Journal of Machine Learning Research 27 (2026) 1-42 Submitted 1/25; Published 3/26

Cheap Bootstrap for Fast Uncertainty Quantification of Stochastic Gradient Descent

Henry Lam HENRY.LAM@COLUMBIA.EDU
Industrial Engineering and Operations Research
Columbia University
500 West 120th Street
New York, NY 10027, USA

Zitong Wang ZW2690@COLUMBIA.EDU
Industrial Engineering and Operations Research
Columbia University
500 West 120th Street
New York, NY 10027, USA

Editor: Jianfeng Lu

Abstract

Stochastic gradient descent (SGD) or stochastic approximation has been widely used in model training and stochastic optimization. While there is a huge literature on analyzing its convergence, inference on the obtained solutions from SGD has only been recently studied, yet it is important due to the growing need for uncertainty quantification. We investigate two computationally cheap resampling-based methods to construct confidence intervals for SGD solutions. One uses multiple, but few, SGDs in parallel via resampling with replacement from the data, and another operates this in an online fashion. Our methods can be regarded as enhancements of established bootstrap schemes to substantially reduce the computation effort in terms of resampling requirements, while bypassing the intricate mixing conditions in existing batching methods. We achieve these via a recent so-called cheap bootstrap idea and refinement of a Berry-Esseen-type bound for SGD.

Keywords: stochastic gradient descent, bootstrap resampling, confidence intervals, Berry-Esseen bounds, statistical inference

1. Introduction

Stochastic optimization commonly arises in many applications across machine learning, operations research, and scientific analysis. The problem can be formulated as:

$$\min_{x\in\mathbb{R}^d} H(x) \triangleq \mathbb{E}_{\zeta\sim P}[h(x,\zeta)], \tag{1}$$

in which P is an underlying data distribution governing the randomness $\zeta \in \Omega$, and h is a known real-valued function. Stochastic gradient descent (SGD) or stochastic approximation is a popular numerical approach to solve (1). With an initial guess $x_0 \in \mathbb{R}^d$, SGD iteratively

updates the solution using

$$x_{t+1} = x_t - \eta_t \nabla h(x_t, \zeta_{t+1}), \quad t = 0, \ldots, n-1, \tag{2}$$

where ζ_t is a sample drawn using a Monte Carlo model generator or real data. The Robbins-Monro procedure (Robbins and Monro, 1951) outputs x_n after a large number of iterations (2). Alternatively, one might take the average $\bar{x}_n \triangleq \frac{1}{n}\sum_{t=1}^{n} x_t$ as the output. This is known as the Polyak-Ruppert-Juditsky averaging (Polyak and Juditsky, 1992), and for convenience in this paper, we call it averaged stochastic gradient descent (ASGD). Both approaches are prevalent, with ASGD known to be more robust with respect to the step size η_t (Rakhlin et al., 2012).

We aim to conduct inference or quantify statistical uncertainty in SGD. More specifically, our goal is to construct a $1-\gamma$ confidence interval for (each component of) the true optimal solution x^* of problem (1) using the iterates (2). Despite the popularity of SGD, to our best knowledge, this problem has been systematically studied only recently, driven by applications in exploration (Lattimore and Szepesvári, 2020) and as stopping criteria (Su and Zhu, 2023; Fang et al., 2018; Chen et al., 2020). In the following, we first review these recent approaches, discuss their main ideas as well as challenges, which then motivate our proposal in this paper.

1.1 Existing Methods and Challenges

One of the primary challenges in SGD inference arises from the serial dependence manifested by the sequence x_t. This dependence makes the construction of a consistent standard error estimator intricate. Several recent works aim to address this issue and, though with its own merits, each of these proposed approaches also encounters limitations. Chen et al. (2020) proposed two methods, one based on the delta method that directly approximates the asymptotic covariance of the gradient $\nabla h(x^*, \zeta)$ and the Hessian $\nabla^2 H(x^*)$ at the optimum. While this method is statistically valid, the required Hessian information is not always available in the context of SGD. For example, backpropagation can only provide first-order gradient information (Rumelhart et al., 1986), and arguably, a major advantage of SGD lies in its Hessian-free nature. Moreover, storing a Hessian matrix requires an expensive $\mathcal{O}(d^2)$ space. These put aside the subtle regularity assumptions needed for consistency as noted by Chen et al. (2020) themselves. Along this vein, Xie et al. (2024) also proposed an inference tool for SA that gives a confidence sequence (Howard et al., 2021; Ramdas et al., 2023) based on the delta method with an asymptotic time-uniform coverage guarantee.

Motivated by the previously mentioned challenges, Chen et al. (2020)'s second method borrows the batch mean idea in stochastic simulation output analysis (Glynn and Iglehart, 1990; Schmeiser, 1982; Schruben, 1983; Glynn and Lam, 2018) and Markov Chain Monte Carlo (Geyer, 1992; Flegal and Jones, 2010; Jones et al., 2006). This approach divides the iterations of SGD into M batches of increasing sizes and aggregates the means of these batches to construct confidence intervals. Nonetheless, the batch mean method introduces the number of batches as a hyperparameter that needs to be tuned. Additionally, experiments show that this method is more sensitive to the quality of convergences of SGD and could underperform other methods. Relatedly, Li et al. (2018) presented a batch mean method for inference in M-estimation by using an SGD trajectory with a constant step size.

Instead of using batches with increasing lengths, they use batches with a fixed length but separated by gaps to overcome the dependence between iterations of SGD. Zhu and Dong (2021) studied a batch mean algorithm by elegantly canceling out the asymptotic covariance matrix of a rescaled SGD using an F-type statistic. The above batching methods require hyperparameter tuning like Chen et al. (2020) such as the batch sizes and gap between batches.

Compared to batch mean methods, Lee et al. (2022) developed a method that directly accounts for the dependency along the SGD trajectory. By leveraging the random scaling technique, they constructed an asymptotically pivotal statistic to produce a confidence interval. This method nicely avoids the hyperparameter tuning faced in batching. However, it needs to update a $d \times d$ matrix on the fly to construct the random scaling matrix that induces some computational and memory overheads. Also using just one pass of data, Chee et al. (2023) proposed a simple and scalable method that gives conservative confidence intervals based on the initial learning rate. The performance of this method relies on the estimation quality of the reciprocal of the smallest eigenvalue of the asymptotic covariance matrix.

Another approach is to use the bootstrap, which, advantageously, does not succumb to the computation load of variance estimation as well as the tuning and sensitivity challenges associated with batch sizes. Fang et al. (2018) developed an online bootstrap method that persistently maintains B perturbed version of SGD estimates, updated upon each data arrival. However, as in other applications of the bootstrap, for their method to be effective, a large value of B is necessary. For linear regression problems of dimensions 10 or 20, they set $B = 200$, which means 200 times more computational cost compared to running the SGD itself or using batch means.

Yet another method, HiGrad, was proposed by Su and Zhu (2023). This approach is rooted in "splitting" an SGD trajectory. The process involves initially running SGD for a set number of steps. Once complete, the result of this iteration is used as a starting point for the next stage, where multiple SGD threads branch off, each utilizing different new data. This branching process continues for the outcome of each thread until all data is exhausted. Confidence intervals are then constructed using all the obtained split outcomes. HiGrad requires a substantial modification to the original SGD runs; in fact, there is no more "original" run of SGD in HiGrad.

Finally, we briefly mention a line of work on quantifying algorithmic randomness. This includes Lunde et al. (2021), which applied the bootstrap on streaming principal component analysis (Oja, 1982), and Chen and Lopes (2020), which investigated randomized Newton methods. Lopes (2019) and Lopes et al. (2020) utilized bootstrap methods to estimate the algorithmic variability for random ensembles such as bagging and random forests. Furthermore, Nesterov and Vial (2008) gave a complexity bound on the number of iterations of their method in relation to the confidence level on reaching the optimal value via SGD. However, all these works focus on assessing the uncertainty from algorithmic randomness and treat the data as fixed. As such, they are less relevant to our focus in this paper.

1.2 Our Contributions

Our discussion above reveals that existing approaches in SGD inference, while being carefully and elegantly designed, encounter either intricate algorithmic tuning that relates to mixing conditions (batching), substantial modification on the SGD itself (HiGrad), or computation and storage challenges (delta method, random scaling method, and online bootstrap). In this paper, we study a methodology designed to surmount these challenges concurrently. More precisely, we adopt the bootstrap approach, which does not require mixing-related tuning nor substantial modification to the original SGD. At the same time, we enhance the bootstrap to make it substantially lighter in terms of resampling cost. The latter is made possible by using a recent "cheap bootstrap" idea (Lam, 2022a; Lam and Liu, 2023; Lam, 2022b) that we will describe in more detail momentarily.

Our methodology can be implemented in both offline and online fashions. The offline version, which we call the *Cheap Offline Bootstrap (COfB)*, reruns the SGD using resampling with replacement from the data B times and constructs confidence intervals from these resampled iterates via an approach similar to the standard error bootstrap. However, while this approach may appear to require heavy resampling effort, our key assertion is that the B in our implementation can be very small (such as 3). In this way, our approach is computationally less demanding than the delta method (Chen et al., 2020) and online bootstrap (Fang et al., 2018), does not require hyperparameter tuning in batch mean (Chen et al., 2020; Zhu and Dong, 2021), and also does not substantially modify the SGD trajectory in HiGrad (Su and Zhu, 2023).

A caveat of COfB is that we can only rerun SGD after all the data becomes available. Thus, it cannot be used in a single-pass streaming fashion. To address this, our online version, *Cheap Online Bootstrap (COnB)*, runs multiple, namely $B+1$, SGDs in parallel on the fly as new data comes in. COnB borrows the idea of Fang et al. (2018) in perturbing the gradient estimate in the SGD iteration. However, like COfB, it is computationally much cheaper than Fang et al. (2018) as it only needs to maintain a very small number of SGD runs. In both our theory and experimentation, we illustrate that using $B = 3$ already produces consistently better coverage than the existing approaches.

Our methodology synthesizes two recent ideas. One, as mentioned earlier, is the recent cheap bootstrap idea. This approach integrates the analysis of the statistical error coming from the original data and the Monte Carlo error in approximating the resample distribution together, in contrast to separated treatment in conventional bootstraps. To explain, while conventional bootstraps rely on the approximation of the sampling distribution by the resample distribution, which is in turn approximated by running Monte Carlo to generate many realized resamples, the cheap bootstrap directly utilizes the joint distribution between the original estimate and each resample estimate to construct pivotal statistics. Subsequently, it allows the use of a minimal number of resample runs, i.e., potentially as low as $B = 1$, while maintaining large-sample exact coverages. However, it also results in longer intervals when B is small. Nonetheless, as discussed in Lam (2022a) and Lam and Liu (2023), the interval length advantageously shrinks quickly as B increases away from 1.

Our second main methodological element is to derive the asymptotic joint distribution, in particular independence, among SGD and resampled SGD's required in invoking the cheap bootstrap idea. More specifically, we prove a joint central limit theorem for both the

Table 1: Comparison among different methods. "delta" denotes the delta method in Chen et al. (2020), "BM" denotes the batch mean method in Chen et al. (2020), "RS" denotes the random scaling method in Lee et al. (2022), and "OB" denotes the online bootstrap method in Fang et al. (2018).

Property \ Method	COfB	COnB	delta	BM	RS	OB	HiGrad
Require substantial procedural modification	No	No	No	No	No	No	Yes
Computational/memory Load	light	light	heavy	light	heavy	heavy	light
Hyperparameter tuning	No	No	No	Yes	No	Yes	Yes
Require second derivative	No	No	Yes	No	No	No	No

original and resampled SGD runs when resampling with replacement. This subsequently guides us in suitably aggregating the outputs to construct asymptotically exact-coverage intervals. To this end, we generalize the recent non-asymptotic bounds for ASGD studied by Shao and Zhang (2022) and Anastasiou et al. (2019) to hold uniformly for both the original and resampled runs, under both SGD and ASGD settings.

Table 1 summarizes the comparisons between our methods and benchmark techniques. HiGrad requires substantial changes to the SGD procedure, while other methods do not involve such changes. The delta method, random scaling, and online bootstrap demand a relatively heavy computation or memory load. The first method requires memorizing a d by d Hessian approximation, the second method requires updating the d by d scaling matrix, and the third method requires maintaining a large number of perturbed trajectories B. Although our methods also introduce B, it can be kept very small, so we consider our methods light in terms of computational and memory load. As discussed in the previous section, the batch mean method, online bootstrap method, and HiGrad introduce hyperparameters that need to be tuned. For our methods, B can be regarded as a hyperparameter as well, but this is typically selected to be the largest integer that fits in the computation budget, keeping in mind that B as low as 1 or 2 already suffices to construct coverage-valid intervals while a larger B would improve the interval width. Lastly, the second derivative is only required by the delta method, which as discussed before can be a challenge since in some application scenarios of SGD, the second-order information may not be available.

Finally, we conduct experiments that support our statements in several aspects. We compare our methods with established methods in the regression experiment regarding coverage probabilities and widths of confidence intervals. The results indicate that our methods generally deliver the most accurate coverage probabilities. Although our methods produce wider confidence intervals, the interval width decreases sharply when B increases even slightly. In addition, our experiments also suggest that our method outperforms others in terms of robustness. Lastly, we analyze and apply our methods in high-dimensional sparse linear regression to enlarge the scope of applicability for our approach.

2. Methodology

Denote the underlying data distribution by P. Let x_t be the solution obtained in the t-th iteration of (2). So the output of SGD is x_n and the output of ASGD is $\bar{x}_n = \frac{1}{n}\sum_{t=1}^{n} x_t$. Let $\hat{P}_n$ denote the empirical distribution from data $\{\zeta_t\}_{t=1}^n$, i.e., $\hat{P}_n(\cdot) = \frac{1}{n}\sum_{t=1}^{n} I(\zeta_t \in \cdot)$,

where $I(\cdot)$ denotes the indicator function. We also use $(\cdot)_i$ to denote the i-th entry of a vector and $(\cdot)_{i,j}$ to denote the (i,j)-th entry of a matrix.

Our first method, COfB, works as follows. After obtaining $\bar{x}_n$ with data $\{\zeta_t\}_{t=1}^n$, we repeatedly resample with replacement from the data (i.e., draw n observations from $\hat{P}_n$) and run ASGD on the resampled data for B times. Denote the b-th resample output by x^{*b}_{COfB}, $b = 1, \ldots, B$. Then, the $1-\gamma$ confidence interval for the i-th entry of x^*, $i = 1, \ldots, d$, is given by

$$\mathcal{I}^{\mathrm{COfB}}_{i,n} = \left[(\bar{x}_n)_i - t_{B,1-\frac{\gamma}{2}} s_i^{\mathrm{ASGD}}, (\bar{x}_n)_i + t_{B,1-\frac{\gamma}{2}} s_i^{\mathrm{ASGD}}\right], \tag{3}$$

where $s_i^{\mathrm{ASGD}} \triangleq \sqrt{\frac{1}{B}\sum_{b=1}^B \left((x^{*b}_{\mathrm{COfB}})_i - (\bar{x}_n)_i\right)^2}$, and the scalar $t_{B,1-\frac{\gamma}{2}}$ denotes the $1-\frac{\gamma}{2}$ quantile of the student-t distribution with degree of freedom B. Importantly, in this construction, the number of reruns B is not necessarily large and can be any integer at least 1. A pseudocode for COfB can be found in Algorithm 1.

Algorithm 1 Cheap Offline Bootstrap (COfB)

1: **Input:** *i.i.d.* data $\{\zeta_t\}_{t=1}^n$, number of bootstrap runs $B \geq 1$, step size sequence $\{\eta_t\}$, initial guess x_0, nominal coverage level $1-\gamma$.
2: **Output:** $\mathcal{I}^{\mathrm{COfB}}_{i,n}$, $i = 1, \ldots, d$.
3: Run ASGD (2) to obtain $\bar{x}_n$.
4: **for** $b \leftarrow [1, 2, \ldots, B]$ **do**
5: Resample with replacement from $\{\zeta_t\}_{t=1}^n$ to obtain $\{\zeta_1^{*b}, \ldots, \zeta_n^{*b}\}$.
6: Run ASGD for n steps on $\{\zeta_t^{*b}\}_{t=1}^n$ with initialization x_0 to obtain x^{*b}_{COfB}.
7: **end for**
8: **for** $i \leftarrow [1, 2, \ldots, d]$ **do**
9: $s_i^{\mathrm{ASGD}} \leftarrow \sqrt{\frac{1}{B}\sum_{b=1}^B \left((x^{*b}_{\mathrm{COfB}})_i - (\bar{x}_n)_i\right)^2}$
10: $\mathcal{I}^{\mathrm{COfB}}_{i,n} \leftarrow \left[(\bar{x}_n)_i - t_{B,1-\frac{\gamma}{2}} s_i^{\mathrm{ASGD}}, (\bar{x}_n)_i + t_{B,1-\frac{\gamma}{2}} s_i^{\mathrm{ASGD}}\right]$
11: **end for**

Moreover, we can replace the ASGD procedure in lines 3 and 6 of Algorithm 1 by standard SGD. In this case, the radius of confidence intervals becomes $t_{B-1,1-\frac{\gamma}{2}} s_i^{\mathrm{SGD}}$, with $s_i^{\mathrm{SGD}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^B \left((x^{*b}_{\mathrm{COfB}})_i - (\overline{x}^*_{\mathrm{COfB}})_i\right)^2}$, $(\overline{x}^*_{\mathrm{COfB}})_i \triangleq \frac{1}{B}\sum_{b=1}^B (x^{*b}_{\mathrm{COfB}})_i$. Here we require B to be at least 2. We will refer to this method by *COfB running SGD* in the following discussions.

Note that COfB is an offline algorithm since resampling from $\{\zeta_i\}_{i=1}^n$ can only be accomplished when all the data points have been obtained. In contrast, our second method, COnB, works by maintaining $B+1$ parallel runs of ASGD starting from the same initialization. One of these trajectories is the original run following exactly (2). The other B trajectories update similarly, except that the gradient estimate $\nabla h(x_t, \zeta_{t+1})$ is perturbed by a factor $W_{t,b}$ following exponential distribution with rate 1. The confidence intervals $\mathcal{I}^{\mathrm{COnB}}_{i,n}$ are constructed in the same way as COfB with $\bar{x}_n$ and $\{x^{*b}_{\mathrm{COnB}}\}_{b=1}^B$. When new data ζ_t arrives, COnB uses only $B+1$ gradient calculations to update the original and resampled outputs. Moreover, like COfB, B is not necessarily large, and in this method, it can be any positive integer. A pseudocode of COnB is in Algorithm 2.

