Fast and Robust Online Inference with Stochastic Gradient Descent via Random Scaling

Sokbae Lee*,[1] **Yuan Liao***,[2] **Myung Hwan Seo***,[3] **Youngki Shin***[4]

[1]Department of Economics, Columbia University, New York, NY 10027, USA
[2]Department of Economics, Rutgers University, New Brunswick, NJ 08901, USA
[3]Corresponding Author, Department of Economics, Seoul National University, Seoul, 08826, Korea
[4]Department of Economics, McMaster University, Hamilton, ON L8S 4L8, Canada
sl3841@columbia.edu, yuan.liao@rutgers.edu, myunghseo@snu.ac.kr, shiny11@mcmaster.ca

Abstract

We develop a new method of online inference for a vector of parameters estimated by the Polyak-Ruppert averaging procedure of stochastic gradient descent (SGD) algorithms. We leverage insights from time series regression in econometrics and construct asymptotically pivotal statistics via random scaling. Our approach is fully operational with online data and is rigorously underpinned by a functional central limit theorem. Our proposed inference method has a couple of key advantages over the existing methods. First, the test statistic is computed in an online fashion with only SGD iterates and the critical values can be obtained without any resampling methods, thereby allowing for efficient implementation suitable for massive online data. Second, there is no need to estimate the asymptotic variance and our inference method is shown to be robust to changes in the tuning parameters for SGD algorithms in simulation experiments with synthetic data.

Introduction

We consider an inference problem for a vector of parameters defined by

$$\beta^* := \arg\min_{\beta\in\mathbb{R}^d} Q(\beta),$$

where $Q(\beta) := \mathbb{E}[q(\beta, Y)]$ is a real-valued population objective function, Y is a random vector, and $\beta \mapsto q(\beta, Y)$ is convex. For a given sample $\{Y_t\}_{t=1}^n$, let β_t denote the stochastic gradient descent (SGD) solution path, that is, for each $t \geq 1$,

$$\beta_t = \beta_{t-1} - \gamma_t \nabla q(\beta_{t-1}, Y_t), \qquad (1)$$

where β_0 is the initial starting value, γ_t is a step size, and $\nabla q(\beta_{t-1}, Y_t)$ denotes the gradient of $q(\beta, Y_t)$ with respect to β at $\beta = \beta_{t-1}$. We study the classical Polyak (1990)-Ruppert (1988) averaging estimator $\bar{\beta}_n := n^{-1}\sum_{t=1}^n \beta_t$. Polyak and Juditsky (1992) established regularity conditions under which the averaging estimator $\bar{\beta}_n$ is asymptotically normal:

$$\sqrt{n}\left(\bar{\beta}_n - \beta^*\right) \xrightarrow{d} \mathcal{N}(0, \Upsilon),$$

where the asymptotic variance Υ has a sandwich form $\Upsilon := H^{-1}SH^{-1}$, $H := \nabla^2 Q(\beta^*)$ is the Hessian matrix and $S := \mathbb{E}\left[\nabla q(\beta^*, Y) \nabla q(\beta^*, Y)'\right]$ is the score variance. The Polyak-Ruppert estimator $\bar{\beta}_n$ can be computed recursively by the updating rule $\bar{\beta}_t = \bar{\beta}_{t-1}\frac{t-1}{t} + \frac{\beta_t}{t}$, which implies that it is well suited to the online setting.

Although the celebrated asymptotic normality result (Polyak and Juditsky 1992) was established about three decades ago, it is only past several years that online inference with $\bar{\beta}_n$ has gained increasing interest in the literature. It is challenging to estimate the asymptotic variance Υ in an online fashion. This is because the naive implementation of estimating it requires storing all data, thereby losing the advantage of online learning. In the seminal work of Chen et al. (2020), the authors addressed this issue by estimating H and S using the online iterated estimator β_t, and recursively updating them whenever a new observation is available. They called this method a *plug-in* estimator and showed that it consistently estimates the asymptotic variance and is ready for inference. However, the plug-in estimator requires that the Hessian matrix be computed to estimate H. In other words, it is necessary to have strictly more inputs than the SGD solution paths β_t to carry out inference. In applications, it can be demanding to compute the Hessian matrix. As an alternative, Chen et al. (2020) proposed a *batch-means* estimator that avoids separately estimating H^{-1} or S. This method directly estimates the variance of the averaged online estimator $\bar{\beta}_n$ by dividing $\{\beta_1, ..., \beta_n\}$ into batches with increasing batch size. The batch-means estimator is based on the idea that correlations among batches that are far apart decay exponentially fast; therefore, one can use nonparametric empirical covariance to estimate Υ. Along this line, Zhu, Chen, and Wu (2021) extended the batch-means approach to allow for real-time recursive updates, which is desirable in the online setting.

The batch-means method produces batches of streaming samples, so that data are weakly correlated when batches far apart. The distance between batches is essential to control dependence among batches so it should be chosen very carefully. In applications, we need to specify a sequence which determines the batch size as well as the speed at which dependence among batches diminish. While this is a new sequence

*These authors contributed equally.

one needs to tune, it also affects the rate of convergence of the estimated covariance matrix. As is shown by Zhu, Chen, and Wu (2021), the optimal choice of this sequence and the batch size is related to the learning rate and could be very slow. Zhu, Chen, and Wu (2021) showed that the batch-mean covariance estimator converges no faster than $O_P(n^{-1/4})$. Simulation results in both Zhu, Chen, and Wu (2021) and this paper show that indeed the coverage probability converges quite slowly.

Instead of estimating the asymptotic variance, Fang, Xu, and Yang (2018) proposed a bootstrap procedure for online inference. Specifically, they proposed to use a large number (say, B) of randomly perturbed SGD solution paths: for all $b = 1, \dots, B$, starting with $\beta_0^{(b)} = \beta_0$ and then iterating

$$\beta_t^{(b)} = \beta_{t-1}^{(b)} - \gamma_t \eta_t^{(b)} \nabla q \left(\beta_{t-1}^{(b)}, Y_t \right), \tag{2}$$

where $\eta_t^{(b)} > 0$ is an independent and identically distributed random variable that has mean one and variance one. The bootstrap procedure needs strictly more inputs than computing $\bar{\beta}_n$ and can be time-consuming.

