

# On the Utility of Equal Batch Sizes for Inference in Stochastic Gradient Descent

**Rahul Singh**

WRAHULSINGH@GMAIL.COM

*Department of Mathematics*

*Indian Institute of Technology Delhi, Delhi - 110016, India;*

*Indian Institute of Technology Delhi - Abu Dhabi*

*Khalifa City B, Abu Dhabi, UAE*

**Abhineek Shukla**

ABHUSHUKLA@GMAIL.COM

*Department of Statistics and Data Science*

*National University of Singapore, Singapore - 117543*

**Dootika Vats**

DOOTIKA@IITK.AC.IN

*Department of Mathematics and Statistics*

*Indian Institute of Technology Kanpur, Kanpur - 208016, India*

**Editor:** Alexandre Bouchard

## Abstract

Stochastic gradient descent (SGD) is an estimation tool for large data employed in machine learning and statistics. Due to the Markovian nature of the SGD process, inference is a challenging problem. An underlying asymptotic normality of the averaged SGD (ASGD) estimator allows for the construction of a batch-means estimator of the asymptotic covariance matrix. Instead of the usual increasing batch-size strategy, we propose a memory efficient equal batch-size strategy and show that under mild conditions, the batch-means estimator is consistent. A key feature of the proposed batching technique is that it allows for bias-correction of the variance, at no additional cost to memory. Further, since joint inference for large dimensional problems may be undesirable, we present marginal-friendly simultaneous confidence intervals, and show through an example on how covariance estimators of ASGD can be employed for improved predictions.

**Keywords:** Batch-means, Bias correction, Covariance estimation, Confidence regions.

## 1. Introduction

Stochastic gradient descent (SGD) is a popular and efficient optimization technique seminally introduced by Robbins and Monro (1951). Given the nature of modern data, the increasing popularity of SGD is natural, owing to computational efficiency for large datasets, and compatibility in online settings (see, e.g., Bottou, 2010; Bottou et al., 2018; Wilson et al., 2017).

We assume data arise from  $\Pi$ , a probability distribution on  $\mathbb{R}^r$ , denoted by  $\zeta \sim \Pi$ . In a model fitting paradigm, a function  $f : \mathbb{R}^d \times \mathbb{R}^r \rightarrow \mathbb{R}$  typically measures empirical loss for estimating a parameter  $\theta$ , having observed the data,  $\zeta$ . Denote the expected loss as  $F(\theta) = \mathbb{E}_{\zeta \sim \Pi} [f(\theta, \zeta)]$ . The main parameter of interest is  $\theta^* \in \mathbb{R}^d$  where

$$\theta^* = \arg \min_{\theta \in \mathbb{R}^d} F(\theta).$$

With data  $\zeta_i \stackrel{\text{iid}}{\sim} \Pi$  for  $i = 1, \dots, n$ , the goal is to estimate  $\theta^*$ , where iid refers to independently and identically distributed. This invariably involves a gradient based technique. Often,  $F(\theta)$  is unavailable and a first approximation step is to replace  $F(\theta)$  with the empirical loss,  $n^{-1} \sum_{i=1}^n f(\theta, \zeta_i)$ . When the data are online or when calculation of the complete gradient vector is expensive, a further adjustment is made by replacing the complete gradient with an unbiased estimate. This yields a large class of stochastic gradient algorithms. Denote  $\nabla f(\theta, \zeta)$  as the gradient vector of  $f(\theta, \zeta)$  with respect to  $\theta$ ,  $\eta_i > 0$  as a learning rate, and  $\theta_0$  as the starting point of an SGD process. The  $i^{\text{th}}$  iterate of SGD is:

$$\theta_i = \theta_{i-1} - \eta_i \nabla f(\theta_{i-1}, \zeta_i), \quad \text{for } i = 1, 2, \dots$$

Despite the approximations introduced in the optimization, SGD estimates of  $\theta$  can have nice statistical properties (Fabian, 1968; Ruppert, 1988; Polyak and Juditsky, 1992), particularly when  $\eta_i$  is appropriately decreasing and the estimator of  $\theta^*$  is chosen to be the averaged SGD (ASGD):

$$\hat{\theta}_n := n^{-1} \sum_{i=1}^n \theta_i.$$

Naturally, a point estimate of  $\theta^*$  alone is not sufficient. The work of Polyak and Juditsky (1992) has particularly been instrumental in building a framework for statistical inference for  $\hat{\theta}_n$ . Let  $A := \nabla^2 F(\theta^*)$  denote the Hessian of  $F(\theta)$  evaluated at  $\theta = \theta^*$  and define  $S := \mathbb{E}_{\Pi} ([\nabla f(\theta^*, \zeta)][\nabla f(\theta^*, \zeta)]^\top)$ . When the derivative and expectation are interchangeable,  $\mathbb{E}_{\Pi} [\nabla f(\theta^*, \zeta)] = \nabla F(\theta^*) = 0$ . Polyak and Juditsky (1992) showed that if  $F$  is strictly convex with a Lipschitz gradient and  $\eta_i = \eta i^{-\alpha}$  with  $\alpha \in (0.5, 1)$ , then  $\hat{\theta}_n$  is a consistent estimator of  $\theta^*$ , and under some additional conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(0, \Sigma) \text{ as } n \rightarrow \infty, \text{ where } \Sigma = A^{-1}SA^{-1}. \quad (1)$$

For a true end-to-end analysis, in addition to estimating  $\theta^*$ , a practitioner would be interested in assessing the quality of this estimator by estimating  $\Sigma$ , employing estimators of  $\Sigma$  for inference, and equipping predictions with uncertainty estimates. Thus, statistical inference for model parameters is a way forward towards robust implementations of machine learning algorithms. Although there is adequate literature devoted to the convergence behavior of the ASGD estimator and its variants (Zhang, 2004; Nemirovski et al., 2008; Agarwal et al., 2012; Zhu and Dong, 2021), estimators of  $\Sigma$  have only recently been developed (Chen et al., 2020, 2022; Fang et al., 2018; Leung and Chan, 2024; Zhu et al., 2021). Robustness and quality of inference depend critically on the quality of estimation of  $\Sigma$ .

Chen et al. (2020) proposed two consistent estimators of  $\Sigma$ : an expensive plug-in estimator that requires repeated computation of the inverse of a Hessian, and a variant of the traditional batch-means estimator of Chen and Seila (1987) that is cheap to implement. In a batch-means estimator, SGD iterates are broken into batches of possibly differing sizes. A weighted sample covariance of the resulting batch-mean vectors yields a batch-means estimator; the quality of estimation is affected by the choice of batch-sizes. Zhu et al. (2021) proposed a novel increasing batch-size strategy where the size of the batches continually increases until saturation, at which point a new batch is created. We refer to this estimator as the increasing batch-size (IBS) estimator.

Despite the novel batching strategy, finite-sample performance for the IBS estimator is underwhelming; we demonstrate this in a variety of examples. A primary reason for the under-performance is that as iteration size increases, the sample mean vectors of all but the last batch cannot improve in quality. We employ equal batch-sizes, that are carefully chosen for both practical utility and theoretical guarantees. Specifically, our proposed batch-sizes are powers of two, where the powers increase as a function of the iteration length. Under mild conditions, our batching strategy yields a consistent estimator and we obtain mean-square-error bounds.

Equal batch-size (EBS) batch-means estimators are common-place in the Markov chain Monte Carlo (MCMC) literature (see, e.g. Geyer, 1992; Jones et al., 2006; Flegal and Jones, 2010). However, the Markov chains generated by MCMC and SGD are fundamentally different; MCMC typically produces a time-homogeneous, stationary, ergodic chain and SGD produces a time-inhomogeneous and non-stationary chain that converges to a Dirac mass distribution (Dieuleveut et al., 2020). Due to these differences, the existing theoretical results of the batch-means estimator of MCMC are not applicable for SGD. However, as we will see, the tools utilized in output analysis for MCMC find use in setting up a workflow for statistical inference in SGD.

The proposed doubling batching structure is developed to allow for finite-sample improvements in the estimation of  $\Sigma$ . As discussed in Chen et al. (2020); Zhu et al. (2021), due to the Markovian structure of  $\{\theta_i\}_{i=1}^n$ , the estimator of  $\Sigma$  is often under-biased for any finite  $n$ ; this bias is one of the primary reasons for the underwhelming inferential performance of most estimators of  $\Sigma$ . Our batching technique allows for a memory-efficient, consistent, and a bias-reduced estimator of  $\Sigma$  using the lugsail technique of Vats and Flegal (2022). Such a bias-reduction technique cannot be applied to the IBS estimator of Zhu et al. (2021).

It is worth considering how practitioners are expected to use estimators of  $\Sigma$ : two potential uses are as follows. First, estimating  $\Sigma$  allows for an interpretation on the quality of estimation of  $\theta^*$ , particularly in smaller dimensions. Secondly, estimators of  $\Sigma$  can then be employed to yield estimators of variability of functions of  $\hat{\theta}_n$ , aiding in uncertainty quantification for predictions. We explore this second point more carefully in Section 6. Further, using the asymptotic normality in (1) and consistent estimators of  $\Sigma$ , it is possible to implement traditional multivariate hypothesis tests. That is, a consistent estimator of  $\Sigma$  can be used to construct an ellipsoidal confidence region for  $\theta^*$  using (1). In high dimensional prediction models, where testing may not be a priority, such confidence regions cease to convey a useful interpretation. Instead, simultaneous hyper-rectangular confidence regions allow easy interpretations for every marginal component, while also retaining coverage of the confidence region. This naturally, comes at the cost of the volume of the confidence region. We term such hyper-rectangular regions as “marginal-friendly”. So far, estimators of  $\Sigma$  have been employed to make either uncorrected marginal confidence intervals, or uninterpretable ellipsoidal confidence regions. Adapting tools developed in stochastic simulation, we construct marginal-friendly confidence regions with simultaneous coverage that utilize consistent estimators of  $\Sigma$ .

The rest of the paper is organized as follows. In Section 2 we present our proposed batching strategy. Assumptions and proof of consistency of the resulting batch-means estimator are in Section 3. Section 4 describes the under-estimation problem in estimating

$\Sigma$ , and discusses bias-correction through a lugsail estimator, for which we obtain the same rate of convergence as the original batch-means estimator. Section 5 presents the structure of the marginal-friendly confidence regions. The performance of our proposed estimator is demonstrated through two simulated data problems in Section 6, where the benefits of our proposed estimator are highlighted. In this section, we also detail how estimators of  $\Sigma$  may be employed to improve predictions in classification problems. The method is applied to four data sets to demonstrate improvements in prediction accuracy. All proofs are presented in the Appendix.

## 2. Proposed batch-means estimator

This section presents some well known batching strategies for estimating variance covariance matrix and a clear motivation for considering the proposed equal batch means estimator in the current setting.

### 2.1 General batch-means estimator

Batch-means estimators and its variants are critical components of output analysis methods in steady-state simulation. A general batch-means estimator can be set up in the following way. For iteration size  $n$ , the SGD iterates (after some user-chosen warm-up) are divided into  $K$  batches with batch-sizes  $b_{n,1}, \dots, b_{n,K}$ . Define  $\tau_0 = 0$  and let  $\tau_k = \sum_{j=1}^k b_{n,j}$  for  $k = 1, 2, \dots, K$ , denote the ending index for the  $k$ th batch. Then the batches are:

$$\underbrace{\{\theta_1, \dots, \theta_{\tau_1}\}}_{1^{\text{st}} \text{ batch}}, \underbrace{\{\theta_{\tau_1+1}, \dots, \theta_{\tau_2}\}}_{2^{\text{nd}} \text{ batch}}, \dots, \underbrace{\{\theta_{\tau_{K-1}+1}, \dots, \theta_{\tau_K}\}}_{K^{\text{th}} \text{ batch}}.$$

Let  $\bar{\theta}_k = b_{n,k}^{-1} \sum_{i=\tau_{k-1}+1}^{\tau_k} \theta_i$  denote the mean vector of the  $k^{\text{th}}$  batch. A general batch-means estimator is

$$\hat{\Sigma}_{\text{gen}} = \frac{1}{K} \sum_{k=1}^K b_{n,k} \left( \bar{\theta}_k - \hat{\theta}_n \right) \left( \bar{\theta}_k - \hat{\theta}_n \right)^{\top}. \quad (2)$$

Batch-means estimators of limiting covariances are commonplace in steady-state simulation (Alexopoulos and Goldsman, 2004; Chen and Seila, 1987; Chien et al., 1997; Glynn and Whitt, 1991; Muñoz and Glynn, 1997; Song and Schmeiser, 1995) and MCMC (Chakraborty et al., 2022; Liu and Flegal, 2018; Vats et al., 2019; Flegal and Jones, 2010). Their performance is critically dependent on the batching structure; this choice is process dependent and much work has gone into their study for ergodic and stationary Markov chains (Damerджи, 1995; Liu et al., 2022).

In the context of SGD, batch-means estimators were recently adopted in the sequence of works by Chen et al. (2020); Zhu and Dong (2021); Zhu et al. (2021). For  $\eta_i = i^{-\alpha}$ ,  $\alpha \in (1/2, 1)$ , the batch-size chosen by Zhu et al. (2021) is

$$b_{n,k} \propto k^{\frac{1+\alpha}{1-\alpha}}. \quad (3)$$

The above choice is motivated by the following argument: if  $b_{n,k}$  is reasonably large, the batch-mean vector is approximately normally distributed. Using Chen et al. (2020, Equation

15), for large  $j$  and  $k$  ( $> j$ ), the strength of correlation between  $\theta_j$  and  $\theta_k$  is

$$\prod_{i=j}^{k-1} \|I_d - \eta_{i+1}A\| \leq \exp\left(-\lambda_{\min}(A) \sum_{i=j}^{k-1} \eta_{i+1}\right), \quad (4)$$

where  $\lambda_{\min}(A)$  is the smallest eigenvalue of  $A$ . Consequently, if  $\sum_{i=j}^{k-1} \eta_{i+1}$  is sufficiently large, the  $j^{\text{th}}$  and  $k^{\text{th}}$  iterates are approximately uncorrelated. Therefore, for large  $b_{n,k}$ 's, batch-mean vectors are approximately independent and normally distributed. The batch-size in (3) is such that  $\sum_{i=j}^{k-1} \eta_{i+1}$  is sufficiently large. This reasoning ignores the dangers of choosing large batch-sizes. For any given iteration length, larger batch-sizes implies smaller number of batches leading to high variance and/or singular estimators of  $\Sigma$ . Consequently, the quality of inference is challenged and multivariate inference becomes difficult.

## 2.2 Proposed batching strategy

Under an equal batch-size strategy,  $b_{n,k} = b_n$  for all  $k$ ; the number of batches is  $a_n := K = \lfloor n/b_n \rfloor$ . With this choice, the estimator in (2) simplifies to

$$\hat{\Sigma}_{b_n} = a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{\theta}_k - \hat{\theta}_n)(\bar{\theta}_k - \hat{\theta}_n)^\top = \frac{b_n}{a_n} \sum_{k=1}^{a_n} \bar{\theta}_k \bar{\theta}_k^\top - b_n \hat{\theta}_n \hat{\theta}_n^\top. \quad (5)$$

Choosing  $b_n \propto \lfloor n^\beta \rfloor$  for some  $\beta \in (0, 1)$  seems natural, and is often considered in stochastic simulation. We consider memory-efficient batch-sizes that are powers of two, and still grow polynomially. That is, for some  $c > 0$  and current iterate  $n$ , we consider batch-sizes of the following form:

$$b_n^* = \min\{2^\gamma : cn^\beta \leq 2^\gamma \text{ for } \gamma \in \mathbb{N}\}. \quad (6)$$

That is,  $b_n^*$  is the smallest power of 2 that is bounded below by  $cn^\beta$ . Naturally,  $cn^\beta \leq b_n^* \leq 2cn^\beta$ , and  $a_n^* := n/b_n^*$  for any given  $n$  is bounded like  $n^{1-\beta}/(2c) \leq a_n^* \leq n^{1-\beta}/c$ . A similar batch-size strategy was hinted at in Gong and Flegal (2016). A pictorial demonstration of the batching strategy is in Figure 1, for the settings discussed in Section 6. Naturally, as  $n \rightarrow \infty$ , both  $b_n^*$  and  $a_n^*$  tend to  $\infty$ . The proposed batching structure reduces storage costs to only  $a_n$  batch-mean vectors at any given stage;  $a_n = O(n^{1-\beta})$  dramatically smaller than  $n$ . Further, for the iterate of  $n$  when batch-size changes, new batches are just made by averaging over adjacent batch-mean vectors; at these moments the number of batches gets halved. Our theoretical results hold for a general of equal batch-sizes and in our simulations we implement both  $b_n^*$  and  $b_n = \lfloor cn^\beta \rfloor$ .

## 3. Main results

Consistency of  $\hat{\Sigma}_{b_n}$  along with the asymptotic normality result of Polyak and Juditsky (1992) in (1), allows for large-sample inferential procedures, similar to traditional maximum likelihood estimation. For this task, we make assumptions that ensure both the asymptotic normality in (1) and consistency of the covariance estimator.

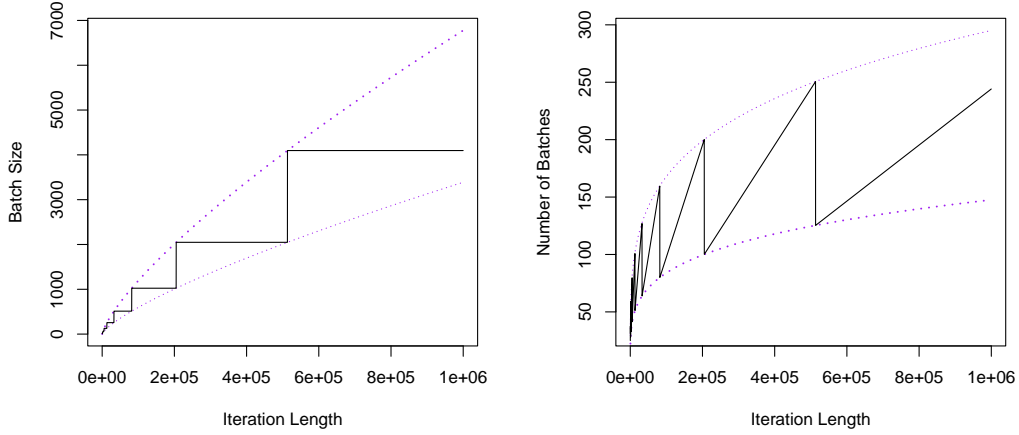


Figure 1: Pictorial demonstration of the proposed batching structure, as a function of  $n$ ; settings are chosen in accordance with the simulations in Section 6. Purple dotted lines are polynomial rates that bound the batch-size (left) and number of batches (right).

