gamma_j^2$. We also notice tha by L'Hospital's Rule, $(\gamma_j - \gamma_{j+1})(\gamma_j^2)^{-1} = O(j^{-(1-r)})$. Then for t large enough,

$$\sum_{j=0}^{t-1} \|S_j^t\| \leq K \sum_{j=0}^{t-1} j^{-(1-r)} \leq K \int_0^t x^{-(1-r)} = O(t^r). \tag{8}$$

Next, we note that

$$\sum_{j=0}^{t-1} \|\beta_j^t\| \leq \sum_{j=0}^{t-1} \exp(-\lambda(t-j)\gamma_t) \leq \frac{1}{1-\exp(-\lambda\gamma_t)} = O(\gamma_t^{-1}) = O(t^r). \tag{9}$$

Combining (8) and (9), we have

$$\sum_{j=0}^{t-1} \|\phi_j^t\| \leq \sum_{j=0}^{t} \|S_j^t\| + \left\|\nabla^2 H(x^*)^{-1}\right\| \cdot \sum_{j=0}^{t} \|\beta_j^t\| = O(t^r).$$

□

Next, we establish the process level convergence of $\bar{X}_t$

Theorem 2. *Under Assumption 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 of Theorem 2. We start by summarizing some useful results from [12]. We first note that $\bar{X}_t$ has the following decomposition:

$$\bar{X}_t - x^* = J^{(1)}(t) + J^{(2)}(t) + J^{(3)}(t),$$

where

$$\begin{aligned} J^{(1)}(t) &= -t^{-1} \sum_{s=0}^{t-1} (\nabla^2 H(x^*) + \phi_s^t)(\nabla H(X_s) - \nabla^2 H(x^*)\Delta_s, \\ J^{(2)}(t) &= t^{-1} \sum_{s=0}^{t-1} \nabla^2 H(x^*)^{-1} \xi_s, \quad J^{(3)}(t) = t^{-1} \sum_{s=0}^{t-1} \phi_s^t \xi_s. \end{aligned}$$

Recall that $\xi_t = \mathcal{G}(X_{t-1}, \zeta_t) - \nabla H(X_{t-1})$ and $\Delta_t = X_t - x^*$. We also have the following properties about 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) $\|\phi_s^t\| \leq K$ 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 [12]. We establish it in Lemma 2.

We are now ready to establish the functional level convergence results for each part of the decomposition. For $J^{(1)}$, we have

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

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

For $J^{(2)}$, let $\mathcal{M}_n(t) := n^{-1/2}\sum_{s=1}^{nt}\xi_s$. We next establish the functional central limit theorem (FCLT) for M_n, i.e. there exists a matrix U such that

$$\mathcal{M}_n(t) \Rightarrow UB(t) \text{ in } D(0,\infty) \text{ as } n\to\infty. \tag{10}$$

Under Assumption 2, ξ_t's form a Martingale-difference sequence. Following Theorem 8.1 in [11], we only need to verify the following two conditions:

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

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

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

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

For C2), we have

$$\begin{aligned}
[\mathcal{M}_{ni},\mathcal{M}_{nj}](t) &= t/(nt)\sum_{s=1}^{nt}\xi_{si}\xi_{sj} \\
&= \underbrace{t/(nt)\sum_{s=1}^{nt}(\xi_{si}\xi_{sj}-E[\xi_{si}\xi_{sj}|\mathcal{F}_{s-1}])}_{(a)} + \underbrace{t/(nt)\sum_{s=1}^{tn}E[\xi_{si}\xi_{sj}|\mathcal{F}_{s-1}]}_{(b)}.
\end{aligned}$$

Under Assumption 2, (a) is again a martingale. We can thus apply martingale law of large number [3], i.e., $(nt)^{-1}\sum_{s=1}^{nt}(\xi_{si}\xi_{sj} - E[\xi_{si}\xi_{sj}]) \Rightarrow 0$ as $n \to \infty$. For (b), under Assumption 2, we have $t/(nt)\sum_{s=1}^{tn} E[\xi_{si}\xi_{sj}|\mathcal{F}_{s-1}] \Rightarrow U_{ij}t$ as $n \to \infty$.

Based on (10), we have for $G = \nabla^2 H(x^*)^{-1}U$,

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

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

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

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 we have $n^{1/2}(\bar{X}_n - x^*) \Rightarrow N(0, G)$ as $n \to \infty$, i.e., the FCLT result we established is stronger than the large sample central limit theorem. We also comment that FCLT is required for batch means and a more general class of cancellation methods known as the standardized time series [5].

