Online bootstrap confidence intervals for the stochastic gradient descent estimator

Fang, Yixin; Xu, Jinfeng; Yang, Lei

Published in:
Journal of Machine Learning Research

Published: 01/12/2018

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Publication details:
Fang, Y., Xu, J., & Yang, L. (2018). Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, *19*. http://jmlr.org/papers/v19/17-370.html

Journal of Machine Learning Research 19 (2018) 1-21 Submitted 6/17; Published 12/18

Online Bootstrap Confidence Intervals for the Stochastic Gradient Descent Estimator

Yixin Fang YIXIN.FANG@NJIT.EDU
Department of Mathematical Sciences
New Jersey Institute of Technology

Jinfeng Xu XUJF@HKU.HK
Department of Statistics and Actuarial Science
The University of Hong Kong

Lei Yang LY888@NYU.EDU
Department of Population Health
New York University School of Medicine

Editor: Gabor Lugosi

Abstract

In many applications involving large dataset or online learning, stochastic gradient descent (SGD) is a scalable algorithm to compute parameter estimates and has gained increasing popularity due to its numerical convenience and memory efficiency. While the asymptotic properties of SGD-based estimators have been well established, statistical inference such as interval estimation remains much unexplored. The classical bootstrap is not directly applicable if the data are not stored in memory. The plug-in method is not applicable when there is no explicit formula for the covariance matrix of the estimator. In this paper, we propose an online bootstrap procedure for the estimation of confidence intervals, which, upon the arrival of each observation, updates the SGD estimate as well as a number of randomly perturbed SGD estimates. The proposed method is easy to implement in practice. We establish its theoretical properties for a general class of models that includes linear regressions, generalized linear models, M-estimators and quantile regressions as special cases. The finite-sample performance and numerical utility is evaluated by simulation studies and real data applications.

Keywords: Bootstrap, Interval estimation, Generalized linear models, Large datasets, M-estimators, Quantile regression, Resampling methods, Stochastic gradient descent

1. Introduction

Big datasets arise frequently in clinical, epidemiological, financial and sociological studies. In such applications, many classical optimization methods for parameter estimation such as Fisher scoring, the EM algorithm or iterated reweighted least squares (Hastie et al. 2009, Nelder and Baker 1972) do not scale well and are computationally less attractive. Due to its computational and memory efficiency, stochastic gradient descent (Robbins and Monro 1951)[SGD] is a scalable algorithm for parameter estimation and has recently drawn a great deal of attention. Unlike other classical methods that evaluate the objective function involving the entire dataset, the SGD method calculates the gradient of the objective function using only one data point at a time and recursively updates the parameter estimate. This

is also numerically appealing and particularly useful in online updating settings such as streaming data where it may not even be feasible to store the entire dataset in memory. Wang et al. (2016) gave a nice review on recent achievements of applying the SGD method to big data and streaming data.

The asymptotic properties of SGD estimators such as consistency and asymptotic normality have been well established; see, for example, Ruppert (1988) and Polyak and Juditsky (1992). However, statistical inference such as confidence interval estimation for SGD estimators has remained largely unexplored. Traditional interval estimation procedures such as the plug-in procedure and the bootstrap are often numerically difficult in the presence of big datasets. The bootstrap repeatedly draws samples from the entire dataset. The plug-in procedure requires an explicit variance-covariance formula. Since the classical bootstrap is not directly applicable if the data are not stored in memory, using the deal from the weighted bootstrap (Rubin 1981), we propose an online bootstrap procedure for the estimation of confidence intervals.

There are only a few papers considering the statistical inference of the SGD method. Chen et al. (2016) proposed a method called the batch-mean procedure. Although computationally efficient and theoretically sound, the batch-means procedure substantially underestimates the variance of the SGD estimator in finite-sample studies, because of the correlations between the batch means. Li et al. (2017) presented a new method for statistical inference in M-estimation problems, based on SGD estimators with a fixed step size. However, this method is limited to M-estimation and fixed step size. Su and Zhu (2018) proposed a new method called HiGrad, short for Hierarchical Incremental GRAdient Descent, which estimates model parameters in an online fashion and provides a confidence interval for the true population value. This method is also computationally efficient and theoretically sound, but it is not applicable to vanilla SGD estimators.

In this paper, we propose an online bootstrap resampling procedure to approximate the distribution of a SGD estimator in a general class of models that includes linear regressions, generalized linear models, M-estimators and quantile regressions as special cases. Our proposal, justified by asymptotic theories, provides a simple way to estimate the covariance matrix and confidence regions. Through numerical experiments, we verify the ability of this procedure to give accurate inference for big datasets.

The rest of the article is organized as follows. In Section 2, we introduce the proposed online bootstrap procedure for constructing confidence regions. In Section 3, we theoretically justify the validity of our proposal for a general class of models, along with some special cases. In Section 4, we demonstrate the performance of the proposed procedures in finite samples via simulation studies and three real data applications. Some concluding remarks are given in Section 5 and all the technical proofs are relegated to the Appendix.

2. The proposed resampling procedure

Parameter estimation by optimizing an objective function is often encountered in statistical practice. Consider the general situation where the optimal model parameter $\theta_0 \in \mathcal{R}^p$ is defined to be the minimizer of the expected loss function,

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ L(\theta) \triangleq \mathbb{E}[\ell(\theta; Z)] \right\}, \tag{1}$$

where $\ell(\theta; z)$ is some loss function and Z denotes one single observation and Θ is the domain on which the loss function is defined, which is assumed to be open. Suppose that the data consist of independent and identically distributed (i.i.d.) copies of Z, denoted by $\mathcal{D}_N = \{Z_1, \ldots, Z_N\}$. Under mild conditions, θ_0 can be consistently estimated by

$$\widetilde{\theta}_N = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell(\theta; Z_i) \right\}. \tag{2}$$

However, the minimization problem (2) for large-scale datasets pose numerical challenges. Furthermore, for applications such as online data where each sample arrives sequentially (e.g., search queries or transactional data), it may not be necessary or feasible to store the entire dataset, leaving alone evaluating the minimand in (2).

As a stochastic approximation method (Robbins and Monro 1951), stochastic gradient descent is a scalable algorithm for parameter estimation with large-scale data. Given an initial estimate $\widehat{\theta}_0$, the SGD method recursively updates the estimate upon the arrival of each data point Z_n, $n = 1, 2, \ldots, N$,

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} - \gamma_n \nabla \ell(\widehat{\theta}_{n-1}; Z_n), \tag{3}$$

where the learning rates are $\gamma_n = \gamma_1 n^{-\alpha}$ with $\gamma_1 > 0$ and $\alpha \in (0.5, 1)$. As suggested by Ruppert (1988) and Polyak and Juditsky (1992), we consider the averaging estimate,

$$\overline{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_i, \tag{4}$$

which can also be recursively updated given that $\overline{\theta}_n = (n-1)\overline{\theta}_{n-1}/n + \widehat{\theta}_n/n$.

In order to conduct statistical inference with the averaging SGD estimator $\overline{\theta}_n$ at any stage, we propose an online bootstrap resampling procedure, which recursively updates the SGD estimate as well as a large number of randomly perturbed SGD estimates, upon the arrival of each data point. Specifically, let $\mathcal{W} = \{W_i, i = 1, \ldots, N\}$ be a set of i.i.d. non-negative random variables with mean and variance equal to one. In parallel with (3) and (4), with $\widehat{\theta}_0^* \equiv \widehat{\theta}_0$, upon observing data point Z_n, we recursively updates randomly perturbed SGD estimates,

$$\widehat{\theta}_n^* = \widehat{\theta}_{n-1}^* - \gamma_n W_n \nabla \ell(\widehat{\theta}_{n-1}^*; Z_n), \tag{5}$$

$$\overline{\theta}_n^* = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_i^*. \tag{6}$$

We will show that $\sqrt{n}(\overline{\theta}_n - \theta_0)$ and $\sqrt{n}(\overline{\theta}_n^* - \overline{\theta}_n)$ converge in distribution to the same limiting distribution. In practice, these results allow us to estimate the distribution of $\sqrt{n}(\overline{\theta}_n - \theta_0)$ by generating a large number, say B, of random samples of $\mathcal{W}$. We obtain $\overline{\theta}_n^{*,b}$ by sequentially updating perturbed SGD estimates for each sample, $b = 1, \ldots, B$,

$$\widehat{\theta}_n^{*,b} = \widehat{\theta}_{n-1}^{*,b} - \gamma_n W_{n,b} \nabla \ell(\widehat{\theta}_{n-1}^{*,b}; Z_n), \tag{7}$$

$$\overline{\theta}_n^{*,b} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_i^{*,b}, \tag{8}$$

and then approximate the sampling distribution of $\overline{\theta}_n - \theta_0$ using the empirical distribution of $\{\overline{\theta}_n^{*,b} - \overline{\theta}_n, b = 1, ..., B\}$. Specifically, the covariance matrix of $\overline{\theta}_n$ can be estimated by the sample covariance matrix constructed from $\{\overline{\theta}_n^{*,b}, b = 1, ..., B\}$. Estimating the distribution of $\sqrt{n}(\overline{\theta}_n - \theta_0)$ based on the distribution of $\sqrt{n}(\overline{\theta}_n^* - \overline{\theta}_n)|\mathcal{D}_n$ leads to the construction of $(1-\alpha)100\%$ confidence regions for θ_0. The resulting inferential procedure retains the numerical simplicity of the SGD method, only using one pass over the data. The proposed inferential procedure scales well for datasets with millions of data points or more, and its theoretical validity can be justified for a general class models with mild regularity conditions as shown in the next section.

3. Theoretical Results

3.1. Main theorems

In this section, we derive some theoretical properties of $\overline{\theta}_n^*$, justifying that the conditional distribution of $\overline{\theta}_n^* - \overline{\theta}_n$ given data $\mathcal{D}_n = \{Z_1, Z_2, \ldots, Z_n\}$ can approximate the sampling distribution of $\overline{\theta}_n - \theta_0$, under the following assumptions. Let $\|\cdot\|$ be the Euclidean norm for vectors and the operator norm for matrices. The proofs are presented in the Appendix.

(A1). The objective function $L(\theta)$ is convex, continuously differentiable over $\theta \in \Theta$, and twice continuously differentiable at $\theta = \theta_0$, where θ_0 is the unique minimizer of $L(\theta)$.

(A2). The gradient of $L(\theta)$, $R(\theta) = \nabla L(\theta)$, is Lipschitz continuous with constant $L_1 > 0$; that is, for any θ_1 and θ_2, $\|R(\theta_1) - R(\theta_2)\| \leq L_1\|\theta_1 - \theta_2\|$.

