
**Technical Report 144**

# Statistical Inference Using Stochastic Gradient Descent: Volumes 1 and 2

**Research Supervisor**
Constantine Caramanis
Wireless Networking & Communications Group

**Project Title:** Statistical Inference Using Stochastic Gradient Descent

September 2018

# Data-Supported Transportation Operations & Planning Center (D-STOP)

A Tier 1 USDOT University Transportation Center at The University of Texas at Austin



D-STOP is a collaborative initiative by researchers at the Center for Transportation Research and the Wireless Networking and Communications Group at The University of Texas at Austin.

Technical Report Documentation Page

| | | |
|---|---|---|
| 1. Report No.<br>D-STOP/2018/144 | 2. Government Accession No. | 3. Recipient's Catalog No. |
| 4. Title and Subtitle<br>Statistical Inference Using Stochastic Gradient Descent: Volumes 1 and 2 | | 5. Report Date<br>October 2018<br>6. Performing Organization Code |
| 7. Author(s)<br>Natalia Ruiz Juri, Stephen D. Boyles, Tengkuo Zhu, Kenneth Perrine, Amber Chen, Yun Li | | 8. Performing Organization Report No.<br>Report 144 |
| 9. Performing Organization Name and Address<br>Data-Supported Transportation Operations & Planning Center (D-STOP)<br>The University of Texas at Austin<br>3925 W. Braker Lane, 4th Floor<br>Austin, Texas 78759 | | 10. Work Unit No. (TRAIS)<br>11. Contract or Grant No.<br>DTRT13-G-UTC58 |
| 12. Sponsoring Agency Name and Address<br>United States Department of Transportation<br>University Transportation Centers<br>1200 New Jersey Avenue, SE<br>Washington, DC 20590 | | 13. Type of Report and Period Covered<br>14. Sponsoring Agency Code |
| 15. Supplementary Notes<br>Supported by a grant from the U.S. Department of Transportation, University Transportation Centers Program. | | |

16. Abstract

**Volume 1:** We present a novel inference framework for convex empirical risk minimization, using approximate stochastic Newton steps. The proposed algorithm is based on the notion of finite differences and allows the approximation of a Hessian-vector product from first-order information. In theory, our method efficiently computes the statistical error covariance in M-estimation, both for unregularized convex learning problems and high-dimensional LASSO regression, without using exact second order information, or resampling the entire data set. In practice, we demonstrate the effectiveness of our framework on large-scale machine learning problems, that go even beyond convexity: as a highlight, our work can be used to detect certain adversarial attacks on neural networks.

**Volume 2:** We present a novel method for frequentist statistical inference in M-estimation problems, based on stochastic gradient descent (SGD) with a fixed step size: we demonstrate that the average of such SGD sequences can be used for statistical inference, after proper scaling. An intuitive analysis using the Ornstein-Uhlenbeck process suggests that such averages are asymptotically normal. From a practical perspective, our SGD-based inference procedure is a first order method, and is well-suited for large scale problems. To show its merits, we apply it to both synthetic and real datasets, and demonstrate that its accuracy is comparable to classical statistical methods, while requiring potentially far less computation.

| | | | |
|---|---|---|---|
| 17. Key Words | 18. Distribution Statement<br>No restrictions. This document is available to the public through NTIS (http://www.ntis.gov):<br>National Technical Information Service<br>5285 Port Royal Road<br>Springfield, Virginia 22161 | | |
| 19. Security Classif.(of this report)<br>Unclassified | 20. Security Classif.(of this page)<br>Unclassified | 21. No. of Pages | 22. Price |

## Disclaimer



## Acknowledgements

The authors recognize that support for this research was provided by a grant from the U.S. Department of Transportation, University Transportation Centers.

# Volume 1: Approximate Newton-based Statistical Inference Using Only Stochastic Gradients

# Approximate Newton-based statistical inference using only stochastic gradients

Tianyang Li[1]
lty@cs.utexas.edu

Liu Liu[1]
liuliu@utexas.edu

Anastasios Kyrillidis[2]
anastasios.kyrillidis@ibm.com

Constantine Caramanis[1]
constantine@utexas.edu

[1] The University of Texas at Austin
[2] IBM T.J. Watson Research Center, Yorktown Heights

**Abstract**

We present a novel inference framework for convex empirical risk minimization, using approximate stochastic Newton steps. The proposed algorithm is based on the notion of finite differences and allows the approximation of a Hessian-vector product from first-order information. In theory, our method efficiently computes the statistical error covariance in $M$-estimation, both for unregularized convex learning problems and high-dimensional LASSO regression, without using exact second order information, or resampling the entire data set. In practice, we demonstrate the effectiveness of our framework on large-scale machine learning problems, that go even beyond convexity: as a highlight, our work can be used to detect certain adversarial attacks on neural networks.

## 1 Introduction

Statistical inference is an important tool for assessing uncertainties, both for estimation and prediction purposes [25, 21]. *E.g.*, in unregularized linear regression and high-dimensional LASSO settings [53, 32, 49], we are interested in computing coordinate-wise confidence intervals and p-values of a $p$-dimensional variable, in order to infer which coordinates are active or not [58]. Traditionally, the inverse Fisher information matrix [20] contains the answer to such inference questions; however it requires storing and computing a $p \times p$ matrix structure, often prohibitive for large-scale applications [52]. Alternatively, the Bootstrap method is a popular statistical inference algorithm, where we solve an optimization problem per dataset replicate, but can be expensive for large data sets [35].

While optimization is mostly used for point estimates, recently it is also used as a means for statistical inference in large scale machine learning [37, 14, 48, 24]. This manuscript follows this path: we propose an inference framework that uses stochastic gradients to approximate second-order, Newton steps. This is enabled by the fact that we only need to compute Hessian-vector products; in math, this can be approximated using $\nabla^2 f(\theta)v \approx \frac{\nabla f(\theta+\delta v)-\nabla f(\theta)}{\delta}$, where $f$ is the objective function, and $\nabla f$, $\nabla^2 f$ denote the gradient and Hessian of $f$. Our method can be interpreted as a generalization of the SVRG approach in optimization [34] (Appendix D); further, it is related to other stochastic Newton methods (e.g. [3]) when $\delta \to 0$. We defer the reader to Section 5 for more details. In this work, we apply our algorithm to unregularized $M$-estimation, and we use a similar approach, with proximal approximate Newton steps, in high-dimensional linear regression.

Our contributions can be summarized as follows; a more detailed discussion is deferred to Section 5:

- ❑ For the case of unregularized $M$-estimation, our method efficiently computes the statistical error covariance, useful for confidence intervals and p-values. Compared to state of the art, our scheme ***(i)*** guarantees consistency of computing the statistical error covariance, ***(ii)*** exploits better the available information (without wasting computational resources to compute quantities that are thereafter discarded), and ***(iii)*** converges to the optimum (without swaying around it).
- ❑ For high-dimensional linear regression, we propose a different estimator (see (13)) than the current literature. It is the result of a different optimization problem that is strongly convex with high probability. This permits the use of linearly convergent proximal algorithms [61, 36] towards the optimum; in contrast, state of the art only guarantees convergence to a neighborhood of the LASSO solution within statistical error. Our model also does not assume that absolute values of the true parameter's non-zero entries are lower bounded.
- ❑ The effectiveness of our framework goes even beyond convexity. As a highlight, we show that our work can be used to detect certain adversarial attacks on neural networks.

## 2 Unregularized $M$-estimation

In unregularized, low-dimensional $M$-estimation problems, we estimate a parameter of interest:

$$\theta^\star = \arg\min_{\theta\in\mathbb{R}^p} \mathbb{E}_{X\sim P}\left[\ell(X;\theta)\right], \quad \text{where } P(X) \text{ is the data distribution,}$$

using *empirical risk minimization* (ERM) on $n > p$ i.i.d. data points $\{X_i\}_{i=1}^n$:

$$\widehat{\theta} = \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \ell(X_i;\theta).$$

Statistical inference, such as computing one-dimensional confidence intervals, gives us information beyond the point estimate $\widehat{\theta}$, when $\widehat{\theta}$ has an asymptotic limit distribution [58]. *E.g.*, under regularity conditions, the $M$-estimator satisfies asymptotic normality [54, Theorem 5.21]. *I.e.*, $\sqrt{n}(\widehat{\theta} - \theta^\star)$ weakly converges to a normal distribution:

$$\sqrt{n}\left(\widehat{\theta} - \theta^\star\right) \to \mathcal{N}\left(0, H^{\star-1} G^\star H^{\star-1}\right),$$

where $H^\star = \mathbb{E}_{X\sim P}[\nabla^2_\theta \ell(X;\theta^\star)]$ and $G^\star = \mathbb{E}_{X\sim P}[\nabla_\theta \ell(X;\theta^\star)\, \nabla_\theta \ell(X;\theta^\star)^\top]$. We can perform statistical inference when we have a good estimate of $H^{\star-1} G^\star H^{\star-1}$. In this work, we use the plug-in covariance estimator $\widehat{H}^{-1}\widehat{G}\widehat{H}^{-1}$ for $H^{\star-1} G^\star H^{\star-1}$, where:

$$\widehat{H} = \tfrac{1}{n}\sum_{i=1}^n \nabla^2_\theta \ell(X_i;\widehat{\theta}), \quad \text{and} \quad \widehat{G} = \tfrac{1}{n}\sum_{i=1}^n \nabla_\theta \ell(X_i;\widehat{\theta})\, \nabla_\theta \ell(X_i;\widehat{\theta})^\top.$$

Observe that, in the naive case of directly computing $\widehat{G}$ and $\widehat{H}^{-1}$, we require both high computational- and space-complexity. Here, instead, we utilize approximate stochastic Newton motions from first order information to compute the quantity $\widehat{H}^{-1}\widehat{G}\widehat{H}^{-1}$.

## 2.1 Statistical inference with approximate Newton steps using only stochastic gradients

Based on the above, we are interested in solving the following $p$-dimensional optimization problem:

$$\widehat{\theta} = \arg\min_{\theta\in\mathbb{R}^p} f(\theta) := \tfrac{1}{n}\sum_{i=1}^n f_i(\theta), \quad \text{where } f_i(\theta) = \ell(X_i;\theta).$$

Notice that $\widehat{H}^{-1}\widehat{G}\widehat{H}^{-1}$ can be written as $\frac{1}{n}\sum_{i=1}^n \left(\widehat{H}^{-1}\nabla_\theta \ell(X_i;\widehat{\theta})\right)\left(\widehat{H}^{-1}\nabla_\theta \ell(X_i;\widehat{\theta})\right)^\top$, which can be interpreted as the covariance of stochastic –inverse-Hessian conditioned– gradients at $\widehat{\theta}$. Thus, the covariance of stochastic Newton steps can be used for statistical inference.

Algorithm 1 approximates each stochastic Newton $\widehat{H}^{-1}\nabla_\theta \ell(X_i;\widehat{\theta})$ step using only first order information. We start from $\theta_0$ which is sufficiently close to $\widehat{\theta}$, which can be effectively achieved using SVRG [34]; a description of the SVRG algorithm can be found in Appendix D. Lines 4, 5 compute a stochastic gradient whose covariance is used as part of statistical inference. Lines 6 to 12 use SGD to solve the Newton step,

$$\min_{g\in\mathbb{R}^p} \left\langle \tfrac{1}{S_o}\sum_{i\in I_o} \nabla f_i(\theta_t), g \right\rangle + \tfrac{1}{2\rho_t}\left\langle g, \nabla^2 f(\theta_t) g\right\rangle, \tag{1}$$

**Algorithm 1** Unregularized M-estimation statistical inference

1: **Parameters:** $S_o, S_i \in \mathbb{Z}_+$; $\rho_0, \tau_0 \in \mathbb{R}_+$; $d_o, d_i \in \left(\frac{1}{2}, 1\right)$ **Initial state:** $\theta_0 \in \mathbb{R}^p$

2: **for** $t = 0$ to $T-1$ **do** // approximate stochastic Newton descent

3: $\quad \rho_t \leftarrow \rho_0 (t+1)^{-d_0}$

4: $\quad I_o \leftarrow$ uniformly sample $S_o$ indices with replacement from $[n]$

5: $\quad g_t^0 \leftarrow -\rho_t \left( \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\theta_t) \right)$

6: $\quad$ **for** $j = 0$ to $L-1$ **do** // solving (1) approximately using SGD

7: $\quad\quad \tau_j \leftarrow \tau_0 (j+1)^{-d_i}$ and $\delta_t^j \leftarrow O(\rho_t^4 \tau_j^4)$

8: $\quad\quad I_i \leftarrow$ uniformly sample $S_i$ indices without replacement from $[n]$

9: $\quad\quad g_t^{j+1} \leftarrow g_t^j - \tau_j \left( \frac{1}{S_i} \sum_{k \in I_i} \frac{\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)}{\delta_t^j} \right) + \tau_j g_t^0$

10: $\quad$ **end for**

11: $\quad$ Use $\sqrt{S_o} \cdot \frac{\bar{g}_t}{\rho_t}$ for statistical inference, where $\bar{g}_t = \frac{1}{L+1} \sum_{j=0}^{L} g_t^j$

12: $\quad \theta_{t+1} \leftarrow \theta_t + g_t^L$

13: **end for**

which can be seen as a generalization of SVRG; this relationship is described in more detail in Appendix D. In particular, these lines correspond to solving (1) using SGD by uniformly sampling a random $f_i$, and approximating:

$$\nabla^2 f(\theta) g \approx \frac{\nabla f(\theta + \delta_t^j g) - \nabla f(\theta)}{\delta_t^j} = \mathbb{E}\left[ \frac{\nabla f_i(\theta + \delta_t^j g) - \nabla f_i(\theta)}{\delta_t^j} \mid \theta \right]. \tag{2}$$

Finally, the outer loop (lines 2 to 13) can be viewed as solving inverse Hessian conditioned stochastic gradient descent, similar to stochastic natural gradient descent [4].

In terms of parameters, similar to [43, 46], we use a decaying step size in Line 8 to control the error of approximating $H^{-1} g$. We set $\delta_t^j = O(\rho_t^4 \tau_j^4)$ to control the error of approximating Hessian vector product using a finite difference of gradients, so that it is smaller than the error of approximating $H^{-1} g$ using stochastic approximation. For similar reasons, we use a decaying step size in the outer loop to control the optimization error.

The following theorem characterizes the behavior of Algorithm 1.

**Theorem 1.** *For a twice continuously differentiable and convex function $f(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)$ where each $f_i$ is also convex and twice continuously differentiable, assume $f$ satisfies*

- ❑ *strong convexity: $\forall \theta_1, \theta_2$, $f(\theta_2) \geq f(\theta_1) + \langle \nabla f(\theta_1), \theta_2 - \theta_1 \rangle + \frac{1}{2} \alpha \|\theta_2 - \theta_1\|_2^2$;*
- ❑ *$\forall \theta$, each $\|\nabla^2 f_i(\theta)\|_2 \leq \beta_i$, which implies that $f_i$ has Lipschitz gradient: $\forall \theta_1, \theta_2$, $\|\nabla f_i(\theta_1) - \nabla f_i(\theta_2)\|_2 \leq \beta_i \|\theta_1 - \theta_2\|_2$;*
- ❑ *each $\nabla^2 f_i$ is Lipschitz continuous: $\forall \theta_1, \theta_2$, $\|\nabla^2 f_i(\theta_2) - \nabla^2 f_i(\theta_1)\|_2 \leq h_i \|\theta_2 - \theta_1\|_2$.*

*In Algorithm 1, we assume that batch sizes $S_o$—in the outer loop—and $S_i$—in the inner loops—are $O(1)$. The outer loop step size is*

$$\rho_t = \rho_0 \cdot (t+1)^{-d_o}, \quad \textit{where } d_o \in \left(\tfrac{1}{2}, 1\right) \textit{ is the decaying rate.} \tag{3}$$

*In each outer loop, the inner loop step size is*

$$\tau_j = \tau_0 \cdot (j+1)^{-d_i}, \quad \textit{where } d_i \in \left(\tfrac{1}{2}, 1\right) \textit{ is the decaying rate.} \tag{4}$$

*The scaling constant for Hessian vector product approximation is*

$$\delta_t^j = \delta_0 \cdot \rho_t^4 \cdot \tau_j^4 = o\left(\tfrac{1}{(t+1)^2(j+1)^2}\right). \tag{5}$$

*Then, for the outer iterate $\theta_t$ we have*

$$\mathbb{E}\left[\|\theta_t - \widehat{\theta}\|_2^2\right] \lesssim t^{-d_o}, \tag{6}$$ *and* $$\mathbb{E}\left[\|\theta_t - \widehat{\theta}\|_2^4\right] \lesssim t^{-2d_o}. \tag{7}$$

*In each outer loop, after $L$ steps of the inner loop, we have:*

$$\mathbb{E}\left[\left\|\tfrac{\bar{g}_t}{\rho_t} - [\nabla^2 f(\theta_t)]^{-1} g_t^0\right\|_2^2 \mid \theta_t\right] \lesssim \tfrac{1}{L}\left\|g_t^0\right\|_2^2, \tag{8}$$

*and at each step of the inner loop, we have:*

$$\mathbb{E}\left[\left\|g_t^{j+1} - [\nabla^2 f(\theta_t)]^{-1} g_t^0\right\|_2^4 \mid \theta_t\right] \lesssim (j+1)^{-2d_i}\left\|g_t^0\right\|_2^4. \tag{9}$$

*After $T$ steps of the outer loop, we have a non-asymptotic bound on the "covariance":*

$$\mathbb{E}\left[\left\|H^{-1}GH^{-1} - \tfrac{S_o}{T}\sum_{t=1}^{T}\tfrac{\bar{g}_t\bar{g}_t^\top}{\rho_t^2}\right\|_2\right] \lesssim T^{-\frac{d_o}{2}} + L^{-\frac{1}{2}}, \tag{10}$$

*where $H = \nabla^2 f(\widehat{\theta})$ and $G = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\widehat{\theta})\nabla f_i(\widehat{\theta})^\top$.*

Some comments on the results in Theorem 1. The main outcome is that (10) provides a non-asymptotic bound and consistency guarantee for computing the estimator covariance using Algorithm 1. This is based on the bound for approximating the inverse-Hessian conditioned stochastic gradient in (8), and the optimization bound in (6). As a side note, the rates in Theorem 1 are very similar to classic results in stochastic approximation [43, 46]; however the nested structure of outer and inner loops is different from standard stochastic approximation algorithms. Heuristically, calibration methods for parameter tuning in subsampling methods ([42], Ch. 9) can be used for hyper-parameter tuning in our algorithm.

In Algorithm 1, $\{\bar{g}_t/\rho_t\}_{i=1}^n$ does not have asymptotic normality. *I.e.*, $\frac{1}{\sqrt{T}}\sum_{t=1}^T \frac{\bar{g}_t}{\rho_t}$ does not weakly converge to $\mathcal{N}\left(0, \frac{1}{S_o}H^{-1}GH^{-1}\right)$; we give an example using mean estimation in Appendix C.1. For a similar algorithm based on SVRG

(Algorithm 5 in Appendix C), we show that we have asymptotic normality and improved bounds for the "covariance"; however, this requires a full gradient evaluation in each outer loop. In Appendix B, we present corollaries for the case where the iterations in the inner loop increase, as the counter in the outer loop increases (*i.e.*, $(L)_t$ is an increasing series). This guarantees consistency (convergence of the covariance estimate to $H^{-1}GH^{-1}$), although it is less efficient than using a constant number of inner loop iterations. Our procedure also serves as a general and flexible framework for using different stochastic gradient optimization algorithms [50, 28, 38, 16] in the inner and outer loop parts.

Finally, we present the following corollary that states that the average of consecutive iterates, in the outer loop, has asymptotic normality, similar to [43, 46].

**Corollary 1.** *In Algorithm 1's outer loop, the average of consecutive iterates satisfies*

$$\mathbb{E}\left[\left\|\frac{\sum_{t=1}^{T}\theta_t}{T}-\widehat{\theta}\right\|_2^2\right]\lesssim\frac{1}{T},\quad(11)\quad\textit{and}\quad\frac{1}{\sqrt{T}}\left(\frac{\sum_{t=1}^{T}\theta_t}{T}-\widehat{\theta}\right)=W+\Delta,\quad(12)$$

*where $W$ weakly converges to $\mathcal{N}(0,\frac{1}{S_o}H^{-1}GH^{-1})$, and $\Delta=o_P(1)$ when $T\to\infty$ and $L\to\infty$ $\left(\mathbb{E}[\|\Delta\|_2^2]\lesssim T^{1-2d_o}+T^{d_o-1}+\frac{1}{L}\right)$.*

Corollary 1 uses $2^{\text{nd}}$ , $4^{\text{th}}$ moment bounds on individual iterates (eqs. (6), (7) in the above theorem), and the approximation of inverse Hessian conditioned stochastic gradient in (9).

# 3 High dimensional LASSO linear regression

In this section, we focus on the case of high-dimensional linear regression. Statistical inference in such settings, where $p \gg n$, is arguably a more difficult task: the bias introduced by the regularizer is of the same order with the estimator's variance. Recent works [63, 53, 32] propose statistical inference via de-biased LASSO estimators. Here, we present a new $\ell_1$-norm regularized objective and propose an approximate stochastic *proximal* Newton algorithm, using only first order information.

We consider the linear model $y_i = \langle \theta^\star, x_i \rangle + \epsilon_i$, for some sparse $\theta^\star \in \mathbb{R}^p$. For each sample, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is i.i.d. noise. And each data point $x_i \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^p$.

- ❏ *Assumptions on $\theta$:* ***(i)*** $\theta^\star$ is $s$-sparse; ***(ii)*** $\|\theta^\star\|_2 = O(1)$, which implies that $\|\theta^\star\|_1 \lesssim \sqrt{s}$.
- ❏ *Assumptions on $\Sigma$:* ***(i)*** $\Sigma$ is sparse, where each column (and row) has at most $b$ non-zero entries;[1] ***(ii)*** $\Sigma$ is well conditioned: all of $\Sigma$'s eigenvalues are $\Theta(1)$;

[1]This is satisfied when $\Sigma$ is block diagonal or banded. Covariance estimation under this sparsity assumption has been extensively studied [7, 8, 13], and soft thresholding is an effective yet simple estimation method [45].

***(iii)*** $\Sigma$ is diagonally dominant ($\Sigma_{ii} - \sum_{j \neq i} |\Sigma_{ij}| \geq D_\Sigma > 0$ for all $1 \leq i \leq p$), and this will be used to bound the $\ell_\infty$ norm of $\widehat{S}^{-1}$ [55]. A commonly used design covariance that satisfies all of our assumptions is $I$.

We estimate $\theta^\star$ using:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \tfrac{1}{2} \left\langle \theta, \left( \widehat{S} - \tfrac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \theta \right\rangle + \tfrac{1}{n} \sum_{i=1}^n \tfrac{1}{2} \left( x_i^\top \theta - y_i \right)^2 + \lambda \|\theta\|_1, \quad (13)$$

where $\widehat{S}_{jk} = \operatorname{sign}\left( \left( \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)_{jk} \right) \left( \left| \left( \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)_{jk} \right| - \omega \right)_+$ is an estimate of $\Sigma$ by soft-thresholding each element of $\frac{1}{n} \sum_{i=1}^n x_i x_i^\top$ with $\omega = \Theta\left( \sqrt{\frac{\log p}{n}} \right)$ [45]. Under our assumptions, $\widehat{S}$ is positive definite with high probability when $n \gg b^2 \log p$ (Lemma 4), and this guarantees that the optimization problem (13) is well defined. *I.e.*, we replace the degenerate Hessian in regular LASSO regression with an estimate, which is positive definite with high probability under our assumptions.

We set the regularization parameter

$$\lambda = \Theta\left( (\sigma + \|\theta^\star\|_1) \sqrt{\tfrac{\log p}{n}} \right),$$

which is similar to LASSO regression [12, 41] and related estimators using thresholded covariance [62, 33].

**Point estimate.** Theorem 2 provides guarantees for our proposed point estimate (13).

**Theorem 2.** *When $n \gg b^2 \log p$, the solution $\widehat{\theta}$ in (13) satisfies*

$$\left\| \widehat{\theta} - \theta^\star \right\|_1 \lesssim s \left( \sigma + \|\theta^\star\|_1 \right) \sqrt{\tfrac{\log p}{n}} \lesssim s \left( \sigma + \sqrt{s} \right) \sqrt{\tfrac{\log p}{n}}, \quad (14)$$

$$\left\| \widehat{\theta} - \theta^\star \right\|_2 \lesssim \sqrt{s} \left( \sigma + \|\theta^\star\|_1 \right) \sqrt{\tfrac{\log p}{n}} \lesssim \sqrt{s} \left( \sigma + \sqrt{s} \right) \sqrt{\tfrac{\log p}{n}}, \quad (15)$$

*with probability at least $1 - p^{-\Theta(1)}$.*

**Confidence intervals.** We next present a de-biased estimator $\widehat{\theta}^{\mathrm{d}}$ (16), based on our proposed estimator. $\widehat{\theta}^{\mathrm{d}}$ can be used to compute confidence intervals and p-values for each coordinate of $\widehat{\theta}^{\mathrm{d}}$, which can be used for false discovery rate control [30]. The estimator satisfies:

$$\widehat{\theta}^{\mathrm{d}} = \widehat{\theta} + \widehat{S}^{-1} \left[ \tfrac{1}{n} \sum_{i=1}^n \left( y_i - x_i^\top \widehat{\theta} \right) x_i \right]. \quad (16)$$

Our de-biased estimator is similar to [63, 53, 31, 32]. however, we have different terms, since we need to de-bias covariance estimation. Our estimator

assumes $n \gg b^2 \log p$, since then $\widehat{S}$ is positive definite with high probability (Lemma 4). The assumption that $\Sigma$ is diagonally dominant guarantees that the $\ell_\infty$ norm $\|\widehat{S}^{-1}\|_\infty$ is bounded by $O\left(\frac{1}{D_\Sigma}\right)$ with high probability when $n \gg \frac{1}{D_\Sigma{}^2} \log p$.

Theorem 3 shows that we can compute valid confidence intervals for each coordinate when $n \gg (\frac{1}{D_\Sigma} s\,(\sigma + \|\theta^\star\|_1) \log p)^2$. This is satisfied when $n \gg (\frac{1}{D_\Sigma} s\,(\sigma + \sqrt{s}) \log p)^2$. And the covariance is similar to the sandwich estimator [29, 59].

**Theorem 3.** *Under our assumptions, when $n \gg \max\{b^2, \frac{1}{D_\Sigma{}^2}\} \log p$, we have:*

$$\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^\star) = Z + R, \tag{17}$$

*where the conditional distribution satisfies $Z \mid \{x_i\}_{i=1}^n \sim \mathcal{N}\left(0, \sigma^2 \cdot \left[\widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\widehat{S}^{-1}\right]\right)$, and $\|R\|_\infty \lesssim \frac{1}{D_\Sigma} s\,(\sigma + \|\theta^\star\|_1)\frac{\log p}{\sqrt{n}} \lesssim \frac{1}{D_\Sigma} s\,(\sigma + \sqrt{s})\frac{\log p}{\sqrt{n}}$ with probability at least $1 - p^{-\Theta(1)}$.*

Our estimate in (13) has similar error rates to the estimator in [62]; however, no confidence interval guarantees are provided, and the estimator is based on inverting a large covariance matrix. Further, although it does not match minimax rates achieved by regular LASSO regression [44], and the sample complexity in Theorem 3 is slightly higher than other methods [53, 31, 32], our criterion is strongly convex with high probability: this allows us to use linearly convergent proximal algorithms [61, 36], whereas provable linearly convergent optimization bounds for LASSO only guarantees convergence to a neighborhood of the LASSO solution within statistical error [1]. This is crucial for computing the de-biased estimator, as we need the optimization error to be much less than the statistical error.

In Appendix A, we present our algorithm for statistical inference in high dimensional linear regression using stochastic gradients. It estimates the statistical error covariance using the plug-in estimator:

$$\widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^n (x_i^\top \widehat{\theta} - y_i)^2 x_i x_i^\top\right)\widehat{S}^{-1},$$

which is related to the empirical sandwich estimator [29, 59]. Algorithm 2 computes the statistical error covariance. Similar to Algorithm 1, Algorithm 2 has an outer loop part and an inner loop part, where the outer loops correspond to approximate proximal Newton steps, and the inner loops solve each proximal Newton step using proximal SVRG [61]. To control the variance, we use SVRG and proximal SVRG to solve the Newton steps. This is because in the high dimensional setting, the variance is too large when we use SGD [40] and proximal SGD [5] for solving Newton steps. However, since we have $p \gg n$ , instead of sampling *by sample*, we sample *by feature*. When we set $L_o^t = \Theta(\log(p) \cdot \log(t))$, we can estimate the statistical error covariance with

element-wise error less than $O\left(\frac{\max\{1,\sigma\}\text{polylog}(n,p)}{\sqrt{T}}\right)$ with high probability, using $O\left(T \cdot n \cdot p^2 \cdot \log(p) \cdot \log(T)\right)$ numerical operations. And Algorithm 3 calculates the de-biased estimator $\widehat{\theta}^{\text{d}}$ (16) via SVRG. For more details, we defer the reader to the appendix.

# 4 Experiments

## 4.1 Synthetic data

The coverage probability is defined as $\frac{1}{p}\sum_{i=1}^{p}\mathbb{P}[\theta_i^\star \in \hat{C}_i]$, where $\hat{C}_i$ is the estimated confidence interval for the $i^{\text{th}}$ coordinate. The average confidence interval length is defined as $\frac{1}{p}\sum_{i=1}^{p}(\hat{C}_i^u - \hat{C}_i^l)$, where $[\hat{C}_i^l, \hat{C}_i^u]$ is the estimated confidence interval for the $i^{\text{th}}$ coordinate. In our experiments, coverage probability and average confidence interval length are estimated through simulation. Result given as a pair $(\alpha, \beta)$ indicates (coverage probability, confidence interval length).

| | Approximate Newton | Bootstrap | Inverse Fisher information | Averaged SGD |
|---|---|---|---|---|
| Lin1 | (0.906, 0.289) | (0.933, 0.294) | (0.918, 0.274) | (0.458, 0.094) |
| Lin2 | (0.915, 0.321) | (0.942, 0.332) | (0.921,0.308) | (0.455 0.103) |

(a) Linear regression

| | Approximate Newton | Jackknife | Inverse Fisher information | Averaged SGD |
|---|---|---|---|---|
| Log1 | (0.902, 0.840) | (0.966 1.018) | (0.938, 0.892) | (0.075 0.044) |
| Log2 | (0.925, 1.006) | (0.979, 1.167) | (0.948, 1.025) | (0.065 0.045) |

(b) Logistic regression

Table 1: Synthetic data average coverage & confidence interval length for low dimensional problems.

**Low dimensional problems.** Table 1 shows 95% confidence interval's coverage and length of 200 simulations for linear and logistic regression. The exact configurations for linear/logistic regression examples are provided in Appendix G.1.1. Compared with Bootstrap and Jackknife [22], Algorithm 1 uses less numerical operations, while achieving similar results. Compared with the averaged SGD method [37, 14], our algorithm performs much better, while using the same amount of computation, and is much less sensitive to the choice hyper-parameters.

**High dimensional linear regression.** We use 600 i.i.d. samples from a model with $\Sigma = I$, $\sigma = 0.7$, $\theta^\star = [1/\sqrt{8}, \cdots, 1/\sqrt{8}, 0, \cdots, 0]^\top \in \mathbb{R}^{1000}$ which is 8-sparse. Figure 1 shows 95% confidence intervals for the first 20 coordinates. The average confidence interval length is 0.14

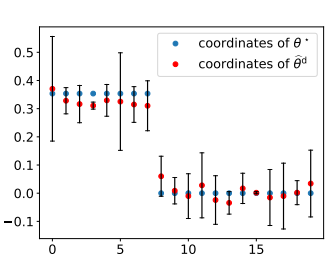



Figure 1: 95% confidence intervals

and average coverage is 0.83. Additional experimental results, including p-value distribution, are presented in Appendix G.1.2.

## 4.2 Real data

**Neural network adversarial attack detection.** Here we use ideas from statistical inference to detect certain adversarial attacks on neural networks. A key observation is that neural networks are effective at representing low dimensional manifolds such as natural images [6, 15], and this causes the risk function's Hessian to be degenerate [47]. From a statistical inference perspective, we interpret this as meaning that the confidence intervals in the null space of $H^+GH^+$ is infinity, where $H^+$ is the pseudo-inverse of the Hessian (see Section 2). When we make a prediction $\Psi(x;\widehat{\theta})$ using a fixed data point $x$ as input (i.e., conditioned on $x$), using the delta method [54], the confidence interval of the prediction can be derived from the asymptotic normality of $\Psi(x;\widehat{\theta})$

$$\sqrt{n}\left(\Psi(x;\widehat{\theta})-\Psi(x;\theta^\star)\right)\to\mathcal{N}\left(0,\nabla_\theta\Psi(x;\widehat{\theta})^\top\left[\widehat{H}^{-1}\widehat{G}\widehat{H}^{-1}\right]\nabla_\theta\Psi(x;\widehat{\theta})\right).$$

To detect adversarial attacks, we use the score

$$\frac{\left\|\left(I-P_{H^+GH^+}\right)\nabla_\theta\Psi(x;\widehat{\theta})\right\|_2}{\left\|\nabla_\theta\Psi(x;\widehat{\theta})\right\|_2},$$

to measure how much $\nabla_\theta\Psi(x;\widehat{\theta})$ lies in null space of $H^+GH^+$, where $P_{H^+GH^+}$ is the projection matrix onto the range of $H^+GH^+$. Conceptually, for the same image, the randomly perturbed image's score should be larger than the original image's score, and the adversarial image's score should be larger than the randomly perturbed image's score.

We train a binary classification neural network with 1 hidden layer and softplus activation function, to distinguish between "Shirt" and "T-shirt/top" in the Fashion MNIST data set [60]. Figure 2 shows distributions of scores of original images, adversarial images generated using the fast gradient sign method [27], and randomly perturbed images. Adversarial and random perturbations have the same $\ell_\infty$ norm. The adversarial perturbations and example images are shown in Appendix G.2.1. Although the scores' values are small, they are still significantly larger than 64-bit floating point precision ($2^{-53}\approx 1.11\times 10^{-16}$). We observe that scores of randomly perturbed images is an order of magnitude larger than scores of original images, and scores of adversarial images is an order of magnitude larger than scores of randomly perturbed images.

