STATISTICAL INFERENCE FOR MODEL PARAMETERS IN STOCHASTIC GRADIENT DESCENT

By Xi Chen[*], Jason D. Lee[†], Xin T. Tong[‡] and Yichen Zhang[*]

New York University[*], University of Southern California[†], and National University of Singapore[‡]

The stochastic gradient descent (SGD) algorithm has been widely used in statistical estimation for large-scale data due to its computational and memory efficiency. While most existing works focus on the convergence of the objective function or the error of the obtained solution, we investigate the problem of statistical inference of true model parameters based on SGD when the population loss function is strongly convex and satisfies certain smoothness conditions.

Our main contributions are two-fold. First, in the fixed dimension setup, we propose two consistent estimators of the asymptotic covariance of the average iterate from SGD: (1) a plug-in estimator, and (2) a batch-means estimator, which is computationally more efficient and only uses the iterates from SGD. Both proposed estimators allow us to construct asymptotically exact confidence intervals and hypothesis tests.

Second, for high-dimensional linear regression, using a variant of the SGD algorithm, we construct a debiased estimator of each regression coefficient that is asymptotically normal. This gives a one-pass algorithm for computing both the sparse regression coefficients and confidence intervals, which is computationally attractive and applicable to online data.

1. Introduction. Estimation of model parameters by minimizing an objective function is a fundamental idea in statistics. Let $x^* \in \mathbb{R}^d$ be the true d-dimensional model parameters. In common models, x^* is the minimizer of a convex objective function $F(x) : \mathbb{R}^d \to \mathbb{R}$, i.e.,

$$(1) \qquad x^* = \operatorname{argmin}\left(F(x) := \mathbb{E}_{\zeta \sim \Pi} f(x, \zeta) = \int f(x, \zeta) \mathrm{d}\Pi(\zeta)\right),$$

where ζ denotes the random sample from a probability distribution Π and $f(x, \zeta)$ is the loss function.

A widely used optimization method for minimizing $F(x)$ is the *stochastic gradient descent (SGD)*, which has a long history in optimization (see, e.g.,

Robbins and Monro (1951); Polyak and Juditsky (1992); Nemirovski et al. (2009)). In particular, let x_0 denote any given starting point. SGD is an iterative algorithm, where the i-th iterate x_i takes the following form,

$$(2) \qquad x_i = x_{i-1} - \eta_i \nabla f(x_{i-1}, \zeta_i).$$

The step size η_i is a decreasing sequence in i, ζ_i is the i-th sample randomly drawn from the distribution Π, and $\nabla f(x_{i-1}, \zeta_i)$ denotes the gradient of $f(x, \zeta_i)$ with respect to x at $x = x_{i-1}$. The algorithm outputs either the last iterate x_n, or the average iterate $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ as the solution to the optimization problem in (1). When $\bar{x}_n$ is adopted as the solution, the algorithm is referred to as the averaged SGD (ASGD), and such an averaging step is known as the Polyak-Ruppert averaging (Ruppert, 1988; Polyak and Juditsky, 1992). SGD has many computational and storage advantages over traditional deterministic optimization methods. First, SGD only uses *one pass* over the data and the per-iteration time complexity of SGD is $O(d)$, which is independent of the sample size. Second, there is no need for SGD to store the dataset, and thus SGD naturally fits in the *online setting*, where each sample arrives sequentially (e.g., search queries or transactional data). Moreover, ASGD is known to achieve the optimal convergence rate in terms of $\mathbb{E}(F(\bar{x}_n) - F(x^*))$ with the rate of $O(1/n)$ (Rakhlin, Shamir and Sridharan, 2012) when $F(x)$ is smooth and strongly convex.It has become the prevailing optimization method for many machine learning tasks (Srebro and Tewari, 2010), e.g., training deep neural networks.

Based on the simple SGD template in (2), there are a large number of variants developed in the optimization and statistical learning literature. Most existing works only focus on the convergence in terms of the objective function or the distance between the obtained solution and the true minimizer x^* of (1). However, the statistical inference (e.g., constructing confidence intervals) for each coordinate of x^* based on SGD has largely remained unexplored. Inference is a core topic in statistics and a confidence interval has been widely used to quantify the uncertainty in the estimation of model parameters. In this paper, we propose computationally efficient methods to conduct the inference for each coordinate of x_j^* for $j = 1, 2, \ldots, d$ based on SGD. With the developed techniques, one can test if $x_j^* = c$ for any number c, and tell a range of values that x_j^* lies within it with a certain probability. These objectives cannot be achieved by deriving deviation inequalities or generalization error bounds (see Section 1.1 for details).

The proposed methods are built on a classical result of ASGD, which characterizes the limiting distribution of $\bar{x}_n$. In particular, let $A = \nabla^2 F(x^*)$

be the Hessian matrix of $F(x)$ at $x = x^*$ and S be the covariance matrix of $\nabla f(x^*, \zeta)$, i.e.,

$$(3) \qquad S = \mathbb{E}\left([\nabla f(x^*, \zeta)][\nabla f(x^*, \zeta)]^T\right).$$

Note that $\mathbb{E}\nabla f(x^*, \zeta) = \nabla F(x^*) = 0$, provided the interchangeability of derivative and expectation. Ruppert (1988) and Polyak and Juditsky (1992) showed that when d is fixed and F is strongly convex with a Lipschitz gradient, by choosing appropriately diminishing step sizes, $\sqrt{n}(\bar{x}_n - x^*)$ converges in distribution to a multivariate normal random vector, i.e.,

$$(4) \qquad \sqrt{n}(\bar{x}_n - x^*) \Rightarrow \mathcal{N}(0, A^{-1}SA^{-1}).$$

However, this asymptotic normality result itself cannot be used to provide confidence intervals. To construct an asymptotically valid confidence interval (or equivalently, an asymptotically valid test that controls the type I error), we need to further construct a consistent estimator of the asymptotic covariance of $\sqrt{n}\bar{x}_n$, i.e., $A^{-1}SA^{-1}$. The standard covariance estimator simply estimates A and S by their sample versions, and replaces the x^* in A and S by $\bar{x}_n$. However, this standard estimator cannot be constructed in an online fashion. In other words, all the data is required to be stored to compute this estimator since $\bar{x}_n$ can only be known when the SGD procedure terminates. This requirement loses the advantage of SGD in terms of data storage.

To address this challenge, we propose two approaches to estimate $A^{-1}SA^{-1}$ without the need of storing the data. The first approach is the *plug-in* estimator. In particular, we propose a thresholding estimator $\widetilde{A}_n$ of A based on the sample estimate $A_n = \frac{1}{n}\sum_{i=1}^{n}\nabla^2 f(x_{i-1}, \zeta_i)$. Note that this is not the standard sample estimate since each term $\nabla^2 f(x_{i-1}, \zeta_i)$ is regarding different SGD iterates x_{i-1} (in contrast to a single $\bar{x}_n$) and thus can be constructed online. This construction facilitates the online computation of A_n, which does not need to store each x_i and ζ_i. Together with the sample estimate S_n of S, the asymptotic covariance $A^{-1}SA^{-1}$ is estimated by $\widetilde{A}_n^{-1}S_n\widetilde{A}_n^{-1}$, which is proven to be a consistent estimator (see Theorem 4.2).

However, the plug-in estimator requires the computation of the Hessian matrix of the loss function f and its inverse, which is usually not available for legacy codes where only the SGD iterates are available. Now a natural question arises: can we estimate the asymptotic covariance *only using the iterates from SGD without requiring additional information*? We provide an affirmative answer to this question by proposing a computationally efficient *batch-means* estimator. Basically, we split the sequence of SGD iterates $\{x_1, x_2, \ldots, x_n\}$ into $M + 1$ batches with batch size $n_0, n_1, \ldots, n_M$. The 0-th batch is discarded since the iterates in that are far from the optimum.

The batch-means estimator is a "weighted" sample covariance matrix that treats each batch-means as a sample.

The idea of batch-means estimator can be traced to Markov Chain Monte Carlo (MCMC), where the batch-means method with equal batch size (see, e.g., Glynn and Iglehart (1990); Glynn and Whitt (1991); Damerdji (1991); Geyer (1992); Fishman (1996); Jones et al. (2006); Flegal and Jones (2010)) is widely used for variance estimation in a time-homogeneous Markov chain. The SGD iterates in (2) indeed form a Markov chain, as x_i only depends on x_{i-1}. However, since the step size sequence η_i is a diminishing sequence, it is a *time-inhomogenous* Markov chain. Moreover, the asymptotic behavior of SGD and MCMC are fundamentally different: while the former converges to the optimum, the latter travels ergodically inside the state space. As a consequence of these important differences, previous literature on batch-means methods is not applicable to our analysis. To address this challenge, our new batch-means method constructs batches of *increasing sizes*. The sizes of batches are chosen to ensure that the correlation decays appropriately among far-apart batches, so that far-apart batch-means can be roughly treated as independent. In Theorem 4.3, we prove that the proposed batch-means method is a consistent estimator of the asymptotic covariance. Further, we believe this new batch-means algorithm with increasing batch sizes is of independent interest since it can be used to estimate the covariance structure of other time-inhomogeneous Markov chains.

As both the plug-in and the batch-means estimator provide asymptotically exact confidence intervals, each of them has its own advantages:

1. The plug-in estimator has a faster convergence rate than the batch-means estimator (see Theorem 4.2 and Corollary 4.5).
2. The plug-in estimator requires the computation of the Hessian matrix of the loss function and its inverse, which can be expensive to obtain for many applications. The batch-means estimator does not require computing any of them. To establish the consistency result, the plug-in estimator requires an additional Lipschitz condition over the Hessian matrix of the loss function (see Assumption 4.1).
3. The plug-in estimator directly computes the entire estimator $\widetilde{A}_n^{-1} S_n \widetilde{A}_n^{-1}$ for the purpose of estimating diagonal elements of $A^{-1} S A^{-1}$. Furthermore, when d is large, storing $\widetilde{A}_n$ and S_n requires $O(d^2)$ bits, which is wasteful since only estimates of the diagonal elements of $A^{-1} S A^{-1}$ are useful for the inference of each x_j^* for $j = 1, 2, \ldots, d$. Meanwhile, the batch-means estimator is able to merely compute and store diagonals.

Practitioners may decide to choose between the plug-in and batch-means

estimators based on their tasks and computing resources. The plug-in estimator has a faster convergence rate, which leads to better performance in practice. However, in some cases when the computation and storage are limited, the batch-means estimator is able to provide an asymptotically exact confidence interval with comparably good performance. Furthermore, the computation of the Hessian matrix in the plug-in estimator is an "intrusive" requirement for SGD (Sullivan, 2015), i.e., it is not available for legacy codes where only the SGD iterates are computed. For example, if one has already obtained SGD iterates and wants to compute confidence intervals afterward, a non-instructive method like batch-means can be directly applied. Such a non-intrusive method that can operate with black-box SGD iterates is more desirable and welcomed by practitioners, as it only uses the existing SGD iterates without the need to change the original SGD code.

For the second part of our contribution, we further study the problem of confidence interval construction for x^* in high-dimensional linear regression based on SGD, where the dimensionality d can be much larger than the sample size n. In a high-dimensional setup, it is natural to solve an ℓ_1-regularized problem, $\min_x F(x) + \lambda\|x\|_1$, where $F(x)$ is defined in (1). A popular approach to solve it is the *proximal stochastic gradient approach* (see, e.g., Ghadimi and Lan (2012) and references therein). However, due to the proximal operator (i.e., the soft-thresholding operator for ℓ_1-regularized problem), the distribution of the average iterate $\bar{x}_n$ no longer converges to a multivariate normal distribution. To address this challenge, we use the recently proposed RADAR algorithm (Agarwal, Negahban and Wainwright, 2012), which is a variant of SGD, together with the debiasing approach (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014). The standard debiasing method relies on solving d convex optimization problems (e.g., node-wise Lasso in van de Geer et al. (2014)) to construct an approximation of the inverse of the design covariance matrix. Each deterministic optimization problem requires a per-iteration complexity $O(nd)$, which is prohibitive when n is large. In contrast, we adopt the stochastic RADAR algorithm to solve these optimization problems, where each problem only requires one pass of the data with the per-iteration complexity $O(d)$. Moreover, since the resulting approximate inverse covariance matrix from the stochastic RADAR is not an exact solution of the corresponding optimization problem, the analysis of van de Geer et al. (2014), which heavily relies on the KKT condition, is no longer applicable. We provide a new analysis to establish the asymptotic normality of the obtained estimator of x^* from the stochastic optimization algorithm.

1.1. *Some related works on SGD.*   There is a large body of literature on stochastic gradient approaches and their applications to statistical learning problems (see, e.g., Zhang (2004); Nesterov and Vial (2008); Xiao (2010); Ghadimi and Lan (2012); Roux, Schmidt and Bach (2012); Agarwal, Negahban and Wainwright (2012); Xiao and Zhang (2014) and references therein). Most works on SGD focus on the convergence rate of the objective function instead of the asymptotic distribution of the obtained solution. Thus, we only review a few closely related works with results on distributions.

Back in 1960s, Fabian (1968) studied the distribution of SGD iterates. However, without averaging, the asymptotic variance is inflated and thus the resulting statistical inference would have a reduced power even if the asymptotic is known. Ruppert (1988); Polyak and Juditsky (1992); Bach and Moulines (2011) studied the averaged SGD (ASGD) and established the asymptotic normality and efficiency of the estimators. However, these works do not discuss the estimation of the asymptotic covariance.

A few works in the SGD literature (e.g., Nesterov and Vial (2008); Nemirovski et al. (2009)) show large deviation results of $\Pr(\|\bar{x}_n - x^*\|_2 > t) \leq C(t)$ by combining the Markov inequality with the expected deviation of $\bar{x}_n$ to x^*. However, we note that large deviation results cannot be used to obtain asymptotically exact confidence intervals, which refer to the exact $1 - q$ coverage as $n \to \infty$. That is, $\Pr(x^* \in \mathrm{CI}_q) \to 1 - q$, where CI_q denotes the confidence interval. Deviation inequalities, which are unable to quantify the exact probability, fail to provide the exact $1 - q$ coverage and will lead to wider confidence intervals. Moreover, note that the ℓ_2 bounds in the SGD literature are generally $O(\sigma\sqrt{\frac{d}{n}})$ (where σ^2 is the variance of the norm of the stochastic gradient) and do not imply a ℓ_∞ bound of size $O(\frac{\sigma}{\sqrt{n}})$, whereas a confidence interval for any single coordinate should be $O(\frac{\sigma}{\sqrt{n}})$ (the $O(\cdot)$ notation here does not depend on d). Therefore, although d is fixed, ℓ_2-norm error bound results still lead to conservative confidence intervals. Instead, we will use the central limit theorem that shows $\lim \Pr(|\bar{x}_{n,j} - x^*| < z_{q/2}\sigma/\sqrt{n}) \to 1 - q$, where $z_{q/2}$ is the $(1-q/2)$-quantile of the standard normal distribution. This allows us to construct an asymptotically exact confidence interval.

We also note that Toulis and Airoldi (2016) established the asymptotic normality for the *averaged implicit SGD* procedure, which is an algorithm different from ASGD. Moreover, this paper does not discuss the estimation of the asymptotic covariance and thus their results cannot be directly used to obtain the confidence intervals.

1.2. *Notations and organization of the paper.*   As a summary of notations, throughout the paper, we use $\|x\|_p$ to denote the vector ℓ_p-norm of x,

$\|x\|_0$ the number of non-zero entries in x, $\|X\|$ the matrix operator norm of X and $\|X\|_\infty$ the element-wise ℓ_∞-norm of X (i.e., $\|X\|_\infty = \max_{i,j} |X_{ij}|$). For a square matrix X, we denote its trace by $\mathrm{tr}(X)$. For a positive semi-definite (PSD) matrix A, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ be its maximum and minimum eigenvalue. For a vector a of length d and any index subset $J \subseteq \{1, \ldots, d\}$, we denote by a_J the sub-vector of a with the elements indexed by J and a_{-J} the sub-vector of a with the elements indexed by $\{1, \ldots, d\} \backslash J$. Similarly, for a $d_1 \times d_2$ matrix X and two index subsets $R \subseteq \{1, \ldots, d_1\}$ and $J \subseteq \{1, \ldots, d_2\}$, we denote by $X_{R,J}$ the $|R| \times |J|$ sub-matrix of X with elements in rows in R and columns in J. When $R = \{1, \ldots, d_1\}$ or $J = \{1, \ldots, d_2\}$, we denote $X_{R,J}$ by $X_{\cdot,J}$ or $X_{R,\cdot}$, respectively. We use I to denote the identity matrix. The function $\Phi(\cdot)$ denotes the CDF of the standard normal distribution.

For any sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \gtrsim b_n$ if $a_n \geq c b_n$ holds for all n large enough and some constant $c > 0$, $a_n \lesssim b_n$ if $b_n \gtrsim a_n$ holds, and $a_n \asymp b_n$ if $a_n \gtrsim b_n$ and $a_n \lesssim b_n$.

The rest of the paper is organized as follows. In Section 2, we provide more background of SGD and detailed results from Polyak and Juditsky (1992). In Section 3, we provide the assumptions and some error bounds on SGD iterates. In Section 4, we propose the plug-in estimator and batch-means estimator for estimating the asymptotic covariance of $\bar{x}_n$ from ASGD. In Section 5, we discuss how to conduct inference for high-dimensional linear regression. In Section 6, we demonstrate the proposed methods by simulated experiments. Further discussions appear in Section 7 and all proofs are given in the Appendix.

2. Background. In the classical work of Polyak and Juditsky (1992), the SGD method was introduced in a form equivalent with (2) to facilitate the analysis. In particular, the iteration is given by

$$(5) \qquad x_n = x_{n-1} - \eta_n \nabla F(x_{n-1}) + \eta_n \xi_n,$$

where $\xi_n := \nabla F(x_{n-1}) - \nabla f(x_{n-1}, \zeta_n)$. The formulation (5) decomposes the descent into two parts: $\nabla F(x_{n-1})$ represents the direction of population gradient which is the major driving force behind the convergence of SGD, and ξ_n is a martingale difference sequence under Assumption 3.2 (see below). That is, $\mathbb{E}_{n-1}[\xi_n] = \nabla F(x_{n-1}) - \mathbb{E}_{n-1} \nabla f(x_{n-1}, \zeta_n) = 0$. Here and in the sequel, $\mathbb{E}_n(\cdot)$ denotes the conditional expectation $\mathbb{E}(\cdot | \mathcal{F}_n)$, where $\mathcal{F}_n$ is the σ-algebra generated by $\{\zeta_1, \ldots, \zeta_n\}$ (ζ_k is the k-th sample). Let $\Delta_n := x_n - x^*$ be the error of the n-th iterate. It it noteworthy that by subtracting x^* from both sides of (5), the recursion (5) is equivalent to

$$(6) \qquad \Delta_n = \Delta_{n-1} - \eta_n \nabla F(x_{n-1}) + \eta_n \xi_n,$$

which will be extensively used throughout the paper.

Given the SGD recursion in the form of (6) and under suitable assumptions (see Section 3 below), Theorem 2 of Polyak and Juditsky (1992) shows that when the step size sequence $\eta_i = \eta i^{-\alpha}, i = 1, 2, \ldots, n$ with $\alpha \in (1/2, 1)$, we have

$$(7) \qquad \sqrt{n} \cdot \overline{\Delta}_n \Rightarrow \mathcal{N}(0, A^{-1}SA^{-1}), \quad \text{if} \quad \alpha \in (\tfrac{1}{2}, 1)$$

where $\overline{\Delta}_n = \frac{1}{n} \sum_{i=1}^n \Delta_i = \bar{x}_n - x^*$. Based on this limiting distribution result, we only need to estimate the asymptotic covariance matrix $A^{-1}SA^{-1}$. Then we can form the confidence interval $\bar{x}_{n,j} \pm z_{q/2}\hat{\sigma}_{jj}$, where $\hat{\sigma}_{jj}$ is a consistent estimator of $(A^{-1}SA^{-1})_{jj}$ and $z_{q/2}$ is the $(1-q/2)$-quantile of the standard normal distribution (i.e., $z_{q/2} = \Phi^{-1}(1 - q/2)$ and $\Phi(\cdot)$ is the CDF of the standard normal distribution). Therefore, the main purpose of the paper is to provide consistent estimators of the asymptotic covariance matrix.

REMARK 2.1. *In the model well-specified case, $\overline{x}_n$ is an asymptotically efficient estimator of the true model parameter x^* according to (7). In particular, suppose ζ comes from the probability distribution Π with density $p_{x^*}(\zeta)$ parameterized by x^*. If the loss function $f(x, \zeta) = -\log p_x(\zeta)$ is the negative log-likelihood, under certain regularity conditions, one can show that*

$$A = \nabla^2 \mathbb{E}\left[-\log p_{x^*}(\zeta)\right] = \mathbb{E}\left(-\nabla \log p_{x^*}(\zeta)\right)\left(-\nabla \log p_{x^*}(\zeta)\right)^T = S = I(x^*)$$

Here $I = I(x^*)$ is the Fisher information matrix. Therefore, the limiting covariance matrix $A^{-1}SA^{-1} = I^{-1}$ achieves the Cramér-Rao lower bound, which indicates that $\bar{x}_n$ is asymptotically efficient. It is worth noting that the asymptotic normality result (7) does not require that the model is well-specified. In a model mis-specified case, the asymptotic distribution of $\bar{x}_n$ is centered at x^*, where x^* is the unique minimizer of $F(x)$ and the asymptotic covariance $A^{-1}SA^{-1}$ is of the so-called "sandwich covariance" form (e.g., see Buja et al. (2013)).

To illustrate this SGD recursion in (6) and the form of A and S, we consider the following two motivating examples.

EXAMPLE 2.1 (Linear Regression). *Under the classical linear regression setup, let the n-th sample be $\zeta_n = (a_n, b_n)$, where the input $a_n \in \mathbb{R}^d$ is a sequence of random vectors independently drawn from the same multivariate distribution and the response $b_n \in \mathbb{R}$ follows a linear model, $b_n = a_n^T x^* + \varepsilon_n$. Here $x^* \in \mathbb{R}^d$ represents the true parameters of the linear model, and $\{\varepsilon_n\}$ are independently and identically distributed (i.i.d.) centered random*

variables, which are uncorrelated with a_n. For simplicity, we assume a_n and ε_n have all moments being finite. Given $\zeta_n = (a_n, b_n)$, the loss function at x is a quadratic one:

$$f(x, \zeta_n) = \frac{1}{2}(a_n^T x - b_n)^2.$$

and the true parameters $x^* = \operatorname{argmin}_x (F(x) := \mathbb{E}f(x, \zeta))$. Given the loss function, the SGD iterates in (2) become, $x_n = x_{n-1} - \eta_n a_n(a_n^T x_{n-1} - b_n)$. This can also be written in the form of (5) as

$$x_n = x_{n-1} - \eta_n A \Delta_{n-1} + \eta_n \xi_n, \quad \xi_n := (A - a_n a_n^T)\Delta_{n-1} + a_n \varepsilon_n,$$

where $A = \mathbb{E}a_n a_n^T$ is the population gram matrix of a_n. It is easy to find that

$$F(x) = \frac{1}{2}(x - x^*)^T A(x - x^*) + \mathbb{E}\varepsilon^2,$$

which implies that $\nabla F(x) = A(x - x^*)$ and $\nabla^2 F(x) = A$ for all x. As for matrix S, it is given by $S := \mathbb{E}\left([\nabla f(x^*, \zeta)][\nabla f(x^*, \zeta)]^T\right) = \mathbb{E}\varepsilon_n^2 a_n a_n^T$.

EXAMPLE 2.2 (Logistic Regression). *One of the most popular applications for general loss in statistics is the logistic regression for binary classification problems. In particular, the logistic model assumes that the binary response $b_n \in \{-1, 1\}$ is generated by the following probabilistic model,*

$$\Pr(b_n | a_n) = \frac{1}{1 + \exp\left(-b_n \langle a_n, x^* \rangle\right)},$$

where a_n is an i.i.d. sequence. The population objective function is given by $F(x) = \mathbb{E}f(x, \zeta_n) = \mathbb{E}\log(1 + \exp\left(-b_n \langle a_n, x \rangle\right))$. Let $\varphi(x) := \frac{1}{1 + \exp(-x)}$ denote the sigmoid function, we have $\nabla f(x, \zeta_n) = -\varphi(-b_n \langle a_n, x \rangle)b_n a_n$. Moreover, we have the formulation of matrix A and S as

$$(8) \qquad A = S = \mathbb{E}\frac{a_n a_n^T}{[1 + \exp(\langle a_n, x^* \rangle)][1 + \exp(-\langle a_n, x^* \rangle)]}.$$

3. Assumptions and Error Bounds. In this section, we provide the assumptions used in the fixed-dimensional case and then provide some useful error bounds on Δ_n. We first make the following standard assumption on the population loss function $F(x)$.

ASSUMPTION 3.1 (Strong convexity and Lipschitz continuity of the gradient). *Assume that the objective function $F(x)$ is continuously differentiable and strongly convex with parameter $\mu > 0$, that is, for any x_1 and x_2,*

$$F(x_2) \geq F(x_1) + \langle \nabla F(x_1), x_2 - x_1 \rangle + \frac{\mu}{2}\|x_1 - x_2\|_2^2.$$

Further, assume that $\nabla^2 F(x^*)$ exists, and $\nabla F(x)$ is Lipschitz continuous with a constant L_F, i.e., for any x_1 and x_2, $\|\nabla F(x_1) - \nabla F(x_2)\|_2 \leq L_F \|x_1 - x_2\|_2$.

Note that the strong convexity of $F(x)$ was adopted by Polyak and Juditsky (1992) (see Assumption 4.1 in Polyak and Juditsky (1992)) to derive the limiting distribution of averaged SGD, which serves as the basis of our work. In fact, the strong convexity of $F(x)$ implies $\lambda_{\min}(A) = \lambda_{\min}(\nabla^2 F(x^*)) \geq \mu$ is an important condition for parameter estimation and inference. There are recent works in optimization on relaxing the strong convexity assumption (e.g., Bach and Moulines (2013)), but they were only able to obtain fast convergence rates in terms of the objective value $F(\bar{x}_n) - F(x^*)$.

We further assume that the martingale difference ξ_n satisfies the following conditions.

ASSUMPTION 3.2.   *The following hold for the sequence $\xi_n = \nabla F(x_{n-1}) - \nabla f(x_{n-1}, \zeta_n)$:*

1. *Assume that $f(x, \zeta)$ is continuously differentiable in x for any ζ and $\|\nabla f(x, \zeta)\|_2$ is uniformly integrable for any x so that $\mathbb{E}_{n-1} \xi_n = 0$.*
2. *The conditional covariance of ξ_n has an expansion around $x = x^*$: $\mathbb{E}_{n-1} \xi_n \xi_n^T = S + \Sigma(\Delta_{n-1})$, and there exists constants Σ_1 and $\Sigma_2 > 0$ such that for any $\Delta \in \mathbb{R}^d$.*

   $$\|\Sigma(\Delta)\| \leq \Sigma_1 \|\Delta\|_2 + \Sigma_2 \|\Delta\|_2^2, \quad |\mathrm{tr}(\Sigma(\Delta))| \leq \Sigma_1 \|\Delta\|_2 + \Sigma_2 \|\Delta\|_2^2.$$

   Note that S is the covariance matrix of $\nabla f(x^*, \zeta)$ defined in (3).
3. *There exists constants Σ_3, Σ_4 such that the fourth conditional moment of ξ_n is bounded by $\mathbb{E}_{n-1} \|\xi_n\|_2^4 \leq \Sigma_3 + \Sigma_4 \|\Delta_{n-1}\|_2^4$.*

For part 1, we note that our assumption on $f(x, \zeta)$ guarantees that Leibniz's integration rule holds, i.e., $\mathbb{E}_{\zeta \sim \Pi} \nabla f(x, \zeta) = \nabla F(x)$ for all x. Therefore, we have $\mathbb{E}_{n-1} \xi_n = 0$, which implies that ξ_n is a martingale difference sequence. Assumption 3.2 is a mild condition over the regularity and boundedness of the loss function. In fact, one can easily verify Assumption 3.2 using the following lemma.

LEMMA 3.1.   *If there is a function $H(\zeta)$ with bounded fourth moment, such that the Hessian of $f(x, \zeta)$ is bounded by*

$$\left\|\nabla^2 f(x, \zeta)\right\| \leq H(\zeta)$$

for all x, and $\nabla f(x^*, \zeta)$ have a bounded fourth moment, then Assumption 3.2 holds with $\Sigma_1 = 2\sqrt{\mathbb{E}\|\nabla f(x^*, \zeta)\|_2^2 \mathbb{E}H(\zeta)^2}$, $\Sigma_2 = 4\mathbb{E}H(\zeta)^2$, $\Sigma_3 = 8\mathbb{E}\|\nabla f(x^*, \zeta)\|_2^4$ and $\Sigma_4 = 64\mathbb{E}H(\zeta)^4$.