In this paper, we propose a novel method of online inference for β^*. While the batch-means estimator aims to mitigate the effect of dependence among the averages of SGD iterates, on the contrary, we embrace dependence among them and propose to build a test statistic via random scaling. We leverage insights from time series regression in econometrics (e.g., Kiefer, Vogelsang, and Bunzel 2000) and use a random transformation of β_t's to construct asymptotically pivotal statistics. Our approach does *not* attempt to estimate the asymptotic variance Υ, but studentize $\sqrt{n}\left(\bar{\beta}_n - \beta^*\right)$ via

$$\widehat{V}_n := \frac{1}{n} \sum_{s=1}^{n} \left\{ \frac{1}{\sqrt{n}} \sum_{t=1}^{s} \left(\beta_t - \bar{\beta}_n \right) \right\} \left\{ \frac{1}{\sqrt{n}} \sum_{t=1}^{s} \left(\beta_t - \bar{\beta}_n \right) \right\}'. \tag{3}$$

The resulting statistic is not asymptotically normal but *asymptotically pivotal* in the sense that its asymptotic distribution is free of any unknown nuisance parameters; thus, its critical values can be easily tabulated. Furthermore, the random scaling quantity $\widehat{V}_n$ does not require any additional inputs other than SGD paths β_t and can be updated recursively. As a result, our proposed inference method has a couple of key advantages over the existing methods. First, the test statistic is computed in an online fashion with only SGD iterates and the critical values can be obtained without any resampling methods, thereby allowing for efficient implementation suitable for massive online data. Second, there is no need to estimate the asymptotic variance and our inference method is shown to be robust to changes in the tuning parameters for SGD algorithms in simulation experiments with synthetic data. Table 1 provides a summary comparison between our proposed method and the existing methods.

Related work on SGD The SGD methods pioneered by Robbins and Monro (1951) are popular in the setting of online learning (e.g., Hoffman, Blei, and Bach 2010; Mairal et al. 2010) and have been studied extensively in the recent decade. Among other things, probability bounds on statistical errors have been derived. For instance, Bach and Moulines (2013) showed that for both the square loss and the logistic loss, one may use the smoothness of the loss function to obtain algorithms that have a fast convergence rate without any strong convexity. See Rakhlin, Shamir, and Sridharan (2012) and Hazan and Kale (2014) for related results on convergence rates. Duchi, Hazan, and Singer (2011) proposed AdaGrad, employing the square root of the inverse diagonal Hessian matrix to adaptively control the gradient steps of SGD and derived regret bounds for the loss function. Kingma and Ba (2015) introduced Adam, computing adaptive learning rates for different parameters from estimates of first and second moments of the gradients. Liang and Su (2019) employed a similar idea as AdaGrad to adjust the gradient direction and showed that the distribution for inference can be simulated iteratively. Toulis and Airoldi (2017) developed implicit SGD procedures and established the resulting estimator's asymptotic normality. Anastasiou, Balasubramanian, and Erdogdu (2019) and Mou et al. (2020) developed some results for non-asymptotic inference.

Related work in econometrics The SGD methods are much less popular in econometrics. An early exception is Chen and White (2002), which studied SGD in a Hilbert Space. On the contrary, the random scaling by means of the partial sum process has been actively employed to estimate the so-called long-run variance, which is the sum of all the autocovariances, in the time series econometrics since it was suggested by Kiefer, Vogelsang, and Bunzel (2000). The literature has documented ample evidence that the random scaling stabilizes the excessive finite sample variation in the traditional consistent estimators of the long-run variance; see, e.g., Velasco and Robinson (2001), Sun, Phillips, and Jin (2008), and a recent review in Lazarus et al. (2018). The insight has proved valid in broader contexts, where the estimation of the asymptotic variance is challenging: e.g., spatially dependent data (Kim and Sun 2011), sieve M estimation with time series data (Chen, Liao, and Sun 2014), and a high-dimensional inference problem (Gupta and Seo 2021). We show in this paper that it is indeed useful in online inference, which has not been explored to the best of our knowledge. While our experiments focus on one of the earlier proposals of the random scaling methods, there are numerous alternatives, see e.g. Sun (2014), which warrants future research on the optimal random scaling method.

Notation Let a' and A', respectively, denote the transpose of vector a and matrix A. Let $|a|$ denote the Euclidean norm of vector a and $\|A\|$ the Frobenius norm of matrix A. Also, let $\ell^\infty [0, 1]$ denote the set of bounded continuous functions on $[0, 1]$.

Online Inference

In this section, we first present asymptotic theory that underpins our inference method and describe our proposed online inference algorithm. Then, we explain our method in comparison with the existing methods using the linear regression model as an example.

Table 1: Criteria for Online Inference Methods

Method	FXY (18) Bootstrap	CLTZ (20) Plug-In	CLTZ (20) Batch Means	ZCW (21) Batch Means	This paper Random Scaling
Is it possible					
to avoid resampling?		✓	✓	✓	✓
to avoid Hessian?	✓		✓	✓	✓
to avoid batches?	✓	✓			✓
to update recursively?	✓	✓		✓	✓

Note. FXY (18), CLTZ (20), and ZCW (21) refer to Fang, Xu, and Yang (2018), Chen et al. (2020), and Zhu, Chen, and Wu (2021), respectively.

Functional central limit theorem for online SGD

We first extend Polyak and Juditsky (1992)'s central limit theorem (CLT) to a *functional* CLT (FCLT), that is,