**High dimensional linear regression.** We apply our high dimensional linear regression statistical inference procedure to a high-throughput genomic data set concerning riboflavin (vitamin B2) production rate [11], which contains $n = 71$ samples of $p = 4088$ genes. We set $\lambda = 4.260$ and $\omega = 0.5$. In Appendix G.2.2, we show that our point estimate is similar to the vanilla LASSO estimate, and compare our statistical inference results with those of [31, 11, 10, 39].

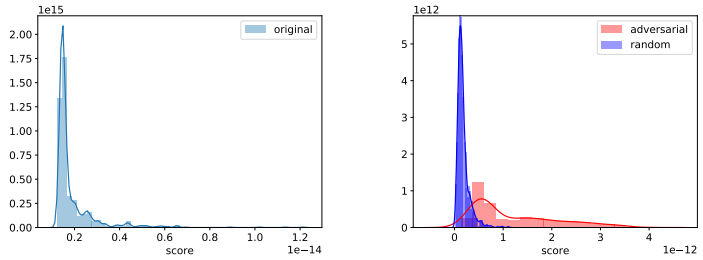



Figure 2: Distribution of scores for original, randomly perturbed, and adversarially perturbed images

# 5 Related work

**Unregularized M-estimation.** This work provides a general, flexible framework for *simultaneous* point estimation and statistical inference, and improves upon previous methods, based on averaged stochastic gradient descent [37, 14].

Compared to [14] (and similar works [48, 24] using SGD with decreasing step size), our method does not need to increase the lengths of "segments" (inner loops) to reduce correlations between different "replicates". Even in that case, if we use $T$ replicates and increasing "segment" length (number of inner loops is $t^{\frac{d_o}{1-d_o}} \cdot L$) with a total of $O(T^{\frac{1}{1-d_o}} \cdot L)$ stochastic gradient steps, [14] guarantees $O(L^{-\frac{1-d_o}{2}} + T^{-\frac{1}{2}} + T^{\max\{\frac{1}{2}-\frac{d_o}{4(1-d_o)},0\}-\frac{1}{2}} \cdot L^{-\frac{d_o}{4}} + T^{\max\{\frac{1-2d_o}{2(1-d_o)},0\}-\frac{1}{2}} \cdot L^{\frac{1-2d_o}{2}})$ , whereas our method guarantees $O(T^{-\frac{d_o}{2}})$. Further, [14] is inconsistent, whereas our scheme guarantees consistency of computing the statistical error covariance.

[37] uses fixed step size SGD for statistical inference, and discards iterates between different "segments" to reduce correlation, whereas we do not discard any iterates in our computations. Although [37] states empirically constant step SGD performs well in statistical inference, it has been empirically shown [17] that averaging consecutive iterates in constant step SGD does not guarantee convergence to the optimal – the average will be "wobbling" around the optimal, whereas decreasing step size stochastic approximation methods ([43, 46] and our work) will converge to the optimal, and averaging consecutive iterates guarantees "fast" rates.

Finally, from an optimization perspective, our method is similar to stochastic Newton methods (e.g. [3]); however, our method only uses first-order information to approximate a Hessian vector product ($\nabla^2 f(\theta)v \approx \frac{\nabla f(\theta+\delta v)-\nabla f(\theta)}{\delta}$). Algorithm 1's outer loops are similar to stochastic natural gradient descent [4]. Also, we demonstrate an intuitive view of SVRG [34] as a special case of approximate stochastic Newton steps using first order information (Appendix D).

**High dimensional linear regression.** [14]'s high dimensional inference algorithm is based on [2], and only guarantees that optimization error is at the same scale as the statistical error. However, proper de-biasing of the LASSO estimator

requires the optimization error to be much less than the statistical error, otherwise the optimization error introduces additional bias that de-biasing cannot handle. Our optimization objective is strongly convex with high probability: this permits the use of linearly convergent proximal algorithms [61, 36] towards the optimum, which guarantees the optimization error to be much smaller than the statistical error.

Our method of de-biasing the LASSO in Section 3 is similar to [63, 53, 31, 32]. Our method uses a new $\ell_1$ regularized objective (13) for high dimensional linear regression, and we have different de-biasing terms, because we also need to de-bias the covariance estimation.

In Algorithm 2, our covariance estimate is similar to the classic *sandwich estimator* [29, 59]. Previous methods require $O(p^2)$ space which unsuitable for large scale problems, whereas our method only requires $O(p)$ space.

Similar to our $\ell_1$-norm regularized objective, [62, 33] shows similar point estimate statistical guarantees for related estimators; however there are no confidence interval results. Further, although [62] is an elementary estimator in closed form, it still requires computing the inverse of the thresholded covariance, which is challenging in high dimensions, and may not computationally outperform optimization approaches.

Finally, for feature selection, we do not assume that absolute values of the true parameter's non-zero entries are lower bounded. [23, 56, 12].

# References


[1] Alekh Agarwal, Sahand Negahban, and Martin Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

[2] 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*, pages 1538–1546, 2012.

[3] Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1):4148–4187, 2017.

[4] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[5] Yves F. Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *J. Mach. Learn. Res*, 18(1):310–342, 2017.

[6] Ronen Basri and David Jacobs. Efficient representation of low-dimensional manifolds using deep networks. *arXiv preprint arXiv:1602.04723*, 2016.

[7] Peter Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604, 2008.

[8] Peter Bickel, Ya'acov Ritov, and Alexandre Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[9] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[10] Peter Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, 2013.

[11] Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.

[12] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[13] Tony Cai and Harrison Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, pages 2389–2420, 2012.

[14] Xi Chen, Jason Lee, Xin Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*, 2016.

[15] Charles K Chui and Hrushikesh Narhar Mhaskar. Deep nets for local manifold learning. *arXiv preprint arXiv:1607.07110*, 2016.

[16] Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Starting small-learning with adaptive sample sizes. In *International conference on machine learning*, pages 1463–1471, 2016.

[17] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains. *arXiv preprint arXiv:1707.06386*, 2017.

[18] Jürgen Dippon. Asymptotic expansions of the Robbins-Monro process. *Mathematical Methods of Statistics*, 17(2):138–145, 2008.

[19] Jürgen Dippon. Edgeworth expansions for stochastic approximation theory. *Mathematical Methods of Statistics*, 17(1):44–65, 2008.

[20] Francis Ysidro Edgeworth. On the probable errors of frequency-constants. *Journal of the Royal Statistical Society*, 71(2):381–397, 1908.

[21] B. Efron and T. Hastie. *Computer age statistical inference*. Cambridge University Press, 2016.

[22] Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[23] Jianqing Fan, Wenyan Gong, Chris Junchi Li, and Qiang Sun. Statistical sparse online regression: A diffusion approximation perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 1017–1026, 2018.

[24] Yixin Fang, Jinfeng Xu, and Lei Yang. On Scalable Inference with Stochastic Gradient Descent. *arXiv preprint arXiv:1707.00192*, 2017.

[25] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[26] Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. *arXiv preprint arXiv:1709.03342*, 2017.

[27] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[28] Reza Harikandeh, Mohamed Osama Ahmed, Alim Virani, Mark Schmidt, Jakub Konecny, and Scott Sallinen. StopWasting My Gradients: Practical SVRG. In *Advances in Neural Information Processing Systems 28*, pages 2251–2259, 2015.

[29] Peter Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on Mathematical Statistics and Probability*, pages 221–233, Berkeley, CA, 1967. University of California Press.

[30] Adel Javanmard and Hamid Javadi. False Discovery Rate Control via Debiased Lasso. *arXiv preprint arXiv:1803.04464*, 2018.

[31] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[32] Adel Javanmard and Andrea Montanari. De-biasing the Lasso: Optimal sample size for Gaussian designs. *arXiv preprint arXiv:1508.02757*, 2015.

[33] Jessie Jeng and John Daye. Sparse covariance thresholding for high-dimensional variable selection. *Statistica Sinica*, pages 625–657, 2011.

[34] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[35] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

[36] Jason Lee, Yuekai Sun, and Michael Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

[37] Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis. Statistical inference using SGD. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[38] Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015.

[39] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. $p$-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

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

[41] Sahand Negahban, Pradeep Ravikumar, Martin Wainwright, and Bin Yu. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, pages 538–557, 2012.

[42] D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer Series in Statistics. Springer New York, 2012.

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

[44] Garvesh Raskutti, Martin Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

[45] Adam Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

[46] David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[47] Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical Analysis of the Hessian of Over-Parametrized Neural Networks. *arXiv preprint arXiv:1706.04454*, 2017.

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

[49] R. Tibshirani, M. Wainwright, and T. Hastie. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

[50] Panos Toulis and Edoardo M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

[51] Joel Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[52] F. Tuerlinckx, F. Rijmen, G. Verbeke, and P. Boeck. Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59(2):225–255, 2006.

[53] Sara van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[54] Aad W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

[55] James Varah. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11(1):3–5, 1975.

[56] Martin Wainwright. Sharp thresholds for High-Dimensional and noisy sparsity recovery using $\ell_1$-Constrained Quadratic Programming (Lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[57] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. To appear, 2017.

[58] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[59] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838, 1980.

[60] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[61] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[62] Eunho Yang, Aurelie Lozano, and Pradeep Ravikumar. Elementary estimators for high-dimensional linear regression. In *International Conference on Machine Learning*, pages 388–396, 2014.

[63] Cun-Hui Zhang and Stephanie Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

# A High dimensional linear regression statistical inference using stochastic gradients (Section 3)

## A.1 Statistical inference using approximate proximal Newton steps with stochastic gradients

Here, we present a statistical inference procedure for high dimensional linear regression via approximate proximal Newton steps using stochastic gradients. It uses the plug-in estimator:

$$\widehat{S}^{-1}\left(\tfrac{1}{n}\sum_{i=1}^{n}(x_i^\top\widehat{\theta}-y_i)^2x_ix_i^\top\right)\widehat{S}^{-1},$$

which is related to the empirical sandwich estimator [29, 59]. Lemma 1 shows this is a good estimate of the covariance when $n \gg \frac{1}{D\Sigma^4}\max\{1,\sigma^2\}s^2(\sigma+\|\theta^\star\|_1)^2$.

Algorithm 2 performs statistical inference in high dimensional linear regression (13), by computing the statistical error covariance in Theorem 3, based on the plug-in estimate in Lemma 1. We denote the soft thresholding of $A$ by $\omega$ as an element-wise procedure $\left(\mathbf{S}_\omega(A)\right)_e = \mathrm{sign}(A_e)(|A_e|-\omega)_+$. For a vector $v$, we write $v$'s $i^{\text{th}}$ coordinate as $v(i)$. The optimization objective (13) is denoted as:

$$\tfrac{1}{2}\theta^\top\left(\widehat{S}-\tfrac{1}{n}\textstyle\sum_{i=1}^{n}x_ix_i^\top\right)\theta+\tfrac{1}{n}\sum_{i=1}^{n}f_i,$$

where $f_i = \frac{1}{2}\left(x_i^\top - y_i\right)^2$. Further,

$$\boldsymbol{g}_{\widehat{S}}(v)=\nabla_v\left[\tfrac{1}{2}v^\top\widehat{S}v\right]=\widehat{S}v=\sum_{j=1}^{p}v(j)\cdot\mathbf{S}_\omega\left(\tfrac{1}{n}\sum_{i=1}^{n}\left[\nabla f_i(\theta+\boldsymbol{e}_j)-\nabla f_i(\theta)\right]\right),$$

where $\boldsymbol{e}_i \in \mathbb{R}^p$ is the basis vector where the $i^{\text{th}}$ coordinate is 1 and others are 0, and $\widehat{S}v$ is computed in a column-wise manner.

For point estimate optimization, the proximal Newton step [36] at $\theta$ solves the optimization problem

$$\min_{\Delta}\tfrac{1}{2\rho}\Delta^\top\widehat{S}\Delta+\left\langle(\widehat{S}-\tfrac{1}{n}\textstyle\sum_{i=1}^{n}x_ix_i^\top)\theta+\tfrac{1}{n}\sum_{i=1}^{n}\nabla f_i(\theta),\Delta\right\rangle+\lambda\|\theta+\Delta\|_1,$$

to determine a descent direction. For statistical inference, we solve a Newton step:

$$\min_{\Delta}\tfrac{1}{2\rho}\Delta^\top\widehat{S}\Delta+\left\langle\tfrac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\theta_t),\Delta\right\rangle$$

to compute $-\widehat{S}^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\theta)$, whose covariance is the statistical error covariance.

To control variance, we solve Newton steps using SVRG and proximal SVRG [61], because in the high dimensional setting, the variance using SGD [40] and

proximal SGD [5] for solving Newton steps is too large. However because $p \gg n$, instead of sampling *by sample*, we sample *by feature*. We start from $\theta_0$ sufficiently close to $\widehat{\theta}$ (see Theorem 4 for details), which can be effectively achieved using proximal SVRG (Appendix A.3). Line 7 corresponds to SVRG's outer loop part that computes the full gradient, and line 12 corresponds to SVRG's inner loop update. Line 8 corresponds to proximal SVRG's outer loop part that computes the full gradient, and line 13 corresponds to proximal SVRG's inner loop update.

The covariance estimate bound, asymptotic normality result, and choice of hyper-parameters are described in Appendix A.4. When $L_o^t = \Theta(\log(p) \cdot \log(t))$, we can estimate the covariance with element-wise error less than $O\left(\frac{\max\{1,\sigma\}\text{polylog}(n,p)}{\sqrt{T}}\right)$ with high probability, using $O\left(T \cdot n \cdot p^2 \cdot \log(p) \cdot \log(T)\right)$ numerical operations. Calculation of the de-biased estimator $\widehat{\theta}^{\mathrm{d}}$ (16) via SVRG is described in Appendix A.2.

**Algorithm 2** High dimensional linear regression statistical inference

1: **Parameters:** $S_o, S_i \in \mathbb{Z}_+$; $\eta, \tau \in \mathbb{R}_+$; **Initial state:** $\theta_0 \in \mathbb{R}^p$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $I_o \leftarrow$ uniformly sample $S_o$ indices with replacement from $[n]$
4: $g_t^0 \leftarrow -\frac{1}{S_o} \sum_{k \in I_o} \nabla f_k(\theta_t)$
5: $d_t^0 \leftarrow -\left(\boldsymbol{g}_{\widehat{S}}(\theta_t) - \frac{1}{n}\sum_{i=1}^n \left[\nabla f_i(\theta_t + \theta_t) - \nabla f_i(\theta_t)\right] + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\theta_t)\right)$
6: **for** $j = 1$ **to** $L_o^t -$ **do** // solving Newton steps using SVRG
7: $u_t^j \leftarrow \boldsymbol{g}_{\widehat{S}}(g_t^{j-1}) - g_t^0$
8: $v_t^j \leftarrow \boldsymbol{g}_{\widehat{S}}(d_t^{j-1}) - d_t^0$
9: $g_t^j \leftarrow g_t^{j-1}$, $d_t^j \leftarrow d_t^{j-1}$
10: **for** $l = 1$ **to** $L_i$ **do**
11: $I_i \leftarrow$ uniformly sample $S_i$ indices without replacement from $[p]$
12: $g_t^j \leftarrow g_t^j - \tau\left[u_t^j + \frac{p}{S_i}\sum_{k \in S_i}\left[g_t^j(k) - g_t^{j-1}(k)\right] \cdot \mathbf{S}_\omega\left(\nabla f_k(\theta_t + \boldsymbol{e}_k) - \nabla f_k(\theta_t)\right)\right]$
13: $d_t^j \leftarrow \mathbf{S}_{\eta\lambda}\left(d_t^j - \eta\left[v_t^j + \frac{p}{S_i}\sum_{k \in S_i}\left[d_t^j(k) - d_t^{j-1}(k)\right] \cdot \mathbf{S}_\omega\left(\nabla f_k(\theta_t + \boldsymbol{e}_k) - \nabla f_k(\theta_t)\right)\right]\right)$
14: **end for**
15: **end for**
16: Use $\sqrt{S_o} \cdot \frac{\bar{g}_t}{\rho_t}$ for statistical inference, where $\bar{g}_t = \frac{1}{L_o^t + 1}\sum_{j=0}^{L_o^t} g_t^j$
17: $\theta_{t+1} = \theta_t + \bar{d}_t$, where $\bar{d}_t = \frac{1}{L_o + 1}\sum_{j=0}^{L_o^t} d_t^j$ // point estimation (optimization)
18: **end for**

## A.2 Computing the de-biased estimator (16) via SVRG

To control variance, we solve each proximal Newton step using SVRG, in stead of SGD as in Algorithm 1. Because However because the number of features is much larger than the number of samples, instead of sampling *by sample*, we sample *by feature*.

The de-biased estimator is

$$\begin{aligned}\widehat{\theta}^{\mathrm{d}} =& \widehat{\theta} + \widehat{S}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\widehat{\theta}\right] \\ =& \widehat{\theta} + \widehat{S}^{-1}\left(-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\widehat{\theta})\right).\end{aligned}$$

And we compute $\widehat{S}^{-1}\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\widehat{\theta})$ using SVRG [34] by solving the following optimization problem using SVRG and sampling by feature

$$\min_u \frac{1}{2}u^\top \widehat{S} u + \left\langle \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\widehat{\theta}), u\right\rangle.$$

**Algorithm 3** Computing the de-biased estimator (16) via SVRG

**for** $i = 0$ **to** $L_o - 1$ **do**
  $d_i^0 \leftarrow -\eta[\boldsymbol{g}_{\widehat{S}}(u_i) + \frac{1}{n}\sum_{k=1}^{n}\nabla f_k(\widehat{\theta})]$
  **for** $j = 0$ **to** $L_i - 1$ **do**
    $I \leftarrow$ sample $S$ indices uniformly from $[p]$ without replacement
    $d_i^{j+1} \leftarrow d_i^j + d_t^0 - \eta\left(\frac{1}{S}\sum_{k\in I} d_i^j(k)\cdot \mathbf{S}_\omega(\nabla f_k(\widehat{\theta}+\boldsymbol{e}_k) - f_k(\widehat{\theta}))\right)$
  **end for**
  $u_{i+1} \leftarrow u_i + \bar{d}_i$, where $\bar{d}_i = \frac{1}{L_i+1}\sum_{j=0}^{L_i} d_i^j$
**end for**

Similar to Algorithm 2, we choose $\eta = \Theta\left(\frac{1}{p}\right)$ and $L_i = \Theta(p)$.

## A.3 Solving the high dimensional linear regression optimization objective (13) using proximal SVRG

We solve our high dimensional linear regression optimization problem using proximal SVRG [61]

$$\widehat{\theta} = \arg\min_\theta \frac{1}{2}\theta^\top\left(\widehat{S} - \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\theta + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(x_i^\top\theta - y_i\right)^2 + \lambda\|\theta\|_1. \quad (18)$$

Similar to Algorithm 2, we choose $\eta = \Theta\left(\frac{1}{p}\right)$ and $L_i = \Theta(p)$.

## A.4 Non-asymptotic covariance estimate bound and asymptotic normality in Algorithm 2

We have a non-asymptotic covariance estimate bound and an asymptotic normality result.

**Algorithm 4** Solving the high dimensional linear regression optimization objective (13) using proximal SVRG

**for** $i = 0$ **to** $L_o - 1$ **do**
  $u_i^0 \leftarrow \theta_i$
  $d_t \leftarrow \boldsymbol{g}_{\widehat{S}}(\theta_i) - \frac{1}{n}\sum_{k=1}^{n}[\nabla f_k(\theta_i + \theta_i) - \nabla f_k(\theta_i)] + \frac{1}{n}\sum_{k=1}^{n}\nabla f_k(\theta_i)$
  **for** $j = 0$ **to** $L_i - 1$ **do**
    $u_i^{j+1} \leftarrow \mathbf{S}_{\eta\lambda}(u_i^j - \eta[d_t + \frac{1}{S}\sum_{k\in I}\left(u_i^j(k) - \theta_i(k)\right) \cdot \mathbf{S}_{\omega}\left(\nabla f_k(\theta_t + \boldsymbol{e}_k) - \nabla f_k(\theta_t)\right)])$
  **end for**
  $\theta_{t+1} \leftarrow \frac{1}{L_i+1}\sum_{j=0}^{L_i} u_i^j$
**end for**

**Theorem 4.** *Under our assumptions, when* $n \gg \max\{b^2, \frac{1}{D_\Sigma{}^2}\}\log p$, $S_o = O(1)$, $S_i = O(1)$, *and conditioned on* $\{x_i\}_{i=1}^n$ *and following events which simultaneously with probability at least* $1 - p^{-\Theta(1)} - n^{-\Theta(1)}$

- ❏ *[A]:* $\max_{1\le i\le n}|\epsilon_i| \lesssim \sigma\sqrt{\log n}$,
- ❏ *[B]:* $\max_{1\le i\le n}\|x_i\|_\infty \lesssim \sqrt{\log p + \log n}$,
- ❏ *[C]:* $\|\widehat{S}^{-1}\|_\infty \lesssim \frac{1}{D_\Sigma}$,

*we choose* $L_i = \Theta(p)$, $\tau = \Theta(\frac{1}{p})$, $\eta = \Theta(\frac{1}{p})$ *in Algorithm 2.*

*Here, we denote the objective function as*

$$P(\theta) = \frac{1}{2}\theta^\top\left(\widehat{S} - \frac{1}{n}\sum_{i=1}^{n}x_i x_i^\top\right)\theta + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(x_i^\top\theta - y_i\right)^2 + \lambda\|\theta\|_1.$$

*Then, we have a non-asymptotic covariance estimate bound*

$$\begin{aligned}&\left\|\tfrac{S_o}{T}\textstyle\sum_{t=1}^{T}\bar{g}_t\bar{g}_t^\top - \widehat{S}^{-1}\left(\tfrac{1}{n}\sum_{i=1}^{n}(x_i^\top\widehat{\theta} - y_i)^2 x_i x_i^\top\right)\widehat{S}^{-1}\right\|_{\max}\\ \lesssim&\sqrt{\left((\log p + \log n)\|\widehat{\theta} - \theta^\star\|_1 + \sigma\sqrt{(\log p + \log n)\log n}\right)\tfrac{\log p}{T}}\\ &+ \tfrac{1}{u}\left[\tfrac{1}{\sqrt{T}}\textstyle\sum_{t=1}^{T}0.95^{L_o^t}(1 + \sqrt{P(\theta_0) - P(\widehat{\theta})}\,0.95^{\sum_{i=0}^{t-1}L_o^t}) + \sqrt{p}(\log p + \log n)\sqrt{P(\theta_0) - P(\widehat{\theta})}\,0.95^{\sum_{i=0}^{t-1}L_o^t}\right],\end{aligned}$$

*where* $\|A\|_{\max} = \max\{1 \le j,k \le p\}|A_{jk}|$ *is the matrix max norm, with probability at least* $1 - p^{-\Theta(-1)} - u$.

*And we have asymptotic normality*

$$\tfrac{1}{\sqrt{t}}\left(\textstyle\sum_{t=1}^{T}\sqrt{S_o}\bar{g}_t + \frac{1}{n}\sum_{i=1}^{n}x_i(x_i^\top\widehat{\theta} - y_i)\right) = W + R,$$

*where* $W$ *weakly converges to* $\mathcal{N}\big(0, \widehat{S}^{-1}\big[\frac{1}{n}\sum_{i=1}^{n}(x_i^\top\widehat{\theta} - y_i)^2 x_i x_i^\top - \big(\frac{1}{n}\sum_{i=1}^{n}x_i(x_i^\top\widehat{\theta} - y_i)\big)\big(\frac{1}{n}\sum_{i=1}^{n}x_i(x_i^\top\widehat{\theta} - y_i)\big)^\top\big]\widehat{S}^{-1}\big)$,
*and* $\mathbb{E}[\|R\|_\infty \mid \{x_i\}_{i=1}^n, [A], [B], [C]] \lesssim \frac{1}{\sqrt{T}}\sum_{t=1}^{T}0.95^{L_o^t}(1 + \sqrt{P(\theta_0) - P(\widehat{\theta})}\,0.95^{\sum_{i=0}^{t-1}L_o^t}) +$
$\sqrt{p}(\log p + \log n)\sqrt{P(\theta_0) - P(\widehat{\theta})}\,0.95^{\sum_{i=0}^{t-1}L_o^t}$.

Note that when we choose $L_o^t = \Theta(\log(p)\cdot\log(t))$, and start from $\theta_0$ satisfying $P(\theta_0) - P(\widehat{\theta}) \lesssim \frac{1}{p(\log p+\log n)^2}$ which can be effectively achieved using proximal SVRG (Appendix A.3), we can estimate the statistical error covariance with element-wise error less than $O\left(\frac{\max\{1,\sigma\}\mathrm{polylog}(n,p)}{\sqrt{T}}\right)$ with high probability, using $O\left(T\cdot n\cdot p^2\cdot\log(p)\cdot\log(T)\right)$ numerical operations.

### A.5 Plug-in statistical error covariance estimate

Algorithm 2 is similar to using plug-in estimator $\frac{1}{n}\sum_{i=1}^n (x_i^\top\widehat{\theta}-y_i)^2 x_i x_i^\top$ for $\sigma^2\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)$ in Theorem 3, similar to the sandwich estimator [29, 59]. Lemma 1 gives a bound on using this plug-in estimator in the statistical error covariance (Theorem 3) for coordinate-wise confidence intervals.

**Lemma 1.** *Under our assumptions, when $n \gg \max\{b^2, \frac{1}{D_\Sigma{}^2}\}\log p$, we have*

$$\left\|\widehat{S}^{-1}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n (x_i^\top\widehat{\theta}-y_i)^2 x_i x_i^\top\right)\widehat{S}^{-1} - \sigma^2\widehat{S}^{-1}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n x_i x_i^\top\right)\widehat{S}^{-1}\right\|_{\max}$$
$$\lesssim \frac{1}{D_\Sigma{}^2}\left(\sigma\sqrt{\log n} + s\left(\sigma+\|\theta^\star\|_1\right)\sqrt{\log p+\log n}\sqrt{\frac{\log p}{n}}\right) s\left(\sigma+\|\theta^\star\|_1\right)(\log p+\log n)^{\frac{3}{2}}\sqrt{\frac{\log p}{n}},$$

*where $\|A\|_{\max} = \max_{1\le j,k\le p}|A_{jk}|$ is the matrix max norm, with probability at least $1 - p^{-\Theta(1)} - n^{-\Theta(1)}$.*

# B Statistical inference via approximate stochastic Newton steps using first order information with increasing inner loop counts

Here, we present corollaries when the number of inner loops increases in the outer loops (i.e., $(L)_t$ is an increasing series). This guarantees convergence of the covariance estimate to $H^{-1}GH^{-1}$, although it is less efficient than using a constant number of inner loops.

### B.1 Unregularized M-estimation

Similar to Theorem 1's proof, we have the following result when the number of inner loop increases in the outer loops.

**Corollary 2.** *In Algorithm 1, if the number of inner loop in each outer loop $(L)_t$ increases in the outer loops, then we have*

$$\mathbb{E}\left[\left\|H^{-1}GH^{-1} - \frac{S_o}{T}\sum_{t=1}^T \frac{\bar{g}_t\bar{g}_t^\top}{\rho_t^2}\right\|_2\right] \lesssim T^{-\frac{d_o}{2}} + \sqrt{\frac{1}{T}\sum_{i=1}^T\frac{1}{(L)_t}}.$$

*For example, when we choose choose $(L)_t = L(t+1)^{d_L}$ for some $d_L > 0$, then $\sqrt{\frac{1}{T}\sum_{i=1}^T\frac{1}{(L)_t}} = O(\frac{1}{\sqrt{L}}T^{-\frac{d_L}{2}})$.*

# C SVRG based statistical inference algorithm in unregularized M-estimation

Here we present a SVRG based statistical inference algorithm in unregularized M-estimation, which has asymptotic normality and improved bounds for the "covariance". Although Algorithm 5 has stronger guarantees than Algorithm 1, Algorithm 5 requires a full gradient evaluation in each outer loop.

**Algorithm 5** SVRG based statistical inference algorithm in unregularized M-estimation

**for** $t \leftarrow 0; t < T; ++t$ **do**
  $d_t^0 \leftarrow -\eta \nabla f(\theta_t) = -\eta \left(\frac{1}{n}\sum_{i=1}^n \nabla f_i(\theta)\right)$ // point estimation via SVRG
  $I_o \leftarrow$ uniformly sample $S_o$ indices with replacement from $[n]$
  $g_t^0 \leftarrow -\rho_t \left(\frac{1}{S_o}\sum_{i\in I_o} \nabla f_i(\theta_t)\right)$ // statistical inference
  **for** $j \leftarrow 0; j < L; ++j$ **do** // solving (1) approximately using SGD
    $I_i \leftarrow$ uniformly sample $S_i$ indices without replacement from $[n]$
    $d_t^{j+1} \leftarrow d_t^j - \eta \left(\frac{1}{S_i}\sum_{k\in I_i}(\nabla f_k(\theta_t + d_t^j) - \nabla f_k(\theta_t)\right) + d_t^0$ // point estimation via SVRG
    $g_t^{j+1} \leftarrow g_t^j - \tau_j \left(\frac{1}{S_i}\sum_{k\in I_i} \frac{1}{\delta_t^j}[\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)]\right) + \tau_j g_t^0$ // statistical inference
  **end for**
  Use $\sqrt{S_o}\cdot\frac{\bar{g}_t}{\rho_t}$ for statistical inference // $\bar{g}_t = \frac{1}{L+1}\sum_{j=0}^L g_t^j$
  $\theta_{t+1} \leftarrow \theta_t + \bar{d}_t$ // $\bar{d}_t = \frac{1}{L+1}\sum_{j=0}^L d_t^j$
**end for**

**Corollary 3.** *In Algorithm 5, when* $L \geq 20\frac{\max_{1\leq i\leq n}\beta_i}{\alpha}$ *and* $\eta = \frac{1}{10\max_{1\leq i\leq n}\beta_i}$, *after* $T$ *steps of the outer loop, we have a non-asymptotic bound on the "covariance"*

$$\mathbb{E}\left[\left\|H^{-1}GH^{-1} - \frac{S_o}{T}\sum_{t=1}^T \frac{\bar{g}_t\bar{g}_t^\top}{\rho_t^2}\right\|_2\right] \lesssim L^{-\frac{1}{2}}, \tag{19}$$

*and asymptotic normality*

$$\frac{1}{\sqrt{T}}\left(\sum_{t=1}^T \frac{\bar{g}_t}{\rho_t}\right) = W + \Delta,$$

*where* $W$ *weakly converges to* $\mathcal{N}(0, \frac{1}{S_o}H^{-1}GH^{-1})$ *and* $\Delta = o_P(1)$ *when* $T \to \infty$ *and* $L \to \infty$ *(*$\mathbb{E}[\|\Delta\|_2] \lesssim \frac{1}{\sqrt{T}} + \frac{1}{L}$*).*

When the number of inner loops increases in the outer loops (i.e., $(L)_t$ is an increasing series), we have a result similar to Corollary 2.

A better understanding of concentration, and Edgeworth expansion of the average consecutive iterates averaged (beyond [18, 19]) in stochastic approximation, would give stronger guarantees for our algorithms, and better compare and understand different algorithms.

## C.1 Lack of asymptotic normality in Algorithm 1 for mean estimation

In mean estimation, we solve the following optimization problem

$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \|\theta - X^{(i)}\|_2^2,$$

where we assume that $\{X^{(i)}\}_{i=1}^n$ are constants.

For ease of explanation we use $S_o = 1$, $\rho_t = \rho$, and $\theta_0 = 0$,and we have

$$\frac{\bar{g}_t}{\rho_t} = -\theta_t + X_t,$$

where $X_t$ is uniformly sampled from $\{X^{(i)}\}_{i=1}^n$.

And for $t \geq 1$ we have

$$\theta_t = \sum_{i=0}^{t-1} \rho(1-\rho)^{t-1-i} X_i.$$

Then, we have

$$\begin{aligned}
&\frac{1}{\sqrt{T}}(\sum_{i=1}^{T} \frac{\bar{g}_t}{\rho_t}) \\
=&\frac{1}{\sqrt{T}}(\sum_{t=1}^{T} X_t - \sum_{t=1}^{T}\sum_{i=0}^{t-1} \rho(1-\rho)^{t-1-i} X_i) \\
=&\frac{1}{\sqrt{T}}(\sum_{t=1}^{T} X_t - \sum_{i=0}^{T-1}(\sum_{t=i+1}^{T} \rho(1-\rho)^{t-1-i}) X_i) \\
=&\frac{1}{\sqrt{T}}(\sum_{t=1}^{T} X_t - \sum_{i=0}^{T-1}(1-(1-\rho)^{T-i}) X_i) \\
=&\frac{1}{\sqrt{T}}(X_T - X_0 + \sum_{i=1}^{T-1}(1-\rho)^{T-i} X_i),
\end{aligned}$$

whose $\ell_2$ norm's expectation converges to 0 when $T \to \infty$, which implies that it converges to 0 with probability 1. Thus, in this setting $\frac{1}{\sqrt{T}} \left( \sum_{t=1}^{T} \frac{\bar{g}_t}{\rho_t} \right)$ does not weakly converge to $\mathcal{N}\left(0, \frac{1}{S_o} H^{-1} G H^{-1}\right)$.

# D An intuitive view of SVRG as approximate stochastic Newton descent

Here we present an intuitive view of SVRG as approximate stochastic Newton descent, which is the inspiration behind our work.

Gradient descent solves the optimization problem $\widehat{\theta} = \arg\min_\theta f(\theta)$, where the function is a sum of $n$ functions $f(\theta) = \frac{1}{n}\sum_{i=1}^n f_i(\theta)$, using

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t),$$

and stochastic gradient descent uniformly samples a random index at each step

$$\theta_{t+1} = \theta_t - \eta_t \nabla f_i(\theta_t).$$

- ❏ **Outer loop:**
- ❏ $g \leftarrow \nabla f(\theta_t) = \sum_{i=1}^n \nabla f_i(\theta_t)$
- ❏ Let $d$ be the descent direction
- ❏ – **Inner loop:**
  - Choose a random index $k$
  - $d \leftarrow d - \eta(\nabla f_k(\theta_t + d) - \nabla f_k(\theta_t) + g)$
- ❏ $\theta_{t+1} = \theta_t + d$

SVRG [34] improves gradient descent and SGD by having an outer loop and an inner loop.