Although we consider the fixed-dimensional case, it is still of practical interest to investigate the dimension dependence in our results. The dimension dependence is rather complicated since our results involve a number of constants in Assumption 3.1 and 3.2 that all depend on the dimension d (e.g., L_F, Σ_1, Σ_2, Σ_3, Σ_4, $\text{tr}(S)$). For example, $\text{tr}(S)$ grows with d. Moreover, the way it grows depends on how S is configured. Therefore, for the ease of presentation, we define the following quantity

$$(9) \qquad C_d := \max\left\{ L_F, \Sigma_1^{\frac{2}{3}}, \sqrt{\Sigma_2}, \sqrt{\Sigma_3}, \Sigma_4^{\frac{1}{4}}, \text{tr}(S) \right\}.$$

In both linear and logistic regression, C_d increases linearly in d (see Appendix A). We will state our results in terms of this single quantity C_d. We also assume $\|x_0 - x^*\|_2^2 = O(C_d)$, and there is a universal constant c such that the step size satisfies $\eta_i C_d \leq c\mu$ for all i. Note that the choice of step sizes does not sacrifice much of generality since when d is a constant, we could always ignore the first a few iterations, which is usually considered as the "burn in" stage. Also, for the starting point x_0, if all the components of $x_0 - x^*$ are bounded by a constant, it naturally satisfies $\|x_0 - x^*\|_2^2 = O(d)$.

In the sequel, we will impose Assumptions 3.1 and 3.2. In Appendix A, we show that Assumptions 3.1 and 3.2 hold on our motivating examples of linear and logistic regression (see Examples 2.1 and 2.2). Under these assumptions, the classical works (Ruppert, 1988; Polyak and Juditsky, 1992) establish the asymptotic normality and efficiency of the $\bar{x}_n$ (see (4) and Remark 2.1). Moreover, we could obtain the following error bounds on the SGD iterates.

LEMMA 3.2. *Under Assumptions 3.1 and 3.2, if the step size is chosen to be $\eta_n = \eta n^{-\alpha}$ with $\alpha \in (0, 1)$, the iterates of error $\Delta_n = x_n - x^*$ satisfy the following.*

$$\mathbb{E}\|\Delta_n\|_2^k \lesssim n^{-k\alpha/2}(C_d^{k/2} + \|\Delta_{n_0}\|_2^k), \quad k = 1, 2, 4.$$

The proof of Lemma 3.2 is provided in Appendix B.2. A result similar to Lemma 3.2 providing the convergence of $\|\Delta_n\|_2$ and $\|\Delta_n\|_2^2$ has been shown in Bach and Moulines (2011) (see Theorem 1 therein). Here, we provide simpler bounds on conditional moments of Δ_n and extend the results in Bach and Moulines (2011) to the fourth moment bound, since we need to access the variance of a variance estimator. This result also tells us how the

error decorrelates in terms of the number of iterations. Our proof strategy is similar to Bach and Moulines (2011) in that we setup up a recursive formula for the Δ_n term, and then show it decays at a certain rate by leveraging the convexity of $F(x)$.

4. Estimators for Asymptotic Covariance. Following the inference procedures illustrated above, when d is fixed and $n \to \infty$, it is essential to estimate the asymptotic covariance matrix $A^{-1}SA^{-1}$. In this section, we will propose two consistent estimators, the plug-in estimator and the batch-means estimator.

4.1. *Plug-in estimator.* The idea of the plug-in estimator is to separately estimate A and S by some $\widehat{A}$ and $\widehat{S}$ and use $\widehat{A}^{-1}\widehat{S}\widehat{A}^{-1}$ as an estimator of $A^{-1}SA^{-1}$. Since x_i converges to x^*, according to the definitions of A and S in (3), an intuitive way to construct $\widehat{A}$ and $\widehat{S}$ is to use the sample estimate

$$A_n := \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(x_{i-1}, \zeta_i), \qquad S_n := \frac{1}{n}\sum_{i=1}^{n} \nabla f(x_{i-1}, \zeta_i)\nabla f(x_{i-1}, \zeta_i)^T,$$

as long as the information of $\nabla^2 f(x_{i-1}, \zeta_i)$ is available. It is worthwhile noting that each summand in A_n and S_n involves different x_{i-1}. Therefore, A_n and S_n can be computed in an online fashion without the need of storing all the data.

Since we are interested in estimating A^{-1}, it is necessary to avoid the possible singularity of A_n from statistical randomness. Therefore, we propose to use thresholding estimator $\widetilde{A}_n$, which is strictly positive definite. In particular, fix $\delta > 0$, and let $\Psi D_n \Psi^T$ be the eigenvalue decomposition of A_n, where D_n is a non-negative diagonal matrix. We construct the thresholding estimator $\widetilde{A}_n$:

$$\widetilde{A}_n = \Psi \widetilde{D}_n \Psi^T, \quad \left(\widetilde{D}_n\right)_{i,i} = \max\left\{\delta, (D_n)_{i,i}\right\}.$$

By construction, it is guaranteed that $\widetilde{A}_n$ is invertible. With the construction of S_n and $\widetilde{A}_n$ in place, we propose the *plug-in estimator* as $\widetilde{A}_n^{-1}S_n\widetilde{A}_n^{-1}$. Our goal is to establish the consistency of the plug-in estimator, i.e.,

$$\mathbb{E} \left\|\widetilde{A}_n^{-1}S_n\widetilde{A}_n^{-1} - A^{-1}SA^{-1}\right\| \longrightarrow 0 \quad \text{as } n \to \infty.$$

Since this estimator relies on the Hessian matrix of the loss function, we need an additional assumption to establish the consistency.

ASSUMPTION 4.1.   *There are constants L_2 and L_4 such that for all x,*

$$(10) \qquad \mathbb{E}\|\nabla^2 f(x,\zeta) - \nabla^2 f(x^*,\zeta)\| \le L_2\|x - x^*\|_2,$$
$$\|\mathbb{E}[\nabla^2 f(x^*,\zeta)]^2 - A^2\| \le L_4.$$

Moreover, we assume that for the choice of δ, we have $\lambda_{\min}(A) > \delta$.

We note that it is easy to verify that (10) holds for the two motivating examples in Section 2. For quadratic loss, the Hessian matrix at any x is A itself, and (51) gives the Hessian for the logistic loss, which is Lipschitz in x and also bounded. In addition, according to Assumption 3.1, we have $\lambda_{\min}(\nabla^2 F(x)) \ge \mu$ for any x and thus $\lambda_{\min}(A) \ge \mu$. Therefore, a valid choice of δ satisfying Assumption 4.1 always exists.

To track the dependence of our results on dimension, we assume L_2 and L_4 are also controlled by C_d in (9) as $L_2 \lesssim C_d^{3/2}, L_4 \lesssim C_d^2$. Lemmas A.1 and A.2 in the Appendix verify this requirement is satisfied in linear and logistic regression.

With this additional assumption, we first establish the consistency of the sample estimate A_n and S_n in the following lemma.

LEMMA 4.1.   *Under Assumptions 3.1, 3.2 and 4.1, the followings hold*

$$\mathbb{E}\|A_n - A\| \lesssim C_d^2 n^{-\frac{\alpha}{2}}, \quad \mathbb{E}\|S_n - S\| \lesssim C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha},$$

where $\alpha \in (0,1)$ is given in the step size sequence $\eta_i = \eta i^{-\alpha}$, $i = 1, 2, \ldots, n$.

The proof of Lemma 4.1 is provided in Appendix C.1. Using Lemma 4.1 and a matrix perturbation inequality for the inverse of a matrix (see Lemma C.1 in Appendix C.2), we obtain the consistency result of the proposed plug-in estimator $\widetilde{A}_n^{-1} S_n \widetilde{A}_n^{-1}$:

THEOREM 4.2 (Error rate of the plug-in estimator).   *Under Assumptions 3.1, 3.2 and 4.1, the thresholded plug-in estimator initialized from any bounded x_0 converges to the asymptotic covariance matrix,*

$$(11) \qquad \mathbb{E}\left\|\widetilde{A}_n^{-1} S_n \widetilde{A}_n^{-1} - A^{-1} S A^{-1}\right\| \lesssim \|S\|(C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha}),$$

where $\alpha \in (0,1)$ is given in the step size sequence $\eta_i = \eta i^{-\alpha}$, $i = 1, 2, \ldots, n$. When C_d is a constant, the right hand side of (11) is dominated by $O(n^{-\frac{\alpha}{2}})$.

REMARK 4.2.   *In practice, we usually do not need to perform the thresholding step, since A_n is positive definite with high probability as A_n is close*

to A. The thresholding step is mainly for obtaining the expected error bound in Theorem 4.2. In fact, without the thresholding step, we are still able to the obtain the following error bound. In our numerical experiments, we do not apply the thresholding procedure and the obtained A_n's are always invertible.

COROLLARY 4.3. *Under Assumptions 3.1, 3.2 and 4.1, as $n \to \infty$,*

$$\|A_n^{-1} S_n A_n^{-1} - A^{-1} S A^{-1}\| = O_p\big(\|S\|(C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha})\big).$$

We also note that since the element-wise ℓ_∞-norm is bounded from above by the matrix operator norm, we have $\mathbb{E} \max_{ij} \big|(\widetilde{A}_n^{-1} S_n \widetilde{A}_n^{-1} - A^{-1} S A^{-1})_{ij}\big|$ converges to zero as $n \to \infty$ according Theorem 4.2. Therefore $(A^{-1} S A^{-1})_{jj}^{1/2}$ can be estimated by $\hat{\sigma}_{n,j}^P = (\widetilde{A}_n^{-1} S_n \widetilde{A}_n^{-1})_{jj}^{1/2}$ for the construction of confidence intervals. In particular, we have the following corollary, which shows that $\bar{x}_{n,j} \pm z_{q/2} \hat{\sigma}_{n,j}^P$ is an asymptotic exact confidence interval.

COROLLARY 4.4. *Under the assumptions of Theorem 4.2, if the step size is chosen to be $\eta_i = \eta i^{-\alpha}$ with $\alpha \in (\frac{1}{2}, 1)$, when d is fixed and $n \to \infty$,*

$$\Pr\left(\bar{x}_{n,j} - z_{q/2} \hat{\sigma}_{n,j}^P \leq x_j^* \leq \bar{x}_{n,j} + z_{q/2} \hat{\sigma}_{n,j}^P\right) \to 1 - q.$$

Proof of Corollary 4.4 is given in Appendix D.5. Note that while Theorem 4.2 holds for all $\alpha \in (0, 1)$, the asymptotic normality in (7) holds only when $\alpha \in (\frac{1}{2}, 1)$. Thus, Corollary 4.4 requires that $\alpha \in (\frac{1}{2}, 1)$.

4.2. *Batch-means estimator.* Although the plug-in estimator is intuitive, it requires the computation of the Hessian matrix and its inverse, as well as an additional Assumption 4.1 on the Lipschitz condition of the Hessian matrix. In this section, we develop the batch-means estimator, which only uses the iterates from SGD without requiring computation of any additional quantities. Intuitively, if all iterates are independent and share the same distribution, the asymptotic covariance can be directly estimated by the sample covariance, $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$. Unfortunately, the SGD iterates are far from independent. To understand the correlation between two consecutive iterates, we note that for sufficiently large n such that x_{n-1} is close to x^*, by the Taylor expansion of $\nabla F(x_{n-1})$ at x^*, we have $\nabla F(x_{n-1}) \approx \nabla F(x^*) + \nabla^2 F(x^*)(x_{n-1} - x^*) = A\Delta_{n-1}$, where $\nabla F(x^*) = 0$ by the first order condition and $A = \nabla^2 F(x^*)$. Combining this with the recursion in (6), we have for sufficiently large n,

$$(12) \qquad\qquad \Delta_n \approx (I_d - \eta_n A)\Delta_{n-1} + \eta_n \xi_n.$$

Based on (12), the strength of correlation between Δ_n and Δ_{n-1} can be approximated by $\|I_d - \eta_n A\|$, which is very close to 1 as $\eta_n \asymp n^{-\alpha}$. To address the challenge of strong correlation among neighboring iterates, we split the entire sequence of iterates into batches with carefully chosen batch sizes. In particular, we split n iterates of SGD $\{x_1, \ldots, x_n\}$ into $M + 1$ batches with sizes $n_0, n_1, \ldots, n_M$:

$$\underbrace{\{x_{s_0}, \ldots, x_{e_0}\}}_{\text{0-th batch}}, \ \underbrace{\{x_{s_1}, \ldots x_{e_1}\}}_{\text{1-st batch}}, \ \ldots, \ \underbrace{\{x_{s_M}, \ldots, x_{e_M}\}}_{M\text{-th batch}}.$$

Here s_k and e_k are the starting and ending index of k-th batch with $s_0 = 1$, $s_k = e_{k-1} + 1$, $n_k = e_k - s_k + 1$, and $e_M = n$. We treat the 0-th batch as the "burn-in stage". More precisely, the iterates $\{x_{s_0}, \ldots, x_{e_0}\}$ will not be used for constructing the batch-means estimator since the step sizes are not small enough and the corresponding iterates in the 0-th batch are far away from the optimum. The batch-means estimator is given by the following:

$$\text{(13)} \qquad \frac{1}{M} \sum_{k=1}^{M} n_k (\overline{X}_{n_k} - \overline{X}_M)(\overline{X}_{n_k} - \overline{X}_M)^T.$$

where $\overline{X}_{n_k} := \frac{1}{n_k} \sum_{i=s_k}^{e_k} x_i$ is the mean of the iterates for the k-th batch and $\overline{X}_M := \frac{1}{e_M - e_0} \sum_{i=s_1}^{e_M} x_i$ is the the mean of all the iterates except for the 0-th batch.

Note that when batch sizes n_k are predetermined, we may rewrite (13) in the following form,

$$\text{(14)} \qquad \frac{1}{M} \sum_{k=1}^{M} n_k \overline{X}_{n_k} \overline{X}_{n_k}^T + \frac{n}{M} \overline{X}_M \overline{X}_M^T - 2 \left(\frac{1}{M} \sum_{k=1}^{M} n_k \overline{X}_{n_k} \right) \overline{X}_M^T.$$

Here, $\overline{X}_M$, $\frac{1}{M} \sum_{k=1}^{M} n_k \overline{X}_{n_k} \overline{X}_{n_k}$ and $\frac{1}{M} \sum_{k=1}^{M} n_k \overline{X}_{n_k}$ can be updated recursively so that there is no need to store all the batch-means $\{\overline{X}_{n_k}\}$. In other words, the memory requirement for the batch-means estimator is only $O(d^2)$ instead of $O(Md^2)$.

Intuitively, the reason why our batch-means estimator with increasing batch size can overcome the strong dependence between iterates is as follows. Although the correlation between neighboring iterates is strong, it decays exponentially for far-apart iterates. Roughly speaking, by (12), for large j and k, the strength of correlation between Δ_j and Δ_k is approximately

$$\text{(15)} \qquad \prod_{i=j}^{k-1} \|I_d - \eta_{i+1} A\| \approx \exp\left(-\lambda_{\min}(A) \sum_{i=j}^{k-1} \eta_{i+1}\right).$$

Therefore, the correlations between the batch-means $\overline{X}_{n_k}$ are close to zero if the batch sizes are large enough, in which case different batch-means can be roughly treated as independent. As a consequence, the sample covariance gathered from the batch-means will serve as a good estimator of the true asymptotic covariance.

The remaining difficulty is how to determine the batch sizes. The approximation of correlation given by (15) provides us a clear clue. If we want the correlation between two neighboring batches to be on the order of $\exp(-cN)$, where N (with $N \to \infty$) is a parameter controlling the amount of decorrelation and c is a constant, we need $\sum_{i=s_k}^{e_k} \eta_i \asymp N$ for every batch k. When $\eta_i = \eta i^{-\alpha}$, $\sum_{i=s_k}^{e_k} \eta_i \asymp (e_k^{1-\alpha} - e_{k-1}^{1-\alpha})$, which leads to the following batch size setting:

$$(16) \qquad e_k = ((k+1)N)^{\frac{1}{1-\alpha}} \quad k = 0, \dots, M,$$

where e_k is the ending point for the k-th batch. There are two practical scenarios to apply the proposed batch-means estimator,

- Total number of iterates n is given: Noting that $e_M = n$, the decorrelation strength factor N takes the following form,

$$(17) \qquad N = \frac{n^{1-\alpha}}{M+1},$$

  where M is the number of batches. Based on the result of Theorem 4.3 below, it is preferable to take $N = n^{\frac{1-\alpha}{2}}$.
- When n is unknown (but sufficiently large): Given a target error bound ϵ, we pick an $N \asymp \epsilon^{-2}$. Then, we receive the online data and batch the SGD iterates according to (16). When the number of batches M is sufficiently large (e.g., the upper bound in (18) below is smaller than ϵ), we stop our SGD procedure and output the batch-means estimator.

Under this setting, the batch-means covariance estimator (13) is consistent as shown in the following theorem.

THEOREM 4.3 (Error rate of the batch-means estimator).   *Under Assumptions 3.1 and 3.2, when d is fixed and the step size is chosen to be $\eta_i = \eta i^{-\alpha}$ with $\alpha \in (\frac{1}{2}, 1)$, the batch-means estimator initialized by any bounded x_0 is a consistent estimator. In particular, for sufficiently large N and M, we have*

$$\mathbb{E} \left\| M^{-1} \sum_{k=1}^{M} n_k (\overline{X}_{n_k} - \overline{X}_M)(\overline{X}_{n_k} - \overline{X}_M)^T - A^{-1}SA^{-1} \right\|$$

$$(18) \lesssim C_d M^{-\frac{1}{2}} + C_d N^{-\frac{1}{2}} + C_d^{\frac{3}{2}}(MN)^{-\frac{\alpha}{4-4\alpha}} + C_d^2 M^{-1} + C_d^3 M^{-1} N^{\frac{1-2\alpha}{1-\alpha}}.$$

As $n \to \infty$, by (17), we can choose $M, N \to \infty$ and thus the right hand side of (18) will converge to zero for any $\alpha \in (1/2, 1)$, which shows the consistency of the proposed covariance estimator. When d is fixed, C_d is a constant, and it is straightforward to see that the right hand side of (18) is dominated by $C_d(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})$. Therefore, according to (17) (i.e., $N(M + 1) = n^{1-\alpha}$), we have the following Corollary 4.5 that suggests the optimal order of M.

COROLLARY 4.5.   *Under Assumptions 3.1 and 3.2, when d is fixed and n is sufficiently large, by choosing the step size $\eta_i = \eta i^{-\alpha}$ with $\alpha \in (\frac{1}{2}, 1)$, $M \asymp n^{\frac{1-\alpha}{2}}$, and $N \asymp n^{\frac{1-\alpha}{2}}$, we have*

$$\mathbb{E} \left\| M^{-1} \sum_{k=1}^{M} n_k (\overline{X}_{n_k} - \overline{X}_M)(\overline{X}_{n_k} - \overline{X}_M)^T - A^{-1} S A^{-1} \right\|$$

(19) $$\lesssim C_d n^{-\frac{1-\alpha}{4}} + C_d^{\frac{3}{2}} n^{-\frac{\alpha}{4}} + C_d^2 n^{-\frac{1-\alpha}{2}} + C_d^3 n^{-\alpha}.$$

When C_d is a constant, the right hand side of (19) is dominated by $O(n^{-\frac{1-\alpha}{4}})$.

As we will show in simulations in Section 6, wide choices between $M = n^{0.2}$ to $M = n^{0.3}$ lead to reasonably good coverage rates when α is close to $1/2$. Moreover, since $\alpha \in (1/2, 1)$, the convergence rate $n^{-\frac{1-\alpha}{4}}$ is slower than the rate of the plug-in estimator $n^{-\frac{\alpha}{2}}$. Although batch-means estimator has a slower convergence rate, the next corollary shows that this method still constructs asymptotic exact confidence intervals.

COROLLARY 4.6.   *Under the assumptions of Theorem 4.3, when d is fixed, $n \to \infty$, and the step size $\eta_i = \eta i^{-\alpha}$ with $\alpha \in (\frac{1}{2}, 1)$, we have that*

$$\Pr \left(\bar{x}_{n,j} - z_{q/2} \hat{\sigma}_{n,j}^B \leq x_j^* \leq \bar{x}_{n,j} + z_{q/2} \hat{\sigma}_{n,j}^B \right) \to 1 - q,$$

where $\hat{\sigma}_{n,j}^B := \left[M^{-1} \sum_{k=1}^{M} n_k (\overline{X}_{n_k} - \overline{X}_M)(\overline{X}_{n_k} - \overline{X}_M)^T \right]_{j,j}^{1/2}.$

The proof is identical to the one of Corollary 4.4 and therefore omitted.

4.3. *Intuition behind the proof.*   Now let us provide the main idea behind the proof of Theorem 4.3. Recall that the SGD recursion in (6) can be approximated by (12): $\Delta_n \approx (I_d - \eta_n A)\Delta_{n-1} + \eta_n \xi_n$. We replace "$\approx$" by the equal sign and define an auxiliary sequence U_n:

(20) $$U_n = U_{n-1} - \eta_n A U_{n-1} + \eta_n \xi_n, \quad U_0 = \Delta_0.$$

Note that $\Delta_n = U_n$ in the linear model setting, but our proof applies to *non-linear models (e.g., generalized linear models).* For a non-linear model, the high-level idea of the proof consists of two steps:

1. Establishing the consistency (and the rate of convergence) of the batch-means estimator based on the sequence U_n;
2. Quantifying the difference between Δ_n and U_n, where Δ_n in (6) is the original sequence of interest generated from SGD for general loss functions, and U_n in (20) is its auxiliary linear approximation sequence.

In fact, the sequence U_n is the so-called "oracle iterate sequence", which has also been considered in Polyak and Juditsky (1992). It can be written in a more explicit form:

$$(21) \qquad U_n = \prod_{k=1}^{n} (I - \eta_k A) U_0 + \sum_{m=1}^{n} \prod_{k=m+1}^{n} (I - \eta_k A) \eta_m \xi_m.$$

Given the sequence U_n, we construct the batch-means estimator based on U_n as $\frac{1}{M} \sum_{k=1}^{M} n_k (\overline{U}_{n_k} - \overline{U}_M)(\overline{U}_{n_k} - \overline{U}_M)^T$, where $\overline{U}_{n_k}$ and $\overline{U}_M$ are defined as in (13) with x_i being replaced by U_i. The analysis of the batch-means estimator from U_n is simpler than that from SGD iterates x_n since the expression of U_n in (21) only involves the product of matrices and the martingale differences ξ_m. In particular, we establish the consistency of the batch-means estimator based on U_n in the following lemma.

LEMMA 4.7.   *Under Assumptions 3.1 and 3.2, when d is fixed and the step size is chosen to be $\eta_i = \eta i^{-\alpha}$ with $\alpha \in (\frac{1}{2}, 1)$, the batch-means estimator based on the sequence U_n with any bounded U_0 satisfies the following inequality for sufficiently large N and M,*

$$\mathbb{E} \left\| M^{-1} \sum_{k=1}^{M} n_k (\overline{U}_{n_k} - \overline{U}_M)(\overline{U}_{n_k} - \overline{U}_M)^T - A^{-1} S A^{-1} \right\|$$
$$\lesssim C_d M^{-\frac{1}{2}} + C_d N^{-\frac{1}{2}} + C_d^{\frac{3}{2}} (MN)^{-\frac{\alpha}{4-4\alpha}}.$$

The proof of Lemma 4.7 is provided in Appendix D.3. With Lemma 4.7 in place, to obtain the desired consistency result in Theorem 4.3, we only need to study the difference between Δ_n and U_n. In particular, let $\delta_n := \Delta_n - U_n$. We have the following recursion:

$$\delta_n = \Delta_{n-1} - U_{n-1} - \eta_n \nabla F(x_{n-1}) - \eta_n A U_{n-1}$$
$$= \delta_{n-1} - \eta_n A \delta_{n-1} + \eta_n (A \Delta_{n-1} - \nabla F(x_{n-1})).$$

Notably, by replacing ξ_n in (20) with $A\Delta_{n-1} - \nabla F(\Delta_{n-1})$, δ_n follows a similar recursion relationship to that of the sequence U_n. Based on this observation, we show that δ_n is a sequence of small numbers, and hence Δ_n and U_n are close to each other. Combining this with Lemma 4.7, we will reach the conclusion in Theorem 4.3 (see Appendix D.4 for the rigorous proof).

5. High-dimensional Linear Regression.   In Section 4.1 and 4.2, we assumed that the dimension d is fixed while $n \to \infty$. However in high-dimensional settings, it is often the case that $d \asymp n$ or $n = o(d)$. In below we consider a high-dimensional linear model $b_i = a_i^T x^* + \varepsilon_i$, where x^* is s-sparse (i.e., $\|x\|_0 \leq s$) and let $S = \{j : x_j^* \neq 0\}$ be the support of true regression coefficients. Each covariate $a_i \in \mathbb{R}^d$ is an $i.i.d.$ sub-Gaussian random vector from a common population a with the covariance matrix A, and $\varepsilon_i \sim N(0, \sigma^2)$. For simplicity, we assume that σ is known. For high-dimensional linear regression, one of the most popular estimators is the Lasso estimator, denoted by $\widehat{x}_{\mathrm{Lasso}}$. That is,

$$\widehat{x}_{\mathrm{Lasso}} = \frac{1}{2n} \operatorname*{argmin}_{x \in \mathbb{R}^d} \|b - Dx\|_2^2 + \lambda \|x\|_1,$$

where $D = [a_1, \ldots, a_n]^T \in \mathbb{R}^{n \times d}$ is the design matrix, $b = [b_1, \ldots, b_n]^T \in \mathbb{R}^{n \times 1}$ is the response vector.

As suggested by earlier work (see, e.g., Meinshausen, Meier and Bühlmann (2009); Wainwright (2009); Bühlmann and van de Geer (2011); Belloni and Chernozhukov (2013)), the Lasso estimator can be used as a screening method to reduce the set of the variables to $\widehat{S}$, a subset which contains the true support S with probability tending to 1. For example, by choosing the regularization parameter λ as (2.12) in Belloni and Chernozhukov (2013) and under certain assumptions, Belloni and Chernozhukov (2013) proved that $S \subseteq \widehat{S}$ and $|\widehat{S} \backslash S| \lesssim s$ with high probability (see Theorems 2 and 3 therein). When s is treated as a constant, the selected model will be of fixed dimension. Based on the selected model, we are able to directly apply our plug-in or batch-means estimator in Section 4 on $\widehat{S}$ to conduct inference for x_j^* for $j \in \widehat{S}$.

However, this approach has several limitations. First, the screening approach requires a strong "beta-min" assumption. In particular, this assumption requires that $\min_{j \in S} |x_j^*| > \max_{j \in S} |\widehat{x}_{\mathrm{Lasso},j} - x_j^*|$, or $\min_{j \in S} |x_j^*| \gtrsim \sqrt{s(\log d)/n}$, e.g., Bühlmann and van de Geer (2011); Bühlmann and Mandozzi (2014); Belloni and Chernozhukov (2013). Other screening methods (e.g., "Sure Independence Screening" (SIS) method (Fan and Lv, 2008)) also require a similar beta-min condition. However, since we are interested in inference of the model parameters instead of the model selection, the "beta-min"

condition should be avoidable. Second, the sparsity level s has to be treated as a constant to apply our theoretical results of either the plug-in or batch-means estimator. Furthermore, when using Lasso as a screening approach, it inevitably requires more than one pass of the data which does not fit our online setting.