### 3.1 Notations and assumptions

For a vector  $x \in \mathbb{R}^d$ , let  $\|x\|$  denote the Euclidean norm and for a matrix  $A$ , let  $\|A\|$  denote its matrix norm. All norms are equivalent in a finite-dimensional Euclidean space, so in the following discussion, we can replace the matrix norm with any other norm.

(A1) (On  $F$ ). Let the objective function  $F$  be such that the following hold:

- (i)  $F(\theta)$  is continuously differentiable and strongly convex with parameter  $M > 0$ . That is, for any  $\theta_1$  and  $\theta_2$ ,

$$F(\theta_2) \geq F(\theta_1) + \nabla F(\theta_1)^\top (\theta_2 - \theta_1) + \frac{M}{2} \|\theta_2 - \theta_1\|^2.$$

- (ii) The gradient vector  $\nabla F(\theta)$  is Lipschitz continuous with constant  $L_F$ , that is, for any  $\theta_1$  and  $\theta_2$ ,

$$\|\nabla F(\theta_1) - \nabla F(\theta_2)\| \leq L_F \|\theta_1 - \theta_2\|.$$

- (iii) The Hessian of  $F$  at  $\theta^*$ ,  $A = \nabla^2 F(\theta^*)$  exists, and there exists  $L_1$  such that

$$\|A(\theta - \theta^*) - \nabla F(\theta)\| \leq L_1 \|\theta - \theta^*\|^2.$$

Assumption (A1) is important for the convergence and asymptotic normality of  $\hat{\theta}_n$  (see Polyak and Juditsky, 1992; Moulines and Bach, 2011; Rakhlin et al., 2012). The strong convexity of  $F$  implies that  $\lambda_{\min}(A) \geq M$ , which is an important condition for parameter estimation (see Chen et al. 2020; Zhu et al. 2021; Zhu and Dong 2021). Further, this will be a key ingredient for proving consistency of the batch-means estimator of  $\Sigma$ .

(A2) (On  $f$  and  $\zeta_i$ ). Let  $D_i = \theta_i - \theta^*$ ,  $\xi_i = \nabla F(\theta_{i-1}) - \nabla f(\theta_{i-1}, \zeta_i)$ , and  $\mathbb{E}_i(\cdot)$  denotes the conditional expectation  $\mathbb{E}(\cdot | \zeta_i, \zeta_{i-1}, \dots, \zeta_1)$ , then the following hold:

- (i) The function  $f(\theta, \zeta)$  is continuously differentiable in  $\theta$  for any  $\zeta$  and  $\|\nabla f(\theta, \zeta)\|$  is uniformly integrable for any  $\theta$  (so that  $\mathbb{E}_{i-1}(\xi_i) = 0$ ).
- (ii) The conditional covariance of  $\xi_i$  has an expansion around  $\theta = \theta^*$ . That is,  $\mathbb{E}_{i-1}[\xi_i \xi_i^\top] = S + \mathcal{H}(D_{i-1})$ , and there exist constants  $\sigma_1$  and  $\sigma_2$  such that for any  $D \in \mathbb{R}^d$ ,

$$\|\mathcal{H}(D)\| \leq \sigma_1 \|D\| + \sigma_2 \|D\|^2 \text{ and } |\text{tr}(\mathcal{H}(D))| \leq \sigma_1 \|D\| + \sigma_2 \|D\|^2.$$

- (iii) There exists constants  $\sigma_3$  and  $\sigma_4$  such that the fourth conditional moment of  $\xi_i$  is bounded, i.e.,  $\mathbb{E}_{i-1} \|\xi_i\|^4 \leq \sigma_3 + \sigma_4 \|D_{i-1}\|^4$ .

Assumption (A2)(i) allows  $\mathbb{E}_{\zeta \sim \Pi} [\nabla f(\theta, \zeta)] = \nabla F(\theta)$ , which implies that the sequence  $\{\xi_i\}$  is a martingale difference process. These assumptions are standard (see Chen et al., 2020; Zhu et al., 2021) and ensure the regularity of the noisy gradients.

Our next set of conditions are on the learning rate and the choice of equal batch-size. Our results are general for any choice of batch-size satisfying the condition below.

(A3) (On  $\eta_i, b_n$ ) The following hold:

- (i) The learning rate is  $\eta_i = \eta i^{-\alpha}$  with  $\alpha \in (0.5, 1)$ ,  $i = 1, 2, 3, \dots$ .
- (ii)  $b_n$  is size of the batch such that  $b_n n^{-\alpha} \rightarrow \infty$  and  $b_n n^{-1} \rightarrow 0$  as  $n \rightarrow \infty$ .

In Assumption (A3)(i), the learning rate is that of Polyak and Juditsky (1992), ensuring asymptotic normality. Assumption (A3)(ii) is the only additional condition added to this statistical inference setup that is specific to our choice of equal batch-size. Our chosen  $b_n = cn^\beta$  satisfies this condition for  $\beta \in (\alpha, 1)$  (and thus  $b_n^*$  satisfies this condition as well). As a consequence of Assumption (A3)(i),

$$\sum_{i=\tau_{k-1}+1}^{\tau_k} \eta_i = \sum_{i=\tau_{k-1}+1}^{\tau_k} \eta i^{-\alpha} > \eta b_n \tau_k^{-\alpha} > \eta b_n n^{-\alpha} =: N. \quad (7)$$

Using Assumption (A3)(ii),  $N \rightarrow \infty$  as  $n \rightarrow \infty$ . This guarantees that the batch-size is larger than the persistent correlation in the SGD iterates. That is, using (4), this ensures fast decay of correlation between batches, which is a critical step in proving consistency of the batch-means estimator.

In the following discussion, for sequences of positive numbers  $p_n$  and  $q_n$ , denote

- $p_n \gtrsim q_n$  if for some  $c > 0$ ,  $cp_n \geq q_n$  for all  $n$  large enough,
- $p_n \lesssim q_n$  if  $q_n \gtrsim p_n$ , and
- $p_n \asymp q_n$  if  $p_n \gtrsim q_n$  and  $p_n \lesssim q_n$ .

For simplicity, we define the following constant,

$$C_d := \max \left\{ L_F, L_1, \sigma_1^{2/3}, \sqrt{\sigma_2}, \sqrt{\sigma_3}, \sigma_4^{1/4}, \text{tr}(S) \right\}.$$

### 3.2 Consistency of the estimator

We now present our main result of consistency of the batch-means estimator for equal batch-sizes satisfying Assumption (A3).

**Theorem 1** *Under the Assumptions (A1), (A2) and (A3), for sufficiently large  $n$*

$$\begin{aligned} \mathbb{E}\|\hat{\Sigma}_{b_n} - \Sigma\| \lesssim & C_d^2 n^{-\alpha/2} a_n^{-1/4} + C_d^3 n^{-\alpha} + C_d a_n^{-1/2} + C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2} \\ & + C_d a_n^{-1} + C_d^4 n^{-2\alpha} b_n. \end{aligned}$$

**Proof** Proof is available in Appendix C. ■

Under Assumption (A3), the bound in Theorem 1 goes to zero, yielding consistency of  $\hat{\Sigma}_{b_n}$ . Chen et al. (2020, Theorem 4.3) and Zhu et al. (2021, Theorem 3.1) provide similar bounds for different IBS batch-means estimators, with Zhu et al. (2021) being an improvement over Chen et al. (2020). One key reason for explicitly writing a bound instead of merely mentioning convergence to zero, is that the bound allows for a reasonable choice for  $b_n$ . Substituting batch-sizes of the form  $b_n = cn^\beta$ , in Theorem 1,

$$\mathbb{E}\|\hat{\Sigma}_{b_n} - \Sigma\| \lesssim n^{-\alpha/2+(\beta-1)/4} + n^{(\beta-1)/2} + n^{-\beta(1-\alpha)} + n^{(\alpha-\beta)/2} + n^{\beta-1} + n^{\beta-2\alpha}. \quad (8)$$

Obtaining a closed-form expression of an optimal  $\beta$  from the right side of the above equation is challenging. A numerical solution is possible, but not interpretable. Instead, we note that by Assumption (A3),  $(\beta-1)/2 > \beta-2\alpha$  and  $-\alpha/2 + (\beta-1)/4 < (\beta-1)/2$ , so among the first, second, fifth, and sixth terms, the second term is dominating. Further,  $-\beta(1-\alpha) < (\alpha-\beta)/2$ , so among the third and fourth terms, the fourth term is dominating. Considering then, only the dominating terms, we have

$$\mathbb{E}\|\hat{\Sigma}_{b_n} - \Sigma\| \lesssim n^{(\beta-1)/2} + n^{(\alpha-\beta)/2}.$$

With this approximation, the optimal choice of  $\beta$  is  $\beta^* = (1+\alpha)/2$ .

**Remark 2** *The bounds we obtain are meaningful only for large  $n$ . For small  $n$ , it is challenging to obtain a meaningful expression of the optimal value of  $\beta$  in (8). Numerically, we observed that for sample size in the thousands and  $\alpha = 0.51$ , the optimal value of  $\beta$  is near 0.66. However, as  $n$  increases, the optimal value of  $\beta$  approaches  $(1+\alpha)/2$ . This agrees with the above mentioned bound. It is also important to remember that (8) is only a bound, optimizing which need not yield a true mean-square-optimal choice of batch-size.*

**Remark 3** *Consistency of  $\hat{\Sigma}_{b_n}$  immediately allows the construction of Wald-like confidence regions (see Section 5). To obtain a consistent estimator of  $\Sigma$ , the number of batches,  $a_n$ , must increase with  $n$ . Naturally, the batch-size also must be large to mimic the limiting Polyak and Juditsky (1992) behavior. This yields a challenging trade-off. Our particular batch-size construction allows finite-sample adjustments for small batch sizes using the lugsail trick (see Section 4). If the goal is only inference, and not the quantification of variance, Zhu and Dong (2021) used cancellation methods to construct valid confidence regions employing batch-means estimators with fixed number of batches.*



**Remark 4** With  $\beta = (1+\alpha)/2$ , the computational complexity of calculating the batch-means estimator is similar to the IBS estimator at  $\mathcal{O}(d^2 n^{(1+\alpha)/2} + dn)$ . On the other hand, an online implementation strategy for EBS estimators remains to be an open problem. Further, as we will discuss in Section 4, EBS estimators allow for bias-reduction strategies which yield significant benefits. Bias-reduction strategies in IBS estimators remain to be an open problem.

#### 4. Bias-reduced estimation

Naturally, the mean-square bound in Theorem 1 is contributed from both the bias and variance of the batch-means estimator. Vats and Flegal (2022) proposed a lugsail batch-means estimator for stochastic simulation that can dramatically reduce bias in variance estimation. Our particular choice of equal batch-size allows an easy and effective implementation of the lugsail technique. Obtaining an exact expression of the bias of the batch-means estimator for a general SGD framework is an open problem. However, the following mean estimation model provides a motivation.

**Example 1** Consider for  $i = 1, 2, \dots, n$ , a mean estimation model  $y_i = \theta^* + \epsilon_i$ , where  $\theta^* \in \mathbb{R}$  and  $\epsilon_i$  are independent mean-zero random error terms. For the squared error loss function  $f(\theta, \zeta) = (y - \theta)^2/2$  for estimating  $\theta^*$ , the  $i^{\text{th}}$  SGD iterate is

$$\theta_i = \theta_{i-1} + \eta_i(y_i - \theta_{i-1}),$$

with  $\eta_i = \eta i^{-\alpha}$ . In Appendix D.1, we show that the bias of  $\hat{\Sigma}_{b_n}$  for this model is:

$$\text{Bias}(\hat{\Sigma}_{b_n}) \approx \frac{-2C_1}{n} \sum_{1 \leq j < k \leq a_n} \sum_{p=\tau_{j-1}+1}^{\tau_j} \sum_{q=\tau_{k-1}+1}^{\tau_k} q^{-\alpha} (1 - q^{-\alpha})^{q-p},$$

where  $C_1$  is a positive constant. The estimator of Zhu et al. (2021) exhibits a similar negative bias expression. For large  $n$ , the bias may be insignificant, however, as Figure 2 exhibits, even in this simple model, the finite-sample bias in the estimator of  $\Sigma$  remains significant.

More than the magnitude of bias, its direction is a larger concern. Variance estimation of any statistical estimator allows us to assess the uncertainty in the problem. Underestimation of this variance leads to a false sense of security and inadequate tests (see Simonoff, 1993). Obtaining bias-free estimators for such long-run variances is a critical and challenging problem in operations research, stochastic simulation, econometrics, and MCMC. A wide range of solutions have been attempted (Kiefer and Vogelsang, 2002, 2005; Liu and Flegal, 2018; Politis and Romano, 1995) to reduce the bias of variance estimators.

By studying linear combinations of lag-windows in spectral variance estimators, Liu and Flegal (2018); Vats and Flegal (2022) develop a family of variance estimators, called *lugsail* estimators, that are essentially obtained by a carefully chosen linear combination of variance estimators. Liu and Flegal (2018) define a batch-means version of this estimator called the weighted batch-means estimator, that seek to combine batch-means estimators obtained through various batch-sizes, using an appropriate weighting strategy. Consider a

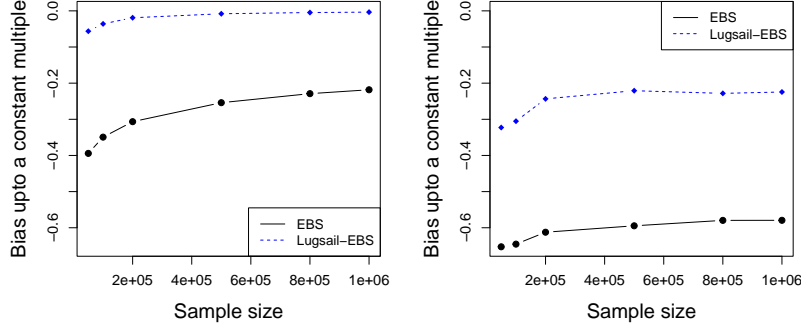


Figure 2: Mean estimation model: Bias of the EBS and Lugsail-EBS estimator for  $\alpha = 0.51$  (left) and  $\alpha = 0.75$  (right) against sample size.

weighting function, called a *lag window*,  $w_n(\cdot)$  such that (i)  $w_n(\cdot)$  is an even function on  $\mathbb{Z}$ , (ii)  $w_n(0) = 1$  for all  $n$ , and (iii)  $w_n(m) = 0$  for all  $m \geq b_n$ . Anderson (2011) provides a comprehensive list of lag windows used in stochastic simulation and time series. We will employ the flat-top lag window of Politis and Romano (1995):

$$w_n(m) = \mathbb{I}\left(|m| \leq \frac{b_n}{2}\right) + 2\left(1 - \frac{|m|}{b_n}\right) \mathbb{I}\left(\frac{b_n}{2} < |m| \leq b_n\right). \quad (9)$$

Let  $\Delta_2 w_n(m) = w_n(m-1) - 2w_n(m) + w_n(m+1)$  denote the second order differencing of  $w_n$  at  $m$ . Consider multiple batch-sizes  $m = 1, 2, \dots, b_n$ , so that the corresponding number of batches are  $\tilde{a}_m := \lfloor n/m \rfloor$ . For each batch size  $m$ , let  $\bar{\theta}_{m,k} = m^{-1} \sum_{t=1}^m \theta_{km+t}$  for  $k = 0, 1, 2, \dots, \tilde{a}_m$ , denote the  $k^{\text{th}}$  batch mean vector. The weighted batch-means estimator of Liu and Flegal (2018) is defined as

$$\hat{\Sigma}_{\text{wBM}} = \sum_{m=1}^{b_n} \frac{m^2 \Delta_2 w_n(m)}{a_m} \sum_{k=0}^{\tilde{a}_m-1} \left( \bar{\theta}_{m,k} - \hat{\theta}_n \right) \left( \bar{\theta}_{m,k} - \hat{\theta}_n \right)^\top. \quad (10)$$

For general lag windows,  $\hat{\Sigma}_{\text{wBM}}$  can be expensive to compute due to the double summation in (10). However, employing piecewise linear lag windows like the flat-top lag window in (9), yields  $\Delta_2 w_n(m) = 0$  everywhere except  $m = b_n/2, b_n$ , making the estimator in (10) computationally viable.

Liu and Flegal (2018); Vats and Flegal (2022) discuss the bias-correction advantages of the weighted batch-means estimators. Employing (9) in (10) and renaming  $b_n \equiv 2b_n$ , we obtain a bias-corrected batch-means estimator, compatible with equal batch-sizes, called the lugsail batch-means estimator.

Specifically, the lugsail batch-means estimator simplifies to

$$\hat{\Sigma}_{L, b_n} := 2\hat{\Sigma}_{2b_n} - \hat{\Sigma}_{b_n}. \quad (11)$$

Our batching strategy,  $b_n^*$ , allows an easy implementation of the lugsail bias-correction strategy since a batch-means estimator of batch-size  $2b_n^*$  can be obtained by collapsing

adjacent batch-means vectors. Thus, the proposed bias-correction does not increase memory costs. We also note that since the lugsail lag-windows rely on equal batch-sizes, such lugsail corrections are not directly possible for the IBS estimator. We call the estimator in (11) the lugsail-EBS estimator.

Define

$$\hat{R}_{b_n} := \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \left[ \left( \bar{\theta}_{2j-1} - \hat{\theta}_n \right) \left( \bar{\theta}_{2j} - \hat{\theta}_n \right)^\top + \left( \bar{\theta}_{2j} - \hat{\theta}_n \right) \left( \bar{\theta}_{2j-1} - \hat{\theta}_n \right)^\top \right]. \quad (12)$$

Then,  $\hat{R}_{b_n}$  summarizes the covariance between adjacent batches. In Appendix D.2 we show that

$$\hat{\Sigma}_{L,b_n} = \hat{\Sigma}_{b_n} + 2\hat{R}_{b_n}. \quad (13)$$

When batch-sizes are not large enough, adjacent batch means will be expected to be positively correlated so that (the diagonals of)  $\hat{R}_{b_n} > 0$ , thereby adjusting some of systematic under-estimation happening due to a small batch-size.

For the mean estimation model in Example 1, Figure 2 presents the bias expression of both EBS and the lugsail-EBS estimators. Although the bias in the estimator depends on the correlation in the process (through  $\alpha$  in this case), in both cases, the lugsail-EBS estimator presents significant bias reduction. As we will see in Section 6, this correction proves to be critical for finite-sample inference.