Here, we give an intuitive explanation of SVRG as stochastic proximal Newton descent, by arguing that

- ❏ each outer loop approximately computes the Newton direction $-(\nabla^2 f)^{-1}\nabla f$
- ❏ the inner loops can be viewed as SGD steps solving a proximal Newton step $\min_d \langle \nabla f, d\rangle + \frac{1}{2} d^\top (\nabla^2 f) d$

First, it is well known [9] that the Newton direction is exactly the solution of

$$\min_d \langle \nabla f(\theta), d\rangle + \frac{1}{2} d^\top [\nabla^2 f(\theta)] d. \tag{20}$$

Next, let's consider solving (20) using gradient descent on a function of $d$, and notice that its gradient with respect to $d$ is

$$\nabla f(\theta) + [\nabla^2 f(\theta)] d,$$

which can be approximated through $f$'s Taylor expansion ($[\nabla^2 f(\theta)] d \approx \nabla f(\theta + d) - \nabla f(\theta)$) as

$$\nabla f(\theta) + [\nabla f(\theta + d) - \nabla f(\theta)].$$

Thus, SVRG's inner loops can be viewed as using SGD to solve proximal Newton steps in outer loops. And it can be viewed as the power series identity for matrix inverse $H^{-1} = \sum_{i=0}^{\infty}(I-\eta H)$, which corresponds to unrolling the gradient descent recursion for the optimization problem $H^{-1} = \arg\min_{\Omega} \operatorname{Tr}\left(\frac{1}{2}\Omega^{\top} H\Omega - \Omega\right)$.

# E Proofs

## E.1 Proof of Theorem 1

Given assumptions about strong convexity, Lipschitz gradient continuity and Hessian Lipschitz continuity in Theorem 1, we denote:

$$\bar{\beta} = \frac{\beta_i}{n}, \quad \bar{h} = \frac{h_i}{n}.$$

Then, $\forall \theta_1, \theta_2$ we have:

$$\|\nabla f(\theta_2) - \nabla f(\theta_1)\|_2 \leq \bar{\beta}\|\theta_2 - \theta_1\|_2, \quad \text{and} \quad \|\nabla^2 f(\theta_2) - \nabla^2 f(\theta_1)\|_2 \leq \bar{h}\|\theta_2 - \theta_1\|_2.$$

and $\forall \theta$:

$$\|\nabla^2 f(\theta)\|_2 \leq \bar{\beta}.$$

In our proof, we also use the following:

$$\bar{h}_2 = \frac{1}{n}\sum_{i=1}^{n} h_i^2, \quad \bar{\beta}_2 = \frac{1}{n}\sum_{i=1}^{n} \beta_i^2, \quad \text{and} \quad \beta = \sup_{\theta} \|\nabla^2 f(\theta)\|_2.$$

Observe that:

$$\bar{h} \leq \sqrt{\bar{h}_2}, \quad \text{and} \quad \alpha \leq \beta \leq \bar{\beta} \leq \sqrt{\bar{\beta}_2}.$$

### E.1.1 Proof of (8)

We first prove (8); the proof is similar to standard SGD convergence proofs (e.g. [37, 14, 43]). For the rest of our discussion, we assume that

$$\delta_t^j \cdot \bar{h} \leq \delta_t^j \cdot \sqrt{\bar{h}_2} \ll 1, \quad \forall t, j.$$

Using $\nabla f(\theta)$'s Taylor series expansion with a Lagrange remainder, we have the following lemma, which bounds the Hessian vector product approximation error.

**Lemma 2.** *$\forall, \theta, g, \delta \in \mathbb{R}^p$, we have:*

$$\left\| \frac{\nabla f_i(\theta+\delta g) - \nabla f_i(\theta)}{\delta} - \nabla^2 f_i(\theta) g \right\|_2 \leq h_i \cdot |\delta| \cdot \|g\|_2,$$
$$\left\| \frac{\nabla f(\theta+\delta g) - \nabla f(\theta)}{\delta} - \nabla^2 f(\theta) g \right\|_2 \leq \bar{h} \cdot |\delta| \cdot \|g\|_2.$$

Denote $H_t = \nabla^2 f(\theta_t)$ and

$$e_t^j = \left( \frac{1}{S_i} \cdot \sum_{k \in I_i} \frac{\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)}{\delta_t^j} \right) - \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j},$$

then we have

$$g_t^{j+1} - H_t^{-1} g_t^0 = g_t^j - H_t^{-1} g_t^0 - \tau_j \cdot \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0 - \tau_j e_t^j. \tag{21}$$

Because $\mathbb{E}[e_j^t \mid g_j^t, \theta_t] = 0$, we have

$$\begin{aligned}
\mathbb{E}\left[ \left\| g_t^{j+1} - H_t^{-1} g_t^0 \right\|_2^2 \mid \theta_t \right] = \mathbb{E}\Bigg[ & \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 - \tau_j \underbrace{\left\langle g_t^j - H_t^{-1} g_t^0, \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - g_t^0 \right\rangle}_{[1]} \\
& + \tau_j^2 \underbrace{\left\| \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - g_t^0 \right\|_2^2}_{[2]} + \tau_j^2 \underbrace{\left\| e_t^j \right\|_2^2}_{[3]} \mid \theta_t \Bigg].
\end{aligned} \tag{22}$$

For term [1], we have

$$\begin{aligned}
& \left\langle g_t^j - H_t^{-1} g_t^0, \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - g_t^0 \right\rangle \\
& \quad = \left( g_t^j - H_t^{-1} g_t^0 \right)^\top H_t \left( g_t^j - H_t^{-1} g_t^0 \right) + \left\langle g_t^j - H_t^{-1} g_t^0, \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - H_t \right\rangle \\
& \quad \geq \left( g_t^j - H_t^{-1} g_t^0 \right)^\top H_t \left( g_t^j - H_t^{-1} g_t^0 \right) - \left| \langle g_t^j - H_t^{-1} g_t^0, \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - H_t \rangle \right| \\
& \quad \text{by Hessian approximation} \\
& \quad \geq \left( g_t^j - H_t^{-1} g_t^0 \right)^\top H_t \left( g_t^j - H_t^{-1} g_t^0 \right) - \delta_t^j \cdot \bar{h} \cdot \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2 \cdot \left\| g_t^j \right\|_2 \\
& \quad \text{by AM-GM inequality} \\
& \quad \geq \left( g_t^j - H_t^{-1} g_t^0 \right)^\top H_t \left( g_t^j - H_t^{-1} g_t^0 \right) - \tfrac{\delta_t^j \cdot \bar{h}}{2} \cdot \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 - \tfrac{\delta_t^j \cdot \bar{h}}{2} \cdot \left\| g_t^j \right\|_2^2 \\
& \quad = \left( g_t^j - H_t^{-1} g_t^0 \right)^\top H_t \left( g_t^j - H_t^{-1} g_t^0 \right) - \tfrac{\delta_t^j \cdot \bar{h}}{2} \cdot \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 - \tfrac{\delta_t^j \cdot \bar{h}}{2} \cdot \left\| g_t^j - H_t^{-1} g_t^0 + H_t^{-1} g_t^0 \right\|_2^2 \\
& \quad \|x + u\|_2^2 \leq 2\|x\|_2^2 + 2\|y\|_2^2 \\
& \quad \geq \left( g_t^j - H_t^{-1} g_t^0 \right)^\top H_t \left( g_t^j - H_t^{-1} g_t^0 \right) - \tfrac{3\delta_t^j \cdot \bar{h}}{2} \cdot \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 - \delta_t^j \bar{h} \cdot \left\| H_t^{-1} g_t^0 \right\|_2^2 \\
& \quad \text{by strong convexity} \\
& \quad \geq \left( g_t^j - H_t^{-1} g_t^0 \right)^\top H_t \left( g_t^j - H_t^{-1} g_t^0 \right) - \tfrac{3\delta_t^j \cdot \bar{h}}{2} \cdot \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 - \tfrac{\delta_t^j \bar{h}}{\alpha^2} \cdot \left\| g_t^0 \right\|_2^2.
\end{aligned} \tag{23}$$

For term [2], by repeatedly applying AM-GM inequality, using $f$'s smoothness and strong convexity, and assuming $\delta_t^j \bar{h} \ll 1$, we have:

$$\left\| \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - g_t^0 \right\|_2^2 = \left\| \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - H_t g_t^j + H_t g_t^j - g_t^0 \right\|_2^2$$

$$
\begin{aligned}
&\leq \left\| \tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} - H_t g_t^j \right\|_2^2 \\
&\quad + 2\left\| \tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} - H_t g_t^j \right\|_2 \cdot \left\| H_t g_t^j - g_t^0 \right\|_2 + \left\| H_t g_t^j - g_t^0 \right\|_2^2 \\
&\leq \left(\delta_t^j \bar{h}\right)^2 \|g_t^j\|_2^2 + 2\delta_t^j \bar{h} \left\| g_t^j \right\|_2 \cdot \left\| H_t g_t^j - g_t^0 \right\|_2 + \left\| H_t g_t^j - g_t^0 \right\|_2^2 \\
&\leq \left(\delta_t^j \bar{h} + \left(\delta_t^j \bar{h}\right)^2\right) \cdot \left\| g_t^j \right\|_2^2 + \left(1 + \delta_t^j \bar{h}\right) \cdot \|H_t g_t^j - g_t^0\|_2^2 \\
&\leq 2\left(\delta_t^j \bar{h} + \left(\delta_t^j \bar{h}\right)^2\right) \cdot \left( \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 + \left\| H_t^{-1} g_t^0 \right\|_2^2 \right) + \left(1 + \delta_t^j \bar{h}\right) \cdot \left\| H_t g_t^j - g_t^0 \right\|_2^2 \\
&\leq \tfrac{2\left(\delta_t^j \bar{h} + \left(\delta_t^j \bar{h}\right)^2\right)}{\alpha^2} \cdot \left\| g_t^0 \right\|_2^2 + \left(1 + 3\delta_t^j \bar{h} + 2\left(\delta_t^j \bar{h}\right)^2\right) \cdot \left\| H_t g_t^j - g_t^0 \right\|_2^2 \\
&\leq \tfrac{4\delta_t^j \bar{h}}{\alpha^2} \cdot \left\| g_t^0 \right\|_2^2 + \left(1 + 5\delta_t^j \bar{h}\right) \cdot \|H_t g_t^j - g_t^0\|_2^2.
\end{aligned}
$$

For term [3], because we sample uniformly without replacement, we obtain:

$$
\mathbb{E}_{I_i}\left[ \left\| e_t^j \right\|_2^2 \mid g_t^j, \theta_t \right] = \tfrac{1}{S_i}\left(1 - \tfrac{S_i - 1}{n-1}\right) \cdot \mathbb{E}_k \left[ \left\| \tfrac{\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)}{\delta_t^j} - \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} \right\|_2^2 \right],
$$

where $k$ is uniformly sampled from $[n]$. Denote $H_t^k = \nabla^2 f_k(\theta_t)$, and by Lipschitz gradient we have $\|H_t^k\|_2 \leq \beta_k$. We can bound the above

$$
\begin{aligned}
&\left\| \tfrac{\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)}{\delta_t^j} - \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} \right\|_2^2 \\
&\qquad = \left\| \tfrac{\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)}{\delta_t^j} - H_t^k g_t^j + H_t^k g_t^j - \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} + H_t g_t^j - H_t g_t^j \right\|_2^2 \\
&\qquad \leq 3\left( \left\| \left(H_t - H_t^k\right) g_t^j \right\|_2^2 + \left\| \tfrac{\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)}{\delta_t^j} - H_t^k g_t^j \right\|_2^2 + \left\| \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - H_t g_t^j \right\| \right. \\
&\qquad \leq 3\left( \left\| H_t - H_t^k \right\|_2^2 + (\delta_t^j)^2 \left(\bar{h}^2 + h_k^2\right) \right) \cdot \left\| g_t^j \right\|_2^2 \\
&\qquad \|H_t - H_t^k\|_2^2 \leq 2(\bar{\beta}^2 + \beta_k^2) \\
&\qquad \leq 3\left( 2\left(\bar{\beta}^2 + \beta_k^2\right) + (\delta_t^j)^2 (\bar{h}^2 + h_k^2) \right) \cdot \left\| g_t^j \right\|_2^2 \\
&\qquad \leq 6\left( 2\left(\bar{\beta}^2 + \beta_k^2\right) + (\delta_t^j)^2 (\bar{h}^2 + h_k^2) \right) \cdot \left( \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 + \left\| H_t^{-1} g_t^0 \right\|_2^2 \right).
\end{aligned}
$$

Taking the expectation over inner loop's random indices, for term [3], we have

$$
\begin{aligned}
\mathbb{E}_{I_i}\left[ \left\| e_t^j \right\|_2^2 \mid g_t^j, \theta_t \right] &\leq 6\left( \tfrac{1}{S_i} \cdot \left(1 - \tfrac{S_i - 1}{n-1}\right) \right) \left( \left(\delta_t^j \bar{h}\right)^2 + 2\bar{\beta}^2 + (\delta_t^j)^2 \bar{h}_2 + 2\bar{\beta}_2 \right) \cdot \left( \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 + \tfrac{1}{\alpha^2} \cdot \left\| g_t^0 \right\|_2^2 \right) \\
&\leq 18\left( \tfrac{1}{S_i}\left(1 - \tfrac{S_i - 1}{n-1}\right) \right) \cdot \left( (\delta_t^j)^2 \bar{h}_2 + \bar{\beta}_2 \right) \cdot \left( \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 + \tfrac{1}{\alpha^2} \left\| g_t^0 \right\|_2^2 \right).
\end{aligned} \tag{24}
$$

Combining all above, we have

$$\begin{aligned}\mathbb{E}\left[\left\|g_t^{j+1}-H_t^{-1}g_t^0\right\|_2^2 \mid g_t^j,\theta_t\right] \le \left\|g_t^j-H_t^{-1}g_t^0\right\|_2^2 \\ -\tau_j\left(g_t^j-H_t^{-1}g_t^0\right)^\top H_t\left(g_t^j-H_t^{-1}g_t^0\right)+\tfrac{3\tau_j\delta_t^j\bar{h}}{2}\left\|g_t^j-H_t^{-1}g_t^0\right\|_2^2+\tfrac{\tau_j\delta_t^j\bar{h}}{\alpha^2}\left\|g_t^0\right\|_2^2 \\ +\tfrac{4\tau_j^2\delta_t^j\bar{h}}{\alpha^2}\cdot\left\|g_t^0\right\|_2^2+\tau_j^2\left(1+5\delta_t^j\bar{h}\right)\cdot\left\|H_tg_t^j-g_t^0\right\|_2^2 \\ +18\tau_j^2\left(\tfrac{1}{S_i}\left(1-\tfrac{S_i-1}{n-1}\right)\right)\cdot\left((\delta_t^j)^2\bar{h}_2+\bar{\beta}_2\right)\cdot\left(\left\|g_t^j-H_t^{-1}g_t^0\right\|_2^2+\tfrac{1}{\alpha^2}\|g_t^0\|_2^2\right).\end{aligned}$$

When we choose the Hessian vector product approximation scaling constant $\delta_t^j$ to be sufficiently small

$$\begin{aligned}\delta_t^j\bar{h}\le\delta_t^j\sqrt{\bar{h}_2}&\le 0.01,\\ \tfrac{3\delta_t^j\bar{h}}{2}&\le 0.01\alpha,\\ \delta_t^j\bar{h}\le\delta_t^j\sqrt{\bar{h}_2}&\le\tfrac{0.01}{S_i}\left(1-\tfrac{S_i-1}{n-1}\right)\bar{\beta}^2\le\tfrac{0.01}{S_i}\left(1-\tfrac{S_i-1}{n-1}\right)\bar{\beta}_2,\\ \delta_t^j\bar{h}\le\delta_t^j\sqrt{\bar{h}_2}&\le\tfrac{0.01\tau_j}{S_i}\left(1-\tfrac{S_i-1}{n-1}\right)\bar{\beta}^2\le\tfrac{0.01\tau_j}{S_i}\left(1-\tfrac{S_i-1}{n-1}\right)\bar{\beta}_2,\\ \delta_t^j\bar{h}\le\delta_t^j\sqrt{\bar{h}_2}&\le 0.01\alpha\le 0.01\bar{\beta}\le 0.01\sqrt{\bar{\beta}_2},\end{aligned}$$

we have

$$\begin{aligned}\mathbb{E}\left[\left\|g_t^{j+1}-H_t^{-1}g_t^0\right\|_2^2 \mid g_t^j,\theta_t\right] \le \left\|g_t^j-H_t^{-1}g_t^0\right\|_2^2 \underbrace{-\tau_j\left(g_t^j-H_t^{-1}g_t^0\right)^\top H_t\left(g_t^j-H_t^{-1}g_t^0\right)+1.05\tau_j^2\|H_tg_t^j-g_t^0\|_2^2}_{[4]} \\ +18.5\tau_j^2\left(\tfrac{1}{S_i}\left(1-\tfrac{S_i-1}{n-1}\right)\right)\bar{\beta}_2\left\|g_t^j-H_t^{-1}g_t^0\right\|_2^2 \\ +18.5\tau_j^2\left(\tfrac{1}{S_i}\left(1-\tfrac{S_i-1}{n-1}\right)\right)\tfrac{\bar{\beta}_2}{\alpha^2}\cdot\left\|g_t^0\right\|_2^2.\end{aligned}$$

For term [4], let us consider the $\alpha$ strongly convex and $\beta$ smooth quadratic function

$$F(g)=\tfrac{1}{2}g^\top H_tg-\langle g_t^0,g\rangle,$$

who attains its minimum at $g=H_t^{-1}g_t^0$. Using a well known property of $\alpha$ strongly convex and $\beta$ smooth functions (Lemma 5), we have

$$\begin{aligned}-\left(g_t^j-H_t^{-1}g_t^0\right)^\top H_t\left(g_t^j-H_t^{-1}g_t^0\right)+\tfrac{1}{2\beta}\|H_tg_t^j-g_t^0\|_2^2 &\le -\left(g_t^j-H_t^{-1}g_t^0\right)^\top H_t\left(g_t^j-H_t^{-1}g_t^0\right)+\tfrac{1}{\alpha+\beta}\|H_tg_t^j-g_t^0\|_2^2\\ &\le -\tfrac{\alpha\beta}{\alpha+\beta}\|g_t^j-H_t^{-1}g_t^0\|_2^2\\ &\le -\tfrac{\alpha}{2}\|g_t^j-H_t^{-1}g_t^0\|_2^2.\end{aligned}$$

Thus, when we choose

$$\tau_j \leq \tfrac{0.476}{\beta},$$

we have

$$\begin{aligned}-\tau_j \left(g_t^j - H_t^{-1} g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1} g_t^0\right) + 1.05\tau_j^2 \cdot \left\| H_t g_t^j - g_t^0 \right\|_2^2 &\leq -\tau_j \left(g_t^j - H_t^{-1} g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1} g_t^0\right) + \tfrac{\tau_j}{2\beta} \left\| H_t g_t^j - g \right. \\ &\leq -\tfrac{\tau_j \alpha}{2} \cdot \|g_t^j - H_t^{-1} g_t^0\|_2^2,\end{aligned}$$

and we have

$$\begin{aligned}\mathbb{E}\left[\left\| g_t^{j+1} - H_t^{-1} g_t^0 \right\|_2^2 \mid g_t^j, \theta_t\right] \leq &\left(1 - \tau_j \alpha + 18.5\tau_j^2 \left(\tfrac{1}{S_i}\left(1 - \tfrac{S_i - 1}{n-1}\right)\right) \bar{\beta}_2\right) \cdot \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 \\ &+ 18.5\tau_j^2 \left(\tfrac{1}{S_i}\left(1 - \tfrac{S_i - 1}{n-1}\right)\right) \cdot \tfrac{\bar{\beta}_2}{\alpha^2} \cdot \|g_t^0\|_2^2.\end{aligned}$$

Next, we set

$$\tau_0 = \min\left\{ \tfrac{0.476}{\beta}, \tfrac{0.025 \cdot \alpha}{\frac{1}{S_i}\left(1 - \frac{S_i - 1}{n-1}\right)\bar{\beta}_2} \right\}, \quad D_j = (j+1)^{-d_i}, \quad \tau_j = \tau_0 D_j, \qquad (25)$$

where $d_i$ is inner loop's step size decay rate, and we have:

$$\begin{aligned}\mathbb{E}\left[\left\| g_t^{j+1} - H_t^{-1} g_t^0 \right\|_2^2 \mid \theta_t\right] \leq &\left(1 - \min\left\{ \tfrac{\alpha}{2\bar{\beta}}, \tfrac{0.013 \cdot \alpha^2}{\frac{1}{S_i}\left(1 - \frac{S_i - 1}{n-1}\right)\bar{\beta}_2} \right\} D_j\right) \cdot \mathbb{E}\left[\left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 \mid \theta_t\right] \\ &+ 18.5 D_j^2 \tau_0^2 \left(\tfrac{1}{S_i}\left(1 - \tfrac{S_i - 1}{n-1}\right)\right) \tfrac{\bar{\beta}_2}{\alpha^2} \cdot \left\| g_t^0 \right\|_2^2.\end{aligned}$$

To satisfy the above requirements, for the Hessian vector product approximation scaling constant, we choose:

$$\begin{aligned}&\delta_t^j = o\left(\min\left\{1, \tfrac{1}{h}\right\} \cdot \min\left\{1, \alpha, \min\left\{1, \tau_0^4 \left(\tfrac{\tau_j}{\tau_0}\right)^4\right\} \tfrac{1}{S_i}\left(1 - \tfrac{S_i - 1}{n-1}\right)\right\}\right) \cdot \delta_t^0 = o\left((j+1)^{-2}\right) \cdot \delta_t^0, \\ &\delta_t^0 = O(\rho_t^4) = o((t+1)^{-2}) = o(1). \qquad (26)\end{aligned}$$

which is trivially satisfied for quadratic functions because all $h_i = 0$.

Note that:

$$18.5\tau_0^2 \left(\tfrac{1}{S_i}\left(1 - \tfrac{S_i - 1}{n-1}\right)\right) \cdot \tfrac{\bar{\beta}_2}{\alpha^2} = \Theta\left(\min\left\{ \left(\tfrac{1}{S_i}\left(1 - \tfrac{S_i - 1}{n-1}\right)\right) \cdot \tfrac{\bar{\beta}_2}{\beta^2 \alpha^2}, \tfrac{1}{\frac{1}{S_i}\left(1 - \frac{S_i - 1}{n-1}\right) \cdot \bar{\beta}_2} \right\}\right).$$

Applying Lemma 6, we have:

$$\mathbb{E}\left[\left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 \mid \theta_t\right] = O\left(t^{-d_i} \cdot \|g_t^0\|_2^2\right), \qquad (27)$$

where we have assumed that $\alpha$, $\beta$, $S_i$, etc. are (data dependent) constants. Further, (27) implies:

$$\mathbb{E}\left[\left\|g_t^j\right\|_2^2\right] \le 2\mathbb{E}\left[\left\|g_t^j - H_t^{-1}g_t^0\right\|_2^2 + \left\|H_t^{-1}g_t^0\right\|_2^2 \mid \theta_t\right] \lesssim \|g_t^0\|_2^2, \quad \text{for all } j. \tag{28}$$

In Algorithm 1, we have

$$g_t^{j+1} - H_t^{-1}g_t^0 = (I - \tau_j H_t)(g_t^j - H_t^{-1}g_t^0) + \tau_j\left(-e_t^j - \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} + H_t g_t^j\right).$$

By unrolling the recursion we have:

$$g_t^{j+1} - H_t^{-1}g_t^0 = \sum_{k=0}^{j}\left(\prod_{l=k+1}^{j}(I - \tau_l H_t)\right)\cdot\tau_k\cdot\left(-e_t^k - \tfrac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k\right). \tag{29}$$

For the average $\bar{g}_t$, we have:

$$\begin{aligned}
\bar{g}_t - H_t^{-1}g_t^0 &= \tfrac{1}{L+1}\sum_{j=0}^{L}(g_t^j - H_t^{-1}g_t^0) \\
&= \tfrac{1}{L+1}\sum_{j=0}^{L}\sum_{k=0}^{j-1}\left(\prod_{l=k+1}^{j-1}(I - \tau_l H_t)\right)\cdot\tau_k\left(-e_t^k - \tfrac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k\right) \\
&= \tfrac{1}{L+1}\underbrace{\sum_{k=0}^{L-1}\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I - \tau_l H_t)}_{[5]}\left(-e_t^k - \tfrac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k\right) \\
&= \underbrace{\tfrac{1}{L+1}\sum_{k=0}^{L-1}\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I - \tau_l H_t)\left(-e_t^k\right)}_{[6]} \\
&\quad + \underbrace{\tfrac{1}{L+1}\sum_{k=0}^{L-1}\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I - \tau_l H_t)\left(-\tfrac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k\right)}_{[7]}.
\end{aligned} \tag{30}$$

For the term [5], we have:

$$\left\|\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I - \tau_l H_t)\right\|_2 \le \tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}\|I - \tau_l H_t\|_2$$

$I - \tau_l H_t$ is positive definite by our choice of $\tau_l$ (25) and $\|I - \tau_l H_t\|_2 \le 1 - \tau_l\alpha$

$$
\begin{aligned}
&\leq \tau_k \sum_{j=k+1}^{L} \prod_{l=k+1}^{j-1} (1-\tau_l\alpha) \\
&\leq \tau_k \sum_{j=k+1}^{L} \prod_{l=k+1}^{j-1} \left(1-\frac{1}{2}\tau_l\alpha\right)^2 \\
&\tau_k \prod_{l=k+1}^{j-1} (1-\frac{1}{2}\tau_l\alpha) \leq \tau_k \exp(-\frac{1}{2}\alpha \sum_{l=k+1}^{j-1} \tau_l) \lesssim k^{-d_i} \exp(\Theta(-j^{1-d_i}+k^{1-d_i})) \lesssim j^{-d_i} \\
&\text{because for a fixed } d_i \; x^{-d_i} e^{\Theta(x^{1-d_i})} \text{ is an increasing function when } x \text{ is sufficiently} \\
&\lesssim \sum_{j=k+1}^{L} \tau_j \prod_{l=k+1}^{j-1} (1-\tfrac{\tau_l\alpha}{2}) = \tfrac{2}{\alpha} \sum_{j=k+1}^{L} \frac{1}{2}\tau_j\alpha \prod_{l=k+1}^{j-1} (1-\tfrac{\tau_l\alpha}{2}) \\
&= \tfrac{2}{\alpha}\left(1-\prod_{j=k+1}^{L}(1-\tfrac{\tau_l\alpha}{2})\right) = O(1), \qquad (31)
\end{aligned}
$$

where we have assumed that $\alpha$, $\beta$, $S_i$, etc. are (data-dependent) constants.

For the term [6], its norm is bounded by:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\tfrac{1}{L+1}\sum_{k=0}^{L-1}\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I-\tau_l H_t)(-e_t^k)\right\|_2^2 \mid \theta_t\right] = \tfrac{1}{(L+1)^2}\mathbb{E}\left[\sum_{k=0}^{L-1}\left\|\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I-\tau_l H_t)(-e_t^k)\right\|_2^2 \mid \theta_t\right] \\
&\qquad \text{using (31)} \\
&\qquad \lesssim \tfrac{1}{(L+1)^2}\mathbb{E}\left[\sum_{k=0}^{L-1}\|e_t^k\|_2^2 \mid \theta_t\right] \\
&\qquad \text{using (24) and (27)} \\
&\qquad \lesssim \tfrac{1}{L}\|g_t^0\|_2^2. \qquad (32)
\end{aligned}
$$

where the first equality is due to $a<b$, $\mathbb{E}[{e_t^a}^\top e_t^b \mid \theta_t]=0$, when we first condition on $b$.

For the term [7], its norm is bounded by:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\tfrac{1}{L+1}\sum_{k=0}^{L-1}\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I-\tau_l H_t)\left(-\tfrac{\nabla f(\theta_t+\delta_t^k g_t^k)-\nabla f(\theta_t)}{\delta_t^k}+H_t g_t^k\right)\right\|_2^2 \mid \theta_t\right] \\
&\quad = \tfrac{1}{(L+1)^2}\mathbb{E}\Bigg[\sum_{0\leq a,b,\leq L-1}\Bigg\langle \tau_a\sum_{j=a+1}^{L}\prod_{l=a+1}^{j-1}(I-\tau_l H_t)\left(-\tfrac{\nabla f(\theta_t+\delta_t^a g_t^a)-\nabla f(\theta_t)}{\delta_t^a}+H_t g_t^a\right), \\
&\qquad\qquad \tau_b\sum_{j=b+1}^{L}\prod_{l=b+1}^{j-1}(I-\tau_l H_t)\left(-\tfrac{\nabla f(\theta_t+\delta_t^b g_t^b)-\nabla f(\theta_t)}{\delta_t^b}+H_t g_t^b\right)\Bigg\rangle \mid \theta_t\Bigg]
\end{aligned}
$$

$$\leq \frac{1}{(L+1)^2}\mathbb{E}\Bigg[\sum_{0\leq a,b,\leq L-1}\Bigg\|\tau_a\sum_{j=a+1}^{L}\prod_{l=a+1}^{j-1}(I-\tau_l H_t)\left(-\frac{\nabla f(\theta_t+\delta_t^a g_t^a)-\nabla f(\theta_t)}{\delta_t^a}+H_t g_t^a\right)\Bigg\|_2$$
$$\cdot\Bigg\|\tau_b\sum_{j=b+1}^{L}\prod_{l=b+1}^{j-1}(I-\tau_l H_t)\left(-\frac{\nabla f(\theta_t+\delta_t^b g_t^b)-\nabla f(\theta_t)}{\delta_t^b}+H_t g_t^b\right)\Bigg\|_2 \mid \theta_t\Bigg]$$

using (31) and Lemma 2

$$\lesssim \frac{1}{(L+1)^2}\mathbb{E}\left[\sum_{0\leq a,b,\leq L-1}\delta_t^a\bar{h}\|g_t^a\|_2\delta_t^b\bar{h}\|g_t^b\|_2 \mid \theta_t\right] \leq \frac{2\bar{h}^2}{(L+1)^2}\sum_{0\leq a,b,\leq L-1}\delta_t^a\delta_t^b\cdot\mathbb{E}\left[\|g_t^a\|_2^2+\|g_t^b\|_2^2 \mid \theta_t\right]$$
$$\lesssim \frac{\|g_t^0\|_2^2}{(L+1)^2}\sum_{0\leq a,b,\leq L-1}\delta_t^a\delta_t^b \lesssim \frac{\|g_t^0\|_2^2}{L^2}\left(\sum_{k=0}^{L}\delta_t^k\right)^2 \tag{33}$$

using (28) and our choice of $\delta_t^k$ (26)

$$\lesssim \frac{1}{L^2}{\delta_t^0}^2\left(\sum_{k=0}^{L}\tau_k\right)^2\cdot\|g_t^0\|_2^2 \lesssim \frac{1}{L^2}{\delta_t^0}^2\left(\sum_{k=0}^{L}(k+1)^{-d_i}\right)^2\cdot\|g_t^0\|_2^2$$

because $\left(\sum_{k=0}^{L}(k+1)^{-d_i}\right)^2 = O\left(L^{1-d_i}\right)$ and $d_i\in\left(\frac{1}{2},1\right)$

$$\ll \frac{1}{L}\|g_t^0\|_2^2. \tag{34}$$

Combining (32) and (34), we have

$$\|\bar{g}_t - H_t^{-1}g_t^0\|_2^2 = O\left(\tfrac{1}{L}\|g_t^0\|_2^2\right).$$

### E.1.2 Proof of (9)

Using (21), we have

$$\mathbb{E}[\|g_t^{j+1}-H_t^{-1}g_t^0\|_2^4 \mid g_t^j]$$
$$=\mathbb{E}[\|g_t^j-H_t^{-1}g_t^0-\tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j}+\tau_j g_t^0-\tau_j e_t^j\|_2^4 \mid g_t^j]$$
$$=\mathbb{E}[(\|g_t^j-H_t^{-1}g_t^0-\tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j}+\tau_j g_t^0\|_2^2$$
$$-2\langle\tau_j e_t^j, g_t^j-H_t^{-1}g_t^0-\tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j}+\tau_j g_t^0\rangle+\tau_j^2\|e_t^j\|_2^2)^2 \mid g_t^j]$$
$$=\mathbb{E}[\|g_t^j-H_t^{-1}g_t^0-\tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j}+\tau_j g_t^0\|_2^4$$
$$+4(\langle\tau_j e_t^j, g_t^j-H_t^{-1}g_t^0-\tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j}+\tau_j g_t^0\rangle)^2+\tau_j^4\|e_t^j\|_2^4$$
$$+2\|g_t^j-H_t^{-1}g_t^0-\tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j}+\tau_j g_t^0\|_2^2\tau_j^2\|e_t^j\|_2^2$$
$$-4\langle\tau_j e_t^j, g_t^j-H_t^{-1}g_t^0-\tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j}+\tau_j g_t^0\rangle\|g_t^j-H_t^{-1}g_t^0-\tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j}+\tau_j g_t^0\|_2^2$$

$$- 4\langle \tau_j e_t^j, g_t^j - H_t^{-1} g_t^0 - \tau_j \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0 \rangle \tau_j^2 \| e_t^j \|_2^2 \mid g_t^j]. \tag{35}$$