5.1. *Debiasing approach.* To relax the strong conditions when using the Lasso as a screening approach, we propose a new approach for conducting inference for high-dimensional linear regression that only uses one pass of the data. Our approach is based on the following debiased Lasso estimator (van de Geer et al., 2014; Javanmard and Montanari, 2014; Zhang and Zhang, 2014),

$$\widehat{x}^d_{\text{Lasso}} = \widehat{x}_{\text{Lasso}} + \frac{1}{n}\widehat{\Omega}D^T(b - D\widehat{x}_{\text{Lasso}}),$$

where $\widehat{\Omega}$ is an estimator of the inverse covariance matrix of the design $\Omega = A^{-1}$. To construct $\widehat{\Omega}$, van de Geer et al. (2014) adopts the node-wise Lasso approach (see also Meinshausen and Bühlmann (2006)), i.e.,

$$(22) \qquad \widehat{\gamma}^j = \underset{\gamma^j \in \mathbb{R}^{d-1}}{\operatorname{argmin}} \frac{1}{2n}\|D_{\cdot,j} - D_{\cdot,-j}\gamma^j\|_2^2 + \lambda_j\|\gamma^j\|_1,$$

where $D_{\cdot,j}$ is the j-th column of the design matrix D and $D_{\cdot,-j}$ is the design submatrix without the j-th column. Further, one can estimate $\Omega_{j,j}$ by

$$\widehat{\tau}_j = \frac{1}{n}(D_{\cdot,j} - D_{\cdot,-j}\widehat{\gamma}^j)^T D_{\cdot,j},$$

Given $\widehat{\gamma}^j$ and $\widehat{\tau}_j$, the matrix Ω is estimated by,

$$(23) \qquad\qquad\qquad\qquad \widehat{\Omega} = \widehat{T}\widehat{C},$$

where $\widehat{T} := \operatorname{diag}(1/\widehat{\tau}_1, \ldots, 1/\widehat{\tau}_d)$ and $\widehat{C} := \begin{pmatrix} 1 & -\widehat{\gamma}_2^1 & \ldots & -\widehat{\gamma}_d^1 \\ -\widehat{\gamma}_1^2 & 1 & \ldots & -\widehat{\gamma}_d^2 \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\gamma}_1^d & -\widehat{\gamma}_2^d & \ldots & 1 \end{pmatrix}.$

Note that in the existing literature, $\widehat{x}_{\text{Lasso}}$ and $\widehat{\Omega}$ are obtained via deterministic convex optimization. Therefore, debiased Lasso approaches cannot be directly applied to the stochastic setting in this work. To address this issue, we propose to compute the estimators for both x^* and Ω using the Regularization Annealed epoch Dual AveRaging (RADAR) algorithm

([Agarwal, Negahban and Wainwright](), [2012]()), which is a variant of SGD. Similar to SGD, RADAR computes the stochastic gradient on one data point at each iteration. Please refer to [Agarwal, Negahban and Wainwright]() ([2012]()) for more details of the RADAR algorithm. The reason why we use RADAR instead of the vanilla SGD is because RADAR provides the optimal convergence rate in terms of the ℓ_1-norm. In particular, we apply RADAR to the following ℓ_1-regularized problem,

$$\min_{x \in \mathbb{R}^d} \mathbb{E}(b - a^T x)^2 + \lambda \|x\|_1 \tag{24}$$

and let $\widehat{x}_n$ be the solution output from RADAR with n iterations. Similarly, we again use stochastic optimization instead of deterministic optimization in the node-wise Lasso in (22), that is, applying RADAR to the following optimization problem for each dimension $1 \le j \le d$,

$$\widehat{\gamma}^j = \operatorname*{argmin}_{\gamma^j \in \mathbb{R}^{d-1}} \mathbb{E}\|a_j - a_{-j}\gamma^j\|_2^2 + \lambda_j \|\gamma^j\|_1, \tag{25}$$

where a_j is the j-th coordinate of the population design vector a and a_{-j} is the subvector of a without the j-th coordinate. Given $\widehat{\gamma}^j$ from solving (25) via the iterative stochastic algorithm, the inverse covariance estimator $\widehat{\Omega}$ is constructed according to (23).

It is noteworthy that although the proposed $\widehat{\Omega}$ is of the same form as the estimator for A^{-1} in [van de Geer et al.]() ([2014]()), our $\widehat{\gamma}^j$ is different from the one in [van de Geer et al.]() ([2014]()). More precisely, our $\widehat{\gamma}^j$ is the output of a stochastic gradient-based algorithm, while [van de Geer et al.]() ([2014]()) obtained $\widehat{\gamma}^j$ from deterministic optimization in (22). With all these ingredients in place, we present the stochastic gradient based construction of the confidence interval for x_j^* for $j \in \{1, \ldots, d\}$ in Algorithm 1. The hypothesis test can also be performed once the estimator of the asymptotic variance of $\widehat{x}_j^d$ is available (see Theorem 5.2). We note that the proposed method is computationally more efficient than the methods based on deterministic optimization. It only requires one pass of the data with the total per-iteration complexity $O(d^2)$ (note that the node-wise Lasso needs to solve d optimization problems) and is applicable to online data (in contrast to multiple passes of data with deterministic optimization used in existing methods). The details of the algorithm are provided in Algorithm 1.

To provide the theoretical justification for Algorithm 1 in terms of constructing valid confidence intervals, we make the following assumptions (which are similar to the assumptions made in [van de Geer et al.]() ([2014]())).

ASSUMPTION 5.1. *The covariate a is a sub-Gaussian random vector with*

Algorithm 1 Stochastic Optimization Based Confidence Interval Construction for High-dimensional Sparse Linear Regression

---

Inputs:
    Regularization parameter $\lambda \asymp \sqrt{\log d/n}$, and $\lambda_j \asymp \sqrt{\log d/n}$ for each dimension j, the noise level σ, confidence level $1 - \alpha$.
for $t = 1$ to n **do**
    Randomly sample the data (a_t, b_t) and update the design $D \leftarrow [D^T, a_t]^T$ and response $b \leftarrow [b^T, b_t]^T$.
    Update x_t by running one iteration of RADAR on the optimization problem (24) using the stochastic gradient $(a_t^T x_{t-1} - b_t)a_t$,
        for $j = 1$ to d **do**
            Update γ_t^j by running one iteration of RADAR on the optimization problem (25) using the stochastic gradient $(a_{t,-j}^T \gamma_{t-1}^j - a_{t,j})a_{t,-j}$.
    end for
end for
Let $\widehat{x}_n = x_n$ and $\widehat{\gamma}^j = \gamma_n^j$ for $j \in \{1, \ldots, d\}$ be the final outputs.
Construct the debiased estimator $\widehat{x}^d$ with $\widehat{\Omega}$ defined in (23).

$$(26) \qquad\qquad \widehat{x}^d = \widehat{x}_n + \frac{1}{n}\widehat{\Omega}D^T(b - D\widehat{x}_n).$$

Outputs:
    The $(1 - \alpha)$ confidence interval for each $x_j^* : \widehat{x}_j^d \pm z_{\alpha/2}\sigma\sqrt{(\widehat{\Omega}\widehat{A}\widehat{\Omega})_{jj}/n}$,
    where $\widehat{A} = \frac{1}{n}D^T D$.

---

variance proxy K^2. The population covariance A has bounded eigenvalues,

$$0 < \mu < \lambda_{\min}(A) < \lambda_{\max}(A) < L_F.$$

Denote the set of parameters by $\mathcal{B}(s) = \{x \in \mathbb{R}^d; \|x\|_0 \leq s$ and $\|x\|_1$ is bounded by a constant$\}$. The true regression parameter $x^* \in \mathcal{B}(s)$ where $s = o(\sqrt{n}/\log d)$. Moreover, the inverse covariance Ω has sparse rows. In particular, define

$$s_j = |\{1 \leq k \leq d : k \neq j, \Omega_{j,k} \neq 0\}|.$$

We assume that $\max_j s_j \leq Cs$ for some constant C.

Under Assumption 5.1, we first present an ℓ_1-bound result as a corollary of Proposition 1 in Agarwal, Negahban and Wainwright (2012),

PROPOSITION 5.1. *Under Assumption 5.1 and using the same algorithm parameters as Proposition 1 in Agarwal, Negahban and Wainwright (2012), there exists a constant c_0, such that $\widehat{x}_n$ in Algorithm 1 satisfies $\|\widehat{x}_n - x^*\|_1 \leq$*

$c_0 s \sqrt{\frac{\log d}{n}}$ uniformly in $x^* \in \mathcal{B}(s)$ with high probability. Further, for each $j = \{1, \ldots, d\}$, we have

$$\|\widehat{\gamma}^j + \Omega_{j,j}^{-1} (\Omega_{j,-j})^T\|_1 \leq c_0 s_j \sqrt{\frac{\log d}{n}}$$

holds with high probability.

The proof of Proposition 5.1 is provided in Appendix E.1. Let $\mathbb{P}_{x^*}$ be the distribution under the high-dimensional linear model $b_i = a_i^T x^* + \epsilon_i$. Given Proposition 5.1, we state the inference result in the next theorem. We note that although the statement of the following theorem is similar to Theorem 2.2 and Corollary 2.1 in van de Geer et al. (2014), the proof is more technically involved. The main challenge is that the existing analysis in van de Geer et al. (2014) starts from the KKT condition of the deterministic optimization for estimating Ω. However, we estimate Ω using the stochastic optimization and thus the corresponding KKT condition no longer holds. Please refer to the proof in Appendix E.2 for more detail.

THEOREM 5.2. *Under Assumption 5.1, for suitable choices of* $\lambda \asymp \sqrt{\log d/n}$ *and* $\lambda_j \asymp \sqrt{\log d/n}$, *we have for all* $j \in \{1, \ldots, d\}$ *and all* $z \in \mathbb{R}$,

$$\sup_{x^* \in \mathcal{B}(s)} \left| \mathbb{P}_{x^*} \left(\frac{\sqrt{n}(\widehat{x}_j^d - x_j^*)}{\sigma \sqrt{(\widehat{\Omega} \widehat{A} \widehat{\Omega}^T)_{jj}}} \leq z \right) - \Phi(z) \right| = o_p(1),$$

where $\widehat{x}^d$ *is the debiased estimator defined in* (25), $\widehat{\Omega}$ *is defined in* (23) *and the sample covariance matrix* $\widehat{A} = \frac{1}{n} D^T D$.

Theorem 5.2 shows that $\frac{1}{\sigma \sqrt{(\widehat{\Omega} \widehat{A} \widehat{\Omega}^T)_{jj}}} \sqrt{n}(\widehat{x}_j^d - x_j^*)$ converges in distribution to $N(0, 1)$ uniformly for any $x^* \in \mathcal{B}(s)$ and $j \in \{1, 2, \ldots, d\}$, which verifies the correctness of the asymptotic pointwise confidence interval in the output of Algorithm 1 for x_j^*. Given the uniform convergence result in Theorem 5.2, we can construct p-values for each single component, and further conduct multiple testing based on component-wise p-values. We also note that a similar uniform convergence result has been established in Ning and Liu (2017) for a score test approach (see Remark 4.6). It is also interesting to investigate the stochastic optimization based score test as a future work.

6. Numerical Simulations. In this section, we investigate the empirical performance of the plug-in estimator and batch-means estimator of the

asymptotic covariance matrix. We consider both linear and logistic regression models, where $\{a_i, b_i\}$ are $i.i.d.$ samples with $a_i \sim \mathcal{N}(0, \Sigma)$ and x^* is the true parameter vector of the model. For both models, we consider three different structures of the $d \times d$ covariance matric Σ:

- Identity: $\Sigma = I_d$;
- Toeplitz: $\Sigma_{i,j} = r^{|i-j|}$;
- Equi Corr: $\Sigma_{i,j} = r$ for all $i \neq j$, $\Sigma_{i,i} = 1$ for all i.

We report $r = 0.5$ for Toeplitz and $r = 0.2$ for equicorrelation (Equi Corr) covariance matrices in the main paper. The experimental results on other settings of r are relegated to Appendix F due to space limitations. The noise ε_n in linear regression is set to $i.i.d.$ $N(0, \sigma^2)$ with $\sigma = 1$. The parameter α in the step size is chosen to be 0.501 (slightly larger than 0.5). All the reported results are obtained by taking the average of 500 independent runs. We consider the finite sample behavior of the plug-in estimator and the batch-means estimator for the inference of each individual regression coefficient x_j, $j \in \{1, 2, \ldots, d\}$.

6.1. *Low-dimensional cases.* In each case, we consider the sample size $n = 10^5$ and the dimension $d = 5, 20, 100, 200$. For each model, the corresponding parameter x^* is a d-dimensional vector linearly spaced between 0 and 1. The thresholding scheme is not used for the plug-in estimator. In fact, we observe that the obtained A_n is always invertible and the results are stable without the thresholding. For the batch-means estimator (BM in short), we consider three different choices of the number of batches: $M = n^{0.2}$, $M = n^{0.25}$, and $M = n^{0.3}$. Note that $\alpha = 0.501$. As we suggested in Corollary 4.5, to achieve a better convergence rate, the number of batches M is chosen around the optimal value $n^{\frac{1-\alpha}{2}} \approx n^{0.25}$.

We set the nominal coverage probability $1 - q$ to 95%. The performance of an estimator is measured by the average coverage rate (Cov Rate) of the confidence intervals and the average length (Avg Len) of the intervals for each individual coefficient. For each setting, we also report the oracle length of the confidence interval with respect to the true covariance matrix $A^{-1}SA^{-1}$ and the corresponding coverage rate when using the same center as the BM.

For linear regression, the asymptotic covariance is $A^{-1}SA^{-1} = \sigma^2 \Sigma^{-1} = \Sigma^{-1}$ and the oracle interval length for each coordinate j will be $\frac{2z_{q/2}(\Sigma^{-1})_{jj}}{\sqrt{n}}$. Table 1 shows the empirical performance of the plug-in and BM under linear models with three different design covariance matrices.

From Table 1, both the plug-in and BM achieve good performance. The

TABLE 1

Linear Regression: The average coverage rate and length of confidence intervals, for the nominal coverage probability 95%. The columns (BM: n^c for c=0.2, 0.25, and 0.3) correspond to the batch-means estimator with $M = n^c$ number of batches. Cov Rate under "Oracle" refers to coverage rates when using the same center as BM but with oracle interval lengths. Standard errors are reported in the brackets.

	d	Plug-in	BM			Oracle
			$M = n^{0.2}$	$M = n^{0.25}$	$M = n^{0.3}$	
Identity Σ						
Cov Rate (%)	5	95.68(0.87)	90.28(0.46)	93.68(0.79)	91.64(0.79)	87.44
Avg Len ($\times 10^{-2}$)		1.49(0.01)	1.39(0.01)	1.47(0.01)	1.43(0.01)	1.24
Cov Rate (%)	20	94.99(0.94)	91.30(1.08)	93.92(1.25)	92.95(1.19)	88.24
Avg Len ($\times 10^{-2}$)		1.44(0.01)	1.35(0.01)	1.41(0.01)	1.38(0.01)	1.24
Cov Rate (%)	100	95.04(1.01)	90.75(1.36)	93.15(1.12)	92.37(1.10)	87.89
Avg Len ($\times 10^{-2}$)		1.41(0.01)	1.32(0.01)	1.35(0.01)	1.35(0.01)	1.24
Cov Rate (%)	200	94.75(1.13)	90.49(1.21)	92.97(1.17)	91.97(1.18)	88.12
Avg Len ($\times 10^{-2}$)		1.39(0.01)	1.30(0.01)	1.31(0.01)	1.32(0.01)	1.24
Toeplitz Σ						
Cov Rate (%)	5	95.24(0.92)	91.16(0.50)	94.28(0.86)	93.04(0.90)	88.31
Avg Len ($\times 10^{-2}$)		1.83(0.10)	1.74(0.10)	1.82(0.11)	1.78(0.12)	1.53
Cov Rate (%)	20	94.84(0.97)	90.97(1.08)	93.75(0.93)	92.77(0.81)	87.26
Avg Len ($\times 10^{-2}$)		1.81(0.05)	1.71(0.06)	1.78(0.06)	1.76(0.06)	1.58
Cov Rate (%)	100	95.01(1.12)	90.36(1.33)	91.83(1.09)	91.52(1.17)	89.11
Avg Len ($\times 10^{-2}$)		1.77(0.02)	1.67(0.03)	1.67(0.03)	1.69(0.02)	1.60
Cov Rate (%)	200	94.69(1.33)	90.01(1.41)	91.65(1.36)	91.24(1.41)	89.43
Avg Len ($\times 10^{-2}$)		1.74(0.02)	1.62(0.02)	1.62(0.02)	1.62(0.02)	1.60
Equi Corr Σ						
Cov Rate (%)	5	94.80(0.88)	90.92(1.09)	93.60(0.92)	92.32(0.68)	86.79
Avg Len ($\times 10^{-2}$)		1.60(0.01)	1.46(0.01)	1.55(0.01)	1.52(0.01)	1.31
Cov Rate (%)	20	95.10(0.99)	91.15(1.14)	93.66(0.99)	92.78(0.92)	88.04
Avg Len ($\times 10^{-2}$)		1.59(0.01)	1.47(0.01)	1.54(0.01)	1.51(0.01)	1.36
Cov Rate (%)	100	94.93(1.06)	90.86(1.26)	93.19(1.15)	92.29(1.10)	87.15
Avg Len ($\times 10^{-2}$)		1.56(0.01)	1.47(0.01)	1.52(0.01)	1.50(0.01)	1.38
Cov Rate (%)	200	94.49(1.09)	90.57(1.45)	92.45(1.27)	91.91(1.13)	87.22
Avg Len ($\times 10^{-2}$)		1.51(0.01)	1.45(0.01)	1.49(0.01)	1.49(0.01)	1.38

plug-in gives better average coverage rate than BM: the average coverage rates in all different settings are nearly 95%. However, the average length of plug-in is usually larger than that of BM and the corresponding oracle interval length. On the other hand, BM achieves about 92% coverage rate when $M = n^{0.25}$ or $M = n^{0.3}$. We further consider the logistic regression. To provide an oracle interval length based on the true asymptotic covariance $A^{-1}SA^{-1} = A^{-1}$, we estimate A in (51) by its empirical version $\widehat{A}$ using one million fresh samples and the oracle interval length of each coordinate j is computed as $\frac{2z_{q/2}(\widehat{A}^{-1})_{jj}}{\sqrt{n}}$. We provide the result in Table 2 for different design covariance matrices. From Table 2, the plug-in still achieves nearly

95% average coverage rate. The BM achieves about 90% coverage rate and the average length is usually smaller than the oracle length. Moreover, as d becomes larger, the interval lengths for both estimators increase. Finally, the performance of BM is insensitive to the choice of the number of batches M: different M's lead to comparable coverage rates. There are two reasons for the undercoverage of BM. First, the obtained center could deviate from x^* that introduces the bias. Second, the BM has a slower convergence rate as compared to the plug-in (especially for the case of logistic regression). However, since the BM only uses the iterates from SGD, it is computationally more efficient than the plug-in estimator which requires the computation of the Hessian matrix $\tilde{A}_n$ and its inverse.

6.2. *High-dimensional cases.* In a high-dimensional setting, we consider the sample size $n = 100$, and the dimension $d = 500$. The active set $S_0 = \{1, 2, \ldots, s_0\}$, where the cardinality $s_0 = |S_0| = 3$ or 15. The non-zero regression coefficients $\{x_j^*\}_{j \in S_0}$ are from a fixed realization of s_0 *i.i.d.* uniform distribution $U[0, c]$ with $c = 2$.

First, we consider the average coverage rate and the average length of the intervals for individual coefficients corresponding to the variables in either S_0 or S_0^c where $S_0^c = \{1, \ldots, d\} \backslash S_0$. Again, we set the nominal coverage probability $1 - q$ to 95%. Our experimental setup follows directly from van de Geer et al. (2014), and we provide the oracle length of the confidence intervals for comparison. For linear regression, the asymptotic covariance is $A^{-1}SA^{-1} = \sigma^2 \Sigma^{-1} = \Sigma^{-1}$ and the oracle interval length for each coordinate j will be $\frac{2z_{q/2}(\Sigma^{-1})_{jj}}{\sqrt{n}}$. We provide the result in Table 3 for different design covariance matrices.

For high-dimensional linear regression, Algorithm 1 achieves good performance, especially in sparse settings ($s_0 = 3$). From Table 3, the average coverage rate is about 90%. For less sparse problems ($s_0 = 15$), our method still achieves about 88% average coverage rate for different design covariance matrices. The coverage rates of the obtained confidence intervals on active sets S_0 are slightly better than those on S_0^c. The average lengths on both sets are slightly smaller than the oracle lengths. The performance of the cases with identity design matrices are better than those with Toeplitz and equicorrelation design matrices (e.g., having smaller standard deviations). It is reasonable since it is easier to estimate the inverse covariance matrix Ω when it is an identity matrix.

In Table 4, we also provide the results using the deterministic optimization (instead of the stochastic RADAR) for constructing $\widehat{\Omega}$ (van de Geer et al., 2014). Both methods achieve comparably reliable coverage rates. From Table

TABLE 2
Logistic Regression: The average coverage rate and length of confidence intervals, for the nominal coverage probability 95%. The columns (BM: n^c for c=0.2, 0.25, and 0.3) correspond to the batch-means estimator with $M = n^c$ number of batches. Cov Rate under "Oracle" refers to coverage rates when using the same center as BM but with oracle interval lengths. Standard errors are reported in the brackets.

	d	Plug-in	BM			Oracle
			$M = n^{0.2}$	$M = n^{0.25}$	$M = n^{0.3}$	
Identity Σ						
Cov Rate (%)	5	95.04(1.13)	89.24(1.55)	90.12(1.70)	89.36(1.97)	91.45
Avg Len ($\times 10^{-2}$)		3.24(0.41)	3.01(0.26)	2.94(0.25)	2.87(0.23)	3.09
Cov Rate (%)	20	95.00(1.34)	89.35(2.00)	90.22(1.67)	89.74(2.11)	90.37
Avg Len ($\times 10^{-2}$)		3.79(0.27)	3.53(0.25)	3.46(0.23)	3.42(0.22)	3.68
Cov Rate (%)	100	94.69(1.06)	89.42(1.66)	90.84(1.68)	90.41(2.01)	91.24
Avg Len ($\times 10^{-2}$)		5.21(0.26)	4.97(0.24)	4.87(0.23)	4.80(0.24)	5.06
Cov Rate (%)	200	94.47(0.91)	89.01(1.41)	90.47(1.49)	90.36(1.74)	92.08
Avg Len ($\times 10^{-2}$)		6.05(0.29)	5.94(0.26)	5.82(0.27)	5.71(0.25)	5.97
Toeplitz Σ						
Cov Rate (%)	5	94.96(1.58)	88.96(2.32)	90.56(2.06)	90.12(2.04)	92.41
Avg Len ($\times 10^{-2}$)		4.06(0.34)	3.75(0.28)	3.73(0.27)	3.61(0.25)	4.04
Cov Rate (%)	20	95.17(1.23)	89.01(1.93)	90.39(1.88)	89.79(1.81)	91.07
Avg Len ($\times 10^{-2}$)		5.74(0.29)	5.57(0.25)	5.22(0.23)	4.95(0.22)	5.59
Cov Rate (%)	100	94.91(0.89)	89.91(1.74)	90.83(1.81)	90.54(1.97)	91.47
Avg Len ($\times 10^{-2}$)		8.47(0.37)	8.01(0.28)	7.71(0.26)	7.37(0.25)	8.28
Cov Rate (%)	200	94.59(1.04)	89.72(1.81)	90.74(1.93)	90.32(2.02)	92.29
Avg Len ($\times 10^{-2}$)		9.81(0.41)	9.24(0.34)	8.95(0.31)	8.78(0.29)	9.84
Equi Corr Σ						
Cov Rate (%)	5	94.80(1.66)	88.08(1.46)	88.64(1.73)	89.48(1.51)	93.79
Avg Len ($\times 10^{-2}$)		3.43(0.35)	3.28(0.28)	3.24(0.25)	3.20(0.24)	3.38
Cov Rate (%)	20	94.54(1.73)	89.27(1.33)	90.64(1.60)	90.31(2.10)	92.50
Avg Len ($\times 10^{-2}$)		5.37(0.31)	4.84(0.26)	4.77(0.24)	4.51(0.21)	5.19
Cov Rate (%)	100	94.79(1.08)	89.01(1.70)	90.27(1.76)	89.42(2.01)	94.92
Avg Len ($\times 10^{-2}$)		10.24(0.51)	10.17(0.47)	9.75(0.42)	9.24(0.40)	10.89
Cov Rate (%)	200	94.24(1.09)	89.13(1.44)	90.01(1.92)	89.23(1.79)	92.40
Avg Len ($\times 10^{-2}$)		15.70(0.62)	14.82(0.57)	14.01(0.55)	13.88(0.52)	15.31

4, the average coverage rates are closer to the nominal levels, better than those in Table 3. The undercovering in Table 3 is due to the estimation error of the diagonals of $\hat{\Omega}\hat{A}\hat{\Omega}$ using stochastic optimization method. Based on the computational and storage requirements, a practitioner may decide to use a one-pass algorithm or a more accurate estimator under deterministic optimization.

7. Conclusions and Future Works. This paper presents two consistent estimators of the asymptotic variance of the average iterate from SGD, especially a computationally more efficient batch-means estimator that only

TABLE 3

High-dimensional linear regression, the average coverage rate and length of confidence intervals, for the nominal coverage probability 95%. Standard errors are reported in the brackets.

Measure	Identity Σ	Toeplitz Σ	Equi Corr Σ
$S_0 = \{1, 2, 3\}$			
Cov Rate $S_0(\%)$	91.93 (3.13)	91.40 (2.39)	90.20 (1.38)
Avg Len S_0	0.387 (0.002)	0.401 (0.019)	0.360 (0.014)
Cov Rate $S_0^c(\%)$	90.80 (1.79)	90.21 (1.98)	89.73 (1.96)
Avg Len S_0^c	0.386 (0.002)	0.417 (0.022)	0.384 (0.023)
Oracle Len	0.392	0.506	0.438
$S_0 = \{1, 2, \ldots, 15\}$			
Cov Rate S_0 (%)	90.48 (1.73)	89.84 (2.61)	89.45 (0.87)
Avg Len S_0	0.379 (0.002)	0.430 (0.024)	0.384 (0.020)
Cov Rate S_0^c (%)	88.43 (2.30)	86.79 (2.10)	87.12 (1.36)
Avg Len S_0^c	0.360 (0.003)	0.425 (0.024)	0.383 (0.022)
Oracle Len	0.392	0.506	0.438

TABLE 4

High-dimensional linear regression using node-wise lasso instead of RADAR for inference, the average coverage rate and length of confidence intervals, for the nominal coverage probability 95%. Standard errors are reported in the brackets.

Measure	Identity Σ	Toeplitz Σ	Equi Corr Σ
$S_0 = \{1, 2, 3\}$			
Cov Rate $S_0(\%)$	94.73 (1.42)	92.60 (1.95)	91.33 (1.99)
Avg Len S_0	0.393 (0.001)	0.472 (0.011)	0.431 (0.007)
Cov Rate $S_0^c(\%)$	96.13 (1.44)	95.17 (1.71)	95.92 (2.04)
Avg Len S_0^c	0.386 (0.001)	0.481 (0.014)	0.429 (0.011)
Oracle Len	0.392	0.506	0.438
$S_0 = \{1, 2, \ldots, 15\}$			
Cov Rate S_0 (%)	91.80 (1.04)	91.07 (2.01)	90.33 (1.92)
Avg Len S_0	0.399 (0.001)	0.512 (0.015)	0.424 (0.008)
Cov Rate S_0^c (%)	95.75 (1.80)	93.17 (1.90)	94.11 (1.73)
Avg Len S_0^c	0.379 (0.001)	0.505 (0.016)	0.453 (0.008)
Oracle Len	0.392	0.506	0.438

uses iterates from SGD. With the proposed estimators, we are able to construct asymptotically exact confidence intervals and hypothesis tests.

We further discuss statistical inference based on SGD for high-dimensional linear regression. An extension to generalized linear models is an interesting problem for future work.

The seminal work by Toulis and Airoldi (2016) develops the averaged implicit SGD procedure and provides the characterization of the limiting distribution. It would be interesting to establish the consistency of the batch-means estimator based on iterates from implicit SGD. It is also interesting to relax the current assumptions and consider SGD for more challenging

optimization problems (e.g., non-convex problems).

A. Verifying assumptions for two examples.   In this section, we verify Assumptions 3.1, 3.2 and 4.1 on Examples 2.1 and 2.2.

LEMMA A.1.   *In Example 2.1, suppose $A = \mathbb{E}a_n a_n^T$ is a positive definite matrix. Assumptions 3.1 and 3.2 hold for $0 < \mu = \lambda_{\min}(A) \leq \lambda_{\max}(A) = L_F$. To track the dependence on dimension d, we consider the case where $a_n \sim \mathcal{N}(0, I_d), \varepsilon \sim \mathcal{N}(0, 1)$, then Assumptions 3.1, 3.2 and 4.1 hold with*

$$L_F = 1, \quad \mathrm{tr}(S) = O(d), \quad \Sigma_1 = 0, \quad \Sigma_2 = O(d^2),$$
$$\Sigma_3 = O(d^2), \quad \Sigma_4 = O(d^4), \quad L_2 = 0, \quad L_4 = O(d^2).$$

Therefore, the dimension constant $C_d = O(d)$.