The following results establish the consistency of the lugsail-EBS estimator, under the same conditions as required in Theorem 1.

**Proposition 5** *Under Assumptions (A1), (A2) and (A3), for sufficiently large  $n$ ,*

$$\begin{aligned} \mathbb{E}\|\hat{R}_{b_n}\| &\lesssim C_d^{1.25} n^{-\alpha/2} a_n^{-1/4} + C_d^2 n^{-\alpha/2} + C_d a_n^{-1/2} + C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2} \\ &\quad + C_d a_n^{-1} + C_d^4 n^{-2\alpha} b_n. \end{aligned}$$

**Proof** Proof is available in Appendix D.3. ■

An immediate consequence of the rate in Proposition 5 is the following result that yields similar rates for the lugsail estimator as the original EBS batch-means estimator. Of course, finite-sample performance is affected by the absorbed constants, and we will see in Section 6 that the finite-sample performance of lugsail estimators is far improved.

**Corollary 6** *Under Assumptions (A1), (A2) and (A3), for sufficiently large  $n$ ,*

$$\mathbb{E}\|\hat{\Sigma}_{L,b_n} - \Sigma\| \asymp \mathbb{E}\|\hat{\Sigma}_{b_n} - \Sigma\|.$$

**Proof** Observe that the only different order term in Proposition 5 as compared to Theorem 1 is  $n^{-\alpha/2}$ . Using Assumption (A3), for all  $n$ ,  $b_n^{1/2} n^{-\alpha} < (b_n n^{-1})^\alpha < 1$ . Consequently,  $b_n^{1/2} n^{-\alpha/2} n^{-\alpha/2} < 1$ , and  $n^{-\alpha/2} < b_n^{-1/2} n^{\alpha/2}$ . Therefore,  $n^{-\alpha/2}$  decays faster than  $b_n^{-1/2} n^{\alpha/2}$ . This illuminates that the rate for the bound on  $\mathbb{E}\|\hat{R}_{b_n}\|$  is the same as that of  $\mathbb{E}\|\hat{\Sigma}_{b_n} - \Sigma\|$ . Using the triangle inequality, the result follows. ■

**Remark 7** *Unlike standard EBS estimators that are guaranteed to be positive semi-definite, lugsail estimators may not retain this property, particularly for small sample sizes. In such a case, users may tune  $c$  so that  $\hat{\Sigma}_{L,b_n}$  is positive-definite. In our simulations,  $c$  was chosen appropriately so as to allow all estimators to be positive-definite.*

## 5. Marginal and simultaneous inference

For problems where SGD is relevant, it is natural to ask what purpose will an estimator of  $\Sigma$  serve? Confidence ellipsoids provide little interpretation in high-dimensional settings, compared to hyper-rectangular regions that are more amenable to marginal-friendly interpretation. Zhu et al. (2021) use the diagonals of  $\Sigma$  to construct uncorrected marginal confidence intervals for each of the  $d$  parameters of interest. The problem of multiple-testing is omnipresent in this case, and corrections like Bonferroni can be crude. Moreover, the potentially complex dependence in  $\Sigma$ , via both  $S$  and  $A$  is completely ignored. Here, we leverage the methods of Robertson et al. (2021) to construct simultaneous marginal-friendly confidence regions that approximately retain the desired coverage probabilities.

For joint inference for  $\theta^*$ , (1) provides a  $100(1-p)\%$  confidence ellipsoid

$$E_p = \left\{ \theta \in \mathbb{R}^d : (\hat{\theta}_n - \theta)^\top \hat{\Sigma}_n^{-1} (\hat{\theta}_n - \theta) \leq \chi_{d,1-p}^2 \right\}, \quad (14)$$

where  $\chi_{d,s}^2$  denotes the  $s^{\text{th}}$  quantile of a chi-squared distribution with  $d$  degrees of freedom. Marginal interpretation of such an ellipsoid confidence region is difficult. Instead, one may study marginal confidence intervals. Let  $\hat{\Sigma}_n$  be any consistent estimator of  $\Sigma$ . Let  $\hat{\theta}_n = (\hat{\theta}_{n1}, \dots, \hat{\theta}_{nd})^\top$ ,  $\theta^* = (\theta_1^*, \dots, \theta_d^*)^\top$  and  $\hat{\Sigma}_n = (\hat{\sigma}_{ij})_{i,j=1,\dots,p}$ . Using (1), for  $0 < p < 1$ , an asymptotic  $100(1-p)\%$  marginal confidence interval of  $\theta_i^*$  is:

$$\hat{\theta}_{ni} \pm z_{1-p/2} \sqrt{\hat{\sigma}_{ii}/n}, \quad (15)$$

where  $z_s$  denotes the  $s^{\text{th}}$  quantile of  $N(0, 1)$ . Fang et al. (2018), Chen et al. (2020), Zhu et al. (2021), and Zhu and Dong (2021) discuss both uncorrected marginal confidence intervals and the ellipsoid joint confidence region. As discussed, both are inconducive for valid and interpretable joint inference. For the general Monte Carlo problem, Robertson et al. (2021) suggest a remedy by using an appropriate hyper-rectangular confidence region, which we now describe. Using the  $d$  uncorrected intervals in (15), an at-most  $100(1-p)\%$  hyper-rectangular confidence region is

$$C_{lb}(z_{p/2}) = \prod_{i=1}^d \left[ \hat{\theta}_{ni} - z_{1-p/2} \sqrt{\hat{\sigma}_{ii}/n}, \hat{\theta}_{ni} + z_{1-p/2} \sqrt{\hat{\sigma}_{ii}/n} \right].$$

Using a Bonferroni approach, an at least  $100(1-p)\%$  hyper-rectangular confidence region is

$$C_{ub}(z_{p/2d}) = \prod_{i=1}^d \left[ \hat{\theta}_{ni} - z_{1-p/2d} \sqrt{\hat{\sigma}_{ii}/n}, \hat{\theta}_{ni} + z_{1-p/2d} \sqrt{\hat{\sigma}_{ii}/n} \right].$$

Clearly,  $C_{lb}(z_{p/2}) \subseteq C_{ub}(z_{p/2d})$ . Robertson et al. (2021) used a quasi Monte-Carlo approach to find a  $z^*$  with  $z_{1-p/2} < z^* < z_{1-p/2d}$  to yield the hyper-rectangular confidence region

$$C(z^*) = \prod_{i=1}^d \left[ \hat{\theta}_{ni} - z^* \sqrt{\hat{\sigma}_{ii}/n}, \hat{\theta}_{ni} + z^* \sqrt{\hat{\sigma}_{ii}/n} \right], \quad (16)$$

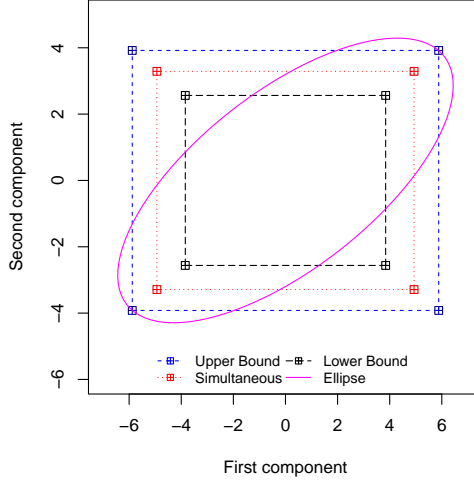


Figure 3: Plot of  $C_{lb}(z_{p/2})$  (black dashed) and  $C_{ub}(z_{p/2d})$  (blue dashed) from a 90% confidence region for a bivariate normal distribution with component variances 9 and 4. The red dashed line gives the corresponding  $C(z^*)$ .

such that  $C_{lb}(z_{p/2}) \subseteq C(z^*) \subseteq C_{ub}(z_{p/2d})$  and  $\mathbb{P}(\theta^* \in C(z^*)) \approx 1 - p$ , under the assumption that  $\hat{\theta}_n \approx N_p(\theta_*, \hat{\Sigma}_n)$ . An illustration is given by Figure 3. The computation of (16) is essentially a quick one-dimensional optimization problem solved by a bisection search over the interval  $(z_{1-p/2}, z_{1-p/2d})$ ; see Robertson et al. (2021) for details. In Section 6, we present coverage properties of both the ellipsoidal region  $E_p$  and the hyper-rectangular region  $C(z^*)$ .

## 6. Numerical implementations

We compare our proposed method with competitors for a few simulated examples and some real data applications. All reproducible codes are provided in the following Github repository: <https://github.com/Abhineek-Shukla/SGD-EBS>.

### 6.1 Setup

We implement our proposed EBS and lugsail-EBS estimators for two simulated models and a real data implementation, for both doubling batch-sizes  $b_n^*$  and polynomial batch-sizes  $b_n = \lfloor cn^\beta \rfloor$ ; we call these EBS and EBS-poly, respectively. Their lugsail versions are L-EBS and L-EBS-poly, respectively. We systematically keep the following settings for our EBS estimator:  $c = 0.1$  so that the number of batches stays reasonably large ensuring that the estimators are positive-definite;  $\beta = (1 + \alpha)/2$  as a reasonable value obtained from the mean-square bounds;  $\alpha = 0.51$  to allow for reasonable exploration. For comparison we implement the IBS estimator of Zhu et al. (2021) with their suggested settings. However, in the event that the IBS estimator is singular, we increase their number of batches to allow positive-definiteness of the IBS estimator. Zhu et al. (2021) showed that their estimator was superior to the estimator of Chen et al. (2020), both theoretically and in simulations,

so for brevity and clarity, we do not present comparisons with the estimator of Chen et al. (2020).

When the true covariance matrix  $\Sigma$  is available, we employ it as an oracle, and use it to calculate the relative Frobenius norm of an estimator  $\hat{\Sigma}$ :  $\|\hat{\Sigma} - \Sigma\|_F / \|\Sigma\|_F$ . Further, we employ  $\Sigma$  in calculating and comparing the coverage probability of the confidence regions discussed in Section 5 for different estimators. Another important feature of confidence regions is their volume, particularly for the hyper-rectangular regions created in (16). Thus, for each estimator, we also report

$$\left( \frac{\text{Volume}(C(z^*))}{\text{Volume}(E_p)} \right)^{1/p},$$

for which a high value indicates an undesired increase in the volume of the hyper-rectangular confidence region.

## 6.2 Linear regression

We simulate data according to the linear regression model, for  $i = 1, 2, \dots, n$ ,  $y_i = x_i^T \beta^* + \epsilon_i$  where for some  $d \times d$  positive-definite matrix  $A$ ,  $x_i \stackrel{\text{iid}}{\sim} N(0, A)$  and  $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, 1)$ , independent of  $x_i$ . We fix  $\beta^* \in \mathbb{R}^d$  to be the  $d$ -vector of equidistant points on the grid  $(0, 1)$ . Here  $\zeta_i = (x_i, y_i)$ . In order to implement ordinary least squares estimation of  $\beta$ , the loss function is

$$f(\beta, \zeta_i) = \frac{(y_i - x_i^\top \beta)^2}{2}.$$

Since the errors are independent and identically distributed (iid), the true  $\Sigma$  is  $A^{-1}$ . We consider the three forms of  $A$  used in Chen et al. (2020): (i) identity ( $A = I_d$ ), (ii) Toeplitz, where element  $A_{i,j} = \rho^{|i-j|}$ , and (iii) equicorrelation, where element  $A_{i,j} = \rho$  for  $i \neq j$  and 1 otherwise. Throughout, we set  $\rho = 0.5$ . We present results of all three settings of  $A$  for  $d = 5$  (here  $\eta = 0.5$ ) and for brevity, only present the identity result for  $d = 20$  (here  $\eta = 1$ ).

Data of size  $5 \times 10^6$  was simulated with the first 1000 SGD iterates being discarded as burn-in. We start the SGD process from  $\mathbf{0}$  and study the statistical performance of various estimators of  $\Sigma$  sequentially as a function of the data. Since the above is done for 1000 replications, for each  $A$  and  $d$ , we present four key comparative plots: (i) the estimated relative Frobenius norm as a function of the sample size, (ii) the estimated ellipsoidal coverage probabilities, (iii) the estimated coverage probabilities of the hyper-rectangular confidence regions, and (iv) the ratio of the volumes of the hyper-rectangular regions to the ellipsoidal regions.

Figure 4 presents the results for  $d = 5$ . Due to the nature of the doubling batching technique  $b_n^*$ , the performance is not monotonic as a function of the sample size for EBS; this is expected. However, the polynomial batch-size in EBS demonstrates the expected monotonic behavior. Further, for each of the three settings, the lugsail-EBS estimators outperforms in all measures with the EBS estimators being competitive with the IBS estimator, when not better. One metric where the IBS estimator suffers drastically, are the marginal confidence regions. As evident from Figure 4, the coverage for the IBS estimator improves drastically when going from the elliptical regions to the hyper-rectangular regions. The bottom row of the plots indicate that this is entirely due to the drastic increase in the volume of region. The hyper-rectangular regions made by IBS are significantly larger than

their ellipsoidal regions. This is likely due to an exaggerated correlation structure captured by the IBS. All the EBS estimators, and particularly the lugsail-EBS estimators do not suffer from this. These problems are further exaggerated for  $d = 20$  as evidenced in Figure 5. The performance of both EBS and lugsail-EBS estimators remain essentially the same. We highlight that even when the relative Frobenius norm is large for the EBS estimators, the simultaneous marginal coverage probabilities are reasonable, with only little cost to the volume.

To understand why the EBS estimators perform significantly better than the IBS estimators, we take a closer look at the batching strategy in  $b_n^*$ . The general idea in batch-means estimators is that each batch-mean vector emulates the sample mean ASGD estimator; thus the empirical sample covariance of these batch-means is a reasonable estimator of  $\Sigma$ . For such a heuristic to hold, the batch-mean vector for each batch must be approximately normally distributed;  $\sqrt{b_{n,k}}\bar{\theta}_k \approx N_d(\theta^*, I_d)$ . For  $d = 5$  with  $A$  being identity, if we accumulate all the components of all batch-mean vectors they should be normally distributed, and thus we may compare them with true Gaussian quantiles. In Figure 6, we present a zoomed-in QQ plot of this for two different data sizes. Figure 6 reveals significant deviation from normality for the IBS estimator, particularly for small sample situations. The EBS estimator, on the other hand, follows the theoretical Gaussian quantiles fairly well.

### 6.3 LAD regression

Assume a similar linear regression model, with non-Gaussian errors:  $y_i = x_i^T \beta^* + \epsilon_i$  where  $x_i \stackrel{\text{iid}}{\sim} N(0, A)$  and  $\epsilon_i \stackrel{\text{iid}}{\sim} \text{DE}(0, 1)$ , here  $\text{DE}(\mu, \lambda)$  denotes the double exponential distribution with median parameter  $\mu$  and scale parameter  $\lambda$ . Instead of ordinary least squares, we consider the least absolute deviation (LAD) loss function,  $f(\beta, \zeta_i) = |y_i - x_i^T \beta|$ . Fang et al. (2018) consider this simulation setup as well and discuss that the true  $\Sigma$  is  $A^{-1}$ .

We repeat the simulation setup of the previous Section with  $d = 20$  for the three different choices of  $A$ . Figure 7 presents the results. The performance of the EBS estimators, particularly lugsail-EBS is significantly superior to that of the IBS estimator. Here again, although the simultaneous coverage of the hyper-rectangular regions is far improved for the IBS estimator, this is purely a consequence of over-inflated volume of the region.

### 6.4 Improving predictions with estimators of $\Sigma$

Consider the binary classification problem where for  $i = 1, 2, \dots$ ,

$$y_i \stackrel{\text{ind}}{\sim} \text{Bernoulli} \left( p_i := \frac{1}{1 + e^{-x_i^T \beta^*}} \right),$$

with  $x_i \in \mathbb{R}^d$  assumed to be iid. For estimation, the loss function is the negative log-likelihood as a consequence of the Bernoulli model assumption. However, estimates of  $\beta^*$  that minimize this loss function are used in predictions, without any focus on accounting for the variability in its estimation. Denote the ASGD estimator of  $\beta^*$  with  $\hat{\beta}_n$ , and for any data point  $j$ , the fitted/predicted probability of success is estimated with

$$\hat{p}_j := \frac{1}{1 + e^{-x_j^T \hat{\beta}_n}}.$$

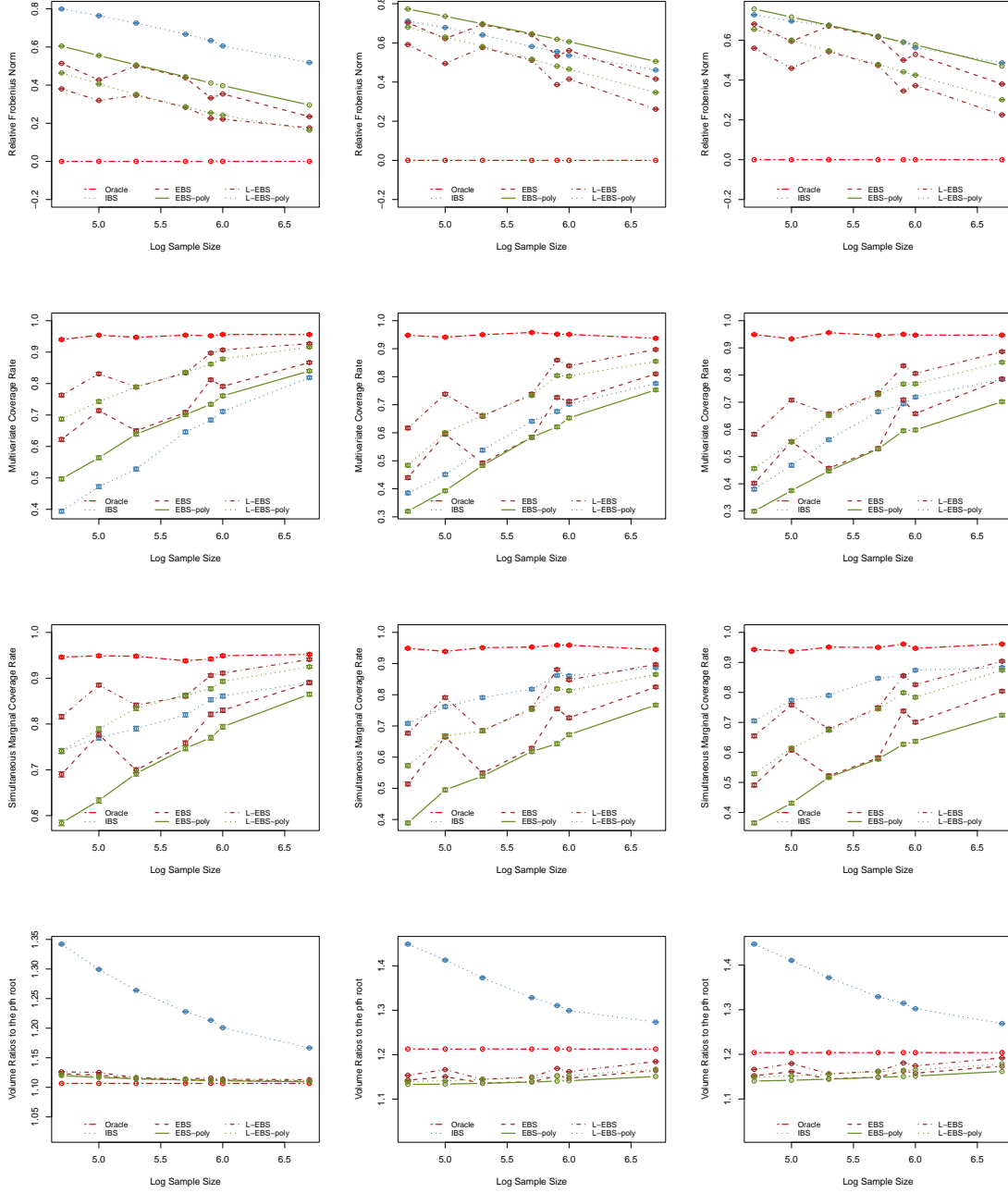


Figure 4: Linear regression  $d = 5$ : Line plots equipped with error bars from 1000 replications. Left plots for  $A$  being identity, middle plots for  $A$  being Toeplitz, and right plots for  $A$  being equicorrelation matrix.