(A3). The Hessian matrix of $L(\theta)$, $S(\theta) = \nabla^2 L(\theta)$, exists and is positive definite at θ_0 with $S_0 = S(\theta_0) > 0$ and is Lipschitz continuous at θ_0 with constant $L_2 > 0$.

(A4). Let $V_0 = \mathbb{E}\{[\nabla \ell(\theta_0; Z)][\nabla \ell(\theta_0; Z)]^{\mathrm{T}}\}$. Assume $\mathbb{E}\|\nabla \ell(\theta; Z)\|^2 \leq C(1 + \|\theta\|^2)$ for some C and $\mathbb{E}\|\nabla \ell(\theta; Z) - \nabla \ell(\theta_0; Z)\|^2 \leq \delta(\|\theta - \theta_0\|)$ for some $\delta(\cdot)$ with $\delta(x) \to 0$ as $x \to 0$.

(A5). The learning rates are chosen as $\gamma_n = \gamma_1 n^{-\alpha}$ with $\gamma_1 > 0$ and $\alpha \in (0.5, 1)$.

(A6). The perturbation variables, $W_1, W_2, \ldots$, are non-negative i.i.d. random variables satisfying that $\mathbb{E}(W_n) = \mathrm{Var}(W_n) = 1$.

Following similar arguments in Ruppert (1988) and Polyak and Juditsky (1992), we can prove the asymptotic normality of the SGD estimator $\overline{\theta}_n$ under the above assumptions.

Lemma 1 *If Assumptions A1-A5 are satisfied, then we have*

$$\sqrt{n}(\overline{\theta}_n - \theta_0) \Rightarrow \mathcal{N}\left(0, S_0^{-1} V_0 S_0^{-1}\right), \text{ in distribution, as } n \to \infty. \tag{9}$$

From Lemma 1, we can conduct statistical inference based on $\overline{\theta}_n$ provided that we can estimate the covariance matrix $S_0^{-1} V_0 S_0^{-1}$, or we can use some resampling procedure to approximate the sampling distribution of $\sqrt{n}(\overline{\theta}_n - \theta_0)$. We first derive the asymptotically linear representation of $\overline{\theta}_n^*$ for any perturbation variables that are i.i.d. random variables satisfying that $\mathbb{E}(W_n) = 1$.

Theorem 2 *If Assumptions A1-A5 hold, and the perturbation variables, $W_1, W_2, \ldots$, are non-negative i.i.d. random variables satisfying that $\mathbb{E}(W_n) = 1$, then we have,*

$$\sqrt{n}(\overline{\theta}_n^* - \theta_0) = -\frac{1}{\sqrt{n}} S_0^{-1} \sum_{i=1}^{n} W_i \nabla \ell(\theta_0; Z_i) + o_p(1). \tag{10}$$

By Theorem 1, letting $W_n \equiv 1$, we derive the following representation for $\overline{\theta}_n$,

$$\sqrt{n}(\overline{\theta}_n - \theta_0) = -\frac{1}{\sqrt{n}} S_0^{-1} \sum_{i=1}^{n} \nabla \ell(\theta_0; Z_i) + o_p(1). \tag{11}$$

Then, considering the difference between (2) and (11), we have

$$\sqrt{n}(\overline{\theta}_n^* - \theta_n) = -\frac{1}{\sqrt{n}} S_0^{-1} \sum_{i=1}^{n} (W_i - 1) \nabla \ell(\theta_0; Z_i) + o_p(1). \tag{12}$$

Let $\mathbb{P}^*$ and $\mathbb{E}^*$ denote the conditional probability and expectation given the data. Starting from (12), we derive the following theorem.

Theorem 3 *If Assumptions A1-A6 hold, then we have*

$$\sup_{v \in \mathcal{R}^p} \left| \mathbb{P}^* \left(\sqrt{n}(\overline{\theta}_n^* - \overline{\theta}_n) \le v \right) - \mathbb{P}\Big(\sqrt{n}(\overline{\theta}_n - \theta_0) \le v \Big) \right| \to 0, \textit{in probability}. \tag{13}$$

By Theorem 2, the Kolmogorow-Smirnov distance between $\sqrt{n}(\overline{\theta}_n^* - \overline{\theta}_n)$ and $\sqrt{n}(\overline{\theta}_n - \theta_0)$ converges to zero in probability. This validates our proposal of the perturbation-based resampling procedure for inference with SGD. In the next section, we consider some special cases where Assumptions A1-A4 are satisfied.

3.2. Special cases

3.2.1. Cases where $\ell(\theta; Z)$ is twice differentiable

If the loss function $\ell(\theta; Z)$ is twice differentiable, we can use the plug-in procedure to estimate the asymptotic covariance matrix of $\sqrt{N}(\overline{\theta}_N - \theta_0)$, $S_0^{-1} V_0 S_0^{-1}$. That is, at the final step, S_0 and V_0 can be estimated respectively by

$$\widehat{S}_N = \frac{1}{N} \sum_{i=1}^{N} \nabla^2 \ell(\overline{\theta}_N; Z_i) \;\text{ and }\; \widehat{V}_N = \frac{1}{N} \sum_{i=1}^{N} [\nabla \ell(\overline{\theta}_N; Z_i)][\nabla \ell(\overline{\theta}_N; Z_i)]^{\mathrm{T}}. \tag{14}$$

However, the above final-step plug-in estimation is impractical for large-scale data or streaming data, because it requires that the whole dataset be stored. To overcome this problem, in practice we can estimate S_0 and V_0 recursively, for $n = 1, 2, \ldots$, using

$$\widehat{S}_n = \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell(\overline{\theta}_i; Z_i) \;\text{ and }\; \widehat{V}_n = \frac{1}{n} \sum_{i=1}^{n} [\nabla \ell(\overline{\theta}_i; Z_i)][\nabla \ell(\overline{\theta}_i; Z_i)]^{\mathrm{T}}. \tag{15}$$

In the following, we examine two examples where $\ell(\theta; Z)$ is twice differentiable. Example 1 is linear regression, where the loss function $\ell(\theta; Z)$ is twice differentiable and the objective

function $L(\theta)$ is strongly convex. Example 2 is logistic regression, where $\ell(\theta; Z)$ is twice differentiable but $L(\theta)$ is non-strongly convex. They are two examples of generalized linear models, one for quantitative outcome and the other for binary outcome. In these two examples, both the plug-in procedures and the proposed perturbation resampling procedure are robust to model mis-specification.

Example 1 (Linear regression) Suppose that $Z_n = (Y_n, X_n)$, $n = 1, 2, \ldots$, are i.i.d. copies of $Z = (Y, X)$, where Y is quantitative and X is p-dim with $\mathbb{E}\|X\|^2 < \infty$. Let

$$\theta_0 = \arg\min_{\theta \in \mathcal{R}^p} \mathbb{E}\left(Y - X^{\mathrm{T}}\theta\right)^2, \tag{16}$$

where $\ell(\theta; Z) = (Y - X^{\mathrm{T}}\theta)^2$ is twice differentiable and $L(\theta) = \mathbb{E}\left(Y - X^{\mathrm{T}}\theta\right)^2$ is strongly convex. Moveover, $\nabla\ell(\theta; Z) = -2(Y - X^{\mathrm{T}}\theta)X$, $\nabla^2\ell(\theta; Z) = 2X^{\mathrm{T}}X$, $\nabla L(\theta) = 2\mathbb{E}\{XX^{\mathrm{T}}\}\theta - 2\mathbb{E}\{XY\}$, and $\nabla^2 L(\theta) = \mathbb{E}\{\nabla l(\theta; Z)\} = 2\mathbb{E}\{XX^{\mathrm{T}}\}$. Letting $V_0 = 4\mathbb{E}\{(Y - X^{\mathrm{T}}\theta_0)^2 XX^{\mathrm{T}}\}$ and $S_0 = 2\mathbb{E}\{XX^{\mathrm{T}}\}$, we can easily verify that Assumptions A1-A4 hold. The SGD and perturbed SGD updates for θ_0, as defined in (3) and (5) respectively, are

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} + 2\gamma_n(Y_n - X_n^{\mathrm{T}}\widehat{\theta}_{n-1})X_n, \tag{17}$$

$$\widehat{\theta}^*_n = \widehat{\theta}^*_{n-1} + 2\gamma_n W_n(Y_n - X_n^{\mathrm{T}}\widehat{\theta}^*_{n-1})X_n. \tag{18}$$

Example 2 (Logistic regression) Suppose that $Z_n = (Y_n, X_n)$, $n = 1, 2, \ldots$, are i.i.d. copies of $Z = (Y, X)$, where $Y = \pm 1$ and X is p-dim with $\mathbb{E}\|X\|^2 < \infty$. Let

$$\theta_0 = \arg\min_{\theta \in \mathcal{R}^p} \mathbb{E}\left\{-\log\left(\frac{1}{1 + \exp(-YX^{\mathrm{T}}\theta)}\right)\right\}, \tag{19}$$

where $\ell(\theta; Z) = \log\left(1 + \exp(-YX^{\mathrm{T}}\theta)\right)$ is twice differentiable and $L(\theta) = \mathbb{E}\{\ell(\theta; Z)\}$ is non-strongly convex. Moveover, $\nabla\ell(\theta; Z) = -\frac{1}{1+\exp(YX^{\mathrm{T}}\theta)}XY$, $\nabla^2\ell(\theta; Z) = \frac{\exp(YX^{\mathrm{T}}\theta)}{[1+\exp(YX^{\mathrm{T}}\theta)]^2}XX^{\mathrm{T}}$, $\nabla L(\theta) = \mathbb{E}\left\{\nabla\ell(\theta; Z)\right\}$, and $\nabla^2 L(\theta) = \mathbb{E}\left\{\nabla^2\ell(\theta; Z)\right\}$. Letting $V_0 = \mathbb{E}\left\{\frac{1}{[1+\exp(YX^{\mathrm{T}}\theta)]^2}XX^{\mathrm{T}}\right\}$ and $S_0 = \mathbb{E}\left\{\frac{\exp(YX^{\mathrm{T}}\theta)}{[1+\exp(YX^{\mathrm{T}}\theta)]^2}XX^{\mathrm{T}}\right\}$, we can easily verify that Assumptions A1-A4 hold. The SGD and perturbed SGD updates for θ_0, as defined in (3) and (5) respectively, are

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} + \gamma_n XY/[1 + \exp(YX^{\mathrm{T}}\widehat{\theta}_{n-1})], \tag{20}$$

$$\widehat{\theta}^*_n = \widehat{\theta}^*_{n-1} + \gamma_n W_n XY/[1 + \exp(YX^{\mathrm{T}}\widehat{\theta}^*_{n-1})]. \tag{21}$$