Algorithm 2 Cheap Online Bootstrap (COnB)

Input: *i.i.d.* data $\{\zeta_t\}_{t=1}^n$, number of bootstrap runs $B \geq 1$, step size sequence $\{\eta_t\}$, initial guess x_0, nominal coverage level $1-\gamma$.
Output: $\mathcal{I}_{i,n}^{\mathrm{COnB}},\ i = 1, \ldots, d.$
for $t \leftarrow [1, 2, \ldots, n]$ **do**
  $x_t \leftarrow x_{t-1} - \eta_t \nabla h(x_{t-1}, \zeta_t)$
  for $b \leftarrow [1, 2, \ldots, B]$ **do**
    Randomly generate $W_{t,b}$ from exponential distribution with rate 1.
    $x_t^{*b} \leftarrow x_{t-1}^{*b} - \eta_t W_{t,b} \nabla h(x_{t-1}^{*b}, \zeta_t)$
  end for
end for
for $b \leftarrow [1, 2, \ldots, B]$ **do**
  $x_{\mathrm{COnB}}^{*b} \leftarrow \frac{1}{n}\sum_{t=1}^n x_t^{*b}$
end for
for $i \leftarrow [1, 2, \ldots, d]$ **do**
  $s_i^{\mathrm{ASGD}} \leftarrow \sqrt{\frac{1}{B}\sum_{b=1}^B \left((x_{\mathrm{COnB}}^{*b})_i - (\bar{x}_n)_i\right)^2}$
  $\mathcal{I}_{i,n}^{\mathrm{COnB}} \leftarrow \left[(\bar{x}_n)_i - t_{B,1-\frac{\gamma}{2}} s_i^{\mathrm{ASGD}}, (\bar{x}_n)_i + t_{B,1-\frac{\gamma}{2}} s_i^{\mathrm{ASGD}}\right]$
end for

3. Main Theoretical Guarantees

Our main theoretical guarantees on COnB and COfB is on coverage exactness, asymptotically as n increases, for B fixed to be as low as either one or two. To explain and state this result, let $H_n(\cdot) = \frac{1}{n}\sum_{i=1}^n h(x, \zeta_i)$ denote the sample average approximation (SAA) of (1) and $\hat{x}_n$ the minimizer of $H_n(\cdot)$. $\|x\|_p$ denotes $(\mathbb{E}[\|x\|^p])^{\frac{1}{p}}$ for a random variable x and $\|\cdot\|$ denotes the standard Euclidean 2-norm for vectors. Let $\mathcal{X}$ be a closed and bounded neighborhood of x^*. For each i, j, define the function class $\mathcal{F}_{i,j} = \{\partial_{i,j}^2 h(x, \zeta) | x \in \mathcal{X}\}$. These function classes represent the scopes of the higher-order terms of the Taylor expansion of H at x^*. Let $G(x) = \nabla^2 H(x)$ and $S(x) = \mathbb{E}[\nabla h(x, \zeta)(\nabla h(x, \zeta))^\top]$ be the Hessian of H and covariance matrix of $\nabla h(x, \zeta)$ respectively. Given n data points, define $G_n(x) = \frac{1}{n}\sum_{i=1}^n \nabla^2 h(x, \zeta_i)$ and $S_n(x) = \frac{1}{n}\sum_{i=1}^n \nabla h(x, \zeta_i)(\nabla h(x, \zeta_i))^\top$.

Assumption 1 *h and H are twice continuously differentiable in x. The eigenvalues of $\nabla^2 h(x, \zeta)$ lie in $[l, L]$ for some positive real numbers $0 < l < L$ for all x, ζ. Moreover, $\nabla^2 h$ is Lipschitz continuous in x with constant l_1 on $\mathcal{X}$ uniform in ζ, i.e., for all ζ*

$$\|\nabla^2 h(x_1, \zeta) - \nabla^2 h(x_2, \zeta)\| \leq l_1 \|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathcal{X}. \tag{4}$$

Assumption 2 *Consider the filtration $\{\mathcal{F}_t = \sigma(\zeta_k | k \leq t)\}_{t \geq 0}$. The noise of estimated gradient is a martingale-difference process, i.e.,*

$$\mathbb{E}[\nabla h(x_{t-1}, \zeta_t) - \nabla H(x_{t-1}) | \mathcal{F}_{t-1}] = 0 \quad a.s..$$

Assumption 3 *There are $\tau_0, \tau > 0$ such that $\|x_0 - x^*\| \leq \tau_0$ and $\|\nabla h(x^*, \zeta)\|_4 \leq \tau$. The eigenvalues of $S(x^*) = \mathbb{E}[\nabla h(x^*, \zeta)(\nabla h(x^*, \zeta))^\top]$ lie in the interval $[\lambda_1, \lambda_2]$ for some positive constants $\lambda_1 < \lambda_2$.*

Assumption 1 specifies that the objective function h exhibits strong convexity along with a bounded Hessian, which implies the same property holds for H, in particular its strong convexity. Thus, it guarantees the existence and uniqueness of x^* that satisfies the first-order optimality condition $\nabla H(x^*) = 0$. The role of the sample-wise strong convexity assumption is to guarantee that the empirical Hessians for both the original and bootstrap samples are uniformly well-conditioned, which ensures sufficient control on the aggregate errors exhibited in the (A)SGD trajectory. This assumption could potentially be replaced by high probability lower bounds on the smallest eigenvalues of G_n and its bootstrap versions, although we do not pursue this relaxation for this work (nonetheless, see Appendix C for some discussions).

(4) is introduced to prove that the estimation error of SGD induced by non-linearity of the gradient of objective function vanishes. In fact, G and G_n, conditional on data, being Lipschitz at optimum is sufficient for this purpose.

Assumption 2 stipulates that the evaluation noise in the first-order gradient oracle is unbiased, which is a standard assumption to ensure the convergence of (A)SGD. Assumption 3 limits the variability of $\nabla h(x, \zeta)$, which is required to establish asymptotic normality. A short discussion on how Assumptions 1 to 3 imply the assumptions in Shao and Zhang (2022), which we utilize later in this paper, can be found in Appendix A.3.

We also make the following assumption regarding the uniform convergence of H_n to H, which is mildly stronger than pointwise convergence given that $\mathcal{X}$ is compact. This assumption, together with Assumption 1, will guarantee that the SAA solution $\hat{x}_n$ is consistent in the sense that $\hat{x}_n$ converges to x^* almost surely.

Assumption 4 *$\hat{x}_n \in \mathcal{X}$ with probability 1 for n large enough, and*

$$\sup_{x \in \mathcal{X}} |H_n(x) - H(x)| \to 0 \quad w.p.1.$$

The following two assumptions are specialized for ASGD and SGD considered in this work respectively. The specific choice of step size guarantees the convergence of (A)SGD in distribution. The Glivenko-Cantelli assumption helps us bridge the gap between the asymptotic distributions of the residual of the original output and the resampled output.

Assumption 5 *The step size satisfies $\eta_t = \eta t^{-\alpha}$ for some $\alpha \in (\frac{1}{2}, 1]$. For each i, j, function class $\mathcal{F}_{i,j}$ is P-Glivenko-Cantelli.*

Assumption 6 *The step size is $\eta_t = \eta t^{-1}$, and the initial step size η satisfies $\eta l > \frac{1}{2}$. For each i, j, function class $\mathcal{F}_{i,j}$ is P-Glivenko-Cantelli.*

Essentially, a function class is Glivenko-Cantelli if the law of large numbers holds uniformly over the class. Given the continuity of $\nabla^2 h$ in x and compactness of $\mathcal{X}$, $\mathcal{F}_{i,j}$ is Glivenko-Cantelli if it admits an integrable envelope function.

With the above assumptions, we have the following theorem:

Theorem 1 *Under Assumptions 1 to 5, for any fixed $B \geq 1$, $i = 1, \ldots, d$, and $\gamma \in (0, 1)$, the confidence intervals for the i-th entry generated by Algorithms 1 and 2 are asymptotically exact in the sense*

$$\lim_{n \to \infty} \mathbb{P}(x_i^* \in \mathcal{I}_{i,n}^{COfB}) = 1 - \gamma, \quad \lim_{n \to \infty} \mathbb{P}(x_i^* \in \mathcal{I}_{i,n}^{COnB}) = 1 - \gamma. \tag{5}$$

Moreover, under Assumptions 1 to 4 and 6, for any fixed $B \geq 2$, $i = 1, \ldots, d$, and $\gamma \in (0, 1)$, the confidence interval for the i-th entry generated by COfB running SGD is also asymptotically exact in the sense

$$\lim_{n\to\infty} \mathbb{P}(x_i^* \in \mathcal{I}_{i,n}^{COfB}) = 1 - \gamma. \tag{6}$$

Theorem 1 states that COfB and COnB attain asymptotically exact coverage as the sample size $n \to \infty$, *regardless of any fixed choice of* B. Note the subtlety that COfB running SGD requires $B \geq 2$, but our methods running ASGD are valid even for B as small as 1. This discrepancy comes from the slight difference in the joint asymptotic limits among the original and resample runs of SGD and ASGD, which will be discussed in Theorem 4 in the following section. Moreover, note that COnB works only for ASGD. Whether it will work for SGD is still open to us due to the delicacy of the asymptotic behavior for SGD in this case.

We conclude this section by remarking on the expected width of the confidence intervals generated by our methods. As will be evidenced by our subsequent analyses, our intervals are based on t-statistic construction and thus follow the behavior of t-intervals. Specifically, our widths are larger than those of normality intervals, but shrink rapidly as B increases. In our experiments, it appears that using a small B, such as $B = 3$ or 5, already gives a good balance in execution time and expected interval width. In addition, this moderate compensation in the interval width significantly improves coverage probability over the benchmark methods. Our experimental discussions in Section 6.1.3 provide more details on these observations.

4. Ideas behind the Main Guarantees

In this section, we delineate the development of Theorem 1 in three layers. First, we establish the conditional convergence for the error of the resample runs of our methods, which is widely utilized in classical bootstraps. For COnB, we borrow this result from Fang et al. (2018). For COfB, we generalize a newly developed Berry-Esseen type bound from recent work by Shao and Zhang (2022). Second, we show a translation from conditional convergence to the asymptotic independence between the error of the original estimate and the resample estimates. Finally, we leverage the cheap bootstrap method by Lam (2022a) to convert the asymptotic independence above into large-sample coverage-exact interval construction.

4.1 Conditional Convergence via a Uniform Non-Asymptotic Bound

We start with the following asymptotic result, which describes the resemblance between the errors of the original and resample runs.

Theorem 2 *Under the same assumptions as in Theorem 1, we have*

$$\sqrt{n}(\bar{x}_n - x^*) \xrightarrow[n\to\infty]{d} Z^{ASGD}, \quad \sqrt{n}(x_n - x^*) \xrightarrow[n\to\infty]{d} Z^{SGD}, \tag{7}$$

where both Z^{ASGD} and Z^{SGD} are d-dimensional Gaussian random variable with mean 0.

The covariance matrix of Z^{ASGD} *is given by* $G(x^*)^{-1}S(x^*)G(x^*)^{-1}$*, where* $G(x^*) = \nabla^2 H(x^*)$ *and* $S(x^*) = \mathbb{E}[\nabla h(x^*,\zeta)(\nabla h(x^*,\zeta))^\top]$.

For the covariance matrix of Z^{SGD}*, consider the singular value decomposition* $G(x^*) = QDQ^\top$ *with* $D = diag(d_1,\dots,d_d)$*, where* $d_1,\dots,d_d$ *are eigenvalues of* $G(x^*)$ *in decreasing order and* Q *the matrix consisting of eigenvectors.*

Let x^{*b} *denote* x^{*b}_{COfB} *or* x^{*b}_{COnB} *in Algorithms 1 and 2 respectively. We have*

$$\sqrt{n}(x^{*b} - \bar{x}_n) \xrightarrow[n\to\infty]{d} Z^{ASGD} \quad \textit{conditional on } \zeta_1,\zeta_2,\dots \textit{ in probability.} \tag{8}$$

For COfB running SGD, we have

$$\sqrt{n}(x^{*b}_{COfB} - \hat{x}_n) \xrightarrow[n\to\infty]{d} Z^{SGD} \quad \textit{conditional on } \zeta_1,\zeta_2,\dots \textit{ in probability.} \tag{9}$$

Partial results in Theorem 2 have already been established in the literature. Note that (7) is the classical asymptotic normality of (A)SGD guaranteed by our assumptions (Chung, 1954; Sacks, 1958). On the other hand, the type of conditional convergence in (8) and (9) is the main driver of classical bootstrap methods that allow the approximation of a sampling distribution using its resampled counterpart. In the case of COnB, the desired conditional convergence result (8) is well established in Appendix A.2 of Fang et al. (2018).

For COfB, nonetheless, it remains to prove (8) and (9). First, notice that the ASGD output $\bar{x}_n$ and the SAA solution $\hat{x}_n$ have an asymptotically negligible discrepancy (see Appendix A.2), so that it suffices to prove the modified version of (8) that replaces the center $\bar{x}_n$ by $\hat{x}_n$ for COfB running ASGD to establish (8). That is, it suffices to show that

$$\sqrt{n}(x^{*b}_{\mathrm{COfB}} - \hat{x}_n) \xrightarrow[n\to\infty]{d} Z^{\mathrm{ASGD}} \quad \text{conditional on } \zeta_1,\zeta_2,\dots \text{ in probability,} \tag{10}$$

for COfB running ASGD. In the rest of this subsection, we outline the sketch of proof for (9) and (10).

To elucidate our proof idea, we denote $\psi(P)$ as the minimizer for (1) with data ζ following distribution P, where ψ is viewed as a mapping from the data distribution to $\mathbb{R}^d$. Correspondingly, define ψ_n as the mapping from the data distribution to the outcome of (A)SGD. Then $\psi_n(P) \in \mathbb{R}^d$ is the (random) outcome of (A)SGD after n iterations, as a function of data distribution P with h and $\{\eta_t\}_{t=1}^n$ implicitly chosen. With the introduced notation, (7) can be restated as the weak limit of $\sqrt{n}(\psi_n(P)-\psi(P))$ being equal to Z_0. This Z_0, depending on the context, denotes the Gaussian variable Z^{SGD} or Z^{ASGD} as described in Theorem 2, whose variance depends on P. Correspondingly, let $\hat{Z}_m$ denote a normal variable that replaces P in its variance with $\hat{P}_m$, conditional on the collected data. With these new notations, (9) and (10) hold if for any Borel measurable set $D \subset \mathbb{R}^d$, we have

$$\lim_{n\to\infty} |\mathbb{P}^*(\sqrt{n}\left(\psi_n(\hat{P}_n) - \psi(\hat{P}_n)\right) \in D) - \mathbb{P}(Z_0 \in D)| = 0 \quad \text{in probability,} \tag{11}$$

where $\mathbb{P}^*$ denotes the probability conditional on the data. By the triangle inequality, one can obtain

$$\begin{aligned}
&|\mathbb{P}^*(\sqrt{n}\left(\psi_n(\hat{P}_n) - \psi(\hat{P}_n)\right) \in D) - \mathbb{P}(Z_0 \in D)| \\
\le &|\mathbb{P}^*(\sqrt{n}\left(\psi_n(\hat{P}_n) - \psi(\hat{P}_n)\right) \in D) - \mathbb{P}^*(\hat{Z}_n \in D)| + |\mathbb{P}^*(\hat{Z}_n \in D) - \mathbb{P}(Z_0 \in D)|.
\end{aligned}$$

It can be proved that the second term above vanishes in probability; see Lemma 6 in Appendix A.1 for details. On the other hand, we have the following theorem for the first term:

Theorem 3 *Under the same assumptions as in Theorem 1, for any Borel measurable set D, we have*

$$\lim_{n\to\infty} |\mathbb{P}^*\left(\sqrt{n}(\psi_n(\hat{P}_n) - \psi(\hat{P}_n)) \in D\right) - \mathbb{P}^*(\hat{Z}_n \in D)| = 0 \quad \textit{in probability.} \tag{12}$$

The proof invokes an expansive analysis on the behavior of the (A)SGD output. From the iterative scheme (2), we obtain

$$x_{n+1} = B_{0n}x_1 - \sum_{m=1}^{n} \eta_m B_{mn}\delta_m - \sum_{m=1}^{n} \eta_m B_{mn}E_m, \tag{13}$$

where $\delta_k \triangleq \delta(x_k) = \nabla H(x_k) - G(x^*)(x_k - x^*)$ is the second-order residual of the Taylor expansion of ∇H at x^*, $E_{k-1} = \nabla h(x_{k-1}, \zeta_k) - \nabla H(x_{k-1})$, and $B_{mn} = \prod_{j=m+1}^{n}(I - \eta_j G(x^*)) \in \mathbb{R}^{d\times d}$.

In the ASGD case, from (13) we show that there exist $\hat{\tau}_0, \hat{\tau}$, and $\hat{C}$ such that for any $\delta > 0$, there is integer N satisfying

$$\begin{aligned}\sup_{n>N} |\mathbb{P}^*(\sqrt{n}\left(\psi_n(\hat{P}_n) - \psi(\hat{P}_n)\right) \in D) - \mathbb{P}^*(\hat{Z}_n \in D)| \\ \leq \hat{C}(d^{3/2} + \hat{\tau}^3 + \hat{\tau}_0^3)d^{\frac{1}{2}}n^{-\alpha+\frac{1}{2}+\epsilon},\end{aligned} \tag{14}$$

with probability at least $1 - \delta$, for any measurable D and $\epsilon > 0$. To achieve this, we generalize the result in Shao and Zhang (2022), who gave an inequality similar to (14) but with a fixed distribution instead of a varying distribution, to establish uniform rates across all data distributions including the empirical distribution $\hat{P}_n$. Detailed proof for (14) can be found in Appendix A.4.

In the SGD case, the first two terms in (13) correspond to the interaction of the error of the initial solution and the second-order residual in the Taylor expansion of ∇H. One can show the following vanishing property

$$\lim_{n\to\infty} \mathbb{E}[\|B_{0n}x_1 - \sum_{m=1}^{n} \eta_m B_{mn}\delta_m\|] = 0.$$

On the other hand, the last term $\sum_{m=1}^{n} \eta_m B_{mn}E_m$ consists of the difference between sample gradient ∇h and true gradient ∇H, which converges to a normal distribution. We use the Berry-Esseen-type result from Corollary 2.3 in Shao and Zhang (2022) to give a non-asymptotic convergence result for this term and thus establish a bound similar to (14) that is uniform across all data distributions including the empirical distribution. Details can be found in Appendix A.5.

4.2 From Conditional Convergence to Asymptotic Independence

The results in Theorem 2 imply that the error of the original estimate is asymptotically independent of the errors of the resample estimates. This implication, which follows a similar idea in the cheap bootstrap (Lam, 2022a), can be stated as follows:

Theorem 4 *Supposing* (7) *and* (8) *hold, we have*

$$\sqrt{n}\begin{pmatrix} \bar{x}_n - x^* \\ x^{*1} - \bar{x}_n \\ \vdots \\ x^{*B} - \bar{x}_n \end{pmatrix} \xrightarrow[n\to\infty]{d} \begin{pmatrix} Z_0^{ASGD} \\ Z_1^{ASGD} \\ \vdots \\ Z_B^{ASGD} \end{pmatrix}. \tag{15}$$

For COfB running SGD, supposing (7) *and* (9) *hold, we have*

$$\sqrt{n}\begin{pmatrix} x_n - x^* \\ x_{COfB}^{*1} - \hat{x}_n \\ \vdots \\ x_{COfB}^{*B} - \hat{x}_n \end{pmatrix} \xrightarrow[n\to\infty]{d} \begin{pmatrix} Z_0^{SGD} \\ Z_1^{SGD} \\ \vdots \\ Z_B^{SGD} \end{pmatrix}. \tag{16}$$

$\{Z_b^{ASGD}\}_{b=0}^B$ *and* $\{Z_b^{SGD}\}_{b=0}^B$ *are i.i.d. copies of* Z^{ASGD} *and* Z^{SGD} *described in Theorem 2, respectively.*

Proof Let Δ and Δ^* denote $\sqrt{n}(\bar{x}_n - x^*)$ and $\sqrt{n}(x^{*1} - \bar{x}_n)$ respectively. Define $\Phi(z) = P\left((Z^{\text{ASGD}})_1 \le z_1, \ldots, (Z^{\text{ASGD}})_d \le z_d\right)$ where $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$. To simplify notation, $\Delta \le x$ is defined componentwise, i.e., $\Delta_i \le x_i,\ \forall i = 1, \ldots, d$. For any $x, y \in \mathbb{R}^d$,

$$\begin{aligned}
&\left|\mathbb{P}(\Delta \le x, \Delta^* \le y) - \Phi(x)\Phi(y)\right| \\
=&\left|\mathbb{E}\left[\mathbb{E}[I(\Delta \le x, \Delta^* \le y)|\{\zeta_i\}_{i=1}^n]\right] - \Phi(x)\Phi(y)\right| \\
=&\left|\mathbb{E}\left[I(\Delta \le x)\mathbb{E}[I(\Delta^* \le y)|\{\zeta_i\}_{i=1}^n]\right] - \Phi(x)\Phi(y)\right| \\
\le&\left|\mathbb{E}\left[I(\Delta \le x)\left[\mathbb{E}[I(\Delta^* \le y)|\{\zeta_i\}_{i=1}^n] - \Phi(y)\right]\right]\right| + \left|\mathbb{P}(\Delta \le x) - \Phi(x)\right|\Phi(y) \\
\le&\mathbb{E}\left[\left|\mathbb{E}[I(\Delta^* \le y)|\{\zeta_i\}_{i=1}^n] - \Phi(y)\right|\right] + \left|\mathbb{P}(\Delta \le x) - \Phi(x)\right|.
\end{aligned}$$

The last line in the above relationship vanishes since we assume (7) and (8). Thus $\mathbb{P}(\Delta \le x, \Delta^* \le y) \to \Phi(x)\Phi(y)$ for any $x, y \in \mathbb{R}^d$. Notice that conditional on $\{\zeta_i\}_{i=1}^n$, the bootstrap replications are independent. So we can generalize the above arguments for $B > 1$ by replacing $\mathbb{E}[I(\Delta^* \le y)|\{\zeta_i\}_{i=1}^n]$ by a product of B conditional expectations, each of them on a bootstrap replication of $I(\Delta^* \le y)$ for a different y. Hence we obtain (15). Lastly, the second case in the theorem follows the same argument as above. ■

The reason why we introduce the implications (15) and (16) from Theorem 2 is to utilize them in a different manner from the classical bootstrap. In particular, instead of using Theorem 2 to reveal the closeness between the sampling and resample distributions followed by Monte Carlo resample approximation, we will use the sample-resample joint distribution in (15) and (16) to construct pivotal statistics directly for inference. This latter idea, which comes from the so-called cheap bootstrap, can substantially reduce the requirement on B as we describe in the next subsection.