Because we have

$$\mathbb{E}[e_t^j \mid g_t^j] = 0,$$

$$\begin{aligned}
&\| g_t^j - H_t^{-1} g_t^0 - \tau_j \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0 \|_2^4 \\
&= \| (I - \tau_j H_t)(g_t^j - H_t^{-1} g_t^0) + \tau_j (- \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} + H_t g_t^j) \|_2^4 \\
&= (\| (I - \tau_j H_t)(g_t^j - H_t^{-1} g_t^0) \|_2^2 \\
&+ 2\tau_j \langle (I - \tau_j H_t)(g_t^j - H_t^{-1} g_t^0), - \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} + H_t g_t^j \rangle \\
&+ \tau_j^2 \| - \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} + H_t g_t^j \|_2^2)^2
\end{aligned}$$

using Lemma 2

$$\begin{aligned}
&\leq (\| (I - \tau_j H_t)(g_t^j - H_t^{-1} g_t^0) \|_2^2 + 2\tau_j \| I - \tau_j H_t \|_2 \| g_t^j - H_t^{-1} g_t^0 \|_2 \delta_t^j \| g_t^j \|_2 + \tau_j^2 {\delta_t^j}^2 \| g_t^j \|_2^2)^2 \\
&= \| (I - \tau_j H_t)(g_t^j - H_t^{-1} g_t^0) \|_2^4 \\
&+ 2\tau_j \| (I - \tau_j H_t)(g_t^j - H_t^{-1} g_t^0) \|_2^2 (2\delta_t^j \| I - \tau_j H_t \|_2 \| g_t^j - H_t^{-1} g_t^0 \|_2 \| g_t^j \|_2 + \tau_j {\delta_t^j}^2 \| g_t^j \|_2^2) \\
&+ \tau_j^2 (2\delta_t^j \| I - \tau_j H_t \|_2 \| g_t^j - H_t^{-1} g_t^0 \|_2 \| g_t^j \|_2 + \tau_j {\delta_t^j}^2 \| g_t^j \|_2^2)^2
\end{aligned}$$

by our choice of $\tau_j = \Theta((j+1)^{-d_i}) = o(1)$ (25)

and using $\| g_t^j \|_2 \leq \| g_t^j - H_t^{-1} g_t^0 \|_2 + \| H_t^{-1} g_t^0 \|_2 \lesssim \| g_t^j - H_t^{-1} g_t^0 \|_2 +^4 \| g_t^0 \|_2$

$$\begin{aligned}
&= (1 - \Theta(\tau_j)) \| g_t^j - H_t^{-1} g_t^0 \|_2^4 \\
&+ O(\tau_j \delta_t^j (\| g_t^j - H_t^{-1} g_t^0 \|_2^4 + \| g_t^j - H_t^{-1} g_t^0 \|_2^3 \| g_t^0 \|_2) + 2\tau_j^2 {\delta_t^j}^3 (\| g_t^j - H_t^{-1} g_t^0 \|_2^4 + \| g_t^j - H_t^{-1} g_t^0 \|_2^2 \| g_t^0 \|_2^2) \\
&\quad + \tau_j^2 {\delta_t^j}^2 (\| g_t^j - H_t^{-1} g_t^0 \|_2^4 + \| g_t^j - H_t^{-1} g_t^0 \|_2^2 \| g_t^0 \|_2^2 + \tau_j \delta_t^j (\| g_t^j - H_t^{-1} g_t^0 \|_2^4 + \| g_t^0 \|_2^4))),
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}[\| e_t^j \|_2^4 \mid g_t^j] \\
&= \mathbb{E}[\| \left( \frac{1}{S_i} \frac{1}{\delta_t^j} \sum_{k \in I_i} (\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)) \right) - \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} \|_2^4 \mid g_t^j] \\
&= \mathbb{E}[\| \left( \frac{1}{S_i} \frac{1}{\delta_t^j} \sum_{k \in I_i} ((\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)) - H_t^k g_t^j + H_t^k g_t^j) \right) \\
&\quad - (\frac{1}{\delta_t^j} (\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)) - H_t g_t^j + H_t g_t^j) \|_2^4 \mid g_t^j]
\end{aligned}$$

using Lemma 2 and repeatedly applying the AM-GM inequality

$$\begin{aligned}
&\lesssim (1 + {\delta_t^j}^4) \| g_t^j \|_2^4 \\
&\lesssim (1 + {\delta_t^j}^4) {\delta_t^j}^4 (\| g_t^j - H_t^{-1} g_t^0 \|_2^4 + \| g_t^0 \|_2^4),
\end{aligned}$$

and by our choice of $\tau_j = \Theta((j+1)^{-d_i}) = o(1)$ (25) and $\delta_t^j = O(\tau_j^4)$ (26), after repeatedly applying the AM-GM inequality, Lemma 2, triangle inequality, and (27), we can bound (35) by

$$\begin{aligned}&\mathbb{E}[\|g_t^{j+1} - H_t^{-1} g_t^0\|_2^4 \mid g_t^j] \\ \leq &(1 - \Theta(\tau_j))\|g_t^j - H_t^{-1} g_t^0\|_2^4 + O(\tau_j^3 \|g_t^0\|_2^4)).\end{aligned} \tag{36}$$

Applying Lemma 6, we have

$$\mathbb{E}[\|g_t^{j+1} - H_t^{-1} g_t^0\|_2^4 \mid \theta_t] = O((j+1)^{-2d_i} \|g_t^0\|_2^4), \tag{37}$$

and using the AM-GM in equality we have

$$\mathbb{E}[\|g_t^{j+1}\|_2^4 \mid \theta_t] = O(\|g_t^0\|_2^4). \tag{38}$$

### E.1.3 Proof of (6)

To prove bounds on $\|\theta_t - \widehat{\theta}\|_2^2$, we will use the following lemma

**Lemma 3.**

$$\begin{aligned}\mathbb{E}[\langle \nabla f(\theta_t), -g_t^L \rangle \mid \theta_t] \gtrsim &\rho_t \|\nabla f(\theta_t)\|_2^2 - \delta_t^0 \|\nabla f(\theta_t)\|_2 \|g_t^0\|_2 \\ \gtrsim &\rho_t \|\nabla f(\theta_t)\|_2^2 - \delta_t^{0^2} \|g_t^0\|_2^2.\end{aligned}$$

*Proof.* Using (29), and because $\mathbb{E}[e_t^j \mid \theta_t = 0]$, we have

$$\begin{aligned}&\mathbb{E}[\langle \nabla f(\theta_t), -g_t^L \rangle \mid \theta_t] \\ &= \rho_t \nabla f(\theta_t)^\top H_t^{-1} \nabla f(\theta_t) - \mathbb{E}\left[\left\langle \nabla f(\theta_t), \sum_{k=0}^{L-1} \Big( \prod_{l=k+1}^{L-1} (I - \tau_l H_t) \Big) \tau_k \big( \tfrac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} - H_t g_t^k \big) \right\rangle \,\middle|\, \theta_t \right]\end{aligned}$$

using strong convexity and Lemma 2

$$\geq \frac{1}{\beta} \rho_t \|\nabla f(\theta_t)\|_2^2 - \|\nabla f(\theta_t)\|_2 \underbrace{\mathbb{E}\left[ \sum_{k=0}^{L-1} \prod_{l=k+1}^{L-1} \|I - \tau_l H_t\|_2 \tau_k \delta_t^k \|g_t^k\|_2 \,\middle|\, \theta_t \right]}_{[8]}.$$

By our choice of $\tau_j = \Theta((j+1)^{-d_i}) = o(1)$ (25) and $\delta_t^j = O(\delta_t^0 \tau_j^4)$ (26), and using (28), term [8] is bounded by

$$\begin{aligned}&\mathbb{E}\left[ \sum_{k=0}^{L-1} \prod_{l=k+1}^{L-1} \|I - \tau_l H_t\|_2 \tau_k \delta_t^k \|g_t^k\|_2 \mid \theta_t \right] \\ \lesssim &\sum_{k=0}^{L-1} \tau_k \delta_t^k \\ \lesssim &\|g_t^0\|_2 \delta_t^0 \underbrace{\sum_{k=0}^{L-1} \tau_k^5}_{=O(1)}.\end{aligned}$$

And we can conclude

$$\begin{aligned}
&\mathbb{E}[\langle \nabla f(\theta_t), -g_t^L \rangle \mid \theta_t] \\
&\geq C_1 \rho_t \|\nabla f(\theta_t)\|_2^2 - C_2 \delta_t^0 \|\nabla f(\theta_t)\|_2 \|g_t^0\|_2 \\
&= C_1 \rho_t \|\nabla f(\theta_t)\|_2^2 - \frac{C_1}{2} {\delta_t^0}^2 \left[ 2 \frac{\|\nabla f(\theta_t)\|_2}{\delta_t^0} \frac{C_2}{C_1} \|g_t^0\|_2 \right] \\
&\geq C_1 \rho_t \|\nabla f(\theta_t)\|_2^2 - \frac{C_1}{2} {\delta_t^0}^2 \left( \left( \frac{\|\nabla f(\theta_t)\|_2}{\delta_t^0} \right)^2 + \left( \frac{C_2}{C_1} \|g_t^0\|_2 \right)^2 \right) \\
&= \frac{C_1}{2} \rho_t \|\nabla f(\theta_t)\|_2^2 - \frac{C_2^2}{2C_1} {\delta_t^0}^2 \|g_t^0\|_2^2,
\end{aligned}$$

for some (data dependent) positive constants $C_1, C_2$.

□

Now, we continue our proof of (6).
In Algorithm 1, because $f$ is $\beta$ smooth, we have

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1}) - f(\widehat{\theta}) \mid \theta_t] \\
&= \mathbb{E}[f(\theta_t + g_t^L) - f(\widehat{\theta}) \mid \theta_t] \\
\leq & f(\theta_t) - f(\widehat{\theta}) + \mathbb{E}\left[ \langle \nabla f(\theta_t), g_t^L \rangle + \frac{\beta}{2} \|g_t^L\|_2^2 \mid \theta_t \right] \\
& \text{using Lemma 3 and (28)} \\
\leq & f(\theta_t) - f(\widehat{\theta}) - \Omega(\rho_t \|\nabla f(\theta_t)\|_2^2) + \mathbb{E}[O(\|g_t^0\|_2^2 + \delta_t^0 \|g_t^0\|_2 \|\nabla f(\theta_t)\|_2) \mid \theta_t]. \quad (39)
\end{aligned}$$

For $g_t^0$, we have

$$\begin{aligned}
\frac{g_t^0}{\rho_t} &= \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\theta_t) \\
&= \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\widehat{\theta}) + \frac{1}{S_o} \sum_{i \in I_o} (\nabla f_i(\theta_t) - \nabla f_i(\widehat{\theta})), \quad (40)
\end{aligned}$$

which implies that

$$\begin{aligned}
&\mathbb{E}\left[ \left\| \frac{g_t^0}{\rho_t} \right\|_2^2 \mid \theta_t \right] \\
\leq & 2\mathbb{E}\left[ \left\| \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\widehat{\theta}) \|_2^2 \mid \theta_t \right] + 2\mathbb{E}[\| \frac{1}{S_o} \sum_{i \in I_o} (\nabla f_i(\theta_t) - \nabla f_i(\widehat{\theta})) \right\|_2^2 \mid \theta_t \Bigg] \\
& \text{because we sample uniformly with replacement and } \nabla f(\widehat{\theta}) = 0 \\
\leq & \frac{2}{S_o} \sum_{i=1}^{n} \|\nabla f_i(\widehat{\theta})\|_2^2 + \mathbb{E}[\|\nabla f_i(\theta_t) - \nabla f_i(\widehat{\theta})\|_2^2 \mid \theta_t]
\end{aligned}$$

$$\leq \frac{2}{S_o}\sum_{i=1}^{n}\|\nabla f_i(\widehat{\theta})\|_2^2 + \|\theta_t-\widehat{\theta}\|_2^2\mathbb{E}[\beta_i^2 \mid \theta_t]$$
$$\lesssim 1+\|\theta_t-\widehat{\theta}\|_2^2. \tag{41}$$

Thus, continuing (39), using (41) and strong convexity $\alpha^2\|\theta_t-\widehat{\theta}\|_2^2 \leq \|\nabla f(\theta_t)\|_2^2$, we have

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1})-f(\widehat{\theta}) \mid \theta_t]\\
\leq& f(\theta_t)-f(\widehat{\theta}) - C_1\rho_t\|\nabla f(\theta_t)\|_2^2 + C_2\rho_t\delta_t^0(1+\|\nabla f(\theta_t)\|_2)\|\nabla f(\theta_t)\|_2 + C_3\rho_t^2(1+\|\nabla f(\theta_t)\|_2^2)\\
=& f(\theta_t)-f(\widehat{\theta}) - \rho_t(C_1 - C_2\delta_t^0 - C_3\rho_t)\|\nabla f(\theta_t)\|_2^2 + C_3\rho_t^2 + C_2\rho_t\delta_t^0\|\nabla f(\theta_t)\|_2\\
&\text{because we have } C_2\rho_t\delta_t^0\|\nabla f(\theta_t)\|_2 = \frac{1}{2}C_1\rho_t\delta_t^{0\,2}\frac{\frac{C_2}{C_1}\|\nabla f(\theta_t)\|_2}{\delta_t^0} \leq \frac{1}{2}C_1\rho_t\delta_t^{0\,2}\left(\left(\frac{C_2}{C_1}\right)^2 + \left(\frac{\|\nabla f(\theta_t)\|_2}{\delta_t^0}\right)^2\right)\\
\leq& f(\theta_t)-f(\widehat{\theta}) - \rho_t(\tfrac{1}{2}C_1 - C_2\delta_t^0 - C_3\rho_t)\|\nabla f(\theta_t)\|_2^2 + C_3\rho_t^2 + \tfrac{C_2^2}{C_1}\rho_t\delta_t^{0\,2}\\
&\text{using strong convexity } \frac{1}{2\alpha}\|\nabla f(\theta_t)\|_2^2 \geq f(\theta_t)-f(\widehat{\theta}) \text{ and smoothness } \frac{1}{2\beta}\|\nabla f(\theta_t)\|_2^2 \leq f(\theta_t)-f(\widehat{\theta})\\
\leq& [f(\theta_t)-f(\widehat{\theta})] - \rho_t(\tfrac{1}{2}C_1 - C_2\delta_t^0 - C_3\rho_t)\tfrac{1}{2\alpha}[f(\theta_t)-f(\widehat{\theta})] + C_3\rho_t^2 + \tfrac{C_2^2}{C_1}\rho_t\delta_t^{0\,2}\\
&\text{when we set } \delta_t^0 = O(\rho_t) \text{ in (26)}\\
\leq& [f(\theta_t)-f(\widehat{\theta})] - \rho_t(\tfrac{1}{2}C_1 - C_2\delta_t^0 - C_3\rho_t)\tfrac{1}{2\alpha}[f(\theta_t)-f(\widehat{\theta})] + (C_3+O(1))\rho_t^2,
\end{aligned} \tag{42}$$

for some (data dependent) positive constants $C_1, C_2, C_3$.

In (42) we choose $\rho_t = \Theta((t+1)^{-d_o})$ for some $d_o \in (\frac{1}{2}, 1)$, and after applying Lemma 6 we have

$$\begin{aligned}
&\mathbb{E}[\|\theta_t-\widehat{\theta}\|_2^2]\\
\leq& \mathbb{E}[\tfrac{2}{\alpha}(f(\theta_t)-f(\widehat{\theta}))]\\
\lesssim& t^{-d_o} + e^{-\Theta(t^{1-d_o})}\|\theta_0-\widehat{\theta}\|_2^2,
\end{aligned} \tag{43}$$

which is $O(t^{-d_o})$ when $\|\theta_0-\widehat{\theta}\|_2 = O(1)$.

#### E.1.4 Proof of (7)

In Algorithm 1, because $f$ is $\beta$ smooth, and $\forall\theta\ f(\theta)-f(\widehat{\theta}) \geq 0$, we have

$$\begin{aligned}
&(f(\theta_{t+1})-f(\widehat{\theta}))^2\\
=& (f(\theta_t+g_t^L)-f(\widehat{\theta}))^2\\
\leq& (f(\theta_t)-f(\widehat{\theta}) + \langle\nabla f(\theta_t), g_t^L\rangle + \tfrac{\beta}{2}\|g_t^L\|_2^2)^2\\
=& (f(\theta_t)-f(\widehat{\theta}))^2 + 2\langle\nabla f(\theta_t), g_t^L\rangle(f(\theta_t)-f(\widehat{\theta}))\\
&+ \langle\nabla f(\theta_t), g_t^L\rangle^2 + \tfrac{\beta^2}{4}\|g_t^L\|_2^4 + 2(f(\theta_t)-f(\widehat{\theta}) + \langle\nabla f(\theta_t), g_t^L\rangle)\tfrac{\beta}{2}\|g_t^L\|_2^2.
\end{aligned}$$

Because we have

$$\begin{aligned}&\mathbb{E}[\langle \nabla f(\theta_t), g_t^L\rangle (f(\theta_t) - f(\widehat{\theta})) \mid \theta_t]\\ \lesssim& -\rho_t \|\nabla f(\theta_t)\|_2^2 (f(\theta_t) - f(\widehat{\theta})) + \delta_t^0 \|g_t^0\|_2^2 (f(\theta_t) - f(\widehat{\theta})),\end{aligned}$$

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{g_t^0}{\rho_t}\right\|_2^4 \mid \theta_t\right]\\ &= \mathbb{E}\left[\left\|\frac{1}{S_o}\sum_{i\in I_o}(\nabla f_i(\theta_t) - \nabla f_i(\widehat{\theta}) + \nabla f_i(\widehat{\theta}))\right\|_2^4 \mid \theta_t\right]\\ \lesssim& 1 + \|\theta_t - \widehat{\theta}\|_2^4,\end{aligned}$$

$$f(\theta_t) - f(\widehat{\theta}) = \Theta(\|\theta_t - \widehat{\theta}\|_2^2) = \Theta(\|\nabla f(\theta_t)\|_2^2),$$

and by our choice of $\rho_t = \Theta((t+1)^{-d_o}) = o(1)$ and $\delta_t^0 = O(\rho_t^4)$ (26), after repeatedly applying the AM-GM inequality and (43), we have

$$\begin{aligned}&\mathbb{E}[(f(\theta_{t+1}) - f(\widehat{\theta}))^2 \mid \theta_t]\\ \leq& (1 - \Theta(\rho_t))(f(\theta_t) - f(\widehat{\theta}))^2 + O(\rho_t^3).\end{aligned}$$

Applying Lemma 6, we have

$$\begin{aligned}&\mathbb{E}[\|\theta_t - \widehat{\theta}\|_2^4]\\ \leq& \mathbb{E}\left[\tfrac{4}{\alpha^2}(f(\theta_t) - f(\widehat{\theta}))^2\right]\\ \lesssim& t^{-2d_o}.\end{aligned} \tag{44}$$

#### E.1.5 Proof of (10)

For $\frac{\bar{g}_t}{\rho_t}$, we have

$$\begin{aligned}\frac{\bar{g}_t}{\rho_t} =& \underbrace{-H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})}_{[1]}\\ &\underbrace{+ H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta}) - H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta}) + H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta}) - H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\theta_t)}_{[2]}\\ &\underbrace{-H_t^{-1}\frac{g_t^0}{\rho_t} + \frac{\bar{g}_t}{\rho_t}}_{[3]}.\end{aligned} \tag{45}$$

Thus, for the "covariance" of our replicates, we have

$$
\begin{aligned}
&\left\| H^{-1}GH^{-1} - \frac{S_o}{T}\sum_{t=1}^{T}\frac{\bar{g}_t\bar{g}_t^\top}{\rho_t^2}\right\|_2 \\
\lesssim &\left\| H^{-1}GH^{-1} - \frac{S_o}{T}\sum_{t=1}^{T}[1]_t[1]_t^\top\right\|_2 \\
&+ \left\|\frac{S_o}{T}\sum_{t=1}^{T}[1]_t([2]_t+[3]_t)^\top\right\|_2 + \left\|\frac{S_o}{T}\sum_{t=1}^{T}([2]_t+[3]_t)[1]_t^\top\right\|_2 \\
&+ \left\|\frac{S_o}{T}\sum_{t=1}^{T}([2]_t+[3]_t)([2]_t+[3]_t)^\top\right\|_2 \\
&\text{because for two vectors } a,b \text{ the operator norm } \|ab^\top\|_2 \le \|a\|_2\|b\|_2 \\
\lesssim &\left\| H^{-1}GH^{-1} - \frac{S_o}{T}\sum_{t=1}^{T}[1]_t[1]_t^\top\right\|_2 \\
&+ \frac{1}{T}\sum_{t=1}^{T}\|[1]_t\|_2(\|[2]_t\|_2+\|[3]_t\|_2) \\
&+ \frac{1}{T}\sum_{t=1}^{T}(\|[2]_t\|_2^2+\|[3]_t\|_2^2).
\end{aligned}
$$

Because $\sum_{t=1}^{T}[1]_t$ consists of $S_o\cdot T$ i.i.d. samples from $\{H^{-1}\nabla f_i(htheta)\}_{i=1}^n$ and the mean $H^{-1}\nabla f(\widehat{\theta})=0$, using matrix concentration [51], we know that

$$
\mathbb{E}\left[\left\| H^{-1}GH^{-1} - \frac{S_o}{T}\sum_{t=1}^{T}[1]_t[1]_t^\top\right\|_2\right] \lesssim \frac{1}{\sqrt{T}}.
$$

For term [3], using (30), because we have

$$
\begin{aligned}
&\sum_{t=1}^{T}[3]_t \\
&= \underbrace{\sum_{t=1}^{T}\frac{1}{L+1}\sum_{k=0}^{L-1}\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I-\tau_l H_t)(-e_t^k)}_{\text{when } a\neq b\ \mathbb{E}[\langle e_t^a, e_t^b\rangle]=0} \\
&+ \sum_{t=1}^{T}\frac{1}{L+1}\sum_{k=0}^{L-1}\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I-\tau_l H_t)(-\tfrac{\nabla f(\theta_t+\delta_t^k g_t^k)-\nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k),
\end{aligned}
$$

by using (31) and (33), we have

$$
\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}[3]_t\right\|_2^2\right]
$$

$$\lesssim \mathbb{E}\left[\frac{1}{T}\Big(\sum_{t=1}^{T}\frac{1}{L}+\Big(\sum_{t=1}^{T}\frac{\sum_{k=0}^{L}\delta_t^k}{L}\Big)^2\Big)\left\|\frac{g_t^0}{\rho_t}\right\|_2^2\right]$$

using (41), and by our choice of $\delta_t^k = \delta_t^0 o((k+1)^{-2})$ and $\delta_t^0 = o((t+1)^{-2})$ (26)

$$\lesssim \mathbb{E}\left[\left(\frac{1}{L}+\frac{\sum_{t=1}^{T}\delta_t^{0\,2}}{T}\right)\left(1+\|\theta_t-\widehat{\theta}\|_2^2\right)\right]$$

$$\lesssim \frac{1}{L}+\frac{1}{T}. \tag{46}$$

And because we have

$$\mathbb{E}[\|[1]_t\|_2] = \mathbb{E}[\|-H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})\|_2] = O(1),$$

$$\mathbb{E}[\|[2]_t\|_2^2 \mid \theta_t]$$

$$\lesssim \mathbb{E}\left[\left\|(H^{-1}-H_t^{-1})\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})\right\|_2^2+\left\|H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}(\nabla f_i(\widehat{\theta})-\nabla f_i(\theta_t))\right\|_2^2 \mid \theta_t\right]$$

because $H^{-1}-H_t^{-1} = H^{-1}(H_t-H)H_t^{-1}$ and using Lemma 2 (47)

$$\lesssim \mathbb{E}[\|\theta_t-\widehat{\theta}\|_2^2 \mid \theta_t]$$

$$\lesssim (t+1)^{-d_o}, \tag{48}$$

by repeatedly applying Cauchy-Schwarz inequality and AM-GM inequality, we can conclude that

$$\mathbb{E}\left[\left\|H^{-1}GH^{-1}-\frac{S_o}{T}\sum_{t=1}^{T}\frac{\bar{g}_t\bar{g}_t^\top}{\rho_t^2}\right\|_2\right]$$

$$\lesssim \frac{1}{\sqrt{T}}+\frac{1}{T}\sum_{t=1}^{T}(t+1)^{-\frac{d_o}{2}}+\frac{1}{T}\sum_{t=1}^{T}(t+1)^{-d_o}+\frac{1}{\sqrt{L}}+\frac{1}{L}$$

because $\sum_{t=1}^{T}(t+1)^{-\frac{d_o}{2}} = T^{1-\frac{d_o}{2}}$ for $d_o \in (\frac{1}{2},1)$

$$\lesssim \frac{1}{T^{\frac{d_o}{2}}}+\frac{1}{\sqrt{L}}.$$

## E.2 Proof of Corollary 1

For $\frac{g_t^L}{\rho_t}$, we have

$$\frac{g_t^L}{\rho_t} = \underbrace{-H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})}_{[1]}$$

$$\underbrace{+H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})-H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})+H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})-H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\theta_t)+H_t^{-1}\nabla f(\theta_t)}_{[2]}$$

$$\underbrace{-H_t^{-1}\nabla f(\theta_t)+(\theta_t-\widehat{\theta})}_{[3]}\underbrace{-H_t^{-1}\frac{g_t^0}{\rho_t}+\frac{g_t^L}{\rho_t}}_{[4]}-(\theta_t-\widehat{\theta}), \tag{49}$$

which gives

$$\begin{aligned}
&\theta_t-\widehat{\theta}\\
&=(1-\rho_{t-1})(\theta_{t-1}-\widehat{\theta})+\rho_{t-1}([1]_{t-1}+[2]_{t-1}+[3]_{t-1}+[4]_{t-1})\\
&=(\prod_{i=0}^{t-1}(1-\rho_i))(\theta_0-\widehat{\theta})+\sum_{i=0}^{t-1}(\prod_{j=i+1}^{t-1}(1-\rho_j))\rho_i([1]_i+[2]_i+[3]_i+[4]_i).
\end{aligned}$$

And we have

$$\begin{aligned}
&\sqrt{T}(\frac{\sum_{t=1}^T\theta_t}{T}-\widehat{\theta})\\
&=\frac{1}{\sqrt{T}}(\sum_{t=1}^T\prod_{i=0}^{t-1}(1-\rho_i))(\theta_0-\widehat{\theta})+\frac{1}{\sqrt{T}}\sum_{t=1}^T\sum_{i=0}^{t-1}(\prod_{j=i+1}^{t-1}(1-\rho_j))\rho_i([1]_i+[2]_i+[3]_i+[4]_i)\\
&=\frac{1}{\sqrt{T}}(\sum_{t=1}^T\prod_{i=0}^{t-1}(1-\rho_i))(\theta_0-\widehat{\theta})+\frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}\sum_{t=i+1}^{T}(\prod_{j=i+1}^{t-1}(1-\rho_j))\rho_i([1]_i+[2]_i+[3]_i+[4]_i).
\end{aligned} \tag{50}$$

For the first term in (50), which is non-stochastic, we have

$$\left\|\frac{1}{\sqrt{T}}(\sum_{t=1}^T\prod_{i=0}^{t-1}(1-\rho_i))(\theta_0-\widehat{\theta})\right\|_2\lesssim\frac{1}{\sqrt{T}}.$$

For the second term in (50), which is stochastic, we first consider $\rho_i\sum_{t=i+1}^T\prod_{j=i+1}^{t-1}(1-\rho_j)$, which is $O(1)$ (similar to (31)) and satisfies

$$\begin{aligned}
&\rho_i\sum_{t=i+1}^T\prod_{j=i+1}^{t-1}(1-\rho_j)\\
&=\sum_{t=i+1}^T\frac{\rho_i}{\rho_t}\rho_t\prod_{j=i+1}^{t-1}(1-\rho_j)\\
\le&\frac{\rho_i}{\rho_s}\sum_{t=i+1}^s\rho_t\prod_{j=i+1}^{t-1}(1-\rho_j)+\rho_i(\prod_{j=i+1}^s(1-\rho_j))\sum_{t=s+1}^T\prod_{j=s+1}^{t-1}(1-\rho_j)\\
&=(1+\frac{\rho_i-\rho_s}{\rho_s})(1-\prod_{t=i+1}^s(1-\rho_t))+\rho_i(\prod_{j=i+1}^s(1-\rho_j))\sum_{t=s+1}^T\prod_{j=s+1}^{t-1}(1-\rho_j)
\end{aligned}$$

$$\leq (1+\frac{\rho_i-\rho_s}{\rho_s})(1-(1-\rho_s)^{s-i})+\rho_i(1-\rho_s)^{s-i}\sum_{t=s+1}^{T}\prod_{j=s+1}^{t-1}(1-\rho_j)$$
$$\leq 1+((1+\frac{s-i}{i+1})^{d_o}-1)+\rho_i e^{-(s-i)\rho_s}\sum_{t=s+1}^{\infty}\prod_{j=s+1}^{t-1}(1-\rho_j)$$
$$\leq 1+\frac{s-i}{i}+\rho_i e^{-(s-i)\rho_s}\sum_{t=s+1}^{\infty}\prod_{j=s+1}^{t-1}(1-\rho_j),$$

for all $i \leq s \leq T$, and

$$\begin{aligned}
&\rho_i \sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}(1-\rho_j)\\
&\geq \sum_{t=i+1}^{T}(\prod_{j=i+1}^{t-1}(1-\rho_j))\rho_t\\
&= 1-\prod_{t=i+1}^{T}(1-\rho_t)\\
&\geq 1-\exp(-\sum_{t=i+1}^{T}\rho_t)\\
&\geq 1-\exp(-\frac{1}{1-d_o}((T+2)^{1-d_o}-(i+2)^{1-d_o}))
\end{aligned}$$

When we choose $s = i+\lceil (i+1)^{\frac{d_o+1}{2}}\rceil$, we have $\frac{s-i}{i} \lesssim i^{\frac{-1+d_o}{2}}$, $(s-i)\rho_s \gtrsim (i+1)^{\frac{1-d_o}{2}}$, and $\rho_i e^{-\frac{1}{2}(s-i)\rho_s} \lesssim \rho_s$. And these imply $|\rho_i \sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}(1-\rho_j)-1| = O(\max\{(i+1)^{\frac{-1+d_o}{2}}, \exp(-\frac{1}{1-d_o}((T+2)^{1-d_o}-(i+2)^{1-d_o})\})$. Thus, for term [1], we have

$$\frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}(1-\rho_j)\rho_i[1]_i = \frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}[1]_i + \frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}(\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}(1-\rho_j)\rho_i-1)[1]_i,$$

where the first term weakly converges to $\mathcal{N}(0, \frac{1}{S_o}H^{-1}GH^{-1})$ by Central Limit Theorem, and the second term satisfies $\mathbb{E}[\|\frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}(\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}(1-\rho_j))\rho_i - 1)[1]_i\|_2^2] = \mathbb{E}[\frac{1}{T}\sum_{i=0}^{T-1}|(\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}(1-\rho_j))\rho_i-1)|^2\|[1]_i\|_2^2] \lesssim T^{d_o-1}+\frac{1}{T}$.

For term [2], we have

$$\|[2]_t\|_2 \lesssim \|\theta_t - \widehat{\theta}\|_2,$$

and $\mathbb{E}[\langle [2]_a, [2]_b\rangle] = 0$ when $a \neq b$. Thus

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}(1-\rho_j)\rho_i[2]_i\right\|_2^2\right] \lesssim \frac{1}{T}\sum_{i=0}^{T-1}\|\theta_t-\widehat{\theta}\|_2^2 \lesssim T^{-d_o}.$$

For term [3], we have

$$\| - H_t^{-1}\nabla f(\theta_t) + (\theta_t - \widehat{\theta})\|_2 \lesssim \|\theta_t - \widehat{\theta}\|_2^2.$$

By using (7) and Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}(1-\rho_j)\rho_i[3]_i\right\|_2^2\right] \lesssim T^{1-2d_o}.$$

For term [4], similar to similar to (46), we have

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}(1-\rho_j)\rho_i[4]_i\right\|_2^2\right] \lesssim \frac{1}{T} + \frac{1}{L}.$$

## E.3 Proof of Corollary 3

Using Theorem 6.5 of [9], we have

$$\mathbb{E}[\|\theta_t - \widehat{\theta}\|_2^2] \lesssim 0.9^t.$$

Similar to (8) in Theorem 1 (Appendix E.1.1), we have

$$\mathbb{E}\left[\left\|\frac{\bar{g}_t}{\rho_t} - [\nabla^2 f(\theta_t)]^{-1} g_t^0\right\|_2^2 \mid \theta_t\right] \lesssim \tfrac{1}{L}\|g_t^0\|_2^2.$$

Similar to the proof of (10) in Theorem 1 (Appendix E.1.5), using (45), we have

$$\mathbb{E}\left[\left\|H^{-1}GH^{-1} - \frac{S_o}{T}\sum_{t=1}^{T}\frac{\bar{g}_t\bar{g}_t^\top}{\rho_t^2}\right\|_2\right] \lesssim L^{-\frac{1}{2}}.$$

For $\frac{g_t^L}{\rho_t}$, we have

$$\begin{aligned}\frac{\bar{g}_t}{\rho_t} =& \underbrace{-H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})}_{[1]} \\ &+ \underbrace{H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta}) - H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta}) + H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta}) - H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\theta_t) + H_t^{-1}\nabla f(\theta_t)}_{[2]} \\ &\underbrace{-H_t^{-1}\nabla f(\theta_t)}_{[3]}\underbrace{-H_t^{-1}\frac{g_t^0}{\rho_t} + \frac{g_t^L}{\rho_t}}_{[4]}. \end{aligned} \tag{51}$$

For term [1], we have

$$\frac{1}{\sqrt{T}}\sum_{i=1}^{T}[1]_t = \frac{1}{\sqrt{T}}\sum_{i=1}^{T}\left(-H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})\right)_t,$$

which consists of $S_o \cot T$ i.i.d samples from 0 mean set $\{H_{-1}\nabla f_i(\widehat{\theta})\}_{i=1}^n$, and weakly converges to $\mathcal{N}(0, \frac{1}{S_o}H^{-1}GH^{-1})$ by the Central Limit Theorem.