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{[nr]}\left(\beta_t-\beta^*\right)\Rightarrow\Upsilon^{1/2}W\left(r\right),\quad r\in\left[0,1\right],\qquad (4)$$

where $\Rightarrow$ stands for the weak convergence in $\ell^\infty\left[0,1\right]$ and $W\left(r\right)$ stands for a vector of the independent standard Wiener processes on $[0,1]$. That is, the partial sum of the online updated estimates β_t converges weakly to a rescaled Wiener process, with scaling equal to the square root asymptotic variance of the Polyak-Ruppert average. The CLT proved in Polyak and Juditsky (1992) is then a special case with $r=1$. Building on this extension, we propose an online inference procedure. Specifically, using the random scaling matrix $\widehat{V}_n$ defined in (3), we consider the following t-statistic

$$\frac{\sqrt{n}\left(\bar{\beta}_{n,j}-\beta_j^*\right)}{\sqrt{\widehat{V}_{n,jj}}},\qquad (5)$$

where the subscripts j and jj, respectively, denote the j-th element of a vector and the (j,j) element of a matrix. Then, the FCLT yields that the t-statistic is asymptotically pivotal. Note that instead of using an estimate of Υ, we use the random scaling $\widehat{V}_n$ for the proposed t-statistic. As a result, the limit is not conventional standard normal but a mixed normal. It can be utilized to construct confidence intervals for β_j^* for each j. A substantial advantage of this random scaling is that it does not have to estimate an analytic asymptotic variance formula and tends to be more robust in finite samples. As mentioned in the introduction, this random scaling idea has been widely used in the literature known as the fixed bandwidth heteroskedasticity and autocorrelation robust (HAR) inference (e.g., Kiefer, Vogelsang, and Bunzel 2000; Lazarus et al. 2018).

More generally, for any $\ell\leq d$ linear restrictions

$$H_0:R\beta^*=c,$$

where R is an $(\ell\times d)$-dimensional known matrix of rank ℓ and c is an ℓ-dimensional known vector, the conventional Wald test based on $\widehat{V}_n$ becomes asymptotically pivotal. To establish this result formally, we make the following assumptions à la Polyak and Juditsky (1992).

Assumption 1. *(i) There exists a function $\Psi(\beta):\mathbb{R}^d\rightarrow\mathbb{R}$ such that for some $\lambda>0$, $\alpha>0$, $\varepsilon>0$, $L>0$, and all $x,y\in\mathbb{R}^d$, $\Psi(x)\geq\alpha|x|^2$, $|\nabla\Psi(x)-\nabla\Psi(y)|\leq L|x-y|$, $\Psi(\beta^*)=0$, and $\nabla\Psi(\beta-\beta^*)^T\nabla Q(\beta)>0$ for $\beta\neq\beta^*$ hold true. Moreover, $\nabla\Psi(\beta-\beta^*)^T\nabla Q(\beta)\geq\lambda\Psi(\beta)$ for all $|\beta-\beta^*|\leq\varepsilon$.*

(ii) The Hessian matrix H is positive definite and there exist $K_1<\infty$, $\varepsilon>0$, $0<\lambda\leq1$ such that $|\nabla Q(\beta)-H(\beta-\beta^*)|\leq K_1|\beta-\beta^*|^{1+\lambda}$, for all $|\beta-\beta^*|\leq\varepsilon$.

(iii) The sequence $\{\xi_t:=\nabla Q\left(\beta_{t-1}\right)-\nabla q\left(\beta_{t-1},Y_t\right)\}_{t\geq1}$ is a martingale-difference sequence (mds), defined on a probability space $(\Omega,\mathcal{F},\mathcal{F}_t,P)$, i.e., $\mathbb{E}(\xi_t|\mathcal{F}_{t-1})=0$ almost surely, and for some K_2, $\mathbb{E}(|\xi_t|^2|\mathcal{F}_{t-1})+|\nabla Q(\beta_{t-1})|^2\leq K_2(1+|\beta_{t-1}|^2)$ a.s. for all $t\geq1$. Then, the following decomposition takes place: $\xi_t=\xi_t(0)+\zeta_t(\beta_{t-1})$, where $\mathbb{E}(\xi_t(0)|\mathcal{F}_{t-1})=0$ a.s., $\mathbb{E}(\xi_t(0)\xi_t(0)'|\mathcal{F}_{t-1})\xrightarrow{P}S$ as $t\rightarrow\infty$; $S>0$ (S is symmetrical and positive definite), $\sup_t\mathbb{E}(|\xi_t(0)|^2I(|\xi_t(0)|>C|\mathcal{F}_{t-1})\xrightarrow{P}0$ as $C\rightarrow\infty$, and for all t large enough, $\mathbb{E}(|\zeta_t(\beta_{t-1})|^2|\mathcal{F}_{t-1})\leq\delta(\beta_{t-1})$ a.s. with $\delta(\beta)\rightarrow0$ as $\beta\rightarrow0$.

(iv) It holds that $\gamma_t=\gamma_0t^{-a}$ for some $1/2<a<1$.

(v) For $p\geq(1-a)^{-1}$, $\mathbb{E}\left\|\xi_t\right\|^{2p}$ is bounded.

Assumptions 1 (i)–(iii) are identical to Assumptions 3.1–3.3 of Polyak and Juditsky (1992) and Assumption 1 (iv) is the standard learning rate. Assumption 1(v) adds the moment condition to enhance the results for uniform convergence, which is needed to prove the functional CLT.

Given these assumptions, The following theorem is a formal statement under the conditions stated above. Note that the proof of the FCLT requires bounding some processes, indexed by r, uniformly over r. Our proof is new, as some of these processes *cannot* be written as partial sums of martingale differences. Hence we cannot simply apply results such as Doob's inequalities. Recent works, as in Zhu and Dong (2020), developed FCLT using bounded sequences. Our proof extends theirs to possibly unbounded sequences but with finite moments, and uses new technical arguments.

Theorem 1. *Suppose rank$(R)=\ell$. Under Assumption 1 and H_0,*

$$n\left(R\bar{\beta}_n-c\right)'\left(R\widehat{V}_nR'\right)^{-1}\left(R\bar{\beta}_n-c\right)$$
$$\xrightarrow{d}W\left(1\right)'\left(\int_0^1\bar{W}(r)\bar{W}(r)'dr\right)^{-1}W\left(1\right),$$

where W is an ℓ-dimensional vector of the standard Wiener processes and $\bar{W}\left(r\right) := W\left(r\right) - rW\left(1\right)$.

Proof of Theorem 1. Rewrite (1) as

$$\beta_t = \beta_{t-1} - \gamma_t \nabla Q\left(\beta_{t-1}\right) + \gamma_t \xi_t. \tag{6}$$

Let $\Delta_t := \beta_t - \beta^*$ and $\bar{\Delta}_t := \bar{\beta}_t - \beta^*$ to denote the errors in the t-th iterate and that in the average estimate at t, respectively. Then, subtracting β^* from both sides of (6) yields that

$$\Delta_t = \Delta_{t-1} - \gamma_t \nabla Q\left(\beta_{t-1}\right) + \gamma_t \xi_t.$$

Furthermore, for $r \in [0,1]$, introduce a partial sum process

$$\bar{\Delta}_t\left(r\right) := t^{-1} \sum_{i=1}^{[tr]} \Delta_i,$$

whose weak convergence we shall establish.