4.3 From Asymptotic Independence to the Cheap Bootstrap

By putting together Theorems 2 and 4, we immediately conclude that, under the assumptions in Theorem 1, the error of a resample run (compared against $\bar{x}_n$ or $\hat{x}_n$) and the error of the original (A)SGD run (compared against x^*) are asymptotically independent and follow the same Gaussian distribution. This subsequently allows us to construct asymptotic pivotal t-statistics that can be converted into coverage-exact confidence intervals. We present this argument in the proof of Theorem 1 below.

Proof of Theorem 1 Under the assumptions in Theorem 1, we have from Theorems 2 and 4 that (15) and (16) hold. Now, let x^{*b} denote x^{*b}_{COfB} and x^{*b}_{COnB} in Algorithms 1 and 2 respectively. In these cases (15) applies. Observe that

$$\frac{(\bar{x}_n)_i - x_i^*}{s_i^{\mathrm{ASGD}}} = \frac{\sqrt{n}\left((\bar{x}_n)_i - x_i^*\right)}{\sqrt{\frac{\sum_{b=1}^{B}\left(\sqrt{n}\left(x_i^{*b} - (\bar{x}_n)_i\right)\right)^2}{B}}}.$$

Taking $n \to \infty$, we have

$$\frac{\sqrt{n}\left((\bar{x}_n)_i - x_i^*\right)}{\sqrt{\frac{\sum_{b=1}^{B}\left(\sqrt{n}\left(x_i^{*b} - (\bar{x}_n)_i\right)\right)^2}{B}}} \xrightarrow{d} \frac{(Z_0^{\mathrm{ASGD}})_i}{\sqrt{\frac{\sum_{b=1}^{B}(Z_b^{\mathrm{ASGD}})_i^2}{B}}} \overset{d}{=} \frac{N}{\sqrt{\frac{\chi_B^2}{B}}} \overset{d}{=} t_B,$$

where N stands for a standard normal variable, χ_B^2 a χ^2-variable with B degree of freedom, t_B a student t-variable with B degree of freedom, and "$\overset{d}{=}$" equality in distribution. The convergence in distribution above comes from the continuous mapping theorem. The first equality in distribution comes from the *i.i.d.* normality limit in Theorem 4 and the elementary relation between χ^2 and normal. The second equality in distribution comes from the elementary construction of a t-variable. Thus, by a pivotal argument, we obtain the asymptotic exact coverage of the confidence intervals generated from Algorithms 1 and 2.

A similar argument works for COfB using SGD. In this case (16) applies, and we use the average of x^{*b}_{COfB} across b, denoted $\overline{x}^*_{\mathrm{COfB}}$, in our pivotal construction. This would result in a student t-distribution with degree of freedom $B-1$. More precisely, we have

$$\begin{aligned}\frac{(x_n)_i - x_i^*}{s_i^{\mathrm{SGD}}} &= \frac{\sqrt{n}\left((x_n)_i - x_i^*\right)}{\sqrt{n \times \frac{\sum_{b=1}^{B}\left((x^{*b}_{\mathrm{COfB}})_i - (\overline{x}^*_{\mathrm{COfB}})_i\right)^2}{B-1}}} \\ &= \frac{\sqrt{n}\left((x_n)_i - x_i^*\right)}{\sqrt{\frac{\sum_{b=1}^{B}\left(\sqrt{n}\left((x^{*b}_{\mathrm{COfB}})_i - (\hat{x}_n)_i\right) - \sqrt{n}\left((\overline{x}^*_{\mathrm{COfB}})_i - (\hat{x}_n)_i\right)\right)^2}{B-1}}}.\end{aligned}$$

As $n \to \infty$, we have

$$\frac{\sqrt{n}\left((x_n)_i - x_i^*\right)}{\sqrt{\frac{\sum_{b=1}^{B}\left(\sqrt{n}\left((x^{*b}_{\mathrm{COfB}})_i - (\hat{x}_n)_i\right) - \sqrt{n}\left((\overline{x}^*_{\mathrm{COfB}})_i - (\hat{x}_n)_i\right)\right)^2}{B-1}}} \xrightarrow{d} \frac{(Z_0^{\mathrm{SGD}})_i}{\sqrt{\frac{\sum_{b=1}^{B}\left((Z_b^{\mathrm{SGD}})_i - (\bar{Z})_i\right)^2}{B-1}}} \overset{d}{=} \frac{N}{\sqrt{\frac{\chi_{B-1}^2}{B-1}}} \overset{d}{=} t_{B-1},$$

where we note that the χ^2 and t-distributions now have degrees of freedom $B-1$, and $\bar{Z}$ denotes $(1/B)\sum_{b=1}^{B} Z_b^{\mathrm{SGD}}$. A pivotal argument gives rise to the asymptotic exactness of the confidence interval generated from COfB using SGD. ■

We note that the distinction between using t_B and t_{B-1}, in the intervals constructed in Algorithms 1 and 2, and in COfB using SGD respectively, stems from the subtle difference in the large-sample asymptotics described in Theorem 4. For Algorithms 1 and 2, the center in the sample variance is $\bar{x}_n$, which is the known outcome from the original SGD run. In contrast, for COfB using SGD, $\hat{x}_n$ is the optimizer of SAA, which is unknown. Thus, when calculating the sample variance for confidence intervals, COfB using SGD needs to set the sample mean as its center and consumes one degree of freedom. Consequently, Algorithms 1 and 2 can use B at least 1, while COfB using SGD requires B to be at least 2.

Our COnB procedure leverages the online bootstrap method of Fang et al. (2018), notably in maintaining multiple ASGD trajectories in parallel, and that each bootstrap run is generated by perturbing the stochastic gradient at each iteration t by a random multiplicative factor $W_{t,b}$. However, there is a fundamental difference in the way we construct the confidence intervals that crucially allows us to substantially reduce computation effort. Fang et al. (2018) approximate the sampling distribution of the ASGD estimator through the empirical distribution of the B bootstrap outputs, and construct confidence intervals using quantiles or variance estimates from this empirical distribution. This way of constructing intervals follows the conventional bootstrap route. Consequently, the coverage accuracy of their online bootstrap method depends not only on the asymptotic validity of the bootstrap approximation (when n is large), but also on the Monte Carlo accuracy of bootstrapping, which in turn requires a large B. In contrast, our COnB leverages the cheap bootstrap principle, using the asymptotic independence among original and bootstrap runs to construct a pivotal t-statistic, thus allowing a fixed (and small) B while still achieving asymptotic exact coverage. In this case, B only serves to reduce the width of the interval, rather than being required for asymptotic validity.

5. Application in Sparse Linear Regression

In the preceding sections, we discussed the case when the dimension of the stochastic optimization problem, d, is fixed as the sample size $n \to \infty$. In many modern applications, the ambient dimension d of the parameter space can be very large relative to the available data. In this section, we discuss a way to extend our methods to high-dimensional but sparse settings.

Throughout this section, we distinguish between the ambient dimension d and the effective dimension p. The ambient dimension $d = d(n)$ can be much larger than n, and may grow with n. The true underlying parameter is assumed to be sparse, with only a fixed number p of non-zero components. Under this setting, our methods proceed by first reducing the problem to a lower-dimensional subspace of size p via model selection, then applying our theory to this reduced space.

To be more concrete, we consider a sequence of regression problems indexed by n, where the ambient dimension $d = d(n)$ scales in the number of observations, whereas the effective

dimension stays constant at p:

$$b^{(n)} = A^{(n)}x^{(n)} + \epsilon^{(n)}.$$

Here, superscript (n) indicates dependence on the sample size n. $\epsilon^{(n)} \in \mathbb{R}^n$ consists of n *i.i.d.* entries, $b^{(n)} \in \mathbb{R}^n$ represents the vector of n responses, $A^{(n)} \in \mathbb{R}^{n\times d}$ encodes the matrix of n feature vectors each of length d, and $x^{(n)} = [x_1^{(n)}, \ldots, x_d^{(n)}]^\top \in \mathbb{R}^d$ has p nonzero entries. Given an index set T and a vector v, $(v)_T$ denotes the $|T|$-dimensional subvector of v consisting of entries of v with indices in T. Let $x^* := (x^{(n)})_{T^*} = [x_1^*, \ldots, x_p^*]^\top$ be the non-zero entries of the true model coefficients, and we assume x^* is not dependent on n. For matrix A and index sets T_1 and T_2, A_{T_1,T_2} denotes the $|T_1| \times |T_2|$ submatrix of A with row indices inside T_1 and column indices inside T_2. Then, $A_1^{(n)} := A_{\mathbb{N},T^*} \in \mathbb{R}^{n\times p}$ is the submatrix of $A^{(n)}$ consisting of columns with indices in T^* and $A_2^{(n)} := A_{\mathbb{N},T^{*C}} \in \mathbb{R}^{n\times(d-p)}$ consists of columns with indices outside T^*.

Our methods described in the previous sections under a fixed-dimensional asymptotic regime are not directly applicable in this setting, as the dimension d of this problem is no longer fixed as $n \to \infty$. In particular, the non-asymptotic bound in (14) deteriorates with increasing d. Inspecting (14), the leading dimension dependent term in the bound scales as $d^2 n^{-\alpha+\frac{1}{2}+\epsilon}$. Consequently, the right hand side of (14) vanishes only if $d(n)$ grows sufficiently slowly relative to n, i.e., $\lim_{n\to\infty} d(n)^2 n^{-\alpha+\frac{1}{2}+\epsilon} = 0$. If this latter condition fails, the bound in (14) no longer guarantees small resampling error, and in this regime we should not expect theoretical control from our previous arguments. This in particular rules out validity when d grows faster than $n^{\alpha/2-1/4}$.

A feasible approach to handle this challenge is to first transform the problem into a lower dimension by Lasso model selection techniques (Zhao and Yu, 2006; Belloni and Chernozhukov, 2013), which produces a sparse solution by solving the following problem:

$$\hat{x}^{(n)}(\lambda) = \arg\min_x \|b^{(n)} - A^{(n)}x\|^2 + \lambda \sum_{i=1}^{d} |x_i|, \tag{17}$$

where $\lambda \geq 0$ is the regularization parameter that controls the sparsity. Let $\mathcal{T}$ be the support function of a d-dimensional vectors, i.e., for a vector $x \in \mathbb{R}^d$, $\mathcal{T}(x) := \{i : x_i \neq 0\}$. The estimated support corresponding to parameter λ is $\mathcal{T}(\hat{x}^{(n)}(\lambda))$, and the true support $T^* = \mathcal{T}(x^{(n)})$. Under assumptions to be specified, the difference between estimated support and true support vanishes (Zhao and Yu, 2006). After the model selection step, our methods can be adapted and applied to the problem confined to the support of $\hat{x}^{(n)}$. Specifically, apply COnB or COfB to the problem given by (1) and (2). In this context, the i-th data is the vector corresponding to $b_i^{(n)}$ and entries in $\mathcal{T}(x^{(n)})$ for the i-th row of $A^{(n)}$. A pseudocode can be found in Algorithm 3.

We put forth the following assumptions to support the efficacy of Algorithm 3.

Assumption 7 *Entries of $\epsilon^{(n)}$ are i.i.d. and bounded. Diagonal entries of the sample covariance matrix $\frac{1}{n}A^{(n)^\top}A^{(n)}$ are upper bounded by constant M_1 not depending on n. The submatrix $\frac{1}{n}A_1^{(n)^\top}A_1^{(n)}$ is positive definite. And as $n \to \infty$, the regularization parameter satisfies $\lambda_n/n \to 0$ and $\lambda_n/\sqrt{n} \to \infty$.*

Algorithm 3 Two-stage Method for Sparse Linear Regression

Input: *i.i.d.* data $\{\zeta_t\}_{t=1}^n$, number of bootstrap runs $B \geq 1$, step size sequence $\{\eta_t\}$, initial guess x_0, nominal coverage level $1-\gamma$, Lasso regularization parameter λ_n.
Output: confidence intervals $\mathcal{I}_{i,n},\ i = 1, \ldots, d$.
Solve (17) to obtain $\hat{x}^{(n)}(\lambda_n)$.
$T_n \leftarrow \{i : (\hat{x}^{(n)}(\lambda_n))_i \neq 0\}$
$\mathcal{I}_{i,n} \leftarrow \{0\}$ for all $i \notin T_n$
Obtain confidence intervals on T_n by running Algorithms 1 and 2 with input: *i.i.d.* data $\{(\zeta_t)_{T_n}\}_{t=1}^n$, number of bootstrap runs B, step size sequence $\{\eta_t\}$, initial guess x_0, nominal coverage level $1-\gamma$.

These assumptions are mild in the context of linear regression. The boundedness of ϵ and the diagonal entries of the covariance matrix are typical results of the common data normalization procedure. Positive definiteness of $\frac{1}{n}{A_1^{(n)}}^\top A_1^{(n)}$ ensures the identifiability of the model. In addition to Assumption 7, we need one further assumption to guarantee the correctness of the selected model by Lasso.

Assumption 8 *There exists a positive constant η such that for all $i = 1, ..., d-p$, we have*

$$|(A_2^{(n)\top} A_1^{(n)} ({A_1^{(n)}}^\top A_1^{(n)})^{-1} sign(x^*))_i| \leq 1 - \eta,$$

where $sign(x^*) \in \mathbb{R}^p$ denotes the entrywise sign function for x^*.

Assumption 8 is referred to as the irrepresentable condition in Zhao and Yu (2006), which turns out to relate closely to the consistency of Lasso. Under Assumptions 7 and 8, the probability of Lasso selecting the correct model converges to 1 as data size $n \to \infty$ (Zhao and Yu, 2006).

Assumption 9 *Rows of $A_1^{(n)}$ are bounded and i.i.d. following distribution P_a such that $\mathbb{E}_{a\sim P_a}[aa^\top] \in \mathbb{R}^{p\times p}$ is positive definite with eigenvalues greater than $l > 0$. Rows of $A_2^{(n)}$ are bounded and i.i.d..*

The above assumptions are sufficient to guarantee Assumptions 1 to 4 for the linear regression problem confined to T^* which, together with one of the following two assumptions, guarantees the correctness of coverage probability of confidence intervals generated after the model selection step.

Assumption 10 *For COfB running ASGD or COnB, the step size $\eta_t = \eta t^{-\alpha}$ for some $\alpha \in (\frac{1}{2}, 1]$.*

Assumption 11 *For COfB running SGD, the step size $\eta_t = \frac{\eta}{t}$ such that $\eta l > \frac{1}{2}$.*

With these assumptions in place, we are ready to state the theoretical guarantee for Algorithm 3.

Theorem 5 *In the sparse linear regression setting, suppose Assumptions 7 to 9, and one of Assumptions 10 and 11 hold. For $i = 1, \ldots, d$, the $1-\gamma$ confidence interval produced by Algorithm 3 satisfies*

$$\lim_{n\to\infty} \mathbb{P}(x_i^{(n)} \in \mathcal{I}_{i,n}) = \begin{cases} 1 & \text{if } \ i \notin T^* \\ 1-\gamma & \text{if } \ i \in T^* \end{cases}.$$

Theorem 5 states that, as $n \to \infty$, Algorithm 3 correctly identifies the support of the true model parameter T^* and provides confidence intervals with exact coverage for non-zero entries of $x^{(n)}$. It is worth noticing that, although COnB could work as the second stage of Algorithm 3, the whole procedure is no longer suitable for the online setting as the model selection part requires all the n data. If new data comes, one needs to solve the Lasso again. The proof for Theorem 5 can be found in Appendix A.6.

In Chen et al. (2020), the authors also discussed inference for high-dimensional linear regression setting. In their work, they developed a plug-in method that estimates the true coefficient and the inverse of the covariance matrix at the same time, using the Regularization Annaled epoch Dual AveRaging (RADAR) algorithm (Agarwal et al., 2012), a variant of SGD. Confidence intervals with a nominal level $1-\gamma$ for each entry of the true coefficient are the outputs of their algorithm, and they have a theoretical guarantee similar to that of Theorem 1. In comparison, our approach for the sparse linear regression provides more information since it also gives asymptotically correct support of the relevant coefficients, as described in Theorem 5. On the other hand, the method in Chen et al. (2020) could handle new data with small additional computations, while ours, as discussed earlier, only works under the offline setting.

6. Experiments

In this section, we illustrate the numerical performance of our approaches and compare them with the other methods in the regression setting. The code to reproduce all experiments in this paper is publicly available at https://github.com/BeetrootWang/Cheap-Bootstrap-for-Fast-Uncertainty-Quantification-of-Stochastic-Gradient-Descent.

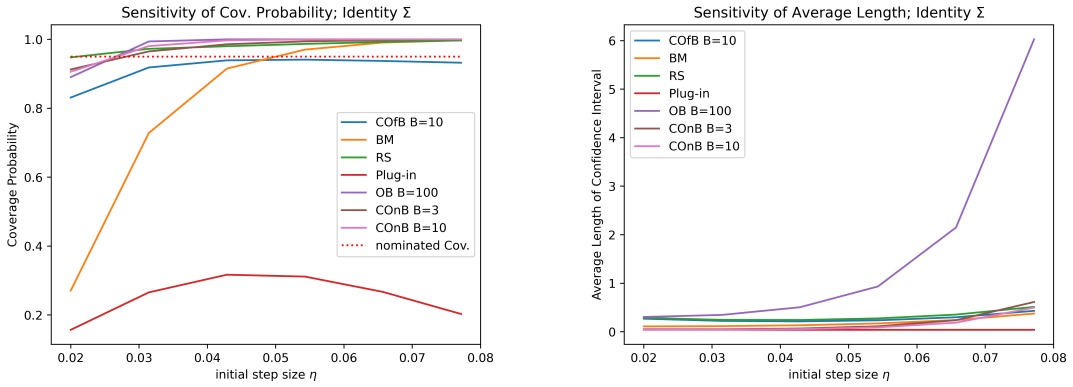
6.1 Regression Problems with Fixed Dimensionality

We consider two sets of problems:

Linear Regression We consider the linear regression problem of dimension d. The data $\zeta = (a, b)$ with distribution P consists of the independent variable $a \in \mathbb{R}^d$ and dependent variable $b \in \mathbb{R}$. In this case, $h(x, \zeta) = \frac{1}{2}(a^\top x - b)^2$. a follows a multivariate normal distribution $\mathcal{N}(0, \Sigma)$. Let x^* be the true regression coefficient. Then, b satisfies the model $b = a^\top x^* + \epsilon$ for some error term ϵ that is assumed to have normal distribution $\mathcal{N}(0, \sigma^2)$. In this experiment, ϵ and a are independent. And we have $\psi(P) = x^*$, $S = \sigma^2\Sigma$, and $G = \Sigma$.