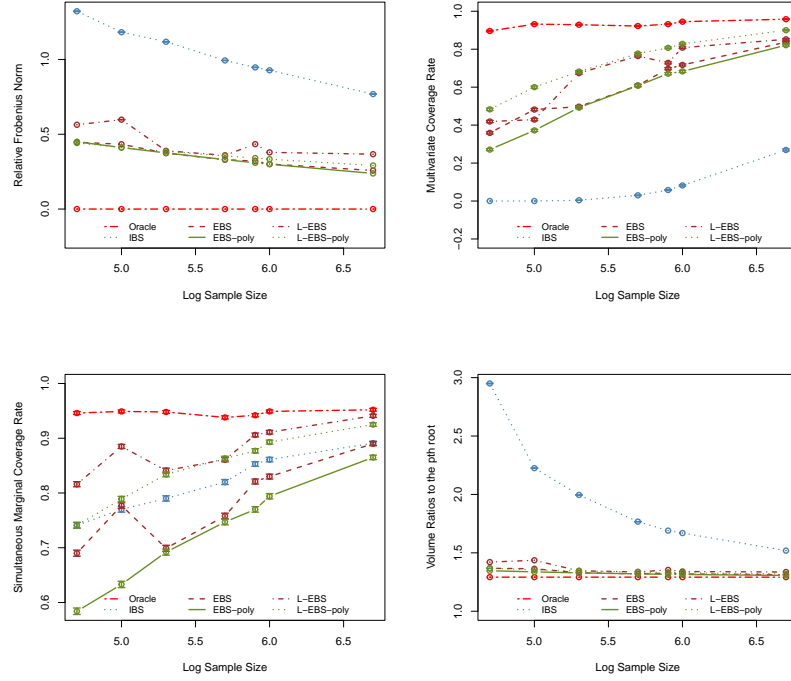


Figure 5: Linear regression  $d = 20$  for identity A: Line plots equipped with error bars from 1000 replications

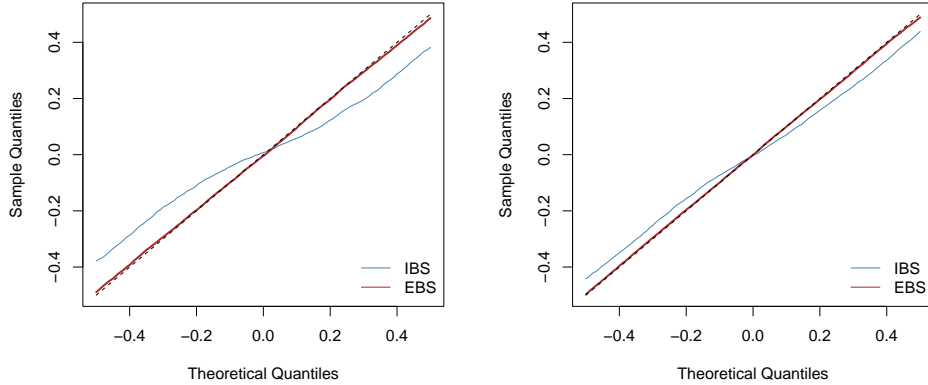


Figure 6: Linear regression for  $d = 5$ : QQ plot of all components of the batch-means vectors in the IBS and the EBS estimators, with black lines indicating theoretical quantiles. Left panel is corresponding to  $n = 50000$  and right panel is corresponding to  $n = 10^6$ .

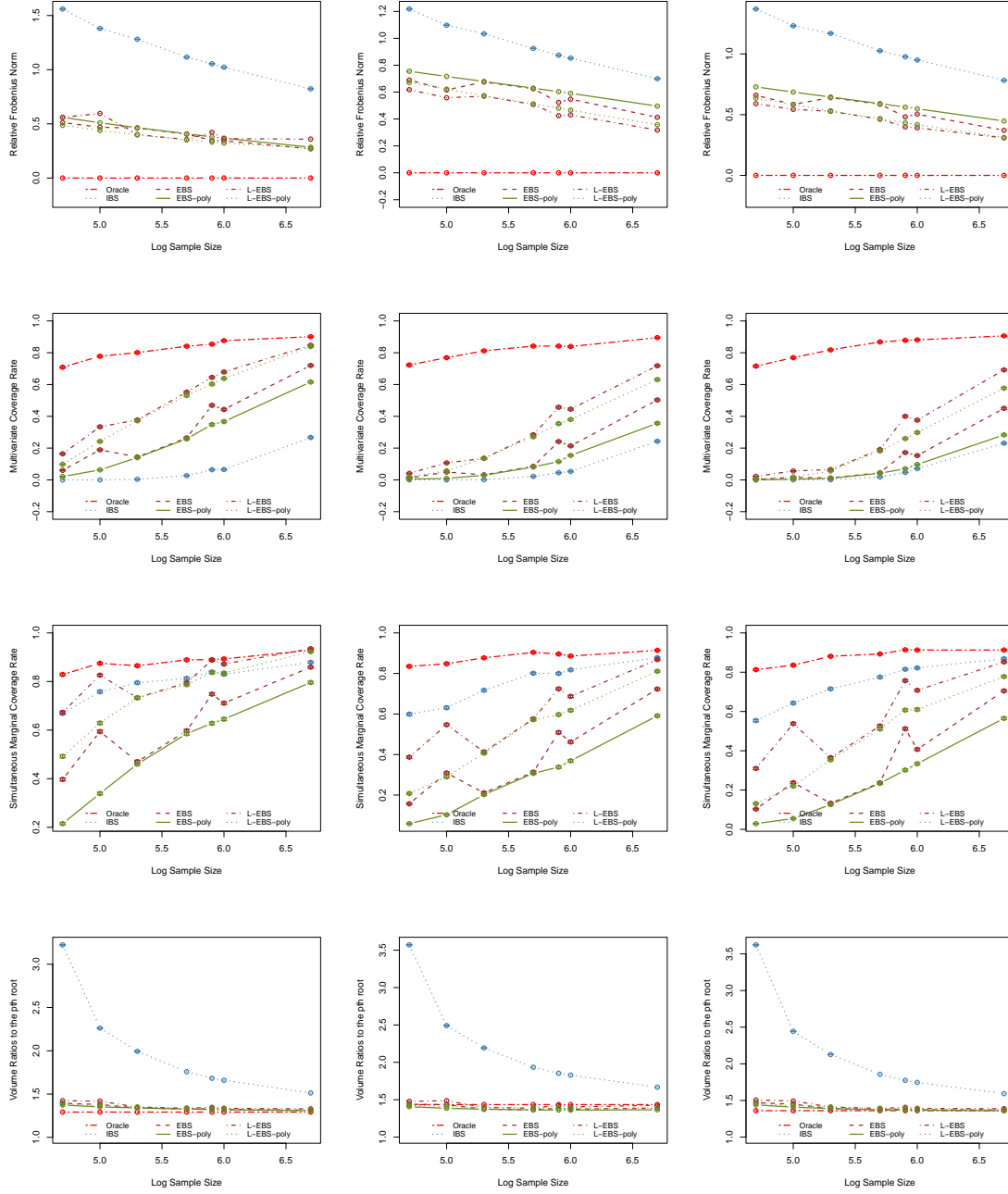


Figure 7: LAD regression for  $d = 20$ : Line plots equipped with error bars from 1000 replications. Left plots for  $A$  being identity, middle plots for  $A$  being Toeplitz, and right plots for  $A$  being equicorrelation matrix.

A thresholding is then typically used to obtain a binary prediction of this  $j$ th observation, based on  $\hat{p}_j$ . That is, for some user-chosen threshold  $q$ ,  $\hat{y}_j = \mathbb{I}(\hat{p}_j > q)$ . When employing a test data set, misclassification rates can then be obtained for model building and comparisons. Due to (1) and the delta method, as  $n \rightarrow \infty$ ,

$$\sqrt{n}(\hat{p}_j - p_j) \xrightarrow{d} N(0, (p_j(1 - p_j))^2 x_j^\top \Sigma x_j) .$$

Since our proposed EBS estimator of  $\Sigma$  is consistent, we can obtain a confidence interval for each  $\hat{p}_j \pm z_{.975} \text{se}_j$ , where  $\text{se}_j$  is the standard error for the  $j$ th response and is calculated using the plug-in estimators of  $\Sigma$  and  $p_j$ . In order to account for the estimation variability in  $\hat{p}_j$ , we employ an alternative estimator of  $y_j$ :  $\tilde{y}_j = \mathbb{I}(\hat{p}_j - z_{.975} \text{se}_j > q)$ . That is, our logistic classifier, classifies the observation as a success if the lower-bound on the confidence interval for  $\hat{p}_j$  is larger than the cutoff,  $q$ ; different observations will have different  $\text{se}_j$ .

We implement this strategy on four data sets (i) Santander customer transaction data set<sup>1</sup>, (ii) Covertypes data set of Blackard (1998), (iii) Spambase data set of Hopkins et al. (1999), and (iv) the diabetes health data set<sup>2</sup>. Implementation details for each of them is provided in Appendix E. For each data set, since the true  $\beta^*$  and  $\Sigma$  are unknown, comparison of confidence regions is unreasonable here. We can obtain marginal simultaneous confidence intervals, however due to the nature of the data, classical hypothesis testing may not be of interest. Instead, we utilize the estimator of  $\Sigma$  for prediction.

Figure 8 demonstrates the test data misclassification rate for various values of  $q$ , with and without confidence interval, for all data sets. Clearly, the blue curve which employs the EBS estimator to obtain  $\text{se}_j$  yields a lower misclassification rate. Due to consistency of the estimator,  $\text{se}_j$  is expected to converge to 0 as  $n \rightarrow \infty$ , and thus, for a large enough training data, we would expect the blue and the black curves to merge into one, for all data sets.

## 7. Discussion

We present EBS batching-strategies for batch-means estimators in order to estimate the limiting variance of SGD estimates. A substantial body of work on batching methods exists for stationary and ergodic processes in the stochastic simulation literature (Alexopoulos and Seila, 1996; Alexopoulos and Goldsman, 2004; Glynn and Whitt, 1992; Schmeiser and Song, 1996; Su et al., 2024). It would be worthwhile studying these more generally for application in inference for SGD. For instance, our proposed EBS batching-strategy can be extended to averaging over  $k$ -neighbouring batches for any fixed positive integer  $k$ , and all the theoretical results discussed will hold true. However, large values of  $k$  will reduce the number of batches, thereby reducing the efficiency of the covariance estimator. Another alternative is to adopt an overlapping batch-means estimator with an EBS strategy, adapting the IBS estimator of Zhu et al. (2021). Overlapping batch-means estimators have found reasonable success in stochastic simulation (Meketon and Schmeiser, 1984) as they allow higher number of batches, albeit correlated. However, the computational complexity of these estimators is  $\mathcal{O}(d^2 n)$  since the number of batches are on the order of the number of samples. Nonetheless, similar theoretical results should be possible for overlapping batch-means estimators with

---

1. <https://www.kaggle.com/competitions/santander-customer-transaction-prediction/overview>  
 2. <https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators>

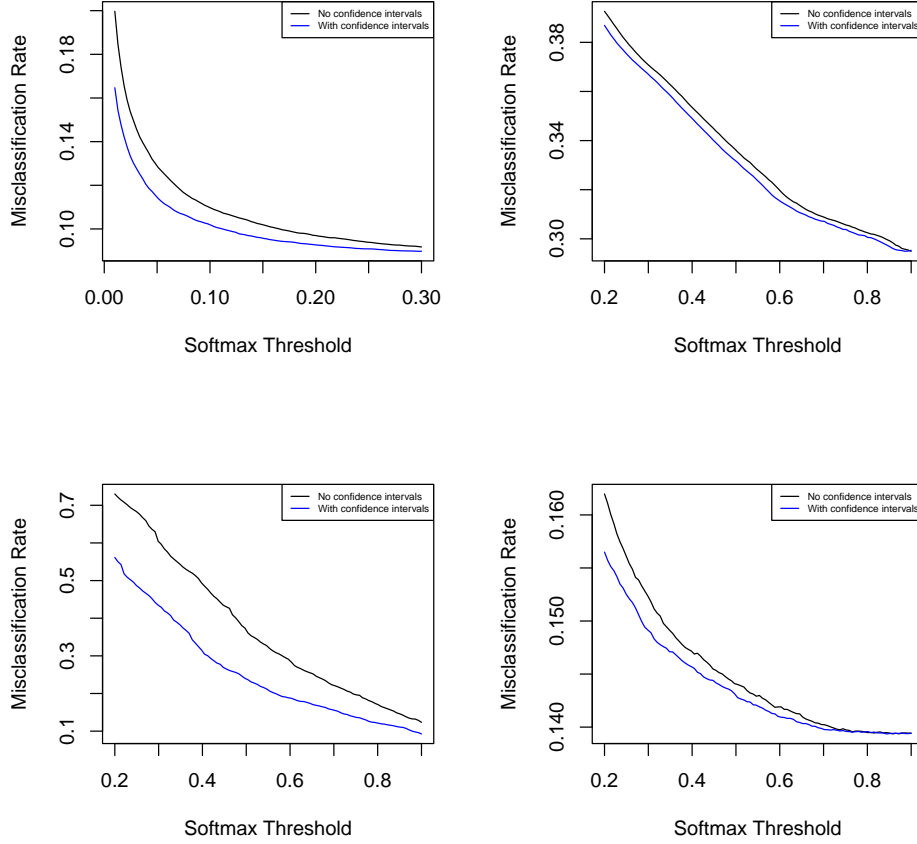


Figure 8: Misclassification rate on the testing set for various values of the cutoff for (i) Santander (topleft) (ii) covtype (topright) (iii) spambase (bottomleft) (iv) diabetes (bottomright) data sets.

an EBS strategy. Leung and Chan (2024) discuss variants of the IBS estimator and find that their performances are quite similar. A study similar to Leung and Chan (2024) for the EBS estimator would make a useful follow-up of our work. Building a statistical inference framework for SGD is an active area of research in recent times. This includes the recent works of Fang et al. (2018); Xie et al. (2023) who use bootstrap techniques to estimate the limiting covariance structure. Fang et al. (2018) heuristically argued to use the (perturbed) bootstrap ASGD outputs to estimate the covariance matrix of  $\hat{\theta}_n$ . The theoretical properties of the estimator are not known and computational demands of the estimator is considerable. Zhu and Dong (2021) present a method of consistent inference for SGD without using a consistent estimator of  $\Sigma$ . It remains unclear if methods of marginal inference and delta method arguments can be used in their framework. Li et al. (2023); Liu et al. (2023) estimate the limiting covariance under situations when the iid assumption on

the data is violated. Our marginal-friendly confidence interval construction, and utilization of  $\Sigma$  in improving predictions are directly applicable to this literature.

There are numerous other variants of the SGD (see, e.g., Konečný et al., 2016; Toulis and Airolidi, 2017; Loizou and Richtárik, 2020; Yuan and Ma, 2020), and the fundamental framework remains essentially the same for other variants of SGD, as long as the results of Polyak and Juditsky (1992) applies to them. For example, Toulis and Airolidi (2017) obtained asymptotic normality of averaged implicit SGD, and the framework we present here can be seamlessly transferred to that setup.

Finally, we employ the estimator of  $\Sigma$  in two tasks: (i) the construction of marginal-friendly simultaneous confidence intervals that favor interpretability over ellipsoidal regions, and (ii) construct confidence intervals around predictions for new observations. The classification example we present in Section 6.4 demonstrates this feature. A similar argument can yield prediction intervals for regression as well, one which accounts for the multivariate estimation error in the SGD estimates.

## Acknowledgments

The authors sincerely appreciate the time and effort of the action editor and two anonymous Reviewers in reviewing our paper and providing insightful feedback, which has significantly improved the presentation of our work. The authors are thankful to Prof. Jing Dong for useful conversations. Dootika Vats is supported by SERB (SPG/2021/001322) and Google Asia Pacific Pte Ltd.

## Appendix A. Preliminary results

**Lemma 8** *Under the Assumption (A3), we have*

1.  $\tau_{k-1} = (k-1)b_n \asymp kb_n$ ,  $\tau_k = kb_n \asymp (k-1)b_n$  and  $\tau_{k-1} \asymp \tau_k \asymp kn^\beta$ .
2.  $\sum_{i=\tau_{k-1}+1}^{\tau_k} \eta_i > \eta b_n \tau_k^{-\alpha} > \eta b_n n^{-\alpha} =: N$ .
3. For any fixed  $\gamma > 0$  and  $\gamma \neq 1$ ,  $T^{-1} \sum_{k=1}^T k^{-\gamma} \lesssim T^{-1} \int_1^{T+1} x^{-\gamma} dx \lesssim T^{-\gamma}$ .
4. For any fixed  $\gamma > 0$ ,  $\sum_{k=1}^T k^\gamma \leq \int_1^T (x+1)^\gamma dx \lesssim T^{1+\gamma}$ .