Specifically, we extend Theorem 2 in Polyak and Juditsky (1992, PJ hereafter) to an FCLT. The first step is a uniform approximation of the partial sum process to another partial sum process $\bar{\Delta}_t^1\left(r\right)$ of

$$\Delta_t^1 := \Delta_{t-1}^1 - \gamma_t H \Delta_{t-1}^1 + \gamma_t \xi_t \quad \text{and} \quad \Delta_0^1 = \Delta_0.$$

That is, we need to show that $\sqrt{t}\sup_r \left|\bar{\Delta}_t\left(r\right) - \bar{\Delta}_t^1\left(r\right)\right| = o_p\left(1\right)$. According to Part 4 in the proof of PJ's Theorem 2, this is indeed the case.

Turning to the weak convergence of $\sqrt{t}\bar{\Delta}_t^1\left(r\right)$, we extend PJ's Theorem 1. Following its decomposition in (A10), write

$$\sqrt{t}\bar{\Delta}_t^1\left(r\right) = I^{(1)}\left(r\right) + I^{(2)}\left(r\right) + I^{(3)}\left(r\right),$$

where

$$I^{(1)}\left(r\right) := \frac{1}{\gamma_0\sqrt{t}}\alpha_{[tr]}\Delta_0,$$

$$I^{(2)}\left(r\right) := \frac{1}{\sqrt{t}}\sum_{j=1}^{[tr]} H^{-1}\xi_j,$$

$$I^{(3)}\left(r\right) := \frac{1}{\sqrt{t}}\sum_{j=1}^{[tr]} w_j^{[tr]}\xi_j,$$

where $\alpha_t = \left(t\gamma_t\right)^{-1} \leq K$ and $\left\{w_j^{[tr]}\right\}$ is a bounded sequence such that $t^{-1}\sum_{j=1}^{t}\left\|w_j^t\right\| \to 0$. Then, $\sup_r\left\|I^{(1)}\left(r\right)\right\| = o_p\left(1\right)$. Suppose for now that $\mathbb{E}\sup_r \|I^{(3)}\|^p = o(1)$ for some $p \geq 1$. The bound for $I^{(3)}$ requires sophisticated arguments, as $w_j\xi_j$ is *not* mds, even though ξ_j is. So we develop new technical arguments to bound this term, whose proof is left at the end of the proof.

Then the FCLT for mds, see e.g. Theorem 4.2 in Hall and Heyde (1980), applies to $I^{(2)}\left(r\right)$, whose regularity conditions are verified in the proof (specifically, Part 1) of the theorem in PJ to apply the mds CLT for $I^{(2)}\left(1\right)$. This shows that $I^{(2)}$ converges weakly to a rescaled Wiener process $\Upsilon^{1/2}W\left(r\right)$. This establishes the FCLT in (4).

Now let $C_n(r) := R\frac{1}{\sqrt{n}}\sum_{t=1}^{[nr]}\left(\beta_t - \beta^*\right)$. Also let $\Lambda = (R\Upsilon R')^{1/2}$, which exists and is invertible as long as $l \leq d$. (4) then shows that for some vector of independent standard Wiener process $W^*(r)$,

$$C_n(r) \Rightarrow \Lambda W^*(r).$$

In addition, $R\widehat{V}_nR' = \frac{1}{n}\sum_{s=1}^{n}[C_n(\frac{s}{n}) - \frac{s}{n}C_n(1)][C_n(\frac{s}{n}) - \frac{s}{n}C_n(1)]'$, where the sum is also an integral over r as $C_n(r)$ is a partial sum process, and $R(\bar{\beta}_n - \beta^*) = \frac{1}{\sqrt{n}}C_n(1)$. Hence $n\left(R\bar{\beta}_n - c\right)'\left(R\widehat{V}_nR'\right)^{-1}\left(R\bar{\beta}_n - c\right)$ is a continuous functional of $C_n(\cdot)$. The desired result in the theorem then follows from the continuous mapping theorem.

It remains to bound $I^{(3)}(r)$ uniformly. Let $S_t = \sum_{j=1}^{t-1} w_j^t\xi_j$. Let $p > (1-a)^{-1}$ and note that

$$\mathbb{E}\sup_r\left\|I^{(3)}\left(r\right)\right\|^{2p} \leq t^{-p}\mathbb{E}\sup_r\left\|S_{[tr]}\right\|^{2p} \leq t^{-p}\sum_{m=1}^{t}\mathbb{E}\left\|S_m\right\|^{2p}$$

and due to Burkholder's inequality (e.g., Hall and Heyde 1980),

$$\begin{aligned}\mathbb{E}\left\|S_m\right\|^{2p} &\leq C_p\mathbb{E}\left|\sum_{j=1}^{m-1}\left\|w_j^m\right\|^2\left\|\xi_j\right\|^2\right|^p \\ &= \sum_{j_1,\ldots,j_p=1}^{m-1}\left\|w_{j_1}^m\right\|^2\cdots\left\|w_{j_p}^m\right\|^2\mathbb{E}\left\|\xi_{j_1}\right\|^2\cdots\left\|\xi_{j_p}\right\|^2,\end{aligned}$$

where the universal constant C_p depends only on p. Note that $\mathbb{E}\left\|\xi_{j_1}\right\|^2\cdots\left\|\xi_{j_p}\right\|^2$ is bounded since $\mathbb{E}\left\|\xi_j\right\|^{2p}$ is bounded. Also, $\sum_{j=1}^{m}\left\|w_j^m\right\|^b = O\left(\sum_{j=1}^{m}\left\|w_j^m\right\|\right)$ for any b due to the boundedness of $\left\|w_j^m\right\|$. According to Lemma 2 in Zhu and Dong (2020), $\sum_{j=1}^{m}\left\|w_j^m\right\| = o\left(m^a\right)$. These facts yield that

$$\mathbb{E}\left\|S_m\right\|^{2p} = O\left(\left(\sum_{j=1}^{m-1}\left\|w_j^m\right\|\right)^p\right) = o(m^{ap}), \tag{7}$$