Logistic Regression Similar to the linear regression setup, the data $\zeta = (a, b)$ coming from distribution P consists of the independent variable $a \in \mathbb{R}^d$ and dependent variable $b \in \{-1, 1\}$. $h(x, \zeta) = \log\left(1 + e^{-b \times a^\top x}\right)$, where a follows a multivariate normal distribution

$\mathcal{N}(0,\Sigma)$, and $b=1$ with probability $\frac{1}{1+e^{-a^\top x^*}}$, where x^* denotes the true regression coefficient. In this case, we have $\nabla h(x,\zeta)=\frac{-b\times a}{1+e^{b\times a^\top x}}$ and $\nabla^2 h(x,\zeta)=\frac{aa^\top}{(1+e^{a^\top x})(1+e^{-a^\top x})}$. The Hessian information $\nabla^2 h(x,\zeta)$ above will only be used in the delta method.



(a) (b)

Figure 1: Performance of methods concerning different choices of initial step size η, with sample size $n=10^4$. We compare the delta method, batch mean method with $M=n^{0.25}$, COfB method with $B=10$, COnB with $B=3,10$, and the online bootstrap method with $B=100$, when $n=10^4$. The left figure shows the sensitivity of the average coverage probability against η. The right figure shows the sensitivity of the average length of confidence interval against η. We report the results for identity Σ.

6.1.1 Baselines

In both experiments, we compare with the batch mean method (BM) and delta method (delta) in Chen et al. (2020), the random scaling method (RS) in Lee et al. (2022), the online bootstrap (OB) in Fang et al. (2018), and the HiGrad method (HiGrad) in Su and Zhu (2023).

The batch mean method splits $\{x_i\}_{i=0}^n$ into M batches, where M is an extra hyperparameter, with e_i and s_i denoting the ending index and starting index of the i-th batch respectively. n_k denotes the number of iterates in k-th batch, and the estimator is defined to be $\frac{1}{M}\sum_{k=1}^M n_k(\bar{x}_{n_k}-\bar{x}_M)(\bar{x}_{n_k}-\bar{x}_M)^\top$, where $\bar{x}_{n_k}=\frac{1}{n_k}\sum_{i=s_k}^{e_k}x_i$ and $\bar{x}_M=\frac{1}{e_M-e_0}\sum_{i=s_1}^{e_M}x_i$. Let $N=\frac{n^{1-\alpha}}{M+1}$, and e_k to be the closest integer to $((k+1)N)^{\frac{1}{1-\alpha}}$ for each $k=0,\ldots,M$ as suggested in Chen et al. (2020). The confidence interval for each entry of x^* is constructed using diagonal entries of the batch mean estimator and a normal quantile.

The delta method (Chen et al., 2020) generates confidence intervals using normal quantiles and $\tilde{\Sigma}_n^2=\tilde{G}_n^{-1}\tilde{S}_n\tilde{G}_n^{-1}$, where $\tilde{G}_n=\frac{1}{n}\sum_{i=1}^n\nabla^2 h(x_{i-1},\zeta_i)$, and $\tilde{S}_n=\frac{1}{n}\sum_{i=1}^n\nabla h(x_{i-1},\zeta_i)(\nabla h(x_{i-1},\zeta_i))^\top$ are computed on the fly.

The random scaling method (Lee et al., 2022) updates two quantities $A_t\in\mathbb{R}^{d\times d}$ and $b_t\in\mathbb{R}^d$ upon arrival of the t-th data with the update rule $A_t=A_{t-1}+t^2\bar{x}_t\bar{x}_t^\top$ and $b_t=b_{t-1}+t^2\bar{x}_t$, respectively. Then, construct the random scaling matrix $\hat{V}_n=$

Dimension (d)	delta	BM	RS	OB ($B = 100$)	HiGrad$_{(2,2)}$	COfB ASGD ($B = 3$)	COnB ($B = 3$)
10	0.32	0.34	0.80	45.90	0.32	0.88	1.75
100	0.59	0.56	2.92	46.74	0.49	1.16	1.94
1000	8.92	3.49	171.41	76.57	2.78	4.76	5.15
2000	44.79	11.35	799.56	116.08	9.72	13.81	13.53

Table 2: Average runtimes (seconds) for different methods for linear regression with $n = 10^5$, identity Σ, and dimension $d \in \{10, 100, 1000, 2000\}$.

$n^{-2}\left(A_n - \bar{x}_n b_n^\top - b_n \bar{x}_n^\top + \bar{x}_n \bar{x}_n^\top \sum_{s=1}^n s^2\right)$. The center of confidence interval for i-th entry of x^* is $(\bar{x}_n)_i$ and the half-width is $cv_{(1-\alpha/2)}\sqrt{\frac{(\hat{V}_n)_{i,i}}{n}}$, where $cv_{(1-\alpha/2)}$ denotes the critical value of the studentized statistic. We use $cv_{0.975} = 6.747$ in our experiment, as suggested by Lee et al. (2022).

The online bootstrap method (Fang et al., 2018) runs $B + 1$ ASGD threads in parallel. For each data ζ_t, the update step for b-th thread is $x_t^{(b)} = x_{t-1}^{(b)} - \eta_t W_t^{(b)} \nabla h(x_{t-1}^{(b)}, \zeta_t),\ b = 0, 1, \ldots, B$, and $W_t^{(0)} = 1,\ \forall t$. Then obtain $\{x^{(b)}\}_{b=0}^B$ by taking average along each thread of SGD. The sample variance for each coordinate $\sigma_i,\ i = 1, \ldots, d$ is then calculated for $\{x^{(b)}\}_{b=1}^B$ with a known mean $x^{(0)}$. Then, using normal quantile and σ_i, one can construct confidence intervals for each entry of x^* centered at $x^{(0)}$.

The HiGrad method (Su and Zhu, 2023) takes two tuples $(B_1, B_2, \ldots, B_K)$ and $(n_0, n_1, \ldots, n_K)$ as hyperparameters describing when to break the SGD thread into how many branches. B_i describes the number of branches a single branch divides into at i-th breaking. n_i describes the number of data each thread uses between i and $i+1$-th breaking. After all the breaking, there will be $T = \prod_{i=1}^K B_i$ threads and one obtains $\{x^{(j)}\}_{j=1}^T$ by averaging each thread. The confidence interval for i-th entry of x^* is calculated by aggregating $\{x_i^{(j)}\}_{j=1}^T$. It is worth pointing out that the total number of data used in HiGrad should be no more than n. i.e. $n_0 + \sum_{i=1}^K B_i n_i \le n$. Thus, the length of a single thread in HiGrad is typically shorter than that of other methods discussed in this paper.

6.1.2 Hyperparameters

The choices of hyperparameters in both experiments are listed here. The nominal coverage probability we consider is 95%. Dimension of the problem $d \in \{5, 20, 200\}$ and we report the result for three choices of covariance matrix Σ of a. Namely, identity $\Sigma = I_d$, Toeplitz $\Sigma_{i,j} = 0.5^{|i-j|}$ and equicorrelation case $\Sigma_{i,j} = 0.2$ if $i \neq j$ and $\Sigma_{i,i} = 1$. The decay rate for learning rate $\alpha = 0.501$. The optimal solution $x^* = [0, \frac{1}{d-1}, \frac{2}{d-1}, \ldots, 1]^\top$ and we set the initial choice $x_0 = [0, 0, \ldots, 0]^\top$.

For each set of hyperparameters, we run 500 independent trials and report the mean and standard deviation of the coverage probabilities and the average length of the intervals across d dimensions. We tune the initial step size η within the range $[0.2, 0.7]$ and report the result with the most accurate average coverage probability. For the batch mean method, M is selected to be the nearest integer to $n^{0.25}$ as suggested in Chen et al. (2020). For HiGrad, the architecture we experimented with is $((2, 2), (n/7, n/7, n/7))$. As mentioned in Su and Zhu (2023), this choice is desirable as it balances accuracy, coverage, and informativeness. We report the performance of COfB and COnB with $B = 3, 5, 10$. For the online bootstrap

method, $B = 200$ is the suggested choice in Fang et al. (2018). We also consider $B = 10$ and 100 based on the observation that online bootstrap with $B = 100$ and 200 perform similarly which prompts the question of whether smaller B's would work.

6.1.3 Results

Result for Toeplitz Σ are presented in Table 3. For full numeric, please refer to Appendix B. In the table, **bold** numbers highlight favorable results, where the coverage probability lies between 92% and 98%. Conversely, *italic* numbers indicate poor results, specifically those with coverage probabilities below 80%.

Coverage Probability Our method gives accurate coverage probability across all values of d and Σ, typically between $94\% - 96\%$, regardless of the choice of B. The delta method and the batch mean method consistently fall behind other techniques. The random scaling method achieves accurate coverage probability in the linear regression experiment, comparable to our methods. However, it fails in some of the logistic regression experiments when $d = 200$. The HiGrad method displays a stark decline in performance at $d = 200$, which is potentially attributed to its abbreviated SGD trajectory. Specifically, when $d = 200$, the available data or iterations do not suffice for proper SGD convergence. Similarly, the performance of the delta method drops significantly as d increases. This is largely due to its dependence on a Monte Carlo estimation of a full Hessian matrix, the precision of which diminishes as dimensionality grows. In contrast, the merits of data reuse in bootstrap-type approaches become increasingly evident in higher dimensions. To this end, note that while the online bootstrap method yields comparable coverage probabilities, our methodologies are significantly faster.

Interval Width Our approach results in a longer average confidence interval relative to the delta method, batch mean method, random scaling method, and HiGrad. This outcome stems from the t-quantile with a degree of freedom defined by either $B - 1$ for our COfB running SGD or B for our other methods. However, given that entries of x^* span $[0, 1]$ and the confidence intervals are of magnitude 10^{-2}, the inflation in interval widths produced by our methods seems secondary compared to our computational gain. In addition, our average width shrinks rapidly as B increases, which follows the behavior of t-interval. With 95% confidence level, the t-interval is 49.6% wider than the normal interval when $B = 3$, and only 10.9% wider when $B = 10$. On the other hand, the execution time grows approximately linearly in B. Furthermore, we note that even though other methods result in narrower intervals, they can fall short significantly in attaining enough coverage probabilities, a performance measure often considered more important than the interval widths. Indeed, we can see from Table 3 that although the delta method, batch mean method, random scaling method, and HiGrad constantly give narrower intervals, their coverage probabilities can lie in the range of 80%, 70% to as low as 30% under certain configurations. Nonetheless, despite the above desirable conclusions on our approach, we caution that in some problems the inflation in interval width could be a significant price relative to the magnitude of the solution, in which case one needs to consider a larger B in our approach to rightly balance interval width with computation effort.

Execution Time Running times for the linear regression experiments across varied methodologies are available in Table 2. These experiments were conducted on a single thread of an Apple M2 Pro processor implemented with Python. We report the average execution time of 10 independent repetitions for each setting. Both HiGrad and the batch mean method emerged as the swiftest, attributed to their avoidance of additional gradient steps or Hessian calculations. Our COfB and COnB are slightly slower compared with HiGrad and the batch mean method. However, our methods are only marginally slower, especially when dimension d gets larger. In contrast, the online bootstrap method is much slower, with around 100 times longer runtimes than the batch mean method under all settings. The execution time for the delta method and the random scaling method escalates significantly as dimensions increase. As discussed in the introduction, this is due to the manipulation of $d \times d$ matrices. Specifically, the random scaling method updates matrix A_i at each iteration, while the delta method updates $\tilde{G}_i$ and $\tilde{S}_i$ during gradient steps and requires the $d \times d$ matrix inversion when constructing the confidence interval. Roughly speaking, the random scaling method has time complexity quadratic in d, and the delta method is cubic in d. The apparent faster performance of the delta method, compared to the random scaling method, might be attributed to the way they are implemented.

Sensitivity Analysis In Figure 1, we compare the performance of our methods, the delta method, the batch mean method, and the online bootstrap method with the number of samples $n = 10^4$ for the linear regression problem. Observe that the coverage probability of COfB methods remains stable at around 95% regardless of changes in the initial step size. The random scaling method also performs well in this experiment, although it has a slight tendency of over-coverage. On the other hand, the batch mean method requires a careful choice of the initial step size to give a comparable coverage rate, and the optimal choice is not the same across different problems. The delta method suffers from a huge under-coverage and fails to give a valid confidence interval. The delta method and the batch mean method have smaller average lengths, which can be associated with their under-coverages. It can be observed that the coverage probability and the average length of the batch mean estimator both increase as η increases. Our COnB method has a similar sensitivity as the online bootstrap method. Nonetheless, as mentioned earlier, COnB is substantially faster. Additionally, the average length of our COnB becomes almost the same as that of the online bootstrap method when increasing B to 10.

6.2 Sparse Linear Regression with Increasing Dimensionality

6.2.1 Problem Setting and Hyperparameters

We evaluate the performance of Algorithm 3 under the setting discussed in Section 5. Specifically, we consider $n = 100$ and $d = 500$. The columns of the $n \times d$ design matrix are *i.i.d.*, generated from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$, with choices of Σ the same as in the fixed dimensionality experiment. We set the sparsity parameter p to be 3 or 15, with the first p entries of the true coefficients non-zero and each uniformly drawn over the interval $[0, 2]$. Consequently, the true model coefficient $x^{(100)} = [x_1^*, \ldots, x_p^*, 0, \ldots, 0] \in \mathbb{R}^{500}$. The nominal coverage probability for the entries in T^* is still 95%.

For the model selection stage, we set the penalty parameter λ to $0.001 \times \frac{\log d}{n}$. We tested both COfB and COnB as the second stage method. Choices of the number of resampling runs are $B \in \{3, 10\}$. In all the experiments, the decay rate of learning rate is fixed at $\alpha = 0.501$, and the initial guess for SGD is $x_0 = [0, \ldots, 0]^\top \in \mathbb{R}^{500}$. The initial learning rate η, ranging within $[0.02, 0.5]$, is tuned to ensure proper convergence of SGD.

6.2.2 Results

We conducted 500 independent trials for each hyperparameter configuration, reporting the mean and standard deviation of the coverage probability and average interval length for coefficients within and outside T^*, respectively. The results for Toeplitz Σ can be found in Table 4. Please refer to Appendix B for the full table. In the table, **bold** numbers represent good results, with coverage probability lying between 92% and 98% on T^* or over 99% outside T^*. As indicated by Theorem 5, the ideal coverage probability for coefficients in T^* would be close to 95%. For coefficients not in T^* (rows with $\notin T^*$ in the table), a coverage close to 1 is desirable.

We conclude from the table that our methods, regardless of the configuration, achieve accurate coverage probability both within and outside T^*. The average length of confidence intervals is generally reasonable compared with the scale of the problem. Notably, confidence intervals for coefficients outside T^* are significantly shorter than those inside T^*. This is due to the accurate model selection in the first stage, as our method will simply output singleton $\{0\}$ for entries not selected in the first stage. We also observe that the average length shrinks when B increases while maintaining high accuracy in coverage probabilities across all B values.

Finally, we provide further sensitivity analyses of our approach with respect to ill-conditionedness and learning rate scheme in Appendix C.

Acknowledgements

We gratefully acknowledge support from the InnoHK initiative, the Government of the HKSAR, Laboratory for AI-Powered Financial Technologies, and the Columbia Innovation Hub Award. We thank Sokbae (Simon) Lee for the helpful suggestions that have greatly improved our manuscript.

Appendix A. Useful Lemmas and Missing Proofs

A.1 Convergence of Normal Variables with Empirical Variances

Lemma 6 *For a fixed n and given $\hat{P}_n$, let $\hat{Z}_n$ be the weak limit of $\sqrt{m}(\psi_m(\hat{P}_n) - \psi(\hat{P}_n))$ as $m \to \infty$. Let Z_0 be the weak limit of $\sqrt{m}(\psi_m(P) - \psi(P))$. Under the same assumptions as Theorem 1, for any Borel set D,*

$$|\mathbb{P}^*(\hat{Z}_n \in D) - \mathbb{P}(Z_0 \in D)| \to 0 \quad \textit{in probability},$$

as $n \to \infty$.