PROOF.   It is easy to see that $\nabla F(x) = A(x - x^*)$ and $\nabla^2 F(x) = A$ for all x. $S := \mathbb{E}\left([\nabla f(x^*, \zeta)][\nabla f(x^*, \zeta)]^T\right) = \mathbb{E}\varepsilon_n^2 a_n a_n^T$. Therefore, Assumption 3.1 holds for $0 < \mu = \lambda_{\min}(A) \leq \lambda_{\max}(A) = L_F$. For Assumption 3.2, the sequence $\xi_n = (A - a_n a_n^T)\Delta_{n-1} + a_n \varepsilon_n$. First, we notice it is indeed a martingale sequence, since $\mathbb{E}_{n-1}(A - a_n a_n^T) = 0$ and $\mathbb{E}_{n-1}\varepsilon_n = 0$. Second, by $S = \mathbb{E}\varepsilon_n^2 a_n a_n^T$,

$$\Sigma(\Delta_{n-1}) = \mathbb{E}_{n-1}\xi_n \xi_n^T - S = \mathbb{E}_{n-1}[(a_n a_n^T - A)\Delta_{n-1}\Delta_{n-1}^T(a_n a_n^T - A)].$$

Assumption 3.2 holds because

$$\|\Sigma(\Delta)\| \leq |\mathrm{tr}(\Sigma(\Delta))| \leq \|\Delta\|_2^2 \mathbb{E}_{n-1}\left\|(a_n a_n^T - A)\right\|^2.$$

and by Hölder's inequality:

$$\mathbb{E}_{n-1}\|(A - a_n a_n^T)\Delta_{n-1} + a_n \varepsilon_n\|_2^4$$
$$\leq 8\|\Delta_{n-1}\|_2^4 \mathbb{E}_{n-1}\left\|A - a_n a_n^T\right\|^4 + 8\mathbb{E}_{n-1}(\varepsilon_n^4\|a_n\|_2^4).$$

To track the dependence on dimension d, we consider the case where $a_n \sim \mathcal{N}(0, I_d), \varepsilon \sim \mathcal{N}(0, 1)$. Lemma A.1 shows that the dimension constant can be chosen as $C_d = d$.

To track the dependence on dimension d, note that $\nabla^2 f(x, \zeta_n) = a_n a_n^T$, so $\nabla^2 F(x) = I_d$, and $L_F = 1$. Next note that $S = \mathbb{E}\varepsilon_n^2 a_n a_n^T = \text{var}(\varepsilon_n)I_d$, so $\text{tr}(S) = O(d)$. Moreover $\Sigma_1 = 0$, while $A = I_d$, so

$$\Sigma_2 = \mathbb{E}\left\|(a_n a_n^T - A)\right\|^2 = \mathbb{E}(\|a_n\|_2^2 - 1)^2 = O(d^2), \quad \Sigma_3 = 8\mathbb{E}(\varepsilon_n^4 \|a_n\|_2^4) = O(d^2),$$

and

$$\Sigma_3 = 8\mathbb{E}(\varepsilon_n^4 \|a_n\|_2^4) = O(d^2), \quad \Sigma_4 = 64\mathbb{E}\left\|A - a_n a_n^T\right\|^4 \asymp \mathbb{E}(\|a_n\|_2^2 - 1)^4 = O(d^4)$$

And for the Assumption 4.1, because $\nabla^2 f(x, \zeta_n) = a_n a_n^T$, therefore $L_2 = 0$, and $L_4 = \|\mathbb{E}(a_n a_n^T)^2 - I_d\| \le \mathbb{E}\|a_n\|_2^4 + 1 = O(d^2)$. □

LEMMA A.2. *In Example 2.2, suppose $\|a_n\|_2$ has bounded eighth moment. Furthermore, assuming a_n has a strictly positive density with respect to Lebesgue measure and that the iterates are bounded. Assumptions 3.1 and 3.2 hold. To track the dependence on dimension d, we consider the case where $a_n \sim \mathcal{N}(0, I_d)$, then Assumptions 3.1, 3.2 and 4.1 hold with*

$$L_F = O(d), \quad \text{tr}(S) = O(d), \quad \Sigma_1 = O(d^{\frac{3}{2}}), \quad \Sigma_2 = O(d^2),$$
$$\Sigma_3 = O(d^2), \quad \Sigma_4 = O(d^4), \quad L_2 = O(d^{\frac{3}{2}}), \quad L_4 = O(d^2).$$

Therefore, the dimension constant $C_d = O(d)$.

PROOF. Using the fact that $|\varphi'(x)| \le \frac{1}{4}$ for any x, $\nabla F(x)$ is Lipschitz continuous with the Lipschitz constant $L_F = \frac{1}{4}\mathbb{E}\|a_n\|_2^2$. Moreover,

$$(51) \qquad A = \nabla^2 F(x^*) = \mathbb{E}\frac{a_n a_n^T}{[1 + \exp(\langle a_n, x^* \rangle)][1 + \exp(-\langle a_n, x^* \rangle)]},$$

is positive semi-definite. Under assumption that a_n has a strictly positive density with respect to Lebesgue measure, and that the iterates are bounded, Lemma A.3 below guarantees that $\nabla^2 F(x)$ has strictly positive minimum eigenvalue.

For the martingale difference assumption in Assumption 3.2, note that

$$\nabla^2 f(x, \zeta_n) = \frac{a_n a_n^T}{[1 + \exp(\langle a_n, x \rangle)][1 + \exp(-\langle a_n, x \rangle)]},$$

which implies that $\left\|\nabla^2 f(x, \zeta_n)\right\| \le \|a_n\|_2^2$. As long as $\|a_n\|_2$ has bounded eighth moment, Lemma 3.1 applies, which establishes Assumption 3.2.

To track the dependence on dimension d, recall that

$$\nabla f(x, \zeta_n) = \frac{-b_n a_n}{1 + \exp(-b_n \langle x, a_n \rangle)},$$

$$\nabla^2 f(x, \zeta_n) = \frac{a_n a_n^T}{[1 + \exp(\langle a_n, x \rangle)][1 + \exp(-\langle a_n, x \rangle)]}.$$

So $H(\zeta_n) = \|\nabla^2 f(x, \zeta_n)\| \leq \|a_n\|_2^2$. This leads to $L_F \leq \mathbb{E}H(\zeta_n) \leq \mathbb{E}\|a_n\|_2^2 = O(d)$, and for any integer m, $\mathbb{E}H(\zeta)^m = O(d^m)$. Further more, for any integer m,

$$\mathbb{E}\|\nabla f(x^*, \zeta)\|^m \leq \mathbb{E}\|a_n\|^m = O(d^{\frac{m}{2}}).$$

So by Lemma 3.1

$$\text{tr}(S) = O(d), \quad \Sigma_1 = O(d^{\frac{3}{2}}), \quad \Sigma_2 = O(d^2), \quad \Sigma_3 = O(d^2), \quad \Sigma_4 = O(d^4).$$

And for Assumption 4.1, we check the Frechet derivative of $\nabla^2 f$. Given a uninorm vector $v \in \mathbb{R}^d$,

$$\lim_\epsilon \frac{1}{\epsilon} \nabla^2 f(x + \epsilon v, \zeta_n) - \nabla^2 f(x^*, \zeta_n) = a_n a_n^T \frac{\langle a_n, v \rangle (\exp(\langle a_n, x \rangle) - \exp(2\langle a_n, x \rangle))]}{[1 + \exp(\langle a_n, x \rangle)]^3}.$$

It is bounded by $\|a_n\|^3$, so $L_2 = O(d^{\frac{3}{2}})$. Then $\nabla^2 f(x, \zeta_n) \preceq \|a_n\|^2 I_d$, so $L_4 \lesssim O(d^2)$. $\square$

LEMMA A.3. *Let a be a random variable on $\mathbb{R}^d$ with strictly positive density with respect to the Lebesgue measure, and f be a continuous strictly positive function on $\mathbb{R}^d$. Then*

$$A := \mathbb{E}f(a)aa^T$$

is a PSD matrix with the minimum eigenvalue strictly above zero.

PROOF. Let x be a unit-norm eigenvector of A associated with the minimum eigenvalue of A. It suffices to show that $x^T A x > \mu_0$. Let B be a neighborhood of x such that for any $y \in B$, $\langle y, x \rangle > \frac{1}{2}$, and $f(y) > m$ for a constant $m > 0$. Then if we denote the density of a as $\mu(dy)$, we have

$$x^T A x = \int \mu(dy) f(y) x y y^T x \geq \int_B \mu(dy) f(y) x y y^T x \geq \frac{1}{4}\mu(B)m =: \mu_0.$$

$\square$

B. Supplement to Section 3.

B.1. *Proof of Lemma 3.1.* Before we prove Lemma 3.1, we introduce the following Lemma B.1,

LEMMA B.1. *The following holds for any vectors:*

$$n^3(\|x_1\|_2^4 + \ldots + \|x_n\|_2^4) \geq \|x_1 + \ldots + x_n\|_2^4$$

PROOF. Apply Cauchy-Schwartz inequality twice:

$$n(\|x_1\|_2^4 + \ldots + \|x_n\|_2^4) \geq (\|x_1\|_2^2 + \ldots + \|x_n\|_2^2)^2$$

and

$$n(\|x_1\|_2^2 + \ldots + \|x_n\|_2^2) \geq (\|x_1\|_2 + \ldots + \|x_n\|_2)^2.$$

Note that $\|x_1\|_2 + \ldots + \|x_n\|_2 \geq \|x_1 + \ldots + x_n\|_2$. □

LEMMA 3.1. *If there is a function $H(\zeta)$ with bounded fourth moment, such that the Hessian of $f(x,\zeta)$ is bounded by*

$$(52) \qquad \qquad \left\|\nabla^2 f(x,\zeta)\right\| \leq H(\zeta)$$

for all x, and $\nabla f(x^*,\zeta)$ have a bounded fourth moment, then Assumption 3.2 holds, where

$$\Sigma_1 = 2\sqrt{\mathbb{E}\|\nabla f(x^*,\zeta)\|_2^2 \mathbb{E}H(\zeta)^2}, \quad \Sigma_2 = 4\mathbb{E}H(\zeta)^2.$$

$$\Sigma_4 = \max\{64\mathbb{E}H(\zeta)^4, 8\mathbb{E}\|\nabla f(x^*,\zeta)\|_2^4\}.$$

PROOF. The first item of Assumption 3.2 holds by definition. For notational simplicity, define

$$L(\Delta_{n-1},\zeta_n) = (\nabla F(x_{n-1}) - \nabla F(x^*)) - (f(x_{n-1},\zeta_n) - \nabla f(x^*,\zeta_n)),$$

By the definition of ξ_n and notice that $\nabla F(x^*) = 0$, we have,

$$(53) \qquad \qquad \xi_n = L(\Delta_{n-1},\zeta_n) - \nabla f(x^*,\zeta_n).$$

To prove Lemma 3.1, we need the fourth moment bounds on both $L(\Delta_{n-1},\zeta_n)$ and $\nabla f(x^*,\zeta_n)$. By our assumption, $\nabla f(x^*,\zeta_n)$ has the bounded fourth moment. Now we upper bound $\mathbb{E}_{n-1}\|L(\Delta_{n-1},\zeta_n)\|_2^m$ for $m = 2$ and $m = 4$. By the mean-value theorem and our assumption Lemma 3.1, for $m = 2$ and 4, we have

$$\mathbb{E}_{n-1}\|\nabla f(x_{n-1},\zeta_n) - \nabla f(x^*,\zeta_n)\|_2^m \leq \mathbb{E}_{n-1}\left[\sup_x \left\|\nabla^2 f(x,\zeta_n)\right\|^m\right] \|\Delta_{n-1}\|_2^m$$

$$\leq \|\Delta_{n-1}\|_2^m \mathbb{E}H(\zeta)^m.$$

By the convexity of $\|x\|_2^m$ and Jensen's inequality,

$$
\begin{aligned}
\|\nabla F(x_{n-1}) - \nabla F(x^*)\|_2^m &= \|\mathbb{E}_{n-1}\left(\nabla f(x_{n-1}, \zeta_n) - \nabla f(x^*, \zeta_n)\right)\|_2^m \\
&\leq \mathbb{E}_{n-1}\|\nabla f(x_{n-1}, \zeta_n) - \nabla f(x^*, \zeta_n)\|_2^m \\
&\leq \|\Delta_{n-1}\|_2^m \mathbb{E}H(\zeta)^m.
\end{aligned}
$$

By Lemma B.1, for $m = 2$ or 4,

$$
\mathbb{E}_{n-1}\|L(\Delta_{n-1}, \zeta_n)\|_2^m \leq 2^m \|\Delta_{n-1}\|_2^m \mathbb{E}H(\zeta)^m.
$$

Therefore, by the decomposition of ξ_n in (53),

$$
\begin{aligned}
\mathbb{E}_{n-1}\xi_n \xi_n^T = S &- \mathbb{E}_{n-1}\nabla f(x^*, \zeta_n)L(\Delta_{n-1}, \zeta_n)^T - \mathbb{E}_{n-1}L(\Delta_{n-1}, \zeta_n)\nabla f(x^*, \zeta_n)^T \\
&+ \mathbb{E}_{n-1}L(\Delta_{n-1}, \zeta_n)L(\Delta_{n-1}, \zeta_n)^T.
\end{aligned}
$$

By Jensen's and Cauchy-Schwartz inequality,

$$
\begin{aligned}
&\left\|\mathbb{E}_{n-1}\nabla f(x^*, \zeta_n)L(\Delta_{n-1}, \zeta_n)^T\right\| \\
&\leq \mathbb{E}_{n-1}\left\|\nabla f(x^*, \zeta_n)L(\Delta_{n-1}, \zeta_n)^T\right\| \\
&= \mathbb{E}_{n-1}\|\nabla f(x^*, \zeta_n)\|_2 \|L(\Delta_{n-1}, \zeta_n)^T\|_2 \\
&\leq [\mathbb{E}_{n-1}\|\nabla f(x^*, \zeta_n)\|_2^2 \mathbb{E}_{n-1}\|L(\Delta_{n-1}, \zeta_n)\|_2^2]^{\frac{1}{2}} \\
&\leq 2\sqrt{\mathbb{E}\|\nabla f(x^*, \zeta)\|_2^2 \mathbb{E}H(\zeta)^2}\|\Delta_{n-1}\|_2
\end{aligned}
$$

A similar bound holds for the trace, since

$$
\begin{aligned}
&\left|\text{tr}(\mathbb{E}_{n-1}\nabla f(x^*, \zeta_n)L(\Delta_{n-1}, \zeta_n)^T)\right| \\
&\leq \mathbb{E}_{n-1}\left|\text{tr}(\nabla f(x^*, \zeta_n)L(\Delta_{n-1}, \zeta_n)^T)\right| \\
&\leq \mathbb{E}_{n-1}\|\nabla f(x^*, \zeta_n)\|_2 \|L(\Delta_{n-1}, \zeta_n)^T\|_2.
\end{aligned}
$$

For the second order term, note that

$$
\begin{aligned}
&\left\|\mathbb{E}_{n-1}L(\Delta_{n-1}, \zeta_n)L(\Delta_{n-1}, \zeta_n)^T\right\| \\
&\leq \mathbb{E}_{n-1}\left\|L(\Delta_{n-1}, \zeta_n)L(\Delta_{n-1}, \zeta_n)^T\right\| = \mathbb{E}_{n-1}\|L(\Delta_{n-1}, \zeta_n)\|_2^2 \\
&\leq 4\mathbb{E}H(\zeta)^2\|\Delta_{n-1}\|_2^2.
\end{aligned}
$$

$$
\begin{aligned}
&\left|\text{tr}(\mathbb{E}_{n-1}L(\Delta_{n-1}, \zeta_n)L(\Delta_{n-1}, \zeta_n)^T)\right| \\
&\leq \mathbb{E}_{n-1}\left|\text{tr}(L(\Delta_{n-1}, \zeta_n)L(\Delta_{n-1}, \zeta_n)^T)\right| = \mathbb{E}_{n-1}\|L(\Delta_{n-1}, \zeta_n)\|_2^2.
\end{aligned}
$$

Combining the above inequalities, we arrive at Assumption 3.2 Claim 2.

Using the decomposition $\xi_n = L(\Delta_{n-1}, \zeta_n) - \nabla f(x^*, \zeta_n)$ and the fourth moment bounds on $L(\Delta_{n-1}, \zeta_n)$ and $\nabla f(x^*, \zeta_n)$, applying Lemma B.1 we get Claim 3. $\quad\square$

B.2. *Proof of Lemma 3.2.*

LEMMA 3.2 (Generalized version).   *Under Assumptions 3.1 and 3.2, if the step size is chosen to be $\eta_n = \eta n^{-\alpha}$ with $\alpha \in (0, 1)$, the iterates of error $\Delta_n = x_n - x^*$ satisfy the following.*

1) There exist universal constants $C \asymp C_d$ and n_0, such that for $n > m \geq n_0$, the following hold:

$$\mathbb{E}_m \|\Delta_n\|_2 \leq \exp\left(-\tfrac{1}{4}\mu \sum_{i=m}^n \eta_i\right) \|\Delta_m\|_2 + \sqrt{C} m^{-\alpha/2},$$

$$\mathbb{E}_m \|\Delta_n\|_2^2 \leq \exp\left(-\tfrac{1}{2}\mu \sum_{i=m}^n \eta_i\right) \|\Delta_m\|_2^2 + C m^{-\alpha},$$

$$\mathbb{E}_m \|\Delta_n\|_2^4 \leq \exp\left(-\tfrac{1}{2}\mu \sum_{i=m}^n \eta_i\right) \|\Delta_m\|_2^4 + C^2 m^{-2\alpha}.$$

2) As a consequence,

$$\mathbb{E}\|\Delta_n\|_2 \lesssim n^{-\alpha/2}(\sqrt{C} + \|\Delta_{n_0}\|_2),$$
$$\mathbb{E}\|\Delta_n\|_2^2 \lesssim n^{-\alpha}(C + \|\Delta_{n_0}\|_2^2),$$
$$\mathbb{E}\|\Delta_n\|_2^4 \lesssim n^{-2\alpha}(C^2 + \|\Delta_{n_0}\|_2^4).$$

To prove Lemma 3.2, we need a simple lemma.

LEMMA B.2.   *Let z_n be a sequence in $\mathbb{R}^+$ that satisfies the recursion*

$$z_n \leq (1 - \lambda \eta_n) z_{n-1} + D \eta_n^{2+\beta},$$

where $D, \lambda, \beta > 0$ are fixed constants, and the sequence η_n is decreasing. Then for any $m \leq n - 1$,

$$z_n \leq \exp(-\lambda(n - m)\eta_n) z_m + D \eta_m^{1+\beta} \lambda^{-1}.$$

PROOF. Using the recursion, we can find that

$$z_n \leq \prod_{k=m+1}^n (1 - \lambda \eta_k) z_m + D \sum_{k=m+1}^n \prod_{j=k+1}^n (1 - \lambda \eta_j) \eta_k^2.$$

Note that

$$\lambda \sum_{k=m+1}^n \prod_{j=k+1}^n (1 - \lambda \eta_j) \eta_k = 1 - \prod_{j=m+1}^n (1 - \lambda \eta_j) \leq 1.$$

Since $\eta_k \leq \eta_m$ for all $k \geq m + 1$,

$$D \sum_{k=m+1}^{n} \prod_{j=k+1}^{n} (1-\lambda\eta_j)\eta_k^{2+\beta} \leq D\lambda^{-1}\eta_m^{1+\beta}\lambda \sum_{k=m+1}^{n} \prod_{j=k+1}^{n} (1-\lambda\eta_j)\eta_k \leq D\lambda^{-1}\eta_m^{1+\beta}.$$

On the other hand, since $1 - x \leq \exp(-x)$ for $x \in [0, 1]$, so

$$\prod_{k=m+1}^{n} (1 - \lambda\eta_k) \leq \exp\left(-\lambda \sum_{k=m+1}^{n} \eta_k\right) \leq \exp(-\lambda(n - m)\eta_n).$$

Putting them back to the first inequality of this proof, we obtain our claim. $\qquad\square$

With Lemma B.2 in place, we are ready to prove Lemma 3.2.

PROOF OF LEMMA 3.2. From the recursion equation (5), we find that

$$\Delta_n = \Delta_{n-1} - \eta_n \nabla \widetilde{F}(\Delta_{n-1}) + \eta_n \xi_n,$$

where the shifted averaged loss function is $\widetilde{F}(\Delta) = F(\Delta + x^*)$. Therefore
(54)
$$\|\Delta_n\|_2^2 = \|\Delta_{n-1}\|_2^2 - 2\eta_n\langle\nabla\widetilde{F}(\Delta_{n-1}), \Delta_{n-1}\rangle + 2\eta_n\langle\xi_n, \Delta_{n-1}\rangle + \eta_n^2\|\xi_n - \nabla\widetilde{F}(\Delta_{n-1})\|_2^2.$$

The square of (54) is

$$\begin{aligned}
&\|\Delta_n\|_2^4 \\
=&\|\Delta_{n-1}\|_2^4 - 4\eta_n\langle\nabla\widetilde{F}(\Delta_{n-1}), \Delta_{n-1}\rangle\|\Delta_{n-1}\|_2^2 + 4\eta_n\langle\xi_n, \Delta_{n-1}\rangle\|\Delta_{n-1}\|_2^2 \\
&+4\eta_n^2\langle\nabla\widetilde{F}(\Delta_{n-1}) - \xi_n, \Delta_{n-1}\rangle^2 + 2\eta_n^2\|\Delta_{n-1}\|^2\|\xi_n - \nabla\widetilde{F}(\Delta_{n-1})\|_2^2 \\
&-4\eta_n^3\langle\nabla\widetilde{F}(\Delta_{n-1}) - \xi_n, \Delta_{n-1}\rangle\|\xi_n - \nabla\widetilde{F}(\Delta_{n-1})\|^2 + \eta_n^4\|\xi_n - \nabla\widetilde{F}(\Delta_{n-1})\|_2^4.
\end{aligned}$$

By the property of the inner product, the fourth term above is less than twice of the fifth, which can be bounded by the following using Young's inequality,

$$2\eta_n^2\|\Delta_{n-1}\|_2^2\|\xi_n - \nabla\widetilde{F}(\Delta_{n-1})\|_2^2 \leq \frac{1}{3}\mu\eta_n\|\Delta_{n-1}\|_2^4 + 3\mu^{-1}\eta_n^3\|\xi_n - \nabla\widetilde{F}(\Delta_{n-1})\|_2^4;$$

Moreover, Hölder's inequality gives the following bound for the sixth term:

$$-4\eta_n^3\langle\nabla\widetilde{F}(\Delta_{n-1}) - \xi_n, \Delta_{n-1}\rangle\|\xi_n - \nabla\widetilde{F}(\Delta_{n-1})\|_2^2 \leq \eta_n^3\|\Delta_{n-1}\|_2^4 + 3\eta_n^3\|\xi_n - \nabla\widetilde{F}(\Delta_{n-1})\|_2^4.$$

Combining all the inequality above back to the expansion of $\|\Delta_n\|_2^4$, we obtain

$$\|\Delta_n\|_2^4 \quad\leq\quad \|\Delta_{n-1}\|_2^4 - 4\eta_n\langle\nabla\widetilde{F}(\Delta_{n-1}), \Delta_{n-1}\rangle\|\Delta_{n-1}\|_2^2 + 4\eta_n\langle\xi_n, \Delta_{n-1}\rangle\|\Delta_{n-1}\|_2^2$$

$$(55) \quad +\mu\eta_n\|\Delta_{n-1}\|_2^4 + \eta_n^3\|\Delta_{n-1}\|_2^4 + \eta_n^3(9\mu^{-1} + \eta_n + 3)\|\xi_n - \nabla\widetilde{F}(\Delta_{n-1})\|_2^4.$$

Then using strong convexity of $\widetilde{F}$,

$$\langle\nabla\widetilde{F}(\Delta_{n-1}), \Delta_{n-1}\rangle \geq \widetilde{F}(\Delta_{n-1}) + \frac{\mu}{2}\|\Delta_{n-1}\|_2^2 \geq \frac{\mu}{2}\|\Delta_{n-1}\|_2^2.$$

By the fact that ξ_n is a martingale difference,

$$\mathbb{E}_{n-1}\langle\xi_n, \Delta_{n-1}\rangle\|\Delta_{n-1}\|_2^2 = 0.$$

Applying Young's inequality to Assumption 3.2 over ξ_n, using the Lipschitzness of ΔF in Assumption 3.1

$$\begin{aligned}
\mathbb{E}_{n-1}\|\xi_n - \nabla\widetilde{F}(\Delta_{n-1})\|_2^2 &\leq 2\mathbb{E}_{n-1}\|\xi_n\|_2^2 + 2\|\nabla\widetilde{F}(\Delta_{n-1})\|_2^2 \\
&\leq 2\mathrm{tr}(S) + 2\Sigma_2\|\Delta_{n-1}\|_2^2 + 2\Sigma_1\|\Delta_{n-1}\|_2 + 2L_F^2\|\Delta_{n-1}\|_2^2 \\
&\leq 2\mathrm{tr}(S) + \Sigma_1^{\frac{2}{3}} + (2\Sigma_2 + \Sigma_1^{\frac{4}{3}})\|\Delta_{n-1}\|_2^2 + 2L_F^2\|\Delta_{n-1}\|_2^2,
\end{aligned}$$

and likewise with Lemma B.1

$$\begin{aligned}
\mathbb{E}_{n-1}\|\xi_n - \nabla\widetilde{F}(\Delta_{n-1})\|_2^4 &\leq 4\mathbb{E}_{n-1}\|\xi_n\|_2^4 + 4\|\nabla\widetilde{F}(\Delta_{n-1})\|_2^4 \\
&\leq 4\Sigma_3 + 4(\Sigma_4 + L_F^4)\|\Delta_{n-1}\|_2^4.
\end{aligned}$$

Replacing propers terms in (54) and (55) with upper bounds above, we can find a constant $C_1 \asymp C_d$,

$$(56) \qquad \mathbb{E}_{n-1}\|\Delta_n\|_2^2 \leq (1 - \mu\eta_n + C_1^2\eta_n^2)\|\Delta_{n-1}\|_2^2 + C_1\eta_n^2.$$

$$(57) \qquad \mathbb{E}_{n-1}\|\Delta_n\|_2^4 \leq (1 - \mu\eta_n + C_1^4(\eta_n^3 + \eta_n^4))\|\Delta_{n-1}\|_2^4 + C_1^2\eta_n^3(\eta_n + 1).$$

Let

$$n_0 = \min\left\{n : C_1^2\eta_n^2 \leq \frac{1}{2}\mu\eta_n, C_1^4\eta_n^3(\eta_n + 1) \leq \frac{1}{2}\mu\eta_n, \mu\eta_n(1 - \alpha) \geq 8\alpha\log n\right\},$$

we have $n_0 \lesssim (C_d^2\eta)^{\frac{1}{\alpha}} \asymp 1$. Then for $n \geq n_0$, and a $C_2 \asymp C_1$, the inequalities can be simplified by

$$\begin{aligned}
\mathbb{E}_{n-1}\|\Delta_n\|_2^2 &\leq \left(1 - \frac{1}{2}\mu\eta_n\right)\|\Delta_{n-1}\|_2^2 + C_2\eta_n^2, \\
\mathbb{E}_{n-1}\|\Delta_n\|_2^4 &\leq \left(1 - \frac{1}{2}\mu\eta_n\right)\|\Delta_{n-1}\|_2^4 + C_2^2\eta_n^3.
\end{aligned}$$

Then Lemma B.2 produces claim 1) for the second and forth moment, while we use Cauchy-Schwartz inequality it gives the bound for the first moment, as

$$\mathbb{E}_m \|\Delta_n\|_2 \leq \sqrt{\mathbb{E}_m \|\Delta_n\|_2^2}.$$

To see the claim 2), note that

$$\max\{C_1^2 \eta_i^2, C_1^4(\eta_i^3 + \eta_i^4)\} \leq C_1^4 \eta_i^2(\eta_i + 1)^2,$$

therefore applying discrete Gronwall's inequality to (56) and (57), we find that

$$\mathbb{E}\|\Delta_{n_0}\|_2^2 \leq C_3(1 + \|\Delta_0\|_2^2), \quad \mathbb{E}\|\Delta_{n_0}\|_2^4 \leq C_3(1 + \|\Delta_0\|_2^4),$$

where

$$C_3 = \left(\prod_{i=1}^{n_0} (1 + C_1^4 \eta_i^2(\eta_i + 1)^2) \right) \left(C_1^4 \sum_{i=1}^{n_0} \eta_i^2(\eta_i + 1)^2 \right) \asymp 1.$$

Recall that we assume $C_d^2 \eta_i \lesssim 1$, so $n_0 \asymp 1$. Then for $n \geq 2n_0$, using the claims from 1), there is a constant $C_4 \asymp C_1$ such that

$$\mathbb{E}_{n_0}\|\Delta_{\frac{n}{2}}\|_2^2 \leq \|\Delta_{n_0}\|_2^2 + C_4 \eta_{n_0}, \quad \mathbb{E}_{n_0}\|\Delta_{\frac{n}{2}}\|_2^4 \leq \|\Delta_{n_0}\|_2^4 + C_4^2 \eta_{n_0}^2,$$

and

$$\mathbb{E}_{\frac{n}{2}}\|\Delta_n\|_2^2 \leq \exp\left(-\tfrac{1}{4}n\mu\eta_n\right) \|\Delta_{\frac{n}{2}}\|_2^2 + C_4(\tfrac{n}{2})^{-\alpha},$$
$$\mathbb{E}_{\frac{n}{2}}\|\Delta_n\|_2^4 \leq \exp\left(-\tfrac{1}{2}n\mu\eta_n\right) \|\Delta_{\frac{n}{2}}\|_2^4 + C_4^2(\tfrac{n}{2})^{-2\alpha}.$$

So using the law of iterated expectation, and that

$$\exp(-\tfrac{1}{4}n\mu\eta_n) \leq \exp(-\tfrac{1}{4}\mu\eta n^{1-\alpha}) \leq n^{-2\alpha} \leq n^{-\alpha}$$

there is a constant $C_5 \asymp 1$ that

$$\mathbb{E}\|\Delta_n\|_2^2 \leq n^{-\alpha} C_5 \left(\|\Delta_0\|_2^2 + C_4\right), \quad \mathbb{E}\|\Delta_n\|_2^4 \leq n^{-2\alpha} C_5 \left(\|\Delta_0\|_2^4 + C_4^2\right).$$

Applying Cauchy-Schwartz inequality again yields the claim for the first moment. □

B.3. *Lemma B.3 and its proof.*   Before we move on, let us apply the same arguments and prove a similar simpler result for the linear oracle sequence U_n.