**Proof** Proofs of 1) and 2) are obvious. For the proof of 3), observe that

$$T^{-1} \sum_{k=1}^T k^{-\gamma} \leq T^{-1} \left[ 1 + \int_1^{T+1} x^{-\gamma} dx \right] \lesssim T^{-1} \int_1^{T+1} x^{-\gamma} dx \lesssim T^{-\gamma}.$$

Next, for the proof of 4), we have

$$\sum_{k=1}^T k^\gamma \leq \int_1^T (x+1)^\gamma dx \leq \frac{1}{1+\gamma} (T+1)^{\gamma+1} \lesssim T^{1+\gamma}.$$

■

The following three results from Chen et al. (2020) will be used in our proofs.

**Result 1 (Chen et al. (2020), Lemma D.2)** For each positive integer  $j$ , let  $Y_j^k$  be a sequence of matrices with  $Y_j^j = I$  and for  $k \geq j$ ,  $Y_j^k = (I - \eta_k A)Y_j^{k-1} = \prod_{i=j}^k (I - \eta_i A)$ , where  $A$  is a positive semidefinite matrix with eigen values bounded below by  $\lambda_A > 0$ . Then, under the Assumption (A3), we have

1) For  $i \in \{j, \dots, k\}$ , the following holds

$$\|Y_j^k\| \leq \exp \left( -\lambda_A \sum_{i=j+1}^k \eta_i \right) \leq \exp(-\lambda_A(k-j)\eta_k).$$

2) Let  $S_j^k = \sum_{i=j+1}^k Y_j^i$  and  $Z_j^k = \eta_j S_j^k - A^{-1}$ , then

$$\|S_j^k\| \lesssim k^\alpha, \|Z_j^k\| \lesssim k^\alpha j^{-1} + \exp \left( -\lambda_A \sum_{i=j}^k \eta_i \right).$$

3) When  $1 \leq j < k \leq \tau_{a_n}$ , for sufficiently large  $N$ , there exists a constant  $c$  such that

$$\|S_j^k\| \lesssim j^\alpha, \|Z_j^k\| \lesssim j^{\alpha-1} + \exp(-c\lambda_A(k-j)\eta_j), \|\eta_j S_j^k\| \lesssim 1,$$

as long as  $a_n \leq N^q$  for certain positive integer  $q$ .

**Result 2 (Chen et al. (2020), Lemma D.3)** If  $S$  and  $\Sigma$  are two  $d \times d$  matrices, and  $S$  is positive definite, then  $\text{tr}(S\Sigma) \leq \text{tr}(S)\|\Sigma\|$ .

**Result 3 (Chen et al. (2020), Lemma 3.2)** Under the Assumptions (A1) and (A2), there exists  $n_0$  such that the iterate error  $D_n$  satisfy the following:

1) For  $n > m \geq n_0$ ,

$$\begin{aligned} \mathbb{E}_m \|D_n\| &\lesssim \exp \left( \frac{-1}{4} M \sum_{i=m}^n \eta_i \right) \|D_m\| + \sqrt{C_d} m^{-\alpha/2}, \\ \mathbb{E}_m \|D_n\|^2 &\lesssim \exp \left( \frac{-1}{4} M \sum_{i=m}^n \eta_i \right) \|D_m\|^2 + C_d m^{-\alpha}, \text{ and} \\ \mathbb{E}_m \|D_n\|^4 &\lesssim \exp \left( \frac{-1}{4} M \sum_{i=m}^n \eta_i \right) \|D_m\|^4 + C_d^2 m^{-2\alpha}. \end{aligned}$$

$$\begin{aligned} 2) \quad \mathbb{E} \|D_n\| &\lesssim n^{-\alpha/2} (\sqrt{C_d} + \|D_{n_0}\|), \\ \mathbb{E} \|D_n\|^2 &\lesssim n^{-\alpha} (C_d + \|D_{n_0}\|^2), \text{ and} \\ \mathbb{E} \|D_n\|^4 &\lesssim n^{-2\alpha} (C_d^2 + \|D_{n_0}\|^4). \end{aligned}$$

## Appendix B. Auxiliary sequence

An approximation of the SGD iterate is  $D_i \simeq (I_d - \eta_i A)D_{i-1} + \eta_i \xi_i$ . We replace  $\simeq$  by  $=$  to obtain an approximated iterate sequence,

$$U_i := (I_d - \eta_i A)U_{i-1} + \eta_i \xi_i, \quad U_0 := D_0.$$

The sequence  $\{U_i\}$  is known as the auxiliary sequence. The following result for the sequence  $\{U_i\}$  from Chen et al. (2020) is helpful.

**Result 4 (Chen et al. (2020), Lemma B.3)** *For the sequence  $U_i$ , under the Assumptions (A1) and (A2), we have*

$$\mathbb{E}\|U_n\|^2 \lesssim n^{-\alpha}(C_d + \|U_0\|^2 + \|D_0\|^2).$$

**Lemma 9** *Under the Assumptions (A1), (A2) and (A3), we have*

$$\hat{S} := \frac{1}{a_n} \sum_{k=1}^{a_n} \frac{1}{b_n} \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \xi_i \right) \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \xi_i \right)^\top$$

*is a consistent estimator of  $S$ . More precisely,*

$$\mathbb{E}\|\hat{S} - S\|^2 \lesssim C_d^4 n^{-\alpha} a_n^{-1/2} + C_d^6 n^{-2\alpha} + C_d^2 a_n^{-1}.$$

**Proof** Let  $\tilde{\xi}_i = -\nabla f(\theta^*, \zeta_i)$ ,  $\hat{\xi}_i = \xi_i - \tilde{\xi}_i$  and

$$\tilde{S} = \frac{1}{a_n} \sum_{k=1}^{a_n} \frac{1}{b_n} \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \tilde{\xi}_i \right) \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \tilde{\xi}_i \right)^\top.$$

Then,  $\xi_i = \hat{\xi}_i + \tilde{\xi}_i$  where  $\{\tilde{\xi}_i\}$  is a sequence of iid zero-mean random variables, and  $\{\hat{\xi}_i\}$  is a martingale difference sequence. Note that

$$\mathbb{E}\|\hat{S} - S\|^2 \lesssim \mathbb{E}\|\tilde{S} - S\|^2 + \mathbb{E}\|\hat{S} - \tilde{S}\|^2, \quad (17)$$

and

$$\mathbb{E}\|\tilde{S} - S\|^2 \leq \mathbb{E} \operatorname{tr}(\tilde{S} - S)^2 = \operatorname{tr} \mathbb{E}(\tilde{S} - S)^2. \quad (18)$$

Further

$$\mathbb{E}(\tilde{S}) = \frac{1}{a_n} \sum_{k=1}^{a_n} \frac{1}{b_n} \sum_{i=\tau_{k-1}+1}^{\tau_k} \mathbb{E}(\tilde{\xi}_i \tilde{\xi}_i^\top) = S,$$

therefore  $\mathbb{E}(\tilde{S} - S)^2 = \mathbb{E}(\tilde{S}^2) - S^2$ . Next

$$\mathbb{E}\tilde{S}^2 = a_n^{-2} \sum_{j=1}^{a_n} \sum_{k=1}^{a_n} b_n^{-2} \sum_{i_1, i_2=\tau_{j-1}+1}^{\tau_j} \sum_{i_3, i_4=\tau_{k-1}+1}^{\tau_k} \mathbb{E}(\tilde{\xi}_{i_1} \tilde{\xi}_{i_2}^\top \tilde{\xi}_{i_3} \tilde{\xi}_{i_4}^\top).$$

Since  $\{\tilde{\xi}_i\}$  is an iid sequence of zero-mean random variables, on the RHS of the above expression, the terms are zero unless  $i_1 = i_2$ ,  $i_3 = i_4$  or  $i_1 = i_3$ ,  $i_2 = i_4$  or  $i_1 = i_4$ ,  $i_2 = i_3$ . Also, the latter two only happen when all the indices are in the same batch, thus

$$\mathbb{E}\tilde{S}^2 = a_n^{-2} \sum_{j=1}^{a_n} \sum_{k=1}^{a_n} b_n^{-2} \sum_{i_1=\tau_{j-1}+1}^{\tau_j} \sum_{i_3=\tau_{k-1}+1}^{\tau_k} \mathbb{E}(\tilde{\xi}_{i_1} \tilde{\xi}_{i_1}^\top \tilde{\xi}_{i_3} \tilde{\xi}_{i_3}^\top)$$

$$+ a_n^{-2} \sum_{j=1}^{a_n} b_n^{-2} \sum_{i_1, i_3 = \tau_{j-1} + 1, i_1 \neq i_3}^{\tau_j} [\mathbb{E}(\tilde{\xi}_{i_1} \tilde{\xi}_{i_3}^\top \tilde{\xi}_{i_1} \tilde{\xi}_{i_3}^\top) + \mathbb{E}(\tilde{\xi}_{i_1} \tilde{\xi}_{i_3}^\top \tilde{\xi}_{i_3} \tilde{\xi}_{i_1}^\top)].$$

Using  $\mathbb{E}(\tilde{\xi}_{i_1} \tilde{\xi}_{i_1}^\top \tilde{\xi}_{i_3} \tilde{\xi}_{i_3}^\top) = S^2$  if  $i_1 \neq i_3$ , we get

$$\text{tr}(\mathbb{E}\tilde{S}^2 - S^2) = a_n^{-2} \sum_{j=1}^{a_n} b_n^{-2} \sum_{i_1, i_3 = \tau_{j-1} + 1, i_1 \neq i_3}^{\tau_j} [\mathbb{E}(\tilde{\xi}_{i_1} \tilde{\xi}_{i_3}^\top \tilde{\xi}_{i_1} \tilde{\xi}_{i_3}^\top) + \mathbb{E}(\tilde{\xi}_{i_1} \tilde{\xi}_{i_3}^\top \tilde{\xi}_{i_3} \tilde{\xi}_{i_1}^\top)].$$

Next using  $\mathbb{E}(\tilde{\xi}_{i_1} \tilde{\xi}_{i_1}^\top \tilde{\xi}_{i_3} \tilde{\xi}_{i_3}^\top) \lesssim C_d^2 + C_d^6 i_1^{-2\alpha}$  if  $i_1 = i_3$ , we obtain

$$\begin{aligned} \text{tr}(\mathbb{E}\tilde{S}^2 - S^2) &= a_n^{-2} \sum_{j=1}^{a_n} b_n^{-2} \left( \sum_{i_1 = \tau_{j-1} + 1}^{\tau_j} \mathbb{E}\|\tilde{\xi}_{i_1}\|^4 + \sum_{i_1, i_3 = \tau_{j-1} + 1, i_1 \neq i_3}^{\tau_j} (\mathbb{E}(\tilde{\xi}_{i_1}^\top \tilde{\xi}_{i_3})^2 + \mathbb{E}(\|\tilde{\xi}_{i_1}\|^2 \|\tilde{\xi}_{i_3}\|^2)) \right) \\ &\lesssim a_n^{-2} \sum_{j=1}^{a_n} \left( b_n^{-1} (C_d^2 + C_d^6 \tau_j^{-2\alpha}) + (\text{tr}(S))^2 \right). \end{aligned}$$

Consequently using Lemma 8

$$\text{tr}(\mathbb{E}\tilde{S}^2 - S^2) \lesssim C_d^2 n^{-1} + C_d^6 a_n^{-(1+2\alpha)} + C_d^2 a_n^{-1} \lesssim C_d^2 a_n^{-1}. \quad (19)$$

Next, we denote

$$\check{S} = \frac{1}{a_n} \sum_{k=1}^{a_n} \frac{1}{b_n} \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \hat{\xi}_i \right) \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \hat{\xi}_i \right)^\top.$$

We notice that

$$\begin{aligned} \mathbb{E}\|\hat{S} - \check{S}\|^2 &= \mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n^{-1} \left( \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \tilde{\xi}_i \right) \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \hat{\xi}_i \right)^\top + \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \hat{\xi}_i \right) \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \tilde{\xi}_i \right)^\top \right) + 2\check{S} \right\|^2 \\ &\lesssim \mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n^{-1} \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \tilde{\xi}_i \right) \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \hat{\xi}_i \right)^\top \right\|^2 + \mathbb{E}\|\check{S}\|^2 \\ &\lesssim \sqrt{\mathbb{E}(\text{tr}(\check{S}^2))} \sqrt{\mathbb{E}(\text{tr}(\check{S}^2))} + \mathbb{E}(\text{tr}(\check{S}^2)). \end{aligned} \quad (20)$$

Now we need to find a bound for  $\mathbb{E}(\text{tr}(\check{S}^2))$ . Let  $\bar{\xi}_k = b_n^{-1} \sum_{i=\tau_{k-1}+1}^{\tau_k} \hat{\xi}_i$ . Then using Cauchy-Schwartz inequality, we have

$$\begin{aligned} \mathbb{E}(\text{tr}(\check{S}^2)) &= a_n^{-2} \sum_{k=1}^{a_n} \sum_{l=1}^{a_n} b_n^{-2} \mathbb{E}(\text{tr}(\bar{\xi}_k \bar{\xi}_k^\top \bar{\xi}_l \bar{\xi}_l^\top)) \lesssim a_n^{-2} \sum_{k=1}^{a_n} \sum_{l=1}^{a_n} b_n^{-2} \sqrt{\mathbb{E}\|\bar{\xi}_k\|^4} \sqrt{\mathbb{E}\|\bar{\xi}_l\|^4} \\ &\lesssim a_n^{-2} \sum_{k=1}^{a_n} \sum_{l=1}^{a_n} b_n^{-2} \sum_{i=\tau_{k-1}+1}^{\tau_k} \sqrt{\mathbb{E}\|\hat{\xi}_i\|^4} \sum_{j=\tau_{l-1}+1}^{\tau_l} \sqrt{\mathbb{E}\|\hat{\xi}_j\|^4}, \text{ using Theorem 2.1 in Rio (2009)}. \end{aligned}$$



Using Result 3, we have  $\mathbb{E}\|\hat{\xi}_j\|_2^4 \lesssim C_d^6 i^{-2\alpha}$ . Consequently using Lemma 8 we get

$$\mathbb{E}(\text{tr}(\check{S}^2)) \lesssim C_d^6 (n^{-1} \sum_{i=1}^n i^{-\alpha}) (n^{-1} \sum_{j=1}^n j^{-\alpha}) \lesssim C_d^6 n^{-2\alpha}.$$

Further substituting in (20) we have

$$\mathbb{E}\|\hat{S} - \tilde{S}\|^2 \lesssim C_d^4 n^{-\alpha} a_n^{-1/2} + C_d^6 n^{-2\alpha}. \quad (21)$$

Combining (17), (18), (19), (20) and (21), completes the proof.  $\blacksquare$

Define the overall mean and batch means of  $U_i$  as follows:

$$\bar{U}_n = \frac{1}{\tau_{a_n}} \sum_{i=1}^{\tau_{a_n}} U_i \quad \text{and} \quad \bar{U}_k = \frac{1}{b_n} \sum_{i=\tau_{k-1}+1}^{\tau_k} U_i \quad \text{for } k = 1, 2, \dots, a_n.$$

**Lemma 10** *Under the Assumptions (A1), (A2) and (A3), we have*

$$\mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n \bar{U}_k \bar{U}_k^\top - A^{-1} S A^{-1} \right\| \lesssim C_d^2 n^{-\alpha/2} a_n^{-1/4} + C_d^3 n^{-\alpha} + C_d a_n^{-1/2} + C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2}.$$

**Proof** The recursion of  $U_j$  can be written as

$$\begin{aligned} U_i &= Y_{i-1}^i U_{i-1} + Y_i^i \eta_i \xi_i = Y_{i-2}^i U_{i-2} + Y_{i-1}^i \eta_{i-1} \xi_{i-1} + Y_i^i \eta_i \xi_i \\ &= Y_j^i U_j + \sum_{l=j+1}^i Y_l^i \eta_l \xi_l. \end{aligned}$$

Therefore, the  $k^{\text{th}}$  batch mean  $\bar{U}_k$  can be written as

$$\begin{aligned} \bar{U}_k &= b_n^{-1} \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} Y_{\tau_{k-1}}^i U_{\tau_{k-1}} + \sum_{i=\tau_{k-1}+1}^{\tau_k} \sum_{l=\tau_{k-1}+1}^i Y_l^i \eta_l \xi_l \right) \\ &= b_n^{-1} \left[ \sum_{i=\tau_{k-1}+1}^{\tau_k} Y_{\tau_{k-1}}^i U_{\tau_{k-1}} + \sum_{l=\tau_{k-1}+1}^{\tau_k} \left( \sum_{i=l}^{\tau_k} Y_l^i \right) \eta_l \xi_l \right] \\ &= b_n^{-1} \left[ \sum_{i=\tau_{k-1}+1}^{\tau_k} Y_{\tau_{k-1}}^i U_{\tau_{k-1}} + \sum_{l=\tau_{k-1}+1}^{\tau_k} (S_l^{\tau_k} + I) \eta_l \xi_l \right] \\ &= b_n^{-1} S_{\tau_{k-1}}^{\tau_k} U_{\tau_{k-1}} + b_n^{-1} A^{-1} \sum_{l=\tau_{k-1}+1}^{\tau_k} \xi_l + b_n^{-1} \sum_{l=\tau_{k-1}+1}^{\tau_k} \xi_l Z_l^{\tau_k} + b_n^{-1} \sum_{l=\tau_{k-1}+1}^{\tau_k} \eta_l \xi_l. \end{aligned}$$

Let us denote

$$A_k := -A^{-1} \sum_{l=\tau_{k-1}+1}^{\tau_k} \xi_l \quad \text{and} \quad B_k := S_{\tau_{k-1}}^{\tau_k} U_{\tau_{k-1}} + \sum_{l=\tau_{k-1}+1}^{\tau_k} \xi_l Z_l^{\tau_k} + \sum_{l=\tau_{k-1}+1}^{\tau_k} \eta_l \xi_l.$$

Then,

$$a_n^{-1} \sum_{k=1}^{a_n} b_n \bar{U}_k \bar{U}_k^\top = a_n^{-1} \sum_{k=1}^{a_n} b_n^{-1} A_k A_k^\top + a_n^{-1} \sum_{k=1}^{a_n} b_n^{-1} [A_k B_k^\top + B_k A_k^\top + B_k B_k^\top]. \quad (22)$$