Proof Consider the ASGD case now, and the SGD case can be treated similarly. By the classical convergence result for ASGD (see Polyak and Juditsky (1992)), we have the

	$d = 5$		$d = 20$		$d = 200$	
	Cov (%)	Len ($\times 10^{-2}$)	Cov (%)	Len ($\times 10^{-2}$)	Cov (%)	Len ($\times 10^{-2}$)
Linear Regression						
delta	**94.20** (0.07)	1.53 (0.00)	**93.23** (0.08)	1.58 (0.00)	*37.16* (0.15)	1.60 (0.00)
BM	91.80 (0.09)	1.53 (0.00)	88.51 (0.10)	1.51 (0.00)	**97.37** (0.05)	6.54 (0.01)
RS	**92.83** (0.08)	1.94 (0.01)	**93.38** (0.08)	2.15 (0.01)	**97.02** (0.05)	8.60 (0.06)
OB $B = 10$	**92.48** (0.08)	1.66 (0.00)	**93.68** (0.08)	1.81 (0.00)	**93.33** (0.25)	13.73 (0.01)
OB $B = 100$	**95.48** (0.07)	1.72 (0.00)	**95.84** (0.06)	1.97 (0.00)	**96.58** (0.18)	14.26 (0.01)
$\text{HiGrad}_{(2,2)}$	**94.33** (0.07)	2.69 (0.01)	**94.88** (0.07)	2.82 (0.01)	**94.09** (0.07)	2.83 (0.01)
COfB ASGD $B = 3$	**94.72** (0.07)	3.06 (0.00)	**94.95** (0.07)	3.30 (0.00)	**94.41** (0.07)	14.44 (0.02)
COfB ASGD $B = 5$	**95.24** (0.07)	2.21 (0.00)	**94.89** (0.07)	2.26 (0.00)	**93.91** (0.08)	9.96 (0.01)
COfB ASGD $B = 10$	**95.28** (0.07)	1.74 (0.00)	**94.95** (0.07)	2.12 (0.00)	**94.01** (0.08)	8.44 (0.01)
COfB SGD $B = 3$	**95.16** (0.07)	6.93 (0.01)	**94.76** (0.07)	4.34 (0.00)	**95.16** (0.07)	6.90 (0.01)
COfB SGD $B = 5$	**95.84** (0.06)	4.70 (0.00)	**94.20** (0.07)	2.97 (0.00)	**95.04** (0.07)	4.72 (0.00)
COfB SGD $B = 10$	**95.36** (0.07)	3.95 (0.00)	**94.20** (0.07)	2.50 (0.00)	**95.00** (0.07)	3.98 (0.00)
COnB $B = 3$	**95.00** (0.07)	2.49 (0.00)	**95.36** (0.07)	2.94 (0.00)	**95.41** (0.07)	21.04 (0.02)
COnB $B = 5$	**94.60** (0.07)	1.96 (0.00)	**95.43** (0.07)	2.44 (0.00)	**95.48** (0.07)	17.42 (0.02)
COnB $B = 10$	**94.92** (0.07)	1.77 (0.00)	**95.42** (0.07)	2.18 (0.00)	**95.78** (0.06)	15.61 (0.01)
Logistic Regression						
delta	**94.83** (0.07)	4.05 (0.00)	**93.29** (0.08)	5.59 (0.00)	*53.69* (0.16)	9.56 (0.00)
BM	84.00 (0.12)	3.16 (0.01)	*75.25* (0.14)	3.75 (0.01)	*34.93* (0.15)	7.30 (0.03)
RS	**92.67** (0.08)	5.14 (0.02)	90.88 (0.09)	7.26 (0.03)	76.38 (0.13)	17.45 (0.10)
OB ($B = 100$)	**95.00** (0.07)	4.22 (0.00)	**94.04** (0.07)	6.65 (0.01)	99.78 (0.01)	69.55 (0.26)
OB ($B = 200$)	**95.00** (0.07)	4.24 (0.00)	**94.67** (0.07)	6.70 (0.01)	99.78 (0.01)	69.28 (0.26)
$\text{HiGrad}_{(2,2)}$	**95.33** (0.07)	7.18 (0.03)	**93.38** (0.08)	8.92 (0.03)	*57.02* (0.16)	10.27 (0.04)
COfB ASGD $B = 3$	**94.12** (0.07)	5.70 (0.00)	**94.77** (0.07)	11.49 (0.01)	**93.79** (0.08)	42.27 (0.05)
COfB ASGD $B = 5$	**94.32** (0.07)	4.82 (0.00)	**94.81** (0.07)	7.97 (0.01)	**93.23** (0.08)	28.81 (0.02)
COfB ASGD $B = 10$	**94.88** (0.07)	4.61 (0.00)	**94.65** (0.07)	6.69 (0.00)	**93.09** (0.08)	24.43 (0.01)
COfB SGD $B = 3$	**95.36** (0.07)	9.48 (0.01)	**94.80** (0.07)	9.65 (0.01)	**94.41** (0.07)	30.53 (0.03)
COfB SGD $B = 5$	**95.40** (0.07)	7.98 (0.00)	**94.37** (0.07)	8.16 (0.00)	**93.79** (0.08)	20.76 (0.02)
COfB SGD $B = 10$	**95.40** (0.07)	7.98 (0.00)	**94.37** (0.07)	8.16 (0.00)	**93.62** (0.08)	17.56 (0.01)
COnB ($B = 3$)	**94.00** (0.08)	6.22 (0.02)	**94.71** (0.07)	10.27 (0.05)	**97.82** (0.05)	99.61 (0.60)
COnB ($B = 5$)	**95.00** (0.07)	5.25 (0.02)	**94.71** (0.07)	8.20 (0.03)	98.75 (0.04)	85.15 (0.45)
COnB ($B = 10$)	**94.83** (0.07)	4.84 (0.01)	**94.29** (0.07)	7.25 (0.02)	99.31 (0.03)	78.02 (0.37)

Table 3: Results for the linear and logistic regression with Toeplitz Σ, $n = 10^5$.

		$p = 3$		$p = 15$	
		Cov (%)	Len ($\times 10^{-2}$)	Cov (%)	Len($\times 10^{-2}$)
COfB ASGD $B = 3$	$\in T^*$	**95.20** (21.38)	2.75 (0.01)	**95.12** (21.54)	34.87 (0.11)
	$\notin T^*$	**99.88** (3.42)	0.05 (0.00)	**99.09** (9.52)	8.62 (0.07)
COfB ASGD $B = 10$	$\in T^*$	**95.07** (21.66)	0.86 (0.00)	**93.13** (25.29)	4.33 (0.01)
	$\notin T^*$	**99.86** (3.74)	0.02 (0.00)	**99.12** (9.32)	1.00 (0.01)
COnB ASGD $B = 3$	$\in T^*$	**94.73** (22.34)	1.28 (0.01)	**97.19** (16.54)	10.68 (0.03)
	$\notin T^*$	**99.93** (2.73)	0.03 (0.00)	**99.80** (4.47)	2.47 (0.02)
COnB ASGD $B = 10$	$\in T^*$	**95.40** (20.95)	1.07 (0.00)	98.85 (10.65)	10.27 (0.04)
	$\notin T^*$	**99.96** (2.00)	0.02 (0.00)	**99.98** (1.51)	2.41 (0.02)

Table 4: Results for sparse linear regression with Toeplitz Σ, $n = 100$.

following:

$$\hat{Z}_n \sim N(0, \sigma_n^2), \quad \sigma_n^2 = G_n(\hat{x}_n)^{-1} S_n(\hat{x}_n) G_n(\hat{x}_n)^{-1},$$
$$Z_0 \sim N(0, \sigma^2), \quad \sigma^2 = G(x^*)^{-1} S(x^*) G(x^*)^{-1}.$$

As a result, $|\mathbb{P}^*(\hat{Z}_n \in D) - \mathbb{P}(Z_0 \in D)|$ is a continuous function of entries of $G_n(\hat{x}_n)$ and $S_n(\hat{x}_n)$. Therefore, it suffices to prove that $(G_n(\hat{x}_n))_{i,j} \to (G(x^*))_{i,j}$ and $(S_n(\hat{x}_n))_{i,j} \to (S(x^*))_{i,j}$ in probability, for each $i, j \in \{1, \ldots, d\}$.

To simplify the notation, let us define $Qf(x) \triangleq \int f(x,\zeta) Q(d\zeta)$ for a function f of ζ parameterized by x and a probability measure Q. Then, $(G_n(\hat{x}_n))_{i,j} = \hat{P}_n \partial^2_{i,j} h(\hat{x}_n)$ and $(G(x^*))_{i,j} = P\partial^2_{i,j} h(x^*)$. Consider the following decomposition:

$$\begin{aligned} &(G_n(\hat{x}_n))_{i,j} - (G(x^*))_{i,j} \\ &= \left(\hat{P}_n \partial^2_{i,j} h(\hat{x}_n) - P\partial^2_{i,j} h(\hat{x}_n)\right) + \left(P\partial^2_{i,j} h(\hat{x}_n) - P\partial^2_{i,j} h(x^*)\right). \end{aligned} \tag{18}$$

By Assumption 1, $P\partial^2_{i,j} h(x)$ is continuous in x. Combined with the fact that $\hat{x}_n$ converges to x^* in probability, the second term in (18) converges to 0 in probability.

By Assumption 5, $\mathcal{F}_{i,j}$ is P-Glivenko-Cantelli, and the first term in (18) vanishes

$$\left|\hat{P}_n \partial^2_{i,j} h(\hat{x}_n) - P\partial^2_{i,j} h(\hat{x}_n)\right| \le \sup_{f \in \mathcal{F}_{i,j}} |\hat{P}_n f - Pf| \xrightarrow{as*} 0. \tag{19}$$

So we conclude that $(G_n(\hat{x}_n))_{i,j} - (G(x^*))_{i,j}$ converges to 0 in probability.

By definition, $(S_n(\hat{x}_n))_{i,j} = \hat{P}_n \partial_i h(\hat{x}_n) \partial_j h(\hat{x}_n)$ and $(S(x^*))_{i,j} = P\partial_i h(x^*) \partial_j h(x^*)$. Define $\tilde{\mathcal{F}}_{i,j} = \{\partial_i h(x,\zeta)\partial_j h(x,\zeta) : x \in \mathcal{X}\}$. By Assumption 1, there is a constant m depending on L such that $|\partial_i h(x_1,\zeta)\partial_j h(x_1,\zeta) - \partial_i h(x_2,\zeta)\partial_j h(x_2,\zeta)| < m\|x_1 - x_2\|$, for all x_1, x_2, and ζ. Therefore, $\tilde{\mathcal{F}}_{i,j}$ is P-Glivenko-Cantelli. The remaining proof follows the same arguments as in the proof of $(G_n(\hat{x}_n))_{i,j} - (G(x^*))_{i,j} \xrightarrow{P} 0$. Hence, we conclude that $|\mathbb{P}^*(\hat{Z}_n \in D) - \mathbb{P}(Z_0 \in D)| \to 0$ in probability. ■

A.2 ASGD and SAA Have Asymptotically Negligible Discrepancy

Theorem 7 *Under the same assumptions as Theorem 1, the outputs of ASGD and SAA have an asymptotically negligible discrepancy in the sense that*

$$\sqrt{n}(\bar{x}_n - \hat{x}_n) \to 0 \quad \textit{in probability.}$$

Proof Let $\bar{\Delta}_n = \bar{x}_n - x^*$, the residual of ASGD. Define $\bar{\Delta}^1_n$ by

$$\begin{aligned} &\Delta^1_0 = x_0 - x^*, \\ &\Delta^1_t = \Delta^1_{t-1} - \eta_t G(x^*)\Delta^1_{t-1} + \eta_t \xi_t, \ \xi_t = \nabla h(x_{t-1},\zeta_t) - \nabla H(x_{t-1}), \ \forall t = 1, \ldots, n, \\ &\bar{\Delta}^1_n = \frac{1}{n}\sum_{t=1}^n \Delta^1_t. \end{aligned}$$

We use superscript ^{1} to indicate that $\bar{\Delta}_n^1$ is an approximated version of $\bar{\Delta}_n$. Moreover, $\bar{\Delta}_n^1$ is also the residual of the ASGD solution for minimizing $\frac{1}{2}G(x^*)(x-x^*)^2$ with gradient noise sequence $\{\xi_t\}_t$. Consider the following decomposition:

$$\begin{aligned}&\sqrt{n}(\bar{x}_n - \hat{x}_n)\\ =&\sqrt{n}(\bar{\Delta}_n - \bar{\Delta}_n^1) + \sqrt{n}(\bar{\Delta}_n^1 - \frac{1}{n}\sum_{t=1}^n G(x^*)^{-1}\xi_t) - \sqrt{n}(\hat{x}_n - x^* - \frac{1}{n}\sum_{t=1}^n G(x^*)^{-1}\xi_t).\end{aligned}$$

It suffices to show that all three terms in the above decomposition vanish. In the decomposition, the first term describes the closeness of $\bar{\Delta}_n$ and its approximation with the same gradient noises but a surrogate objective. By the proof of Theorem 2, part 4 in Polyak and Juditsky (1992), $\sqrt{n}(\bar{\Delta}_n - \bar{\Delta}_n^1) \xrightarrow{P} 0$. The second term follows from the classical stochastic approximation result for linear problem. In particular, by the proof of Theorem 1, part 1 in Polyak and Juditsky (1992), $\sqrt{n}\left(\bar{\Delta}_n^1 - \frac{1}{n}\sum_{t=1}^n G(x^*)^{-1}\xi_t\right) \xrightarrow{P} 0$. The last term converges to 0 in probability by standard M-estimator theory; see, for example, Theorem 5.21 in Van der Vaart (2000). ■

A.3 Discussion on Assumptions

The following set of assumptions is listed in Shao and Zhang (2022) to guarantee a Berry-Esseen type bound for the residual of ASGD.

Assumption 12 *There exists a constant $\tau_0 > 0$ such that $\|x_0 - x^*\|_4 \le \tau_0$.*

Assumption 13 *The sequence $\{\nabla h(x_{t-1}, \zeta_t) - \nabla H(x_{t-1})\}_t$ is independent of x_0, and for each $t \ge 1$, $\nabla h(x_{t-1}, \zeta_t) - \nabla H(x_{t-1})$ admits the following decomposition:*

$$\nabla h(x_{t-1}, \zeta_t) - \nabla H(x_{t-1}) = \xi_t + \gamma_t,$$

where:

1. *$\{\xi_t\}$ is a sequence of independent random variables and $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}[\xi_i\xi_i^\top] = \Sigma_i$; there exist positive numbers λ_1 and λ_2 such that for any $i \ge 1$, $\lambda_1 \le \lambda_{min}(\Sigma_i) \le \lambda_{max}(\Sigma_i) \le \lambda_2$; moreover, there exists a positive number τ such that $\max_i \|\xi_i\|_4 \le \tau$.*

2. *Let $\mathcal{F}_0 = \sigma(x_0)$, and for each $t \ge 0$, $\mathcal{F}_t = \sigma(x_0, \zeta_k | k \le t)$; let $g(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, and let the random variable $\gamma_t = g(x_{t-1}, \zeta_t)$ satisfy $\mathbb{E}[\gamma_t|\mathcal{F}_{t-1}] = 0$. For any x and x', there exists a non-negative constant $c_1 \ge 0$ such that*

$$g(x,\zeta) - g(x',\zeta) \le c_1\|x - x'\| \quad \text{and} \quad g(x^*,\zeta) = 0,$$

for all $\zeta \in \mathbb{R}^d$.

Assumption 14 *The function H is L-smooth and strongly convex with convexity constant $\mu > 0$.*

Assumption 15 *There exist positive constants c_2 and β such that for all x with $\|x-x^*\| \leq \beta$:*

$$\|G(x) - G(x^*)\| \leq c_2\|x - x^*\|.$$

Lemma 8 *Consider the ASGD procedure* (2)*. Under Assumptions 12 to 15, for any Borel set D, we have:*

1. *if $\alpha \in (\frac{1}{2}, 1)$,*

$$|\mathbb{P}(\sqrt{n}\Sigma_n^{-\frac{1}{2}}(\bar{x}_n - x^*) \in D) - \mathbb{P}(Z^{ASGD} \in D)| \leq C(d^{3/2} + \tau^3 + \tau_0^3)(d^{1/2}n^{-1/2} + n^{-\alpha+1/2});$$

2. *if $\alpha = 1$, for any $\epsilon > 0$,*

$$|\mathbb{P}(\sqrt{n}\Sigma_n^{-\frac{1}{2}}(\bar{x}_n - x^*) \in D) - \mathbb{P}(Z^{ASGD} \in D)| \leq C(d^{3/2} + \tau^3 + \tau_0^3)n^{-1/2+\epsilon}d^{1/2}.$$

Our Assumptions 1 to 3 imply the above set of assumptions. Specifically, Assumption 1 implies Assumption 15. Assumptions 12 and 14 are implied by Assumptions 1 and 3. To see that our assumptions imply Assumption 13, let $g(x, \zeta) = \nabla h(x, \zeta) - \nabla h(x^*, \zeta) - \nabla H(x)$, and $\xi_t = \nabla h(x^*, \zeta_t)$. Given Assumption 3, and with ξ_t and $\gamma_t = g(x_{t-1}, \zeta_t)$, it follows that $\mathbb{E}[\xi_t\xi_t^\top] = \mathbb{E}[\nabla h(x^*, \zeta_t)\nabla h(x^*, \zeta_t)^\top] \equiv S(x^*)$. Consequently, the first condition of Assumption 13 holds. The martingale-difference structure of g, which is part of the second condition of Assumption 13, stems from Assumption 2. One can also verify that $g(x^*, \zeta) = 0$ by direct calculation. For the Lipschitz property of g, notice that

$$\|g(x, \zeta) - g(x', \zeta)\| \leq \|\nabla h(x, \zeta) - \nabla h(x', \zeta)\| + \|\nabla H(x) - \nabla H(x')\|,$$

where the right-hand-side is bounded by a constant multiple of $\|x - x'\|$ since the Hessian of both h and H have bounded spectrums, as stated in Assumption 1.

To conclude, the two bounds in Lemma 8 hold under our assumptions.

A.4 Proof of Theorem 2 in the ASGD Case

Proof Let D be a Borel set. Consider $\alpha \in (\frac{1}{2}, 1)$ first. As discussed in the main body. It is sufficient to establish the relationship (14). Specifically, we will show that there exist $(\hat{\tau}_0, \hat{\tau}, \hat{C})$ such that for any $\delta > 0$, there is an integer N (depending on δ), satisfying

$$\begin{aligned} &\sup_{n>N} |\mathbb{P}^*(\sqrt{n}(\psi_n(\hat{P}_n) - \psi(\hat{P}_n)) \in D) - \mathbb{P}^*(\hat{Z}_n \in D)| \\ &\quad \leq \hat{C}(d^{3/2} + \hat{\tau}^3 + \hat{\tau}_0^3)(d^{1/2}n^{-1/2} + n^{-\alpha+1/2}), \end{aligned} \tag{20}$$

with probability at least $1 - \delta$.

By Lemma 8, we have the following Berry-Esseen bound for fixed data distribution P:

$$\begin{aligned} &|\mathbb{P}(\sqrt{n}(\psi_n(P) - \psi(P)) \in D) - \mathbb{P}(Z_0 \in D)| \\ \leq &C(d^{3/2} + \tau^3 + \tau_0^3)(d^{1/2}n^{-1/2} + n^{-\alpha+1/2}), \end{aligned} \tag{21}$$

where $C = C(\eta, \lambda_1, \lambda_2, \alpha, l, L) > 0$ is a constant independent of D, d, τ, and τ_0. It remains to specify $\hat{C}$, $\hat{\tau}$, and $\hat{\tau}_0$ in (20).

Consider $\hat{\tau}_0$ first. Let $\epsilon_0, \delta_0 > 0$, and set $\hat{\tau}_0 = (1+\epsilon_0)\tau_0$. By Assumptions 1 and 4 and M-estimator theory (see Van der Vaart (2000) for example), we have $\psi(\hat{P}_n) \xrightarrow{P} \psi(P)$. As a result, there is $N_0 = N_0(\epsilon_0, \delta_0)$ such that for any $n > N_0$, we have $\|\psi(P) - \psi(\hat{P}_n)\| < \epsilon_0\tau_0$ with probability at least $1-\delta_0$. Consider the decomposition $\|x_0 - \psi(\hat{P}_n)\| \le \|x_0 - \psi(P)\| + \|\psi(P) - \psi(\hat{P}_n)\|$. Therefore, for any $n > N_0$, $\mathbb{P}(\|x_0 - \psi(\hat{P}_n)\| \le \hat{\tau}_0) > 1 - \delta_0$.

Next, consider $\hat{\tau}$. For any $n > N_0$,

$$\begin{aligned}\|\nabla h(\psi(\hat{P}_n), \hat{\zeta})\|_4 &\le \|\nabla h(\psi(P), \hat{\zeta})\|_4 + \|\nabla h(\psi(P), \hat{\zeta}) - \nabla h(\psi(\hat{P}_n), \hat{\zeta})\|_4 \\ &\le \|\nabla h(\psi(P), \hat{\zeta})\|_4 + L\tau_0,\end{aligned}$$

where $\hat{\zeta}$ follows distribution $\hat{P}_n$. By the law of large numbers, for the first term, we have

$$\|\nabla h(\psi(P), \hat{\zeta})\|_4 \xrightarrow{n\to\infty} \|\nabla h(\psi(P), \zeta)\|_4 \le \tau_0.$$

As a result, $\|\nabla h(\psi(P), \hat{\zeta})\|_4$ is bounded by a constant not depending on n. So we conclude that there exists such a $\hat{\tau}$ that $\|\nabla h(\psi(\hat{P}_n), \hat{\zeta})\|_4 \le \hat{\tau}$ for all $n > N_0$.

To determine $\hat{C}$, it suffices to find $\hat{\lambda}_1, \hat{\lambda}_2, \hat{l}$ and $\hat{L}$, since α and η are not affected by the underlying data distribution P.