which holds uniformly for m, k. It in turn implies that

$$\mathbb{E}\sup_r\left\|I^{(3)}\left(r\right)\right\|^{2p} \leq t^{-p}\sum_{m=1}^{t}o\left(m^{ap}\right) = o\left(t^{1+ap-p}\right) = o\left(1\right),$$

as required. □

As an important special case of Theorem 1, the t-statistic defined in (5) converges in distribution to the following pivotal limiting distribution: for each $j = 1, \ldots, d$,

$$\frac{\sqrt{n}\left(\bar{\beta}_{n,j} - \beta_j^*\right)}{\sqrt{\widehat{V}_{n,jj}}} \xrightarrow{d} W_1\left(1\right)\left[\int_0^1\left\{W_1\left(r\right) - rW_1\left(1\right)\right\}^2 dr\right]^{-1/2}, \tag{8}$$

where W_1 is a one-dimensional standard Wiener process.

Related work on Polyak and Juditsky (1992) There exist papers that have extended Polyak and Juditsky (1992) to more general forms (e.g., Kushner and Yang 1993; Godichon-Baggioni 2017; Su and Zhu 2018; Zhu and Dong 2020). The stochastic process defined in Kushner and Yang (1993, equation (2.2)) is different from the partial sum process in (4). Godichon-Baggioni (2017) considers parameters taking values in a separable Hilbert space and as such it considers a generalization of Polyak-Juditsky to more of an empirical process type while our FCLT concerns the partial sum processes. Su and Zhu (2018)'s HiGrad tree divide updates into levels, with the idea that correlations among distant SGD iterates decay rapidly. Their Lemma 2.5 considers the joint asymptotic normality of certain K partial sums for a finite K while our FCLT is for the partial sum process indexed by real numbers on the $[0, 1]$ interval. Zhu and Dong (2020) appears closer than the others to our FCLT for the partial sums, although the set of sufficient conditions is not the same as ours. However, we emphasize that the more innovative part of our work is the way how we utilize the FCLT than the FCLT itself. Indeed, it appears that prior to this paper, there is no other work in the literature that makes use of the FCLT as this paper does in order to conduct online inference with SGD.

An Algorithm for Online Inference

Just as the average SGD estimator can be updated recursively via $\bar\beta_t = \bar\beta_{t-1}\frac{t-1}{t} + \frac{\beta_t}{t}$, the statistic $\widehat{V}_t$ can also be updated in an online fashion. To state the online updating rule for $\widehat{V}_t$, note that

$$\begin{aligned}
t^2\widehat{V}_t &= \sum_{s=1}^{t}\left(\sum_{j=1}^{s}\beta_j - s\bar\beta_t\right)\left(\sum_{j=1}^{s}\beta_j - s\bar\beta_t\right)' \\
&= \sum_{s=1}^{t}\sum_{j=1}^{s}\beta_j\sum_{j=1}^{s}\beta_j' - \bar\beta_t\sum_{s=1}^{t}s\sum_{j=1}^{s}\beta_j' \\
&\quad - \sum_{s=1}^{t}s\sum_{j=1}^{s}\beta_j\bar\beta_t' + \bar\beta_t\bar\beta_t'\sum_{s=1}^{t}s^2
\end{aligned}$$

and

$$\begin{aligned}
\sum_{s=1}^{t}\sum_{j=1}^{s}\beta_j\sum_{j=1}^{s}\beta_j' &= \sum_{s=1}^{t-1}\sum_{j=1}^{s}\beta_j\sum_{j=1}^{s}\beta_j' + t^2\bar\beta_t\bar\beta_t' \\
\sum_{s=1}^{t}s\sum_{j=1}^{s}\beta_j &= \sum_{s=1}^{t-1}s\sum_{j=1}^{s}\beta_j + t^2\bar\beta_t.
\end{aligned}$$

Thus, at step $t-1$, we only need to keep the three quantities, $\bar\beta_{t-1}$,

$$A_{t-1} = \sum_{s=1}^{t-1}\sum_{j=1}^{s}\beta_j\sum_{j=1}^{s}\beta_j', \quad \text{and} \quad b_{t-1} = \sum_{s=1}^{t-1}s\sum_{j=1}^{s}\beta_j,$$

to update $\widehat{V}_{t-1}$ to $\widehat{V}_t$ using the new observation β_t. The following algorithm summarizes the arguments above.

Algorithm 1: Online Inference with SGD via Random Scaling

```
Input: function $q(\cdot)$, parameters $(\gamma_0, a)$ for step size
       $\gamma_t = \gamma_0 t^{-a}$ for $t \geq 1$
Initialize: set initial values for $\beta_0, \bar\beta_0, A_0, b_0$
for $t = 1, 2, \ldots$ do
    Receive: new observation $Y_t$
    $\beta_t = \beta_{t-1} - \gamma_t \nabla q(\beta_{t-1}, Y_t)$
    $\bar\beta_t = \bar\beta_{t-1}\frac{t-1}{t} + \frac{\beta_t}{t}$
    $A_t = A_{t-1} + t^2\bar\beta_t\bar\beta_t'$
    $b_t = b_{t-1} + t^2\bar\beta_t$
    Obtain $\widehat{V}_t$ by
        $\widehat{V}_t = t^{-2}\left(A_t - \bar\beta_t b_t' - b_t\bar\beta_t' + \bar\beta_t\bar\beta_t'\sum_{s=1}^{t}s^2\right)$
    Output: $\bar\beta_t$, $\widehat{V}_t$
end
```

Once $\bar\beta_n$ and $\widehat{V}_n$ are obtained, it is straightforward to carry out inference. For example, we can use the t-statistic in (5) to construct the $(1-\alpha)$ asymptotic confidence interval for the j-th element β_j^* of β^* by

$$\left[\bar\beta_{n,j} - \text{cv}(1-\alpha/2)\sqrt{\frac{\widehat{V}_{n,jj}}{n}},\ \bar\beta_{n,j} + \text{cv}(1-\alpha/2)\sqrt{\frac{\widehat{V}_{n,jj}}{n}}\ \right],$$

where the critical value cv$(1-\alpha/2)$ is tabulated in Abadir and Paruolo (1997, Table I). The limiting distribution in (8) is mixed normal and symmetric around zero. For easy reference, we reproduce the critical values in Table 2. When $\alpha = 0.05$, the critical value is 6.747. Critical values for testing linear restrictions $H_0 : R\beta^* = c$ are given in Kiefer, Vogelsang, and Bunzel (2000, Table II).