For term [2], similar to (47), we have

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{i=1}^{T}[2]_t\right\|_2^2\right] = \frac{1}{T}\mathbb{E}[\sum_{i=1}^{T}\|[2]_t\|_2^2] \lesssim \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\theta_t - \widehat{\theta}\|_2^2] \lesssim \frac{1}{T}.$$

For term [3], we have

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{i=1}^{T}[3]_t\right\|_2\right] \lesssim \frac{1}{\sqrt{T}}\mathbb{E}[\|\theta_t - \widehat{\theta}\|_2] \lesssim \frac{1}{\sqrt{T}}.$$

For term [4], similar to (46), we have

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{i=1}^{T}[4]_t\right\|_2\right] \lesssim \frac{1}{\sqrt{T}} + \frac{1}{\sqrt{L}}.$$

## E.4 Proof of Theorem 2

The error bound proof is similar to standard LASSO proofs [12, 41].

We will use Lemma 4 for the covariance estimate using soft thresholding.

We denote "soft thresholding by $\omega$" as an element-wise procedure $\mathbf{S}_\omega(A) = \text{sign}(A)(|A| - \omega)_+$ , where $A$ is an arbitrary number, vector, or matrix, and $\omega$ is non-negative.

**Lemma 4.** *Under our assumptions in Section 3, we choose soft threshold* $\frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ *using*

$$\omega = \Theta\left(\sqrt{\frac{\log p}{n}}\right).$$

*When* $n \gg \log p$*, the matrix max norm of* $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top - \Sigma$ *is bounded by*

$$\max_{1\le i,j\le p}\left|\left(\frac{1}{n}\sum_{i=1}^{n}x_i x_i^\top\right)_{ij} - \Sigma_{ij}\right| \lesssim \sqrt{\frac{\log p}{n}},$$

*with probability at least* $1 - p^{-\Theta(1)}$.

*Under this event, $\ell_2$ operator norm of $\widehat{S} - \Sigma$ satisfies*

$$\|\widehat{S} - \Sigma\|_2 \lesssim b\sqrt{\frac{\log p}{n}},$$

*$\ell_1$ and $\ell_\infty$ operator norm of $\widehat{S} - \Sigma$ satisfies*

$$\|\widehat{S} - \Sigma\|_\infty = \|\widehat{S} - \Sigma\|_1 \lesssim b\sqrt{\frac{\log p}{n}}.$$

*Proof.* The proof is similar to that of Theorem 1, [7].

Our assumption that $\Sigma$ is well conditioned implies that each off diagonal entry is bounded, and each diagonal entry is $\Theta(1)$ and positive.

Omitting the subscript for the $i^{\text{th}}$ sample, for each i.i.d. sample $x = [x(1), x(2), \ldots, x(p)]^\top \sim \mathcal{N}(0, \Sigma)$, each $x(j)x(k)$ satisfies

$$x(j)x(k) = \frac{1}{4}(x(j) + x(k))^2 - \frac{1}{4}(x(j) - x(k))^2,$$

where $x(j) \pm x(k)$ are Gaussian random variables with variance $\Sigma_{jj} \pm 2\Sigma_{jk} + \Sigma_{kk} = \Theta(1)$, because all of $\Sigma$'s eigenvalues are upper and lower bounded. Thus, $x(j) \pm x(k)$ are $\chi_1^2$ random variables scaled by $\Sigma_{jj} \pm 2\Sigma_{jk} + \Sigma_{kk} = \Theta(1)$, and they are sub-exponential with parameters that are $\Theta(1)$ [57]. And this implies that, $x(j)x(k) - \Sigma_{jk}$ is sub-exponential

$$\mathbb{P}[|x(j)x(k) - \Sigma_{jk}| > t] \lesssim \exp(-\Theta(\min\{t^2, t\})),$$

for all $1 \leq j, k \leq p$.

Using Bernstein inequality [57], we have

$$\mathbb{P}\left[\left|\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)_{jk} - \Sigma_{jk}\right| > t\right] \lesssim \exp(-n\Theta(\min\{t^2, t\})),$$

for all $1 \leq j, k \leq p$.

Taking a union bound over all matrix entries, and using $n \gg \log p$, we have

$$\max_{1\leq j,k\leq p}\left|\left((\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)_{jk} - \Sigma_{jk}\right| \lesssim \sqrt{\frac{\log p + \log\frac{1}{\delta}}{n}},$$

with probability at least $1 - \delta$.

Under this event, the soft thresholding estimate $\mathbf{S}_\omega(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top)_{ij}$ with $\omega = \Theta(\sqrt{\frac{\log p}{n}})$ is 0 when $\Sigma_{ij} = 0$, and $|\Sigma_{ij} - \mathbf{S}_\omega(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top)_{ij}| \leq \omega$ (even when $|\Sigma_{ij}| \leq \omega$). And this implies our bounds.

□

Lemma 4 guarantees that the optimization problem (13) is well defined with high probability when $n \gg b\sqrt{\frac{\log p}{n}}$. Because the $\ell_2$ operator norm $\|\widehat{S} - \Sigma\|_2 \lesssim b\sqrt{\frac{\log p}{n}} \ll 1$, and the positive definite matrix $\Sigma$'s eigenvalues are all $\Theta(1)$, the symmetric matrix $\widehat{S}$ is positive definite, and $\widehat{S}$'s eigenvalues are all $\Theta(1)$, and for all $v \in \mathbb{R}^p$ we have

$$0 \leq v^\top \widehat{S} v = \Theta(\|v\|_2^2). \tag{52}$$

Because $\widehat{\theta}$ attains the minimum, by definition, we have

$$\begin{aligned}
&\tfrac{1}{2}\widehat{\theta}^\top \left(\widehat{S} - \tfrac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\widehat{\theta} + \tfrac{1}{n}\sum_{i=1}^n \tfrac{1}{2}(x_i^\top \widehat{\theta} - y_i)^2 + \lambda\|\widehat{\theta}\|_1 \\
\leq &\tfrac{1}{2}\theta^{\star\top}\left(\widehat{S} - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\theta^\star + \tfrac{1}{n}\sum_{i=1}^n \tfrac{1}{2}\left(x_i^\top\theta^\star - y_i\right)^2 + \lambda\|\theta^\star\|_1,
\end{aligned}$$

which, after rearranging terms, is equivalent to

$$\tfrac{1}{2}(\widehat{\theta} - \theta^\star)^\top \widehat{S}(\widehat{\theta} - \theta^\star) + \left\langle \left(\widehat{S} - \tfrac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\theta^\star + \tfrac{1}{n}\sum_{i=1}^n \epsilon_i x_i, \widehat{\theta} - \theta^\star \right\rangle \leq \lambda(\|\theta^\star\|_1 - \|\widehat{\theta}\|_1). \tag{53}$$

Because $\widehat{S} = \mathbf{S}_\omega(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top)$ soft thresholds each entry of $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ with $\omega = \Theta(\sqrt{\frac{\log p}{n}})$, each entry of $\widehat{S} - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ will lie in the interval $[-\omega, \omega]$. And this implies , with probability at least $1 - p^{-\Theta(1)}$, we have

$$\left\| \left(\widehat{S} - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\theta^\star \right\|_\infty \lesssim \sqrt{\frac{\log p}{n}}\|\theta^\star\|_1 \lesssim \sqrt{\frac{s\log p}{n}},$$

where we used the assumption that $\theta^\star$ is $s$ sparse and $\|\theta^\star\|_2 = O(1)$, which implies $\|\theta^\star\|_1 \lesssim \sqrt{s}$.

For the $j^{\text{th}}$ coordinate of $\epsilon_i x_i$, because $\epsilon_i$ and $x_i$ are independent Gaussian random variables, we know that it is sub-exponential [57]

$$\mathbb{P}[|\epsilon_i x_i(j)| > t] \lesssim \exp\left(-\Theta\left(\min\left\{\tfrac{t^2}{\sigma^2}, \tfrac{t}{\sigma}\right\}\right)\right), \tag{54}$$

for all $1 \leq i \leq n$ and $1 \leq j \leq p$.

Using Bernstein inequality, we have

$$\mathbb{P}[|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i(j)| > t] \lesssim \exp\left(-\Theta\left(n\min\left\{\tfrac{t^2}{\sigma^2}, \tfrac{t}{\sigma}\right\}\right)\right),$$

for all $1 \leq j \leq p$.

Taking a union bound over all $p$ coordinates, with probability at least $1 - p^{-\Theta(1)}$, we have

$$\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i x_i\|_\infty \lesssim \sigma\sqrt{\frac{\log p}{n}}, \tag{55}$$

when $n \gg \log p$.

Thus, we set the regularization parameter

$$\begin{aligned}\lambda =&\Theta\left((\sigma + \|\theta^\star\|_1)\sqrt{\frac{\log p}{n}}\right)\\ \geq& 2\left\|\left(\widehat{S} - \frac{1}{n}\sum_{i=1}^{n}x_i x_i^\top\right)\theta^\star + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i x_i\right\|_\infty,\end{aligned} \tag{56}$$

which holds under the events in Lemma 4 and (55).

For a vector $v \in \mathbb{R}^p$, let $v^S$ indicate the sub-vector of on the support of $\theta^\star$, and $v^{\bar{S}}$ the sub-vector not on the support of $\theta^\star$.

(53) and (56) implies that

$$-\tfrac{1}{2}\lambda(\|(\theta-\theta^\star)^S\|_1 + \|\theta^{\bar{S}}\|_1) = -\tfrac{1}{2}\lambda\|\theta-\theta^\star\|_1 \leq \lambda(\|\theta^\star\|_1 - \|\widehat{\theta}\|_1) \leq \lambda(\|(\theta-\theta^\star)^S\|_1 - \|\theta^{\bar{S}}\|_1),$$

which is equivalent to

$$\|\theta^{\bar{S}}\|_1 \leq 3\|(\theta-\theta^\star)^S\|_1, \tag{57}$$

because $\lambda > 0$.

For any vector $v \in \mathbb{R}^p$, it satisfies $\|v\|_2^2 \geq \|v^S\|_2^2 \geq \frac{1}{s}\|v^S\|_1^2$. Using this in (53), we have

$$\tfrac{1}{s}\|(\theta-\theta^\star)^S\|_1^2 \lesssim \lambda\|(\theta-\theta^\star)^S\|_1,$$

which implies that

$$\|(\theta-\theta^\star)^S\|_1 \lesssim s(\sigma + \|\theta^\star\|_1)\sqrt{\frac{\log p}{n}}. \tag{58}$$

Combining (58) and (57), we have proven (14)

$$\|\theta-\theta^\star\|_1 \lesssim s\left(\sigma + \|\theta^\star\|_1\right)\sqrt{\frac{\log p}{n}} \lesssim s\left(\sigma + \sqrt{s}\right)\sqrt{\frac{\log p}{n}}.$$

In (53) because $\langle(\widehat{S} - \frac{1}{n}\sum_{i=1}^{n}x_i x_i^\top)\theta^\star + \frac{1}{n}\sum_{i=1}\epsilon_i x_i, \widehat{\theta} - \theta^\star\rangle \geq 0$ by convexity, and using (52), we have proven (15)

$$\|\theta-\theta^\star\|_2^2 \lesssim \lambda\|(\theta-\theta^\star)^S\|_1 \lesssim s\left(\sigma + \|\theta^\star\|_1\right)^2\frac{\log p}{n} \lesssim s\left(\sigma + \sqrt{s}\right)^2\frac{\log p}{n}.$$

## E.5 Proof of Theorem 3

At the solution $\widehat{\theta}$ of the optimization problem (13), using the KKT condition, we have

$$\left(\widehat{S} - \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\widehat{\theta} + \frac{1}{n}\sum_{i=1}^{n} x_i(x_i^\top\widehat{\theta} - y_i) + \lambda\widehat{g} = 0, \tag{59}$$

where $\widehat{g} \in \partial\|\widehat{\theta}\|_1$. And this is equivalent to

$$\widehat{S}\widehat{\theta} - \frac{1}{n}\sum_{i=1}^{n} x_i(x_i^\top\theta^\star + \epsilon_i) + \lambda\widehat{g} = 0,. \tag{60}$$

By Lemma 4, we know that $\widehat{S}$ is invertible when $n \gg b^2 \log p$.

Plugging (16) into (60), we have

$$\widehat{S}(\widehat{\theta}^{\mathrm{d}} - \widehat{S}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\widehat{\theta}\right]) - \frac{1}{n}\sum_{i=1}^{n} x_i(x_i^\top\theta^\star + \epsilon_i) + \lambda\widehat{g} = 0,$$

which is equivalent to

$$\widehat{S}(\widehat{\theta}^{\mathrm{d}} - \theta^\star) - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i x_i + \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top - \widehat{S}\right)(\widehat{\theta} - \theta^\star) = 0, \tag{61}$$

where we used the fact that $\lambda\widehat{g} = -\widehat{S}\widehat{\theta} + \frac{1}{n}\sum_{i=1}^{n} x_i(x_i^\top\theta^\star + \epsilon_i)$.

Rewriting (61), we have

$$\widehat{\theta}^{\mathrm{d}} - \theta^\star = \widehat{S}^{-1}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i x_i + \left(I - \widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\right)(\widehat{\theta} - \theta^\star). \tag{62}$$

For $\max_{1\le j,k\le p}\left|\left(I - \widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\right)_{jk}\right|$, we have

$$\begin{aligned}
&\max_{1\le j,k\le p}\left|\left(I - \widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\right)_{jk}\right| \\
=&\max_{1\le j,k\le p}\left|\left(\widehat{S}^{-1}\left(S - \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\right)_{jk}\right| \\
\le&\|\widehat{S}^{-1}\|_\infty \max_{1\le j,k\le p}\left|\left(S - \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)_{jk}\right|.
\end{aligned} \tag{63}$$

Under the event in Lemma 4, we have

$$\max_{1\le j,k\le p}\left|\left(S - \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)_{jk}\right| \lesssim \sqrt{\frac{\log p}{n}}. \tag{64}$$

Also under the event in Lemma 4, we have

$$\widehat{S}_{ii} - \sum_{j\neq i} |\widehat{S}_{ij}| \geq \Sigma_{ii} - \Theta\left(\sqrt{\tfrac{\log p}{n}}\right) - \sum_{j\neq i} |\Sigma_{ij}| \geq D_\Sigma - \Theta\left(\sqrt{\tfrac{\log p}{n}}\right),$$

where we used $\widehat{S}_{ii} > 0$ and $|\Sigma_{ij}| \geq |\widehat{S}_{ij}|$ by definition of the soft thresholding operation.

Thus, when $n \gg \frac{1}{{D_\Sigma}^2} \log p$, we have

$$\widehat{S}_{ii} - \sum_{j\neq i} |\widehat{S}_{ij}| \gtrsim D_\Sigma,$$

which implies that $\widehat{S}$ is also diagonally dominant. Thus, using Theorem 1, [55], when $n \gg \frac{1}{{D_\Sigma}^2} \log p$, we have

$$\|\widehat{S}\|_\infty \lesssim \frac{1}{D_\Sigma}, \tag{65}$$

with probability at least $1 - p^{-\Theta(1)}$

And using (64) and (65) in (63), we have

$$\max_{1\leq j,k\leq p} \left| \left( I - \widehat{S}^{-1} (\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)_{jk} \right| \lesssim \frac{1}{D_\Sigma} \sqrt{\frac{\log p}{n}}. \tag{66}$$

Using (66) and the bound on $\|\widehat{\theta} - \theta^\star\|_1$ (14), in (62), we have

$$\left\| \left( I - \widehat{S}^{-1} \left( \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \right) (\widehat{\theta} - \theta^\star) \right\|_\infty \lesssim \frac{1}{D_\Sigma} s \left(\sigma + \|\theta^\star\|_1\right) \frac{\log p}{n} \lesssim \frac{1}{D_\Sigma} s \left(\sigma + \sqrt{s}\right) \frac{\log p}{n}. \tag{67}$$

Combining (67) and (62), we have proven Theorem 3, when $n \gg \max\{b^2, \frac{1}{{D_\Sigma}^2}\} \log p$, we have

$$\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^\star) = Z + R,$$

where $Z \mid \{x_i\}_{i=1}^n \sim \mathcal{N}\left(0, \sigma^2 \widehat{S}^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top\right) \widehat{S}^{-1}\right)$, and $\|R\|_\infty \lesssim \frac{1}{D_\Sigma} s \left(\sigma + \|\theta^\star\|_1\right) \frac{\log p}{\sqrt{n}} \lesssim \frac{1}{D_\Sigma} s \left(\sigma + \sqrt{s}\right) \frac{\log p}{\sqrt{n}}$ with probability at least $1 - p^{-\Theta(1)}$.

## E.6 Proof of Theorem 4

We analyze the optimization problem conditioned on the data set $\{x_i\}_{i=1}^n$, which satisfies Lemma 4 with probability at least $1 - p^{\Theta(-1)}$ when $n \gg b^2 \log p$.

Here, we denote the objective function as

$$P(\theta) = \frac{1}{2} \theta^\top \left( \widehat{S} - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \theta + \frac{1}{n} \sum_{i=1}^n \frac{1}{2} \left( x_i^\top \theta - y_i \right)^2 + \lambda \|\theta\|_1.$$

In Algorithm 2, lines 6 to 15 are using SVRG [34] to solve the Newton step

$$\min_{\Delta} \frac{1}{2}\Delta^\top \widehat{S}\Delta + \left\langle \frac{1}{S_o}\sum_{k\in I_o} \nabla f_k(\theta_t), \Delta \right\rangle, \tag{68}$$

and using proximal SVRG [61] to solve the proximal Newton step

$$\min_{\Delta} \frac{1}{2}\Delta^\top \widehat{S}\Delta + \left\langle \frac{1}{n}\sum_{k=1}^{n} \nabla f_k(\theta_t), \Delta \right\rangle + \lambda\|\theta+\Delta\|_1. \tag{69}$$

The gradient of (68) is

$$\widehat{S}\Delta + \frac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\theta_t) = \underbrace{\frac{1}{p}\sum_{k=1}^{p}\left[p\widehat{S}_k\right]\Delta(k)}_{\text{sample } \textit{by feature} \text{ in SVRG}} + \underbrace{\frac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\theta_t)}_{\text{compute exactly in SVRG}},$$

where $\widehat{S}_k$ is the $k^{\text{th}}$ column of $\widehat{S}$ and $\Delta(k)$ is the $k^{\text{th}}$ coordinate of $\Delta$.

Line 7 corresponds to SVRG's outer loop part that computes the full gradient. Line 12 corresponds to SVRG's inner loop update.

By Lemma 4, when $n \gg b^2 \log p$, the $\ell_2$ operator norm of $\|\widehat{S}\|_2 = O(1)$. And this implies $\|\widehat{S}^\top\widehat{S}\|_2 = O(1)$. Because $\|\widehat{S}_k\|_2^2$ is the $k^{\text{th}}$ diagonal element of $\widehat{S}^\top\widehat{S}$, we have $\|\widehat{S}_k\|_2^2 = O(1)$ for all $1 \le k \le p$. Thus, each $\left[p\widehat{S}_k\right]\Delta(k)$ is a $O(p)$-Lipschitz function.

By Theorem 6.5 of [9], when conditioned on $\theta_t$, and choosing

$$\tau = \Theta\left(\tfrac{1}{p}\right),$$
$$L_i \gtrsim p,$$

after $L_o^t$ SVRG outer steps, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{g}_t + \widehat{S}^{-1}\left(\frac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\theta_t)\right)\right\|_2^2 \middle| \theta_t, \{x_i\}_{i=1}^n\right] \lesssim & 0.9^{L_o^t}\left\|\frac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\theta_t)\right\|_2^2 \\ \lesssim & 0.9^{L_o^t}(1+\|\theta_t-\widehat{\theta}\|_2),\end{aligned}$$

where $\bar{g}_t = \frac{1}{L_o^t}\sum_{j=0}^{L_o^t} g_t^j$.

The gradient of the smooth component $\frac{1}{2}\Delta^\top\widehat{S}\Delta + \left\langle \frac{1}{n}\sum_{k=1}^n \nabla f_k(\theta_t), \Delta\right\rangle$ in (69) is

$$\widehat{S}\Delta + \frac{1}{n}\sum_{k=1}^{n}\nabla f_k(\theta_t) = \underbrace{\frac{1}{p}\sum_{k=1}^{p}\left[p\widehat{S}_k\right]\Delta(k)}_{\text{sample } \textit{by feature} \text{ in proximal SVRG}} + \underbrace{\frac{1}{n}\sum_{k=1}^{n}\nabla f_k(\theta_t)}_{\text{compute exactly in proximal SVRG}}.$$

Line 8 corresponds to proximal SVRG's outer loop part that computes the full gradient. Line 13 corresponds to proximal SVRG's inner loop update.

By Theorem 3.1 of [61], when conditioned on $\theta_t$, and choosing

$$\begin{aligned}\eta &= \Theta\left(\tfrac{1}{p}\right),\\ L_i &\gtrsim p,\end{aligned}$$

after $L_o^t$ proximal SVRG outer steps, we have

$$\begin{aligned}\mathbb{E}[P(\theta_{t+1} - P(\widehat{\theta})) \mid \theta_t] =& \mathbb{E}\left[P(\theta_t + \bar{d}_t - \widehat{\theta}) - P(\widehat{\theta}) \;\middle|\; \theta_t, \{x_i\}_{i=1}^n\right]\\ \lesssim& 0.9^{L_o^t}(P(\theta_t) - P(\widehat{\theta})),\end{aligned}$$

where $\bar{d}_t = \frac{1}{L_o^t}\sum_{j=0}^{L_o^t} d_t^j$. And this implies

$$\mathbb{E}[\|\theta_t - \widehat{\theta}\|_2^2] \lesssim 0.9^{\sum_{i=0}^{t-1} L_o^t}(P(\theta_0) - P(\widehat{\theta})).$$

At each $\theta_t$, we have

$$x_i(x_i^\top \theta_t - y_i) = x_i x_i^\top(\theta_t - \widehat{\theta}) + x_i(x_i^\top \widehat{\theta} - y_i).$$

For the first term, we have

$$\begin{aligned}\|x_i x_i^\top(\theta_t - \widehat{\theta})\|_\infty \leq& |x_i^\top(\theta_t - \widehat{\theta})|\|x_i\|_\infty\\ \leq& \|x_i\|_2\|\theta_t - \widehat{\theta}\|_2\|x_i\|_\infty\\ \leq& \sqrt{p}\|x_i\|_\infty^2\|\theta_t - \widehat{\theta}\|_2,\end{aligned}$$

which implies that

$$\begin{aligned}\max_{1\leq j,k\leq p}\left|\left[\left(x_i x_i^\top(\theta_t - \widehat{\theta})\right)\left(x_i x_i^\top(\theta_t - \widehat{\theta})\right)^\top\right]_{jk}\right| \leq& \|x_i x_i^\top(\theta_t - \widehat{\theta})\|_\infty^2\\ \leq& p\|x_i\|_\infty^4\|\theta_t - \widehat{\theta}\|_2^2.\end{aligned}$$

For the second term, we have

$$\begin{aligned}\|x_i(x_i^\top \widehat{\theta} - y_i)\|_\infty \leq& \|x_i x_i^\top(\widehat{\theta} - \theta^\star)\|_\infty + \|x_i\epsilon_i\|_\infty\\ \leq& \|x_i\|_\infty^2\|\widehat{\theta} - \theta^\star\|_1 + |\epsilon_i|\|x_i\|_\infty\end{aligned}$$

Because when $n \gg \log p$, from (72) we have with probability at least $1 - p^{-\Theta(1)}$

$$\max_{1\leq i\leq n}\|x_i\|_\infty \lesssim \sqrt{\log p + \log n},$$

and from (74) we have with probability at least $1 - n^{-\Theta(1)}$

$$\max_{1\leq i\leq n}|\epsilon_i| \lesssim \sigma\sqrt{\log n},$$

when conditioned on $\theta_t$ (and the data set $\{x_i\}_{i=1}^n$) we have

$$\begin{aligned}
&\max_{1\le j,k\le p}\left|\left[\left(\widehat{S}^{-1}g_t^0\right)\left(\widehat{S}^{-1}g_t^0\right)^\top-\left(\widehat{S}^{-1}\tfrac{1}{S_o}\textstyle\sum_{k\in I_o}\nabla f_k(\theta_t)\right)\left(\widehat{S}^{-1}\tfrac{1}{S_o}\textstyle\sum_{k\in I_o}\nabla f_k(\theta_t)\right)^\top\right]_{jk}\right|\\
\lesssim&\frac{1}{{D_\Sigma}^2}(\|x_ix_i^\top(\theta_t-\widehat{\theta})\|_\infty^2+2\|x_ix_i^\top(\theta_t-\widehat{\theta})\|_\infty\|x_i(x_i^\top\widehat{\theta}-y_i)\|_\infty)\\
\lesssim&\frac{1}{{D_\Sigma}^2}(p(\log p+\log n)^2\|\theta_t-\widehat{\theta}\|_2^2+\sqrt{p}(\log p+\log n)\|\theta_t-\widehat{\theta}\|_2((\log p+\log n)\|\widehat{\theta}-\theta^\star\|_1+\sigma\sqrt{(\log p+\log n)\log n}\\
\lesssim&\frac{1}{{D_\Sigma}^2}(p\|\theta_t-\widehat{\theta}\|_2^2+\sqrt{p}\|\theta-\widehat{\theta}\|_2(\sigma+\|\widehat{\theta}-\theta^\star\|_1))\mathrm{polylog}(p,n)
\end{aligned}$$

under the events of (72), (65) , and (74), where we used the fact (65) that the $\ell_\infty$ operator norm $\|\widehat{S}^{-1}\|_\infty\lesssim\frac{1}{D_\Sigma}$ with probability at least $1-p^{-\Theta(1)}$ when $n\gg\max\{b^2,\frac{1}{{D_\Sigma}^2}\}\log p$.

Thus, we can conclude that, conditioned on the data set $\{x_i\}_{i=1}^n$, and the events (72), (74), and (65), we have we have an asymptotic normality result

$$\tfrac{1}{\sqrt{t}}\left(\textstyle\sum_{t=1}^T\sqrt{S_o}\bar{g}_t+\frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\widehat{\theta}-y_i)\right)=W+R,$$

where $W$ weakly converges to $\mathcal{N}\Big(0,\widehat{S}^{-1}\Big[\frac{1}{n}\sum_{i=1}^n(x_i^\top\widehat{\theta}-y_i)^2x_ix_i^\top-\Big(\frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\widehat{\theta}-y_i)\Big)\Big(\frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\widehat{\theta}-y_i)\Big)^\top\Big]\widehat{S}^{-1}\Big)$, and

$$\begin{aligned}
\|R\|_\infty\le&\frac{1}{\sqrt{t}}\sum_{t=1}^T\left(\|\bar{g}_t-\widehat{S}^{-1}g_t^0\|_\infty+\|\widehat{S}^{-1}g_t^0-\tfrac{1}{S_o}\textstyle\sum_{k\in I_o}\nabla f_k(\widehat{\theta})\|_\infty\right)\\
\le&\frac{1}{\sqrt{t}}\sum_{t=1}^T\left(\|\bar{g}_t-\widehat{S}^{-1}g_t^0\|_2+\|\widehat{S}^{-1}g_t^0-\tfrac{1}{S_o}\textstyle\sum_{k\in I_o}\nabla f_k(\widehat{\theta})\|_\infty\right),
\end{aligned}$$

which implies

$$\begin{aligned}
&\mathbb{E}\left[\|R\|_\infty\mid\{x_i\}_{i=1}^n,(72),(74),(65)\right]\\
\lesssim&\mathbb{E}\left[\frac{1}{\sqrt{t}}\sum_{t=1}^T0.95^{L_o^t}(1+\|\theta_t-\widehat{\theta}\|_2)+\sqrt{p}(\log p+\log n)\|\theta_t-\widehat{\theta}\|_2\mid\{x_i\}_{i=1}^n,(72),(74),(65)\right]\\
\lesssim&\frac{1}{\sqrt{T}}\sum_{t=1}^T0.95^{L_o^t}(1+\sqrt{P(\theta_0)-P(\widehat{\theta})}0.95^{\sum_{i=0}^{t-1}L_o^t})+\sqrt{p}(\log p+\log n)\sqrt{P(\theta_0)-P(\widehat{\theta})}0.95^{\sum_{i=0}^{t-1}L_o^t}.
\end{aligned}$$

And, because $\left(\frac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\widehat{\theta})\right)_t$ are i.i.d., and bounded when conditioned on the data set $\{x_i\}_{i=1}^n$, and the events (72), (74), and (65), using a union bound over all matrix entries, and sub-Gaussian concentration inequalities [57] similar to Lemma 1's proof, when $T\gg\left((\log p+\log n)\|\widehat{\theta}-\theta^\star\|_1+\sigma\sqrt{(\log p+\log n)\log n}\right)\log p$, we also have

$$\left\|\tfrac{S_o}{T}\textstyle\sum_{t=1}^T\bar{g}_t\bar{g}_t^\top-\widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^n(x_i^\top\widehat{\theta}-y_i)^2x_ix_i^\top\right)\widehat{S}^{-1}\right\|_{\max}$$

$$\lesssim\sqrt{\left((\log p+\log n)\|\widehat{\theta}-\theta^\star\|_1+\sigma\sqrt{(\log p+\log n)\log n}\right)\frac{\log p}{T}}$$
$$+\frac{1}{u}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}0.95^{L_o^t}(1+\sqrt{P(\theta_0)-P(\widehat{\theta})}0.95^{\sum_{i=0}^{t-1}L_o^t})+\sqrt{p}(\log p+\log n)\sqrt{P(\theta_0)-P(\widehat{\theta})}0.95^{\sum_{i=0}^{t-1}L_o^t}\right],$$

with probability at least $1-p^{-\Theta(-1)}-u$, where we used Markov inequality for the remainder term.

## E.7 Proof of Lemma 1

We analyze the optimization problem conditioned on the data set $\{x_i\}_{i=1}^n$, which satisfies Lemma 4 with probability at least $1-p^{\Theta(-1)}$ when $n\gg b^2\log p$.

Because we have

$$\begin{aligned}&(x_i^\top\widehat{\theta}-y_i)^2\\=&(x_i^\top(\widehat{\theta}-\theta^\star)-\epsilon_i)^2\\=&\epsilon_i^2-2\epsilon_i x_i^\top(\widehat{\theta}-\theta^\star)+(x_i^\top(\widehat{\theta}-\theta^\star))^2,\end{aligned}$$

we can write

$$\begin{aligned}&\sigma^2\tfrac{1}{n}\textstyle\sum_{i=1}^n x_ix_i^\top-\tfrac{1}{n}\sum_{i=1}^n(x_i^\top\widehat{\theta}-y_i)^2x_ix_i^\top\\=&\tfrac{1}{n}\textstyle\sum_{i=1}^n(\sigma^2-\epsilon_i^2)x_ix_i^\top+\tfrac{1}{n}\sum_{i=1}^n(2\epsilon_ix_i^\top(\widehat{\theta}-\theta^\star)-(x_i^\top(\widehat{\theta}-\theta^\star))^2)x_ix_i^\top.\end{aligned}\quad(70)$$

Conditioned on $\{x_i\}_{i=1}^n$, because $\epsilon_i\sim\mathcal{N}(0,\sigma^2)$ are i.i.d., and $\epsilon_i^2$ is sub-exponential, using Bernstein inequality [57], we have

$$\begin{aligned}&\mathbb{P}\left[\left|\tfrac{1}{n}\textstyle\sum_{i=1}^n\left(1-\frac{\epsilon_i^2}{\sigma^2}\right)x_i(j)x_i(k)\right|>t\mid\{x_i\}_{i=1}^n\right]\\\lesssim&\exp\left(-n\min\left\{\frac{t}{\max_{1\le i\le n}|x_i(j)x_i(k)|},\left(\frac{t}{\max_{1\le i\le n}|x_i(j)x_i(k)|}\right)^2\right\}\right),\end{aligned}\quad(71)$$

for $1\le j,k\le p$, where $x_i(j)$ is the $j^{\text{th}}$ coordinate of $x_i$.

Because each $x_i(j)$ is $\mathcal{N}(0,\Theta(1))$ by our assumptions, using a union bound over all samples' coordinates we have

$$\max_{\substack{1\le i\le n\\1\le j\le p}}|x_i(j)|\lesssim\sqrt{\log p+\log n},\quad(72)$$

with probability at least $1-(pn)^{-\Theta(1)}$ .

Combining (71) and (72), and taking a union bound over all entries of the matrix $\frac{1}{n}\sum_{i=1}^n(\sigma^2-\epsilon_i^2)x_ix_i^\top$, when $n\gg\log p$, we have

$$\max_{1\le j,k\le p}|(\tfrac{1}{n}\textstyle\sum_{i=1}^n(\sigma^2-\epsilon_i^2)x_ix_i^\top)|_{jk}\lesssim\sigma^2(\log p+\log n)\sqrt{\frac{\log p}{n}},\quad(73)$$

with probability at least $(1-(pn)^{-\Theta(1)})(1-p^{-\Theta(1)})=1-(pn)^{-\Theta(1)}-p^{-\Theta(1)}$.

Because $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, by a union bound, we have

$$\max_{1\le i\le n} |\epsilon_i| \lesssim \sigma\sqrt{\log n}, \tag{74}$$

with probability at least $1 - n^{-\Theta(1)}$.

Using (72), we have

$$\begin{aligned}
&\max_{1\le i\le n} |x_i^\top(\widehat{\theta} - \theta^\star)| \\
\le&\|\widehat{\theta} - \theta^\star\|_1 \max_{1\le i\le n}\max_{1\le j\le p}|x_i(j)| \\
\lesssim& s\left(\sigma + \|\theta^\star\|_1\right)\sqrt{\tfrac{\log p}{n}(\log p + \log n)} \lesssim s\left(\sigma + \sqrt{s}\right)\sqrt{\tfrac{\log p}{n}(\log p + \log n)},
\end{aligned} \tag{75}$$

with probability at least $1 - p^{-\Theta(1)} - (pn)^{-\Theta(1)}$.