We have from Lemma 9,

$$\begin{aligned} \mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n^{-1} A_k A_k^\top - A^{-1} S A^{-1} \right\|^2 &= \mathbb{E} \left\| A^{-1} (\hat{S} - S) A^{-1} \right\|^2 \\ &\lesssim C_d^4 n^{-\alpha} a_n^{-1/2} + C_d^6 n^{-2\alpha} + C_d^2 a_n^{-1}. \end{aligned} \quad (23)$$

Using Cauchy Schwartz inequality, we have

$$\mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n^{-1} A_k A_k^\top - A^{-1} S A^{-1} \right\| \lesssim C_d^2 n^{-\alpha/2} a_n^{-1/4} + C_d^3 n^{-\alpha} + C_d a_n^{-1/2}.$$

Using  $\|B_k\|^2 = \|B_k B_k^\top\| \leq \text{tr}(B_k B_k^\top)$  and  $\xi_i$  being a martingale, we have

$$\begin{aligned} &\mathbb{E} B_k B_k^\top \\ &= \mathbb{E} \left[ S_{\tau_{k-1}}^{\tau_k} U_{\tau_{k-1}} + \sum_{l=\tau_{k-1}+1}^{\tau_k} \xi_l Z_l^{\tau_k} + \sum_{l=\tau_{k-1}+1}^{\tau_k} \eta_l \xi_l \right] \left[ S_{\tau_{k-1}}^{\tau_k} U_{\tau_{k-1}} + \sum_{l=\tau_{k-1}+1}^{\tau_k} \xi_l Z_l^{\tau_k} + \sum_{l=\tau_{k-1}+1}^{\tau_k} \eta_l \xi_l \right]^\top \\ &= S_{\tau_{k-1}}^{\tau_k} \mathbb{E} \left[ U_{\tau_{k-1}} U_{\tau_{k-1}}^\top \right] (S_{\tau_{k-1}}^{\tau_k})^\top + \sum_{l=\tau_{k-1}+1}^{\tau_k} (Z_l^{\tau_k} + \eta_l I) \mathbb{E}(\xi_l \xi_l^\top) (Z_l^{\tau_k} + \eta_l I)^\top. \\ \Rightarrow \text{tr}(\mathbb{E} B_k B_k^\top) &\leq \|S_{\tau_{k-1}}^{\tau_k}\|^2 \mathbb{E} \|U_{\tau_{k-1}}\|^2 + \sum_{l=\tau_{k-1}+1}^{\tau_k} \|Z_l^{\tau_k} + \eta_l I\|^2 \mathbb{E} \|\xi_l\|^2. \end{aligned}$$

Now, using Result 1, we have

$$\|Z_l^{\tau_k} + \eta_l I\| \leq \|Z_l^{\tau_k}\| + \eta l^{-\alpha} \lesssim l^{\alpha-1} + \exp(-c\lambda_A(\tau_k - l)\eta) + l^{-\alpha}.$$

Using  $l^{-\alpha} < l^{\alpha-1}$  for  $\alpha \in (0.5, 1)$  and Cauchy-Schwartz inequality, we get

$$\|Z_l^{\tau_k} + \eta_l I\|^2 \lesssim l^{2\alpha-2} + \exp(-2c\lambda_A(\tau_k - l)\eta).$$

Next, using the bound of  $\|U_n\|^2$  in Result 4, and  $\mathbb{E}\|\xi_l\|^2 \lesssim C_d$ , we obtain

$$\begin{aligned} \frac{1}{C_d} \text{tr}(\mathbb{E} B_k B_k^\top) &\lesssim (\tau_{k-1} + 1)^\alpha + \sum_{l=\tau_{k-1}+1}^{\tau_k} \|Z_l^{\tau_k} + \eta_l I\|^2 \\ &\lesssim (\tau_{k-1} + 1)^\alpha + b_n (\tau_{k-1} + 1)^{2\alpha-2} + \sum_{l=\tau_{k-1}+1}^{\tau_k} \exp(-2c\lambda_A(\tau_k - l)\eta_{\tau_{k-1}+1}) \\ &\lesssim (\tau_{k-1} + 1)^\alpha + b_n (\tau_{k-1} + 1)^{2\alpha-2} + \sum_{l=0}^{\infty} \exp(-2c\lambda_A l \eta_{\tau_{k-1}+1}) \\ &\lesssim (\tau_{k-1} + 1)^\alpha + b_n (\tau_{k-1} + 1)^{2\alpha-2} + (2c\lambda_A \eta_{\tau_{k-1}+1})^{-1}. \end{aligned}$$

Using

$$\begin{aligned} b_n(\tau_{k-1} + 1)^{2\alpha-2} + (2c\lambda_A\eta_{\tau_{k-1}+1})^{-1} &= b_n(k-1)^{2\alpha-2}b_n^{2\alpha-2} + (k-1)^{-\alpha}b_n^{-\alpha}\frac{1}{2c\lambda_A} \\ &\lesssim (k-1)^{2\alpha-2}b_n^{2\alpha-1} + (k-1)^{-\alpha}b_n^{-\alpha} \lesssim b_n^{2\alpha-1}, \end{aligned}$$

and Lemma 8 (claim 4), we have

$$\begin{aligned} a_n^{-1} \sum_{k=1}^{a_n} b_n^{-1} \mathbb{E} \|B_k\|^2 &\lesssim C_d a_n^{-1} \sum_{k=1}^{a_n} b_n^{-1} (b_n^{2\alpha-1} + (\tau_{k-1} + 1)^\alpha) \\ &\leq C_d b_n^{2\alpha-2} + C_d a_n^{-1} \sum_{k=1}^{a_n} b_n^{-1} (\tau_{k-1} + 1)^\alpha \\ &\leq C_d b_n^{2\alpha-2} + C_d b_n^{-1} n^\alpha. \end{aligned} \quad (24)$$

On the other hand, using martingale property of  $\{\xi_i\}$ , we have

$$\begin{aligned} b_n^{-1} \mathbb{E} \|A_k\|^2 &= b_n^{-1} \text{tr}(\mathbb{E} A_k A_k^\top) = b_n^{-1} \text{tr} \left( A^{-1} \mathbb{E} \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \xi_i \right) \left( \sum_{i=\tau_{k-1}+1}^{\tau_k} \xi_i \right)^\top A^{-1} \right) \\ &\leq b_n^{-1} \|A^{-1}\|^2 \left[ b_n \text{tr}(S) + 2 \sum_{i=\tau_{k-1}+1}^{\tau_k} \mathbb{E}(\sigma_1 \|D_i\| + \sigma_2 \|D_i\|^2) \right] \lesssim C_d. \end{aligned} \quad (25)$$

So, using (24), (25) and Cauchy-Schwartz inequality, we get

$$b_n^{-1} \mathbb{E} \|A_k\| \|B_k\| \lesssim \sqrt{b_n^{-1} \mathbb{E} \|A_k\|^2} \sqrt{b_n^{-1} \mathbb{E} \|B_k\|^2} \lesssim C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2}. \quad (26)$$

Thus, using (22), (23) and (26), we obtain

$$\begin{aligned} &\mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n \bar{U}_k \bar{U}_k^\top - A^{-1} S A^{-1} \right\| \\ &\lesssim \mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n^{-1} A_k A_k^\top - A^{-1} S A^{-1} \right\| + 2a_n^{-1} \sum_{k=1}^{a_n} b_n^{-1} \left( \mathbb{E} \|A_k\| \|B_k\| + \mathbb{E} \|B_k\|^2 \right) \\ &\lesssim C_d^2 n^{-\alpha/2} a_n^{-1/4} + C_d^3 n^{-\alpha} + C_d a_n^{-1/2} + C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2}. \end{aligned}$$

■

**Lemma 11** *Under the Assumptions (A1), (A2) and (A3), and for sufficiently large  $n$ , we have*

$$\begin{aligned} &\mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{U}_k - \bar{U}_n) (\bar{U}_k - \bar{U}_n)^\top - A^{-1} S A^{-1} \right\| \\ &\lesssim C_d^2 n^{-\alpha/2} a_n^{-1/4} + C_d^3 n^{-\alpha} + C_d a_n^{-1/2} + C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2} + C_d a_n^{-1}. \end{aligned}$$

**Proof** Observe that

$$\begin{aligned} a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{U}_k - \bar{U}_n) (\bar{U}_k - \bar{U}_n)^\top &= a_n^{-1} \sum_{k=1}^{a_n} b_n \bar{U}_k \bar{U}_k^\top - a_n^{-1} \sum_{k=1}^{a_n} b_n \bar{U}_n \bar{U}_n^\top \\ &= a_n^{-1} \sum_{k=1}^{a_n} b_n \bar{U}_k \bar{U}_k^\top - b_n \bar{U}_n \bar{U}_n^\top. \end{aligned} \quad (27)$$

From Lemma 10, we have

$$\mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n \bar{U}_k \bar{U}_k^\top - A^{-1} S A^{-1} \right\| \lesssim C_d^2 n^{-\alpha/2} a_n^{-1/4} + C_d^3 n^{-\alpha} + C_d a_n^{-1/2} + C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2}.$$

Next, using

$$\bar{U}_n = n^{-1} S_0^{\tau_{a_n}} U_0 + n^{-1} \sum_{i=1}^{\tau_{a_n}} (S_i^{\tau_{a_n}} + I) \eta_i \xi_i,$$

and martingale property of  $\xi_i$ , we get

$$\begin{aligned} &\mathbb{E} \|\bar{U}_n\|^2 \\ &\leq n^{-2} \|S_0^{\tau_{a_n}}\|^2 \mathbb{E} \|U_0\|^2 + n^{-2} \sum_{i=1}^{\tau_{a_n}} \|(S_i^{\tau_{a_n}} + I)\|^2 \eta_i^2 \mathbb{E} \|\xi_i\|^2 \\ &\leq n^{-2} \|S_0^{\tau_{a_n}}\|^2 \mathbb{E} \|U_0\|^2 + n^{-2} \sum_{i=1}^{\tau_{a_n}} \eta_i^2 \|(S_i^{\tau_{a_n}} + I)\|^2 [\text{tr}(S) + \mathbb{E}(\sigma_1 \|D_i\| + \sigma_2 \|D_i\|^2)]. \end{aligned} \quad (28)$$

Further, using Result 1 and Lemma 9, we have  $\|S_0^{\tau_{a_n}}\|^2 \mathbb{E} \|U_0\|^2 \lesssim C_d$ ,  $\eta_i^2 \|(S_i^{\tau_{a_n}} + I)\|$  is uniformly bounded, and  $[\text{tr}(S) + \mathbb{E}(\sigma_1 \|D_i\| + \sigma_2 \|D_i\|^2)] \lesssim C_d$ . Therefore,

$$n^{-2} \sum_{i=1}^{\tau_{a_n}} \eta_i^2 \|(S_i^{\tau_{a_n}} + I)\|^2 [\text{tr}(S) + \mathbb{E}(\sigma_1 \|D_i\| + \sigma_2 \|D_i\|^2)] \lesssim n^{-1} C_d. \quad (29)$$

Hence, using (27)-(29) and Lemma 10, we get

$$\begin{aligned} &\mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{U}_k - \bar{U}_n) (\bar{U}_k - \bar{U}_n)^\top - A^{-1} S A^{-1} \right\| \\ &\leq \mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n \bar{U}_k \bar{U}_k^\top - A^{-1} S A^{-1} \right\| + \mathbb{E} \|b_n \bar{U}_n \bar{U}_n^\top\| \\ &\lesssim C_d^2 n^{-\alpha/2} a_n^{-1/4} + C_d^3 n^{-\alpha} + C_d a_n^{-1/2} + C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2} + C_d a_n^{-1}. \end{aligned}$$

■

### Appendix C. Consistency of batch-means estimator

We denote overall mean and batch means of  $D_i$  as follows:

$$\bar{D}_n = \frac{1}{\tau_{a_n}} \sum_{i=1}^{\tau_{a_n}} D_i \text{ and } \bar{D}_k = \frac{1}{b_n} \sum_{i=\tau_{k-1}+1}^{\tau_k} D_i \text{ for } k = 1, 2, \dots, a_n.$$

**Proof (Proof of Theorem 1)** We have the linear auxiliary iterate sequence  $U_i = (I_d - \eta_i A)U_{i-1} + \eta_i \xi_i$ ,  $U_0 = D_0$ . Let  $\delta_i := D_i - U_i$ , then

$$\delta_i = (I - \eta_i A)\delta_{i-1} + \eta_i(AD_{i-1} - \nabla F(\theta_{i-1})).$$

Define overall mean and batch means of  $\delta_i$  as follows:

$$\bar{\delta}_n = \frac{1}{\tau_{a_n}} \sum_{i=1}^{\tau_{a_n}} \delta_i \text{ and } \bar{\delta}_k = \frac{1}{b_n} \sum_{i=\tau_{k-1}+1}^{\tau_k} \delta_i \text{ for } k = 1, 2, \dots, a_n.$$

Now observe that

$$\begin{aligned} & a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{\theta}_k - \bar{\theta}_n)(\bar{\theta}_k - \bar{\theta}_n)^\top \\ &= a_n^{-1} \sum_{k=1}^{a_n} b_n \bar{D}_k \bar{D}_k^\top - a_n^{-1} \sum_{k=1}^{a_n} b_n \bar{D}_n \bar{D}_n^\top \\ &= a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{U}_k + \bar{\delta}_k)(\bar{U}_k + \bar{\delta}_k)^\top - a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{U}_n + \bar{\delta}_n)(\bar{U}_n + \bar{\delta}_n)^\top \\ &= a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{U}_k - \bar{U}_n)(\bar{U}_k - \bar{U}_n)^\top + a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{U}_k - \bar{U}_n)(\bar{\delta}_k - \bar{\delta}_n)^\top \\ &\quad + a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{\delta}_k - \bar{\delta}_n)(\bar{U}_k - \bar{U}_n)^\top + a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{\delta}_k - \bar{\delta}_n)(\bar{\delta}_k - \bar{\delta}_n)^\top. \end{aligned}$$

Then, by using  $\|C\| \leq \text{tr}(C)$  for a positive semidefinite matrix  $C$ , and Cauchy-Schwartz inequality, we have

$$\begin{aligned} & \mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{\theta}_k - \bar{\theta}_n)(\bar{\theta}_k - \bar{\theta}_n)^\top - A^{-1} S A^{-1} \right\| \\ & \leq \mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{U}_k - \bar{U}_n)(\bar{U}_k - \bar{U}_n)^\top - A^{-1} S A^{-1} \right\| \\ & \quad + a_n^{-1} \sum_{k=1}^{a_n} b_n \mathbb{E} \text{tr}[(\bar{\delta}_k - \bar{\delta}_n)(\bar{\delta}_k - \bar{\delta}_n)^\top] \\ & \quad + \frac{2}{a_n} \sqrt{\sum_{k=1}^{a_n} b_n \mathbb{E} \text{tr}[(\bar{U}_k - \bar{U}_n)(\bar{U}_k - \bar{U}_n)^\top] \sum_{k=1}^{a_n} b_n \mathbb{E} \text{tr}[(\bar{\delta}_k - \bar{\delta}_n)(\bar{\delta}_k - \bar{\delta}_n)^\top]}, \quad (30) \end{aligned}$$

and  $\sum_{k=1}^{a_n} b_n \mathbb{E} \text{tr}[(\bar{\delta}_k - \bar{\delta}_n)(\bar{\delta}_k - \bar{\delta}_n)^\top] \leq \sum_{k=1}^{a_n} b_n \mathbb{E} \text{tr}[\bar{\delta}_k \bar{\delta}_k^\top]$ . Further, using the notations  $Y_j^k$  and  $S_j^k$  in Result 1, we get for  $i > \tau_{k-1}$

$$\begin{aligned}
\delta_i &= (I - \eta_i A) \delta_{i-1} + \eta_i (AD_{i-1} - \nabla F(\theta_{i-1})) \\
&= Y_{\tau_{k-1}}^i \delta_{\tau_{k-1}} + \sum_{j=\tau_{k-1}+1}^i Y_j^i \eta_j (AD_{j-1} - \nabla F(\theta_{j-1})) \\
\Rightarrow \bar{\delta}_k &= \frac{1}{b_n} \sum_{i=\tau_{k-1}+1}^{\tau_k} Y_{\tau_{k-1}}^i \delta_{\tau_{k-1}} + \frac{1}{b_n} \sum_{i=\tau_{k-1}+1}^{\tau_k} \sum_{j=\tau_{k-1}+1}^i Y_j^i \eta_j (AD_{j-1} - \nabla F(\theta_{j-1})) \\
&= \frac{1}{b_n} S_{\tau_{k-1}}^{\tau_k} \delta_{\tau_{k-1}} + \frac{1}{b_n} \sum_{j=\tau_{k-1}+1}^{\tau_k} \sum_{i=j}^{\tau_k} Y_j^i \eta_j (AD_{j-1} - \nabla F(\theta_{j-1})) \\
&= \frac{1}{b_n} S_{\tau_{k-1}}^{\tau_k} \delta_{\tau_{k-1}} + \frac{1}{b_n} \sum_{j=\tau_{k-1}+1}^{\tau_k} (I + S_j^{\tau_k}) \eta_j (AD_{j-1} - \nabla F(\theta_{j-1})).
\end{aligned}$$

Now, by Cauchy-Schwartz inequality,

$$\begin{aligned}
\mathbb{E} \|\bar{\delta}_k\|^2 &\leq \frac{2}{b_n^2} \|S_{\tau_{k-1}}^{\tau_k}\|^2 \mathbb{E} \|\delta_{\tau_{k-1}}\|^2 + \\
&\quad + \frac{2}{b_n^2} \left( \sum_{j=\tau_{k-1}+1}^{\tau_k} \|(I + S_j^{\tau_k}) \eta_j\|^2 \right) \mathbb{E} \left( \sum_{j=\tau_{k-1}+1}^{\tau_k} \|AD_{j-1} - \nabla F(\theta_{j-1})\|^2 \right). \quad (31)
\end{aligned}$$

Using Result 1 and Result 3, we have

$$\|S_{\tau_{k-1}}^{\tau_k}\|^2 \lesssim (\tau_{k-1} + 1)^{2\alpha}, \quad \sum_{j=\tau_{k-1}+1}^{\tau_k} \|S_j^{\tau_k} \eta_j\|^2 \lesssim b_n \quad (32)$$

$$\text{and } \mathbb{E} \left( \sum_{j=\tau_{k-1}+1}^{\tau_k} \|AD_{j-1} - \nabla F(\theta_{j-1})\|^2 \right) \lesssim \sum_{j=\tau_{k-1}+1}^{\tau_k} L_F^2 \mathbb{E} \|D_{j-1}\|^4 \lesssim C_d^4 b_n \tau_{k-1}^{-2\alpha}. \quad (33)$$

Notice that Result 4 holds for both  $U_i$  and  $D_i$ , so

$$\mathbb{E} \|\delta_{\tau_{k-1}}\|^2 \leq 2\mathbb{E} \|D_{\tau_{k-1}}\|^2 + 2\mathbb{E} \|U_{\tau_{k-1}}\|^2 \lesssim \tau_{k-1}^{-\alpha} C_d. \quad (34)$$

Therefore, using (31)-(34) and Lemma 8, we have

$$\begin{aligned}
\mathbb{E} \|\bar{\delta}_k\|^2 &\lesssim \frac{1}{b_n^2} \tau_{k-1}^\alpha + \frac{1}{b_n^2} b_n C_d^4 b_n \tau_{k-1}^{-2\alpha} \lesssim \frac{1}{b_n^2} \tau_{k-1}^\alpha + C_d^4 \tau_{k-1}^{-2\alpha} \\
\Rightarrow a_n^{-1} \sum_{k=1}^{a_n} b_n \mathbb{E} \|\bar{\delta}_k\|^2 &\lesssim C_d^4 b_n a_n^{-1} \sum_{k=1}^{a_n} \tau_{k-1}^{-2\alpha} \lesssim C_d^4 b_n^{1-2\alpha} a_n^{-1} \sum_{k=1}^{a_n} k^{-2\alpha} \asymp C_d^4 b_n^{1-2\alpha} a_n^{-2\alpha} = C_d^4 n^{-2\alpha} b_n. \quad (35)
\end{aligned}$$

Thus, using (30) and (35), we obtain

$$\mathbb{E} \left\| a_n^{-1} \sum_{k=1}^{a_n} b_n (\bar{\theta}_k - \bar{\theta}_n) (\bar{\theta}_k - \bar{\theta}_n)^\top - A^{-1} S A^{-1} \right\|$$

$$\lesssim C_d^2 n^{-\alpha/2} a_n^{-1/4} + C_d^3 n^{-\alpha} + C_d a_n^{-1/2} + C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2} + C_d a_n^{-1} + C_d^4 n^{-2\alpha} b_n.$$

■

## Appendix D. Lugsail estimator

This section provides some additional details regarding Lugsail estimator and proof of Proposition 5.