LEMMA B.3.   *For the linear oracle sequence U_n defined by (20), it satisfies the following:*

$$\mathbb{E}\|U_n\|_2^2 \lesssim n^{-\alpha}(C_d + \|U_0\|_2^2 + \|\Delta_0\|_2^2).$$

PROOF.   Clearly, we have the following recursion

$$\|U_n\|_2^2 = \|U_{n-1}\|_2^2 - 2\eta_n\langle AU_{n-1}, U_{n-1}\rangle + 2\eta_n\langle \xi_n, U_{n-1}\rangle + \eta_n^2\|\xi_n - AU_{n-1}\|_2^2.$$

Note that

$$\langle AU_{n-1}, U_{n-1}\rangle \geq \frac{\mu}{2}\|U_{n-1}\|_2^2,$$

and also that ξ_n is a martingale difference, so

$$\mathbb{E}_{n-1}\langle \xi_n, U_{n-1}\rangle = 0.$$

Moreover by Assumption 3.2, there is a constant $C_1 \asymp C_d$ such that

$$\begin{aligned}
\mathbb{E}_{n-1}\|\xi_n - AU_{n-1}\|_2^2 &\leq 2\mathbb{E}_{n-1}\|\xi_n\|_2^2 + 2\|A\|^2\|U_{n-1}\|_2^2 \\
&\leq 2(\mathrm{tr}(S) + \Sigma_1\|\Delta_{n-1}\|_2 + \Sigma_2\|\Delta_{n-1}\|_2^2) + 2L_F^2\|U_{n-1}\|_2^2 \\
&\leq C_1 + C_1^2(\|\Delta_{n-1}\|_2^2 + \|U_{n-1}\|_2^2).
\end{aligned}$$

Here we have used that ∇F being Lipschitz implies $\|A\| = \|\nabla^2 F(x^*)\| \leq L_F$.
We find

$$\mathbb{E}_{n-1}\|U_n\|_2^2 \leq (1 - \mu\eta_n + C_1^2\eta_n^2)\|U_{n-1}\|_2^2 + C_1(\eta_n^2 + C_1\|\Delta_{n-1}\|_2^2).$$

Summing this inequality with (56) yields the following with $z_n := \|U_n\|_2^2 + \|\Delta_n\|_2^2$ and $C_2 = 2C_1$

$$\mathbb{E}_{n-1}z_n \leq (1 - \mu\eta_n + C_2^2\eta_n^2)z_{n-1} + C_2\eta_n^2.$$

Then redoing the analysis in the proof of Lemma 3.2 after (56) yields that

$$\mathbb{E}\|U_n\|_2^2 \leq \mathbb{E}z_n \lesssim n^{-\alpha}(C_2 + \|\Delta_0\|_2^2 + \|U_0\|_2^2).$$

$\square$

C. Consistency Proofs of the Plug-in Estimator in Section 4.1.

C.1. *Proof of Lemma 4.1 for the consistency of A_n and S_n.*

LEMMA 4.1. *Under Assumptions 3.1, 3.2 and 4.1, the followings hold*

$$\mathbb{E}\|A_n - A\| \lesssim C_d^2 n^{-\frac{\alpha}{2}}, \quad \mathbb{E}\|S_n - S\| \lesssim C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha},$$

where α is given in the step size sequence $\eta_i = \eta i^{-\alpha}$, $i = 1, 2, \ldots, n$.

PROOF. We decompose $A_n - A$ into the following terms:

$$
\begin{aligned}
A_n - A &= \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(x_{i-1}, \zeta_i) - A \\
(58) \qquad &= \left(\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(x^*, \zeta_i) - A\right) + \frac{1}{n}\sum_{i=1}^{n} \left(\nabla^2 f(x_{i-1}, \zeta_i) - \nabla^2 f(x^*, \zeta_i)\right).
\end{aligned}
$$

The first part can be bounded using the second moment bound, since by independence of ζ_i and ζ_j,

$$\mathbb{E}(\nabla^2 f(x^*, \zeta_i) - A)(\nabla^2 f(x^*, \zeta_j) - A)^T = \mathbb{1}_{i=j}[\mathbb{E}[\nabla^2 f(x^*, \zeta_i)][\nabla^2 f(x^*, \zeta_i)]^T - AA^T].$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(x^*, \zeta_i) - A\right)&\left(\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(x^*, \zeta_i) - A\right)^T \\
&= \frac{1}{n^2}\sum_{i=1}^{n}[\mathbb{E}[\nabla^2 f(x^*, \zeta_i)][\nabla^2 f(x^*, \zeta_i)]^T - AA^T]
\end{aligned}
$$

which has its norm bounded by $\frac{1}{n}L_4$. Note that by Cauchy-Schwartz inequality, for any matrix B

$$\mathbb{E}\left\|BB^T\right\| = \mathbb{E}\|B\|^2 \geq [\mathbb{E}\|B\|]^2.$$

Therefore we have

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(x^*, \zeta_i) - A\right\| \leq \frac{\sqrt{L_4}}{\sqrt{n}} \lesssim C_d n^{-\frac{1}{2}}.$$

The second term in (58) can be bounded using Assumption 4.1,

$$
\begin{aligned}
&\left\|\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(\nabla^2 f(x_{i-1}, \zeta_i) - \nabla^2 f(x^*, \zeta_i))\right\| \\
&\leq \frac{1}{n}\mathbb{E}\sum_{i=1}^{n}\left\|\nabla^2 f(x_{i-1}, \zeta_i) - \nabla^2 f(x^*, \zeta_i)\right\| \leq \frac{L_2}{n}\sum_{i=1}^{n}\mathbb{E}\|x_i - x^*\|_2.
\end{aligned}
$$

Using Lemma 3.2 2), $\mathbb{E}\|x_i - x^*\| \lesssim \sqrt{C_d}i^{-\frac{\alpha}{2}}$. Note that by Lemma D.1 5) in below

$$\sum_{i=1}^{n} i^{-\frac{\alpha}{2}} \lesssim n^{1-\frac{\alpha}{2}}.$$

So the second term in (58) is of order $C_d^2 n^{-\frac{\alpha}{2}}$, which concludes the consistency proof for A_n.

For S_n, we decompose

$$\nabla f(x_{i-1}, \zeta_i) = \nabla f(x^*, \zeta_i) + (\nabla f(x_{i-1}, \zeta_i) - \nabla f(x^*, \zeta_i)) := X_i + Y_i,$$

then

$$S_n - S = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T - S\right) + \frac{1}{n}\sum_{i=1}^{n} X_i Y_i^T + \frac{1}{n}\sum_{i=1}^{n} Y_i X_i^T + \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^T.$$

Note that
$$X_i = \nabla f(x^*, \zeta_i) = \nabla f(x^*, \zeta_i) - \nabla F(x^*)$$

which is ξ_n when $\Delta_{n-1} = 0$, so by part 3 of Assumption 3.2, we know the fourth moment of $\|X_i\|$ is bounded by Σ_3. Then by $\mathbb{E}X_i X_i^T = S$ and the independence between the X_i's,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T - S\right\|^2 \leq \operatorname{tr}\left(\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T - S\right)^2\right) = \frac{1}{n}\mathbb{E}\operatorname{tr}((X_i X_i^T)^2 - S^2).$$

Note that $\mathbb{E}\operatorname{tr}(X_i X_i^T)^2 = \mathbb{E}(\operatorname{tr}(X_i X_i^T))^2 \leq \Sigma_3$. By Cauchy-Schwartz inequality,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T - S\right\| \lesssim \sqrt{\Sigma_3} n^{-\frac{1}{2}}.$$

Further, by Assumption 4.1 and Lemma 3.2,

$$\mathbb{E}\left\|Y_i Y_i^T\right\| \leq L_F^2 \mathbb{E}\left\|\Delta_{i-1}\right\|^2 \lesssim C_d^3 i^{-\alpha}$$
$$\mathbb{E}\left\|X_i X_i^T\right\| \lesssim C_d$$
$$\mathbb{E}\left\|X_i Y_i^T\right\| \leq [\mathbb{E}\left\|X_i\right\|^2 \mathbb{E}\left\|Y_i\right\|^2]^{\frac{1}{2}} \lesssim C_d^2 i^{-\frac{\alpha}{2}}.$$

Summing up all the terms, we obtain $\mathbb{E}\|S_n - S\| \lesssim C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha}$. $\square$

C.2. *Proof of Theorem 4.2 for consistency of the plug-in estimator.* Before we provide the proof the consistency result of plug-in estimator in Theorem 4.2, we first provide the following lemma on matrix perturbation inequality for the inverse of a matrix.

LEMMA C.1 (Matrix perturbation inequality for the inverse). *Let* $B = A + E$, *where* A *and* B *are assumed to be invertible, and* $\left\| A^{-1}E \right\| < \frac{1}{2}$. *We have*

$$(59) \qquad \left\| B^{-1} - A^{-1} \right\| \leq 2 \left\| E \right\| \left\| A^{-1} \right\|^2$$

PROOF. Using (Chang, 2006, Eq. 3), we have

$$B^{-1} - A^{-1} = -A^{-1}E(I + A^{-1}E)^{-1}A^{-1},$$

which implies that

$$\begin{aligned}
\left\| B^{-1} - A^{-1} \right\| &= \left\| A^{-1}E(I + A^{-1}E)^{-1}A^{-1} \right\| \\
&\leq \left\| A^{-1} \right\|^2 \left\| E \right\| \left\| (I + A^{-1}E)^{-1} \right\| \\
&\leq \left\| A^{-1} \right\|^2 \left\| E \right\| \frac{1}{\lambda_{\min}(I + A^{-1}E)} \\
&\leq \left\| A^{-1} \right\|^2 \left\| E \right\| \frac{1}{1 - \left\| A^{-1}E \right\|}
\end{aligned}$$

Here $\lambda_{\min}$ is the minimum eigenvalue, and the last inequality uses Weyl's inequality $\lambda_{\min}(A + B) \geq \lambda_{\min}(A) - \|B\|$. The final result in (59) follows the assumption that $\left\| A^{-1}E \right\| < \frac{1}{2}$. □

THEOREM 4.2. *Under Assumptions 3.1, 3.2 and 4.1, the thresholded plug-in estimator initialized from any bounded x_0 converges to the asymptotic covariance matrix,*

$$(11) : \mathbb{E} \left\| \widetilde{A}_n^{-1} S_n \widetilde{A}_n^{-1} - A^{-1}SA^{-1} \right\| \lesssim \|S\|(C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha}),$$

where $\alpha \in (0,1)$ is given in the step size sequence $\eta_i = \eta i^{-\alpha}$, $i = 1, 2, \ldots, n$. When C_d is a constant, the right hand side of (11) is dominated by $O(n^{-\frac{\alpha}{2}})$.

PROOF. Let us define

$$E_A = \widetilde{A}_n - A, \quad F_A = \widetilde{A}_n^{-1} - A^{-1}, \quad F_S = S_n - S,$$

Note that by Lemma 4.1, $\mathbb{E} \|F_S\| \lesssim C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha}$ while by Markov inequality,

$$
\begin{aligned}
\mathbb{E} \|E_A\| &\leq \mathbb{E} \|A_n - A\| + \|A\| \, \mathbb{P}(A_n \prec \delta I) \\
&\lesssim \mathbb{E} \|A_n - A\| + \mathbb{P}(\|A_n - A\| \geq \lambda_{\min}(A) - \delta) \\
&\leq \mathbb{E} \|A_n - A\| + \frac{1}{\lambda_{\min}(A) - \delta} \mathbb{E} \|A_n - A\| \lesssim C_d^2 n^{-\alpha/2}.
\end{aligned}
$$

Then Markov inequality yields further that

$$
\mathbb{P}(\|A^{-1}E_A\| \geq 1/2) \leq 2(\mathbb{E} \|E_A\|) \lesssim C_d^2 n^{-\alpha/2}.
$$

Note that by construction, the following hold almost surely

$$
\left\| \widetilde{A}_n^{-1} \right\| \leq \delta^{-1}, \quad \|A^{-1}\| \leq \lambda_{\min}^{-1}(A), \quad \|F_A\| \leq \delta^{-1} + \lambda_{\min}^{-1}(A).
$$

By Lemma C.1,

$$
\|F_A\| \leq \mathbb{1}_{\|A^{-1}E_A\| \leq 1/2} 2 \|E_A\| \left\| A^{-1} \right\|^2 + \mathbb{1}_{\|A^{-1}E_A\| \geq 1/2} (\delta^{-1} + \lambda_{\min}^{-1}(A)),
$$

therefore

$$
\mathbb{E} \|F_A\| \leq 2 \left\| A^{-1} \right\|^2 \mathbb{E} \|E_A\| + (\delta^{-1} + \lambda_{\min}^{-1}(A)) \mathbb{P}(\|A^{-1}E_A\| \geq 1/2) \lesssim C_d^2 n^{-\frac{\alpha}{2}}.
$$

Finally, we can decompose the error in the estimation into the following

$$
\begin{aligned}
\widetilde{A}_n^{-1} S_n \widetilde{A}_n^{-1} - A^{-1} S A^{-1} &= (A^{-1} + F_A)(S + F_S)(A^{-1} + F_A) - A^{-1} S A^{-1}, \\
&= (A^{-1} + F_A) F_S (A^{-1} + F_A) + A^{-1} S F_A + F_A S A^{-1} + F_A S F_A.
\end{aligned}
$$

Note that $\|F_A\| \leq \delta^{-1} + \mu^{-1} \asymp 1$, so

$$
\mathbb{E}\|(A^{-1} + F_A) F_S (A^{-1} + F_A)\| \lesssim \mathbb{E}\|F_S\| \lesssim C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha},
$$

and

$$
\|A^{-1} S F_A\| + \|F_A S A^{-1}\| + \|F_A S F_A\| \lesssim \|S\|(C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha}).
$$

In their summation, we have our claim. $\qquad\square$

D. Consistency proof of the batch-means estimator.

D.1. *Technical Lemmas.* As a preparation for the consistency result, we introduce the following technical lemmas. Some variants of them have appeared in Polyak and Juditsky (1992) or Bach and Moulines (2011). Since we need non-asymptotic bounds to investigate the relationship between the estimation error and M, N, we redo some key steps in Polyak and Juditsky (1992); Bach and Moulines (2011) with simpler notations.

Let us start with some approximations of various quantities related to the number of batches and batch sizes. These will simplify our discussion afterwards.

LEMMA D.1. *Under the formulation of* (16), *the followings hold:*

1) The starting and ending indices of each non starting $(k \neq 0)$ batches, and their neighbors $s_k - 1, s_k$ and $e_k, e_k + 1$ are of the same order,

$$s_k - 1 \asymp s_k \asymp e_k \asymp e_k + 1 \asymp (kN)^{\frac{1}{1-\alpha}}.$$

2) The total number of iterations used in our estimator can be estimated by:

$$S_M \asymp e_M \asymp (MN)^{\frac{1}{1-\alpha}}.$$

3) The number of iterations in each batch can be estimated by:

$$n_k = e_k - e_{k-1} \asymp k^{\frac{\alpha}{1-\alpha}} N^{\frac{1}{1-\alpha}}.$$

4) The sum of stepsize in each batch, which represents the log of decorrelation across that batch, is of order

$$\sum_{i=s_k}^{e_k} \eta_i \geq \eta_{e_k} n_k \gtrsim (kN)^{-\frac{\alpha}{1-\alpha}} k^{\frac{\alpha}{1-\alpha}} N^{\frac{1}{1-\alpha}} = N.$$

5) For any fixed $\gamma > 0$, the following estimate holds for sum of geometric sequence

$$M^{-1} \sum_{k=1}^{M} k^{-\gamma} \lesssim M^{-1} \int_{1}^{M+1} x^{-\gamma} dx \lesssim M^{(1-\gamma)^+ - 1}.$$

PROOF. The first two inequalities comes from direct verifications, since $\frac{k+1}{k} \leq 2$ for $k \geq 1$. The third one comes from the fact that

$$e_k - e_{k-1} = [(k+1)^{\frac{1}{1-\alpha}} - k^{\frac{1}{1-\alpha}}]N^{\frac{1}{1-\alpha}} = \frac{N^{\frac{1}{1-\alpha}}}{1-\alpha} \int_{k}^{k+1} x^{\frac{\alpha}{1-\alpha}} dx,$$

and then observe that $k^{\frac{\alpha}{1-\alpha}} \leq x^{\frac{\alpha}{1-\alpha}} \leq 2^{\frac{\alpha}{1-\alpha}} k^{\frac{\alpha}{1-\alpha}}$. The fourth one is a corollary of the first and third. The last inequality comes from that $k + 1 \leq 2k$ for $k \geq 1$ again. □

Recall the linear oracle sequence defined by (20): $U_n = (I - \eta_n A)\Delta_{n-1} + \eta_n \xi_n$. We now introduce Lemma D.2.

LEMMA D.2.   *For each positive integer j, let Y_j^k be a matrix sequence with $Y_j^j = I$ and for $k \geq j+1$*

$$Y_j^k = (I - \eta_k A)Y_j^{k-1} = \prod_{i=j+1}^{k} (I - \eta_i A),$$

where $\eta_k = \eta k^{-\alpha}$ with $\alpha \in (1/2, 1)$ and A is a PSD matrix with $\lambda_A > 0$ be any lower bound on the minimum eigenvalue of A. Then

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

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

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

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

3) When $s_1 \leq j < k \leq e_M$, for sufficiently large N, there exists a constant c such that

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

as long as $M \leq N^m$ for certain positive integer m.

PROOF.   Assume $Q\Lambda Q^T$ is the eigenvalue decomposition of A, then

$$\prod_{i=j+1}^{k} (I - \eta_i A) = Q \prod_{i=j+1}^{k} (I - \eta_i \Lambda)Q^T$$

which immediately gives us the claim 1) since $1 - x \leq \exp(-x)$ for $x \in [0,1]$. The first part of 2) comes as an immediate consequence since:

$$\|S_j^k\| \leq \sum_{i=j+1}^{k} \|Y_j^i\| \leq \sum_{i=j+1}^{k} \exp(-\lambda_A(i-j)\eta_k) \leq \lambda_A^{-1}\eta_k^{-1},$$

while $\eta_k \asymp k^{-\alpha}$. To show the second part of 2), we sum the identity $Y_j^i = Y_j^{i-1} - \eta_k A Y_j^{i-1}$ from $i = j+1$ to $i = k+1$, which gives us:

$$Y_j^{k+1} = I - \sum_{i=j}^{k} \eta_k A Y_j^i.$$

This can be rewritten as

$$\sum_{i=j}^{k} \eta_i Y_j^i = A^{-1} - A^{-1} Y_j^{k+1}.$$

Therefore,

$$\left\| Z_j^k \right\| \leq \left\| \eta_i S_i^n - \sum_{i=j}^{k} \eta_i Y_j^i \right\| + \left\| \sum_{i=j}^{k} \eta_i Y_j^i - A^{-1} \right\|$$

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

In order for us to have claim 2), it suffices to bound the first part. Notice that for $j \leq m \leq i$,

$$\eta_{m-1} - \eta_m = \eta((m-1)^{-\alpha} - m^{-\alpha}) = \frac{\eta}{\alpha} \int_{m-1}^{m} x^{-\alpha-1} dx$$

$$\leq \frac{2^{\alpha+1}\eta}{\alpha} m^{-\alpha-1} \leq 4(\eta\alpha)^{-1} \eta_m \eta_{i+1} (i+1)^\alpha j^{-1},$$

therefore

$$\eta_j - \eta_i = \sum_{m=j+1}^{i} (\eta_{m-1} - \eta_m) \leq \sum_{m=j+1}^{i} 4(\eta\alpha)^{-1} \eta_m \eta_i (i+1)^\alpha j^{-1}$$

$$\leq 4(\eta\alpha)^{-1}(i+1)^\alpha j^{-1} \eta_{i+1} \sum_{m=j+1}^{i} \eta_m.$$

Using $S_j^k = \sum_{i=j+1}^k Y_j^i$,

$$\left\| \eta_j S_j^k - \sum_{i=j}^k \eta_i Y_j^i \right\|$$

$$\leq \sum_{i=j}^k (\eta_j - \eta_i) \left\| Y_j^i \right\| + \eta_j$$

$$\leq 4(\eta\alpha)^{-1} k^\alpha j^{-1} \sum_{i=j}^k \eta_{i+1} \left(\sum_{m=j+1}^i \eta_m \right) \exp\left(-\lambda_A \sum_{m=j+1}^i \eta_m \right) + \eta j^{-\alpha}$$

Note that $xe^{-\lambda_A x} \leq e^{\lambda_A} y e^{-\lambda_A y}$ when $x \leq y \leq x+1$, so

$$\eta_{i+1} \left(\sum_{m=j+1}^i \eta_m \right) \exp\left(-\lambda_A \sum_{m=j+1}^i \eta_m \right)$$

$$= \int_{\sum_{m=j+1}^i \eta_m}^{\sum_{m=j+1}^{i+1} \eta_m} \left(\sum_{m=j+1}^i \eta_m \right) \exp\left(-\lambda_A \sum_{m=j+1}^i \eta_m \right)$$

$$\leq e^{\lambda_A} \int_{\sum_{m=j+1}^i \eta_m}^{\sum_{m=j+1}^{i+1} \eta_m} y e^{-\lambda_A y} dx.$$

So

$$\sum_{i=j}^k \eta_{i+1} \left(\sum_{m=j+1}^i \eta_m \right) \exp\left(-\lambda_A \sum_{m=j+1}^i \eta_m \right) \leq e^{\lambda_A} \int_0^\infty x e^{-\lambda_A x} dx,$$

which is a constant. Note that $j^{-\alpha} \leq j^{\alpha-1} \leq k^\alpha j^{-1}$, therefore

$$\left\| Z_j^k \right\| \leq \left\| \eta_i S_i^n - \sum_{i=j}^k \eta_i Y_j^i \right\| + \left\| \sum_{i=j}^k \eta_i Y_j^i - A^{-1} \right\| \lesssim k^\alpha j^{-1} + \exp\left(-\lambda_A \sum_{i=j+1}^k \eta_i \right).$$

For claim 3), when j and k are in neighboring batches, the second claim already produces the results, since the k/j is bounded by $3^{\frac{1}{1-\alpha}}$. Otherwise there is a batch between j, k, that is for an m the following holds,

$$j \leq e_{m-1} \leq e_m \leq k.$$

Note that

$$S_j^k = S_j^{e_m} + Y_j^{e_m} S_{e_m}^k$$

so
$$Z_j^k = \eta_j S_j^k - A^{-1} = Z_j^{e_m} + Y_{e_m}^k S_{e_m}^k$$

Then because e_m and j are in neighboring batches, by Lemma D.1 claim 3, there is a c such that

$$\left\| Z_j^{e_m} \right\| \lesssim K_1 j^{\alpha-1} + \exp\left(-\lambda_A \sum_{i=j+1}^{e_m} \eta_i \right) \le j^{\alpha-1} + \exp(-c\lambda_A N).$$

Because e_m is at least one batch away from j, so by the first claim of this lemma and Lemma D.1 4)

$$\left\| Y_j^{e_m} \right\| \le \exp\left(-\lambda_A \sum_{i=s_m}^{e_m} \eta_i \right) \le \exp(-\lambda_A c N),$$

while by the second claim,

$$\left\| S_{e_m}^k \right\| \lesssim e_M^\alpha \asymp (NM)^{\frac{\alpha}{1-\alpha}}.$$

Then because M is of polynomial order of N, so for N large enough,

$$\| Y_{e_m}^k S_{e_m}^k \| \lesssim \eta^\alpha (NM)^{\frac{\alpha}{1-\alpha}} \exp(-\lambda_A c N) \le e_M^{\alpha-1} \le j^{\alpha-1}$$

Therefore,

$$\left\| Z_j^k \right\| \lesssim j^{\alpha-1} + \exp(-c\lambda_A N) \asymp j^{\alpha-1}.$$

$\square$

D.2. *A linear framework.* This section is devoted to the linear framework illustrated by Lemma 4.7. As we will see, the following two quantities

$$\widehat{S} := M^{-1} \sum_{k=1}^{M} n_k^{-1} \left(\sum_{i=s_k}^{e_k} \xi_i \right) \left(\sum_{i=s_k}^{e_k} \xi_i \right)^T$$

and $M^{-1} \sum_{k=1}^{M} n_k \overline{U}_{n_k} \overline{U}_{n_k}^T$ converge to the asymptotic covariance $A^{-1}SA^{-1}$. The strategy of the proof below is first showing $\widehat{S}$ converges to the right covariance, then $M^{-1} \sum_{k=1}^{M} n_k \overline{U}_{n_k} \overline{U}_{n_k}^T$ is not far from the $\widehat{S}$, and eventually reaches Lemma 4.7.

LEMMA D.3. *Given two symmetric $d \times d$ matrices S and Σ, suppose S is also positive semidefinite, then*

$$tr(S\Sigma) \le tr(S) \left\| \Sigma \right\|.$$

PROOF. Suppose the eigenvalue decomposition of Σ is $\Phi D \Phi^T$. Suppose also that $Q^T \Lambda Q$ is the eigenvalue decomposition of $\Phi^T S \Phi$. Then

$$\operatorname{tr}(S\Sigma) = \operatorname{tr}(\Phi^T S \Phi D) = \operatorname{tr}(Q^T \Lambda Q D) = \operatorname{tr}(\Lambda^{1/2} Q D Q^T \Lambda^{1/2}).$$

Let e_i be the i-th standard euclidean basis vector, then

$$e_i^T \Lambda^{1/2} Q D Q^T \Lambda^{1/2} e_i = \lambda_i e_i^T Q D Q^T e_i \le \lambda_i \|D\| \|Q^T e_i\|^2 = \lambda_i \|\Sigma\|.$$

Summing over all i yields our claim.

$\square$

LEMMA D.4. *Under Assumptions 3.1 and 3.2, the following*

$$(60) \qquad \widehat{S} := M^{-1} \sum_{k=1}^{M} n_k^{-1} \left(\sum_{i=s_k}^{e_k} \xi_i \right) \left(\sum_{i=s_k}^{e_k} \xi_i \right)^T$$

is a consistent estimator of S, namely

$$\mathbb{E} \|\widehat{S} - S\|^2 \lesssim \left(C_d^3 N^{-\frac{\alpha}{2(1-\alpha)}} M^{-\beta_\alpha} + C_d^2 M^{-1} \right)$$

with $\beta_\alpha := 1 - \left(1 - \frac{\alpha}{2-2\alpha} \right)^+$.