We next carry out the second step. 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 when $m \le d$.

Lemma 3. *For $w > 0$, 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 of Lemma 3. Recall $w_i = c_i - c_{i-1}$, Let $N_i = w_i^{-1/2}B(c_i) - B(c_{i-1})$, and $r_i = \left[-w_1^{1/2}, \ldots, w_i^{-1/2} - w_i^{1/2}, \ldots, -w_m^{1/2}\right]^T$, for $i = 1, 2, \ldots, m$. Then

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

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 $i \neq j$.

In what follows, we shall prove that V has rank $m-1$. We first note that as $\sum_{i=1}^m w_i^{1/2} V_i = 0$, where V_i denotes i-th row of V, $\mathrm{rank}(V) \leq m-1$. 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 $\tilde{w}_i = (m-1)w_i/m$, $\tilde{r}_i = \left[-\tilde{w}_1^{1/2}, \ldots, (\tilde{w}_i)^{-1/2} - (\tilde{w}_i)^{1/2}, \ldots, -(\tilde{w}_{m-1})^{1/2}\right]^T$, 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 $\mathrm{rank}(V) \geq m-1$. Thus, $\mathrm{rank}(V) = m-1$.

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

For $m > d$, define $\mathcal{P}(N) = \det(NN^T)$, which is a polynomial function over entries of N. Notice as all entries of N are independent and identically distributed standard Normal random variables, $\{X \in \mathbb{R}^{d \times m} : \mathcal{P}(X) = 0\}$ has Lebesgue measure 0. Therefore, $\mathrm{pr}(\det(NN^T) = 0) = 0$. This indicates that N has rank d a.s.. Since V is of rank $m-1$ and 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 $NP\Lambda^{1/2} \overset{d}{=} N\Lambda^{1/2} = [\tilde{N}, 0]$ where $\tilde{N} = [\lambda_1 N_1, \ldots, \lambda_{m-1} N_{m-1}]$. It is easy to see that $\tilde{N}$ is again a gaussian random matrix with each independent and identically distributed standard Normal elements. Then $\mathrm{rank}(\tilde{N}) = d$ a.s.. Therefore, $g_m(B, w) = (m-1)^{-1} NVN^T$ has the same distribution as $(m-1)^{-1}\tilde{N}\tilde{N}^T$, having rank d almost surely, i.e., $g_m(B, w)$ is positive definite almost surely. □

Now we are ready to prove Theorem 1.

Proof of Theorem 1. The proof builds on verifying the conditions for Theorem 1 in [9]. We denote B as a d-dimensional Brownian motion. We first show that $g_m(x, c)$ satisfies following four properties:

(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) $\mathrm{pr}(B \in D(g_m(\cdot, w))) = 0$ 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

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

Lastly, (c) follows from Lemma 3. Since $g_m(\cdot,w)$ 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, $\bar{Y}_T(u)\Rightarrow GB(t)$ in $D[0,1]$ as $T\to\infty$. Then, from Theorem 1 in [9], we have

$$\begin{aligned}
&\Gamma_T = m(m-d)/(d(m-1))(\bar{X}_T-x^*)^T S_m^{-1}(T)(\bar{X}_T-x^*) \\
&\Rightarrow m(m-d)/(d(m-1))B^T(1)g_m(B,w)^{-1}B(1) \text{ as } T\to\infty.
\end{aligned}$$

Moreover, we note that

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

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

References

[1] A. Agarwal, M. J. Wainwright, P. L. Bartlett, and P.K. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.

[2] X. Chen, J. D. Lee, X. T. Tong, and Y Zhang. Statistical inference for model parameters in stochastic gradient descent. https://arxiv.org/pdf/1610.08637.pdf, 2018.

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

[4] Y. Fang, J. Xu, and L. Yang. On scalable inference with stochastic gradient descent. arXiv preprint arXiv:1707.00192, 2017.

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

[6] T.P. Hayes. A large-deviation inequality for vector-valued martingales. available at https://www.cs.unm.edu/ hayes/papers/VectorAzuma/, 2005.

[7] M. Hsieh and P.W. Glynn. Confidence region for stochastic approximation algorithms. In *Winter Simulation Conference*, 2002.

[8] D.P. Kingma and J.L. Ba. ADAM: A method for stochastic optimization. In *ICLR*, 2015.

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

[10] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

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

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

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

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

[15] W. J. Su and Y. Zhu. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. arXiv preprint arXiv:1802.04876, 2018.

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

[17] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116. ACM, 2004.