We conclude this subsection with some discussion on the strong convexity of objective function $L(\theta)$, which is strongly convex in Example 1 and is non-strongly convex in Example 2. If $L(\theta)$ is strongly convex, i.e. there exists $\mu > 0$ such that $L(\theta_1) \geq L(\theta_2) + \nabla L(\theta_2)^{\mathrm{T}}(\theta_1 - \theta_2) + \mu\|\theta_1 - \theta_2\|^2$ for any θ_1 and θ_2, Moulines and Bach (2011) derived a non-asymptotic bound for $(\mathbb{E}\|\overline{\theta}_n - \theta_0\|)^{1/2}$. The bound of $(\mathbb{E}\|\overline{\theta}_n - \theta_0\|)^{1/2}$ has several terms; the leading term is of order $O(n^{-1})$ and the next two leading terms have order $O(n^{\alpha-2})$ and $O(n^{-2\alpha})$, suggesting the setting $\alpha = 2/3$ to make them equal. If $L(\theta)$ is non-strongly convex, Moulines and Bach (2011) derived a non-asymptotic bound for $\mathbb{E}[L(\widehat{\theta}_n) - L(\theta_0)]$ and a non-asymptotic bound for $\mathbb{E}[L(\overline{\theta}_n) - L(\theta_0)]$. The bound of $\mathbb{E}[L(\widehat{\theta}_n) - L(\theta_0)]$ is $O(\max\{n^{\alpha-1}, n^{-\alpha/2}\})$, also suggesting the setting $\alpha = 2/3$ to achieve optimal rate $O(n^{-1/3})$. Using the Polyak-Ruppert averaging has allowed the bound of $\mathbb{E}[L(\overline{\theta}_n) - L(\theta_0)]$ to go from $O(\max\{n^{\alpha-1}, n^{-\alpha/2}\})$ to $O(n^{-\alpha})$. Therefore, we use $\alpha = 2/3$ in the numerical results.

3.2.2. Cases where $\ell(\theta; Z)$ is not twice-differentiable

If the loss function $\ell(\theta; Z)$ is not twice-differentiable, neither the final-step plug-in estimation (14) nor the recursive plug-in estimation (15) is applicable. Fortunately, our proposal of scalable inference based on perturbation resampling is still applicable because it only depends on the first order derivative $\nabla\ell(\theta; Z)$. To understand this explicitly, consider the following example of robust regression via ψ-type M-estimator where the loss function may be not twice-differentiable.

Example 3 (Robust regression via ψ-type M-estimator) Suppose that $Z_n = (Y_n, X_n)$, $n = 1, 2, \ldots$, are i.i.d. copies of $Z = (Y, X)$, where Y is quantitative and X is p-dim with $\mathbb{E}\|X\|^2 < \infty$. Let $\rho(\cdot)$ be some convex function with $\rho(0) = 0$ and we attempt to estimate

$$\theta_0 = \arg\min_{\theta\in\mathcal{R}^p} \mathbb{E}\rho\left(Y - X^{\mathrm{T}}\theta\right), \tag{22}$$

where $\ell(\theta; Z) = \rho(Y - X^{\mathrm{T}}\theta)$ and $L(\theta) = \mathbb{E}\rho\left(Y - X^{\mathrm{T}}\theta\right)$. This is robust regression via ρ-type M-estimator. If $\rho(\cdot)$ is differentiable with derivative $\dot\rho(\cdot) = \psi(\cdot)$, we can solve it via ψ-type M-estimator, solving the following equation,

$$\mathbb{E}\left\{\psi\left(Y - X^{\mathrm{T}}\theta_0\right)X\right\} = 0, \tag{23}$$

where $\nabla\ell(\theta; Z) = -\psi(Y - X^{\mathrm{T}}\theta)X$ and $\nabla L(\theta) = -\mathbb{E}\left\{\psi\left(Y - X^{\mathrm{T}}\theta\right)X\right\}$. Hence, the SGD and perturbed SGD updates for θ_0, as defined in (3) and (5) respectively, are

$$\widehat\theta_n = \widehat\theta_{n-1} + \gamma_n\psi(Y_n - X_n^{\mathrm{T}}\widehat\theta_{n-1})X_n, \tag{24}$$
$$\widehat\theta_n^* = \widehat\theta_{n-1}^* + \gamma_n W_n\psi(Y_n - X_n^{\mathrm{T}}\widehat\theta_{n-1}^*)X_n. \tag{25}$$

If $\psi(\cdot)$ is not differentiable, neither the final-step plug-in estimation (14) nor the recursive plug-in estimation (15) is applicable. However, if the corresponding $\ell(\theta; Z)$ and $L(\theta)$ satisfy Assumptions A1-A4, the perturbation resampling procedure is applicable. Next we consider a special setting where the following Assumptions B1-B4 hold.

(B1). Assume that $\rho(u)$ is a convex function on $\mathcal{R}$ with the right derivative being $\psi_+(u)$ and left derivative being $\psi_-(u)$. Let $\psi(u)$ be a function such that $\psi_-(u) \le \psi(u) \le \psi_+(u)$. There exists constant $C_1 > 0$ such that $|\psi(u)| \le C_1(1 + |u|)$.

(B2). Let $\varepsilon_n = Y_n - X_n^{\mathrm{T}}\theta_0$. Assume that $(X_n, \varepsilon_n), n = 1, 2, ...,$ are i.i.d. copies of (X, ε), with $\mathbb{E}\|X\|^4 < \infty$ and $\mathbb{E}\|\varepsilon\|^2 < \infty$. Let $V_0 = \mathbb{E}\{\psi^2(\varepsilon)XX^{\mathrm{T}}\} > 0$.

(B3). Let $\phi(u|X) = \mathbb{E}\{\psi(u + \varepsilon)|X\}$. Assume that $\phi(0|X) = 0$, $u\phi(u|X) > 0$ for any $u \ne 0$, and $\phi(u|X)$ has a derivative at $u = 0$ with $\dot\phi(0|X) \ge \sigma > 0$ uniformly over X. Let $S_0 = \mathbb{E}\{\dot\phi(0|X)XX^{\mathrm{T}}\} > 0$.

(B4). Assume that $\dot\phi(u|X)$ is uniformly Lipschitz at $u = 0$. That is, there exist constants $C_2 > 0$ and $\delta > 0$ such that $\left|\dot\phi(u|X) - \dot\phi(0|X)\right| \le C_2|u|$ for $|u| \le \delta$ uniformly over X.

We derive the asymptotic properties of the ψ-type M-estimator as follows.

Lemma 4 *If Assumptions B1-B4 and A5 are satisfied, then we have*

$$\sqrt{n}(\overline{\theta}_n - \theta_0) \Rightarrow \mathcal{N}\left(0, S_0^{-1} V_0 S_0^{-1}\right), \text{ in distribution, as } n \to \infty. \tag{26}$$

Theorem 5 *If Assumptions B1-B4 and A5-A6 are satisfied, then we have*

$$\sqrt{n}(\overline{\theta}_n^* - \theta_0) = \frac{1}{\sqrt{n}} S_0^{-1} \sum_{i=1}^{n} W_i \psi(\varepsilon_i) X_i + o_p(1). \tag{27}$$

From Lemma 2 we see that the plug-in procedures are not applicable for estimating the asymptotic covariance matrix, because although they are applicable for estimating V_0, they are not applicable for estimating S_0, which involves $\dot{\phi}(0|X)$. Moreover, by Theorem 3, we can show that the Kolmogorow-Smirnov distance between $\sqrt{n}(\overline{\theta}_n^* - \overline{\theta}_n)$ and $\sqrt{n}(\overline{\theta}_n - \theta_0)$ converges to zero in probability, as stated in Theorem 2. This validates our proposal of the perturbation-based resampling procedure for robust regression. To further understand Assumptions B1-B4, we examine the following example of quantile regression, which is a special case of the above robust regression.

Example 4 (Quantile regression). Assume the τ-quantile of Y given X is $X^{\mathrm{T}}\theta_0$, with

$$\theta_0 = \arg\min_{\theta \in \mathcal{R}^p} \mathbb{E}\rho_\tau(Y - X^{\mathrm{T}}\theta), \tag{28}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ with a given $0 < \tau < 1$. Let $\varepsilon = Y - X^{\mathrm{T}}\theta_0$ and $\psi_\tau(u) = \tau - I(u < 0)$. Thus $\mathbb{E}\{\psi_\tau(\varepsilon)|X\} = \tau - F_\varepsilon(0|X) = 0$, where $F_\varepsilon(u|X)$ is the conditional distribution function of ε. Let $p_\varepsilon(u|X)$ be the conditional density function of ε. Note that $\phi(u|X) = \mathbb{E}\{\psi_\tau(u+\varepsilon)|X\} = \tau - F_\varepsilon(-u|X)$, $\dot{\phi}(0|X) = p_\varepsilon(0|X)$, $V_0 = \mathbb{E}\{\psi_\tau^2(\varepsilon)XX^{\mathrm{T}}\} = \tau(1-\tau)\mathbb{E}\{XX^{\mathrm{T}}\}$ and $S_0 = \mathbb{E}\{p_\varepsilon(0|X)XX^{\mathrm{T}}\}$. Therefore, we can easily verify that if $p_\varepsilon(0|X)$ is uniformly bounded away from 0 and $p_\varepsilon(u|X)$ is uniformly Lipschitz continuous at $u = 0$, then Assumptions B1-B4 hold. Thus, the SGD and perturbed SGD updates for θ_0, as defined in (3) and (5) respectively, are

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} + \gamma_n \left\{\tau - I(Y_n - X_n^{\mathrm{T}}\widehat{\theta}_{n-1} < 0)\right\} X_n, \tag{29}$$

$$\widehat{\theta}_n^* = \widehat{\theta}_{n-1}^* + \gamma_n W_n \left\{\tau - I(Y_n - X_n^{\mathrm{T}}\widehat{\theta}_{n-1}^* < 0)\right\} X_n, \tag{30}$$

and the asymptotic results stated in Lemma 2 and Theorem 3 follow directly.

4. Numerical results

4.1. Simulation studies

To assess the performance of the proposed online bootstrap resampling procedure (a.k.a. random weighting procedure; RW) for SGD estimators, we conduct simulation studies for those four examples discussed in Section 3. We compare the proposed procedure with the recursive plug-in procedure (RPI).