Table 2: Asymptotic critical values of the t-statistic

Probability	90%	95%	97.5%	99%
Critical Value	3.875	5.323	6.747	8.613

Note. The table gives one-sided asymptotic critical values that satisfy $\Pr(\hat{t} \leq c) = p$ asymptotically, where $p \in \{0.9, 0.95, 0.975, 0.99\}$. Source: Abadir and Paruolo (1997, Table I).

Estimation of the linear regression model

In this subsection, we consider the least squares estimation of the linear regression model $y_t = x_t'\beta^* + \varepsilon_t$. In this example, the stochastic gradient sequence β_t is given by

$$\beta_t = \beta_{t-1} - \gamma_t x_t\left(x_t'\beta_{t-1} - y_t\right),$$

where $\gamma_t = \gamma_0 t^{-\alpha}$ with $1/2 < \alpha < 1$ is the step size. This linear regression model satisfies Assumption 1, as shown by Polyak and Juditsky (1992, section 5). The asymptotic variance of $\bar{\beta}_n$ is $\Upsilon = H^{-1}SH^{-1}$ where $H = \mathbb{E}x_t x_t'$ and $S = \mathbb{E}x_t x_t' \epsilon_t^2$.

Our proposed method would standardize using $\widehat{V}_n$ according to (3), which does *not* consistently estimate Υ. We use critical values as tabulated in Table 2, whereas the existing methods (except for the bootstrap) would seek for consistent estimation of Υ. For instance, the plug-in method respectively estimates H and S by $\widehat{H} = \frac{1}{n}\sum_{t=1}^n x_t x_t'$, and $\widehat{S} = \frac{1}{n}\sum_{t=1}^n x_t x_t' \widehat{\epsilon}_t^2$, where $\widehat{\epsilon}_t = y_t - x_t'\beta_{t-1}$. Note that $\widehat{H}^{-1}$ does not rely on the updated β_t but may not be easy to compute if $\dim(x_t)$ is moderately large. Alternatively, the batch-mean method first splits the iterates β_t's into $M+1$ batches, discarding the first batch as the burn-in stage, and estimates Υ directly by

$$\widehat{\Upsilon}_1 = \frac{1}{M}\sum_{k=1}^{M} n_k(\widehat{\beta}_k - \bar{\beta}_n)(\widehat{\beta}_k - \bar{\beta}_n)',$$

where $\widehat{\beta}_k$ is the mean of β_t's for the k-th batch and n_k is the batch size. One may also discard the first batch when calculating $\bar{\beta}_n$ in $\widehat{\Upsilon}_1$. As noted by Zhu, Chen, and Wu (2021), a serious drawback of this approach is that one needs to know the total number of n as a priori, so one needs to recalculate $\widehat{\Upsilon}_1$ whenever a new observation arrives. Instead, Zhu, Chen, and Wu (2021) proposed a "fully online-fashion" covariance estimator, which splits the iterates β_t into n batches $B_1, ..., B_n$, and estimates the covariance by

$$\widehat{\Upsilon}_2 = \frac{1}{\sum_{k=1}^n |B_k|}\sum_{k=1}^{n}(S_k - |B_k|\bar{\beta}_n)(S_k - |B_k|\bar{\beta}_n)',$$

where S_k denotes the sum of all elements in B_k and $|B_k|$ denotes the size of the k-th batch. The batches are overlapped. For instance, fix a pre-determined sequence $\{1, 3, 5, 7, ...\}$, we can set

$$\begin{aligned} B_1 &= \{\beta_1\}, \quad B_2 = \{\beta_1, \beta_2\}, \\ B_3 &= \{\beta_3\}, \quad B_4 = \{\beta_3, \beta_4\}, \\ B_5 &= \{\beta_5\}, \quad B_6 = \{\beta_5, \beta_6\}, \end{aligned}$$

and subsequent B_t's are defined analogously. Our proposed scaling $\widehat{V}_n$ is similar to $\widehat{\Upsilon}_2$ in the sense that it can be formulated as:

$$\widehat{V}_n = \frac{1}{n^2}\sum_{k=1}^{n}(S_k - |B_k^*|\bar{\beta}_n)(S_k - |B_k^*|\bar{\beta}_n)'$$

with a particular choice of batches being:

$$\begin{aligned} B_1^* &= \{\beta_1\}, B_2^* = \{\beta_1, \beta_2\}, B_3^* = \{\beta_1, \beta_2, \beta_3\}, ..., \\ B_k^* &= B_{k-1}^* \cup \{\beta_k\}, ... \end{aligned}$$

However, there is a key difference between $\widehat{V}_n$ and $\widehat{\Upsilon}_2$: the batches used by $\widehat{\Upsilon}_2$, though they can be overlapped, are required to be weakly correlated as they become far apart. In contrast, B_k^* are strongly correlated and strictly nested. Thus, we embrace dependences among B_k^*, and reach the scaling $\widehat{V}_n$ that does not consistently estimate Υ. The important advantage of our approach is that there is no need to choose the batch size.

In the next section, we provide results of experiments that compare different methods in the linear regression model.

Experiments

In this section we investigate the numerical performance of the random scaling method via Monte Carlo experiments. We consider two baseline models: linear regression and logistic regression. We use the Compute Canada Graham cluster composed of Intel CPUs (Broadwell, Skylake, and Cascade Lake at 2.1GHz–2.5GHz) and they are assigned with 3GB memory. The replication code is available at https://github.com/SGDinference-Lab/AAAI-22.