PROOF. Since

$$\mathbb{E} \left\| (\widehat{S} - S)^2 \right\| \le \mathbb{E} \operatorname{tr}(\widehat{S} - S)^2 = \operatorname{tr} \mathbb{E} (\widehat{S} - S)^2,$$

where $\operatorname{tr}(\cdot)$ denotes the trace of a matrix, it suffices to show $\operatorname{tr} \mathbb{E} (\widehat{S} - S)^2$ follows the claimed bound. Notice that

$$\mathbb{E} \widehat{S} = M^{-1} \sum_{k=1}^{M} n_k^{-1} \sum_{i=s_k}^{e_k} \mathbb{E} \xi_i \xi_i^T = M^{-1} \sum_{k=1}^{M} n_k^{-1} \sum_{i=s_k}^{e_k} (S + \mathbb{E} \Sigma(\Delta_{i-1}))$$

So
(61)
$$\mathbb{E}(\widehat{S} - S)^2 = \mathbb{E} \widehat{S}^2 - S^2 + M^{-1} S \sum_{k=1}^{M} n_k^{-1} \sum_{i=s_k}^{e_k} \mathbb{E} \Sigma(\Delta_{i-1}) + M^{-1} \sum_{k=1}^{M} n_k^{-1} \sum_{i=s_k}^{e_k} \mathbb{E} \Sigma(\Delta_{i-1}) S.$$

The third term can be bounded by the following

$$\left| \mathrm{tr}(M^{-1}S \sum_{k=1}^{M} n_k^{-1} \sum_{i=s_k}^{e_k} \mathbb{E}\Sigma(\Delta_{i-1})) \right|$$

$$\leq M^{-1} |\mathrm{tr}(S)| \sum_{k=1}^{M} n_k^{-1} \sum_{i=s_k}^{e_k} \mathbb{E}(\Sigma_1 \|\Delta_{i-1}\|_2 + \Sigma_2 \|\Delta_{i-1}\|_2^2)$$

$$\overset{(a)}{\lesssim} \mathrm{tr}(S) M^{-1} \sum_{k=1}^{M} (s_k^{-\alpha/2} \Sigma_1 \sqrt{C_d} + s_k^{-\alpha} \Sigma_2 C_d)$$

$$\lesssim C_d^3 M^{-1} \sum_{k=1}^{M} s_k^{-\alpha/2} + C_d s_k^{-\alpha}$$

$$\lesssim C_d^3 M^{-1} \sum_{k=1}^{M} s_k^{-\alpha/2} \overset{(b)}{\lesssim} C_d^3 N^{-\frac{\alpha}{2(1-\alpha)}} M^{[1-\frac{\alpha}{2(1-\alpha)}]^+ - 1}.$$

where (a) uses Lemma 3.2 and (b) uses Lemma D.1 1) and 5). Likewise, we can bound the fourth term in (61) as well. For the computation of $\mathbb{E}\widehat{S}^2$, which is

$$\mathbb{E}\widehat{S}^2 = M^{-2} \sum_{j=1}^{M} \sum_{k=1}^{M} n_j^{-1} n_k^{-1} \sum_{i_1,i_2=s_j}^{e_j} \sum_{i_3,i_4=s_k}^{e_k} \mathbb{E}\xi_{i_1}\xi_{i_2}^T\xi_{i_3}\xi_{i_4}^T,$$

using the marginal difference nature of ξ_i, one can see that the terms are 0 unless indices come in pairs, i.e. $i_1 = i_2, i_3 = i_4$ or $i_1 = i_3, i_2 = i_4$ or $i_1 = i_4, i_2 = i_3$. Also notice that the latter two only happens when all the indices are in the same batch. So

$$\mathbb{E}\widehat{S}^2 = \frac{1}{M^2} \left[\sum_{j=1}^{M} \sum_{k=1}^{M} n_j^{-1} n_k^{-1} \sum_{i_1=s_j}^{e_j} \sum_{i_3=s_k}^{e_k} \mathbb{E}\xi_{i_1}\xi_{i_1}^T\xi_{i_3}\xi_{i_3}^T \right.$$

$$\left. + \sum_{j=1}^{M} n_j^{-2} \sum_{i_1,i_3=s_j, i_1 \neq i_3}^{e_j} \mathbb{E}\xi_{i_1}\xi_{i_3}^T\xi_{i_1}\xi_{i_3}^T + \mathbb{E}\xi_{i_1}\xi_{i_3}^T\xi_{i_3}\xi_{i_1}^T \right],$$

Then using that

$$S^2 = M^{-2} \sum_{j,k=1}^{M} n_j^{-1} n_k^{-1} \sum_{i_1=s_j}^{e_j} \sum_{i_3=s_k}^{e_k} S^2$$

we have

$$
\begin{aligned}
\mathrm{tr}(\mathbb{E}\widehat{S}^2 - S^2) =& M^{-2} \sum_{j=1}^{M} \sum_{k=1}^{M} n_j^{-1} n_k^{-1} \sum_{i_1=s_j}^{e_j} \sum_{i_3=s_k}^{e_k} \mathrm{tr}(\mathbb{E}\xi_{i_1}\xi_{i_1}^T \xi_{i_3}\xi_{i_3}^T - S^2) \\
& + M^{-2} \sum_{j=1}^{M} n_j^{-2} \sum_{i_1,i_2=s_j}^{e_j} \mathrm{tr}(\mathbb{E}\xi_{i_1}\xi_{i_2}^T \xi_{i_1}\xi_{i_2}^T) + \mathrm{tr}(\mathbb{E}\xi_{i_1}\xi_{i_2}^T \xi_{i_2}\xi_{i_1}^T).
\end{aligned}
$$

(62)

First, we setup a upper bound of $|\mathrm{tr}(\mathbb{E}\xi_{i_1}\xi_{i_1}^T \xi_{i_3}\xi_{i_3}^T - S^2)|$ when $|j - k| > 1$, that is when i_1 and i_3 are from two non-neighboring batches. Without lost of generality, we assume that $i_3 > i_1$:

$$
\begin{aligned}
\left|\mathrm{tr}(\mathbb{E}(\xi_{i_1}\xi_{i_1}^T)(\xi_{i_3}\xi_{i_3}^T) - S^2)\right| &= \left|\mathrm{tr}(\mathbb{E}(\xi_{i_1}\xi_{i_1}^T)(S + \mathbb{E}_{i_1}\Sigma(\Delta_{i_3-1})) - S^2)\right| \\
&\stackrel{(a)}{=} \left|\mathrm{tr}(S^2 + \mathbb{E}\Sigma(\Delta_{i_1-1})S + \mathbb{E}\xi_{i_1}\xi_{i_1}^T \mathbb{E}_{i_1}\Sigma(\Delta_{i_3-1}) - S^2)\right| \\
&\leq |\mathrm{tr}(S)|\,\mathbb{E}(\Sigma_1\|\Delta_{i_1-1}\|_2 + \Sigma_2\|\Delta_{i_1-1}\|_2^2) + \left|\mathrm{tr}(\mathbb{E}\xi_{i_1}\xi_{i_1}^T \mathbb{E}_{i_1}\Sigma(\Delta_{i_3-1}))\right|,
\end{aligned}
$$

where (a) uses Assumption 3.2. Also notice that, by Lemma D.3

$$
\begin{aligned}
\left|\mathrm{tr}(\mathbb{E}\xi_{i_1}\xi_{i_1}^T \mathbb{E}_{i_1}\Sigma(\Delta_{i_3-1}))\right| &\leq \mathbb{E}\left|\mathrm{tr}(\xi_{i_1}\xi_{i_1}^T)\right| \mathbb{E}_{i_1}\|\Sigma(\Delta_{i_3-1})\| \\
&\leq \mathbb{E}\|\xi_{i_1}\|_2^2 \mathbb{E}_{i_1}(\Sigma_1\|\Delta_{i_3-1}\|_2 + \Sigma_2\|\Delta_{i_3-1}\|_2^2) \\
&\stackrel{(a)}{\lesssim} \mathbb{E}\|\xi_{i_1}\|_2^2 \Sigma_1\left[\exp\left(-\tfrac{1}{4}\mu \sum_{k=i_1+1}^{i_3-1} \eta_k\right)\|\Delta_{i_1}\|_2 + \sqrt{C_d}\,i_1^{-\alpha/2}\right] \\
&\quad + \Sigma_2 \mathbb{E}\|\xi_{i_1}\|_2^2\left[\exp\left(-\tfrac{1}{2}\mu \sum_{k=i_1+1}^{i_3-1} \eta_k\right)\|\Delta_{i_1}\|_2^2 + C_d i_1^{-\alpha}\right] \\
&\stackrel{(b)}{\lesssim} \mathbb{E}\|\xi_{i_1}\|_2^2\left[C_d^{3/2}\|\Delta_{i_1}\|_2 + C_d^2 i_1^{-\alpha/2}\right] + \mathbb{E}\|\xi_{i_1}\|_2^2\left[C_d^2\|\Delta_{i_1}\|_2^2 + C_d^2 i_1^{-\alpha}\right] \\
&\leq 2C_d^2 i_1^{-\alpha/2}\mathbb{E}\|\xi_{i_1}\|_2^2 + C_d^{3/2}\sqrt{\mathbb{E}\|\xi_{i_1}\|_2^4 \mathbb{E}\|\Delta_{i_1}\|_2^2} + C_d^2\sqrt{\mathbb{E}\|\xi_{i_1}\|_2^4 \mathbb{E}\|\Delta_{i_1}\|_2^4} \\
&\stackrel{(c)}{\lesssim} C_d^3 i_1^{-\alpha/2} + C_d^6 i_1^{-2\alpha}
\end{aligned}
$$

where (a) uses the second claim in Lemma 3.2, (b) uses the fourth claim in Lemma D.1, and (c) uses the second and third claim in Lemma 3.2, while

$$
\mathbb{E}\|\xi_{i_1}\|_2^2 \leq \mathrm{tr}(S) + \Sigma_1\mathbb{E}\|\Delta_{i_1}\|_2 + \Sigma_2\mathbb{E}\|\Delta_{i_1}\|_2^2 \lesssim C_d + C_d^2 i_1^{-\alpha/2} + C_d^3 i_1^{-\alpha},
$$

$$
\mathbb{E}\|\xi_{i_1}\|_2^4 \leq \Sigma_3 + \Sigma_4\mathbb{E}\|\Delta_{i_1}\|_2^4 \lesssim C_d^2 + C_d^6 i_1^{-2\alpha}.
$$

Now using the second claim in Lemma 3.2, we obtain

$$|\text{tr}(S)|\,\mathbb{E}(\Sigma_1\|\Delta_{i_1-1}\|_2+\Sigma_2\|\Delta_{i_1-1}\|_2^2)\lesssim C_d^3 i_1^{-\alpha/2}+C_d^4 i_1^{-\alpha}\lesssim C_d^3 i_1^{-\alpha/2}+C_d^6 i_1^{-2\alpha}.$$

and the following bound for i_1, i_3 from non-neighboring batches:

$$\left|\text{tr}(\mathbb{E}(\xi_{i_1}\xi_{i_1}^T)(\xi_{i_3}\xi_{i_3}^T)-S^2)\right|\lesssim C_d^3 i_1^{-\alpha/2}+C_d^6 i_1^{-2\alpha}\leq C_d^3(i_1^{-\alpha/2}+i_3^{-\alpha/2})+C_d^6(i_1^{-2\alpha}+i_3^{-2\alpha}).$$

We added $i_3^{-\alpha/2}$ so the bound holds also for the case that $i_1 > i_3$. On the other hand, the following bound holds for general choice of $i_1, \ldots, i_4$. Note that $\xi_{i_1}\xi_{i_2}^T\xi_{i_3}\xi_{i_4}^T$ is rank one, so

$$\left|\text{tr}(\xi_{i_1}\xi_{i_2}^T\xi_{i_3}\xi_{i_4}^T)\right|=\left\|\xi_{i_1}\xi_{i_2}^T\xi_{i_3}\xi_{i_4}^T\right\|.$$

By Lemma B.1

$$\mathbb{E}\left|\text{tr}(\xi_{i_1}\xi_{i_2}^T\xi_{i_3}\xi_{i_4}^T)\right|\leq\frac{1}{4}\mathbb{E}[\|\xi_{i_1}\|_2^4+\|\xi_{i_2}\|_2^4+\|\xi_{i_3}\|_2^4+\|\xi_{i_4}\|_2^4]\leq\Sigma_3+\Sigma_4\mathbb{E}\|\Delta_{i_1}\|^4\lesssim C_d^2+C_d^6 i_1^{-2\alpha}.$$

and also

$$\left|\text{tr}(\mathbb{E}\xi_{i_1}\xi_{i_2}^T\xi_{i_3}\xi_{i_4}^T-S^2)\right|\leq\mathbb{E}\left|\text{tr}(\xi_{i_1}\xi_{i_2}^T\xi_{i_3}\xi_{i_4}^T)\right|+\left|\text{tr}(S^2)\right|\lesssim C_d^2+C_d^6 i_1^{-2\alpha}.$$

Therefore the inequality (62) gives the following:

$$\begin{aligned}\left|\text{tr}(\mathbb{E}\widehat{S}^2-S^2)\right|\lesssim&M^{-2}\sum_{|j-k|>1}n_j^{-1}n_k^{-1}\sum_{i_1=s_j}^{e_j}\sum_{i_3=s_k}^{e_k}C_d^3(i_1^{-\alpha/2}+i_3^{-\alpha/2})\\
&+2M^{-2}\sum_{|j-k|\leq1}n_j^{-1}n_k^{-1}\sum_{i_1=s_j}^{e_j}\sum_{i_3=s_k}^{e_k}C_d^2\\
&+2M^{-2}\sum_{j=1}^{M}n_j^{-2}\sum_{i_1=s_j}^{e_j}C_d^2\\
\leq&C_d^3 M^{-1}\left[\sum_{k=1}^{M}s_k^{-\alpha/2}\right]+6C_d^2 M^{-1}\\
\lesssim&C_d^3 N^{-\frac{\alpha}{2-2\alpha}}M^{[1-\frac{\alpha}{2(1-\alpha)}]^++-1}+C_d^2 M^{-1},\end{aligned}$$

where we used estimations from Lemma D.1 1) and 5).   □

LEMMA D.5.   *Under Assumptions 3.1 and 3.2, the following holds for the linear oracle sequence*

$$\mathbb{E}\left\|M^{-1}\sum_{k=1}^{M}n_k\overline{U}_{n_k}\overline{U}_{n_k}^T-A^{-1}SA^{-1}\right\|\lesssim C_d M^{-\frac{1}{2}}+C_d N^{-\frac{1}{2}}+C_d^{\frac{3}{2}}(MN)^{-\frac{\alpha}{4-4\alpha}}.$$

PROOF. We define the following matrices sequences through recursion as in Lemma D.2,

$$Y_i^j = (I - \eta_j A)Y_i^{j-1}, \quad Y_i^i = I_d.$$

Then the recursion of U_n can be rewritten as

$$
\begin{aligned}
U_n &= Y_{n-1}^n U_{n-1} - Y_n^n \eta_n \xi_n \\
&= Y_{n-2}^n U_{n-2} - Y_{n-1}^n \eta_{n-1} \xi_{n-1} - Y_n^n \eta_n \xi_n \\
&= Y_{s-1}^n U_{s-1} - \sum_{i=s}^{n} Y_i^n \eta_i \xi_i.
\end{aligned}
$$

Therefore $\overline{U}_{n_k}$ can be written as

$$
\begin{aligned}
\overline{U}_{n_k} &= n_k^{-1} \left[\sum_{l=s_k}^{e_k} Y_{s_k-1}^l U_{s_k-1} - \sum_{l=s_k}^{e_k} \sum_{i=s_k}^{l} Y_i^l \eta_i \xi_i \right] \\
&= n_k^{-1} \left[\sum_{l=s_k}^{e_k} Y_{s_k-1}^l U_{s_k-1} - \sum_{i=s_k}^{e_k} \left(\sum_{l=i}^{e_k} Y_i^l \right) \eta_i \xi_i \right] \\
&= n_k^{-1} \left[\sum_{l=s_k}^{e_k} Y_{s_k-1}^l U_{s_k-1} - \sum_{i=s_k}^{e_k} \left(S_i^{e_k} + I \right) \eta_i \xi_i \right] \\
&= n_k^{-1} S_{s_k-1}^{e_k} U_{s_k-1} - n_k^{-1} A^{-1} \sum_{i=s_k}^{e_k} \xi_i - n_k^{-1} \sum_{i=s_k}^{e_k} Z_i^{e_k} \xi_i - n_k^{-1} \sum_{i=s_k}^{e_k} \eta_i \xi_i
\end{aligned}
$$

Here we denoted

$$S_m^{e_k} = \sum_{l=m+1}^{e_k} Y_m^l = \sum_{l=m}^{e_k} Y_m^l - I, \quad Z_m^{e_k} = \eta_m S_m^{e_k} - A^{-1},$$

as in Lemma D.2. Denote

$$A_k := -A^{-1} \sum_{i=s_k}^{e_k} \xi_i, \quad B_k := S_{s_k-1}^{e_k} U_{s_k-1} - \sum_{i=s_k}^{e_k} Z_i^{e_k} \xi_i - \sum_{i=s_k}^{e_k} \eta_i \xi_i,$$

then

$$
\begin{aligned}
M^{-1} \sum_{k=1}^{M} n_k \overline{U}_{n_k} \overline{U}_{n_k}^T &= M^{-1} \sum_{k=1}^{M} n_k^{-1} A_k A_k^T \\
&\quad + M^{-1} \sum_{k=1}^{M} n_k^{-1} \left[A_k B_k^T + B_k A_k^T + B_k B_k^T \right].
\end{aligned}
$$

(63)

Denote

$$\widehat{S} := M^{-1} \sum_{k=1}^{M} n_k^{-1} \left(\sum_{i=s_k}^{e_k} \xi_i \right) \left(\sum_{i=s_k}^{e_k} \xi_i \right)^T.$$

By Lemma D.4, we have for sufficiently large N

$$\mathbb{E} \left\| M^{-1} \sum_{k=1}^{M} n_k^{-1} A_k A_k^T - A^{-1} S A^{-1} \right\|^2 = \mathbb{E} \left\| A^{-1}(\widehat{S} - S) A^{-1} \right\|^2$$
$$\lesssim \left[C_d^3 N^{-\frac{\alpha}{2(1-\alpha)}} M^{-\beta\alpha} + C_d^2 M^{-1} \right].$$

By Cauchy-Schwartz inequlaity,

$$\mathbb{E} \left\| M^{-1} \sum_{k=1}^{M} n_k^{-1} A_k A_k^T - A^{-1} S A^{-1} \right\| \lesssim C_d^{3/2} N^{-\frac{\alpha}{4-4\alpha}} M^{-\frac{\beta\alpha}{2}} + C_d M^{-\frac{1}{2}}$$

Next, note that

$$\mathbb{E}\|B_k\|_2^2 = \mathbb{E} \left\| B_k B_k^T \right\| \leq \mathbb{E}\text{tr}[B_k B_k^T].$$

Using that ξ_n are martingale differences

$$\mathbb{E} B_k B_k^T$$
$$= \mathbb{E} \left[S_{s_k-1}^{e_k} U_{s_k-1} - \sum_{i=s_k}^{e_k} Z_i^{e_k} \xi_i - \sum_{i=s_k}^{e_k} \eta_i \xi_i \right] \left[S_{s_k-1}^{e_k} U_{s_k-1} - \sum_{i=s_k}^{e_k} Z_i^{e_k} \xi_i - \sum_{i=s_k}^{e_k} \eta_i \xi_i \right]^T,$$
$$= S_{s_k-1}^{e_k} \mathbb{E} \left[U_{s_k-1} U_{s_k-1}^T \right] (S_{s_k-1}^{e_k})^T + \sum_{i=s_k}^{e_k} (Z_i^{e_k} + \eta_i I) \mathbb{E}(\xi_i \xi_i^T)(Z_i^{e_k} + \eta_i I)^T,$$

From D.2 claim 3,

$$\text{tr}(\mathbb{E} B_k B_k^T) \leq \left\| S_{s_k-1}^{e_k} \right\|^2 \mathbb{E}\|U_{s_k-1}\|_2^2 + \sum_{i=s_k}^{e_k} \left\| Z_i^{e_k} + \eta_i I \right\|^2 \mathbb{E}\|\xi_i\|_2^2,$$

Note that from Lemma D.2 claim 3

$$\|Z_i^{e_k} + \eta_i I\| \leq \|Z_i^{e_k}\| + \eta i^{-\alpha} \lesssim i^{\alpha-1} + \exp\left(-c\lambda_A(e^k - i)\eta_i \right) + i^{-\alpha}.$$

Using Young's inequality and the estimate from Lemma D.1, and $i^{-\alpha} \leq i^{\alpha-1}$ with $\alpha \in [1/2, 1)$,

$$\|Z_i^{e_k} + \eta_i I\|^2 \lesssim s_k^{2\alpha-2} + \exp(-2c\lambda_A(e_k - i)\eta_{e_k}).$$

Thus from the bounds of $\|U_n\|_2^2$ in Lemma B.3, and that $\mathbb{E}\|\xi_i\|_2^2 \lesssim C_d$,

$$\frac{1}{C_d}\text{tr}(\mathbb{E}B_k B_k^T) \lesssim s_k^\alpha + \sum_{i=s_k}^{e_k} \|Z_i^{e_k} + \eta_i I\|^2 \lesssim s_k^\alpha + n_k s_k^{2\alpha-2} + \sum_{i=s_k}^{e_k} \exp(-2c\lambda_A(e_k-i)\eta_{e_k})$$

Note that

$$\sum_{i=s_k}^{e_k} \exp(-2c\lambda_A(e_k - i)\eta_{e_k}) \leq \sum_{i=0}^{\infty} \exp(-2c\lambda_A i\eta_{e_k}) = [2c\lambda_A \eta_{e_k}]^{-1}.$$

Then based on Lemma D.1,

$$n_k s_k^{2\alpha-2} + \eta_{e_k}^{-1} \lesssim N^{-\frac{1-2\alpha}{1-\alpha}} k^{-\frac{2-\alpha}{1-\alpha}} + N^{-\frac{\alpha}{1-\alpha}} k^{-\frac{\alpha}{1-\alpha}} \lesssim s_k^\alpha.$$

Using the bound that $e_k^\alpha(s_k - 1)^{-\alpha} \leq 4$,

$$M^{-1} \sum_{k=1}^{M} n_k^{-1}\mathbb{E}\|B_k\|_2^2 \lesssim C_d M^{-1} \sum_{k=1}^{M} n_k^{-1} s_k^\alpha.$$

By Lemma D.1,

$$n_k^{-1} s_k^\alpha \lesssim N^{-\frac{1}{1-\alpha}} k^{-\frac{\alpha}{1-\alpha}} N^{\frac{\alpha}{1-\alpha}} k^{\frac{\alpha}{1-\alpha}} = N^{-1}.$$

So

$$M^{-1} \sum_{k=1}^{M} n_k^{-1}\mathbb{E}\|B_k\|_2^2 \lesssim C_d N^{-1}.$$

On the other hand, by Lemma 3.2,

$$n_k^{-1}\mathbb{E}\|A_k\|_2^2 = n_k^{-1}\text{tr}\mathbb{E}A_k A_k^T$$
$$\leq n_k^{-1}\|A^{-1}\|^2 \left[n_k\text{tr}(S) + 2\sum_{i=s_k}^{e_k} \mathbb{E}(\Sigma_1\|\Delta_i\|_2 + \Sigma_2\|\Delta_i\|_2^2)\right] \lesssim C_d.$$

So

$$n_k^{-1}\mathbb{E}\|A_k\|_2\|B_k\|_2 \lesssim \sqrt{n_k^{-1}\mathbb{E}\|A_k\|_2^2}\sqrt{n_k^{-1}\mathbb{E}\|B_k\|_2^2} \lesssim C_d N^{-\frac{1}{2}}.$$

Therefore, in conclusion,

$$
\mathbb{E}\left\|M^{-1}\sum_{k=1}^{M}\overline{U}_{n_k}\overline{U}_{n_k}^{T}-A^{-1}SA^{-1}\right\|
$$

$$
\leq \mathbb{E}\left\|M^{-1}\sum_{k=1}^{M}n_k^{-1}A_kA_k^{T}-A^{-1}SA^{-1}\right\|
$$

$$
+2M^{-1}\sum_{k=1}^{M}n_k^{-1}(\mathbb{E}\|A_k\|_2\|B_k\|_2+\mathbb{E}\|B_k\|_2^2)
$$

$$
\lesssim C_d^{3/2}N^{-\frac{\alpha}{4-4\alpha}}M^{-\frac{\beta\alpha}{2}}+C_dM^{-\frac{1}{2}}+C_dN^{-\frac{1}{2}}.
$$

Finally, if $\frac{1}{2}\leq\frac{\alpha}{4-4\alpha}$, then

$$
C_d^{3/2}N^{-\frac{\alpha}{4-4\alpha}}M^{-\frac{\beta\alpha}{2}}\lesssim C_d^{3/2}N^{-\frac{\alpha}{4-4\alpha}}M^{-\frac{1}{2}}\lesssim C_dM^{-\frac{1}{2}},
$$

so it is an insignificant term. When $\frac{1}{2}\geq\frac{\alpha}{4-4\alpha}$, then the first term is of order $(MN)^{-\frac{\alpha}{4-4\alpha}}$.

$\square$

D.3. *Proof of Lemma 4.7.*

LEMMA 4.7. *Under Assumptions 3.1 and 3.2, when d is fixed and the step size is chosen to be $\eta_i=\eta i^{-\alpha}$ with $\alpha\in(\frac{1}{2},1)$, the batch-means estimator based on the sequence U_n with any bounded U_0 satisfies the following inequality for sufficiently large N and M,*

$$
\mathbb{E}\left\|M^{-1}\sum_{k=1}^{M}n_k(\overline{U}_{n_k}-\overline{U}_M)(\overline{U}_{n_k}-\overline{U}_M)^{T}-A^{-1}SA^{-1}\right\|
$$

$$
\lesssim C_dM^{-\frac{1}{2}}+C_dN^{-\frac{1}{2}}+C_d^{\frac{3}{2}}(NM)^{-\frac{\alpha}{4-4\alpha}}.
$$

PROOF OF LEMMA 4.7. Using the identity $\sum_{k=1}^{M}n_k\overline{U}_M=\sum_{k=1}^{M}n_k\overline{U}_{n_k}$, we can easily find

$$
M^{-1}\sum_{k=1}^{M}n_k(\overline{U}_{n_k}-\overline{U}_M)(\overline{U}_{n_k}-\overline{U}_M)^{T}
$$

$$
=M^{-1}\sum_{k=1}^{M}n_k\overline{U}_{n_k}\overline{U}_{n_k}^{T}-M^{-1}\sum_{k=1}^{M}n_k\overline{U}_M\overline{U}_M^{T}.
$$

Lemma D.5 already shows that

$$\mathbb{E}\left\|M^{-1}\sum_{k=1}^{M}n_k\overline{U}_{n_k}\overline{U}_{n_k}^T - A^{-1}SA^{-1}\right\| \lesssim C_dM^{-\frac{1}{2}} + C_dN^{-\frac{1}{2}} + C_d^{\frac{3}{2}}(NM)^{-\frac{\alpha}{4-4\alpha}}$$

holds, so in order to complete our proof, it suffices to show $\mathbb{E}\|M^{-1}S_M\overline{U}_M\overline{U}_M^T\|$ is of the order C_dM^{-1}. Note that

$$\overline{U}_M = S_M^{-1}S_{s_1-1}^{e_M}U_{s_1-1} - S_M^{-1}\sum_{i=s_1}^{e_M}\left(S_i^{e_M}+I\right)\eta_i\xi_i$$

so using the fact that ξ_i are martingale differences

$$\mathbb{E}\|\overline{U}_M\|_2^2$$

$$\leq S_M^{-2}\|S_{s_1-1}^{e_M}\|^2\mathbb{E}\|U_{s_1-1}\|_2^2 + S_M^{-2}\sum_{i=s_1}^{e_M}\eta_i^2\|S_i^{e_M}+I\|^2\left[\operatorname{tr}(S) + \mathbb{E}(\Sigma_1\|\Delta_i\|_2 + \Sigma_2\|\Delta_i\|_2^2)\right].$$

Lemma D.2 claim 3 provides the following bound using the approximations in Lemma D.1

$$M^{-1}S_M^{-1}\|S_{s_1-1}^{e_M}\|^2\mathbb{E}\|U_{s_1-1}\|_2^2 \lesssim M^{-1}S_M^{-1}s_1^{\alpha}C_d \lesssim C_dM^{-1}N^{-1}.$$

On the other hand, Lemma D.2 3) also shows that $\eta_i^2\|S_i^{e_M}+I\|^2$ is uniformly bounded, and by Lemma 3.2,

$$[\operatorname{tr}(S) + \mathbb{E}(\Sigma_1\|\Delta_i\|_2 + \Sigma_2\|\Delta_i\|_2^2)] \lesssim C_d.$$

Therefore

$$M^{-1}S_M^{-1}\sum_{i=s_1}^{e_M}\eta_i^2\|S_i^{e_M}+I\|^2\left[\operatorname{tr}(S) + \mathbb{E}(\Sigma_1\|\Delta_i\|_2 + \Sigma_2\|\Delta_i\|_2^2)\right] \lesssim M^{-1}C_d.$$

Collecting the terms with lower order produces the claimed result. □

D.4. *Consistency proof of Batch-Means Estimator.*

THEOREM 4.3. *Under Assumptions 3.1 and 3.2, when d is fixed and the step size is chosen to be $\eta_i = \eta i^{-\alpha}$ with $\alpha \in (\frac{1}{2}, 1)$, the batch-means estimator initialized by any bounded x_0 is a consistent estimator. In particular, for sufficiently large N and M, we have*

$$\mathbb{E}\left\|M^{-1}\sum_{k=1}^{M}n_k(\overline{X}_{n_k}-\overline{X}_M)(\overline{X}_{n_k}-\overline{X}_M)^T - A^{-1}SA^{-1}\right\|$$

$$\lesssim C_dM^{-\frac{1}{2}} + C_dN^{-\frac{1}{2}} + C_d^{\frac{3}{2}}(MN)^{-\frac{\alpha}{4-4\alpha}} + C_d^2M^{-1} + C_d^3M^{-1}N^{\frac{1-2\alpha}{1-\alpha}}.$$