Setting 1 (Linear regression): Consider linear regression (16), where covariates $X^{(j)}$, $j = 1, \ldots, p$, and residual $\varepsilon = Y - X^{\mathrm{T}}\theta_0$, are independently generated from standard normal

$N(0,1)$. Here $X^{(j)}$ indicates the j-th dimension of X. Let $\theta_0 = (\mu \mathbf{1}_{q/2}^{\mathrm{T}}, -\mu \mathbf{1}_{q/2}^{\mathrm{T}}, \mathbf{0}_{p-q}^{\mathrm{T}})^{\mathrm{T}}$ (same for the other three settings). Consider the corresponding SGD estimators (17) and (18).

Setting 2 (Logistic regression): Consider logistic regression (19), where covariates $X^{(j)}$ are independently from $N(0,1)$ and response Y from Bernoulli distribution with $\mathrm{logit}\{P(Y=1|X)\} = X^{\mathrm{T}}\theta_0$. Consider the corresponding SGD estimators (20) and (21).

Setting 3 (LAD regression): Consider least absolute deviation (LAD) regression, which is a special case of robust regression (22) with $\rho(x) = |x|$ and quantile regression (28) with $\tau = 0.5$, where covariates $X^{(j)}$, $j = 1, \ldots, p$, are i.i.d. with $N(0,1)$ and residual ε, defined as $Y - X^{\mathrm{T}}\theta_0$, is independently from double exponential distribution $DE(0,1)$. Consider the corresponding SGD estimators (29) and (30) with $\tau = 0.5$.

Setting 4 (LAD regression for data with outliers): Consider LAD regression for the data generated from Setting 1 but contaminated with 10% outliers. The contaminated data are obtained by transforming the outcome variable in the data generated from Setting 1 using $Y \leftarrow Y + 10$ if $|X^{(1)}| \geq 1.96$ and $|X^{(2)}| < 1.96$; and $Y \leftarrow Y - 10$ if $|X^{(1)}| < 1.96$ and $|X^{(2)}| \geq 1.96$. In this setting, covariate vector X and residual $\varepsilon = Y - X^{\mathrm{T}}\theta_0$ are not independent, but $\mathrm{median}\{\varepsilon|X\} = 0$. Consider the corresponding SGD estimators defined in (29) and (30) with $\tau = 0.5$.

For each simulation setting, we consider four scenarios, as described by (N, p, q, μ), where sample size $N = 10,000$ or $20,000$, number of covariates $p = 10$ or 20, number of informative covariates $q = 6$, and effect size $\mu = 0.1$ or 0.2. For each example, we repeat the data generation 1000 times. For each data repetition, we use $W_{nb} \sim \exp(1)$ as random weights and generate $B = 200$ copies of random weights whenever a new data point is read. Then, for each data repetition, we obtain the SGD estimate (4), and apply the following procedures to construct 95% confidence intervals: the proposed random weighting procedure to obtain 2.5% and 97.5% quantile (RW-Q), the proposed random weighting procedure to estimate its standard error (RW-σ) and then construct "estimate $\pm$ 1.96 $\times$ $\widehat{\mathrm{SE}}$", and the recursive plug-in procedures (RPI), if applicable, to estimate its standard error. We consider the learning rate $\alpha = 2/3$. When we calculate the average SGD estimators (4) and (6), the first 2000 and 4000 estimates are excluded for $N = 10,000$ and $N = 20,000$, respectively. We also obtain the empirical standard error based on 1000 repeated SGD estimates, as a benchmark approximation to the true standard error.

The coverage probabilities of the 95% confidence interval estimates are summarized in Table 1 for linear regression (Setting 1), Table 2 for logistic regression (Setting 2) and Table 3 for LAD regression (Settings 3-4), respectively. We only report results corresponding to the first, fourth and seventh covariates (that is, $X^{(1)}$, $X^{(q/2+1)}$ and $X^{(q+1)}$). The plug-in procedures are not applicable for LAD regression in Settings 3-4.

From Tables 1 and 2, we see that, for linear regression and logistic regression, the coverage probabilities using the RW-Q, RW-σ and RPI are close to 95%. Therefore, the proposed random weighting procedures (both RW-Q and RW-σ) perform well for linear regression and logistic regression, and if we choose to use the plug-in procedure, which involves matrix inverse, we can use RPI. Since the point estimate is $\sum_{i=N_0+1}^{N} \widehat{\theta}_i/(N - N_0)$, where N_0 is the number of excluded estimates and which involves a pass of SGD estimates, its standard error should also involve a pass of SGD estimates, instead of involving only the final-step estimate or the true parameter value. We can understand this clearer if we see the "SE" column, where the standard errors from RW-σ and RPI are close to the

empirical standard error. Moveover, from Table 3, we see that, for LAD regression and for both Settings 3 and 4, the coverage probabilities using either of the two proposed random weighting procedures are close to 95%, and the standard errors using RW-σ are close to the empirical standard errors.

Table 1: Coverage probabilities of 95% confidence intervals for linear regression

(N, p, q, μ)	Method	Dim 1		Dim $q/2+1$		Dim $q+1$	
		Cover	SE	Cover	SE	Cover	SE
(10000,10,6,0.1)	RW-Q	0.947	–	0.946	–	0.950	–
	RW-σ	0.956	0.012	0.950	0.012	0.959	0.012
	RPI	0.953	0.011	0.946	0.011	0.956	0.011
	Empirical	–	0.011	–	0.011	–	0.011
(10000,10,6,0.2)	RW-Q	0.947	–	0.946	–	0.950	–
	RW-σ	0.956	0.012	0.950	0.012	0.959	0.012
	RPI	0.953	0.011	0.946	0.011	0.956	0.011
	Empirical	–	0.011	–	0.011	–	0.011
(20000,20,6,0.1)	RW-Q	0.939	–	0.951	–	0.945	–
	RW-σ	0.949	0.008	0.953	0.008	0.960	0.008
	RPI	0.956	0.008	0.948	0.008	0.957	0.008
	Empirical	–	0.008	–	0.008	–	0.008
(20000,20,6,0.2)	RW-Q	0.939	–	0.951	–	0.945	–
	RW-σ	0.949	0.008	0.953	0.008	0.960	0.008
	RPI	0.956	0.008	0.948	0.008	0.957	0.008
	Empirical	–	0.008	–	0.008	–	0.008

4.2. Real data applications

In this section, we apply the proposed procedures to conduct inference for linear regression analysis for the individual household electric power consumption data (POWER) and logistic regression analysis for the skin segmentation dataset (SKIN) and gas sensors for the home Activity monitoring data (GAS). All the three datasets are publicly available on UCI machine learning repository.

The POWER dataset contains 2,075,259 observations and we fit a linear model to investigate how the time of a day influences the response variable "sub-metering-1", the energy sub-metering No. 1, in watt-hour of active energy corresponding to kitchen. The observations with missing value are deleted, and the time of a day is divided into 8 categories, "*Time 0-2*", "*Time 3-5*", ..., and "*Time 21-23*". The SKIN dataset contains 245,057 observations, out of which 50,859 is the skin samples and 194,198 is non-skin samples. We fit a logistic model to examine the relationship between the indicator of skin and three predictors, B, G and R. The GAS dataset contains 919,438 observations and we only use a subset containing 652,024 observations with response variable being either "banana" or "wine". We fit a logistic model to examine the association between the response variable and 11 explanatory variables, *Time*, *R1* to *R8*, *Temperature* and *Humidity*.

Table 2: Coverage probabilities of 95% confidence intervals for logistic regression

(N, p, q, μ)	Method	Dim 1		Dim $q/2+1$		Dim $q+1$	
		Cover	SE	Cover	SE	Cover	SE
(10000,10,6,0.1)	RW-Q	0.950	–	0.948	–	0.928	–
	RW-σ	0.954	0.024	0.954	0.024	0.948	0.023
	RPI	0.954	0.023	0.952	0.023	0.942	0.023
	Empirical	–	0.022	–	0.022	–	0.023
(10000,10,6,0.2)	RW-Q	0.945	–	0.945	–	0.944	–
	RW-σ	0.959	0.025	0.954	0.025	0.954	0.024
	RPI	0.955	0.023	0.952	0.023	0.948	0.023
	Empirical	–	0.023	–	0.023	–	0.023
(20000,20,6,0.1)	RW-Q	0.944	–	0.939	–	0.934	–
	RW-σ	0.951	0.016	0.953	0.016	0.952	0.016
	RPI	0.948	0.016	0.953	0.016	0.949	0.016
	Empirical	–	0.016	–	0.016	–	0.016
(20000,20,6,0.2)	RW-Q	0.946	–	0.941	–	0.939	–
	RW-σ	0.958	0.017	0.952	0.017	0.957	0.016
	RPI	0.950	0.016	0.944	0.016	0.952	0.015
	Empirical	–	0.016	–	0.016	–	0.016

Although standard softwares such as SAS and R can fit linear and logistic regression to such datasets without difficulty, for the illustration purpose, we use the SGD as in Examples 1 and 2 to fit linear and logistic regression and use the proposed online boostrap procedure to construct confidence intervals. The point estimates and the 95% confidence intervals of the coefficients are showed in Table 4. From the left-top panel of Table 4, we see that the electronic power consumption from kitchen is relatively high in the evening and night. From the left-bottom panel of Table 4, we see that variable B is positively associated with the response while the other two variables G and R are negatively associated. From the right panel of Table 4, we see that all the variables but *R4* are statistical significantly associated with the response. Furthermore, we display the histogram of $B = 1000$ perturbation-based SGD estimates for each coefficient in Figures 4.2-4.2 for the POWER data, the SKIN data and the GAS data, respectively. The blue triangle in each figure indicates the corresponding point estimate the red triangles indicate 2.5 and 97.5 quantiles. From these figures, we see the the perturbation-based procedure can be used to approximate the sampling distribution of the corresponding SGD estimator, which might be skewed. For example, for the GAS data, the sampling distributions are very skewed, and therefore the proposed resampling procedure is able to display such skewness.