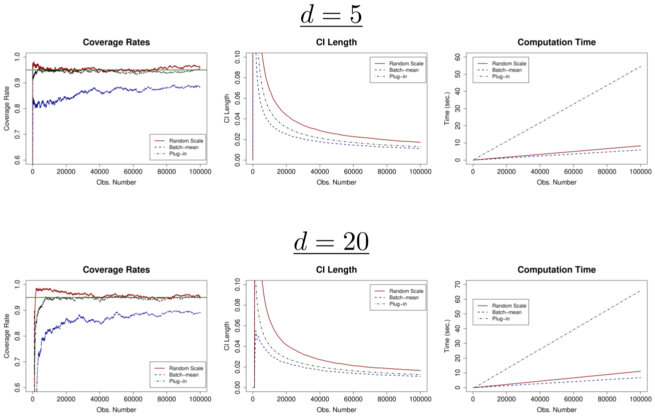
Linear Regression: The data are generated from

$$y_t = x_t'\beta^* + \varepsilon_t \quad \text{for} \quad t = 1, \dots, n,$$

where x_t is a d-dimensional vector of covariates generated from the multivariate normal distribution $\mathcal{N}(0, I_d)$, ε_t is from $N(0,1)$, and β^* is equi-spaced on the interval $[0,1]$. This experimental design is the same as that of Zhu, Chen, and Wu (2021). The dimension of x is set to $d = 5, 20$. We consider different combination of the learning rate $\gamma_t = \gamma_0 t^{-a}$ by setting $\gamma_0 = 0.5, 1$ and $a = 0.505, 0.667$. The sample size set to be $n = 100000$. The initial value β_0 is set to be zero. In case of $d = 20$, we burn in around 1% of observations and start to estimate $\bar{\beta}_t$ from $t = 1000$. Finally, the simulation results are based on 1000 replications.

We compare the performance of the proposed random scaling method with the state-of-the-art methods in the literature, especially the plug-in method in Chen et al. (2020) and the recursive batch-mean method in Zhu, Chen, and Wu (2021).The performance is measured by three statistics: the coverage rate, the average length of the 95% confidence interval, and the average computation time. Note that the nominal coverage probability is set at 0.95. For brevity, we focus on the first coefficient β_1 hereafter. The results are similar across different coefficients.

Figure 1: Linear Regression: $d \in \{5, 20\}$, $\gamma_0 = 0.5$, $a = 0.505$ for $\gamma_t = \gamma_0 t^{-a}$

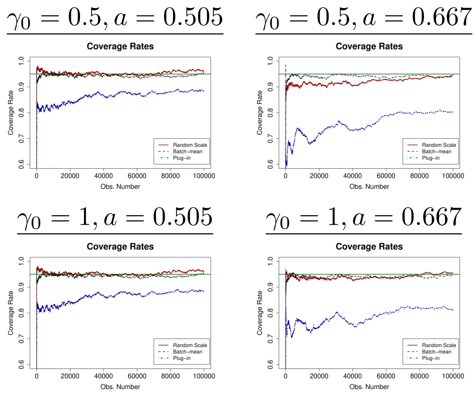


$\underline{d = 5}$

$\underline{d = 20}$

Figures 1–2 summarize the simulation results. The complete set of simulation results are reported in the Appendix. In Figure 1, we adopt the same learning rate parameters as in Zhu, Chen, and Wu (2021): $\gamma_0 = 0.5$ and $a = 0.505$. Overall, the performance of the random scaling method is satisfactory. First, the random scaling and plug-in methods show

Figure 2: Linear Regression: $d = 5$, $\gamma_0 \in \{0.5, 1\}$, $a \in \{0.505, 0.667\}$ for $\gamma_t = \gamma_0 t^{-a}$


$\gamma_0 = 0.5, a = 0.505$ $\gamma_0 = 0.5, a = 0.667$

$\gamma_0 = 1, a = 0.505$ $\gamma_0 = 1, a = 0.667$

better coverage rates. The coverage rate of the batch-mean method deviates more than 5% from the nominal rate even at $n = 100000$. Second, the batch-mean method shows the smallest average length of the confidence interval followed by the plug-in and the random scaling methods. Third, the plug-in method requires substantially more time for computation than the other two methods. The random scaling method takes slightly more computation time than the batch-mean method. Finally, we check the robustness of the performance by changing the learning rates in Figure 2, focusing on the case of $d = 5$. Both the random scaling method and the plug-in method are robust to the changes in the learning rates in terms of the coverage rate. However, the batch-mean method converges slowly when $a = 0.667$ and it deviates from the nominal rates about 15% even at $n = 100000$.

Logistic Regression: We next turn our attention to the following logistic regression model:

$$y_t = 1(x_t'\beta^* - \varepsilon_t \geq 0) \text{ for } t = 1, \ldots, n,$$

where ε_t follows the standard logistic distribution and $1(\cdot)$ is the indicator function. We consider a large dimension of x_t ($d = 200$) as well as $d = 5, 20$. All other settings are the same as the linear model.

Table 3: Logistic Regression, $n = 10^5$, $\gamma_0 = 0.5$, $a = 0.505$ for $\gamma_t = \gamma_0 t^{-a}$

	$d = 5$	$d = 20$	$d = 200$
Random Scale			
Coverage	0.930	0.929	0.919
Length	0.036	0.043	0.066
Time (sec.)	8.4	11.4	170.3
Batch-mean			
Coverage	0.824	0.772	0.644
Length	0.022	0.024	0.027
Time (sec.)	6.0	7.0	10.7
Plug-in			
Coverage	0.953	0.946	0.944
Length	0.029	0.035	0.053
Time (sec.)	55.2	66.8	955.0

Overall, the simulation results are similar to those in linear regression. Table 3 summarizes the simulation results of a single design. The coverage rates of Random Scale and Plug-in are satisfactory while that of Batch-mean is 30% lower when $d = 200$. Random Scale requires more computation time than Batch-mean but is still much faster than Plug-in. The computation time of Random Scale can be substantially reduced when we are interested in the inference of a single parameter. In such a case, we need to update only a single element of $\hat{V}$ rather than the whole $d \times d$ matrix. In Table 4, we show that Random Scale can be easily scaled up to $d = 800$ with only 11.7 seconds computation time when we are interested in the inference of a single parameter. Finally, the results in the appendix reinforce our findings from the linear regression design that the performance of Random-scale is less sensitive to the choice of tuning parameters than Batch-mean.