Combining (72), (73), (74), (75), and using (70), when $n \gg \log p$, we have

$$\begin{aligned}
&\max_{1\le j,k\le p}\left|\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n (x_i^\top\widehat{\theta} - y_i)^2 x_i x_i^\top - \sigma^2\tfrac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)_{jk}\right| \\
\lesssim&\sigma^2(\log p + \log n)\sqrt{\tfrac{\log p}{n}} + \sigma s\left(\sigma + \|\theta^\star\|_1\right)(\log p + \log n)^{\frac{3}{2}}\sqrt{\tfrac{\log p\cdot\log n}{n}} \\
&+ s^2\left(\sigma + \|\theta^\star\|_1\right)^2(\log p + \log n)^2\tfrac{\log p}{n},
\end{aligned} \tag{76}$$

with probability at least $1 - p^{-\Theta(1)} - n^{-\Theta(1)}$.

Combining (76) and (65), when $n \gg \max\{b^2, \frac{1}{D_{\Sigma^2}}\}\log p$, we have

$$\begin{aligned}
&\max_{1\le j,k\le p}\left|\left(\widehat{S}^{-1}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n (x_i^\top\widehat{\theta} - y_i)^2 x_i x_i^\top\right)\widehat{S}^{-1} - \sigma^2\widehat{S}^{-1}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n x_i x_i^\top\right)\widehat{S}^{-1}\right)_{jk}\right| \\
\lesssim&\tfrac{1}{D_{\Sigma^2}}\left(\sigma^2 + \sigma s\left(\sigma + \|\theta^\star\|_1\right)\sqrt{\log p + \log n}\sqrt{\log n} + s^2\left(\sigma + \|\theta^\star\|_1\right)^2(\log p + \log n)\sqrt{\tfrac{\log p}{n}}\right)(\log p + \log n)\sqrt{\tfrac{\log p}{n}}
\end{aligned}$$

with probability at least $1 - p^{-\Theta(1)} - n^{-\Theta(1)}$.

# F Technical lemmas

## F.1 Lemma 5

Next lemma is a well known property of convex functions (Lemma 3.11 of [9]).

**Lemma 5.** *For a $\alpha$ strongly convex and $\beta$ smooth function $F(x)$, we have*

$$\begin{aligned}
\langle\nabla F(x_1) - \nabla F(x_2), x_1 - x_2\rangle \ge&\frac{\alpha\beta}{\alpha+\beta}\|x_1 - x_2\|_2^2 + \frac{1}{\beta+\alpha}\|\nabla F(x_1) - \nabla F(x_2)\|_2^2 \\
\ge&\frac{1}{2}\alpha\|x_1 - x_2\|_2^2 + \frac{1}{2\beta}\|\nabla F(x_1) - \nabla F(x_2)\|_2^2.
\end{aligned}$$

## F.2 Lemma 6

Next lemma provides a bound on a geometric-like sequence.

**Lemma 6.** *Suppose we have a sequence*

$$a_{t+1} = (1 - \kappa t^{-d})a_t + Ct^{-pd},$$

*where* $a_1 \geq 0$, $0 < \kappa < 1$, $p \geq 2$ *and* $d \in (\frac{1}{2}, 1)$ *is the decaying rate.*

*Then,* $\forall 1 \leq s \leq t$ *we have*

$$a_t \leq C\frac{1}{pd-1}(1 - t^{1-pd}) \exp\left(-\kappa \frac{1}{1-d}\left((t+1)^{1-d} - (s+1)^{1-d}\right)\right) + a_1 s^{-(p-1)d}\frac{1}{\kappa}.$$

*When we assume that* $a_1, C, \kappa, p, d$ *are all constants, we have*

$$a_t = O(t^{-(p-1)d}).$$

*Proof.* Unrolling the recursion, we have

$$a_t = C\underbrace{\sum_{i=1}^{t-1}\Big(\prod_{j=i+1}^{t-1}(1 - \kappa j^{-d})\Big) i^{-pd}}_{[1]} + a_1 \underbrace{\prod_{i=1}^{t-1}(1 - \kappa i^{-d})}_{[2]}.$$

Splitting term [1] into two parts, we have

$$\begin{aligned}
&\sum_{i=1}^{t-1}\left(\prod_{j=i+1}^{t-1}(1 - \kappa j^{-d})\right) i^{-pd} \\
=&\sum_{i=1}^{s-1}\left(\prod_{j=i+1}^{t-1}(1 - \kappa j^{-d})\right) i^{-pd} + \sum_{i=s}^{t-1}\left(\prod_{j=i+1}^{t-1}(1 - \kappa j^{-d})\right) i^{-pd}.
\end{aligned}$$

For the first part, we have

$$\begin{aligned}
&\sum_{i=1}^{s-1}\Big(\prod_{j=i+1}^{t-1}(1 - \kappa j^{-d})\Big) i^{-pd} \\
\leq&\left(\prod_{j=s}^{t-1}(1 - \kappa j^{-d})\right)\sum_{i=1}^{s-1} i^{-pd} \\
\leq&\frac{1}{pd-1}(1 - t^{1-pd}) \exp\left(-\kappa \frac{1}{1-d}((t+1)^{1-d} - (s+1)^{1-d})\right)
\end{aligned}$$

where we used

$$\sum_{i=r}^{s} i^{-pd}$$

$$
\begin{aligned}
&\leq \int_r^{s+1} u^{-pd}\, du \\
&\leq \frac{1}{pd-1}(r^{1-pd} - (s+1)^{1-pd}).
\end{aligned}
$$

For term [2], notice that for $1 \leq r \leq s$, using $1-x \leq \exp(-x)$ when $x \in [0,1]$, we have

$$
\prod_{i=r}^{s}(1-\kappa i^{-d}) \leq \exp(-\kappa \textstyle\sum_{i=r}^{s} i^{-d}),
$$

and using the fact that

$$
\begin{aligned}
\sum_{i=r}^{s} i^{-d} &\geq \int_r^{s+1} (u+1)^{-d}\, du \\
&= \frac{1}{1-d}\left((s+2)^{1-d} - (r+1)^{1-d}\right),
\end{aligned}
$$

we have

$$
\prod_{i=1}^{t-1}(1-\kappa i^{-d}) \leq \exp\left(-\kappa \tfrac{1}{1-d}(t^{1-d} - 2^{1-d})\right).
$$

For the second part, we have

$$
\begin{aligned}
&\sum_{i=s}^{t-1}\left(\prod_{j=i+1}^{t-1}(1-\kappa j^{-d})\right) i^{-pd} \\
\leq & s^{-(p-1)d} \sum_{i=s}^{t-1}\left(\prod_{j=i+1}^{t-1}(1-\kappa j^{-d})\right) i^{-d} \\
= & s^{-(p-1)d} \frac{1}{\kappa} \sum_{i=s}^{t-1}\left(\prod_{j=i+1}^{t-1}(1-\kappa j^{-d})\right) \kappa i^{-d} \\
= & s^{-(p-1)d} \frac{1}{\kappa}\left(1 - \prod_{i=s}^{t-1}(1-\kappa i^{-d})\right) \\
\leq & s^{-(p-1)d} \frac{1}{\kappa},
\end{aligned}
$$

where we used the fact that

$$
\begin{aligned}
&\sum_{i=s}^{t-1} \kappa i^{-d} \prod_{j=i+1}^{t-1}(1-\kappa j^{-d}) \\
= & 1 - \prod_{i=s}^{t-1}(1-\kappa j^{-d})
\end{aligned}
$$

$<1$.

When we assume that $a_1, C, \kappa, p, d$ are all constants, setting $s = \lfloor \frac{n}{2} \rfloor$, we have

$$a_t = O(t^{-(p-1)d}).$$

□

# G Experiments

## G.1 Synthetic data

### G.1.1 Low dimensional problems

Here, we provide the exact configurations for linear/logistic regression examples provided in Table 1.

**Linear regression.** We consider the model $y = \langle [1, \cdots, 1]^\top / \sqrt{10}, x \rangle + \epsilon$, where $x \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^{10}$ and $\epsilon \sim \mathcal{N}(0, 0.7^2)$, with 100 i.i.d. data points.

***Lin1:*** We used $\Sigma = I$. For Algorithm 1, we set $T = 100$, $d_o = d_i = 2/3$, $\rho_0 = 0.1$, $L = 200$, $\tau_0 = 20$, $S_o = S_i = 10$. In bootstrap we used 100 replicates. For averaged SGD, we used 100 averages each of length 50, with step size $0.7 \cdot (t+1)^{-2/3}$ and batch size 10.

***Lin2:*** We used $\Sigma_{jk} = 0.4^{|j-k|}$. For Algorithm 1, we set $T = 100$, $d_o = d_i = 2/3$, $\rho_0 = 0.7$, $L = 100$, $\tau_0 = 1$, $S_o = S_i = 10$. In bootstrap we used 100 replicates. For averaged SGD, we used 100 averages each of length 50, with step size $(t+1)^{-2/3}$ and batch size 10.

**Logistic regression.** Although logistic regression does not satisfy strong convexity, experimentally Algorithm 1 still gives valid confidence intervals ([26] recently has shown that SGD in logistic regression behaves similar to strongly convex problems). We consider the model $\mathbb{P}[y = 1] = \mathbb{P}[y = 0] = 1/2$ and $x \mid y \sim \mathcal{N}(0.1/\sqrt{10} \cdot [1, \cdots, 1]^\top, \Sigma) \in \mathbb{R}^{10}$, with 100 i.i.d. data points. Because in bootstrap resampling the Hessian is singular for some replicates, we use jackknife and solve each replicate using Newton's method, which approximately needs 25 steps per replicate.

***Log1:*** We used $\Sigma = I$. For Algorithm 1, we set $T = 50$, $d_o = d_i = 2/3$, $\rho_0 = 0.1$, $L = 100$, $\tau_0 = 2$, $S_o = S_i = 10$, $\delta_0 = 0.01$. For averaged SGD, we used 50 averages each of length 100, with step size $2 \cdot (t+1)^{-2/3}$ and batch size 10.

***Log2:*** We used $\Sigma_{jk} = 0.4^{|j-k|}$. For Algorithm 1, we set $T = 50$, $d_o = d_i = 2/3$, $\rho_0 = 0.1$, $L = 100$, $\tau_0 = 5$, $S_o = S_i = 10$, $\delta_0 = 0.01$ For averaged SGD, we used 50 averages each of length 100, with step size $5 \cdot (t+1)^{-2/3}$ and batch size 10.

### G.1.2 High dimensional linear regression

For comparison with de-biased LASSO [32, 53], we use the *oracle* de-biased LASSO estimator

$$\widehat{\theta}^{\mathrm{d}}_{\mathrm{oracle}} = \widehat{\theta}_{\mathrm{LASSO}} + \tfrac{1}{n} \cdot \Sigma^{-1} \left( \tfrac{1}{n} \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i x_i^\top \widehat{\theta}_{\mathrm{LASSO}} \right),$$

and its corresponding statistical error covariance estimate

$$\sigma^2 \cdot \Sigma^{-1} \left( \tfrac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \right) \Sigma^{-1},$$

which assumes that the true inverse covariance $\Sigma^{-1}$ and observation noise variance $\sigma^2$ are known.

**Experiment 1.** We use 600 i.i.d. samples from a model with $\Sigma = I$, $\sigma = 0.7$, $\theta^\star = [1/\sqrt{8}, \cdots, 1/\sqrt{8}, 0, \cdots, 0]^\top \in \mathbb{R}^{1000}$ which is 8-sparse.

For our method, the average confidence interval length is 0.14 and average coverage is 0.83. For the oracle de-biased LASSO estimator, the average confidence interval length is 0.11 and average coverage is 0.98.

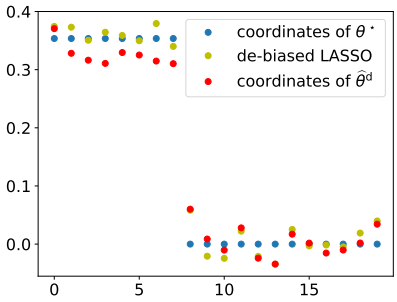



Figure 3: Comparison of our de-biased estimator and oracle de-biased LASSO estimator

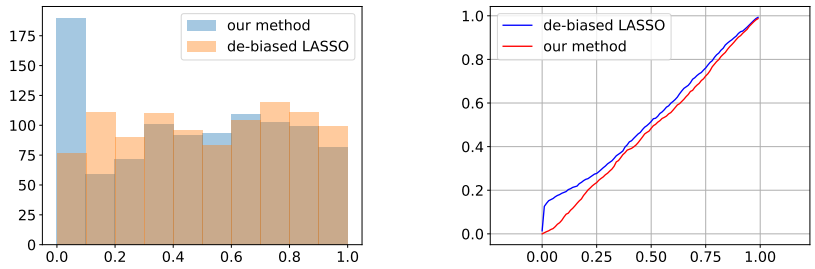
**Experiment 2.** We use 600 i.i.d. samples from a model with $\Sigma = I$, $\sigma = 0.7$, $\theta^\star = \mathbf{0} \in \mathbb{R}^{1000}$ which is 8-sparse.

Figure 4 shows p-value distribution for our method and the oracle de-biased LASSO estimator.

## G.2 Real data

### G.2.1 Neural network adversarial attack detection

The adversarial perturbation used in our experiments is shown in Figure 7. It is generated using the fast gradient sign method [27] Figure 5 shows images in a




(a) PDF (b) CDF

Figure 4: Distribution of two-sided Z-test p-values

"Shirt" example. Figure 6 shows images in a "T-shirt/top" example.

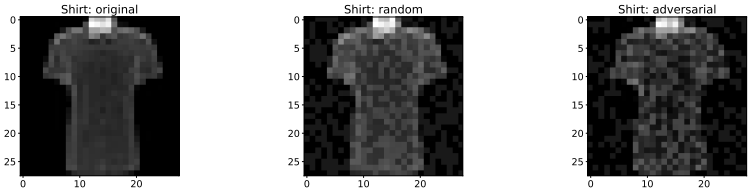



Figure 5: "Shirt" example

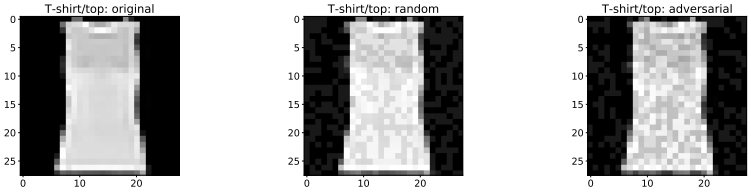



Figure 6: "T-shirt/top" example

### G.2.2 High dimensional linear regression

For the vanilla LASSO estimate on the high-throughput genomic data set concerning riboflavin (vitamin B2) production rate [11], we set $\lambda = 0.021864$. Figure 8, and we see that our point estimate is similar to the vanilla LASSO point estimate.

For statistical inference, in our method, we compute p-values using two-sided Z-test. Adjusting FWER to 5% signifi-cance level, our method does not find

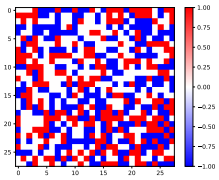


Figure 7: Adversarial perturbation generated using the fast gradient sign method [27]

any significant gene. [31, 11] report that [10] also does not find any significant gene, whereas [39] finds one significant gene (YXLD-at), and [31] finds two significant genes (YXLD-at and YXLE-at). This indicates that our method is more conservative than [31, 39].

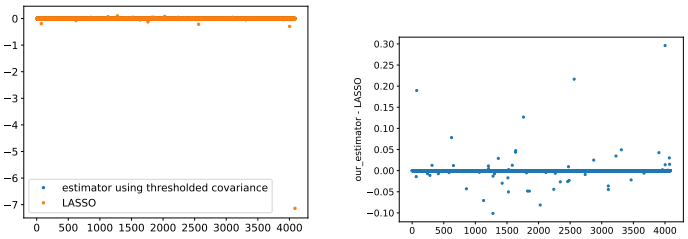



Figure 8: Comparison of our high dimensional linear regression point estimate with the vanilla LASSO estimate

# Volume 2: Statistical Inference Using SGD

# Statistical inference using SGD

Tianyang Li Liu Liu Anastasios Kyrillidis Constantine Caramanis

October 22, 2018

**Abstract**

We present a novel method for frequentist statistical inference in $M$-estimation problems, based on stochastic gradient descent (SGD) *with a fixed step size*: we demonstrate that the average of such SGD sequences can be used for statistical inference, after proper scaling. An intuitive analysis using the Ornstein-Uhlenbeck process suggests that such averages are asymptotically normal. From a practical perspective, our SGD-based inference procedure is a first order method, and is well-suited for large scale problems. To show its merits, we apply it to both synthetic and real datasets, and demonstrate that its accuracy is comparable to classical statistical methods, while requiring potentially far less computation.

## 1 Introduction

In $M$-estimation, the minimization of empirical risk functions (RFs) provides point estimates of the model parameters. Statistical inference then seeks to assess the quality of these estimates, for example, obtaining confidence intervals or solving hypothesis testing problems. Within this context, a classical result in statistics states that the asymptotic distribution of the empirical RF's minimizer is normal, centered around the population RF's minimizer [24]. Thus, given the mean and covariance of this normal distribution, we can infer a range of values, along with probabilities, that allows us to quantify the probability that this interval includes the true minimizer.

The Bootstrap [8, 9] is a classical tool for obtaining estimates of the mean and covariance of this distribution. The Bootstrap operates by generating samples from this distribution (usually, by re-sampling with or without replacement from the entire data set) and repeating the estimation procedure over these different re-samplings. As the data dimensionality and size grow, the Bootstrap becomes increasingly –even prohibitively– expensive.

In this context, we follow a different path: we show that inference can also be accomplished by directly using stochastic gradient descent (SGD) *with a fixed step size over the data set*. It is well-established that fixed step-size SGD is by and large the dominant method used for large scale data analysis. We prove, and also demonstrate empirically, that *the average of SGD sequences, obtained by minimizing RFs, can also be used for statistical inference.* Unlike the Bootstrap, our approach does not require creating many large-size subsamples from the data, neither re-running SGD from scratch for each of these subsamples. Our method only uses first order information from gradient computations, and does not require any second order information. Both of these are important for large scale problems, where re-sampling many times, or computing Hessians, may be computationally prohibitive.

**Outline and main contributions:** This paper studies and analyzes a simple, *fixed step size*[1], SGD-based algorithm for inference in $M$-estimation problems. Our algorithm produces samples, whose covariance converges to the covariance of the $M$-estimate, without relying on bootstrap-based schemes, and also avoiding direct and costly computation of second order information. Much work has been done on the asymptotic normality of SGD, as well as on the Stochastic Gradient Langevin Dynamics (and variants) in the Bayesian setting. As we discuss in detail in Section 4, this is the first work to provide finite sample inference results, using fixed step size, and without imposing overly restrictive assumptions on the convergence of fixed step size SGD.

[1] *Fixed step size* means we use the same step size every iteration, but the step size is smaller with more total number of iterations. *Constant step size* means the step size is constant no matter how many iterations taken.

The remainder of the paper is organized as follows. In the next section, we define the inference problem for $M$-estimation, and recall basic results of asymptotic normality and how these are used. Section 3 is the main body of the paper: we provide the algorithm for creating bootstrap-like samples, and also provide the main theorem of this work. As the details are involved, we provide an intuitive analysis of our algorithm and explanation of our main results, using an asymptotic Ornstein-Uhlenbeck process approximation for the SGD process [12, 18, 4, 13, 15], and we postpone the full proof until the appendix. We specialize our main theorem to the case of linear regression (see supplementary material), and also that of logistic regression. For logistic regression in particular, we require a somewhat different approach, as the logistic regression objective is not strongly convex. In Section 4, we present related work and elaborate how this work differs from existing research in the literature. Finally, in the experimental section, we provide parts of our numerical experiments that illustrate the behavior of our algorithm, and corroborate our theoretical findings. We do this using synthetic data for linear and logistic regression, and also by considering the Higgs detection [3] and the LIBSVM Splice data sets. A considerably expanded set of empirical results is deferred to the appendix.

Supporting our theoretical results, our empirical findings suggest that the SGD inference procedure produces results similar to bootstrap while using far fewer operations, thereby producing a more efficient inference procedure applicable in large scale settings where other approaches fail.

## 2 Statistical inference for $M$-estimators

Consider the problem of estimating a set of parameters $\theta^\star \in \mathbb{R}^p$ using $n$ samples $\{X_i\}_{i=1}^n$, drawn from some distribution $P$ on the sample space $\mathcal{X}$. In frequentist inference, we are interested in estimating the minimizer $\theta^\star$ of the population risk:

$$\theta^\star = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \mathbb{E}_P[f(\theta; X)] = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \int_x f(\theta; x)\, \mathrm{d}P(x), \tag{1}$$

where we assume that $f(\cdot; x) : \mathbb{R}^p \to \mathbb{R}$ is real-valued and convex; further, we use $\mathbb{E} \equiv \mathbb{E}_P$, unless otherwise stated. In practice, the distribution $P$ is unknown. We thus estimate $\theta^\star$ by solving an empirical risk minimization (ERM) problem, where we use the estimate $\widehat{\theta}$:

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n f(\theta; X_i). \tag{2}$$

Statistical inference consists of techniques for obtaining information beyond point estimates $\widehat{\theta}$, such as confidence intervals. These can be performed if there is an asymptotic limiting distribution associated with $\widehat{\theta}$ [25]. Indeed, under standard and well-understood regularity conditions, the solution to $M$-estimation problems satisfies asymptotic normality. That is, the distribution $\sqrt{n}(\widehat{\theta} - \theta^\star)$ converges weakly to a normal:

$$\sqrt{n}(\widehat{\theta} - \theta^\star) \longrightarrow \mathcal{N}(0, H^{\star -1} G^\star H^{\star -1}), \tag{3}$$

where $H^\star = \mathbb{E}[\nabla^2 f(\theta^\star; X)]$, and $G^\star = \mathbb{E}[\nabla f(\theta^\star; X) \nabla f(\theta^\star; X)^\top]$ (Theorem 5.21, [24]). We can therefore use this result, as long as we have a good estimate of the covariance matrix: $H^{\star -1} G^\star H^{\star -1}$. The central goal of this paper is obtaining accurate estimates for $H^{\star -1} G^\star H^{\star -1}$.

A naive way to estimate $H^{\star -1} G^\star H^{\star -1}$ is through the empirical estimator $\widehat{H}^{-1} \widehat{G} \widehat{H}^{-1}$ where:

$$\widehat{H} = \frac{1}{n} \sum_{i=1}^n \nabla^2 f(\widehat{\theta}; X_i) \qquad \text{and}$$

$$\widehat{G} = \frac{1}{n} \sum_{i=1}^n \nabla f(\widehat{\theta}; X_i) \nabla f(\widehat{\theta}; X_i)^\top. \tag{4}$$

Beyond calculating $\widehat{H}$ and $\widehat{G}$,[2] this computation requires an inversion of $\widehat{H}$ and matrix-matrix multiplications in order to compute $\widehat{H}^{-1} \widehat{G} \widehat{H}^{-1}$ – a key computational bottleneck in high dimensions. Instead, our method uses SGD to directly estimate $\widehat{H}^{-1} \hat{G} \widehat{H}^{-1}$.

[2] In the case of maximum likelihood estimation, we have $H^\star = G^\star$ which is called Fisher information, thus the covariance of interest is $H^{\star -1} = G^{\star -1}$. This can be estimated either using $\widehat{H}$ or $\widehat{G}$.

# 3 Statistical inference using SGD

In this section, we provide our main results, including the algorithm and its theoretical guarantees. We also describe its specialization to linear regression and logistic regression.

Consider the optimization problem in (2). For instance, in maximum likelihood estimation (MLE), $f_i(\theta; X_i)$ is a negative log-likelihood function. For simplicity of notation, we use $f_i(\theta)$ and $f(\theta)$ in the rest of the paper.

The SGD algorithm with a fixed step size $\eta$, is given by the iteration

$$\theta_{t+1} = \theta_t - \eta g_s(\theta_t), \tag{5}$$

where $g_s(\cdot)$ is an unbiased estimator of the gradient, *i.e.*, $\mathbb{E}[g_s(\theta) \mid \theta] = \nabla f(\theta)$, where the expectation is w.r.t. the stochasticity in the $g_s(\cdot)$ calculation. A classical example of an unbiased estimator of the gradient is $g_s(\cdot) \equiv \nabla f_j(\cdot)$, where $j$ is a uniformly random index over the samples $X_j$.

*Our inference procedure uses the average of $t$ SGD iterations.* Denote such sequences as $\bar{\theta}_t$:

$$\bar{\theta}_t = \frac{1}{t}\sum_{i=1}^{t} \theta_i. \tag{6}$$

The algorithm proceeds as follows: Given a sequence of SGD iterates, we use the first SGD iterates $\theta_{-b}, \theta_{-b+1}, \ldots, \theta_0$ as a burn in period; we discard these iterates. Next, for each "segment" of $t + d$ iterates, we use the first $t$ iterates to compute $\bar{\theta}_t^{(i)} = \frac{1}{t}\sum_{j=1}^{t} \theta_j^{(i)}$ and discard the last $d$ iterates, where $i$ indicates the $i$-th segment. This procedure is illustrated in Figure 1.

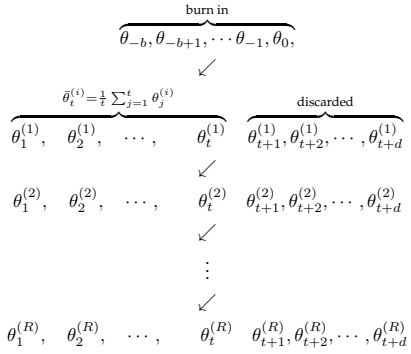



Figure 1: Our SGD inference procedure

Similar to ensemble learning [17], we use $i = 1, 2, \ldots, R$ estimators for statistical inference.

$$\theta^{(i)} = \widehat{\theta} + \frac{\sqrt{K_s}\sqrt{t}}{\sqrt{n}}(\bar{\theta}_t^{(i)} - \widehat{\theta}). \tag{7}$$

Here, $K_s$ is a scaling factor that depends on how the stochastic gradient $g_s$ is computed. We show examples of $K_s$ for mini batch SGD in linear regression and logistic regression in the corresponding sections. In practice, we can use $\widehat{\theta} \approx \frac{1}{R}\sum_{i=1}^{R} \bar{\theta}_t^{(i)}$ [5].

*Step size $\eta$ selection and length $t$:* Theorem 1 below is consistent only for SGD with fixed step size that depends on the number of samples taken. Our experiments, however, demonstrate that choosing a constant (large) $\eta$ gives equally accurate results with significantly reduced running time. A better understanding of $t$'s and $\eta$'s influence requires (conjectured) stronger bounds for SGD with constant step size. Heuristically, calibration methods for parameter tuning in subsampling methods ([19], Ch. 9) could be used for hyper-parameter tuning in our SGD procedure. We leave the problem of finding maximal (provable) learning rates for future work.

*Discarded length $d$:* Based on the analysis of mean estimation, if we discard $d$ SGD iterates in every segment, the correlation between consecutive $\theta^{(i)}$ and $\theta^{(i+1)}$ is on the order of $C_1 e^{-C_2 \eta d}$, where $C_1$ and $C_2$ are data dependent constants. This can be used as a rule of thumb to reduce correlation between samples from our SGD inference procedure.

*Burn-in period $b$:* The purpose of the burn-in period $b$, is to ensure that samples are generated when SGD iterates are sufficiently close to the optimum. This can be determined using heuristics for SGD convergence diagnostics. Another approach is to use other methods (*e.g.*, SVRG [11]) to find the optimum, and use a relatively small $b$ for SGD to reach stationarity, similar to Markov Chain Monte Carlo burn in.

## 3.1 Theoretical guarantees

Next, we provide the main theorem of our paper. Essentially, this provides conditions under which our algorithm is guaranteed to succeed, and hence has inference capabilities.

**Theorem 1.** *For a differentiable convex function $f(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$, with gradient $\nabla f(\theta)$, let $\widehat{\theta} \in \mathbb{R}^p$ be its minimizer, according to* (2)*, and denote its Hessian at $\widehat{\theta}$ by $H := \nabla^2 f(\widehat{\theta})$ . Assume that $\forall \theta \in \mathbb{R}^p$, $f$ satisfies:*

*($F_1$) Weak strong convexity: $(\theta - \widehat{\theta})^\top \nabla f(\theta) \geq \alpha \|\theta - \widehat{\theta}\|_2^2$, for constant $\alpha > 0$,*

*($F_2$) Lipschitz gradient continuity: $\|\nabla f(\theta)\|_2 \leq L\|\theta - \widehat{\theta}\|_2$, for constant $L > 0$,*

*($F_3$) Bounded Taylor remainder: $\|\nabla f(\theta) - H(\theta - \widehat{\theta})\|_2 \leq E\|\theta - \widehat{\theta}\|_2^2$, for constant $E > 0$,*

*($F_4$) Bounded Hessian spectrum at $\widehat{\theta}$: $0 < \lambda_L \leq \lambda_i(H) \leq \lambda_U < \infty, \forall i$.*

*Furthermore, let $g_s(\theta)$ be a stochastic gradient of $f$, satisfying:*

*($G_1$)* $\mathbb{E}\left[g_s(\theta) \mid \theta\right] = \nabla f(\theta)$,

*($G_2$)* $\mathbb{E}\left[\|g_s(\theta)\|_2^2 \mid \theta\right] \leq A\|\theta - \widehat{\theta}\|_2^2 + B$,

*($G_3$)* $\mathbb{E}\left[\|g_s(\theta)\|_2^4 \mid \theta\right] \leq C\|\theta - \widehat{\theta}\|_2^4 + D$,

*($G_4$)* $\left\|\mathbb{E}\left[g_s(\theta)g_s(\theta)^\top \mid \theta\right] - G\right\|_2 \leq A_1\|\theta - \widehat{\theta}\|_2 + A_2\|\theta - \widehat{\theta}\|_2^2 + A_3\|\theta - \widehat{\theta}\|_2^3 + A_4\|\theta - \widehat{\theta}\|_2^4$,

*for positive, data dependent constants $A, B, C, D, A_i$, for $i = 1, \ldots, 4$. Assume that $\|\theta_1 - \widehat{\theta}\|_2^2 = O(\eta)$; then for sufficiently small step size $\eta > 0$, the average SGD sequence in* (6) *satisfies:*

$$\left\| t\mathbb{E}[(\bar{\theta}_t - \widehat{\theta})(\bar{\theta}_t - \widehat{\theta})^\top] - H^{-1}GH^{-1} \right\|_2 \lesssim \sqrt{\eta} + \sqrt{\tfrac{1}{t\eta} + t\eta^2},$$

*where $G = \mathbb{E}[g_s(\widehat{\theta})g_s(\widehat{\theta})^\top \mid \widehat{\theta}]$.*

We provide the full proof in the appendix, and also we give precise (data-dependent) formulas for the above constants. For ease of exposition, we leave them as constants in the expressions above.

*Discussion*. For linear regression, assumptions ($F_1$), ($F_2$), ($F_3$), and ($F_4$) are satisfied when the empirical risk function is not degenerate. In mini batch SGD using sampling with replacement, assumptions ($G_1$), ($G_2$), ($G_3$), and ($G_4$) are satisfied. Linear regression's result is presented in Corollary 1.

For logistic regression, assumption ($F_1$) is not satisfied because the empirical risk function in this case is strictly but not strongly convex. Thus, we cannot apply Theorem 1 directly. Instead, we consider the use of SGD on the *square of the empirical risk function plus a constant; see eq.* (13) *below*. When the empirical risk function is not degenerate, (13) satisfies assumptions ($F_1$), ($F_2$), ($F_3$), and ($F_4$). We cannot directly use vanilla SGD to minimize (13), instead we describe a modified SGD procedure for minimizing (13) in Section 3.5, which satisfies assumptions ($G_1$), ($G_2$), ($G_3$), and ($G_4$). We believe that this result is of interest by its own. We present the result specialized for logistic regression in Corollary 2.

Note that Theorem 1 proves consistency for SGD with fixed step size, requiring $\eta \to 0$ when $t \to \infty$. However, we empirically observe in our experiments that a sufficiently large *constant* $\eta$ gives better results. We conjecture that the average of consecutive iterates in SGD with *larger constant step size* converges to the optimum and we consider it for future work.