5. Discussion

Online updating is a useful strategy for analyzing big data and streaming data, and recently stochastic gradient decent has become a popular method for doing online updating. Although the asymptotic properties of SGD have been well studied, there is little research

Table 3: Coverage probabilities of 95% confidence intervals for LAD regression

(N, p, q, μ)	Method	Dim 1		Dim $q/2+1$		Dim $q+1$	
		Cover	SE	Cover	SE	Cover	SE
Simulation Setting 3							
(10000,10,6,0.1)	RW-Q	0.965	–	0.965	–	0.969	–
	RW-σ	0.969	0.029	0.965	0.029	0.965	0.029
	Empirical	–	0.026	–	0.027	–	0.026
(10000,10,6,0.2)	RW-Q	0.972	–	0.965	–	0.967	–
	RW-σ	0.973	0.029	0.966	0.029	0.968	0.029
	Empirical	–	0.026	–	0.027	–	0.026
(20000,20,6,0.1)	RW-Q	0.970	–	0.973	–	0.966	–
	RW-σ	0.966	0.010	0.969	0.010	0.965	0.010
	Empirical	–	0.009	–	0.009	–	0.009
(20000,20,6,0.2)	RW-Q	0.967	–	0.974	–	0.966	–
	RW-σ	0.966	0.010	0.969	0.010	0.969	0.010
	Empirical	–	0.009	–	0.009	–	0.009
Simulation Setting 4							
(10000,10,6,0.1)	RW-Q	0.954	–	0.971	–	0.950	–
	RW-σ	0.958	0.035	0.969	0.035	0.960	0.035
	Empirical	–	0.032	–	0.031	–	0.033
(10000,10,6,0.2)	RW-Q	0.960	–	0.966	–	0.955	–
	RW-σ	0.958	0.035	0.966	0.035	0.956	0.035
	Empirical	–	0.032	–	0.031	–	0.033
(20000,20,6,0.1)	RW-Q	0.948	–	0.953	–	0.965	–
	RW-σ	0.963	0.047	0.958	0.047	0.964	0.047
	Empirical	–	0.043	–	0.045	–	0.044
(20000,20,6,0.2)	RW-Q	0.944	–	0.957	–	0.961	–
	RW-σ	0.961	0.047	0.961	0.047	0.961	0.047
	Empirical	–	0.044	–	0.045	–	0.044

on conducting statistical inference based on SGD estimators. In this paper, we propose the perturbation-based resampling procedure, which can be applied to estimate the sampling distribution of an SGD estimator. The offline version of perturbation-based resamping procedure was first proposed by Rubin (1981) and was also discussed in Shao and Tu (2012).

The proposed resampling procedure is in essence an online version of the bootstrap. Recall that the data points, $Z_1, Z_2, \ldots, Z_N$, are arriving one at a time and an SGD estimate updates itself from $\widehat{\theta}_{n-1}$ to $\widehat{\theta}_n$ whenever a new data point Z_n arrives. If we are forced to apply the bootstrap, then we should have many bootstrap samples; the data points of each bootstrap sample, $Z_1^*, Z_2^*, \ldots, Z_N^*$, are assumed to be arriving one at a time and the bootstrapped SGD estimate updates itself from $\widehat{\theta}_{n-1}^*$ to $\widehat{\theta}_n^*$ whenever a new data point Z_n^* arrives. Of course such bootstrap is impractical here because in online updating we will not obtain and store all the data points and then generate bootstrap samples. Now if we rearrange hypothetical bootstrap sample $Z_1^*, Z_2^*, \ldots, Z_N^*$ as

Table 4: Point estimates and 95% confidence intervals for three real datasets; POWER data on the left-top panel, SKIN data on the left-bottom panel and GAS data on the right panel

Variable	Est.	95% CI	Variable	Est.	95% CI
Time 0-2	2.265	$(2.254, 2.275)$	*Time*	-0.158	$(-0.178, -0.139)$
Time 3-5	2.045	$(2.040, 2.049)$	*R1*	-0.202	$(-0.215, -0.190)$
Time 6-8	2.623	$(2.608, 2.639)$	*R2*	0.176	$(0.160, 0.191)$
Time 9-11	3.323	$(3.298, 3.347)$	*R3*	-0.907	$(-0.932, -0.882)$
Time 12-14	3.445	$(3.420, 3.470)$	*R4*	-0.007	$(-0.018, 0.004)$
Time 15-17	3.059	$(3.037, 3.082)$	*R5*	-0.450	$(-0.467, -0.432)$
Time 18-20	4.176	$(4.143, 4.208)$	*R6*	1.772	$(1.759, 1.785)$
Time 21-23	4.053	$(4.024, 4.082)$	*R7*	0.173	$(0.139, 0.207)$
B	1.501	$(1.441, 1.569)$	*R8*	0.302	$(0.272, 0.332)$
G	-0.242	$(-0.319, -0.166)$	*Temp.*	-0.175	$(-0.191, -0.160)$
R	-1.956	$(-1.999, -1.918)$	*Humi.*	-0.551	$(-0.560, -0.542)$

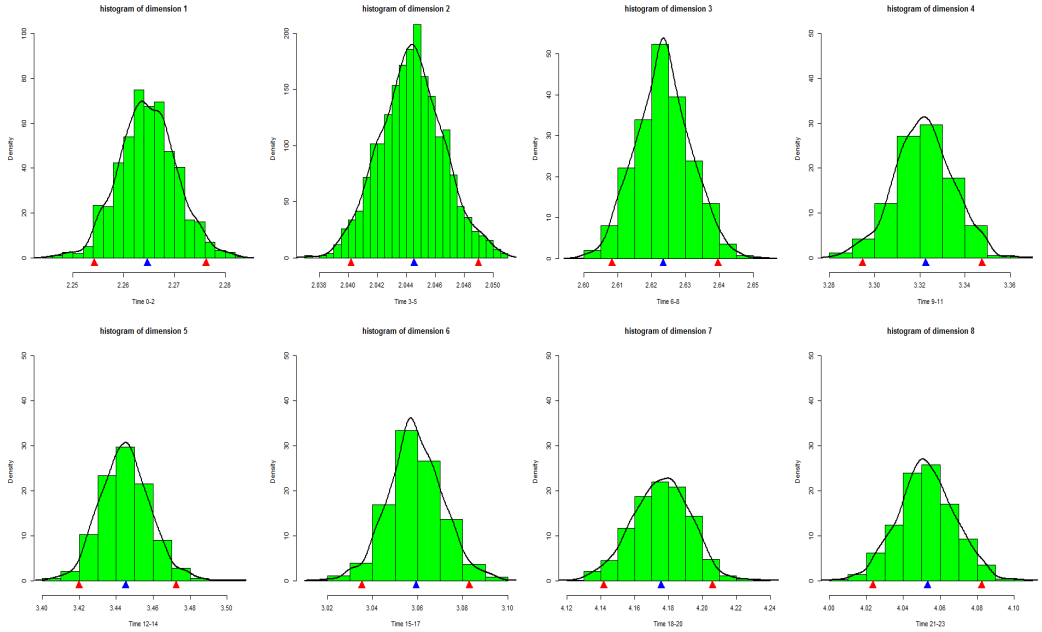

Figure 1: Histograms of $B = 1000$ perturbation-based SGD estimates for the POWER data.

$\{K_1$ copies Z_1, K_2 copies $Z_2, \ldots, K_N$ copies $Z_N\}$, where K_n follows binomial distribution $B(N, 1/N)$, then the SGD estimator updates itself from $\widehat{\theta}^*_{n-1}$ to $\widehat{\theta}^*_n$ whenever a new batch of data points, K_n copies of Z_n, arrives. Noting that binomial distribution $B(N, 1/N)$ approximates to Poisson distribution $P(1)$ as $N \to \infty$, we see that the aforementioned hypothetical

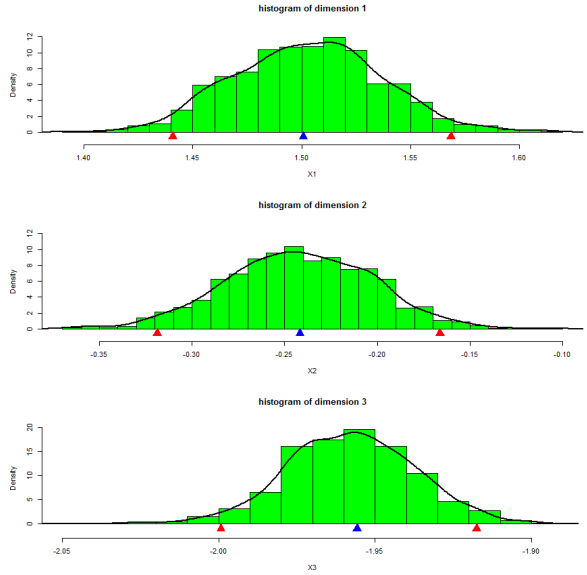

Figure 2: Histograms of $B = 1000$ perturbation-based SGD estimates for the SKIN data.

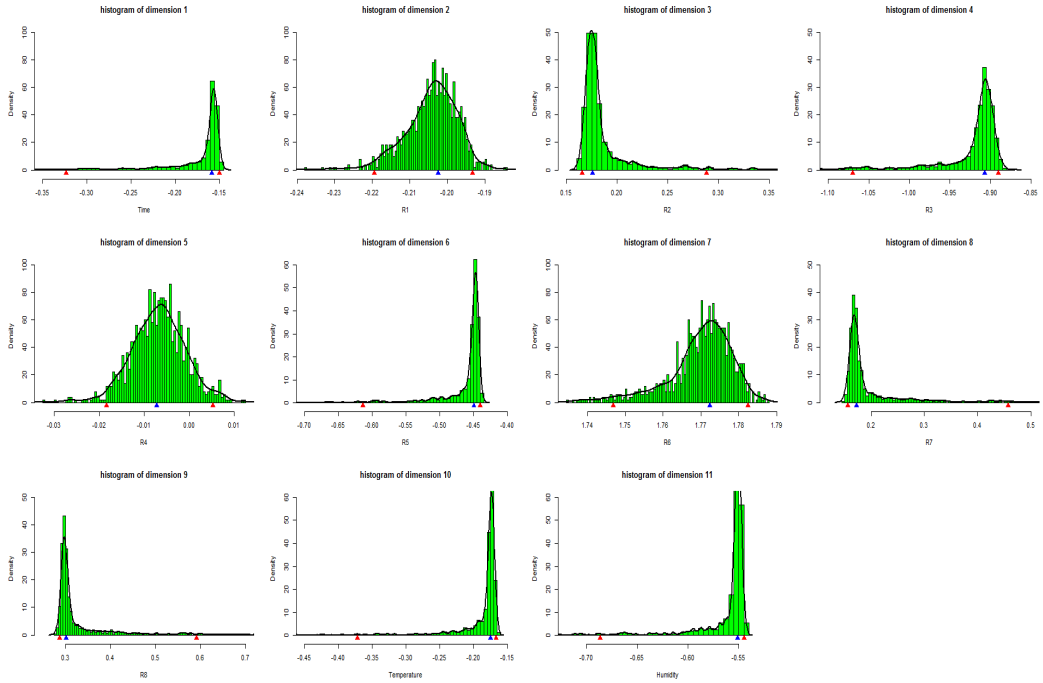

Figure 3: Histograms of $B = 1000$ perturbation-based SGD estimates for the GAS data.

bootstrap is equivalent to our proposed online bootstrap procedure with $W_n \sim P(1)$, whose mean and variance are both equal to one.