Consider $\hat{\lambda}_1$ and $\hat{\lambda}_2$ first. $S(x) \triangleq \mathbb{E}_{\zeta\sim P}[\nabla h(x,\zeta)(\nabla h(x,\zeta))^\top]$ and the corresponding empirical version $S_n(x) \triangleq \mathbb{E}_{\hat{\zeta}\sim\hat{P}_n}[\nabla h(x,\hat{\zeta})(\nabla h(x,\hat{\zeta}))^\top]$. We have:

$$\|(S_n(\psi(\hat{P}_n)) - S(x^*))\| \le \|(S_n(\psi(\hat{P}_n)) - S_n(x^*))\| + \|(S_n(x^*) - S(x^*))\|. \tag{22}$$

Consider the first term on the right-hand side of (22). We have:

$$\begin{aligned}&\|(S_n(\psi(\hat{P}_n)) - S_n(x^*))\| \\ =&\|\mathbb{E}_{\hat{\zeta}\sim\hat{P}_n}[\nabla h(\psi(\hat{P}_n),\hat{\zeta})(\nabla h(\psi(\hat{P}_n),\hat{\zeta}))^\top - \nabla h(x^*,\hat{\zeta})(\nabla h(x^*,\hat{\zeta}))^\top]\| \\ \le&\|\mathbb{E}_{\hat{\zeta}\sim\hat{P}_n}[\nabla h(\psi(\hat{P}_n),\hat{\zeta})(\nabla h(\psi(\hat{P}_n),\hat{\zeta}))^\top - \nabla h(x^*,\hat{\zeta})(\nabla h(\psi(\hat{P}_n),\hat{\zeta}))^\top]\| \\ &+ \|\mathbb{E}_{\hat{\zeta}\sim\hat{P}_n}[\nabla h(\psi(\hat{P}_n),\hat{\zeta})(\nabla h(x^*,\hat{\zeta}))^\top - \nabla h(\psi(P),\hat{\zeta})(\nabla h(x^*,\hat{\zeta}))^\top]\| \\ \le&\mathbb{E}_{\hat{\zeta}\sim\hat{P}_n}[\|\nabla h(\psi(\hat{P}_n),\hat{\zeta}) - \nabla h(x^*,\hat{\zeta})\|\|\nabla h(\psi(\hat{P}_n),\hat{\zeta})\|] \\ &+ \mathbb{E}_{\hat{\zeta}\sim\hat{P}_n}[\|\nabla h(\psi(\hat{P}_n),\hat{\zeta}) - \nabla h(x^*,\hat{\zeta})\|\|\nabla h(x^*,\hat{\zeta})\|] \\ \le&\mathbb{E}[L\|\psi(\hat{P}_n) - x^*\|\|\nabla h(\psi(\hat{P}_n),\hat{\zeta})\|] + \mathbb{E}[L\|\psi(\hat{P}_n) - x^*\|\|\nabla h(\psi(P),\hat{\zeta})\|] \xrightarrow{P} 0.\end{aligned}$$

In the above derivations, the first and second inequalities follow standard inequalities. The third inequality follows from Assumption 1. The last convergence holds since $\|\psi(\hat{P}_n) - x^*\|$ converges to 0 in probability and the remaining factors are bounded.

The second term of (22), $\|(S_n(x^*) - S(x^*))\|$, converges to 0 almost surely by the law of large numbers. As a result, for any $\epsilon_1, \delta_1 > 0$, there is $N = N(\epsilon_1, \delta_1, \epsilon_0, \delta_0) > N_0$ such that, $\forall n > N$, eigenvalues of $S_n(\hat{\psi}(\hat{P}_n))$ lies in $[\frac{1}{1+\epsilon}\lambda_1, (1+\epsilon)\lambda_2]$ with probability at least $1-\delta_1$. So we can choose $\hat{\lambda}_1 = \frac{1}{1+\epsilon}\lambda_1$ and $\hat{\lambda}_2 = (1+\epsilon)\lambda_2$.

In a similar manner, we can obtain desired constants $\hat{l}$ and $\hat{L}$ and conclude the proof for (20).

Now, we consider $\alpha = 1$. Similar to the case when $\alpha \in (\frac{1}{2}, 1)$, it suffices to show that there exists a 4-tuple $(\hat{\tau}_0, \hat{\tau}, \hat{C}, N)$ such that for any $\epsilon > 0$:

$$\sup_{n>N} |\mathbb{P}^*(\sqrt{n}(\psi_n(\hat{P}_n) - \psi(\hat{P}_n)) \in D) - \mathbb{P}^*(\hat{Z}_n \in D)| \leq \hat{C}(d^{\frac{3}{2}} + \hat{\tau}^3 + \hat{\tau}_0^3)n^{-\frac{1}{2}+\epsilon}d^{\frac{1}{2}}. \tag{23}$$

Theorem 3.4 in Shao and Zhang (2022) gives that when $\alpha = 1$, for all $\epsilon > 0$,

$$|\mathbb{P}(\sqrt{n}(\psi_n(P) - \psi(P)) \in D) - \mathbb{P}(Z_0 \in D)| \leq C(d^{3/2} + \tau^3 + \tau_0^3)n^{-\frac{1}{2}+\epsilon}d^{\frac{1}{2}}, \tag{24}$$

where $C = C(\eta, \lambda_1, \lambda_2, \alpha, l, L) > 0$ is a constant independent of D, d, τ, and τ_0. By the same arguments, we can derive (23) from (24). ■

A.5 Proof of Theorem 2 in the SGD Case

Before proving Theorem 2 in the SGD case, we state Corollary 2.3 from Shao and Zhang (2022) that plays a critical role in our proof.

Lemma 9 *(Shao and Zhang, 2022) Let T be a d-dimensional statistic. Suppose that $T = T(\zeta_1, \ldots, \zeta_n)$ admits the decomposition $T = W + D$, where $W = \sum_{i=1}^n f_i(\zeta_i)$ satisfies:*

$$\mathbb{E}[f_i(\zeta_i)] = 0, \forall i, \quad \textit{and} \ \sum_{i=1}^n \mathbb{E}[f_i(\zeta_i)(f_i(\zeta_i))^\top] = \Sigma_n,$$

where $0 < \sigma_n = \lambda_{min}(\Sigma_n)$, and f is some mapping to $\mathbb{R}^d$. Define $\gamma_n = \sum_{i=1}^n \mathbb{E}\|f_i(\zeta_i)\|^3$. Let Δ and $(\Delta^{(i)})1 \leq i \leq n$ be random variables such that $\Delta \geq \|D\|$ and $\Delta^{(i)}$ is independent of ζ_i. Let $Z_0 \sim \mathcal{N}(0, I_d)$. Then, for any Borel set D,

$$\begin{aligned} &|\mathbb{P}(\Sigma_n^{-1/2}T \in D) - \mathbb{P}(Z_0 \in D)| \\ &\leq 259\sigma_n^{-3/2}d^{1/2}\gamma_n + 2\sigma_n^{-1}\mathbb{E}[\|W\|\Delta] + 2\sigma_n^{-1}\sum_{i=1}^n \mathbb{E}[\|f_i(\zeta_i)\||\Delta - \Delta^{(i)}|]. \end{aligned}$$

The above lemma gives the Berry-Esseen bound for any statistic T that has a specific decomposition. Now we are ready to prove Theorem 2 in the SGD case.

Proof We start with the following recursive formula derived from (2):

$$x_{k+1} - x^* = x_k - x^* - \eta_k G(x^*)(x_k - x^*) - \eta_k\delta_k - \eta_k E_k, \tag{25}$$

where $\delta_k \triangleq \delta(x_k) = \nabla H(x_k) - G(x^*)(x_k - x^*)$, and $E_{k-1} = \nabla h(x_{k-1}, \zeta_k) - \nabla H(x_{k-1})$.

Let $G(x^*) = U\Lambda U^\top$ be the singular value decomposition of $G(x^*)$, where Λ is the diagonal matrix consisting of eigenvalues of $G(x^*)$. Then, (25) becomes d parallel updates with respect to $x_n' = U^\top x_n$, $\nabla h' = U^\top \nabla h$. As a result, we can assume $d = 1$ without loss of generality. In this case, $G(x^*)$ becomes a scalar, and we let $\alpha_1 = G(x^*)$, using the notation α_1 later on so that it is not recognized as a matrix.

The goal is to show that

$$\sup_t |\mathbb{P}^*(\sqrt{n}(\psi_n(\hat{P}_n) - \psi(\hat{P}_n)) \leq t) - \mathbb{P}^*(\hat{Z}_n \leq t)| \to 0, \text{ in probability as } n \to \infty,$$

where $\hat{Z}_n \sim \mathcal{N}(0, \tilde{\sigma}^2(\hat{P}_n))$, and $\tilde{\sigma}^2(P) = \eta^2(2\eta\alpha_1 - 1)^{-1}\mathrm{Var}_{\zeta\sim P}(\nabla h(\psi(P), \zeta))$ for generic distribution P.

Define $\beta_{mn} = \prod_{j=m+1}^{n}(1 - \eta_j\alpha_1) \in \mathbb{R}$ and $h_n = (\sum_{m=1}^{n}\alpha_1^2\eta_m^2\beta_{mn}^2)^{-\frac{1}{2}} \in \mathbb{R}$. We have the following closed-form expression of $h_n x_{n+1}$,

$$h_n(x_{n+1} - x^*) = h_n\beta_{0n}(x_1 - x^*) - h_n\sum_{m=1}^{n}\eta_m\beta_{mn}\delta_m - h_n\sum_{m=1}^{n}\eta_m\beta_{mn}E_m.$$

To simplify the notation later on, let $I_n^{(1)}$, $I_n^{(2)}$, and $I_n^{(3)}$ represent the three terms on the right-hand side respectively. Namely, $I_n^{(1)} = h_n\beta_{0,n}(x_1 - x^*)$, $I_n^{(2)} = h_n\sum_{m=1}^{n}\eta_m\beta_{mn}\delta_m$ and $I_n^{(3)} = h_n\sum_{m=1}^{n}\eta_m\beta_{mn}E_m$.

Let $Y_n(P) = \sqrt{n}(\psi_n(P) - \psi(P))$, and define $I_n^{(i)}(P)$ in a similar manner when addressing the underlying data distribution. Note that $I_n^{(1)}$ is deterministic and does not depend on P. Consider the following decomposition:

$$\begin{aligned}
&|\mathbb{P}^*(\sqrt{n}(\psi_n(\hat{P}_n) - \psi(\hat{P}_n)) \le t) - \mathbb{P}^*(Z_n \le t)| \\
\le&|\mathbb{P}^*(Y_n(\hat{P}_n) \le t) - \mathbb{P}^*(Y_n(\hat{P}_n) - \frac{\sqrt{n}}{h_n}(I_n^{(1)} - I_n^{(2)}(\hat{P}_n)) \le t)| \\
&+ |\mathbb{P}^*(Y_n(\hat{P}_n) - \frac{\sqrt{n}}{h_n}(I_n^{(1)} - I_n^{(2)}(\hat{P}_n)) \le t) - \mathbb{P}^*(Z_n \le t)|.
\end{aligned} \tag{26}$$

The remainder of this proof will show that both terms on the right-hand side converge to 0 in probability.

To see that $|\mathbb{P}^*(Y_n(\hat{P}_n) \le t) - \mathbb{P}^*(Y_n(\hat{P}_n) - \frac{\sqrt{n}}{h_n}(I_n^{(1)}(\hat{P}_n) - I_n^{(2)}(\hat{P}_n)) \le t)| \xrightarrow{P} 0$, it suffices to verify that $I_n^{(1)} \to 0$ and $I_n^{(2)}(\hat{P}_n) \to 0$ conditional on $\zeta_1, \zeta_2, \ldots$, in probability. This follows from the (3.9a) and (3.9b) in Sacks (1958).

It remains to show that $|\mathbb{P}^*(Y_n(\hat{P}_n) - \frac{\sqrt{n}}{h_n}(I_n^{(1)}(\hat{P}_n) - I_n^{(2)}(\hat{P}_n)) \le t) - \mathbb{P}^*(Z_n \le t)| \xrightarrow{P} 0$. We will make use of Lemma 9 to prove this claim. Following the framework for proving the ASGD case, it suffices to establish a bound similar to (21). To be specific, we need to show that, for any Borel set D,

$$|P(-\frac{\sqrt{n}}{h_n}I_n^{(3)}(P) \in D) - P(\hat{Z}_n \in D)| \le C_n,$$

where C_n converges to 0 as $n \to \infty$, and C_n (to be specified) depends on $\eta, \lambda_1, \lambda_2, l, L, \tau$, and τ_0. To simplify the notation, let

$$A_n = \nabla h(x^*, \zeta_n), \quad B_n = E_n - A_n = \nabla h(x_{n-1}, \zeta_n) - \nabla h(x^*, \zeta_n) - \nabla H(x_{n-1}).$$

We have

$$I_n^{(3)} = h_n\sum_{m=1}^{n}\eta_m\beta_{mn}A_m + h_n\sum_{m=1}^{n}\eta_m\beta_{mn}B_m.$$

Let $Q_{mn} = \eta_m\beta_{mn}$, and $\bar{\Sigma}_n = h_n^2\sum_{m=1}^{n}Q_{mn}^2\tilde{\sigma}(P)$. Define

$$T_n \triangleq \bar{\Sigma}_n^{-\frac{1}{2}}I_n^{(3)} = W_n + D_n,$$

where $W_n = \sum_{m=1}^n Y_{mn}$, $Y_{mn} = h_n \bar{\Sigma}_n^{-\frac{1}{2}} Q_{mn} A_m$, and $D_n = h_n \sum_{m=1}^n Q_{mn} B_m \bar{\Sigma}_n^{-\frac{1}{2}}$. Notice that $\{Y_{mn}\}_{m=1}^n$ are independent and that

$$\begin{cases} \mathbb{E}[W_n] = 0 \\ Var(W_n) = 1 \\ T_n = W_n + D_n \end{cases}$$

Now, we construct Δ and $\{\Delta^{(i)}\}$ that will later be useful when applying Lemma 9. The idea is borrowed from Shao and Zhang (2022). Let $(\zeta_1', \ldots, \zeta_n')$ be an independent copy of $(\zeta_1, \ldots, \zeta_n)$ and define $(A_1', \ldots, A_n')$ according to the relationship $A_i' = \nabla h(x^*, \zeta_i')$, $i = 1, \ldots, n$. For each i, construct $x_1^i, \ldots, x_n^i$ as follows:

- If $j < i$, $x_j^{(i)} = x_j$
- If $j = i$, $x_j^{(i)} = x_{j-1}^{(i)} - \eta_j(\nabla H(x_j^{(i)}) + A_j' + B(x_j^{(i)}, A_j'))$
- If $j > i$, $x_j^{(i)} = x_{j-1}^{(i)} - \eta_j(\nabla H(x_j^{(i)}) + A_j + B(x_j^{(i)}, A_j))$

That is, $(x_1^i, \ldots, x_n^i)$ is obtained by running SGD with only the i-th data replaced by ζ_i'. It is worth noting that this construction is only for the sake of establishing the Berry-Esseen type bound, and it is not required in real applications. For each $i = 1, \ldots, n$, we define $T_n^{(i)}$, $W_n^{(i)}$ and $D_n^{(i)}$ following the same procedure described above. And notice that $\forall 1 \le i \le n$, $D_n^{(i)} \perp\!\!\!\perp \zeta_i$. Let $\Delta \triangleq \|D_n\|$ and $\Delta^{(i)} \triangleq \|D_n^{(i)}\|$. Then $\Delta^{(i)} \perp\!\!\!\perp \zeta_i$, $\forall i = 1, \ldots, n$.

Applying Lemma 9 to $T_n = W_n + D_n$, Δ, and $\{\Delta^{(i)}\}$ defined above, we obtain the following inequality:

$$|\mathbb{P}(T_n \in D) - \mathbb{P}(Z_0 \in D)| \le 259\gamma_n + 2\mathbb{E}[\|W_n\|\Delta] + 2\sum_{i=1}^n \mathbb{E}[\|Y_{in}\||\Delta - \Delta^{(i)}|], \tag{27}$$

where $\gamma_n = \sum_{i=1}^n \mathbb{E}|Y_{in}|^3 = h_n \bar{\Sigma}_n^{-\frac{3}{2}} Q_{in}^3 |A_m|^3$ and Z_0 follows a standard normal distribution. To establish the desired conclusion, it remains to prove that the following three terms vanish as $n \to \infty$:

- γ_n
- $\mathbb{E}[\|W_n\|\Delta]$
- $\sum_{i=1}^n \mathbb{E}[\|Y_{in}\||\Delta - \Delta^{(i)}|]$

Consider γ_n first. By the definition of Y_{mn} and $\bar{\Sigma}_n$, we have

$$\gamma_n = \sum_{m=1}^n \mathbb{E}[|Y_{mn}|^3] = (\sum_{i=1}^n Q_{in}^2 \bar{\Sigma}^2)^{-\frac{3}{2}} \sum_{m=1}^n Q_{mn}^3 \mathbb{E}[|A_m|^3].$$

Since $Q_{mn} = \eta_m \beta_{mn} = \eta m^{-1} \prod_{j=m+1}^n (1 - \alpha_1 \eta j^{-1})$, it is not hard to derive the following sandwich inequality (see, e.g. (2.3) in Sacks (1958))

$$(1 - \epsilon_m') m^{\alpha_1 \eta - 1} n^{-\alpha_1 \eta} \le Q_{mn} \le (1 + \epsilon_m') m^{\alpha_1 \eta - 1} n^{-\alpha_1 \eta},$$

for any $n \geq m$, where $\epsilon'_m \to 0$ as $m \to \infty$. By Assumption 3, $\mathbb{E}[|A_m|^3] \leq \tau^3$, so there exists constant $C_\gamma > 0$ such that

$$\gamma_n \leq C_\gamma \sum_{j=1}^{n} j^{3\alpha_1\eta-3}(\sum_{m=1}^{n} m^{2\alpha_1\eta-2})^{-\frac{3}{2}}.$$

The magnitude of the above expression depends on the value of $\alpha_1\eta$. To be specific, we have

$$\gamma_n \leq \begin{cases} C_\gamma n^{-\frac{1}{2}} & \text{if } \alpha_1\eta > \frac{2}{3} \\ C_\gamma n^{-\frac{1}{2}} \log n & \text{if } \alpha_1\eta = \frac{2}{3} \\ C_\gamma n^{-3\alpha_1\eta+\frac{3}{2}} & \text{if } \frac{1}{2} < \alpha_1\eta < \frac{2}{3} \\ C_\gamma (\log n)^{-\frac{3}{2}} & \text{if } \alpha_1\eta = \frac{1}{2} \\ C_\gamma & \text{if } \alpha_1\eta < \frac{1}{2} \end{cases}$$

for some C_γ depending on $\eta, \lambda_1, \lambda_2, \alpha, l, L$.