## 3.2 Intuitive interpretation via the Ornstein-Uhlenbeck process approximation

Here, we describe a continuous approximation of the discrete SGD process and relate it to the Ornstein-Uhlenbeck process [21], to give an intuitive explanation of our results—the complete proofs appear in the appendix. In particular, under regularity conditions, the stochastic process $\Delta_t = \theta_t - \widehat{\theta}$ asymptotically converges to an Ornstein-Uhlenbeck process $\Delta(t)$, [12, 18, 4, 13, 15] that satisfies:

$$\mathrm{d}\Delta(T) = -H\Delta(T)\,\mathrm{d}T + \sqrt{\eta}G^{\frac{1}{2}}\,\mathrm{d}B(T), \tag{8}$$

where $B(T)$ is a standard Brownian motion. Given (8), $\sqrt{t}(\bar{\theta}_t - \widehat{\theta})$ can be approximated as

$$\begin{aligned} \sqrt{t}(\bar{\theta}_t - \widehat{\theta}) &= \tfrac{1}{\sqrt{t}}\sum_{i=1}^{t}(\theta_i - \widehat{\theta}) \\ &= \tfrac{1}{\eta\sqrt{t}}\sum_{i=1}^{t}(\theta_i - \widehat{\theta})\eta \approx \tfrac{1}{\eta\sqrt{t}}\int_0^{t\eta}\Delta(T)\,\mathrm{d}T, \end{aligned} \tag{9}$$

where we use the approximation that $\eta \approx \mathrm{d}T$. By rearranging terms in (8) and multiplying both sides by $H^{-1}$, we can rewrite the stochastic differential equation (8) as $\Delta(T)\,\mathrm{d}T = -H^{-1}\,\mathrm{d}\Delta(T) + \sqrt{\eta}H^{-1}G^{\frac{1}{2}}\,\mathrm{d}B(T)$. Thus, we have

$$\int_0^{t\eta} \Delta(T)\,\mathrm{d}T = \\ -H^{-1}(\Delta(t\eta) - \Delta(0)) + \sqrt{\eta}H^{-1}G^{\frac{1}{2}}B(t\eta). \tag{10}$$

After plugging (10) into (9) we have

$$\sqrt{t}\left(\bar{\theta}_t - \widehat{\theta}\right) \approx \\ -\tfrac{1}{\eta\sqrt{t}}H^{-1}\left(\Delta(t\eta) - \Delta(0)\right) + \tfrac{1}{\sqrt{t\eta}}H^{-1}G^{\frac{1}{2}}B(t\eta).$$

When $\Delta(0) = 0$, the variance $\mathrm{Var}\left[-1/\eta\sqrt{t}\cdot H^{-1}\left(\Delta(t\eta) - \Delta(0)\right)\right] = O\left(1/t\eta\right)$. Since $1/\sqrt{t\eta}\cdot H^{-1}G^{\frac{1}{2}}B(t\eta) \sim \mathcal{N}(0, H^{-1}GH^{-1})$, when $\eta \to 0$ and $\eta t \to \infty$, we conclude that

$$\sqrt{t}(\bar{\theta}_t - \widehat{\theta}) \sim \mathcal{N}(0, H^{-1}GH^{-1}).$$

### 3.3 Exact analysis of mean estimation

In this section, we give an exact analysis of our method in the least squares, mean estimation problem. For $n$ i.i.d. samples $X_1, X_2, \ldots, X_n$, the mean is estimated by solving the following optimization problem

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \tfrac{1}{2}\|X_i - \theta\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

In the case of mini-batch SGD, we sample $S = O(1)$ indexes uniformly randomly with replacement from $[n]$; denote that index set as $I_t$. For convenience, we write $Y_t = \frac{1}{S}\sum_{i\in I_t} X_i$, Then, in the $t^{\text{th}}$ mini batch SGD step, the update step is

$$\theta_{t+1} = \theta_t - \eta(\theta_t - Y_t) = (1-\eta)\theta_t + \eta Y_t, \tag{11}$$

which is the same as the exponential moving average. And we have

$$\sqrt{t}\widehat{\theta}_t = -\frac{1}{\eta\sqrt{t}}(\theta_{t+1} - \theta_1) + \frac{1}{\sqrt{t}}\sum_{i=1}^{n} Y_i. \tag{12}$$

Assume that $\|\theta_1 - \widehat{\theta}\|_2^2 = O(\eta)$, then from Chebyshev's inequality $-\frac{1}{\eta\sqrt{t}}(\theta_{t+1} - \theta_1) \to 0$ almost surely when $t\eta \to \infty$. By the central limit theorem, $\frac{1}{\sqrt{t}}\sum_{i=1}^{n} Y_i$ converges weakly to $\mathcal{N}(\widehat{\theta}, \frac{1}{S}\hat{\Sigma})$ with $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\theta})(X_i - \widehat{\theta})^\top$. From (11), we have $\|\operatorname{Cov}(\theta_a, \theta_b)\|_2 = O(\eta(1-\eta)^{|a-b|})$ uniformly for all $a, b$, where the constant is data dependent. Thus, for our SGD inference procedure, we have $\|\operatorname{Cov}(\theta^{(i)}, \theta^{(j)})\|_2 = O(\eta(1-\eta)^{d+t|i-j|})$. Our SGD inference procedure does not generate samples that are independent conditioned on the data, whereas replicates are independent conditioned on the data in bootstrap, but this suggests that our SGD inference procedure can produce "almost independent" samples if we discard sufficient number of SGD iterates in each segment.

When estimating a mean using our SGD inference procedure where each mini batch is $S$ elements sampled with replacement, we set $K_s = S$ in (7).

### 3.4 Linear Regression

In linear regression, the empirical risk function is given by:

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} \tfrac{1}{2}(\theta^\top x_i - y_i)^2,$$

where $y_i$ denotes the observations of the linear model and $x_i$ are the regressors. To find an estimate to $\theta^\star$, one can use SGD with stochastic gradient given by:

$$g_s[\theta_t] = \frac{1}{S}\sum_{i\in I_t} \nabla f_i(\theta_t),$$

where $I_t$ are $S$ indices uniformly sampled from $[n]$ with replacement.

Next, we state a special case of Theorem 1. Because the Taylor remainder $\nabla f(\theta) - H(\theta - \widehat{\theta}) = 0$, linear regression has a stronger result than general $M$-estimation problems.

**Corollary 1.** *Assume that $\|\theta_1 - \widehat{\theta}\|_2^2 = O(\eta)$. We have*

$$\left\| t\mathbb{E}[(\bar{\theta}_t - \widehat{\theta})(\bar{\theta}_t - \widehat{\theta})^\top] - H^{-1}GH^{-1} \right\|_2 \lesssim \sqrt{\eta} + \frac{1}{\sqrt{t\eta}},$$

*where $H = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ and $G = \frac{1}{S}\frac{1}{n}\sum_{i=1}^n (x_i^\top\widehat{\theta} - y_i)^2 x_i x_i^\top$.*

*We assume that $S = O(1)$ is bounded, and quantities other than $t$ and $\eta$ are data dependent constants.*

As with our main theorem, in the appendix we provide explicit data-dependent expressions for the constants in the result.

Because in linear regression the estimate's covariance is $\frac{1}{n}(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top)^{-1})(\frac{1}{n}(x_i^\top\widehat{\theta} - y_i)(x_i^\top\widehat{\theta} - y_i)^\top)(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top)^{-1})$, we set the scaling factor $K_s = S$ in (7) for statistical inference.

### 3.5 Logistic regression

We next apply our method to logistic regression. We have $n$ samples $(X_1, y_1), (X_2, y_2), \ldots (X_n, y_n)$ where $X_i \in \mathbb{R}^p$ consists of features and $y_i \in \{+1, -1\}$ is the label. We estimate $\theta$ of a linear classifier $\mathrm{sign}(\theta^T X)$ by:

$$\widehat{\theta} = \underset{\theta\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(-y_i\theta^\top X_i)).$$

We cannot apply Theorem 1 directly because the empirical logistic risk is not strongly convex; it does not satisfy assumption ($F_1$). Instead, we consider the convex function

$$f(\theta) = \frac{1}{2}\left(c + \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp(-y_i\theta^\top X_i)\right)\right)^2, \quad \text{where } c > 0 \text{ (e.g., } c = 1). \tag{13}$$

The gradient of $f(\theta)$ is a product of two terms

$$\nabla f(\theta) = \underbrace{\left(c + \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp(-y_i\theta^\top X_i)\right)\right)}_{\Psi} \times \underbrace{\nabla\left(\frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp(-y_i\theta^\top X_i)\right)\right)}_{\Upsilon}.$$

Therefore, we can compute $g_s = \Psi_s\Upsilon_s$, *using two independent random variables satisfying* $\mathbb{E}[\Psi_s \mid \theta] = \Psi$ and $\mathbb{E}[\Upsilon_s \mid \theta] = \Upsilon$. For $\Upsilon_s$, we have $\Upsilon_s = \frac{1}{S_\Upsilon}\sum_{i\in I_t^\Upsilon} \nabla\log(1 + \exp(-y_i\theta^\top X_i))$, where $I_t^\Upsilon$ are $S_\Upsilon$ indices sampled from $[n]$ uniformly at random with replacement. For $\Psi_s$, we have $\Psi_s = c + \frac{1}{S_\Psi}\sum_{i\in I_t^\Psi} \log(1 + \exp(-y_i\theta^\top X_i))$, where $I_t^\Psi$ are $S_\Psi$ indices uniformly sampled from $[n]$ with or without replacement. Given the above, we have $\nabla f(\theta)^\top(\theta - \widehat{\theta}) \geq \alpha\|\theta - \widehat{\theta}\|_2^2$ for some constant $\alpha$ by the generalized self-concordance of logistic regression [1, 2], and therefore the assumptions are now satisfied.

For convenience, we write $k(\theta) = \frac{1}{n}\sum_{i=1}^n k_i(\theta)$ where $k_i(\theta) = \log(1 + \exp(-y_i\theta^\top X_i))$. Thus $f(\theta) = (k(\theta + c)^2$, $\mathbb{E}[\Psi_s \mid \theta] = k(\theta) + c$, and $\mathbb{E}[\Upsilon_s \mid \theta] = \nabla k(\theta)$.

**Corollary 2.** *Assume $\|\theta_1 - \widehat{\theta}\|_2^2 = O(\eta)$; also $S_\Psi = O(1)$, $S_\Upsilon = O(1)$ are bounded. Then, we have*

$$\left\| t\mathbb{E}\left[(\bar{\theta}_t - \widehat{\theta})(\bar{\theta}_t - \widehat{\theta})^\top\right] - H^{-1}GH^{-1} \right\|_2 \lesssim \sqrt{\eta} + \sqrt{\tfrac{1}{t\eta} + t\eta^2},$$

*where $H = \nabla^2 f(\widehat{\theta}) = (c + k(\widehat{\theta}))\nabla^2 k(\widehat{\theta})$. Here, $G = \frac{1}{S_\Upsilon} K_G(\widehat{\theta}) \frac{1}{n}\sum_{i=1}^n \nabla k_i(\widehat{\theta}) k_i(\widehat{\theta})^\top$ with $K_G(\theta) = \mathbb{E}[\Psi(\theta)^2]$ depending on how indexes are sampled to compute $\Psi_s$:*

- *with replacement:* $K_G(\theta) = \frac{1}{S_\Psi}(\frac{1}{n}\sum_{i=1}^n (c + k_i(\theta))^2) + \frac{S_\Psi - 1}{S_\Psi}(c + k(\theta))^2$ ,
- *no replacement:* $K_G(\theta) = \frac{1 - \frac{S_\Psi - 1}{n-1}}{S_\Psi}(\frac{1}{n}\sum_{i=1}^n (c + k_i(\theta))^2) + \frac{S_\Psi - 1}{S_\Psi}\frac{n}{n-1}(c + k(\theta))^2$.

*Quantities other than $t$ and $\eta$ are data dependent constants.*

As with the results above, in the appendix we give data-dependent expressions for the constants. Simulations suggest that the term $t\eta^2$ in our bound is an artifact of our analysis. Because in logistic regression the estimate's covariance is $\frac{\left(\nabla^2 k(\widehat{\theta})\right)^{-1}}{n}\left(\frac{\sum_{i=1}^n \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top}{n}\right)\left(\nabla^2 k(\widehat{\theta})\right)^{-1}$, we set the scaling factor $K_s = \frac{(c+k(\widehat{\theta}))^2}{K_G(\widehat{\theta})}$ in (7) for statistical inference. Note that $K_s \approx 1$ for sufficiently large $S_\Psi$.

# 4 Related work

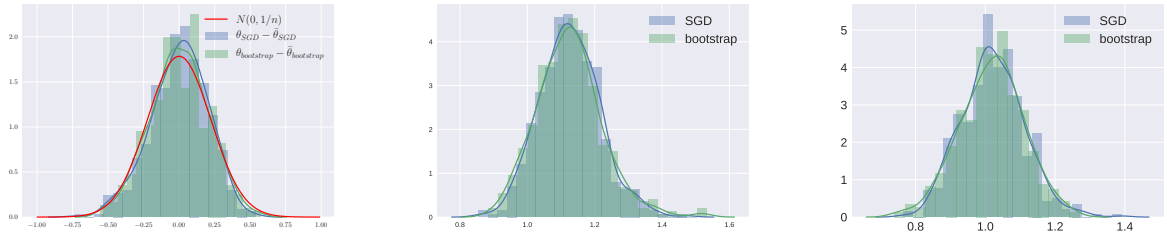
*Bayesian inference:* First and second order iterative optimization algorithms – including SGD, gradient descent, and variants – naturally define a Markov chain. Based on this principle, most related to this work is the case of stochastic gradient Langevin dynamics (SGLD) for Bayesian inference – namely, for sampling from the posterior distributions – using a variant of SGD [26, 6, 15, 16]. We note that, here as well, the vast majority of the results rely on using a decreasing step size. Very recently, [16] uses a heuristic approximation for Bayesian inference, and provides results for fixed step size.

Our problem is different in important ways from the Bayesian inference problem. In such likelihood parameter estimation problems, the covariance of the estimator only depends on the gradient of the likelihood function. This is not the case, however, in general frequentist $M$-estimation problems (*e.g.*, linear regression). In these cases, the covariance of the estimator depends both on the gradient and Hessian of the empirical risk function. For this reason, without second order information, SGLD methods are poorly suited for general $M$-estimation problems in frequentist inference. In contrast, our method exploits properties of averaged SGD, and computes the estimator's covariance without second order information.

*Connection with Bootstrap methods:* The classical approach for statistical inference is to use the bootstrap [9, 22]. Bootstrap samples are generated by replicating the entire data set by resampling, and then solving the optimization problem on each generated set of the data. We identify our algorithm and its analysis as an alternative to bootstrap methods. Our analysis is also specific to SGD, and thus sheds light on the statistical properties of this very widely used algorithm.

*Connection with stochastic approximation methods:* It has been long observed in stochastic approximation that under certain conditions, SGD displays asymptotic normality for both the setting of *decreasing step size*, e.g., [14, 20], and more recently, [23, 7]; and also for *fixed step size*, e.g., [4], Chapter 4. All of these results, however, provide their guarantees with the requirement that the stochastic approximation iterate converges to the optimum. For decreasing step size, this is not an overly burdensome assumption, since with mild assumptions it can be shown directly. As far as we know, however, it is not clear if this holds in the fixed step size regime. To side-step this issue, [4] provides results only when the (constant) step-size approaches 0 (see Section 4.4 and 4.6, and in particular Theorem 7 in [4]). Similarly, while [13] has asymptotic results on the average of consecutive stochastic approximation iterates with constant step size, it assumes convergence of iterates (assumption A1.7 in Ch. 10) – an assumption we are unable to justify in even simple settings.

Beyond the critical difference in the assumptions, the majority of the "classical" subject matter seeks to prove asymptotic results about different flavors of SGD, but does not properly consider its use for inference. Key exceptions are the recent work in [23] and [7], which follow up on [20]. Both of these rely on decreasing step size, for reasons mentioned above. The work in [7] uses SGD with decreasing step size for estimating




(a) Normal. (b) Exponential. (c) Poisson.

Figure 2: Estimation in univariate models.

an $M$-estimate's covariance. Work in [23] studies implicit SGD with decreasing step size and proves results similar to [20], however it does not use SGD to compute confidence intervals.

Overall, to the best of our knowledge, there are no prior results establishing asymptotic normality for SGD with fixed step size for general M-estimation problems (that do not rely on overly restrictive assumptions, as discussed).

# 5 Experiments

## 5.1 Synthetic data

The coverage probability is defined as $\frac{1}{p}\sum_{i=1}^{p}\mathbb{P}[\theta_i^\star \in \hat{C}_i]$ where $\theta^\star = \arg\min_\theta \mathbb{E}[f(\theta, X)] \in \mathbb{R}^p$, and $\hat{C}_i$ is the estimated confidence interval for the $i^{\text{th}}$ coordinate. The average confidence interval width is defined as $\frac{1}{p}\sum_{i=1}^{p}(\hat{C}_i^u - \hat{C}_i^l)$ where $[\hat{C}_i^l, \hat{C}_i^u]$ is the estimated confidence interval for the $i^{\text{th}}$ coordinate. In our experiments, coverage probability and average confidence interval width are estimated through simulation. We use the empirical quantile of our SGD inference procedure and bootstrap to compute the 95% confidence intervals for each coordinate of the parameter. Because theoretical justifications of our SGD inference procedure do not yet deal with pivotal quantities, here we have not included such comparisons. For results given as a pair $(\alpha, \beta)$, it usually indicates (coverage probability, confidence interval length).

### 5.1.1 Univariate models

In Figure 2, we compare our SGD inference procedure with $(i)$ Bootstrap and $(ii)$ normal approximation with inverse Fisher information in univariate models. We observe that our method and Bootstrap have similar statistical properties. Figure 8 in the appendix shows Q-Q plots of samples from our SGD inference procedure. *Normal distribution mean estimation:* Figure 2a compares 500 samples from SGD inference procedure and Bootstrap versus the distribution $\mathcal{N}(0, 1/n)$, using $n = 20$ i.i.d. samples from $\mathcal{N}(0, 1)$. We used mini batch SGD described in Sec. 3.3. For the parameters, we used $\eta = 0.8$, $t = 5$, $d = 10$, $b = 20$, and mini batch size of 2. Our SGD inference procedure gives (0.916 , 0.806), Bootstrap gives (0.926 , 0.841), and normal approximation gives (0.922 , 0.851). *Exponential distribution parameter estimation:* Figure 2b compares 500 samples from inference procedure and Bootstrap, using $n = 100$ samples from an exponential distribution with PDF $\lambda e^{-\lambda x}$ where $\lambda = 1$. We used SGD for MLE with mini batch sampled with replacement. For the parameters, we used $\eta = 0.1$, $t = 100$, $d = 5$, $b = 100$, and mini batch size of 5. Our SGD inference procedure gives (0.922, 0.364), Bootstrap gives (0.942 , 0.392), and normal approximation gives (0.922, 0.393). *Poisson distribution parameter estimation:* Figure 2c compares 500 samples from inference procedure and Bootstrap, using $n = 100$ samples from a Poisson distribution with PDF $\lambda^x e^{-\lambda x}$ where $\lambda = 1$. We used SGD for MLE with mini batch sampled with replacement. For the parameters, we used $\eta = 0.1$, $t = 100$, $d = 5$, $b = 100$, and mini batch size of 5. Our SGD inference procedure gives (0.942 , 0.364), Bootstrap gives (0.946 , 0.386), and normal approximation gives (0.960 , 0.393).

| $\eta$ | $t = 100$ | $t = 500$ | $t = 2500$ |
|---|---|---|---|
| 0.1 | (0.957, 4.41) | (0.955, 4.51) | (0.960, 4.53) |
| 0.02 | (0.869, 3.30) | (0.923, 3.77) | (0.918, 3.87) |
| 0.004 | (0.634, 2.01) | (0.862, 3.20) | (0.916, 3.70) |

(a) Bootstrap (0.941, 4.14), normal approximation (0.928, 3.87)

| $\eta$ | $t = 100$ | $t = 500$ | $t = 2500$ |
|---|---|---|---|
| 0.1 | (0.949, 4.74) | (0.962, 4.91) | (0.963, 4.94) |
| 0.02 | (0.845, 3.37) | (0.916, 4.01) | (0.927, 4.17) |
| 0.004 | (0.616, 2.00) | (0.832, 3.30) | (0.897, 3.93) |

(b) Bootstrap (0.938, 4.47), normal approximation (0.925, 4.18)

Table 1: Linear regression. *Left*: Experiment 1, *Right*: Experiment 2.

| $\eta$ | $t = 100$ | $t = 500$ | $t = 2500$ |
|---|---|---|---|
| 0.1 | (0.872, 0.204) | (0.937, 0.249) | (0.939, 0.258) |
| 0.02 | (0.610, 0.112) | (0.871, 0.196) | (0.926, 0.237) |
| 0.004 | (0.312, 0.051) | (0.596, 0.111) | (0.86, 0.194) |

(a) Bootstrap (0.932, 0.253), normal approximation (0.957, 0.264)

| $\eta$ | $t = 100$ | $t = 500$ | $t = 2500$ |
|---|---|---|---|
| 0.1 | (0.859, 0.206) | (0.931, 0.255) | (0.947, 0.266) |
| 0.02 | (0.600, 0.112) | (0.847, 0.197) | (0.931, 0.244) |
| 0.004 | (0.302, 0.051) | (0.583, 0.111) | (0.851, 0.195) |

(b) Bootstrap (0.932, 0.245), normal approximation (0.954, 0.256)

Table 2: Logistic regression. *Left*: Experiment 1, *Right*: Experiment 2.

#### 5.1.2 Multivariate models

In these experiments, we set $d = 100$, used mini-batch size of $4$, and used 200 SGD samples. In all cases, we compared with Bootstrap using 200 replicates. We computed the coverage probabilities using 500 simulations. Also, we denote $1_p = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^\top \in \mathbb{R}^p$. Additional simulations comparing covariance matrix computed with different methods are given in Sec. B.1.2.

**Linear regression:** *Experiment 1:* Results for the case where $X \sim \mathcal{N}(0, I) \in \mathbb{R}^{10}$, $Y = {w^*}^T X + \epsilon$, $w^* = 1_p/\sqrt{p}$, and $\epsilon \sim \mathcal{N}(0, \sigma^2 = 10^2)$ with $n = 100$ samples is given in Table 1a. Bootstrap gives (0.941, 4.14), and confidence intervals computed using the error covariance and normal approximation gives (0.928, 3.87). *Experiment 2:* Results for the case where $X \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^{10}$, $\Sigma_{ij} = 0.3^{|i-j|}$, $Y = {w^*}^T X + \epsilon$, $w^* = 1_p/\sqrt{p}$, and $\epsilon \sim \mathcal{N}(0, \sigma^2 = 10^2)$ with $n = 100$ samples is given in Table 1b. Bootstrap gives (0.938, 4.47), and confidence intervals computed using the error covariance and normal approximation gives (0.925, 4.18).

**Logistic regression:** Here we show results for logistic regression trained using vanilla SGD with mini batch sampled with replacement. Results for modified SGD (Sec. 3.5) are given in Sec. B.1.2. *Experiment 1:* Results for the case where $\mathbb{P}[Y = +1] = \mathbb{P}[Y = -1] = 1/2$, $X \mid Y \sim \mathcal{N}(0.01 Y 1_p/\sqrt{p}, I) \in \mathbb{R}^{10}$ with $n = 1000$ samples is given in Table 2a. Bootstrap gives (0.932, 0.245), and confidence intervals computed using inverse Fisher matrix as the error covariance and normal approximation gives (0.954, 0.256). *Experiment 2:* Results for the case where $\mathbb{P}[Y = +1] = \mathbb{P}[Y = -1] = 1/2$, $X \mid Y \sim \mathcal{N}(0.01 Y 1_p/\sqrt{p}, \Sigma) \in \mathbb{R}^{10}$, $\Sigma_{ij} = 0.2^{|i-j|}$ with $n = 1000$ samples is given in Table 2b. Bootstrap gives (0.932, 0.253), and confidence intervals computed using inverse Fisher matrix as the error covariance and normal approximation gives (0.957, 0.264).

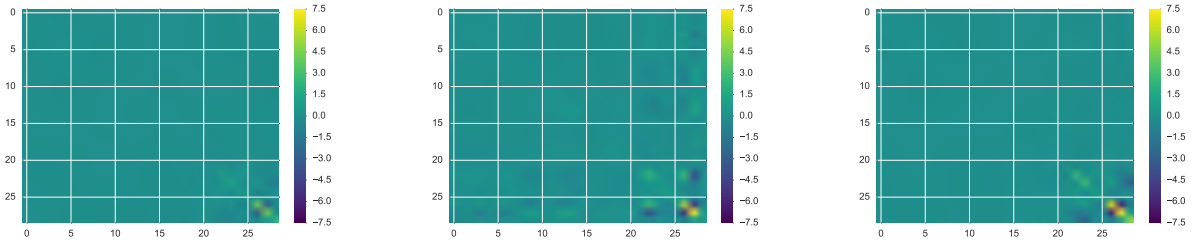
### 5.2 Real data

Here, we compare covariance matrix computed using our SGD inference procedure, bootstrap, and inverse Fisher information matrix on the Higgs data set [3] and the LIBSVM Splice data set, and we observe that they have similar statistical properties.

#### 5.2.1 Higgs data set

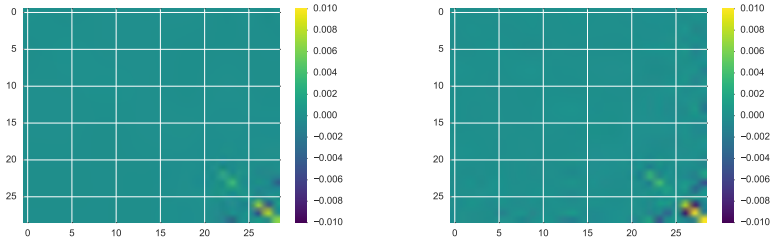
The Higgs data set [3] [3] contains 28 distinct features with 11,000,000 data samples. This is a classification problem between two types of physical processes: one produces Higgs bosons and the other is a background process that does not. We use a logistic regression model, trained using vanilla SGD, instead of the modified SGD described in Section 3.5.

[3] `https://archive.ics.uci.edu/ml/datasets/HIGGS`



(a) Inverse Fisher information (b) SGD inference covariance (c) Bootstrap estimated covariance

Figure 3: Higgs data set with $n = 200$



(a) Inverse Fisher information (b) SGD inference covariance

Figure 4: Higgs data set with $n = 50000$

To understand different settings of sample size, we subsampled the data set with different sample size levels: $n = 200$ and $n = 50000$. We investigate the empirical performance of SGD inference on this subsampled data set. In all experiments below, the batch size of the mini batch SGD is 10.

In the case $n = 200$, the asymptotic normality for the MLE is not a good enough approximation. Hence, in this small-sample inference, we compare the SGD inference covariance matrix with the one obtained by inverse Fisher information matrix and bootstrap in Figure 3.

For our SGD inference procedure, we use $t = 100$ samples to average, and discard $d = 50$ samples. We use $R = 20$ averages from 20 segments (as in Figure 1). For bootstrap, we use 2000 replicates, which is much larger than the sample size $n = 200$.

Figure 3 shows that the covariance matrix obtained by SGD inference is comparable to the estimation given by bootstrap and inverse Fisher information.

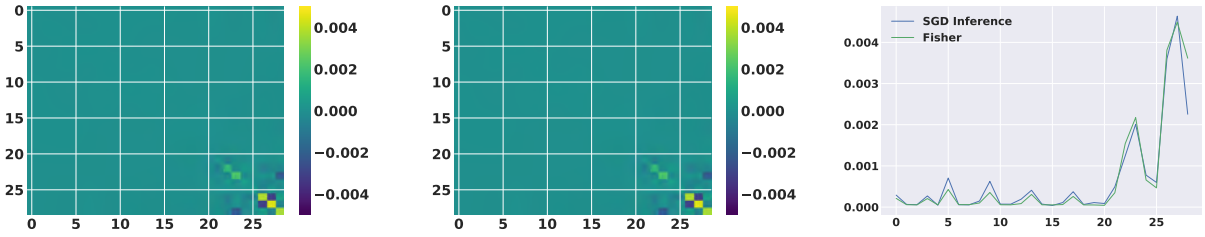
In the case $n = 50000$, we use $t = 5000$ samples to average, and discard $d = 500$ samples. We use $R = 20$ averages from 20 segments (as in Figure 1). For this large data set, we present the estimated covariance of SGD inference procedure and inverse Fisher information (the asymptotic covariance) in Figure 4 because bootstrap is computationally prohibitive. Similar to the small sample result in Figure 3, the covariance of our SGD inference procedure is comparable to the inverse Fisher information.

In Figure 5, we compare the covariance matrix computed using our SGD inference procedure and inverse Fisher information with $n = 90000$ samples . We used 25 samples from our SGD inference procedure with $t = 5000$, $d = 1000$, $\eta = 0.2$, and mini batch size of 10.

### 5.2.2 Splice data set

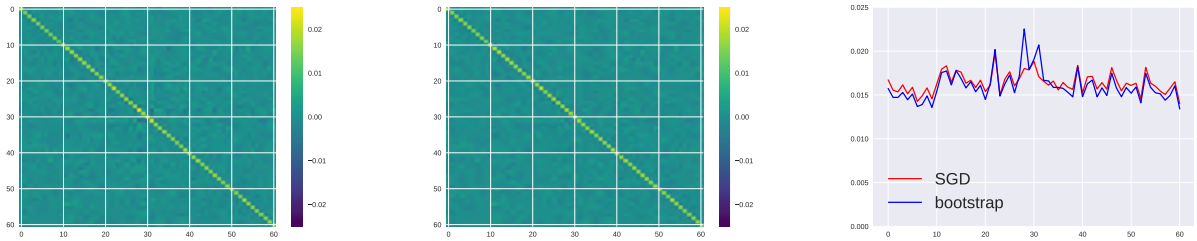
The Splice data set [4] contains 60 distinct features with 1000 data samples. This is a classification problem between two classes of splice junctions in a DNA sequence. Similar to the Higgs data set, we use a logistic regression model, trained using vanilla SGD.

[4] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html




(a) Inverse Fisher information (b) SGD inference covariance (c) Diagonal terms

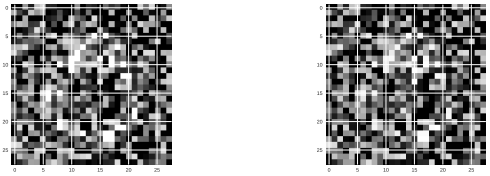
Figure 5: Higgs data set with $n = 90000$




(a) Bootstrap (b) SGD inference covariance (c) Diagonal terms

Figure 6: Splice data set

In Figure 6, we compare the covariance matrix computed using our SGD inference procedure and bootstrap $n = 1000$ samples. We used 10000 samples from both bootstrap and our SGD inference procedure with $t = 500$, $d = 100$, $\eta = 0.2$, and mini batch size of 6.



(a) Original "0": logit -46.3, CI (-64.2, -27.9) (b) Adversarial "0": logit 16.5, CI (-10.9, 30.5)

Figure 7: MNIST

### 5.2.3 MNIST

Here, we train a binary logistic regression classifier to classify 0/1 using perturbed MNIST data set, and demonstrate that certain adversarial examples (e.g. [10]) can be detected using prediction confidence intervals. For each image, where each original pixel is either 0 or 1, we randomly changed 70% pixels to random numbers uniformly on $[0, 0.9]$. Figure 7 shows each image's logit value ($\log \frac{\mathbb{P}[1|\text{image}]}{\mathbb{P}[0|\text{image}]}$) and its 95% confidence interval (CI) computed using our SGD inference procedure.

### 5.3 Discussion

In our experiments, we observed that using a larger step size $\eta$ produces accurate results with significantly accelerated convergence time. This might imply that the $\eta$ term in Theorem 1's bound is an artifact of our analysis. Indeed, although Theorem 1 only applies to SGD with fixed step size, where $\eta t \to \infty$ and $\eta^2 t \to 0$ imply that the step size should be smaller when the number of consecutive iterates used for the average is larger, our experiments suggest that we can use a (data dependent) constant step size $\eta$ and only require $\eta t \to \infty$.

In the experiments, our SGD inference procedure uses $(t + d) \cdot S \cdot p$ operations to produce a sample, and Newton method uses $n \cdot$(matrix inversion complexity $= \Omega(p^2)) \cdot$(number of Newton iterations $t$) operations to produce a sample. The experiments therefore suggest that our SGD inference procedure produces results similar to Bootstrap while using far fewer operations.

# References


[1] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[2] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

[3] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5, 2014.

[4] A. Benveniste, P. Priouret, and M. Métivier. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag New York, Inc., New York, NY, USA, 1990.

[5] S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, Nov. 2015.

[6] S. Bubeck, R. Eldan, and J. Lehec. Finite-time analysis of projected langevin monte carlo. In *Advances in Neural Information Processing Systems*, pages 1243–1251, 2015.

[7] X. Chen, J. Lee, X. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*, 2016.

[8] B. Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[9] B. Efron and R. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[10] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[11] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[12] H. Kushner and H. Huang. Asymptotic properties of stochastic approximations with constant coefficients. *SIAM Journal on Control and Optimization*, 19(1):87–105, 1981.

[13] H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer New York, 2003.

[14] L. Ljung, G. C. Pflug, and H. Walk. *Stochastic approximation and optimization of random systems*, volume 17. Birkhäuser, 2012.

[15] S. Mandt, M. Hoffman, and D. Blei. A Variational Analysis of Stochastic Gradient Algorithms. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 354–363, 2016.

[16] S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic Gradient Descent as Approximate Bayesian Inference. *arXiv preprint arXiv:1704.04289*, 2017.

[17] D. Opitz and R. Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

[18] G. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

[19] D. Politis, J. Romano, and M. Wolf. *Subsampling*. Springer Series in Statistics. Springer New York, 2012.

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

[21] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[22] J. Shao and D. Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012.

[23] P. Toulis and E. M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *arXiv preprint arXiv:1408.2923*, 2014.

[24] A. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[25] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[26] M. Welling and Y. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

# A Proofs

## A.1 Proof of Theorem 1

We first assume that $\theta_1 = \widehat{\theta}$ for more precise constants in our bounds, the same analysis applies when $\|\theta_1\|_2^2$

For ease of notation, we denote

$$\Delta_t = \theta_t - \widehat{\theta}, \tag{14}$$

and, without loss of generality, we assume that $\widehat{\theta} = 0$. The stochastic gradient descent recursion satisfies:

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta \cdot g_s(\theta_t) \\ &= \theta_t - \eta \cdot (g_s(\theta_t) - \nabla f(\theta_t) + \nabla f(\theta_t)) \\ &= \theta_t - \eta \cdot \nabla f(\theta_t) - \eta \cdot e_t,\end{aligned}$$

where $e_t = g_s(\theta_t) - \nabla f(\theta_t)$. Note that $e_1, e_2, \ldots$ is a martingale difference sequence. We use

$$g_i = \nabla f_i(\widehat{\theta}) \qquad \text{and} \qquad H_i = \nabla^2 f_i(\widehat{\theta}) \tag{15}$$

to denote the gradient component at index $i$, and the Hessian component at index $i$, at optimum $\widehat{\theta}$, respectively. Note that $\sum g_i = 0$ and $\frac{1}{n}\sum H_i = H$.

For each $f_i$, its Taylor expansion around $\widehat{\theta}$ is

$$f_i(\theta) = f_i(\widehat{\theta}) + g_i^\top(\theta - \widehat{\theta}) + \frac{1}{2}(\theta - \widehat{\theta})^\top H_i(\theta - \widehat{\theta}) + R_i(\theta, \widehat{\theta}), \tag{16}$$

where $R_i(\theta, \widehat{\theta})$ is the remainder term. For convenience, we write $R = \frac{1}{n}\sum R_i$.

For the proof, we require the following lemmata. The following lemma states that $\mathbb{E}[\|\Delta_t\|_2^2] = O(\eta)$ as $t \to \infty$ and $\eta \to 0$.