Finally, the SGD method considered in this paper is actually the explicit SGD, in contrast with the implicit SGD considered in Toulis et al (2017). We are working on extending

the proposed perturbation-based resampling procedure for conducting statistical inference for the implicit SGD.

Appendix A. technical proofs

For ease exposition of establishing asymptotic normality of SGD and perturbed SGD estimates, we present the following Proposition 1, adapted from Theorem 2 of Polyak and Juditsky (1992; pp.841). Let $R(\theta) : \mathcal{R}^p \to \mathcal{R}^p$ be some unknown function and $R(\theta_0) = 0$. The dataset consists of $Z_n, n = 1, 2, \ldots,$ which are i.i.d. copies of Z. Stochastic gradients are $\widehat{R}(\theta; Z_i)$ and $\mathbb{E}\{\widehat{R}(\theta; Z_i)\} = R(\theta)$. With an initial point $\widehat{\theta}_0$ and learning rates γ_n, the SGD estimate is defined as

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} - \gamma_n \widehat{R}(\widehat{\theta}_{n-1}; Z_n) = \widehat{\theta}_{n-1} - \gamma_n \left(R(\widehat{\theta}_{n-1}) - D_n \right), \tag{31}$$

where $D_n = R(\widehat{\theta}_{n-1}) - \widehat{R}(\widehat{\theta}_{n-1}; Z_n)$ is a martingale-difference process; that is, $\mathbb{E}\{D_n | \mathfrak{F}_{n-1}\} = 0$, where $\mathfrak{F}_{n-1} = \sigma(\mathcal{D}_{n-1})$. The regularity conditions for Proposition 1 are listed as follows.

(C1). There exists a function $U(\theta) : \mathcal{R}^p \to \mathcal{R}$ such that for some $\lambda > 0$, $\delta > 0$, $l_0 > 0$, $L_0 > 0$, and all $\theta, \theta' \in \mathcal{R}^p$, the conditions $U(\theta) \geq \lambda \|\theta\|^2$, $\|\nabla U(\theta) - \nabla U(\theta')\| \leq L_0 \|\theta - \theta'\|$, $U(0) = 0$, $\nabla U(\theta - \theta_0)^{\mathrm{T}} R(\theta) > 0$ for $\theta \neq \theta_0$ hold true. Moreover, $\nabla U(\theta - \theta_0)^{\mathrm{T}} R(\theta) \geq l_0 U(\theta - \theta_0)$ for all $\|\theta - \theta_0\| \leq \delta$.

(C2). There exists a positive definite matrix $S_0 \in \mathcal{R}^{p \times p}$ such that for some $C > 0$, $0 < \varrho \leq 1$, and $\delta > 0$, the condition $\|R(\theta) - S_0(\theta - \theta_0)\| \leq C\|\theta - \theta_0\|^{1+\varrho}$ for all $\|\theta - \theta_0\| \leq \delta$ holds true.

(C3). $\{D_n\}_{n \geq 1}$ is a martingale difference process with $\mathbb{E}\{D_n | \mathfrak{F}_{n-1}\} = 0$, and for some $C > 0$,

$$\mathbb{E}\left\{ \|D_n\|^2 \,|\, \mathfrak{F}_{n-1} \right\} + \|R(\widehat{\theta}_{n-1})\|^2 \leq C \left(1 + \|\widehat{\theta}_{n-1}\|^2 \right) \text{ a.s.},$$

for all $n \geq 1$. Consider decomposition $D_n = D_n(0) + E_n(\widehat{\theta}_{n-1})$, where $D_n(0) = R(\theta_0) - \widehat{R}(\theta_0; Z_n)$ and $E_n(\widehat{\theta}_{n-1}) = D_n - D_n(0)$. Assume that $\mathbb{E}\{D_n(0) | \mathfrak{F}_{n-1}\} = 0$ a.s.,

$$\mathbb{E}\{D_n(0) D_n(0)^{\mathrm{T}} | \mathfrak{F}_{n-1}\} \xrightarrow{P} V_0 > 0,$$
$$\sup_{n \geq 1} \mathbb{E}\left\{ \|D_n(0)\|^2 I(\|D_n(0)\| > \eta) | \mathfrak{F}_{n-1} \right\} \xrightarrow{P} 0, \text{ as } \eta \to \infty,$$

and there exists $\delta(x) \to 0$ as $x \to 0$ such that, for all n large enough,

$$\mathbb{E}\left\{ \|E_n(\widehat{\theta}_{n-1})\|^2 \,|\, \mathfrak{F}_{n-1} \right\} \leq \delta(\|\widehat{\theta}_{n-1} - \theta_0\|) \text{ a.s..}$$

(C4). It holds that $(\gamma_n - \gamma_{n+1})/\gamma_n = o(\gamma_n)$, $\gamma_n > 0$ for all n, and $\sum_{n=1}^{\infty} \gamma_n^{(1+\varrho)/2} n^{-1/2} < \infty$.

Assumptions C2 and C4 are implied by the following Assumptions C2′ (i.e. Assumption C2 with $\varrho = 1$) and C4′ (i.e. Assumption A5):

(C2′). There exists a positive definite matrix $S \in \mathcal{R}^{p \times p}$ such that for some $C > 0$ and $\delta > 0$, the condition $\|R(\theta) - S_0(\theta - \theta_0)\| \leq C\|\theta - \theta_0\|^2$ for all $\|\theta - \theta_0\| \leq \delta$ holds true.

(C4′). The learning rates are chosen as $\gamma_n = \gamma_1 n^{-\alpha}$ with $\gamma_1 > 0$ and $\alpha \in (0.5, 1)$.

Proposition 1. *If Assumptions C1-C4 are satisfied, then we have* $\overline{\theta}_n \to \theta_0$, *a.s.; and*

$$\sqrt{n}(\overline{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} S^{-1} \sum_{i=1}^{n} D_i + o_p(1), \tag{32}$$

which implies $\sqrt{n}(\overline{\theta}_n - \theta_0) \Rightarrow \mathcal{N}\left(0, S_0^{-1} V_0 S_0^{-1}\right)$, *in distribution.*

A1. Proof of Lemma 1

Proof. By Proposition 1, it suffices to show that Assumptions C1-C4 hold under Assumptions A1-A5. Because C2 and C4 are implied by C2′ and C4′, it suffices to show that Assumptions C1, C2′, C3 and C4′ hold under Assumptions A1-A5.

Verification of Assumption C1: Recall that $R(\theta) = \nabla L(\theta)$ and $\widehat{R}(\theta; Z_i) = \nabla \ell(\theta; Z_i)$. Define $U(\theta) = L(\theta_0 + \theta) - L(\theta_0) + \lambda\|\theta\|^2$ for a given $\lambda > 0$. By definition of $U(\theta)$ and Assumption A1, we have $U(\theta) \geq \lambda\|\theta\|^2$ and $U(0) = 0$. For any θ and θ', since $\nabla U(\theta) - \nabla U(\theta') = R(\theta+\theta_0) - R(\theta'+\theta_0) + 2\lambda(\theta - \theta')$, letting $L_0 = L_1 + 2\lambda$ and by Assumption A2, we have $\|\nabla U(\theta) - \nabla U(\theta')\| \leq L_1\|\theta - \theta'\|$. Since $\nabla U(\theta - \theta_0)^{\mathrm{T}} R(\theta) = \|R(\theta)\|^2 + \lambda(\theta - \theta_0)^{\mathrm{T}} R(\theta)$, by Assumption A1, we have $\nabla U(\theta - \theta_0)^{\mathrm{T}} R(\theta) > 0$ for any $\theta \neq \theta_0$. Last, it remains to verify there exist $l_0 > 0$ and $\delta > 0$ such that $\nabla U(\theta - \theta_0)^{\mathrm{T}} R(\theta) \geq l_0 U(\theta - \theta_0)$ for all $\|\theta - \theta_0\| \leq \delta$. Noting that $U(\theta - \theta_0) = L(\theta + \theta_0) + \lambda\|\theta - \theta_0\|^2$, by Taylor expansion and Assumption A3, we see that there exist $l_1 > 0$ and δ_1 such that $U(\theta - \theta_0) \leq l_1\|\theta - \theta_0\|^2$ for all $\|\theta - \theta_0\| \leq \delta_1$. On the other hand, noting that $\nabla U(\theta - \theta_0)^{\mathrm{T}} R(\theta) = \|R(\theta)\|^2 + \lambda(\theta - \theta_0)^{\mathrm{T}} R(\theta)$, by Assumption A3 also, we see that there exist $l_2 > 0$ and δ_2 such that $(\theta - \theta_0)^{\mathrm{T}} R(\theta) \geq l_2\|\theta - \theta_0\|^2$ for all $\|\theta - \theta_0\| \leq \delta_2$. Selecting $\delta = \min(\delta_1, \delta_2)$ and $l_0 = \lambda l_2 / l_2$, we show that $\nabla U(\theta - \theta_0)^{\mathrm{T}} R(\theta) \geq l_0 U(\theta - \theta_0)$ for all $\|\theta - \theta_0\| \leq \delta$.

Verification of Assumption C2′: Recall that $R(\theta) = \nabla L(\theta)$, $S(\theta) = \nabla R(\theta)$, and $S_0 = S(\theta_0) > 0$. By Assumption A3, there exists $\delta > 0$ such that for any $\|\theta - \theta_0\| < \delta'$, $\|S(\theta) - S(\theta_0)\| < L_2\|\theta - \theta_0\|$. By mean-value theorem, $\|S(\theta) - S_0(\theta - \theta_0)\| = \|S(\widetilde{\theta})(\theta - \theta_0) - S_0(\theta - \theta_0)\|$, where $\widetilde{\theta}$ lies between θ and θ_0. Hence $\|S(\theta) - S_0(\theta - \theta_0)\| \leq L_2\|\theta - \theta_0\|^2$, for any $\|\theta - \theta_0\| \leq \delta$. Letting $C = L_2$, we have verified Assumption C2′.