Now, we consider $\mathbb{E}[\|W_n\|\Delta]$. There exists constant $c' > 0$ such that

$$\begin{aligned}
\mathbb{E}[|W_n|\Delta] \leq & (\mathbb{E}[W_n^2]\mathbb{E}[\Delta^2])^{\frac{1}{2}} \\
= & (\mathbb{E}[\Delta^2])^{\frac{1}{2}} \\
= & (\mathbb{E}[(h_n \sum_{m=1}^{n} Q_{mn} B_m \bar{\Sigma}_n^{-\frac{1}{2}})^2])^{\frac{1}{2}} \\
= & (\mathbb{E}[(\bar{\Sigma}^{-\frac{1}{2}} (\frac{(\sum_{i=1}^{n} Q_{in} B_i)^2}{\sum_m Q_{mn}^2}))^2])^{\frac{1}{2}} \\
= & (\mathbb{E}[\bar{\Sigma}^{-\frac{1}{2}} \frac{\sum_{m=1}^{n} Q_{mn}^2 B_m^2}{\sum_{m=1}^{n} Q_{mn}^2}])^{\frac{1}{2}} \\
= & (\bar{\Sigma}^{-\frac{1}{2}} \frac{\sum_m Q_{mn}^2 \mathbb{E}[B_m^2]}{\sum_m Q_{mn}^2})^{\frac{1}{2}} \\
\leq & c' (\frac{\sum_{i=1}^{n} i^{2\alpha_1\eta-2} \mathbb{E}[B_m^2]}{\sum_{m=1}^{n} m^{2\alpha_1\eta-2}})^{\frac{1}{2}}.
\end{aligned}$$

For any m, by Assumption 1,

$$\begin{aligned}
|B_m| = & |\nabla h(x_{m-1}, \zeta_m) - \nabla h(x^*, \zeta_m) - \nabla H(x_{m-1}) + \nabla H(x^*)| \\
\leq & |\nabla h(x_{m-1}, \zeta_m) - \nabla h(x^*, \zeta_m)| + |\nabla H(x_{m-1}) - \nabla H(x^*)| \\
\leq & 2L|x_{m-1} - x^*|.
\end{aligned}$$

Combining Lemma 5.12 in Shao and Zhang (2022) with the above inequality, we get:

$$\mathbb{E}[B_m^2] \leq 2L\mathbb{E}[|x_{m-1} - x^*|^2] \leq \begin{cases} cn^{-1} & \text{if } \alpha_1\eta > 1, \\ cn^{-1} \log n & \text{if } \alpha_1\eta = 1, \\ cn^{-\alpha_1\eta} & \text{if } \alpha_1\eta < 1, \end{cases}$$

for some constant $c = c(\tau, \tau_0, L) > 0$. Thus, there exists constant $\hat{c} = \hat{c}(\tau, \tau_0, L) > 0$ such that:

$$\mathbb{E}[|W_n|\Delta] \leq \begin{cases} \hat{c}n^{-\frac{1}{2}} & \text{if } \alpha_1\eta > 1, \\ \hat{c}n^{-\frac{1}{2}}(\log^2 n) & \text{if } \alpha_1\eta = 1, \\ \hat{c}n^{-2\alpha_1\eta+1} & \text{if } \alpha_1\eta \in (\frac{1}{2}, 1), \\ \hat{c}(\log n)^{-1} & \text{if } \alpha_1\eta = \frac{1}{2}, \\ \hat{c} & \text{if } \alpha_1\eta < \frac{1}{2}. \end{cases}$$

For the third term, consider $\mathbb{E}[Y_{in}^2]$ first, we have:

$$\begin{aligned} \mathbb{E}[Y_{in}^2] &= \mathbb{E}[\bar{\Sigma}^{-2}(\sum_{m=1}^{n} Q_{mn}^2)^{-1} Q_{in}^2 A_i^2] \\ &= \bar{\Sigma}^{-2}(\sum_{m=1}^{n} Q_{mn}^2)^{-1} Q_{in}^2 \mathbb{E}[A_i^2] \\ &\leq c(\sum_{m=1}^{n} m^{2\alpha_1\eta-2} n^{-2\alpha_1\eta})^{-1} i^{2\alpha_1\eta-2} n^{-2\alpha_1\eta} \\ &= c(\sum_{m=1}^{n} m^{2\alpha_1\eta-2})^{-1} i^{2\alpha_1\eta-2}. \end{aligned}$$

Consider $\mathbb{E}[|\Delta - \Delta^{(i)}|^2]$. By definition of Δ and $\Delta^{(i)}$:

$$\mathbb{E}[|\Delta - \Delta^{(i)}|^2] = \mathbb{E}[\frac{(\sum_{m=1}^{n} Q_{mn}(B_m - B_m^{(i)}))^2}{\sum_{m=1}^{n} Q_{mn}^2}].$$

Since $(B_m - B_m^{(i)})_m$ forms a martingale difference sequence for any i with respect to its canonical filtration, the crossing terms in the above complete square are zero. By definition of B_m and $B_m^{(i)}$, we have:

$$B_m - B_m^{(i)} = \begin{cases} 0 & m < i, \\ B(x_{m-1}, A_m) - B(x_{m-1}, A'_m) & m = i, \\ B(x_{m-1}, A_m) - B(x_{m-1}^{(i)}, A_m) & m > i. \end{cases}$$

As a result,

$$\mathbb{E}[|\Delta - \Delta^{(i)}|^2] = \frac{Q_{in}^2}{\sum_{m=1}^{n} Q_{mn}^2} \mathbb{E}[(B_i - B_i^{(i)})^2] + \frac{1}{\sum_{m=1}^{n} Q_{mn}^2} \sum_{m>n} Q_{mn}^2 \mathbb{E}[(B_m - B_m^{(i)})^2].$$

Lemma 5.13 in Shao and Zhang (2022) provides the following bound for the martingale difference term $(B_j - B_j^{(i)})$:

$$\mathbb{E}[(B_j - B_j^{(i)})^2] \leq c(\tau^2 + \tau_0^2) i^{-2} (\frac{i}{j})^{2\alpha_1\eta}.$$

Combining the above two relationships, there is some constant $\hat{c}$ such that

$$\mathbb{E}[|\Delta - \Delta^{(i)}|^2] \leq \hat{c}(\tau^2 + \tau_0^2)\frac{1}{\sum_m m^{2\alpha_1\eta-2}} i^{2\alpha_1\eta-3}.$$

Now we are ready to bound $\sum_i \mathbb{E}[|Y_{in}||\Delta - \Delta^{(i)}|]$. There exists constant $\hat{c}$, depending on η, λ_1, λ_2, l, L, τ, τ_0, such that:

$$\begin{aligned}\sum_i \mathbb{E}[|Y_{in}||\Delta - \Delta^{(i)}|] &\leq \sum_i (\mathbb{E}[Y_{in}^2]\mathbb{E}[(\Delta - \Delta^{(i)})^2])^{\frac{1}{2}} \\ &\leq \hat{c}(\sum_m m^{2\alpha_1\eta-2})^{-1} \sum_i i^{2\alpha_1\eta-5/2}.\end{aligned}$$

So we have the following vanishing rate for $\sum_i \mathbb{E}[|Y_{in}||\Delta - \Delta^{(i)}|]$, depending on the choice of $\alpha_1\eta$:

$$\sum_i \mathbb{E}[|Y_{in}||\Delta - \Delta^{(i)}|] \leq \begin{cases} \hat{c}n^{-\frac{1}{2}} & \text{if } \alpha_1\eta > \frac{3}{4}, \\ \hat{c}n^{-\frac{1}{2}}\log n & \text{if } \alpha_1\eta = \frac{3}{4}, \\ \hat{c}n^{-2\alpha_1\eta+1} & \text{if } \frac{1}{2} < \alpha_1\eta < \frac{3}{4}, \\ \hat{c}(\log n)^{-1} & \text{if } \alpha_1\eta = \frac{1}{2}, \\ \hat{c} & \text{if } 0 < \alpha_1\eta < \frac{1}{2}. \end{cases}$$

Thus, we have established the vanishing property of the right-hand-side of (27), which we now denote by C_n. Specifically,

$$|P(\bar{\Sigma}_n^{-\frac{1}{2}} I_n^{(3)} \in D) - P(Z_0 \in D)| \leq C_n.$$

Therefore, for any measurable set D,

$$|P(-\frac{\sqrt{n}}{h_n} I_n^{(3)} \in D) - P(-\frac{\sqrt{n}}{h_n}\bar{\Sigma}_n^{\frac{1}{2}} Z_0 \in D)| \leq C_n.$$

By Sacks (1958), $\frac{\sqrt{n}}{h_n}\bar{\Sigma}_n^{\frac{1}{2}} \to \eta(2\eta\alpha_1 - 1)^{-\frac{1}{2}}\tilde{\sigma}(P)^{\frac{1}{2}}$ as $n \to \infty$. This concludes the proof. ■

A.6 Proof of Theorem 5

Theorem 5 is an immediate consequence of the following result regarding the consistency of Lasso.

Lemma 10 *(Zhao and Yu, 2006) Under Assumptions 7 and 8, we have*

$$\lim_{n\to\infty} \mathbb{P}(\hat{x}^{(n)}(\lambda_n) =_s x^{(n)}) = 1,$$

where $=_s$ denotes equality in sign, which holds true if and only if the values on both sides are either both positive, both negative, or both zero.

Let $i \notin T^*$, and notice that $\mathbb{P}(\hat{x}^{(n)}(\lambda_n) \leq \mathbb{P}(x_i^{(n)} = 0)$. Lemma 10 directly implies that $\mathbb{P}(x_i^{(n)} \in \mathcal{I}_{i,n}) \to 1$ for $i \notin T^*$. For $i \in T^*$, $\mathbb{P}(x_i^{(n)} \in \mathcal{I}_{i,n}) \to 1 - \gamma$ follows by combining Lemma 10 and Theorem 1.

Appendix B. Additional Numerical Results

Please refer to Table 9 for full results for linear regression, Table 10 for logistic regression, and Table 11 for sparse linear regression.

Appendix C. Additional Discussion on Assumptions and Further Sensitivity Analyses

Assumption 1 imposes uniform strong convexity of the sample loss $h(\cdot, \zeta)$. This assumption can potentially be relaxed. Essentially, we need, with high probability, that the empirical Hessians for both the original and bootstrap samples to be well-conditioned, i.e.,

$$\lambda_{min}(G_n(x^*)) \geq c, \lambda_{min}(G^{*b}(x^*)) \geq c,$$

for some $c > 0$ for all bootstrap replicates b, together with local smoothness of H and stability of the (A)SGD iterates near x^*. Under these conditions, the nonlinear remainder term involving δ_m in (13) in the error analysis is summable along the (A)SGD trajectory and asymptotically negligible at the $\sqrt{n}$ scale with probability tending to one sufficiently fast, so that the uniform control over all bootstrap replicates required later in the proof remains valid. Establishing our results using these relaxed conditions would be delegated to future work.

In the following subsections, we present additional numerical experiments to illustrate the behavior of the proposed method when key assumptions are violated, mainly to investigate the robustness of our proposed methods.

C.1 Ill-conditioned Curvature

We first examine the effect of ill-conditioned curvature by considering ridge regression with varying levels of conditioning in the feature covariance matrix Σ. Throughout, we quantify conditioning by the condition number

$$\kappa(\Sigma) = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma),$$

where $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ denote the smallest and largest eigenvalues of Σ respectively. Specifically, we consider the model

$$b = a^\top x^* + \varepsilon, \quad a \sim \mathcal{N}(0, \Sigma), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

and minimize the ridge regression objective

$$H_\lambda(x) = \mathbb{E}\big[\frac{1}{2}\mathbb{E}[(a^\top x - b)^2]\big] + \frac{\lambda}{2}\|x\|^2.$$

Under this model, the population Hessian of H_λ equals $\Sigma + \lambda I$. $\kappa(\Sigma)$ governs the anisotropic curvature induced by the data, while λ adds isotropic curvature that improves conditioning by increasing small eigenvalues. In particular, increasing $\kappa(\Sigma)$ makes the problem more ill-conditioned, whereas increasing λ mitigates ill-conditionedness by shrinking $\kappa(\Sigma + \lambda I)$.

We construct Σ as an equicorrelation matrix with unit diagonal to isolate the effect of curvature rather than feature scaling. Coverage is evaluated with respect to x^*_λ that

Table 5: COfB under curvatures with different conditioning (ridge regression). Empirical coverages (nominal 0.95) of componentwise confidence intervals for the ridge minimizer x^*_λ, with average interval lengths in parentheses. Features have unit variance and equicorrelation structure with condition number $\kappa(\Sigma)$.

$\kappa(\Sigma)$	$\lambda = 0$	10^{-4}	10^{-3}	10^{-2}
10	0.975 (0.02)	0.975 (0.02)	0.975 (0.02)	0.975 (0.02)
10^2	0.915 (0.08)	0.915 (0.08)	0.915 (0.08)	0.915 (0.07)
10^3	0.735 (0.67)	0.745 (0.66)	0.765 (0.64)	0.845 (0.47)

Table 6: COnB under curvatures with different conditioning (ridge regression). Empirical coverages (nominal 0.95) of componentwise confidence intervals for the ridge minimizer x^*_λ, with average interval lengths in parentheses. All settings match Table 5.

$\kappa(\Sigma)$	$\lambda = 0$	10^{-4}	10^{-3}	10^{-2}
10	0.965 (0.03)	0.965 (0.03)	0.965 (0.03)	0.965 (0.03)
10^2	0.995 (0.53)	0.995 (0.53)	0.995 (0.52)	0.995 (0.47)
10^3	0.825 (10.34)	0.825 (10.31)	0.835 (9.98)	0.915 (7.29)

minimizes the ridge regression objective. Tables 5 and 6 report empirical coverages with nominal level 95% and average interval widths of COfB and COnB, respectively. The upper and right parts of each table reflect the well-conditioned regime, and the lower-left portion corresponds to the most ill-conditioned setting.

The results illustrate some robustness of both methods to curvature degradation, but also caution the sensitivity to extreme ill-conditionedness. When the problem is well-conditioned ($\kappa(\Sigma) = 10$), both COfB and COnB achieve near nominal coverages for all choices of λ, with short confidence intervals. As $\kappa(\Sigma)$ increases, coverages deteriorate and the interval lengths grow, indicating that ill-conditioned curvature negatively impacts the performance of both methods. Nonetheless, the coverages in this setting are still moderately close to the nominal, with the interval lengths of COfB relatively small but those of COnB larger. In the most ill-conditioned case when $\kappa(\Sigma) = 10^3$ and the ridge regularization parameter λ is small, both methods substantially undercover and COnB in particular produces very wide intervals, although its coverages are better than COfB. These observations put caution on using our methods in extreme ill-conditioned problems.

C.2 Learning Rate Schemes

In this subsection, we examine the stability of our methods to the step size assumptions by varying the learning rate schedule while keeping the objective well-conditioned.

We consider linear regression with a well-conditioned design, fixing $d = 20$, $n = 10^5$, $\kappa(\Sigma) = 10$. We evaluate a range of learning rate schemes that may not be covered by our theory, including power-law decay with $\alpha = 0.501$ (covered by our theory), and $\alpha = 0.25$ and a constant step size. In addition, we also consider piecewise constant schedule and cosine annealing. Empirical coverages and average confidence interval lengths of the solutions are

Table 7: COfB under different learning rate schemes. Empirical coverages (nominal 0.95) of componentwise confidence intervals for the minimizer x^* under different learning rate schemes, with average confidence interval lengths in parentheses. All settings use $n = 10^5$, $d = 20$, equicorrelated Σ with $\kappa(\Sigma) = 10$, noise variance $\sigma^2 = 1$, and $B = 5$ bootstrap replicates.

Learning-rate scheme	Coverage (Avg. length)
Power-law ($\alpha = 0.501$)	0.985 (0.0188)
Power-law ($\alpha = 0.25$)	0.955 (0.0187)
Constant step size	0.940 (0.0187)
Piecewise constant	0.980 (0.0216)
Cosine annealing	0.965 (0.0192)

Table 8: COnB under different learning rate schemes. Empirical coverages (nominal 0.95) of componentwise confidence intervals for the minimizer x^* under different learning rate schemes, with average confidence interval lengths in parentheses. All settings match Table 7.

Learning-rate scheme	Coverage (Avg. length)
Power-law ($\alpha = 0.501$)	0.960 (0.0218)
Power-law ($\alpha = 0.25$)	0.960 (0.0169)
Constant step size	0.955 (0.0190)
Piecewise constant	0.960 (0.0217)
Cosine annealing	0.955 (0.0182)

reported for both COfB and COnB. Table 7 and Table 8 report empirical coverages with nominal level 95% and average interval widths of COfB and COnB, respectively.

Across all learning rate schemes considered, both COfB and COnB achieve near nominal empirical coverages, with only modest variation in confidence interval lengths. These results indicate that while the step size conditions imposed in the theory are sufficient for establishing asymptotic validity, the proposed methods exhibit substantial empirical robustness to learning rate scheme in well-conditioned, large-sample settings.

	$d=5$		$d=20$		$d=200$	
	Cov (%)	Len ($\times 10^{-2}$)	Cov (%)	Len ($\times 10^{-2}$)	Cov (%)	Len ($\times 10^{-2}$)
Identity Σ, $n=10^5$						
delta	**94.84** (0.07)	1.24 (0.00)	**94.38** (0.07)	1.24 (0.00)	*71.37* (0.14)	1.24 (0.00)
BM	**92.72** (0.08)	1.23 (0.00)	**92.92** (0.08)	1.29 (0.00)	98.34 (0.04)	3.65 (0.00)
RS	**94.50** (0.07)	1.61 (0.01)	**93.92** (0.08)	1.62 (0.01)	**96.94** (0.05)	3.37 (0.02)
OB $B=10$	**92.36** (0.08)	1.26 (0.00)	**92.42** (0.08)	1.26 (0.00)	99.78 (0.05)	6.43 (0.01)
OB $B=100$	**94.88** (0.07)	1.29 (0.00)	**95.16** (0.07)	1.29 (0.00)	100.00 (0.00)	7.30 (0.01)
$\text{HiGrad}_{(2,2)}$	**95.67** (0.06)	2.56 (0.01)	**95.29** (0.07)	2.66 (0.01)	**96.03** (0.06)	2.67 (0.01)
COfB ASGD $B=3$	**95.44** (0.07)	2.44 (0.00)	**94.49** (0.07)	2.46 (0.00)	**92.31** (0.08)	20.71 (16.16)
COfB ASGD $B=5$	**95.24** (0.07)	1.68 (0.00)	**94.55** (0.07)	1.69 (0.00)	**94.59** (0.07)	3.06 (1.18)
COfB ASGD $B=10$	**95.04** (0.07)	1.40 (0.00)	**94.82** (0.07)	1.42 (0.00)	**94.30** (0.07)	2.56 (0.00)
COfB SGD $B=3$	**94.96** (0.07)	3.16 (0.00)	**94.53** (0.07)	3.18 (0.00)	**94.59** (0.07)	3.20 (0.00)
COfB SGD $B=5$	**94.44** (0.07)	2.17 (0.00)	**94.30** (0.07)	2.17 (0.00)	**94.42** (0.07)	2.19 (0.00)
COfB SGD $B=10$	**94.68** (0.07)	1.82 (0.00)	**93.97** (0.08)	1.83 (0.00)	**94.33** (0.07)	1.84 (0.00)
COnB $B=3$	**95.32** (0.07)	1.94 (0.00)	**94.63** (0.07)	1.97 (0.00)	99.17 (0.03)	9.15 (0.02)
COnB $B=5$	**95.04** (0.07)	1.61 (0.00)	**95.06** (0.07)	1.61 (0.00)	99.62 (0.02)	8.02 (0.01)
COnB $B=10$	**95.16** (0.07)	1.40 (0.00)	**95.52** (0.07)	1.43 (0.00)	99.88 (0.01)	7.31 (0.01)
Toeplitz Σ, $n=10^5$						
delta	**94.20** (0.07)	1.53 (0.00)	**93.23** (0.08)	1.58 (0.00)	*37.16* (0.15)	1.60 (0.00)
BM	91.80 (0.09)	1.53 (0.00)	88.51 (0.10)	1.51 (0.00)	**97.37** (0.05)	6.54 (0.01)
RS	**92.83** (0.08)	1.94 (0.01)	**93.38** (0.08)	2.15 (0.01)	**97.02** (0.05)	8.60 (0.06)
OB $B=10$	**92.48** (0.08)	1.66 (0.00)	**93.68** (0.08)	1.81 (0.00)	**93.33** (0.25)	13.73 (0.01)
OB $B=100$	**95.48** (0.07)	1.72 (0.00)	**95.84** (0.06)	1.97 (0.00)	**96.58** (0.18)	14.26 (0.01)
$\text{HiGrad}_{(2,2)}$	**94.33** (0.07)	2.69 (0.01)	**94.88** (0.07)	2.82 (0.01)	**94.09** (0.07)	2.83 (0.01)
COfB ASGD $B=3$	**94.72** (0.07)	3.06 (0.00)	**94.95** (0.07)	3.30 (0.00)	**94.41** (0.07)	14.44 (0.02)
COfB ASGD $B=5$	**95.24** (0.07)	2.21 (0.00)	**94.89** (0.07)	2.26 (0.00)	**93.91** (0.08)	9.96 (0.01)
COfB ASGD $B=10$	**95.28** (0.07)	1.74 (0.00)	**94.95** (0.07)	2.12 (0.00)	**94.01** (0.08)	8.44 (0.01)
COfB SGD $B=3$	**95.16** (0.07)	6.93 (0.01)	**94.76** (0.07)	4.34 (0.00)	**95.16** (0.07)	6.90 (0.01)
COfB SGD $B=5$	**95.84** (0.06)	4.70 (0.00)	**94.20** (0.07)	2.97 (0.00)	**95.04** (0.07)	4.72 (0.00)
COfB SGD $B=10$	**95.36** (0.07)	3.95 (0.00)	**94.20** (0.07)	2.50 (0.00)	**95.00** (0.07)	3.98 (0.00)
COnB $B=3$	**95.00** (0.07)	2.49 (0.00)	**95.36** (0.07)	2.94 (0.00)	**95.41** (0.07)	21.04 (0.02)
COnB $B=5$	**94.60** (0.07)	1.96 (0.00)	**95.43** (0.07)	2.44 (0.00)	**95.48** (0.07)	17.42 (0.02)
COnB $B=10$	**94.92** (0.07)	1.77 (0.00)	**95.42** (0.07)	2.18 (0.00)	**95.78** (0.06)	15.61 (0.01)
equicorrelation Σ, $n=10^5$						
delta	**94.76** (0.07)	1.31 (0.00)	**93.94** (0.08)	1.36 (0.00)	*25.95* (0.14)	1.38 (0.00)
BM	**93.20** (0.08)	1.30 (0.00)	**92.10** (0.09)	1.44 (0.00)	99.74 (0.02)	34.65 (0.22)
RS	**94.50** (0.07)	1.69 (0.01)	**94.42** (0.07)	1.82 (0.01)	98.61 (0.04)	13.90 (0.13)
OB $B=10$	**92.88** (0.08)	1.35 (0.00)	**92.78** (0.08)	1.45 (0.00)	89.23 (0.31)	19.42 (0.04)
OB $B=100$	**95.28** (0.07)	1.39 (0.00)	**94.73** (0.07)	1.52 (0.00)	**94.29** (0.23)	20.78 (0.03)
$\text{HiGrad}_{(2,2)}$	**95.00** (0.07)	2.59 (0.01)	**95.12** (0.07)	2.78 (0.01)	**95.93** (0.06)	2.83 (0.01)
COfB ASGD $B=3$	**94.76** (0.07)	2.58 (0.00)	**94.66** (0.07)	2.74 (0.00)	**93.23** (0.08)	32.56 (0.04)
COfB ASGD $B=5$	**95.52** (0.07)	1.77 (0.00)	**95.02** (0.07)	1.88 (0.00)	90.80 (0.09)	52.72 (0.29)
COfB ASGD $B=10$	**95.44** (0.07)	1.48 (0.00)	**94.95** (0.07)	1.58 (0.00)	91.87 (0.09)	49.84 (0.22)
COfB SGD $B=3$	**93.28** (0.08)	3.19 (0.00)	**93.85** (0.08)	3.26 (0.00)	**94.53** (0.07)	4.43 (0.00)
COfB SGD $B=5$	**93.16** (0.08)	2.18 (0.00)	**93.45** (0.08)	2.23 (0.00)	**94.20** (0.07)	3.06 (0.00)
COfB SGD $B=10$	**95.20** (0.07)	3.94 (0.00)	**94.78** (0.07)	3.94 (0.00)	**94.08** (0.07)	2.57 (0.00)
COnB $B=3$	**95.16** (0.07)	2.05 (0.00)	**95.00** (0.07)	2.28 (0.00)	**92.55** (0.08)	31.28 (0.16)
COnB $B=5$	**95.16** (0.07)	1.65 (0.00)	**94.60** (0.07)	1.89 (0.00)	**92.26** (0.08)	23.83 (0.04)
COnB $B=10$	**95.60** (0.06)	1.48 (0.00)	**95.07** (0.07)	1.68 (0.00)	**92.54** (0.08)	22.08 (0.04)