PROOF OF THEOREM 4.3. Recall the linear oracle sequence defined by (20): $U_n = (I - \eta_n A) U_{n-1} + \eta_n \xi_n$, $U_0 = \Delta_0$. From Lemma 4.7

$$\mathbb{E} \left\| M^{-1} \sum_{k=1}^{M} n_k (\overline{U}_{n_k} - \overline{U}_M)(\overline{U}_{n_k} - \overline{U}_M)^T - A^{-1}SA^{-1} \right\|$$

$$\lesssim C_d M^{-\frac{1}{2}} + C_d N^{-\frac{1}{2}} + C_d^{\frac{3}{2}} (MN)^{-\frac{\alpha}{4-4\alpha}}.$$

Let $\delta_n := \Delta_n - U_n$, which follows a recursion:

$$\delta_n = \Delta_{n-1} - U_{n-1} - \eta_n \nabla F(x_{n-1}) - \eta_n A U_{n-1}$$
$$= (I - \eta_n A)\delta_{n-1} + \eta_n (A\Delta_{n-1} - \nabla F(x_{n-1})).$$

Define the mean differences for batches and overall:

$$\overline{\delta}_{n_k} := n_k^{-1} \sum_{i=s_k}^{e_k} (\Delta_i - U_i), \quad \overline{\delta}_M := S_M^{-1} \sum_{i=s_1}^{e_M} (\Delta_i - U_i).$$

Using $\sum_{k=1}^{M} n_k \overline{U}_M = \sum_{k=1}^{M} n_k \overline{U}_{n_k}$ and $\sum_{k=1}^{M} n_k \overline{X}_M = \sum_{k=1}^{M} n_k \overline{X}_{n_k}$, we have the following decomposition:

$$\frac{1}{M} \sum_{k=1}^{M} n_k (\overline{X}_{n_k} - \overline{X}_M)(\overline{X}_{n_k} - \overline{X}_M)^T = \frac{1}{M} \sum_{k=1}^{M} n_k \overline{X}_{n_k} \overline{X}_{n_k}^T - \frac{1}{M} \sum_{k=1}^{M} n_k \overline{X}_M \overline{X}_M^T$$

$$= \frac{1}{M} \sum_{k=1}^{M} n_k \overline{\Delta}_{n_k} \overline{\Delta}_{n_k}^T - \frac{1}{M} \sum_{k=1}^{M} n_k x^* x^{*T} + \frac{1}{M} \sum_{k=1}^{M} n_k \overline{\Delta}_M x^{*T}$$

$$+ \frac{1}{M} \sum_{k=1}^{M} n_k x^* \overline{\Delta}_M^T - \frac{1}{M} \sum_{k=1}^{M} n_k (x^* + \overline{\Delta}_M)(x^* + \overline{\Delta}_M)^T$$

$$= \frac{1}{M} \sum_{k=1}^{M} n_k \overline{\Delta}_{n_k} \overline{\Delta}_{n_k}^T - \frac{1}{M} \sum_{k=1}^{M} n_k \overline{\Delta}_M \overline{\Delta}_M^T$$

$$= \frac{1}{M} \sum_{k=1}^{M} n_k (\overline{U}_{n_k} + \overline{\delta}_{n_k})(\overline{U}_{n_k} + \overline{\delta}_{n_k})^T - \frac{1}{M} \sum_{k=1}^{M} n_k (\overline{U}_M + \overline{\delta}_M)(\overline{U}_M + \overline{\delta}_M)^T$$

$$= \frac{1}{M} \sum_{k=1}^{M} n_k (\overline{U}_{n_k} - \overline{U}_M)(\overline{U}_{n_k} - \overline{U}_M)^T + \frac{1}{M} \sum_{k=1}^{M} n_k (\overline{U}_{n_k} - \overline{U}_M)(\overline{\delta}_{n_k} - \overline{\delta}_M)^T$$

$$+ \frac{1}{M} \sum_{k=1}^{M} n_k (\overline{\delta}_{n_k} - \overline{\delta}_M)(\overline{U}_{n_k} - \overline{U}_M)^T + \frac{1}{M} \sum_{k=1}^{M} n_k (\overline{\delta}_{n_k} - \overline{\delta}_M)(\overline{\delta}_{n_k} - \overline{\delta}_M)^T.$$

Then by Young's inequality and $\|C\| \le \operatorname{tr}(C)$ for any PSD matrix C,

$$
\mathbb{E}\left\| \frac{1}{M} \sum_{k=1}^{M} n_k (\overline{X}_{n_k} - \overline{X}_M)(\overline{X}_{n_k} - \overline{X}_M)^T - A^{-1} S A^{-1} \right\|
$$

$$
\le \mathbb{E}\left\| \frac{1}{M} \sum_{k=1}^{M} n_k (\overline{U}_{n_k} - \overline{U}_M)(\overline{U}_{n_k} - \overline{U}_M)^T - A^{-1} S A^{-1} \right\|
$$

$$
+ \frac{1}{M} \sum_{k=1}^{M} n_k \mathbb{E} \operatorname{tr}\left[(\overline{\delta}_{n_k} - \overline{\delta}_M)(\overline{\delta}_{n_k} - \overline{\delta}_M)^T \right]
$$

$$
+ \frac{2}{M} \sqrt{ \sum_{k=1}^{M} n_k \mathbb{E} \operatorname{tr}[(\overline{U}_{n_k} - \overline{U}_M)(\overline{U}_{n_k} - \overline{U}_M)^T] \sum_{k=1}^{M} n_k \mathbb{E} \operatorname{tr}\left[(\overline{\delta}_{n_k} - \overline{\delta}_M)(\overline{\delta}_{n_k} - \overline{\delta}_M)^T \right] }
$$

Note that

$$
\sum_{k=1}^{M} n_k \operatorname{tr}(\overline{\delta}_{n_k} - \overline{\delta}_M)(\overline{\delta}_{n_k} - \overline{\delta}_M)^T
$$

$$
= \sum_{k=1}^{M} n_k \left[\operatorname{tr}(\overline{\delta}_{n_k} \overline{\delta}_{n_k}^T) - \operatorname{tr}(\overline{\delta}_M \overline{\delta}_M^T) \right] \le \sum_{k=1}^{M} n_k \operatorname{tr}(\overline{\delta}_{n_k} \overline{\delta}_{n_k}^T).
$$

Therefore, a proper bound for $\frac{1}{M} \sum_{k=1}^{M} n_k \mathbb{E} \|\overline{\delta}_{n_k}\|_2^2$ will conclude our proof. Using the definition of Y_j^k and S_j^k in Lemma D.2, we find that

$$
\delta_n = (I - \eta_n A)\delta_{n-1} + \eta_n (A\Delta_{n-1} - \nabla F(x_{n-1}))
$$

$$
= \left[\prod_{i=s_k}^{n} (I - \eta_i A) \right] \delta_{s_k - 1} + \sum_{i=s_k}^{n} \left[\prod_{j=i+1}^{n} (I - \eta_j A) \right] \eta_i (A\Delta_{i-1} - \nabla F(x_{i-1}))
$$

$$
= Y_{s_k - 1}^{n} \delta_{s_k - 1} + \sum_{i=s_k}^{n} Y_i^n \eta_i (A\Delta_{i-1} - \nabla F(x_{i-1}))
$$

$$
\overline{\delta}_{n_k} = \frac{1}{n_k} \sum_{j=s_k}^{e_k} \delta_j = \frac{1}{n_k} \sum_{j=s_k}^{e_k} Y_{s_k - 1}^{j} \delta_{s_k - 1} + \frac{1}{n_k} \sum_{j=s_k}^{e_k} \sum_{i=s_k}^{j} Y_i^j \eta_i (A\Delta_{i-1} - \nabla F(x_{i-1}))
$$

$$
= \frac{1}{n_k} S_{s_k - 1}^{e_k} \delta_{s_k - 1} + \frac{1}{n_k} \sum_{i=s_k}^{e_k} \sum_{j=i}^{e_k} Y_i^{e_k} \eta_i (A\Delta_{i-1} - \nabla F(x_{i-1}))
$$

$$
= \frac{1}{n_k} S_{s_k - 1}^{e_k} \delta_{s_k - 1} + \frac{1}{n_k} \sum_{n=s_k}^{e_k} (I + S_n^{e_k}) \eta_n (A\Delta_{n-1} - \nabla F(x_{n-1})).
$$

So by Young's inequality and Cauchy-Schwartz inequality

$$
\begin{aligned}
\mathbb{E}\|\overline{\delta}_{n_k}\|_2^2 \leq{} & 2n_k^{-2}\|S_{s_k-1}^{e_k}\|^2 \mathbb{E}\|\delta_{s_k-1}\|_2^2 \\
& + 2n_k^{-2}\mathbb{E}\left(\sum_{n=s_k}^{e_k}\|\eta_n(I+S_n^{e_k})\|\|(A\Delta_{n-1}-\nabla F(x_{n-1})\|\right)^2 \\
\leq{} & 2n_k^{-2}\|S_{s_k-1}^{e_k}\|^2\mathbb{E}\|\delta_{s_k-1}\|_2^2 \\
& + 2n_k^{-2}\sum_{n=s_k}^{e_k}\|\eta_n(S_n^{e_k}+I)\|^2\mathbb{E}\sum_{n=s_k}^{e_k}\|A\Delta_{n-1}-\nabla F(x_{n-1})\|_2^2
\end{aligned}
$$

Apply Lemma 3.2 and Lemma D.2,

$$
\|S_{s_k-1}^{e_k}\|^2 \lesssim s_k^{2\alpha}, \quad \sum_{n=s_k}^{e_k}\|\eta_n S_n^{e_k}\|^2 \lesssim n_k.
$$

$$
\mathbb{E}\sum_{n=s_k}^{e_k}\|A\Delta_{n-1}-\nabla F(x_{n-1})\|_2^2 \lesssim \sum_{n=s_k}^{e_k}\mathbb{E}L_F^2\|\Delta_{n-1}\|_2^2 \lesssim n_k s_k^{-\alpha}C_d^3.
$$

Notice that Lemma 3.2 can be applied to both U_n and Δ_n, so

$$
\mathbb{E}\|\delta_{s_k-1}\|_2^2 \leq 2\mathbb{E}\|\Delta_{s_k-1}\|_2^2 + 2\mathbb{E}\|U_{s_k-1}\|_2^2 \lesssim s_k^{-\alpha}C_d.
$$

Furthermore, recall the batch size $e_k = ((k+1)N)^{\frac{1}{1-\alpha}}$,

$$
n_k = \left[(k+1)^{\frac{1}{1-\alpha}}-k^{\frac{1}{1-\alpha}}\right]N^{\frac{1}{1-\alpha}} = \frac{N^{\frac{1}{1-\alpha}}}{1-\alpha}\int_k^{k+1}x^{\frac{\alpha}{1-\alpha}}dx \asymp k^{\frac{\alpha}{1-\alpha}}N^{\frac{1}{1-\alpha}},
$$

$$
n_k\mathbb{E}\|\overline{\delta}_{n_k}\|_2^2 \lesssim C_d n_k^{-1}s_k^{\alpha} + n_k s_k^{-2\alpha}C_d^3 \lesssim C_d N^{-1} + C_d^3 k^{-\frac{\alpha}{1-\alpha}}N^{\frac{1-2\alpha}{1-\alpha}}.
$$

For the sum of geometric sequence in the last term, the following holds

$$
\frac{1}{M}\sum_{k=1}^{M}k^{-\frac{\alpha}{1-\alpha}} \lesssim \frac{1}{M}\int_1^{M+1}x^{-\frac{\alpha}{1-\alpha}}dx \lesssim M^{(1-\frac{\alpha}{1-\alpha})^+-1},
$$

which implies

$$
\frac{1}{M}\sum_{k=1}^{M}n_k\mathbb{E}\|\overline{\delta}_{n_k}\|_2^2 \lesssim C_d N^{-1} + C_d^3 M^{(\frac{1-2\alpha}{1-\alpha})^+-1}N^{\frac{1-2\alpha}{1-\alpha}}.
$$

But since $\alpha > 1/2$, so

$$
M^{(\frac{1-2\alpha}{1-\alpha})^+-1}N^{\frac{1-2\alpha}{1-\alpha}} = M^{-1}N^{\frac{1-2\alpha}{1-\alpha}} \leq M^{-1}.
$$

In conclusion,

$$\mathbb{E}\left\|M^{-1}\sum_{k=1}^{M}n_k(\overline{X}_{n_k}-\overline{X}_M)(\overline{X}_{n_k}-\overline{X}_M)^T - A^{-1}SA^{-1}\right\|$$
$$\lesssim C_d M^{-\frac{1}{2}} + C_d N^{-\frac{1}{2}} + C_d^{\frac{3}{2}}(MN)^{-\frac{\alpha}{4-4\alpha}} + C_d^2 M^{-1} + C_d^3 M^{-1} N^{\frac{1-2\alpha}{1-\alpha}}.$$

$\square$

D.5. *Proof of Corollaries 4.3 and 4.4 .*

COROLLARY 4.3. *Under Assumptions 3.1, 3.2 and 4.1, as $n \to \infty$,*

$$\left\|A_n^{-1}S_nA_n^{-1} - A^{-1}SA^{-1}\right\| = O_p\big(\|S\|(C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha})\big).$$

PROOF. Let $\mathcal{Q} := \{\lambda_{\min}(A_n) > \delta\} = \{A_n = \widetilde{A}_n\}$, by Markov inequality and Lemma 4.1,

$$\mathbb{P}(\lambda_{\min}(A_n) \le \tfrac{\mu}{2}) \le \mathbb{P}(\|A_n - A\| \ge \tfrac{\mu}{2}) \le \frac{2\mathbb{E}\|A_n - A\|}{\mu} \lesssim C_d^2 n^{-\alpha/2}.$$

then $\mathbb{P}(\mathcal{Q}^c) \le c_1 C_d^2 n^{-\frac{\alpha}{2}}$ for some c_1. Then by Theorem 4.2, for a constant c_2,

$$\mathbb{E}\mathbb{1}_{\mathcal{Q}}\left\|A_n^{-1}S_nA_n^{-1} - A^{-1}SA^{-1}\right\| \le c_2\|S\|(C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha}).$$

By Markov inequality again, for any $c_3 > 0$,

$$\mathbb{P}\big(\left\|A_n^{-1}S_nA_n^{-1} - A^{-1}SA^{-1}\right\| > c_3 c_2\|S\|(C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha})\big) \le c_1 C_d^2 n^{-\frac{\alpha}{2}} + c_3^{-1}.$$

Therefore we can choose a large c_3, so when $n \to \infty$,

$$\left\|A_n^{-1}S_nA_n^{-1} - A^{-1}SA^{-1}\right\| = O_p(\|S\|(C_d^2 n^{-\frac{\alpha}{2}} + C_d^3 n^{-\alpha})).$$

$\square$

COROLLARY 4.4. *Under the assumptions of Theorem 4.2, if the step size is chosen to be $\eta_i = \eta i^{-\alpha}$ with $\alpha \in (\frac{1}{2}, 1)$, when d is fixed and $n \to \infty$,*

$$\Pr\left(\bar{x}_{n,j} - z_{q/2}\hat{\sigma}_{n,j}^P \le x_j^* \le \bar{x}_{n,j} + z_{q/2}\hat{\sigma}_{n,j}^P\right) \to 1 - q.$$

PROOF OF COROLLARY 4.4. Denote $\sigma_j = \left(A^{-1}SA^{-1}\right)_{jj}^{1/2}$. Fix any $\epsilon > 0$, note that

$$
\begin{aligned}
&\Pr\left(\bar{x}_{n,j} - z_{q/2}\hat{\sigma}_{n,j}^P \le x_j^* \le \bar{x}_{n,j} + z_{q/2}\hat{\sigma}_{n,j}^P\right) \\
&\le \Pr\left(\bar{x}_{n,j} - z_{q/2-\epsilon}\sigma_j \le x_j^* \le \bar{x}_{n,j} + z_{q/2-\epsilon}\sigma_j\right) + \Pr(z_{q/2-\epsilon}\sigma_j \le z_{q/2}\hat{\sigma}_{n,j}^P).
\end{aligned}
$$

The first term by theorem 2 of Polyak and Juditsky (1992) converges to $1 - q + 2\epsilon$ when $n \to \infty$, the second term by Theorem 4.2 and Markov inequality converges to 0 when $n \to \infty$. Therefore we have

$$
\Pr\left(\bar{x}_{n,j} - z_{q/2}\hat{\sigma}_{n,j}^P \le x_j^* \le \bar{x}_{n,j} + z_{q/2}\hat{\sigma}_{n,j}^P\right) \le 1 - q + 2\epsilon, \quad n \to \infty.
$$

Likewise we can have the opposite direction. Since this holds for any $\epsilon > 0$, we have our result. □

E. Supplement to the High Dimensional Section (Section 5).

E.1. *Proof of Proposition 5.1.*

PROPOSITION 5.1.   *Under Assumption 5.1 and using the same algorithm parameters as Proposition 1 in* Agarwal, Negahban and Wainwright (2012), *there exists a constant c_0, such that $\widehat{x}_n$ in Algorithm 1 satisfies*

$$
\|\widehat{x}_n - x^*\|_1 \le c_0 s \sqrt{\frac{\log d}{n}} \tag{64}
$$

uniformly in $x^* \in \mathcal{B}(s)$ with high probability. Further, for each $j = \{1, \dots, d\}$,

$$
\|\widehat{\gamma}^j + \Omega_{j,j}^{-1}(\Omega_{j,-j})^T\|_1 \le c_0 s_j \sqrt{\frac{\log d}{n}} \tag{65}
$$

holds with high probability.

PROOF. where $R_i = \frac{R_1}{2^{i/2}}$, $T_i \ge c_1 \frac{s^2}{R_i^2}\log d$ for a universal constant c_1, $A_\psi \le e\log d$, $\omega_i^2 = \omega^2 + 24\log i$, and K^2 is the variance proxy of a in Assumption E.2. Therefore, let $c_0 = (c_1\log d)^{-\frac{1}{4}}(2\mu)^{\frac{1}{2}}\left[e(\log d)(L_F^2 + K^2) + \omega_i^2 K^2\right]^{\frac{1}{4}}$. We have

$$
\lambda_{K_n} \le \frac{c_0 R_{K_n}}{\sqrt{2}s} \le c_0 \frac{R_1}{2^{K_n/2}s}.
$$

By Eq. (32) in Agarwal, Negahban and Wainwright (2012), the total number of iterations $n \le s^2(\log d)2^{K_n}$. Therefore, $\|\widehat{x}_n - x^*\|_1 \le \frac{9c_0 R_1}{\mu} \cdot s\sqrt{\frac{\log n}{d}}$. The constant c_0 is uniform for $x^* \in \mathcal{B}(s)$. Consider the set of parameters $\mathcal{B}(s) =$

$\{x \in \mathbb{R}^d; \ |\{j : x_j \neq 0\}| \leq s\}$. The convergence of $\|\widehat{x}_n - x^*\|_1$ in (64) is a direct consequence of Proposition 1 in Agarwal, Negahban and Wainwright (2012). To see this, from Lemma 1 and Proposition 1 of Agarwal, Negahban and Wainwright (2012), By Lemma 1 (Eq. (39)) of Agarwal, Negahban and Wainwright (2012), we have

$$(66) \qquad \|\widehat{x}_n - x^*\|_1 \leq \frac{9}{\mu} s \lambda_{K_n},$$

where K_n is the number of epochs and λ_{K_n} is the regularization parameters used in the K_n-th epoch. By Assumption 5.1, there exists positive constants $\mu, L_F > 0$ such that $\mu < \lambda_{\min}(A) < \lambda_{\max}(A) < L_F$, and $\mathbb{E}(b - a^T x)^2$ is Lipschitz continuous with constant L_F. By the setting of λ_i for the i-th epoch in Eq. (34) in Agarwal, Negahban and Wainwright (2012), we have

$$(67) \qquad \lambda_i^2 = \frac{R_i \mu}{s \sqrt{T_i}} \sqrt{A_\psi (L_F^2 + K^2) + \omega_i^2 K^2},$$

where $R_i = \frac{R_1}{2^{i/2}}$, $T_i \geq c_1 \frac{s^2}{R_i^2} \log d$ for a universal constant c_1, $A_\psi \leq e \log d$, $\omega_i^2 = \omega^2 + 24 \log i$, and K^2 is the variance proxy of a in Assumption E.2. Therefore, let $c_0 = (c_1 \log d)^{-\frac{1}{4}} (2\mu)^{\frac{1}{2}} \left[e(\log d)(L_F^2 + K^2) + \omega_i^2 K^2 \right]^{\frac{1}{4}}$. We have

$$(68) \qquad \lambda_{K_n} \leq \frac{c_0 R_{K_n}}{\sqrt{2} s} \leq c_0 \frac{R_1}{2^{K_n/2} s}.$$

To further bound λ_{K_n} in (68), we only need a lower bound on the number of epochs K_n. By the setting of T_i in Eq. (32) in Agarwal, Negahban and Wainwright (2012), we can easily derive that,

$$n = \sum_{i=1}^{K_n} T_i \geq s^2 (\log d) 2^{K_n},$$

which implies that

$$\lambda_{K_n} \leq \frac{c_0 R_1}{2^{K_n/2} s} \leq c_0 R_1 \sqrt{\frac{\log d}{n}},$$

which combining with (66) completes the proof of (64) with the notice that R_1 is bounded uniformly by Assumption 5.1.

For (65), recall that $\widehat{\gamma}^j$ is the solution of RADAR for solving the problem in (25). We can apply the Proposition 1 from (Agarwal, Negahban and Wainwright,

2012) to provide the bound on $\|\widehat{\gamma}^j - (\gamma^j)^*\|_1$. Here $(\gamma^j)^*$ is the optimum for the un-regularized population loss, i.e.,

$$(\gamma^j)^* = \operatorname{argmin} \mathbb{E}\left((a_j - a_{-j}\gamma^j)^2\right)$$
$$= -\Omega_{j,j}^{-1}\left(\Omega_{j,-j}\right)^T$$

We further note that by the row sparsity assumption of Ω in Assumption 5.1, $(\gamma^j)^*$ is an s_j-sparse vector. By exactly the same argument as for proving (66), we obtain the result in (65). □

E.2. *Proof of Theorem 5.2.*   We first prove two lemmas about estimating Ω_{jj} and Ω,

LEMMA E.1.   *Let* $\hat{\tau}_j \equiv \frac{1}{n}\|D_{\cdot,j} - D_{\cdot,-j}\hat{\gamma}_j\|_2^2$ *and* $\gamma_j := \arg\min_{\gamma \in \mathbb{R}^{p-1}} \mathbb{E}[\|a_j - a_{-j}\gamma\|_2^2]$. *Assuming* $\|\hat{\gamma}_j - \gamma_j\|_2 \leq C\sqrt{\frac{s_j \log d}{n}}$, $\|\hat{\gamma}_j - \gamma_j\|_1 \leq Cs_j\sqrt{\frac{\log d}{n}}$, *and* $n > s_j \log d$, *we have*

$$(69) \qquad \left|\frac{\hat{\tau}_j}{\Omega_{jj}^{-1}} - 1\right| \leq C\sqrt{\frac{\log d}{n}} + 2\lambda_{\max}(A)C\frac{s_j \log d}{n}$$

PROOF. Fix an index j. Let $z = \sqrt{\Omega_{j,j}}(D_{\cdot,j} - D_{\cdot,-j}\gamma_j)$. The entries of z are *i.i.d.* realizations of $\sqrt{\Omega_{j,j}}(a_j - a_{-j}\gamma_j) = \sqrt{\Omega_{j,j}}(a_j - \Omega_{jj}^{-1}\Omega_{j,-j}a_{-j})$. Since a is sub-Gaussian, z is also sub-Gaussian and the entries of z are *i.i.d.* with mean 0 and variance 1.

$$\sqrt{\hat{\tau}_j} = \frac{1}{\sqrt{n}}\|D_{\cdot,j} - D_{\cdot,-j}\hat{\gamma}_j\|_2 = \frac{1}{\sqrt{n}}\|D_{\cdot,j} - D_{\cdot,-j}\gamma_j + D_{\cdot,-j}(\hat{\gamma}_j - \gamma_j)\|_2$$
$$= \frac{1}{\sqrt{n}}\left\|\sqrt{\Omega_{jj}^{-1}}z + D_{\cdot,-j}(\hat{\gamma}_j - \gamma_j)\right\|_2$$

Therefore

$$(70) \qquad \left|\sqrt{\hat{\tau}_j} - \frac{1}{\sqrt{n}}\left\|\sqrt{\Omega_{jj}^{-1}}z\right\|_2\right| \leq \frac{1}{\sqrt{n}}\|D_{\cdot,-j}(\hat{\gamma}_j - \gamma_j)\|_2$$

By Rudelson and Zhou (2012), $\frac{1}{n}D_{\cdot,-j}^T D_{\cdot,-j}$ satisfies the Upper-RE condition in Loh et al. (2012, Equation 2.13), we have that

$$\frac{1}{n}\|D_{\cdot,-j}(\hat{\gamma}_j - \gamma_j)\|_2^2 \leq 2\lambda_{\max}(A)\|\hat{\gamma}_j - \gamma_j\|_2^2 + \tau(n,d)\|\hat{\gamma}_j - \gamma_j\|_1^2,$$
$$\frac{1}{\sqrt{n}}\|D_{\cdot,-j}(\hat{\gamma}_j - \gamma_j)\|_2 \leq \sqrt{2\lambda_{\max}(A)}\|\hat{\gamma}_j - \gamma_j\|_2 + \sqrt{\tau(n,d)}\|\hat{\gamma}_j - \gamma_j\|_1$$

for all j, where $\gamma(n,d) \asymp \sqrt{(\log d)/n}$. For a sub-Gaussian random variable X with parameter K, it's easy to see X^2 is sub-exponential with parameter K^2 (Vershynin, 2010, Lemma 5.14). By the concentration inequality for K^2 sub-exponential random variables (Vershynin, 2010, Proposition 5.16),

$$\Pr\left(\left|\frac{1}{n}\|z\|_2^2 - 1\right| > t\right) \leq 2\exp\left(-\min\left(\frac{nt^2}{2K^4}, \frac{nt}{2K^2}\right)\right).$$

In below, let C_K be a dimension independent constant that may change its value through derivation. By setting $t \asymp \sqrt{\frac{\log d}{n}}$ with probability at least $1 - \frac{1}{d^C}$,

$$\left|\frac{1}{n}\|z\|^2 - 1\right| \leq C_K\sqrt{\frac{\log d}{n}}.$$

Using $\sqrt{a+b} < \sqrt{a} + \sqrt{b}$ and re-arranging,

$$\sqrt{1 - C_K\sqrt{\frac{\log d}{n}}} \leq \frac{\|z\|_2}{\sqrt{n}} \leq \sqrt{1 + C_K\sqrt{\frac{\log d}{n}}},$$

$$1 - C_K\left(\frac{\log d}{n}\right)^{.25} \leq \frac{\|z\|_2}{\sqrt{n}} \leq 1 + C_K\left(\sqrt{\frac{\log d}{n}}\right)^{.25}.$$

Using these two estimates with (70),

$$\sqrt{\hat{\tau}_j} \leq \sqrt{\Omega_{jj}^{-1}}\left(1 + C_K\left(\frac{\log d}{n}\right)^{.25}\right) + \sqrt{2\lambda_{\max}(A)}\sqrt{\frac{s_j \log d}{n}} + Cs_j\frac{\log d}{n},$$

$$\sqrt{\hat{\tau}_j} \geq \sqrt{\Omega_{jj}^{-1}}\left(1 - C_K\left(\frac{\log d}{n}\right)^{.25}\right) - \sqrt{2\lambda_{\max}(A)}\sqrt{\frac{s_j \log d}{n}} - Cs_j\frac{\log d}{n}.$$

Dividing by $\sqrt{\Omega_{jj}^{-1}}$, squaring, and using $\frac{s_j \log d}{n} < 1$ gives

$$\left|\frac{\hat{\tau}_j}{\Omega_{jj}^{-1}} - 1\right| \lesssim \sqrt{\frac{\log d}{n}} + 2\lambda_{\max}(A)\frac{s_j \log d}{n}.$$

$\square$

LEMMA E.2. *Let Ω_j and $\widehat{\Omega}_j$ be the j-th row of Ω and $\widehat{\Omega}$. Under the row sparsity assumption of Ω (see Assumption 5.1), we have with high probability*

$$\max_{j\in\{1,\ldots,d\}}\|\widehat{\Omega}_j - \Omega_j\|_2 \leq C\sqrt{\frac{\max_j s_j \log d}{n}}$$

$$\max_{j\in\{1,\ldots,d\}}\|\widehat{\Omega}_j - \Omega_j\|_1 \leq C\max_j s_j\sqrt{\frac{\log d}{n}}$$

PROOF OF LEMMA E.2. By the definition of $\widehat{\Omega}$ in (23), we have the j-th row of $\widehat{\Omega}$ takes the following form,

$$\widehat{\Omega}_j = \widehat{C}_j / \widehat{\tau}_j, \qquad \widehat{C}_j = (-\widehat{\gamma}_1^j, \ldots, -\widehat{\gamma}_{j-1}^j, 1, \ldots, -\widehat{\gamma}_d^j).$$

By Proposition 5.1, we have for any j,

$$(71) \qquad \|\widehat{C}_j - \Omega_{j,j}^{-1}\Omega_j\|_1 \lesssim s_j\sqrt{\frac{\log d}{n}},$$

holding with high probability.