Verification of Assumption C3: $D_n = R(\widehat{\theta}_{n-1}) - \widehat{R}(\widehat{\theta}_{n-1}; Z_n)$. Consider decomposition $D_n = D_n(0) + E_n(\widehat{\theta}_{n-1})$, where $D_n(0) = -\nabla \ell(\theta_0, Z_n)$ and $E_n(\widehat{\theta}_{n-1}) = [R(\widehat{\theta}_{n-1}) - R(\theta_0)] - [\nabla \ell(\widehat{\theta}_{n-1}; Z_n) - \nabla \ell(\theta_0; Z_n)]$. By Assumption A2, $\|R(\widehat{\theta}_{n-1})\|^2 \leq L_1^2\|\widehat{\theta}_{n-1} - \theta_0\|^2$. In addition, Cauchy-Schwartz inequality implies that $\mathbb{E}\left\{\|\nabla \ell(\theta; Z_n) - \nabla \ell(\theta_0; Z_n)\|^2\right\} = 2\mathbb{E}\|\nabla \ell(\theta; Z_n)\|^2 + 2\mathbb{E}\|\nabla \ell(\theta_0; Z_n)\|^2$; we have $\mathbb{E}\left\{\|\nabla \ell(\widehat{\theta}_{n-1}; Z_n) - \nabla \ell(\theta_0; Z_n)\|^2 \mid\right\} \leq C'(1 + \|\theta\|^2)$ for some $C' > 0$ by Assumption A4, . Together, we have $\mathbb{E}\left\{\|D_n\|^2 \mid \mathfrak{F}_{n-1}\right\} + \|R(\widehat{\theta}_{n-1})\|^2 \leq C\left(1 + \|\widehat{\theta}_{n-1}\|^2\right)$ for some $C > 0$. Moreover, because $D_n(0)$'s are i.i.d., we have $\mathbb{E}\{D_n(0) D_n(0)^{\mathrm{T}} | \mathfrak{F}_{n-1}\} = V_0 > 0$ and $\sup_{n \geq 1} \mathbb{E}\left\{\|D_n(0)\|^2 I(\|D_n(0)\| > \eta) | \mathfrak{F}_{n-1}\right\} \xrightarrow{P} 0$, as $\eta \to \infty$. Finally, note that $\mathbb{E}\|E_n(\theta)\|^2 \leq L_1^2\|\overline{\theta} - \theta_0\|^2 + \mathbb{E}\|\nabla \ell(\theta; Z_n) - \nabla \ell(\theta_0; Z_n)\|^2$. By Assumption A3, $\mathbb{E}\|\nabla \ell(\theta; Z_n) - \nabla \ell(\theta_0; Z_n)\|^2 \leq \delta'(\|\theta - \theta_0\|)$ for some $\delta'(\cdot)$ with $\delta'(x) \to 0$ as $x \to 0$. Define $\delta(x) = L_1^2 x^2 + \delta'(x)$, we show $\mathbb{E}\|E_n(\theta)\|^2 \leq \delta(\|\theta - \theta_0\|)$. This complete the verification of Assumption C3.

Obviously Assumption A5 is the same as Assumption C4′. Therefore, we have verified that Assumptions A1-A5 imply Assumptions C1, C2′, C3 and C4′. By Proposition 1, we complete the proof of Lemma 1. □

A2. Proof of Theorem 1

Proof. Rewrite $\widehat{\theta}_n^*$ as

$$\widehat{\theta}_n^* = \widehat{\theta}_{n-1}^* - \gamma_n R(\widehat{\theta}_{n-1}^*) + \gamma_n D_n^*, \tag{33}$$

where $D_n^* = R(\widehat{\theta}_{n-1}^*) - W_n \nabla \ell(\widehat{\theta}_{n-1}^*; Z_n)$. Then let $\mathfrak{F}_{n-1}^*$ be the Borel field generated by $\{(Z_i, W_i), i \le n-1\}$. Since $\mathbb{E}\{W_n|\mathfrak{F}_{n-1}^*\} = 1$ and $R(\theta) = \mathbb{E}\{\nabla \ell(\theta; Z_n)\}$, we have $\mathbb{E}\{D_n^*|\mathfrak{F}_{n-1}\} = 0$. Thus D_n^* is a martingale-difference process. Let $D_n^*(\theta) = R(\theta) - W_n \nabla \ell(\theta; Z_n)$. Consider decomposition $D_n^* = D_n^*(0) + E_n^*(\widehat{\theta}_{n-1}^*)$, where

$$D_n^*(0) = -W_n \nabla \ell(\theta_0; Z_n) \tag{34}$$

and

$$E_n^*(\theta) = [R(\theta) - R(\theta_0)] - W_n[\nabla \ell(\theta; Z_n) - \nabla \ell(\theta_0; Z_n)]. \tag{35}$$

Noting that $\mathbb{E}\{D_n^*(0)\} = 0$ and $\mathbb{E}(W_n^2) = 2$ under Assumption A6, by Assumption A4, we have

$$\mathbb{E}\{[D_n^*(0)][D_n^*(0)]^{\mathrm{T}}\} = 2\mathbb{E}\left\{[\ell(\theta_0; Z_n)][\ell(\theta_0; Z_n)]^{\mathrm{T}}\right\} = 2V_0. \tag{36}$$

By Cauchy-Schwartz inequality and Assumptions A2 and A4, we have

$$\mathbb{E}\{\|E_n^*(\theta)\|^2\} \le 2\|R(\theta)\|^2 + 4\mathbb{E}\left\{\|\nabla \ell(\theta, Z) - \nabla \ell(\theta_0, Z)\|^2\right\} \le \delta'(\|\theta - \theta_0\|), \tag{37}$$

where $\delta'(x) = 2L_1^2 x^2 + 2\delta(x)$ satisfying that $\delta'(x) \to 0$ as $x \to 0$. Also by Cauchy-Schwartz inequality, $\mathbb{E}\{\|E_n^*(\theta)\|^2\} \le 2\|R(\theta)\|^2 + 2\mathbb{E}\{\|\nabla \ell(\theta, Z)\|^2\}$. Further, by Assumptions A2 and A4,

$$\mathbb{E}\{\|D_n^*\|^2|\mathfrak{F}_{n-1}\} + \|R(\widehat{\theta}_{n-1}^*)\|^2 \le 3L_1^2\|\widehat{\theta}_{n-1}^* - \theta_0\|^2 + 2C\|\widehat{\theta}_{n-1}^*\|^2 \le C'(1 + \|\widehat{\theta}_{n-1}^* - \theta_0\|^2), \tag{38}$$

for some large enough $C' > 0$. Combining results (36)-(38), we have verified Assumption C3. Moreover, noting that $\mathbb{E}W_n = 1$, we can easily verify that Assumptions A1 and A2 imply that Assumption C1 holds, and Assumption A3 implies that Assumption C2 holds. By Proposition 1, we show that $\widehat{\theta}_n^* \to \theta_0$ almost surely, and

$$\begin{aligned}
\sqrt{n}(\overline{\theta}_n^* - \theta_0) &= \frac{1}{\sqrt{n}} S_0^{-1} \sum_{i=1}^n D_i^* + o_p(1) \\
&= -\frac{1}{\sqrt{n}} S_0^{-1} \sum_{i=1}^n W_i \nabla \ell(\theta_0; Z_i) + \frac{1}{\sqrt{n}} S_0^{-1} \sum_{i=1}^n E_n^*(\widehat{\theta}_{n-1}^*) + o_p(1).
\end{aligned} \tag{39}$$

Note that $\mathbb{E}\{\|E_n(\widehat{\theta}_{n-1}^*)\|^2|\mathfrak{F}_{n-1}\} \le \delta'(\|\widehat{\theta}_{n-1}^* - \theta_0\|)$, following (37). Since $\widehat{\theta}_n^* \to \theta_0$ a.s., we have $\delta(\|\widehat{\theta}_{n-1}^* - \theta_0\|) \to 0$ a.s.. Thus, $S_0^{-1} \sum_{i=1}^n E_n(\widehat{\theta}_{n-1}^*)/\sqrt{n} = o_p(1)$. Therefore, by (39), this completes the proof of Theorem 1. □

A3. Proof of Theorem 2

Proof. Let

$$V_n = -\frac{1}{\sqrt{n}} S_0^{-1}(W_i - 1)\nabla\ell(\theta_0, Z_i). \tag{40}$$

By Theorem 1, we have $\sqrt{n}(\overline{\theta}_n^* - \overline{\theta}_n) = V_n + o_p(1)$. We first show that, for any $\beta \in \mathbf{B} \triangleq \{\beta \in \mathcal{R}^p : \|\beta\| = 1\}$ and $u \in \mathcal{R}$,

$$\mathbb{P}^*\left(\beta^{\mathrm{T}} V_n \le u\right) \to \Phi(u/\sigma_\beta), \text{ in probability}, \tag{41}$$

where $\Phi(u)$ is the distribution of $\mathcal{N}(0,1)$ and $\sigma_\beta = \beta^{\mathrm{T}} S_0^{-1} V_0 S_0^{-1}\beta$. In fact,

$$\beta^{\mathrm{T}} V_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (W_i - 1)\xi_i, \tag{42}$$

where $\xi_i = -\beta^{\mathrm{T}} S_0^{-1}\nabla\ell(\theta_0, Z_i)$. Note that, by Assumption A6, $\mathbb{E}W_i = \mathrm{Var}(W_i) = 1$. Hence

$$s_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}^*\{(W_i - 1)\xi_i\} = \frac{1}{n}\sum_{i=1}^{n} \xi_i^2 \to \sigma_\beta^2, \text{ in probability}, \tag{43}$$

and for any $\epsilon > 0$,

$$\frac{1}{n s_n}\sum_{i=1}^{n} \mathbb{E}^*\left\{(W_i - 1)^2 \xi_i^2 I\left[|(W_i - 1)\xi_i| > \sqrt{n} s_n \epsilon\right]\right\} \to 0, \text{ in probability}. \tag{44}$$

Therefore, the Lindeberg's condition is satisfied. By the central limit theorem, (41) holds, which implies that for any $\beta \in \mathbf{B}$,

$$\sup_{u\in\mathcal{R}} \left|\mathbb{P}^*\left(\beta^{\mathrm{T}} V_n \le u\right) - \Phi(u/\sigma_\beta)\right| \to 0, \text{ in probability}. \tag{45}$$

Consider $\mathbf{B}_0 \triangleq \{\beta \in \mathcal{R}^p : \|\beta\| = 1, \text{the components of } \beta \text{ are rational}\}$, which contains only countable many β and is a dense subset of $\mathbf{B}$. For any subsequence $\{n_1\}$, by Cantor's "diagonal method" used in Rao and Zhao (1992), we can show that there exists a subsequence $\{n_2\} \subset \{n_1\}$ such that, with probability one,

$$\sup_{u\in\mathcal{R}} \left|\mathbb{P}^*\left(\beta^{\mathrm{T}} V_{n-1} \le u\right) - \Phi(u/\sigma_\beta)\right| \to 0, \text{ for any } \beta \in \mathbf{B}_0. \tag{46}$$

Hence, we show that

$$\sup_{v\in\mathcal{R}^p} \left|\mathbb{P}^*\left(\sqrt{n}(\overline{\theta}_n^* - \overline{\theta}_n) \le v\right) - \mathbb{P}(\zeta \le v)\right| \to 0, \text{ in probability}, \tag{47}$$

where $\zeta \sim \mathcal{N}(0, S_0^{-1} V_0 S_0^{-1})$. By Lemma 1, we can also show that

$$\sup_{v\in\mathcal{R}^p} \left|\mathbb{P}\left(\sqrt{n}(\overline{\theta}_n - \theta_0) \le v\right) - \mathbb{P}(\zeta \le v)\right| \to 0. \tag{48}$$

Combining (47) and (48), we complete the proof of Theorem 2. □

A4. Proof of Lemma 2

Proof. By Proposition 1, it suffices to show that Assumptions C1, C2′ and C3 hold under Assumptions B1-B4.