Table 4: Logistic Regression: Random Scale Updating a Single Element of $\hat{V}$, $n = 10^5$, $\gamma_0 = 0.5$, $a = 0.505$ for $\gamma_t = \gamma_0 t^{-a}$

	$d = 5$	$d = 20$	$d = 200$	$d = 500$	$d = 800$
Coverage	0.930	0.929	0.919	0.927	0.931
Length	0.037	0.043	0.066	0.133	0.196
Time (sec.)	5.0	5.3	6.7	9.7	11.7

Acknowledgments

We are grateful to the authors of Zhu, Chen, and Wu (2021) for providing us their code. This work was supported in part by the European Research Council (ERC-2014-CoG-646917-ROMIA) and by the UK Economic and Social Research Council (ESRC) through research grant (ES/P008909/1) to the Centre for Microdata Methods and Practice. This research was enabled in part by support provided by McMaster COVID-19 Research Fund (Stream 2), Compute Ontario (https://www.computeontario.ca/) and Compute Canada (http://www.computecanada.ca). Financial support from Korea Bureau of Economic Research in the Institute of Economic Research of Seoul National University is gratefully acknowledged. We are grateful to Xiang Li, Jiadong Liang, Xin Tong, Yichen Zhang, and anonymous referees for their valuable comments.

References

Abadir, K. M.; and Paruolo, P. 1997. Two Mixed Normal Densities from Cointegration Analysis. *Econometrica*, 65(3): 671–680.

Anastasiou, A.; Balasubramanian, K.; and Erdogdu, M. A. 2019. Normal Approximation for Stochastic Gradient Descent via Non-Asymptotic Rates of Martingale CLT. In Beygelzimer, A.; and Hsu, D., eds., *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, 115–137.

Bach, F.; and Moulines, E. 2013. Non-strongly-convex smooth stochastic approximation with convergence rate O

(1/n). In *Advances in Neural Information Processing Systems (NIPS)*.

Chen, X.; Lee, J. D.; Tong, X. T.; and Zhang, Y. 2020. Statistical inference for model parameters in stochastic gradient descent. *Annals of Statistics*, 48(1): 251–273.

Chen, X.; Liao, Z.; and Sun, Y. 2014. Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics*, 178: 639–658.

Chen, X.; and White, H. 2002. Asymptotic properties of some projection-based Robbins-Monro procedures in a Hilbert space. *Studies in Nonlinear Dynamics & Econometrics*, 6(1).

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7): 2121–2159.

Fang, Y.; Xu, J.; and Yang, L. 2018. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19(1): 1–21.

Godichon-Baggioni, A. 2017. A Central Limit Theorem for averaged stochastic gradient algorithms in Hilbert spaces and online estimation of the asymptotic variance. Application to the Geometric Median and Quantiles. *arXiv:1702.00931v1 [math.ST]*, available at https://arxiv.org/pdf/1702.00931v1.pdf.

Gupta, A.; and Seo, M. H. 2021. Robust Inference on Infinite and Growing Dimensional Regression. *arXiv:1911.08637 [econ.EM]*, available at https://arxiv.org/abs/1911.08637v2.

Hall, P.; and Heyde, C. C. 1980. *Martingale Limit Theory and Its Application*. Academic Press, Boston,.

Hazan, E.; and Kale, S. 2014. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1): 2489–2512.

Hoffman, M. D.; Blei, D. M.; and Bach, F. 2010. Online Learning for Latent Dirichlet Allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'10, 856–864.

Kiefer, N. M.; Vogelsang, T. J.; and Bunzel, H. 2000. Simple robust testing of regression hypotheses. *Econometrica*, 68(3): 695–714.

Kim, M. S.; and Sun, Y. 2011. Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix. *Journal of Econometrics*, 160(2): 349–371.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kushner, H. J.; and Yang, J. 1993. Stochastic Approximation with Averaging of the Iterates: Optimal Asymptotic Rate of Convergence for General Processes. *SIAM Journal on Control and Optimization*, 31(4): 1045–1062.

Lazarus, E.; Lewis, D. J.; Stock, J. H.; and Watson, M. W. 2018. HAR inference: Recommendations for practice. *Journal of Business & Economic Statistics*, 36(4): 541–559.

Liang, T.; and Su, W. J. 2019. Statistical inference for the population landscape via moment-adjusted stochastic gradients. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2): 431–456.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2010. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1): 19–60.

Mou, W.; Li, C. J.; Wainwright, M. J.; Bartlett, P. L.; and Jordan, M. I. 2020. On Linear Stochastic Approximation: Fine-grained Polyak-Ruppert and Non-Asymptotic Concentration. In Abernethy, J.; and Agarwal, S., eds., *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, 2947–2997.

Polyak, B. T. 1990. New method of stochastic approximation type. *Automation and Remote Control*, 51(7): 937–946.

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

Rakhlin, A.; Shamir, O.; and Sridharan, K. 2012. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, 1571–1578.

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

Ruppert, D. 1988. Efficient estimations from a slowly convergent Robbins–Monro process. Technical Report 781, Cornell University Operations Research and Industrial Engineering. Available at https://ecommons.cornell.edu/bitstream/handle/1813/8664/TR000781.pdf?sequence=1.

Su, W. J.; and Zhu, Y. 2018. Uncertainty Quantification for Online Learning and Stochastic Approximation via Hierarchical Incremental Gradient Descent. *arXiv:1802.04876 [stat.ML]*, available at https://arxiv.org/abs/1802.04876.

Sun, Y. 2014. Fixed-smoothing asymptotics in a two-step generalized method of moments framework. *Econometrica*, 82(6): 2327–2370.

Sun, Y.; Phillips, P. C.; and Jin, S. 2008. Optimal bandwidth selection in heteroskedasticity–autocorrelation robust testing. *Econometrica*, 76(1): 175–194.

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

Velasco, C.; and Robinson, P. M. 2001. Edgeworth expansions for spectral density estimates and studentized sample mean. *Econometric Theory*, 17(3): 497–539.

Zhu, W.; Chen, X.; and Wu, W. B. 2021. Online Covariance Matrix Estimation in Stochastic Gradient Descent. *Journal of the American Statistical Association*, AHEAD-OF-PRINT: 1–12. Available at https://doi.org/10.1080/01621459.2021.1933498.

Zhu, Y.; and Dong, J. 2020. On Constructing Confidence Region for Model Parameters in Stochastic Gradient Descent via Batch Means. *arXiv:1911.01483 [stat.ML]*, available at https://arxiv.org/abs/1911.01483.