Using Lemma E.1, $\widehat{\tau}_j = O(1)$, and noting that $\frac{s_j \log d}{n} < \sqrt{\frac{\log d}{n}}$, we see that

$$(72) \qquad \left|\frac{\Omega_{jj}^{-1}}{\widehat{\tau}_j} - 1\right| \leq C\sqrt{\frac{\log d}{n}}.$$

Combining (72) with (71) and noticing that

$$\|\Omega_j\|_1 \leq \sqrt{s_j}\|\Omega_j\|_2 \leq \frac{\sqrt{s_j}}{\lambda_{\min}(A)},$$

we have with probability at least $1 - 1/d^2$,

$$\begin{aligned}
\|\widehat{\Omega}_j - \Omega_j\|_1 &= \left\|\left(\widehat{C}_j/\widehat{\tau}_j - \Omega_{j,j}^{-1}\Omega_j/\widehat{\tau}_j\right) + \left(\Omega_{j,j}^{-1}\Omega_j/\widehat{\tau}_j - \Omega_j\right)\right\|_1 \\
&\leq \frac{1}{\widehat{\tau}_j}\|\widehat{C}_j - \Omega_{j,j}^{-1}\Omega_j\|_1 + \left|\frac{\Omega_{j,j}^{-1}}{\widehat{\tau}_j} - 1\right|\|\Omega_j\|_1 \\
&\lesssim s_j\sqrt{\frac{\log d}{n}},
\end{aligned}$$

which completes the proof by taking a union bound over j. The proof for the l_2 bound is analogous. □

With Lemma E.2 in place, we are ready to prove Theorem 5.2.

THEOREM 5.2. *Under Assumption 5.1, for suitable choices of $\lambda \asymp \sqrt{\log d/n}$ and $\lambda_j \asymp \sqrt{\log d/n}$, we have for all $j \in \{1, \ldots, d\}$ and all $z \in \mathbb{R}$,*

$$(73) \qquad \sup_{x^* \in \mathcal{B}(s)}\left|\mathbb{P}_{x^*}\left(\frac{\sqrt{n}(\widehat{x}_j^d - x_j^*)}{\sigma\sqrt{(\widehat{\Omega}\widehat{A}\widehat{\Omega}^T)_{jj}}} \leq z\right) - \Phi(z)\right| = o_p(1).$$

where $\widehat{x}^d$ is the debiased estimator defined in (25), $\widehat{\Omega}$ is defined in (23) and the sample covariance matrix $\widehat{A} = \frac{1}{n}D^T D$.

PROOF. We first plug the linear model $b = Dx^* + \epsilon$ into the definition of the desparsified estimator in (25) and obtain,

(74)
$$
\begin{aligned}
\widehat{x}^d &= \widehat{x}_n + \frac{1}{n}\widehat{\Omega}\widehat{A}\,(x^* - \widehat{x}_n) + \frac{1}{n}\widehat{\Omega}D^T\epsilon \\
&= x^* + \left(\widehat{\Omega}\widehat{A} - I\right)(x^* - \widehat{x}_n) + \frac{1}{n}\widehat{\Omega}D^T\epsilon.
\end{aligned}
$$

It is easy to see that $\frac{1}{\sqrt{n}}\widehat{\Omega}D^T\epsilon \sim N(0, \sigma^2\widehat{\Omega}\widehat{A}\widehat{\Omega}^T)$. Therefore, to prove Theorem 5.2, we only need to show the term

(75)
$$
\Delta := \left(\widehat{\Omega}\widehat{A} - I\right)(x^* - \widehat{x}_n) = o_{\mathbb{P}_{x^*}}(\frac{1}{\sqrt{n}}),
$$

uniformly in $x^* \in \mathcal{B}(s)$.

By (Javanmard and Montanari, 2014, Lemma 23), we have

$$
\Pr\left(\|\Omega\widehat{A} - I\|_\infty \geq a\sqrt{\frac{\log d}{n}}\right) \leq 2d^{-c_2}
$$

with $c_2 = (a^2\lambda_{\min})/(24e^2\kappa^4\lambda_{\max}) - 2$, where κ is the sub-gaussian norm for $A^{-1}a$. By adding and subtracting $\Omega\widehat{A}(\widehat{x}_n - x^*)$,

$$
\begin{aligned}
\|\Delta\|_\infty &\leq \left\|(\widehat{\Omega}\widehat{A} - \Omega\widehat{A})(\widehat{x}_n - x^*)\right\|_\infty + \left\|(\Omega\widehat{A} - I)(\widehat{x}_n - x^*)\right\|_\infty \\
&\leq \left\|(\widehat{\Omega}\widehat{A} - \Omega\widehat{A})(\widehat{x}_n - x^*)\right\|_\infty + \left\|\Omega\widehat{A} - I\right\|_\infty \|\widehat{x}_n - x^*\|_1 \\
&\leq \left\|(\widehat{\Omega}\widehat{A} - \Omega\widehat{A})(\widehat{x}_n - x^*)\right\|_\infty + C\frac{s\log d}{n}
\end{aligned}
$$

The first term can be bounded

$$
\left\| (\widehat{\Omega}\widehat{A} - \Omega\widehat{A})(\hat{x}_n - x^*) \right\|_\infty
$$

$$
= \left\| (\widehat{\Omega} - \Omega)\frac{1}{n}D^T D(\hat{x}_n - x^*) \right\|_\infty
$$

$$
\leq \left[\max_j \frac{1}{\sqrt{n}} \left\| D(\widehat{\Omega}^T - \Omega)e_j \right\|_2 \right] \frac{1}{\sqrt{n}} \left\| D(\hat{x}_n - x^*) \right\|_2
$$

$$
\leq \left[\max_j \sqrt{C_1\lambda_{\max}(A) \left\| (\widehat{\Omega}^T - \Omega)e_j \right\|_2^2 + C_2\frac{\log d}{n} \left\| (\widehat{\Omega}^T - \Omega)e_j \right\|_1^2} \right]
$$

$$
\sqrt{2\lambda_{\max}(A) \left\| \hat{x}_n - x^* \right\|_2^2 + \tau(n,d) \left\| \hat{x}_n - x^* \right\|_1^2}
$$

$$
\leq C_3 \max_j \sqrt{\frac{s_j \log d}{n} + \frac{s_j^2 \log^2 d}{n^2}} \sqrt{\frac{s \log d}{n} + \frac{s^2 \log^2 d}{n^2}}
$$

$$
\leq C_4\frac{s\log d}{n}
$$

where in the second inequality we used (Loh et al., 2012, Equation 2.13) and Rudelson and Zhou (2012). In the third inequality we used

$$
\left\| (\widehat{\Omega}^T - \Omega)e_j \right\|_2 \lesssim \sqrt{\frac{s_j \log d}{n}}, \quad \left\| (\widehat{\Omega}^T - \Omega)e_j \right\|_1 \lesssim s_j\sqrt{\frac{\log d}{n}}
$$

from Lemma E.2, $\|\hat{x}_n - x^*\|_2 \leq c_0'\sqrt{\frac{s \log d}{n}}$, and $\|\hat{x}_n - x^*\|_1 \leq c_0 s\sqrt{\frac{\log d}{n}}$ uniformly for all x^* from Agarwal, Negahban and Wainwright (2012) and Proposition 5.1. In the last inequality, we use $s = o(\sqrt{n}/\log d)$ and $s_j \leq Cs$ by Assumption 5.1.

By the assumption that $s = o(\sqrt{n}/\log d)$, we have that $\|\Delta\|_\infty \leq o_{\mathbb{P}_{x^*}}(\frac{1}{\sqrt{n}})$ uniformly in $x^* \in \mathcal{B}(s)$, which completes the proof.

$\square$

F. More Simulations. In this section, we report more simulation experiments by considering different design matrices Σ (recall that each covariate $a_n \sim N(0, \Sigma)$).

In Table 5, we report the performance of the linear regression case with Toeplitz covariance matrix of $\Sigma_{i,j} = r^{|i-j|}$ for different correlation parameter r's. As the correlation parameter r varies from 0.4 to 0.6, the coverage rates remains similar when d is small. On the other hand, when d is large, the coverage rates slightly decrease as r increases. At the meantime, the average lengths increase, which is consistent with the trend of the oracle lengths

TABLE 5

Linear Regression: the coverage rate, average length of confidence intervals and their standard errors, for the nominal coverage probability 95%. *The covariance matrix is in the Toeplitz setting, where* $\Sigma_{i,j} = r^{|i-j|}$ *for* $r = 0.4, 0.5, 0.6$, $i, j = 1, 2, \ldots, d$. *Standard errors are reported in the brackets.*

	d	Plug-in	BM			Oracle
			$M = n^{0.2}$	$M = n^{0.25}$	$M = n^{0.3}$	
$r = 0.4$						
Cov Rate (%)	5	95.16(0.69)	90.88(1.78)	92.56(0.93)	93.40(1.12)	87.44
Avg Len ($\times 10^{-2}$)		1.70(0.07)	1.60(0.06)	1.65(0.07)	1.69(0.06)	1.42
Cov Rate (%)	20	94.91(0.92)	91.39(1.04)	93.07(1.13)	94.26(0.96)	89.12
Avg Len ($\times 10^{-2}$)		1.66(0.04)	1.56(0.04)	1.61(0.04)	1.63(0.04)	1.45
Cov Rate (%)	100	94.79(1.00)	90.55(1.28)	92.00(1.22)	92.47(1.08)	88.26
Avg Len ($\times 10^{-2}$)		1.64(0.02)	1.53(0.02)	1.55(0.02)	1.54(0.02)	1.45
Cov Rate (%)	200	94.41(1.02)	90.22(1.23)	91.46(1.18)	91.83(1.19)	87.49
Avg Len ($\times 10^{-2}$)		1.61(0.01)	1.50(0.02)	1.50(0.01)	1.49(0.01)	1.46
$r = 0.5$						
Cov Rate (%)	5	95.24(0.92)	91.16(0.50)	94.28(0.86)	93.04(0.90)	88.31
Avg Len ($\times 10^{-2}$)		1.83(0.10)	1.74(0.10)	1.82(0.11)	1.78(0.12)	1.53
Cov Rate (%)	20	94.84(0.97)	90.97(1.08)	93.75(0.93)	92.77(0.81)	87.26
Avg Len ($\times 10^{-2}$)		1.81(0.05)	1.71(0.06)	1.78(0.06)	1.76(0.06)	1.58
Cov Rate (%)	100	95.01(1.12)	90.36(1.33)	91.83(1.09)	91.52(1.17)	89.11
Avg Len ($\times 10^{-2}$)		1.77(0.02)	1.67(0.03)	1.67(0.03)	1.69(0.02)	1.60
Cov Rate (%)	200	94.69(1.33)	90.01(1.41)	91.65(1.36)	91.24(1.41)	89.43
Avg Len ($\times 10^{-2}$)		1.74(0.02)	1.62(0.02)	1.62(0.02)	1.62(0.02)	1.60
$r = 0.6$						
Cov Rate (%)	5	94.92(1.21)	90.72(0.58)	93.20(0.87)	94.28(0.73)	87.34
Avg Len ($\times 10^{-2}$)		0.00(0.14)	1.91(0.14)	1.97(0.15)	2.03(0.17)	1.70
Cov Rate (%)	20	95.11(1.03)	91.15(1.25)	92.54(0.85)	93.46(0.86)	88.59
Avg Len ($\times 10^{-2}$)		0.00(0.07)	1.93(0.09)	1.97(0.09)	2.00(0.09)	1.78
Cov Rate (%)	100	94.76(1.19)	90.18(1.15)	91.18(1.20)	91.31(1.21)	89.11
Avg Len ($\times 10^{-2}$)		0.00(0.03)	1.87(0.04)	1.88(0.03)	1.86(0.03)	1.80
Cov Rate (%)	200	94.59(1.31)	90.02(1.29)	90.87(1.38)	91.09(1.46)	87.46
Avg Len ($\times 10^{-2}$)		0.00(0.02)	1.79(0.03)	1.78(0.02)	1.85(0.03)	1.80

(see the last column). In Table 6, we report the simulations with Equi Corr covariance matrix $\Sigma_{i,j} = r \cdot 1_{\{i \neq j\}} + 1_{\{i=j\}}$ when r varies from 0.1 to 0.3. The observation is similar to the case of Toeplitz covariance matrix. It is worth noting that for each of the cases, the minimum eigenvalue of the covariance matrix Σ decreases when r increases. Therefore Tables 5 and 6 show that our methods are robust with the minimum eigenvalue of covariance matrix Σ. Similar comparisons for logistic regression cases are provided in Tables 7 and 8.

TABLE 6

Linear Regression: the coverage rate, average length of confidence intervals and their standard errors, for the nominal coverage probability 95%. The covariance matrix is in the Equi Corr setting, where $\Sigma_{i,j} = r \cdot 1_{\{i \neq j\}} + 1_{\{i=j\}}$ for $r = 0.1, 0.2, 0.3$, $i, j = 1, 2, \ldots, d$. Standard errors are reported in the brackets.

	d	Plug-in	BM			Oracle
			$M = n^{0.2}$	$M = n^{0.25}$	$M = n^{0.3}$	
$r = 0.1$						
Cov Rate (%)	5	94.88(1.07)	90.84(1.62)	92.48(1.45)	94.40(1.26)	88.19
Avg Len ($\times 10^{-2}$)		1.52(0.01)	1.40(0.01)	1.45(0.01)	1.49(0.01)	1.26
Cov Rate (%)	20	94.98(0.94)	90.65(1.00)	92.49(1.26)	93.46(0.93)	87.41
Avg Len ($\times 10^{-2}$)		1.50(0.01)	1.40(0.01)	1.43(0.01)	1.45(0.01)	1.28
Cov Rate (%)	100	95.62(1.11)	90.67(1.33)	92.14(1.24)	92.92(1.15)	89.11
Avg Len ($\times 10^{-2}$)		1.47(0.01)	1.38(0.01)	1.41(0.01)	1.41(0.01)	1.30
Cov Rate (%)	200	94.61(1.10)	90.42(1.35)	92.05(1.12)	92.49(1.18)	88.40
Avg Len ($\times 10^{-2}$)		1.41(0.01)	1.37(0.01)	1.38(0.01)	1.38(0.01)	1.30
$r = 0.2$						
Cov Rate (%)	5	94.80(0.88)	90.92(1.09)	93.60(0.92)	92.32(0.68)	86.79
Avg Len ($\times 10^{-2}$)		1.60(0.01)	1.46(0.01)	1.55(0.01)	1.52(0.01)	1.31
Cov Rate (%)	20	95.10(0.99)	91.15(1.14)	93.66(0.99)	92.78(0.92)	88.04
Avg Len ($\times 10^{-2}$)		1.59(0.01)	1.47(0.01)	1.54(0.01)	1.51(0.01)	1.36
Cov Rate (%)	100	94.93(1.06)	90.86(1.26)	93.19(1.15)	92.29(1.10)	87.15
Avg Len ($\times 10^{-2}$)		1.56(0.01)	1.47(0.01)	1.52(0.01)	1.50(0.01)	1.38
Cov Rate (%)	200	94.49(1.09)	90.57(1.45)	92.45(1.27)	91.91(1.13)	87.22
Avg Len ($\times 10^{-2}$)		1.51(0.01)	1.45(0.01)	1.49(0.01)	1.49(0.01)	1.38
$r = 0.3$						
Cov Rate (%)	5	95.00(0.96)	91.80(1.34)	92.76(0.75)	94.16(1.21)	87.12
Avg Len ($\times 10^{-2}$)		1.74(0.01)	1.57(0.02)	1.60(0.02)	1.65(0.01)	1.38
Cov Rate (%)	20	95.27(1.00)	91.08(1.32)	92.60(1.28)	93.52(1.16)	86.91
Avg Len ($\times 10^{-2}$)		1.72(0.01)	1.58(0.01)	1.62(0.02)	1.63(0.01)	1.45
Cov Rate (%)	100	94.58(1.14)	90.68(1.25)	92.23(1.14)	93.35(1.11)	87.05
Avg Len ($\times 10^{-2}$)		1.71(0.01)	1.57(0.01)	1.61(0.02)	1.63(0.01)	1.47
Cov Rate (%)	200	94.69(1.25)	90.31(1.46)	91.73(1.27)	93.21(1.11)	88.32
Avg Len ($\times 10^{-2}$)		1.64(0.01)	1.55(0.01)	1.60(0.01)	1.61(0.01)	1.48

References.

AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. In *Proceedings of the Advances in Neural Information Processing Systems*.

BACH, F. and MOULINES, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for Machine learning. In *Proceedings of the Advances in Neural Information Processing Systems*.

BACH, F. and MOULINES, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Proceedings of the Advances in Neural Information Processing Systems*.

BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547.

BÜHLMANN, P. and MANDOZZI, J. (2014). High-dimensional variable screening and bias

TABLE 7

Logistic Regression: the coverage rate, average length of confidence intervals and their standard errors, for the nominal coverage probability 95%. The covariance matrix is in the Toeplitz setting, where $\Sigma_{i,j} = r^{|i-j|}$ for $r = 0.4, 0.5, 0.6$, $i,j = 1, 2, \ldots, d$. Standard errors are reported in the brackets.

	d	Plug-in	BM			Oracle
			$M = n^{0.2}$	$M = n^{0.25}$	$M = n^{0.3}$	
$r = 0.4$						
Cov Rate (%)	5	95.08(1.41)	89.04(2.19)	90.92(2.01)	90.28(1.97)	94.71
Avg Len ($\times 10^{-2}$)		3.74(0.28)	3.44(0.25)	3.36(0.24)	3.26(0.21)	3.70
Cov Rate (%)	20	94.69(1.33)	89.64(1.99)	90.49(1.87)	90.41(1.47)	93.89
Avg Len ($\times 10^{-2}$)		4.99(0.29)	4.65(0.24)	4.49(0.22)	4.41(0.22)	4.90
Cov Rate (%)	100	94.81(1.21)	89.38(1.84)	90.31(1.77)	90.26(1.92)	94.20
Avg Len ($\times 10^{-2}$)		7.21(0.34)	6.51(0.29)	6.33(0.26)	6.16(0.24)	7.15
Cov Rate (%)	200	94.54(1.04)	89.27(1.78)	90.22(1.69)	90.31(1.88)	93.96
Avg Len ($\times 10^{-2}$)		8.68(0.37)	7.88(0.31)	7.82(0.30)	7.71(0.28)	8.47
$r = 0.5$						
Cov Rate (%)	5	94.96(1.58)	88.96(2.32)	90.56(2.06)	90.12(2.04)	92.41
Avg Len ($\times 10^{-2}$)		4.06(0.34)	3.75(0.28)	3.73(0.27)	3.61(0.25)	4.04
Cov Rate (%)	20	95.17(1.23)	89.01(1.93)	90.39(1.88)	89.79(1.81)	91.07
Avg Len ($\times 10^{-2}$)		5.74(0.29)	5.57(0.25)	5.22(0.23)	4.95(0.22)	5.59
Cov Rate (%)	100	94.91(0.89)	89.91(1.74)	90.83(1.81)	90.54(1.97)	91.47
Avg Len ($\times 10^{-2}$)		8.47(0.37)	8.01(0.28)	7.71(0.26)	7.37(0.25)	8.28
Cov Rate (%)	200	94.59(1.04)	89.72(1.81)	90.74(1.93)	90.32(2.02)	92.29
Avg Len ($\times 10^{-2}$)		9.81(0.41)	9.24(0.34)	8.95(0.31)	8.78(0.29)	9.84
$r = 0.6$						
Cov Rate (%)	5	94.72(1.88)	89.32(2.38)	90.16(2.14)	89.96(2.16)	93.77
Avg Len ($\times 10^{-2}$)		4.69(0.47)	4.31(0.41)	4.22(0.39)	4.13(0.36)	4.54
Cov Rate (%)	20	95.11(1.49)	89.07(1.91)	90.44(1.97)	89.91(1.94)	94.62
Avg Len ($\times 10^{-2}$)		6.59(0.41)	6.38(0.36)	6.26(0.33)	6.11(0.31)	6.58
Cov Rate (%)	100	94.29(1.11)	90.14(1.79)	90.31(1.89)	90.42(1.82)	93.15
Avg Len ($\times 10^{-2}$)		10.01(0.39)	9.61(0.36)	9.11(0.34)	8.89(0.32)	9.93
Cov Rate (%)	200	94.47(1.28)	89.63(1.84)	90.17(1.98)	90.13(2.01)	93.27
Avg Len ($\times 10^{-2}$)		12.27(0.41)	11.39(0.37)	10.97(0.34)	10.66(0.31)	11.83

in subsequent inference, with an empirical comparison. *Comput. Statist.* **29** 407–430.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer.

BUJA, A., BERK, R., BROWN, L., GEORGE, E., TRASKIN, M., ZHANG, K. and ZHAO, L. (2013). A conspiracy of random X and model violation against classical inference in linear regression Technical Report, Wharton School of the University of Pennsylvania.

CHANG, F. C. (2006). Inversion of a perturbed matrix. *Appl. Math. Lett.* **19** 169–173.

DAMERDJI, H. (1991). Strong Consistency and Other Properties of the Spectral Variance Estimator. *Management Science* **37** 1424–1440.

FABIAN, V. (1968). On asymptotic normality in stochastic approximation. *Ann. Math. Statist* 1327–1332.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B* **70** 849–911.

TABLE 8

Logistic Regression: the coverage rate, average length of confidence intervals and their standard errors, for the nominal coverage probability 95%. The covariance matrix is in the Equi Corr setting, where $\Sigma_{i,j} = r \cdot 1_{\{i \neq j\}} + 1_{\{i=j\}}$ for $r = 0.1, 0.2, 0.3$, $i, j = 1, 2, \ldots, d$. Standard errors are reported in the brackets.

	d	Plug-in	BM			Oracle
			$M = n^{0.2}$	$M = n^{0.25}$	$M = n^{0.3}$	
$r = 0.1$						
Cov Rate(%)	5	94.72(1.75)	88.80(1.36)	89.68(1.58)	89.56(1.45)	94.09
Avg Len($\times 10^{-2}$)		3.39(0.32)	3.04(0.26)	2.97(0.24)	2.90(0.23)	3.21
Cov Rate(%)	20	94.85(1.82)	89.48(1.48)	90.89(1.48)	90.68(1.56)	94.12
Avg Len($\times 10^{-2}$)		4.94(0.32)	4.07(0.24)	4.00(0.23)	3.82(0.22)	4.48
Cov Rate(%)	100	95.11(1.58)	89.41(1.52)	90.51(1.55)	90.15(1.70)	93.88
Avg Len($\times 10^{-2}$)		9.06(0.47)	8.45(0.42)	8.24(0.40)	8.01(0.37)	8.83
Cov Rate(%)	200	94.55(1.46)	88.95(1.58)	89.87(1.30)	89.60(1.79)	92.46
Avg Len($\times 10^{-2}$)		12.79(0.61)	11.99(0.55)	11.42(0.52)	11.28(0.52)	12.33
$r = 0.2$						
Cov Rate (%)	5	94.80(1.66)	88.08(1.46)	88.64(1.73)	89.48(1.51)	93.79
Avg Len ($\times 10^{-2}$)		3.43(0.35)	3.28(0.28)	3.24(0.25)	3.20(0.24)	3.38
Cov Rate (%)	20	94.54(1.73)	89.27(1.33)	90.64(1.60)	90.31(2.10)	92.50
Avg Len ($\times 10^{-2}$)		5.37(0.31)	4.84(0.26)	4.77(0.24)	4.51(0.21)	5.19
Cov Rate (%)	100	94.79(1.08)	89.01(1.70)	90.27(1.76)	89.42(2.01)	94.92
Avg Len ($\times 10^{-2}$)		10.24(0.51)	10.17(0.47)	9.75(0.42)	9.24(0.40)	10.89
Cov Rate (%)	200	94.24(1.09)	89.13(1.44)	90.01(1.92)	89.23(1.79)	92.40
Avg Len ($\times 10^{-2}$)		15.70(0.62)	14.82(0.57)	14.01(0.55)	13.88(0.52)	15.31
$r = 0.3$						
Cov Rate(%)	5	95.00(1.84)	89.12(1.54)	89.68(1.77)	89.28(1.64)	93.42
Avg Len($\times 10^{-2}$)		3.65(0.33)	3.44(0.26)	3.39(0.24)	3.31(0.23)	3.61
Cov Rate(%)	20	94.86(1.79)	89.25(1.45)	90.42(1.59)	89.84(1.63)	93.69
Avg Len($\times 10^{-2}$)		5.98(0.33)	5.69(0.25)	5.54(0.24)	5.48(0.21)	5.91
Cov Rate(%)	100	94.57(1.44)	89.04(1.71)	90.07(1.65)	89.22(1.68)	94.01
Avg Len($\times 10^{-2}$)		13.07(0.64)	12.15(0.48)	11.80(0.46)	11.64(0.41)	12.77
Cov Rate(%)	200	94.21(1.55)	89.00(1.62)	90.11(1.74)	89.07(1.78)	92.19
Avg Len($\times 10^{-2}$)		19.41(0.69)	17.15(0.60)	16.70(0.59)	16.29(0.55)	18.01

FISHMAN, G. (1996). *Monte Carlo: Concepts, Algorithms and Applications.* Chapman and Hall.

FLEGAL, J. and JONES, G. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.* **38** 1034–1070.

GEYER, C. (1992). Practical Markov Chain Monte Carlo. *Statist. Sci.* **7** 473–483.

GHADIMI, S. and LAN, G. (2012). Optimal Stochastic Approximation Algorithms for Strongly Convex Stochastic Composite Optimization I: A Generic Algorithmic Framework. *SIAM J. Optim.* **22** 1469–1492.

GLYNN, P. W. and IGLEHART, D. L. (1990). Simulation analysis using standardized time series. *Math. Oper. Res.* **15** 1–16.

GLYNN, P. W. and WHITT, W. (1991). Estimating the asymptotic variance with batch means. *Oper. Res. Lett.* **10** 431 - 435.

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing

for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909.

JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-Width Output Analysis for Markov Chain Monte Carlo. *J. Amer. Statist. Assoc.* **101** 1537–1547.

LOH, P.-L., WAINWRIGHT, M. J. et al. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **3** 1436–1462.

MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). P-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681.

NEMIROVSKI, A., JUDITSKY, A., LAN, G. and SHAPIRO, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** 1574–1609.

NESTEROV, Y. and VIAL, J.-P. (2008). Confidence level solutions for stochastic programming. *Automatica* **44** 1559–1568.

NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195.

POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** 838–855.

RAKHLIN, A., SHAMIR, O. and SRIDHARAN, K. (2012). Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. In *Proceedings of the International Conference on Machine Learning*.

ROBBINS, H. and MONRO, S. (1951). A Stochastic Approximation Method. *Ann. Math. Statist.* **22** 400–407.

ROUX, N. L., SCHMIDT, M. and BACH, F. (2012). A Stochastic Gradient Method with an Exponential Convergence Rate for Strongly-Convex Optimization with Finite Training Sets. In *Proceedings of the Advances in Neural Information Processing Systems*.

RUDELSON, M. and ZHOU, S. (2012). Reconstruction from anisotropic random measurements. In *Proceedings of the Conference on Learning Theory*.

RUPPERT, D. (1988). Efficient estimations from a slowly convergent Robbins-Monro process Technical Report, Cornell University Operations Research and Industrial Engineering.

SREBRO, N. and TEWARI, A. (2010). Stochastic Optimization for Machine Learning. Tutorial at International Conference on Machine Learning.

SULLIVAN, T. J. (2015). *Introduction to Uncertainty Quantication. Texts in Applied Mathematics* **63**. Springer.

TOULIS, P. and AIROLDI, E. M. (2016). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Ann. Statist.* **45** 1694-1727.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202.

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *Preprint. Available at arXiv:1011.3027*.

WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programs. *IEEE Trans. Inform. Theory* **55** 2183–2202.

XIAO, L. (2010). Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization. *J. Mach. Learn. Res.* **11** 2543-2596.

XIAO, L. and ZHANG, T. (2014). A proximal stochastic gradient method with progressive variance reduction. *SIAM J. Optim.* **24** 2057–2075.

ZHANG, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the International Conference on Machine*

Learning.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. Roy. Statist. Soc. Ser. B* **76** 217–242.

Information, Operations and Management Sciences
Stern School of Business
New York University
New York, New York 10012
USA
E-mail: xchen3@stern.nyu.edu; yzhang@stern.nyu.edu

Data Sciences and Operations
Marshall School of Business
University of Southern California
Los Angeles, CA 90089-0808
USA
E-mail: jasonlee@marshall.usc.edu

Department of Mathematics
National University of Singapore
Singapore, 119076
Republic of Singapore
E-mail: mattxin@nus.edu.sg