Verification of Assumption C1: Define $\Delta = \theta - \theta_0$. Let $R(\Delta) = \mathbb{E}\{\phi(\Delta^{\mathrm{T}}X|X)X\}$ and $\widehat{R}(\Delta; Z_n) = \psi(\Delta^{\mathrm{T}}X_n+\varepsilon_n)X_n$. Define $U(\Delta) = \Delta^{\mathrm{T}}\Delta$. By definition of $U(\Delta)$, $U(\Delta) \geq \lambda\|\Delta\|^2$ with $\lambda = 1$, $U(0) = 0$ and $\nabla U(\Delta)$ is Lipschitz continuous. Since $\Delta^{\mathrm{T}}\mathbb{E}\{\phi(\Delta^{\mathrm{T}}X|X)\Delta^{\mathrm{T}}X\} > 0$ by Assumption B3, we see that $\nabla U(\Delta)^{\mathrm{T}}R(\Delta) > 0$. By the mean-value theorem and by Assumption B3, there exists δ such that $\Delta^{\mathrm{T}}\mathbb{E}\{\phi(\Delta^{\mathrm{T}}X|X)\Delta^{\mathrm{T}}X\} \geq \Delta^{\mathrm{T}}\mathbb{E}\{\dot{\phi}(0|X)XX^{\mathrm{T}}\}\Delta/2 \geq \lambda_{\min}(S_0)\|\Delta\|^2/2$, for any $\|\Delta\| \leq \delta$, where $\lambda_{\min}(S_0)$ is the minimum eigenvalue of S_0. Hence $\nabla U(\Delta)^{\mathrm{T}}R(\Delta) \geq \lambda_{\min}(S_0)\|\Delta\|^2$ and we have verified Assumption C1.

Verification of Assumption C2′: Note that

$$\|R(\Delta) - S_0\Delta\| = \|\mathbb{E}\phi(\Delta^{\mathrm{T}}X|X) - \mathbb{E}\dot{\phi}(0|X)XX^{\mathrm{T}}\Delta\|.$$

By the mean-value theorem and Assumption B4, there exists δ such that, for any $\|\Delta\| \leq \delta$, $\|\mathbb{E}\{\phi(\Delta^{\mathrm{T}}X|X)X\} - \mathbb{E}\{\dot{\phi}(0|X)XX^{\mathrm{T}}\Delta\}\| \leq C_2\lambda_{\max}(S_0)\|\Delta\|^2$. This implies Assumption C2′.

Verification of Assumption C3: Let $\widehat{\Delta}_n = \widehat{\theta}_n - \theta_0$. Then $D_n = R(\widehat{\Delta}_{n-1}) - \widehat{R}(\widehat{\Delta}_{n-1}; Z_n)$. Consider decomposition $D_n = D_n(0)+E_n(\widehat{\Delta}_{n-1})$, where $D_n(0) = \psi(\varepsilon_n)X_n$ and $E_n(\widehat{\Delta}_{n-1}) = \mathbb{E}\{\psi(\widehat{\Delta}_{n-1}^{\mathrm{T}}X+\varepsilon_n)X_n\}-[\psi(\widehat{\Delta}_{n-1}^{\mathrm{T}}X+\varepsilon_n)X_n-\psi(\varepsilon_n)X_n]$. By Assumption B1 and the Cauchy-Schwartz inequality, we can show that $\mathbb{E}\left\{\|D_n\|^2 \mid \mathfrak{F}_{n-1}\right\}+\|R(\widehat{\Delta}_{n-1})\|^2 \leq 2C_1\left(1 + \|\widehat{\Delta}_{n-1}\|^2\right)$. Moreover, by Assumption B2, $D_n(0)$'s are i.i.d., so $\mathbb{E}\{D_n(0)D_n(0)^{\mathrm{T}}|\mathfrak{F}_{n-1}\} = V_0 > 0$ and $\sup_{n\geq 1}\mathbb{E}\left\{\|D_n(0)\|^2 I(\|D_n(0)\| > \eta)|\mathfrak{F}_{n-1}\right\} \to 0$, as $\eta \to \infty$. Finally, by Cauchy-Schwartz inequality and Assumptions B2-B3, we can show that $\mathbb{E}\|E_n(\Delta)\|^2 \leq \delta(\|\Delta\|)$ for some $\delta(\cdot)$ with $\delta(x) \to 0$ as $x \to 0$. This complete the verification of Assumption C3.

Obviously Assumption A5 is the same as Assumption C4′. Therefore, we have verified that Assumptions B1-B4 and A5 imply Assumptions C1, C2′, C3 and C4′. By Proposition 1, we complete the proof of Lemma 2. □

A5. Proof of Theorem 3

Proof. Recall that $R(\Delta) = \mathbb{E}\{\phi(\Delta^{\mathrm{T}}X|X)X\}$ and $\widehat{\Delta}_n = \widehat{\theta}_n - \theta_0$. Rewrite $\widehat{\theta}_n^*$ as

$$\widehat{\theta}_n^* = \widehat{\theta}_{n-1}^* + \gamma_n R(\widehat{\Delta}_n) + \gamma_n D_n^*, \tag{49}$$

where $D_n^* = W_n\psi(\widehat{\Delta}_n^{\mathrm{T}}X_n+\varepsilon_n)X_n - R(\widehat{\Delta}_n)$ is a martingale-difference process by Assumption B3 and that $\mathbb{E}\{W_n|\mathfrak{F}_{n-1}\} = 1$. Let $D_n^*(\Delta) = W_n\psi\left(\Delta^{\mathrm{T}}X_n + \varepsilon_n\right)X_n - R(\Delta)$ and $D_n^*(\Delta) = D_n^*(0) + E_n^*(\Delta)$, where

$$E_n^*(\Delta) = W_n\left[\psi(\Delta^{\mathrm{T}}X_n + \varepsilon_n) - \psi(\varepsilon_n)\right]X_n - R(\Delta). \tag{50}$$

Since $D_n^*(0) = W_n\psi(\varepsilon_n)X_n$, $\mathbb{E}\{D_n^*(0)\} = 0$ and $\mathbb{E}\{[D_n^*(0)][D_n^*(0)]^{\mathrm{T}}\} = (1 + \mathrm{Var}(W_1))V_0$ by Assumption B2. By Cauchy-Schwartz inequality and Assumptions B2-B3, we can show that $\mathbb{E}\|E_n^*(\Delta)\|^2 \leq \delta(\|\Delta\|)$ for some $\delta(\cdot)$ with $\delta(x) \to 0$ as $x \to 0$. Therefore, using the similar arguments as those in the proof of Lemma 2, we can verify that, under Assumptions B1-B4, Assumptions C1-C4 are satisfied. By Proposition 1, it follows that $\widehat{\theta}_n^* \to \theta_0$ almost

surely, and

$$\sqrt{n}(\overline{\theta}_n^* - \theta_0) = \frac{1}{\sqrt{n}} S_0^{-1} \sum_{i=1}^{n} D_i^* + o_p(1). \tag{51}$$

By the decomposition of D_i^*, we have

$$\sqrt{n}(\overline{\theta}_n^* - \theta_0) = \frac{1}{\sqrt{n}} S_0^{-1} \sum_{i=1}^{n} W_i \psi(\varepsilon_i) X_i + \frac{1}{\sqrt{n}} S_0^{-1} \sum_{i=1}^{n} E_n^*(\widehat{\theta}_{n-1}^* - \theta_0) + o_p(1). \tag{52}$$

By the definition of $\delta(\|\Delta\|)$, $\mathbb{E}\{\|E_n^*(\widehat{\theta}_{n-1}^* - \Delta)\|_2^2 | \mathfrak{F}_{n-1}\} = \delta(\|\widehat{\theta}_{n-1}^* - \theta_0\|)$. Since $\widehat{\theta}_n^* \to \theta_0$ a.s., we have $\delta(\|\widehat{\theta}_{n-1}^* - \theta_0\|) \to 0$ a.s.. Thus, $\sum_{i=1}^{n} E_n^*(\widehat{\theta}_{n-1}^* - \theta_0)/\sqrt{n} = o_p(1)$. By (52), we complete the proof of Theorem 3. □

References

Moulines, Eric and Bach, Francis R. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 451–459, 2011.

Chen, Xi and Lee, Jason D and Tong, Xin T and Zhang, Yichen. Statistical Inference for Model Parameters in Stochastic Gradient Descent. *arXiv preprint arXiv:1610.08637*, 2016.

Efron, Bradley. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Wang, Chun and Chen, Ming-Hui and Schifano, Elizabeth and Wu, Jing and Yan, Jun. Statistical methods and computing for big data. *Statistics and its interface*, 9, 399.

Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome. The elements of statistical learning 2nd edition. *New York: Springer*, 2009.

Kushner, Harold and Yin, G George. Stochastic approximation and recursive algorithms and applications. *Springer Science & Business Media*, 2003.

Li, Tianyang and Liu, Liu and Kyrillidis, Anastasios and Caramanis, Constantine. Statistical methods and computing for big data. *arXiv preprint arXiv:1705.07477*, 2017.

Nelder, John A and Baker, R Jacob. Generalized linear models. *Encyclopedia of statistical sciences*, 1972.

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

Rao, C Radhakrishna and Zhao, LC. Approximation to the distribution of M-estimates in linear models by randomly weighted bootstrap. *Sankhyā: The Indian Journal of Statistics, Series A*, 323–331.

Robbins, Herbert and Monro, Sutton. A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Rubin, Donald B. The bayesian bootstrap. *The annals of statistics*, 9, 130–134.

Ruppert, David. Efficient estimations from a slowly convergent Robbins-Monro process. *Cornell University Operations Research and Industrial Engineering*, 1988.

Shao, Jun and Tu, Dongsheng. The jackknife and bootstrap. *Springer Science & Business Media*, 2012.

Su, Weijie and Zhu, Yuancheng. Statistical Inference for Online Learning and Stochastic Approximation via Hierarchical Incremental Gradient Descent. *arXiv preprint arXiv:1802.04876*, 2018.

Toulis, Panos and Airoldi, Edoardo M and others. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45, 1694–1727.