### D.1 Details for Example 1

The mean estimation model is given by

$$y = \theta^* + \epsilon,$$

where  $\theta^* \in \mathbb{R}$  and  $\epsilon$  is the random error term with mean zero. Let  $y_i$  be a sequence of iid observations from the model. Consider the square error loss function  $F(\theta) = (y - \theta)^2/2$ . Without loss of generality, take  $\theta^* = 0$  and  $\theta_0 = 0$ , then the  $i^{th}$  SGD iterate has the form

$$\theta_i = \theta_{i-1} + \eta_i(y_i - \theta_{i-1}) = (1 - \eta_i)\theta_{i-1} + \eta_i\epsilon_i, \quad i \geq 1. \quad (36)$$

This implies

$$\theta_i = \sum_{p=1}^i \prod_{k=p+1}^i (1 - k^\alpha) p^{-\alpha} \epsilon_p.$$

Now, the estimand is

$$\text{Var}(\sqrt{n} \hat{\theta}_n) = n \text{Var} \left( \frac{1}{a_n} \sum_{k=1}^{a_n} \bar{\theta}_k \right) = \frac{b_n^2}{n} \mathbb{E} \left( \sum_{k=1}^{a_n} \bar{\theta}_k \right)^2,$$

and the proposed estimator is

$$\hat{\Sigma}_n = \frac{1}{a_n} \sum_{k=1}^{a_n} b_n (\bar{\theta}_k - \hat{\theta}_n)^2 = \frac{b_n^2}{n} \sum_{k=1}^{a_n} \bar{\theta}_k^2 + \frac{b_n}{n} O_p(1).$$

Thus, ignoring the second term, which is tending to zero at a higher rate, we have

$$\begin{aligned} \text{Bias}(\hat{\Sigma}_n) &= \mathbb{E}(\hat{\Sigma}_n) - \text{Var}(\sqrt{n} \hat{\theta}_n) \\ &\approx \frac{-2b_n^2}{n} \sum_{1 \leq j < k \leq a_n} \text{Cov}(\bar{\theta}_j, \bar{\theta}_k) = \frac{-2}{n} \sum_{1 \leq j < k \leq a_n} \text{Cov}(b_n \bar{\theta}_j, b_n \bar{\theta}_k). \end{aligned} \quad (37)$$

Further, using the fact that  $\text{Cov}(\theta_p, \theta_q) = C_1 q^{-\alpha} (1 - q^{-\alpha})^{q-p}$  for  $p < q$ , where  $C_1$  is fixed constant, we get for  $j < k$

$$\text{Cov}(b_n \bar{\theta}_j, b_n \bar{\theta}_k) = \sum_{p=\tau_{j-1}+1}^{\tau_j} \sum_{q=\tau_{k-1}+1}^{\tau_k} \text{Cov}(\theta_p, \theta_q) = C_1 \sum_{p=\tau_{j-1}+1}^{\tau_j} \sum_{q=\tau_{k-1}+1}^{\tau_k} q^{-\alpha} (1 - q^{-\alpha})^{q-p}. \quad (38)$$

Using (37) and (38), we get

$$\text{Bias}(\hat{\Sigma}_{b_n}) \approx \frac{-2}{n} \sum_{1 \leq j < k \leq a_n} \text{Cov}(b_n \bar{\theta}_j, b_n \bar{\theta}_k) = \frac{-2C_1}{n} \sum_{1 \leq j < k \leq a_n} \sum_{p=\tau_{j-1}+1}^{\tau_j} \sum_{q=\tau_{k-1}+1}^{\tau_k} q^{-\alpha} (1 - q^{-\alpha})^{q-p}.$$

## D.2 Alternate expression

Recall that for batch size  $b_n$ , the  $k^{\text{th}}$  batch-mean vector is  $\bar{\theta}_k = b_n^{-1} \sum_{i=\tau_{k-1}+1}^{\tau_k} \theta_i$ . Define the mean of adjacent batches as  $\tilde{\theta}_j = (\bar{\theta}_{2j-1} + \bar{\theta}_{2j})/2$ . Then  $\tilde{\theta}_j$  is the  $j^{\text{th}}$  batch-mean vector with batch size  $2b_n$ . Now, with batch means estimator

$$\hat{\Sigma}_{b_n} = \frac{b_n}{a_n} \sum_{k=1}^{a_n} (\bar{\theta}_k - \hat{\theta}_n) (\bar{\theta}_k - \hat{\theta}_n)^\top,$$

we can write  $\hat{\Sigma}_{2b_n}$  as

$$\hat{\Sigma}_{2b_n} = \frac{2b_n}{a_n/2} \sum_{j=1}^{a_n/2} (\tilde{\theta}_j - \hat{\theta}_n) (\tilde{\theta}_j - \hat{\theta}_n)^\top.$$

So,

$$\begin{aligned} \hat{\Sigma}_{2b_n} &= \frac{2b_n}{a_n/2} \sum_{j=1}^{a_n/2} \left( \frac{\bar{\theta}_{2j-1} + \bar{\theta}_{2j}}{2} - \hat{\theta}_n \right) \left( \frac{\bar{\theta}_{2j-1} + \bar{\theta}_{2j}}{2} - \hat{\theta}_n \right)^\top \\ &= \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (\bar{\theta}_{2j-1} + \bar{\theta}_{2j} - 2\hat{\theta}_n) (\bar{\theta}_{2j-1} + \bar{\theta}_{2j} - 2\hat{\theta}_n)^\top, \\ &= \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} [(\bar{\theta}_{2j-1} - \hat{\theta}_n)(\bar{\theta}_{2j-1} - \hat{\theta}_n)^\top + (\bar{\theta}_{2j} - \hat{\theta}_n)(\bar{\theta}_{2j} - \hat{\theta}_n)^\top + (\bar{\theta}_{2j-1} - \hat{\theta}_n)(\bar{\theta}_{2j} - \hat{\theta}_n)^\top \\ &\quad + (\bar{\theta}_{2j} - \hat{\theta}_n)(\bar{\theta}_{2j-1} - \hat{\theta}_n)^\top] \\ &= \hat{\Sigma}_b + \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} [(\bar{\theta}_{2j-1} - \hat{\theta}_n)(\bar{\theta}_{2j} - \hat{\theta}_n)^\top + (\bar{\theta}_{2j} - \hat{\theta}_n)(\bar{\theta}_{2j-1} - \hat{\theta}_n)^\top] \\ &= \hat{\Sigma}_b + \hat{R}_{b_n}, \end{aligned}$$

where  $\hat{R}_{b_n} = \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} [(\bar{\theta}_{2j-1} - \hat{\theta}_n)(\bar{\theta}_{2j} - \hat{\theta}_n)^\top + (\bar{\theta}_{2j} - \hat{\theta}_n)(\bar{\theta}_{2j-1} - \hat{\theta}_n)^\top]$ . Therefore the lugsail estimator can be rewritten as

$$\hat{\Sigma}_{L, b_n} = 2\hat{\Sigma}_{2b_n} - \hat{\Sigma}_{b_n} = \hat{\Sigma}_{b_n} + 2\hat{R}_{b_n}.$$

## D.3 Proof of Proposition 5

First we Simplify  $\hat{R}_{b_n}$ . Using  $D_i = \theta_i - \theta^*$  and  $D_i = U_i + \delta_i$ , we have

$$\hat{R}_{b_n} = \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} [(\bar{D}_{2j-1} - \hat{D}_n)(\bar{D}_{2j} - \hat{D}_n)^\top + (\bar{D}_{2j} - \hat{D}_n)(\bar{D}_{2j-1} - \hat{D}_n)^\top]$$



$$\begin{aligned}
 &= \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} [(\bar{U}_{2j-1} - \hat{U}_n)(\bar{U}_{2j} - \hat{U}_n)^\top + (\bar{U}_{2j-1} - \hat{U}_n)(\bar{\delta}_{2j} - \hat{\delta})^\top + (\bar{\delta}_{2j-1} - \hat{\delta})(\bar{U}_{2j} - \hat{U}_n)^\top \\
 &\quad + (\bar{\delta}_{2j-1} - \hat{\delta})(\bar{\delta}_{2j} - \hat{\delta})^\top + (\bar{U}_{2j} - \hat{U}_n)(\bar{U}_{2j-1} - \hat{U}_n)^\top + (\bar{U}_{2j} - \hat{U}_n)(\bar{\delta}_{2j-1} - \hat{\delta})^\top \\
 &\quad + (\bar{\delta}_{2j} - \hat{\delta})(\bar{U}_{2j-1} - \hat{U}_n)^\top + (\bar{\delta}_{2j} - \hat{\delta})(\bar{\delta}_{2j-1} - \hat{\delta})^\top] \\
 &= \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (\bar{U}_{2j-1} - \hat{U}_n)(\bar{U}_{2j} - \hat{U}_n)^\top + \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (\bar{U}_{2j-1} - \hat{U}_n)(\bar{\delta}_{2j} - \hat{\delta})^\top \\
 &\quad + \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (\bar{\delta}_{2j-1} - \hat{\delta})(\bar{U}_{2j} - \hat{U}_n)^\top + \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (\bar{\delta}_{2j-1} - \hat{\delta})(\bar{\delta}_{2j} - \hat{\delta})^\top \\
 &\quad + \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (\bar{U}_{2j} - \hat{U}_n)(\bar{U}_{2j-1} - \hat{U}_n)^\top + \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (\bar{U}_{2j} - \hat{U}_n)(\bar{\delta}_{2j-1} - \hat{\delta})^\top \\
 &\quad + \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (\bar{\delta}_{2j} - \hat{\delta})(\bar{U}_{2j-1} - \hat{U}_n)^\top + \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (\bar{\delta}_{2j} - \hat{\delta})(\bar{\delta}_{2j-1} - \hat{\delta})^\top. \tag{39}
 \end{aligned}$$

Now we prove two additional results needed to establish Proposition 5.

**Lemma 12** *Under the assumptions of Proposition 5,*

$$\mathbb{E}\|W\|^2 \lesssim C_d^{2.5} n^{-\alpha} a_n^{-1/2} + C_d^4 n^{-\alpha} + C_d^2 a_n^{-1}, \text{ where } W = a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} \begin{pmatrix} \tau_{2j-1} \\ \sum_{k=\tau_{2j-2}+1}^{\tau_{2j-1}} \xi_k \end{pmatrix} \begin{pmatrix} \tau_{2j} \\ \sum_{l=\tau_{2j-1}+1}^{\tau_{2j}} \xi_l \end{pmatrix}^\top.$$

**Proof** Let  $\tilde{\xi}_i = -\nabla f(\theta^*, \zeta_i)$ ,  $\hat{\xi}_i = \xi_i - \tilde{\xi}_i$  and

$$\tilde{W} := a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} \begin{pmatrix} \tau_{2j-1} \\ \sum_{k=\tau_{2j-2}+1}^{\tau_{2j-1}} \tilde{\xi}_k \end{pmatrix} \begin{pmatrix} \tau_{2j} \\ \sum_{l=\tau_{2j-1}+1}^{\tau_{2j}} \tilde{\xi}_l \end{pmatrix}^\top.$$

Then,  $\xi_i = \hat{\xi}_i + \tilde{\xi}_i$  where  $\{\tilde{\xi}_i\}$  is a sequence of iid random variables and  $\{\hat{\xi}_i\}$  is a martingale difference sequence. Note that

$$\mathbb{E}\|\tilde{W}\|^2 \leq \mathbb{E} \text{tr}(\tilde{W})^2 = \text{tr} \mathbb{E}(\tilde{W})^2.$$

and

$$\mathbb{E}(\tilde{W}^2) = a_n^{-2} \sum_{j=1}^{a_n/2} \sum_{k=1}^{a_n/2} b_n^{-2} \sum_{i_1=\tau_{2j-2}+1}^{\tau_{2j-1}} \sum_{i_2=\tau_{2j-1}+1}^{\tau_{2j}} \sum_{i_3=\tau_{2k-2}+1}^{\tau_{2k-1}} \sum_{i_4=\tau_{2k-1}+1}^{\tau_{2k}} \mathbb{E}(\tilde{\xi}_{i_1} \tilde{\xi}_{i_2}^\top \tilde{\xi}_{i_3} \tilde{\xi}_{i_4}^\top).$$

Since  $\{\tilde{\xi}_i\}$  is an iid sequence of zero-mean random variables, on the RHS of the above expression, the terms are zero unless  $i_1 = i_3$ ,  $i_2 = i_4$  or  $i_1 = i_4$ ,  $i_2 = i_3$ . Also, this happens only when all the indices are in the same set of consecutive batches, thus

$$\mathbb{E}(\tilde{W}^2) = a_n^{-2} \sum_{j=1}^{a_n/2} b_n^{-2} \sum_{i_1=\tau_{2j-2}+1}^{\tau_{2j-1}} \sum_{i_3=\tau_{2j-1}+1}^{\tau_{2j}} \left[ \mathbb{E}(\tilde{\xi}_{i_1} \tilde{\xi}_{i_3}^\top \tilde{\xi}_{i_1} \tilde{\xi}_{i_3}^\top) + \mathbb{E}(\tilde{\xi}_{i_1} \tilde{\xi}_{i_3}^\top \tilde{\xi}_{i_3} \tilde{\xi}_{i_1}^\top) \right]$$

and

$$\begin{aligned} \text{tr}(\mathbb{E}\widetilde{W}^2) &= a_n^{-2} \sum_{j=1}^{a_n/2} b_n^{-2} \sum_{i_1=\tau_{2j-2}+1}^{\tau_{2j-1}} \sum_{i_3=\tau_{2j-1}+1}^{\tau_{2j}} \left[ \mathbb{E}(\tilde{\xi}_{i_1}^\top \tilde{\xi}_{i_3})^2 + \mathbb{E}(\|\tilde{\xi}_{i_1}\|^2 \|\tilde{\xi}_{i_3}\|^2) \right] \\ &\lesssim a_n^{-2} \sum_{j=1}^{a_n/2} (\text{tr}(S))^2 \lesssim C_d^2 a_n^{-1}. \end{aligned}$$

Next, we denote

$$\widehat{W} = \frac{1}{a_n} \sum_{k=1}^{a_n/2} \frac{1}{b_n} \left( \sum_{i=\tau_{2k-2}+1}^{\tau_{2k-1}} \hat{\xi}_i \right) \left( \sum_{i=\tau_{2k-1}+1}^{\tau_{2k}} \hat{\xi}_i \right)^\top$$

and using steps similar to those in the proof of Lemma 9, we obtain

$$\mathbb{E}(\|W - \widetilde{W}\|^2) \lesssim \sqrt{\mathbb{E}[\text{tr}(\widehat{W}^2)]} \sqrt{\mathbb{E}[\text{tr}(\widetilde{W}^2)]} + \mathbb{E}[\text{tr}(\widehat{W}^2)]. \quad (40)$$

Now, we need to find a bound for  $\mathbb{E}(\text{tr}(\widehat{W}^2))$ . Let  $\bar{\xi}_k = b_n^{-1} \sum_{i=\tau_{k-1}+1}^{\tau_k} \hat{\xi}_i$ . Then using Cauchy-Schwartz inequality, we have

$$\begin{aligned} \mathbb{E}[\text{tr}(\widehat{W}^2)] &= a_n^{-2} \sum_{k=1}^{a_n/2} \sum_{l=1}^{a_n/2} b_n^{-2} \mathbb{E}[\text{tr}(\bar{\xi}_{2k-1} \bar{\xi}_{2k}^\top \bar{\xi}_{2l-1} \bar{\xi}_{2l}^\top)] \\ &\lesssim a_n^{-2} \sum_{k=1}^{a_n/2} \sum_{l=1}^{a_n/2} b_n^{-2} \sqrt{\mathbb{E}\|\bar{\xi}_{2k-1}\|^2} \sqrt{\mathbb{E}\|\bar{\xi}_{2k}\|^2} \sqrt{\mathbb{E}\|\bar{\xi}_{2l-1}\|^2} \sqrt{\mathbb{E}\|\bar{\xi}_{2l}\|^2} \\ &\lesssim a_n^{-2} \sum_{k=1}^{a_n/2} \sum_{l=1}^{a_n/2} b_n^{-2} \sum_{i_1=\tau_{2k-2}+1}^{\tau_{2k-1}} \sqrt{\mathbb{E}\|\hat{\xi}_{i_1}\|^2} \sum_{i_2=\tau_{2k-1}+1}^{\tau_{2k}} \sqrt{\mathbb{E}\|\hat{\xi}_{i_2}\|^2} \sum_{i_3=\tau_{2l-2}+1}^{\tau_{2l-1}} \sqrt{\mathbb{E}\|\hat{\xi}_{i_3}\|^2} \sum_{i_4=\tau_{2l-1}+1}^{\tau_{2l}} \sqrt{\mathbb{E}\|\hat{\xi}_{i_4}\|^2}, \end{aligned}$$

using Theorem 2.1 in Rio (2009).