Table 9: Full results for the linear regression experiment.

	$d=5$		$d=20$		$d=200$	
	Cov (%)	Len ($\times 10^{-2}$)	Cov (%)	Len ($\times 10^{-2}$)	Cov (%)	Len ($\times 10^{-2}$)
Identity Σ, $n=10^5$						
delta	**95.00** (0.07)	3.10 (0.00)	**94.12** (0.07)	3.68 (0.00)	*61.92* (0.15)	5.85 (0.00)
BM	89.33 (0.10)	2.64 (0.00)	87.29 (0.11)	3.11 (0.01)	*57.47* (0.16)	5.56 (0.02)
RS	**94.17** (0.07)	3.84 (0.01)	**94.04** (0.07)	5.41 (0.02)	76.55 (0.13)	9.74 (0.04)
OB ($B=100$)	**95.00** (0.07)	3.21 (0.00)	**96.71** (0.06)	4.34 (0.01)	99.88 (0.01)	50.95 (0.23)
OB ($B=200$)	**94.83** (0.07)	3.18 (0.00)	**96.96** (0.05)	4.41 (0.01)	99.91 (0.01)	50.71 (0.23)
$\text{HiGrad}_{(2,2)}$	**94.33** (0.07)	5.77 (0.02)	**95.46** (0.07)	7.10 (0.02)	80.61 (0.13)	10.26 (0.04)
COfB ASGD $B=3$	**94.12** (0.07)	4.21 (0.00)	**95.03** (0.07)	7.32 (0.01)	**92.34** (0.08)	19.00 (0.02)
COfB ASGD $B=5$	**95.08** (0.07)	4.14 (0.00)	**94.77** (0.07)	5.01 (0.00)	90.64 (0.09)	13.03 (0.01)
COfB ASGD $B=10$	**94.88** (0.07)	3.52 (0.00)	**94.51** (0.07)	4.22 (0.00)	89.37 (0.10)	11.06 (0.01)
COfB SGD $B=3$	**95.40** (0.07)	9.38 (0.01)	**94.99** (0.07)	9.42 (0.01)	**94.72** (0.07)	25.61 (0.03)
COfB SGD $B=5$	**95.44** (0.07)	7.93 (0.00)	**94.70** (0.07)	7.93 (0.00)	**94.66** (0.07)	17.57 (0.01)
COfB SGD $B=10$	**95.44** (0.07)	7.93 (0.00)	**94.70** (0.07)	7.93 (0.00)	**94.54** (0.07)	14.84 (0.01)
COnB ($B=3$)	**94.33** (0.07)	4.71 (0.02)	**95.62** (0.06)	6.56 (0.03)	99.42 (0.02)	75.25 (0.43)
COnB ($B=5$)	**95.33** (0.07)	3.93 (0.01)	**96.79** (0.06)	5.35 (0.02)	99.50 (0.02)	64.17 (0.35)
COnB ($B=10$)	**94.83** (0.07)	3.55 (0.01)	**97.00** (0.05)	4.85 (0.01)	99.72 (0.02)	57.67 (0.30)
Toeplitz Σ, $n=10^5$						
delta	**94.83** (0.07)	4.05 (0.00)	**93.29** (0.08)	5.59 (0.00)	*53.69* (0.16)	9.56 (0.00)
BM	84.00 (0.12)	3.16 (0.01)	*75.25* (0.14)	3.75 (0.01)	*34.93* (0.15)	7.30 (0.03)
RS	**92.67** (0.08)	5.14 (0.02)	90.88 (0.09)	7.26 (0.03)	76.38 (0.13)	17.45 (0.10)
OB ($B=100$)	**95.00** (0.07)	4.22 (0.00)	**94.04** (0.07)	6.65 (0.01)	99.78 (0.01)	69.55 (0.26)
OB ($B=200$)	**95.00** (0.07)	4.24 (0.00)	**94.67** (0.07)	6.70 (0.01)	99.78 (0.01)	69.28 (0.26)
$\text{HiGrad}_{(2,2)}$	**95.33** (0.07)	7.18 (0.03)	**93.38** (0.08)	8.92 (0.03)	*57.02* (0.16)	10.27 (0.04)
COfB ASGD $B=3$	**94.12** (0.07)	5.70 (0.00)	**94.77** (0.07)	11.49 (0.01)	**93.79** (0.08)	42.27 (0.05)
COfB ASGD $B=5$	**94.32** (0.07)	4.82 (0.00)	**94.81** (0.07)	7.97 (0.01)	**93.23** (0.08)	28.81 (0.02)
COfB ASGD $B=10$	**94.88** (0.07)	4.61 (0.00)	**94.65** (0.07)	6.69 (0.00)	**93.09** (0.08)	24.43 (0.01)
COfB SGD $B=3$	**95.36** (0.07)	9.48 (0.01)	**94.80** (0.07)	9.65 (0.01)	**94.41** (0.07)	30.53 (0.03)
COfB SGD $B=5$	**95.40** (0.07)	7.98 (0.00)	**94.37** (0.07)	8.16 (0.00)	**93.79** (0.08)	20.76 (0.02)
COfB SGD $B=10$	**95.40** (0.07)	7.98 (0.00)	**94.37** (0.07)	8.16 (0.00)	**93.62** (0.08)	17.56 (0.01)
COnB ($B=3$)	**94.00** (0.08)	6.22 (0.02)	**94.71** (0.07)	10.27 (0.05)	**97.82** (0.05)	99.61 (0.60)
COnB ($B=5$)	**95.00** (0.07)	5.25 (0.02)	**94.71** (0.07)	8.20 (0.03)	98.75 (0.04)	85.15 (0.45)
COnB ($B=10$)	**94.83** (0.07)	4.84 (0.01)	**94.29** (0.07)	7.25 (0.02)	99.31 (0.03)	78.02 (0.37)
equicorrelation Σ, $n=10^5$						
delta	**94.83** (0.07)	3.39 (0.00)	**94.12** (0.07)	5.19 (0.00)	*31.55* (0.15)	13.50 (0.01)
BM	90.00 (0.09)	2.92 (0.00)	81.25 (0.12)	3.84 (0.01)	*14.67* (0.11)	7.17 (0.03)
RS	**94.50** (0.07)	4.41 (0.02)	91.08 (0.09)	6.31 (0.02)	74.79 (0.14)	32.13 (0.20)
OB ($B=100$)	**95.00** (0.07)	3.50 (0.00)	**95.46** (0.07)	6.37 (0.02)	98.12 (0.04)	110.74 (0.42)
OB ($B=200$)	**94.67** (0.07)	3.50 (0.00)	**95.92** (0.06)	6.42 (0.02)	98.38 (0.04)	109.89 (0.40)
$\text{HiGrad}_{(2,2)}$	**96.17** (0.06)	5.97 (0.02)	**95.00** (0.07)	9.11 (0.03)	*26.94* (0.14)	11.30 (0.04)
COfB ASGD $B=3$	**94.12** (0.07)	4.68 (0.00)	**94.87** (0.07)	10.42 (0.01)	83.65 (0.12)	64.61 (0.08)
COfB ASGD $B=5$	**93.96** (0.08)	3.92 (0.00)	**94.96** (0.07)	7.11 (0.01)	*76.52* (0.13)	44.29 (0.04)
COfB ASGD $B=10$	**95.12** (0.07)	3.85 (0.00)	**94.78** (0.07)	6.04 (0.00)	*71.44* (0.14)	37.42 (0.02)
COfB SGD $B=3$	**94.96** (0.07)	9.39 (0.01)	**93.85** (0.08)	9.50 (0.01)	*77.71* (0.13)	39.47 (0.05)
COfB SGD $B=5$	**94.76** (0.07)	7.92 (0.00)	**93.59** (0.08)	8.06 (0.00)	*66.67* (0.15)	26.98 (0.03)
COfB SGD $B=10$	**94.76** (0.07)	7.92 (0.00)	**93.59** (0.08)	8.06 (0.00)	*58.94* (0.16)	22.78 (0.02)
COnB ($B=3$)	**94.17** (0.07)	5.18 (0.02)	**95.96** (0.06)	9.51 (0.06)	**93.07** (0.08)	138.99 (1.13)
COnB ($B=5$)	**94.67** (0.07)	4.38 (0.01)	**95.12** (0.07)	7.66 (0.03)	**94.14** (0.07)	115.73 (0.73)
COnB ($B=10$)	**95.00** (0.07)	3.94 (0.01)	**95.83** (0.06)	6.92 (0.02)	**95.90** (0.06)	111.29 (0.60)

Table 10: Full results for the logistic regression experiment.

Identity Σ, $n = 100$					
		$p = 3$		$p = 15$	
		Cov (%)	Len ($\times 10^{-2}$)	Cov (%)	Len($\times 10^{-2}$)
COfB ASGD $B = 3$	$\in T^*$	**94.87** (22.07)	1.52 (0.00)	**94.92** (21.96)	5.57 (0.01)
	$\notin T^*$	**99.87** (3.62)	0.03 (0.00)	**99.76** (4.92)	0.38 (0.00)
COfB ASGD $B = 10$	$\in T^*$	**95.13** (21.52)	0.82 (0.00)	**95.44** (20.86)	3.74 (0.01)
	$\notin T^*$	**99.85** (3.91)	0.02 (0.00)	**99.77** (4.81)	0.26 (0.00)
COnB ASGD $B = 3$	$\in T^*$	**95.40** (20.95)	1.15 (0.00)	**96.07** (19.44)	5.13 (0.01)
	$\notin T^*$	**99.92** (2.83)	0.02 (0.00)	**99.93** (2.69)	0.33 (0.00)
COnB ASGD $B = 10$	$\in T^*$	**95.40** (20.95)	0.92 (0.00)	**96.96** (17.17)	3.99 (0.01)
	$\notin T^*$	**99.93** (2.56)	0.02 (0.00)	**99.97** (1.71)	0.27 (0.00)
Toeplitz Σ, $n = 100$					
COfB ASGD $B = 3$	$\in T^*$	**95.20** (21.38)	2.75 (0.01)	**95.12** (21.54)	34.87 (0.11)
	$\notin T^*$	**99.88** (3.42)	0.05 (0.00)	**99.09** (9.52)	8.62 (0.07)
COfB ASGD $B = 10$	$\in T^*$	**95.07** (21.66)	0.86 (0.00)	**93.13** (25.29)	4.33 (0.01)
	$\notin T^*$	**99.86** (3.74)	0.02 (0.00)	**99.12** (9.32)	1.00 (0.01)
COnB ASGD $B = 3$	$\in T^*$	**94.73** (22.34)	1.28 (0.01)	**97.19** (16.54)	10.68 (0.03)
	$\notin T^*$	**99.93** (2.73)	0.03 (0.00)	**99.80** (4.47)	2.47 (0.02)
COnB ASGD $B = 10$	$\in T^*$	**95.40** (20.95)	1.07 (0.00)	98.85 (10.65)	10.27 (0.04)
	$\notin T^*$	**99.96** (2.00)	0.02 (0.00)	**99.98** (1.51)	2.41 (0.02)
equicorrelation Σ, $n = 100$					
COfB ASGD $B = 3$	$\in T^*$	**94.53** (22.73)	3.36 (0.02)	**95.08** (21.63)	1.94 (21.00)
	$\notin T^*$	**99.89** (3.32)	0.08 (0.00)	98.69 (11.37)	9.35 (20.00)
COfB ASGD $B = 10$	$\in T^*$	**94.87** (22.07)	1.23 (0.01)	**94.81** (22.18)	33.65 (0.30)
	$\notin T^*$	**99.90** (3.23)	0.03 (0.00)	98.74 (11.16)	10.39 (0.17)
COnB ASGD $B = 3$	$\in T^*$	**93.60** (24.48)	1.47 (0.01)	**95.95** (19.72)	20.26 (0.12)
	$\notin T^*$	**99.93** (2.55)	0.03 (0.00)	**99.80** (4.42)	5.52 (0.07)
COnB ASGD $B = 10$	$\in T^*$	**93.93** (23.87)	1.15 (0.00)	**95.55** (20.63)	15.59 (0.17)
	$\notin T^*$	**99.97** (1.82)	0.03 (0.00)	**99.96** (2.00)	4.73 (0.10)

Table 11: Results for sparse linear regression.

References

Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A. Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In *Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 115–137. PMLR, 2019.

Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

Jerry Chee, Hwanwoo Kim, and Panos Toulis. "Plus/minus the learning rate": Easy and scalable statistical inference with SGD. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2285–2309. PMLR, 2023.

Jessie X.T. Chen and Miles Lopes. Estimating the error of randomized Newton methods: A bootstrap approach. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1649–1659. PMLR, 2020.

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

Kai Lai Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3):463–483, 1954.

Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19(78):1–21, 2018.

James M. Flegal and Galin L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034–1070, 2010.

Charles J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.

Peter W. Glynn and Donald L. Iglehart. Simulation output analysis using standardized time series. *Mathematics of Operations Research*, 15(1):1–16, 1990.

Peter W. Glynn and Henry Lam. Constructing simulation output intervals under input uncertainty via data sectioning. In *Proceedings of the 2018 Winter Simulation Conference*, pages 1551–1562. Institute of Electrical and Electronics Engineers, Inc., 2018.

Steven R. Howard, Aaditya Ramdas, Jon D. McAuliffe, and Jasjeet S. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(1):266–299, 2021.

Galin L. Jones, Murali Haran, Brian S. Caffo, and Ronald Neath. Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101(476):1537–1547, 2006.

Henry Lam. A cheap bootstrap method for fast inference. *arXiv preprint arXiv:2202.00090*, 2022a.

Henry Lam. Cheap bootstrap for input uncertainty quantification. In *Proceedings of the 2022 Winter Simulation Conference*, pages 2318–2329. Institute of Electrical and Electronics Engineers, Inc., 2022b.

Henry Lam and Zhenyuan Liu. Bootstrap in high dimension with low computation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18419–18453. PMLR, 2023.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020.

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7381–7389, 2022.

Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis. Statistical inference using SGD. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1): 3571–3578, 2018.

Miles E. Lopes. Estimating the algorithmic variance of randomized ensembles via the bootstrap. *The Annals of Statistics*, 47(2):1088–1112, 2019.

Miles E. Lopes, Suofei Wu, and Thomas C.M. Lee. Measuring the algorithmic convergence of randomized ensembles: The regression setting. *SIAM Journal on Mathematics of Data Science*, 2(4):921–943, 2020.

Robert Lunde, Purnamrita Sarkar, and Rachel Ward. Bootstrapping the error of Oja's algorithm. In *Advances in Neural Information Processing Systems*, volume 34, pages 6240–6252, 2021.

Yurii Nesterov and Jean-Philippe Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6):1559–1568, 2008.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.

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

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1571–1578, Edinburgh, Scotland, 2012. Omnipress.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

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

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Jerome Sacks. Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29(2):373–405, 1958.

Bruce Schmeiser. Batch size effects in the analysis of simulation output. *Operations Research*, 30(3):556–568, 1982.

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

Qi-Man Shao and Zhuo-Song Zhang. Berry–Esseen bounds for multivariate nonlinear statistics with applications to M-estimators and stochastic gradient descent algorithms. *Bernoulli*, 28(3):1548–1576, 2022.

Weijie J. Su and Yuancheng Zhu. HiGrad: Uncertainty quantification for online learning and stochastic approximation. *Journal of Machine Learning Research*, 24(124):1–53, 2023.

Aad W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, Cambridge, UK, 2000.

Chuhan Xie, Kaicheng Jin, Jiadong Liang, and Zhihua Zhang. Asymptotic time-uniform inference for parameters in averaged stochastic approximation. *arXiv preprint arXiv:2410.15057*, 2024.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Yi Zhu and Jing Dong. On constructing confidence region for model parameters in stochastic gradient descent via batch means. In *Proceedings of the 2021 Winter Simulation Conference*, pages 1–12. Institute of Electrical and Electronics Engineers, Inc., 2021.