Using Result 3, we have  $\mathbb{E}\|\hat{\xi}_j\|^2 \lesssim C_d^3 i^{-\alpha}$ . Consequently using Lemma 8 we get

$$\mathbb{E}[\text{tr}(\widehat{W}^2)] \lesssim C_d^3 \left( n^{-1} \sum_{i=1}^n i^{-\alpha/2} \right) \left( n^{-1} \sum_{j=1}^n j^{-\alpha/2} \right) \lesssim C_d^3 n^{-\alpha}.$$

Further substituting in (40) we have

$$\mathbb{E}\|W - \widetilde{W}\|^2 \lesssim C_d^{2.5} n^{-\alpha} a_n^{-1/2} + C_d^4 n^{-\alpha}.$$

Now using  $\mathbb{E}\|W\|^2 \leq \mathbb{E}\|W - \widetilde{W}\|^2 + \mathbb{E}\|\widetilde{W}\|^2$  the proof is complete.  $\blacksquare$

**Lemma 13** *Under the assumptions of Proposition 5,*

$$\mathbb{E} \left\| \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \bar{U}_{2j-1} \bar{U}_{2j}^\top \right\| \lesssim C_d^{1.25} n^{-\alpha/2} a_n^{-1/4} + C_d^2 n^{-\alpha/2} + C_d a_n^{-1/2} + C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2}.$$

**Proof** The  $k^{th}$  batch mean  $\bar{U}_k$  can be written as

$$\bar{U}_k = b_n^{-1} S_{\tau_{k-1}}^{\tau_k} U_{\tau_{k-1}} + b_n^{-1} A^{-1} \sum_{l=\tau_{k-1}+1}^{\tau_k} \xi_l + b_n^{-1} \sum_{l=\tau_{k-1}+1}^{\tau_k} \xi_l Z_l^{\tau_k} + b_n^{-1} \sum_{l=\tau_{k-1}+1}^{\tau_k} \eta_l \xi_l.$$

Let us denote

$$A_k := -A^{-1} \sum_{l=\tau_{k-1}+1}^{\tau_k} \xi_l \text{ and } B_k := S_{\tau_{k-1}}^{\tau_k} U_{\tau_{k-1}} + \sum_{l=\tau_{k-1}+1}^{\tau_k} \xi_l Z_l^{\tau_k} + \sum_{l=\tau_{k-1}+1}^{\tau_k} \eta_l \xi_l.$$

Then,

$$\begin{aligned} \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \bar{U}_{2j-1} \bar{U}_{2j}^\top &= \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (b_n^{-1} A_{2j-1} + b_n^{-1} B_{2j-1}) (b_n^{-1} A_{2j} + b_n^{-1} B_{2j})^\top \\ &= a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} A_{2j-1} A_{2j}^\top + a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} (A_{2j-1} B_{2j}^\top + B_{2j-1} A_{2j}^\top + B_{2j-1} B_{2j}^\top). \end{aligned}$$

Now,

$$a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} A_{2j-1} A_{2j}^\top = A^{-1} W A^{-1}, \text{ where } W = a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} \left( \sum_{k=\tau_{2j-2}+1}^{\tau_{2j-1}} \xi_k \right) \left( \sum_{l=\tau_{2j-1}+1}^{\tau_{2j}} \xi_l \right)^\top,$$

and using Lemma 12 we get

$$\left\| a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} A_{2j-1} A_{2j}^\top \right\|^2 \lesssim C_d^{2.5} n^{-\alpha+(\beta-1)/2} + C_d^4 n^{-\alpha} + C_d^2 n^{\beta-1}.$$

Using Cauchy-Schwartz inequality, we have

$$\left\| a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} A_{2j-1} A_{2j}^\top \right\| \lesssim C_d^{1.25} n^{-\alpha/2+(\beta-1)/4} + C_d^2 n^{-\alpha/2} + C_d n^{(\beta-1)/2}. \quad (41)$$

Next, using Cauchy-Schwartz inequality, we have

$$\left\| a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} B_{2j-1} A_{2j}^\top \right\|^2 \leq \left\| a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} B_{2j-1} \right\|^2 \left\| a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} A_{2j} \right\|^2.$$

Further,  $\|B_k\|^2 = \|B_k B_k^\top\| \leq \text{tr}(B_k B_k^\top)$  and  $\xi_i$  is a martingale, so from Lemma 10, we have

$$\text{tr}(\mathbb{E} B_k B_k^\top) \leq \|S_{\tau_{k-1}}^{\tau_k}\|^2 \mathbb{E} \|U_{\tau_{k-1}}\|^2 + \sum_{l=\tau_{k-1}+1}^{\tau_k} \|Z_l^{\tau_k} + \eta_l I\|^2 \mathbb{E} \|\xi_l\|^2.$$

Now, using steps similar to those in the proof of Lemma 10, we obtain

$$a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} \mathbb{E} \|B_{2j-1} A_{2j}\| \lesssim \sqrt{a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} \mathbb{E} \|B_{2j-1}\|^2} \sqrt{a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} \mathbb{E} \|A_{2j}\|^2} \lesssim C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2}. \quad (42)$$

Thus, using (41) and (42), we obtain

$$\begin{aligned}
& \mathbb{E} \left\| \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \bar{U}_{2j-1} \bar{U}_{2j}^\top \right\| \\
& \lesssim \mathbb{E} \left\| a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} A_{2j-1} A_{2j}^\top \right\| + 2a_n^{-1} \sum_{j=1}^{a_n/2} b_n^{-1} \left( \mathbb{E} \|A_{2j-1} B_{2j}^\top\| + \mathbb{E} \|B_{2j-1} B_{2j}^\top\| \right) \\
& \lesssim C_d^{1.25} n^{-\alpha/2} a_n^{-1/4} + C_d^2 n^{-\alpha/2} + C_d a_n^{-1/2} + C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2}.
\end{aligned}$$

■

Now, we proceed to prove Proposition 5.

**Proof**

Note that

$$\begin{aligned}
\frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (\bar{U}_{2j-1} - \hat{U}_n)(\bar{U}_{2j} - \hat{U}_n)^\top &= \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \bar{U}_{2j-1} \bar{U}_{2j}^\top - \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \bar{U}_{2j-1} \hat{U}_n^\top - \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \hat{U}_n \bar{U}_{2j}^\top \\
&\quad + \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \hat{U}_n \hat{U}_n^\top,
\end{aligned}$$

and we have from the proof of Lemma 11, that  $\mathbb{E} \|b_n \bar{U}_n \bar{U}_n^\top\| \lesssim C_d a_n^{-1}$ . Therefore, using Lemma 13 and Cauchy-Schwartz inequality we have

$$\begin{aligned}
& \left\| \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (\bar{U}_{2j-1} - \hat{U}_n)(\bar{U}_{2j} - \hat{U}_n)^\top \right\| \\
& \leq C_d^{1.25} n^{-\alpha/2} a_n^{-1/4} + C_d^2 n^{-\alpha/2} + C_d a_n^{-1/2} + C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2} + C_d a_n^{-1}.
\end{aligned}$$

Next, from the proof of Theorem 1, we have

$$\begin{aligned}
\mathbb{E} \|\bar{\delta}_k\|^2 &\lesssim \frac{1}{b_n^2} \tau_{k-1}^\alpha + \frac{1}{b_n^2} b_n C_d^4 b_n \tau_{k-1}^{-2\alpha} \lesssim \frac{1}{b_n^2} \tau_{k-1}^\alpha + C_d^4 \tau_{k-1}^{-2\alpha}, \\
\Rightarrow a_n^{-1} \sum_{j=1}^{a_n/2} b_n \mathbb{E} \|\bar{\delta}_{2j-1}\|^2 &\lesssim C_d^4 b_n a_n^{-1} \sum_{k=1}^{a_n} \tau_{k-1}^{-2\alpha} \lesssim C_d^4 b_n^{1-2\alpha} a_n^{-1} \sum_{k=1}^{a_n} k^{-2\alpha} \asymp C_d^4 b_n^{1-2\alpha} a_n^{-2\alpha} = C_d^4 n^{-2\alpha} b_n.
\end{aligned} \tag{43}$$

Furthermore, using Cauchy-Schwartz inequality we have

$$\begin{aligned}
& \mathbb{E} \left\| \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} (\bar{\delta}_{2j} - \hat{\delta})(\bar{U}_{2j-1} - \hat{U}_n)^\top \right\| \leq \frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \mathbb{E} \|(\bar{\delta}_{2j} - \hat{\delta})(\bar{U}_{2j-1} - \hat{U}_n)^\top\| \\
& \leq \sqrt{\frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \mathbb{E} \|(\bar{\delta}_{2j} - \hat{\delta})\|^2} \sqrt{\frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \mathbb{E} \|(\bar{U}_{2j-1} - \hat{U}_n)\|^2}
\end{aligned}$$

$$\begin{aligned}
 &\leq \sqrt{\frac{b_n}{a_n} \sum_{j=1}^a \mathbb{E} \|\bar{\delta}_j - \hat{\delta}\|^2} \sqrt{\frac{b_n}{a_n} \sum_{j=1}^a \mathbb{E} \|\bar{U}_j - \hat{U}_n\|^2} \\
 &\leq \sqrt{\frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \mathbb{E} \|\bar{\delta}_{2j}\|^2} \sqrt{\frac{b_n}{a_n} \sum_{j=1}^{a_n/2} \mathbb{E} \|\bar{U}_{2j-1}\|^2}.
 \end{aligned}$$

Similarly proceeding, we obtain bounds on norm of norms of other terms as well, and consequently, we get

$$\begin{aligned}
 &\mathbb{E} \|\hat{R}_{b_n}\| \\
 &\lesssim 2[C_d^{1.25} n^{-\alpha/2} a_n^{-1/4} + C_d^2 n^{-\alpha/2} + C_d a_n^{-1/2} + C_d b_n^{\alpha-1} + C_d b_n^{-1/2} n^{\alpha/2} + C_d a_n^{-1} + C_d^4 n^{-2\alpha} b_n].
 \end{aligned}$$

■

## Appendix E. Additional data sets

We consider four data sets in Section 6.4. First, consider the Santander customer transaction data set<sup>3</sup>, which contains 200 features on  $2 \times 10^5$  bank transactions. The response is a binary variable indicating whether the transaction is of a certain type. We implement ASGD with  $\eta = .05$  starting the process at the maximum-likelihood estimate of the first 10000 observations. The next 5000 data points were employed in a burn-in, yielding an SGD sequence of length 85000.

The second data set is the covtype data set of Blackard (1998) consisting of 581012 tree observations from few areas of Roosevelt National Forest in Colorado with the target of classifying covtype of trees based on 54 independent variables such as elevation, slope, soil type, distance to nearby landmarks etc. The target variable has seven categories which we dichotomize into two classes based on the sets  $\{1\}$  and  $\{2, 3, 4, 5, 6, 7\}$ . To ensure robustness of gradients, we drop binary predictors with less than 1 % response, set  $\eta = 100$ , burn-in sample size to be 5000 and initial sample size to be 5000.

The third data set is the Spambase data set (Hopkins et al., 1999) which contains 4601 emails as observations with 57 continuous predictors and the target variable is to classify whether the email is spam or not. The data set is taken from UCI repository. The burn-in sample size is 500 and we set  $\eta = 4.5$ .

The final data set is the diabetes health indicators data set is obtained from UCI repository <https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators> where the target variable informs whether the individual is diabetic or not. The aim is to classify diabetic condition based on 21 predictors with the total number of observations 253680. The burn-in sample size is 1000 and we employ  $\eta = 1$ . For all data sets, we divide testing and training equally.

3. [www.kaggle.com/competitions/santander-customer-transaction-prediction/overview](https://www.kaggle.com/competitions/santander-customer-transaction-prediction/overview)

## References

- Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.
- Christos Alexopoulos and David Goldsman. To batch or not to batch? *ACM Transactions on Modeling and Computer Simulation*, 14(1):76–114, 2004.
- Christos Alexopoulos and Andrew F Seila. Implementing the batch means method in simulation experiments. In *Proceedings of the 28th conference on winter simulation*, pages 214–221, 1996.
- Theodore W Anderson. *The Statistical Analysis of Time Series*. John Wiley & Sons, 2011.
- Jock Blackard. Covertime [dataset]. *UCI Machine Learning Repository*, 1998. doi: 10.24432/C50K5N.
- Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010*, pages 177–186. Physica-Verlag/Springer, Heidelberg, 2010.
- Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.
- Saptarshi Chakraborty, Suman K Bhattacharya, and Kshitij Khare. Estimating accuracy of the MCMC variance estimator: Asymptotic normality for batch means estimators. *Statistics & Probability Letters*, 183:109337, 2022.
- Der-Fa R Chen and Andrew F Seila. Multivariate inference in stationary simulation using batch means. In *Proceedings of the 19th conference on Winter simulation*, pages 302–304, 1987.
- Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020. ISSN 0090-5364. doi: 10.1214/18-AOS1801.
- Xi Chen, Zehua Lai, He Li, and Yichen Zhang. Online statistical inference for contextual bandits via stochastic gradient descent. *arXiv preprint arXiv:2212.14883*, 2022.
- Chiahon Chien, David Goldsman, and Benjamin Melamed. Large-sample results for batch means. *Management Science*, 43(9):1288–1295, 1997.
- Halim Damerdj. Mean-square consistency of the variance estimator in steady-state simulation output analysis. *Operations Research*, 43(2):282–291, 1995.
- Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348–1382, 2020.
- Vaclav Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, pages 1327–1332, 1968.

- Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19:Paper No. 78, 21, 2018. ISSN 1532-4435.
- James M. Flegal and Galin L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034–1070, 2010.
- Charles J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.
- Peter W Glynn and Ward Whitt. Estimating the asymptotic variance with batch means. *Operations Research Letters*, 10(8):431–435, 1991.
- Peter W Glynn and Ward Whitt. The asymptotic validity of sequential stopping rules for stochastic simulations. *The Annals of Applied Probability*, 2(1):180–198, 1992.
- Lei Gong and James M Flegal. A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 25(3):684–700, 2016.
- Mark Hopkins, Erik Reeber, George Forman, and Jaap Suermondt. Spambase [dataset]. *UCI Machine Learning Repository*, 1999. doi: 10.24432/C53G6X.
- Galin L. Jones, Murali Haran, Brian S. Caffo, and Ronald Neath. Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101(476):1537–1547, 2006.
- Nicholas M Kiefer and Timothy J Vogelsang. Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation. *Econometrica*, 70:2093–2095, 2002.
- Nicholas M Kiefer and Timothy J Vogelsang. A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, 21:1130–1164, 2005.
- Jakub Konečný, Jie Liu, Peter Richtárik, and Martin Takáč. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):242–255, 2016.
- Man Fung Leung and Kin Wai Chan. Principles of statistical inference in online problems. *arXiv preprint arXiv:2209.05399*, 2024.
- Xiang Li, Jiadong Liang, and Zhihua Zhang. Online statistical inference for nonlinear stochastic approximation with Markovian data. *arXiv preprint arXiv:2302.07690*, 2023.
- Ruiqi Liu, Xi Chen, and Zuofeng Shang. Statistical inference with stochastic gradient methods under  $\phi$ -mixing data. *arXiv preprint arXiv:2302.12717*, 2023.
- Ying Liu and James M Flegal. Weighted batch means estimators in Markov chain Monte Carlo. *Electronic Journal of Statistics*, 12(2):3397–3442, 2018.

- Ying Liu, Dootika Vats, and James M Flegal. Batch size selection for variance estimators in MCMC. *Methodology and Computing in Applied Probability*, 24(1):65–93, 2022.
- Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.
- Marc S Meketon and Bruce Schmeiser. Overlapping batch means: Something for nothing? In *Proceedings of the 16th conference on Winter simulation*, pages 226–230, 1984.
- Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, page 856–864. Curran Associates, Inc., 2011.
- David F Muñoz and Peter W Glynn. A batch means methodology for estimation of a nonlinear function of a steady-state mean. *Management Science*, 43(8):1121–1135, 1997.
- Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2008.
- Dimitris N Politis and Joseph P Romano. Bias-corrected nonparametric spectral estimation. *Journal of Time Series Analysis*, 16(1):67–103, 1995.
- Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. ISSN 0363-0129. doi: 10.1137/0330046.
- Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.
- Emmanuel Rio. Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability*, 22(1):146–163, 2009.
- Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.
- Nathan Robertson, James M. Flegal, Dootika Vats, and Galin L. Jones. Assessing and visualizing simultaneous simulation error. *Journal of Computational and Graphical Statistics*, 30(2):324–334, 2021.
- David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.
- Bruce W Schmeiser and Wheyming Tina Song. Batching methods in simulation output analysis: what we know and what we don’t. In *Proceedings of the 28th conference on Winter simulation*, pages 122–127, 1996.



- Jeffrey S Simonoff. The relative importance of bias and variability in the estimation of the variance of a statistic. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 42:3–7, 1993.
- Wheyming Tina Song and Bruce W Schmeiser. Optimal mean-squared-error batch sizes. *Management Science*, 41(1):110–123, 1995.
- Ziwei Su, Raghu Pasupathy, Yingchieh Yeh, and Peter Glynn. Overlapping batch confidence intervals on statistical functionals constructed from time series: Application to quantiles, optimization, and estimation. *ACM Transactions on Modeling and Computer Simulation*, 34(2):1–43, 2024.
- Panos Toulis and Edoardo M. Airolidi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.
- Dootika Vats and James M Flegal. Lugsail lag windows for estimating time-average covariance matrices. *Biometrika*, 109(3):735–750, 2022.
- Dootika Vats, James M Flegal, and Galin L Jones. Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2):321–337, 2019.
- Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- Jinhan Xie, Enze Shi, Peijun Sang, Zuofeng Shang, Bei Jiang, and Linglong Kong. Scalable inference in functional linear regression with streaming data. *arXiv preprint arXiv:2302.02457*, 2023.
- Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5332–5344. Curran Associates, Inc., 2020.
- Tong 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, 2004.
- Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 0(0):1–12, 2021.
- Yi Zhu and Jing Dong. On constructing confidence region for model parameters in stochastic gradient descent via batch means. In *2021 Winter Simulation Conference (WSC)*, pages 1–12, 2021.