**Lemma 1.** *For data dependent, positive constants $\alpha, A, B$ according to assumptions $(F_1)$ and $(G_2)$ in Theorem 1, and given assumption $(G_1)$, we have*

$$\mathbb{E}\left[\|\Delta_t\|_2^2\right] \leq (1 - 2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \frac{B\eta}{2\alpha - A\eta}, \tag{17}$$

*under the assumption $\eta < \frac{2\alpha}{A}$.*

*Proof.* As already stated, we assume without loss of generality that $\widehat{\theta} = 0$. This further implies that: $g_s(\theta_t) = g_s(\theta_t - \widehat{\theta}) = g_s(\Delta_t)$, and

$$\Delta_{t+1} = \Delta_t - \eta \cdot g_s(\Delta_t).$$

Given the above and assuming expectation $\mathbb{E}[\cdot]$ w.r.t. the selection of a sample from $\{X_i\}_{i=1}^n$, we have:

$$\begin{aligned}\mathbb{E}\left[\|\Delta_{t+1}\|_2^2 \mid \Delta_t\right] &= \mathbb{E}\left[\|\Delta_t - \eta g_s(\Delta_t)\|_2^2 \mid \Delta_t\right] \\ &= \mathbb{E}\left[\|\Delta_t\|_2^2 \mid \Delta_t\right] + \eta^2 \cdot \mathbb{E}\left[\|g_s(\Delta_t)\|_2^2 \mid \Delta_t\right] - 2\eta \cdot \mathbb{E}\left[g_s(\Delta_t)^\top \Delta_t \mid \Delta_t\right] \\ &= \|\Delta_t\|_2^2 + \eta^2 \cdot \mathbb{E}\left[\|g_s(\Delta_t)\|_2^2 \mid \Delta_t\right] - 2\eta \cdot \nabla f(\Delta_t)^\top \Delta_t \\ &\overset{(i)}{\leq} \|\Delta_t\|_2^2 + \eta^2 \cdot \left(A \cdot \|\Delta_t\|_2^2 + B\right) - 2\eta \cdot \alpha \|\Delta_t\|_2^2 \\ &= (1 - 2\alpha\eta + A\eta^2)\|\Delta_t\|_2^2 + \eta^2 B.\end{aligned} \tag{18}$$

where $(i)$ is due to assumptions $(F_1)$ and $(G_2)$ of Theorem 1. Taking expectations for every step $t = 1, \cdot$ over the whole history, we obtain the recursion:

$$\begin{aligned}\mathbb{E}\left[\|\Delta_{t+1}\|_2^2\right] &\leq (1 - 2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \eta^2 B \cdot \sum_{i=0}^{t-1}(1 - 2\alpha\eta + A\eta^2)^i \\ &= (1 - 2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \eta^2 B \cdot \tfrac{1-(1-2\alpha\eta+A\eta^2)^t}{2\alpha\eta - A\eta^2} \\ &\leq (1 - 2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \tfrac{\eta B}{2\alpha - A\eta}.\end{aligned}$$

□

The following lemma states that $\mathbb{E}[\|\Delta_t\|_2^4] = O(\eta^2)$ as $t \to \infty$ and $\eta \to 0$.

**Lemma 2.** *For data dependent, positive constants $\alpha, A, B, C, D$ according to assumptions $(F_1)$, $(G_1)$, $(G_2)$ in Theorem 1, we have:*

$$\begin{aligned}\mathbb{E}[\|\Delta_t\|_2^4] \leq &(1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + B(3\eta + \eta^2) + C(2\eta^3 + \eta^4))^{t-1}\|\Delta_1\|_2^4 \\ &+ \frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - B(3 + \eta) - C(2\eta^2 + \eta^3)}.\end{aligned} \tag{19}$$

*Proof.* Given $\Delta_t$, we have the following sets of (in)equalities:

$$\begin{aligned}&\mathbb{E}\left[\|\Delta_{t+1}\|_2^4 \mid \Delta_t\right] \\ =&\mathbb{E}\left[\|\Delta_t - \eta g_s(\Delta_t)\|_2^4 \mid \Delta_t\right] \\ =&\mathbb{E}\left[(\|\Delta_t\|_2^2 - 2\eta \cdot g_s(\Delta_t)^\top \Delta_t + \eta^2\|g_s(\Delta_t)\|_2^2)^2 \mid \Delta_t\right] \\ =&\mathbb{E}\big[\|\Delta_t\|_2^4 + 4\eta^2(g_s(\Delta_t)^\top \Delta_t)^2 + \eta^4\|g_s(\Delta_t)\|_2^4 - 4\eta \cdot g_s(\Delta_t)^\top \Delta_t\|\Delta_t\|_2^2 \\ &\qquad + 2\eta^2 \cdot \|g_s(\Delta_t)\|_2^2\|\Delta_t\|_2^2 - 4\eta^3 \cdot g_s(\Delta_t)^\top \Delta_t\|g_s(\Delta_t)\|_2^2 \mid \Delta_t\big] \\ \overset{(i)}{\leq}&\mathbb{E}\big[\|\Delta_t\|_2^4 + 4\eta^2 \cdot \|g_s(\Delta_t)\|_2^2 \cdot \|\Delta_t\|_2^2 + \eta^4\|g_s(\Delta_t)\|_2^4 - 4\eta \cdot g_s(\Delta_t)^\top \Delta_t\|\Delta_t\|_2^2 \\ &\qquad + 2\eta^2 \cdot \|g_s(\Delta_t)\|_2^2 \cdot \|\Delta_t\|_2^2 + 2\eta^3 \cdot (\|g_s(\Delta_t)\|_2^2 + \|\Delta_t\|_2^2) \cdot \|g_s(\Delta_t)\|_2^2 \mid \Delta_t\big] \\ \overset{(ii)}{\leq}&\mathbb{E}\left[\|\Delta_t\|_2^4 + (2\eta^3 + \eta^4)\|g_s(\Delta_t)\|_2^4 + (6\eta^2 + 2\eta^3)\|g_s(\Delta_t)\|_2^2\|\Delta_t\|_2^2 \mid \Delta_t\right] - 4\alpha\eta\|\Delta_t\|_2^4 \\ \overset{(iii)}{\leq}&(1 - 4\alpha\eta)\|\Delta_t\|_2^4 + (6\eta^2 + 2\eta^3)(A\|\Delta_t\|_2^2 + B)\|\Delta_t\|_2^2 + (2\eta^3 + \eta^4)(C\|\Delta_t\|_2^4 + D) \\ =&(1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + C(2\eta^3 + \eta^4))\|\Delta_t\|_2^4 + B(6\eta^2 + 2\eta^3)\|\Delta_t\|_2^2 + D(2\eta^3 + \eta^4) \\ \overset{(iv)}{\leq}&(1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + C(2\eta^3 + \eta^4)) \cdot \|\Delta_t\|_2^4 + B(3\eta + \eta^2)(\eta^2 + \|\Delta_t\|_2^4) + D(2\eta^3 + \eta^4) \\ =&(1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + B(3\eta + \eta^2) + C(2\eta^3 + \eta^4)) \cdot \|\Delta_t\|_2^4 + B\eta^2(3\eta + \eta^2) + D(2\eta^3 + \eta^4),\end{aligned} \tag{20}$$

where $(i)$ is due to $(g_s(\Delta_t)^\top \Delta_t)^2 \leq \|g_s(\Delta_t)\|_2^2 \cdot \|\Delta_t\|_2^2$ and $-2g_s(\Delta_t)^\top \Delta_t \leq \|g_s(\Delta_t)\|_2^2 + \|\Delta_t\|_2^2$, $(ii)$ is due to assumptions $(G_1)$ and $(F_1)$ in Theorem 1, $(iii)$ is due to assumptions $(G_2)$ and $(G_3)$ in Theorem 1, and $(iv)$ is due to $2\eta\|\Delta_t\|_2^2 \leq \eta^2 + \|\Delta_t\|_2^4$. Similar to the proof of the previous lemma, applying the above rule recursively and w.r.t. the whole history of estimates, we

obtain:

$$\begin{aligned}\mathbb{E}\left[\|\Delta_{t+1}\|_2^4\right] &\leq (1-4\alpha\eta+A(6\eta^2+2\eta^3)+B(3\eta+\eta^2)+C(2\eta^3+\eta^4))^{t-1}\|\Delta_1\|_2^4 \\ &\quad+\left(B\eta^2(3\eta+\eta^2)+D(2\eta^3+\eta^4)\right)\cdot\sum_{i=0}^{t-1}\left(1-4\alpha\eta+A(6\eta^2+2\eta^3)+B(3\eta+\eta^2)+C(2\eta^3+\eta^4)\right)^i \\ &\leq (1-4\alpha\eta+A(6\eta^2+2\eta^3)+B(3\eta+\eta^2)+C(2\eta^3+\eta^4))^{t-1}\|\Delta_1\|_2^4 \\ &\quad+\frac{B\eta^2(3\eta+\eta^2)+D(2\eta^3+\eta^4)}{4\alpha\eta-A(6\eta^2+2\eta^3)-B(3\eta+\eta^2)-C(2\eta^3+\eta^4)},\end{aligned}$$

which is the target inequality, after simple transformations. □

For SGD, we have

$$\begin{aligned}\Delta_t &= (I-\eta H)\Delta_{t-1}-\eta(\nabla R(\Delta_{t-1})+e_{t-1}) \\ &= (I-\eta H)^{t-1}\Delta_1-\eta\sum_{i=1}^{t-1}(I-\eta H)^{t-1-i}(e_i+\nabla R(\Delta_i)).\end{aligned} \tag{21}$$

For $t\geq 2$,

$$\begin{aligned}t(\bar{\theta}-\widehat{\theta}) &= \sum_{i=1}^{\top}\Delta_i \\ &= (I-(I-\eta H)^t)\frac{H^{-1}}{\eta}\Delta_1-\eta\sum_{j=1}^{t-1}\sum_{i=1}^{j}(I-\eta H)^{j-1-i}(e_i+\nabla R(\Delta_i)).\end{aligned} \tag{22}$$

For the latter term,

$$\begin{aligned}&\eta\sum_{j=1}^{t-1}\sum_{i=1}^{j}(I-\eta H)^{j-i}(e_i+\nabla R(\Delta_i)) \\ =&\eta\sum_{i=1}^{t-1}(\sum_{j=0}^{t-i-1}(I-\eta H)^j)(e_i+\nabla R(\Delta_i)) \\ =&\sum_{i=1}^{t-1}(I-(I-\eta H)^{t-i})H^{-1}(e_i+\nabla R(\Delta_i)) \\ =&H^{-1}\sum_{i=1}^{t-1}e_i+H^{-1}\sum_{i=1}^{t-1}\nabla R(\Delta_i)-H^{-1}\sum_{i=1}^{t-1}(I-\eta H)^{t-i}(e_i+\nabla R(\Delta_i)) \\ \overset{(i)}{=}&H^{-1}\sum_{i=1}^{t-1}e_i+H^{-1}\sum_{i=1}^{t-1}\nabla R(\Delta_i)+H^{-1}(I-\eta H)\frac{1}{\eta}(\Delta_t-(I-\eta H)^{t-1}\Delta_1),\end{aligned} \tag{23}$$

where step (i) follows from the fact $\sum_{i=1}^{t-1}(I-\eta H)^{t-i}(e_i+\nabla R(\Delta_i))=(I-\eta H)\frac{1}{\eta}(\Delta_t-(I-\eta H)^{t-1}\Delta_1)$ .

Thus, we have

$$\begin{aligned}\sqrt{t}\bar{\Delta}_t =&\frac{1}{\sqrt{t}}(I-(I-\eta H)^t)\frac{H^{-1}}{\eta}\Delta_1\\ &-\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}e_i\\ &-\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}\nabla R(\Delta_i)\\ &-\frac{1}{\sqrt{t}}H^{-1}(I-\eta H)\frac{1}{\eta}(\Delta_t-(I-\eta H)^{t-1}\Delta_1).\end{aligned} \tag{24}$$

In the statement of the theorem we have $\Delta_1 = 0$ (however similar bounds will hold if $\|\Delta_1\|_2^2 = O(\eta)$), thus for above terms we have

$$\frac{1}{\sqrt{t}}(I-(I-\eta H)^t)\frac{H^{-1}}{\eta}\Delta_1 = 0, \tag{25}$$

$$\begin{aligned}&\mathbb{E}[\|\frac{1}{\sqrt{t}}H^{-1}(I-\eta H)\frac{1}{\eta}(\Delta_t-(I-\eta H)^{t-1}\Delta_1)\|_2^2]\\ \leq&\frac{1-\eta\lambda_U}{\lambda_L}\mathbb{E}[\frac{\|\Delta_t\|_2^2}{\eta^2 t}]\\ \leq&\frac{1-\eta\lambda_U}{\lambda_L}\frac{1}{\eta^2 t}((1-2\alpha\eta+A\eta^2)^{t-1}\|\Delta_1\|_2^2+\frac{B\eta}{2\alpha-A\eta})\\ \leq&\frac{1-\eta\lambda_U}{\lambda_L}\frac{B}{t\eta(2\alpha-A\eta)}\\ =&O(\frac{1}{t\eta}).\end{aligned} \tag{26}$$

$$
\begin{aligned}
&\mathbb{E}[\|\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}\nabla R(\Delta_i)\|_2^2] \\
\leq&\mathbb{E}[\frac{1}{\lambda_L}\frac{1}{t}(\sum_{i=1}^{t-1}\|\nabla R(\Delta_i)\|_2)^2] \\
\leq&\mathbb{E}[\frac{E^2}{\lambda_L t}(\sum_{i=1}^{t-1}\|\Delta_i\|_2^2)^2] \\
\leq&\frac{E^2}{\lambda_L t}(t-1)\mathbb{E}[\sum_{i=1}^{t-1}\|\Delta_i\|_2^4] \\
\leq&\frac{E^2}{\lambda_L}\frac{t}{t-1}\sum_{i=1}^{t-1}((1-4\alpha\eta+A(6\eta^2+2\eta^3)+C(2\eta^3+\eta^4))^{t-1}\|\Delta_1\|_2^4+\frac{B(3\eta^2+\eta^3)+D(2\eta^2+\eta^3)}{4\alpha-A(6\eta+2\eta^2)-C(2\eta^2+\eta^3)}) \\
=&\frac{E^2}{\lambda_L}t\frac{B(3\eta^2+\eta^3)+D(2\eta^2+\eta^3)}{4\alpha-A(6\eta+2\eta^2)-C(2\eta^2+\eta^3)} \\
=&O(t\eta^2).
\end{aligned} \tag{27}
$$

For the term $-\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}e_i$, we have

$$
\begin{aligned}
&\mathbb{E}[\|-\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}e_i\|_2^2] \\
\overset{(i)}{=}&\frac{1}{t}\sum_{i=1}^{t-1}\mathbb{E}[\|H^{-1}e_i\|_2^2] \\
\leq&\frac{\lambda_U}{t}\sum_{i=1}^{t-1}\mathbb{E}[\|e_i\|_2^2] \\
=&\frac{\lambda_U}{t}\sum_{i=1}^{t-1}\mathbb{E}[\|g_s(\Delta_i)-\nabla f(\Delta_i)\|_2^2] \\
\leq&2\frac{\lambda_U}{t}(\sum_{i=1}^{t-1}\mathbb{E}[\|g_s(\Delta_i)\|_2^2]+\sum_{i=1}^{t-1}\mathbb{E}[\|\nabla f(\Delta_i)\|_2^2]) \\
\leq&2\frac{\lambda_U}{t}((t-1)B+(A+L^2)\sum_{i=1}^{t-1}\|\Delta_i\|_2^2) \\
\leq&2\frac{\lambda_U}{t}((t-1)B+(A+L^2)\sum_{i=1}^{t-1}((1-2\alpha\eta+A\eta^2)^{t-1}\|\Delta_1\|_2^2+\frac{B\eta}{2\alpha-A\eta})) \\
=&2\lambda_U\frac{t-1}{t}(B+(A+L^2)\frac{B\eta}{2\alpha-A\eta}) \\
=&O(1),
\end{aligned} \tag{28}
$$

where step (i) follows from $i \neq j$ leading to $\mathbb{E}[(H^{-1}e_i)^\top H^{-1}e_j] = 0$. We also have

$$\begin{aligned}&\mathbb{E}[(-\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}e_i)(-\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}e_i)^\top]\\=&\frac{1}{t}H^{-1}(\sum_{i=1}^{t-1}\mathbb{E}[e_ie_i^\top])H^{-1}.\end{aligned} \tag{29}$$

For each term $\mathbb{E}[e_ie_i^\top]$, we have

$$\begin{aligned}&\|\mathbb{E}[e_ie_i^\top] - G\|_2\\=&\|\mathbb{E}[g_s(\Delta_i)g_s(\Delta_i)^\top] - \mathbb{E}[(\nabla f(\Delta_i))(\nabla f(\Delta_i))^\top] - G\|_2\\\leq&\mathbb{E}[\|\nabla f(\Delta_i)\|_2^2] + \mathbb{E}[A_1\|\Delta_i\|_2 + A_2\|\Delta_i\|_2^2 + A_3\|\Delta_i\|_2^3 + A_4\|\Delta_i\|_2^4]\\\leq&L^2\mathbb{E}[\|\Delta_i\|_2^2] + A_1\sqrt{\mathbb{E}[\|\Delta_i\|_2^2]} + A_2\mathbb{E}[\|\Delta_i\|_2^2] + \frac{A_3}{2}\mathbb{E}[\|\Delta_i\|_2^2 + \|\Delta_i\|_2^4] + A_4\mathbb{E}[\|\Delta_i\|_2^4]\\=&A_1\sqrt{\mathbb{E}[\|\Delta_i\|_2^2]} + (L^2 + A_2 + \frac{A_3}{2})\mathbb{E}[\|\Delta_i\|_2^2] + (\frac{A_3}{2} + A_4)\mathbb{E}[\|\Delta_i\|_2^4]\\\leq&A_1\sqrt{(1-2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \frac{B\eta}{2\alpha - A\eta}} + (L^2 + A_2 + \frac{A_3}{2})((1-2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \frac{B\eta}{2\alpha - A\eta})\\&+ (\frac{A_3}{2} + A_4)((1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + C(2\eta^3 + \eta^4))^{t-1}\|\Delta_1\|_2^4 + \frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - C(2\eta^2 + \eta^3)})\\=&A_1\sqrt{\frac{B\eta}{2\alpha - A\eta}} + (L^2 + A_2 + \frac{A_3}{2})\frac{B\eta}{2\alpha - A\eta} + (\frac{A_3}{2} + A_4)\frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - C(2\eta^2 + \eta^3)}.\end{aligned} \tag{30}$$

Thus, we have

$$\begin{aligned}&\|\frac{1}{t}H^{-1}(\sum_{i=1}^{t-1}\mathbb{E}[e_ie_i^\top])H^{-1} - H^{-1}GH^{-1}\|_2\\\leq&\frac{1}{t}\|H^{-1}GH^{-1}\|_2\\&+ \frac{t-1}{t}\frac{1}{\lambda_L^2}(A_1\sqrt{\frac{B\eta}{2\alpha - A\eta}} + (L^2 + A_2 + \frac{A_3}{2})\frac{B\eta}{2\alpha - A\eta} + (\frac{A_3}{2} + A_4)\frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - C(2\eta^2 + \eta^3)})\\=&O(\sqrt{\eta}).\end{aligned} \tag{31}$$

For convenience, denote

$$\begin{aligned}
\square_0 &= \frac{1}{\sqrt{t}}(I-(I-\eta H)^t)\frac{H^{-1}}{\eta}\Delta_1,\\
\square_1 &= -\frac{1}{\sqrt{t}}H^{-1}(I-\eta H)\frac{1}{\eta}(\Delta_t-(I-\eta H)^{t-1}\Delta_1),\\
\square_2 &= -\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}\nabla R(\Delta_i),\\
\square_3 &= -\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}e_i,
\end{aligned} \tag{32}$$

and we have $\mathbb{E}[t\bar{\Delta}_t\bar{\Delta}_t]=\mathbb{E}[(\square_0+\square_1+\square_2+\square_3)(\square_0+\square_1+\square_2+\square_3)^\top]$.

Combining above results, we can bound

$$\begin{aligned}
&\|t\mathbb{E}[(\bar{\theta}_t-\widehat{\theta})(\bar{\theta}_t-\widehat{\theta})^\top]-H^{-1}GH^{-1}\|_2\\
=&\|\mathbb{E}[(\square_0+\square_1+\square_2+\square_3)(\square_0+\square_1+\square_2+\square_3)^\top]-H^{-1}GH^{-1}\|_2\\
=&\|\mathbb{E}[\square_3\square_3^\top]-H^{-1}GH^{-1}+\mathbb{E}[\square_3(\square_0+\square_1+\square_2)^\top+(\square_0+\square_1+\square_2)\square_3^\top+(\square_0+\square_1+\square_2)(\square_0+\square_1+\square_2)^\top]\|_2\\
\lesssim&\|\mathbb{E}[\square_3\square_3^\top]-H^{-1}GH^{-1}\|_2+\sqrt{\mathbb{E}[\|\square_3\|_2^2](\mathbb{E}[\|\square_0\|_2^2]+\mathbb{E}[\|\square_1\|_2^2]+\mathbb{E}[\|\square_2\|_2^2])}+\mathbb{E}[\|\square_0\|_2^2]+\mathbb{E}[\|\square_1\|_2^2]+\mathbb{E}[\|\square_2\|_2^2]\\
\lesssim&\sqrt{\eta}+\sqrt{\frac{1}{t\eta}+t\eta^2}.
\end{aligned} \tag{33}$$

Here we have used the fact that for two $p$-dimensional random vectors $a$ and $b$, the expectation of the matrix $ab^\top$ satisfies

$$\|\mathbb{E}[ab^\top]\|_2\le\sqrt{\mathbb{E}[\|a\|_2^2]\mathbb{E}[\|b\|_2^2]}\le\frac{1}{2}\mathbb{E}[\|a\|_2^2]+\mathbb{E}[\|b\|_2^2]. \tag{34}$$

Indeed, for any fixed unit vector $u$ we have $\|\mathbb{E}[ab^\top]u\|_2=\|\mathbb{E}[a(b^\top u)]\|_2\le\mathbb{E}[\|a\|_2|b^\top u|]\le\mathbb{E}[\|a\|_2\|b\|_2]\le\sqrt{\mathbb{E}[\|a\|_2^2]\mathbb{E}[\|b\|_2^2]}$. Here we used the fact $\|\mathbb{E}[x]\|_2\le\mathbb{E}[\|x\|_2]$ because $\|x\|_2$ is convex.

■

## A.2 Proof of Corollary 1

*Proof of Corollary 1.* Here we use the same notations as the proof of Theorem 1.

Because linear regression satisfies $\nabla f(\theta)-H(\theta-\widehat{\theta})=0$, we do not have to consider the Taylor remainder term in our analysis. And we do not need 4-th order bound for SGD.

Because the quadratic function is strongly convex, we have $\Delta^\top\nabla f(\Delta+\widehat{\theta})\ge\lambda_L\|\Delta\|_2^2$.

By sampling with replacement, we have

$$
\begin{aligned}
&\mathbb{E}[\|g_s(\theta_t)\|_2^2 \mid \theta_t] \\
=&\|\nabla f(\theta_t)\|_2^2 + \mathbb{E}[\|e_t\|_2^2 \mid \theta_t] \\
=&\|\nabla f(\theta_t)\|_2^2 + \frac{1}{S}(\frac{1}{n}\sum \|\nabla f_i(\theta_t)\|_2^2 - \|\nabla f(\theta_t)\|_2^2) \\
\leq& L^2(1-\frac{1}{S})\|\Delta_t\|_2^2 + \frac{1}{S}\frac{1}{n}\sum \|x_i(x_i^\top \theta_t - y_i)\|_2^2 \\
=& L^2(1-\frac{1}{S})\|\Delta_t\|_2^2 + \frac{1}{S}\frac{1}{n}\sum \|x_i x_i^\top \Delta_t + x_i x_i^\top \widehat{\theta} - y_i x_i\|_2^2 \\
\leq& L^2(1-\frac{1}{S})\|\Delta_t\|_2^2 + 2\frac{1}{S}\frac{1}{n}\sum (\|x_i x_i^\top \Delta_t\|_2^2 + \|x_i x_i^\top \widehat{\theta} - y_i x_i\|_2^2) \\
\leq& (L^2(1-\frac{1}{S}) + 2\frac{1}{S}\frac{1}{n}\sum \|x_i\|_2^4)\|\Delta_t\|_2^2 + 2\frac{1}{S}\frac{1}{n}\sum \|x_i x_i^\top \widehat{\theta} - y_i x_i\|_2^2.
\end{aligned} \tag{35}
$$

We also have

$$
\begin{aligned}
&\|\mathbb{E}[g_s(\theta) g_s(\theta)^\top \mid \theta] - G\|_2 \\
=&\|\frac{1}{S}\frac{1}{n}\sum \nabla f_i(\theta) f_i(\theta)^\top - \nabla f(\theta)\nabla f(\theta)^\top - G\|_2 \\
\leq& \|\nabla f(\theta)\|_2^2 + \frac{1}{S}\|\frac{1}{n}\sum \nabla f_i(\theta) f_i(\theta)^\top - G\|_2 \\
\leq& \|\nabla f(\theta)\|_2^2 + \frac{1}{S}\|\frac{1}{n}\sum (g_i + H_i\Delta)(g_i + H_i\Delta)^\top - G\|_2 \\
\leq& \|\nabla f(\theta)\|_2^2 + \frac{1}{S}\|\frac{1}{n}\sum H_i\Delta g_i^\top + g_i\Delta^\top H_i + H_i\Delta\Delta^\top H_i\|_2 \\
\leq& \|\nabla f(\theta)\|_2^2 + \frac{1}{S}(\frac{2}{n}\|H_i\|_2\|g_i\|_2)\|\Delta\|_2 + \frac{1}{S}(\frac{1}{n}\sum \|H_i\|_2^2)\|\Delta\|_2^2 \\
\leq& \frac{1}{S}(\frac{2}{n}\|H_i\|_2\|g_i\|_2)\|\Delta\|_2 + (L^2 + \frac{1}{S}\frac{1}{n}\sum \|H_i\|_2^2)\|\Delta\|_2^2,
\end{aligned} \tag{36}
$$

where $g_i = x_i(x_i^\top \widehat{\theta} - y_i)$ and $H_i = x_i x_i^\top$.

Following Theorem 1's proof, we have

$$
\|t\mathbb{E}[(\bar{\theta}_t - \widehat{\theta})(\bar{\theta}_t - \widehat{\theta})^\top] - H^{-1}GH^{-1}\|_2 \lesssim \sqrt{\eta} + \frac{1}{\sqrt{t\eta}}. \tag{37}
$$

□

## A.3 Proof of Corollary 2

*Proof of Corollary 2.* Here we use the same notations as the proof of Theorem 1.

Because $\nabla^2 f(\theta) = \nabla k(\theta)\nabla k(\theta)^\top + (k(\theta)+c)\nabla^2 k(\theta)$, $f(\theta)$ is convex.

The following lemma shows that $\nabla f(\theta) = (k(\theta)+c)\nabla k(\theta)$ is Lipschitz.

**Lemma 3.**

$$
\|\nabla f(\theta)\|_2 \leq L\|\Delta\|_2 \tag{38}
$$

*for some data dependent constant $L$.*

*Proof.* First, because

$$\nabla k(\theta) = \frac{1}{n}\sum -\frac{-y_i x_i}{1+\exp(y_i\theta^\top x_i)}, \tag{39}$$

we have

$$\|\nabla k(\theta)\|_2 \leq \frac{1}{n}\sum \|x_i\|_2. \tag{40}$$

Also, we have

$$\begin{aligned}\|\nabla^2 k(\theta)\|_2 =& \|\frac{1}{n}\sum \frac{\exp(y_i\theta^\top x_i)}{(1+\exp(y_i\theta^\top x_i))^2} x_i x_i^\top\|_2 \\ \leq& \frac{1}{4}\frac{1}{n}\sum \|x_i\|_2^2,\end{aligned} \tag{41}$$

which implies

$$\|\nabla k(\theta)\|_2 \leq \frac{1}{4}\frac{1}{n}\sum \|x_i\|_2^2\|\Delta\|_2. \tag{42}$$

And, we have

$$\begin{aligned}k(\theta) =& \frac{1}{n}\sum \log(1+\exp(-y_i\Delta^\top x_i - y_i\widehat{\theta}^\top x_i)) \\ \leq& \frac{1}{n}\sum \log(1+\exp(\|x_i\|_2\|\Delta\|_2 - y_i\widehat{\theta}^\top x_i)) \\ \overset{(i)}{\leq}& \frac{1}{n}\sum (\log(1+\exp(-y_i\widehat{\theta}^\top x_i)) + \|x_i\|_2\|\Delta\|_2)\end{aligned} \tag{43}$$

where step (i) follows from $\log(1+\exp(a+b)) \leq \log(1+e^b)+|a|$. Thus, we have

$$\begin{aligned}&\|\nabla f(\theta)\|_2 \\ =&\|(k(\theta)+c)\nabla k(\theta)\|_2 \\ \leq& k(\theta)\|\nabla k(\theta)\|_2 + c\|\nabla k(\theta)\|_2 \\ \leq& (c+\frac{1}{n}\sum \log(1+\exp(-y_i\widehat{\theta}^\top x_i)))\|\nabla k(\theta)\|_2 + (\frac{1}{n}\sum \|x_i\|_2)^2\|\Delta\|_2,\end{aligned} \tag{44}$$

and we can conclude that $\|\nabla f(\theta)\|_2 \leq L\|\Delta\|_2$ for some data dependent constant $L$. □

Next, we show that $f(\theta)$ has a bounded Taylor remainder.

**Lemma 4.**

$$\|\nabla f(\theta) - H(\theta-\widehat{\theta})\|_2 \leq E\|\theta-\widehat{\theta}\|_2^2, \tag{45}$$

*for some data dependent constant $E$.*

*Proof.* Because $\nabla f(\theta) = (k(\theta) + c)\nabla k(\theta)$, we know that $\|\nabla f(\theta)\|_2 = O(\|\Delta\|_2)$ when $\|\Delta\|_2 = \Omega(1)$ where the constants are data dependent.

Because $f(\theta)$ is infinitely differentiable, by the Taylor expansion we know that $\|\nabla f(\theta) - H(\theta - \widehat{\theta})\|_2 = O(\|\theta - \widehat{\theta}\|_2^2)$ when $\|\Delta\|_2 = O(1)$ where the constants are data dependent.

Combining the above, we can conclude $\|\nabla f(\theta) - H(\theta - \widehat{\theta})\|_2 \leq E\|\theta - \widehat{\theta}\|_2^2$ for some data dependent constant $E$.

□

In the following lemma, we will show that $\nabla f(\theta)^\top(\theta - \widehat{\theta}) \geq \alpha\|\theta - \widehat{\theta}\|_2^2$ for some data dependent constant $\alpha$.

**Lemma 5.**

$$\nabla f(\theta)^\top(\theta - \widehat{\theta}) \geq \alpha\|\theta - \widehat{\theta}\|_2^2, \tag{46}$$

*for some data dependent constant $\alpha$.*

*Proof.*

$$\nabla f(\theta)^\top\Delta = (k(\theta) + c)\nabla k(\theta)^\top\Delta. \tag{47}$$

First, notice that locally (when $\|\Delta\|_2 = O(\frac{\lambda_L}{E})$) we have

$$\nabla k(\theta)^\top\Delta \gtrsim \Delta^\top H\Delta \gtrsim \lambda_L\|\Delta\|_2^2, \tag{48}$$

because of the optimality condition. This lower bounds $\nabla f(\theta)^\top(\theta - \widehat{\theta})$ when $\|\Delta\|_2 = O(\frac{\lambda_L}{E})$. Next we will lower bound it when $\|\Delta\|_2 = \Omega(\frac{\lambda_L}{E})$.

Consider the function for $t \in [0, \infty)$, we have

$$\begin{aligned} g(t) &= \nabla f(\widehat{\theta} + ut)^\top ut \\ &= (k(\widehat{\theta} + ut) + c)\nabla k(\widehat{\theta} + ut)^\top ut \\ &= k(\widehat{\theta} + ut)\nabla k(\widehat{\theta} + ut)^\top ut + c\nabla k(\widehat{\theta} + ut)^\top ut, \end{aligned} \tag{49}$$

where $u = \frac{\Delta}{\|\Delta\|_2}$.

Because $k(\theta)$ is convex, $\nabla k(\widehat{\theta} + ut)^\top u$ is an increasing function in $t$, thus we have $\nabla k(\widehat{\theta} + ut)^\top u = \Omega(\frac{\lambda_L^2}{E})$ when $t = \Omega(\frac{\lambda_L}{E})$. And we can deduce $\nabla k(\widehat{\theta} + ut)^\top ut = \Omega(\frac{\lambda_L^2}{E}t)$ when $t = \Omega(\frac{\lambda_L}{E})$.

Similarly, because $k(\theta)$ is convex, $k(\widehat{\theta} + ut)$ is an increasing function in $t$. Its derivative $\nabla k(\widehat{\theta} + ut)^\top u = \Omega(\frac{\lambda_L^2}{E})$ when $t = \Omega(\frac{\lambda_L}{E})$. So we have $k(\widehat{\theta} + ut) = \Omega(\frac{\lambda_L^2}{E}t)$ when $t = \Omega(\frac{\lambda_L}{E})$.

Thus, we have

$$k(\widehat{\theta} + ut)\nabla k(\widehat{\theta} + ut)^\top ut = \Omega(\frac{\lambda_L^4}{E^2}t^2), \tag{50}$$

when $t = \Omega(\frac{E}{\lambda_L})$.

And we can conclude that $\nabla f(\theta)^\top(\theta - \widehat{\theta}) \geq \alpha\|\theta - \widehat{\theta}\|_2^2$ for some data dependent constant $\alpha = \Omega(\min\{\lambda_L, \frac{\lambda_L^4}{E^2}\})$. □

Next, we will prove properties about $g_s = \Psi_s \Upsilon_s$.

$$\begin{aligned}
&\mathbb{E}[\|\Upsilon\|_2^2 \mid \theta] = \frac{1}{S_\Upsilon}(\frac{1}{n}\sum \|\nabla k_i(\theta)\|_2^2 - \|\nabla k(\theta)\|_2^2) + \|\nabla k(\theta)\|_2^2 \\
\lesssim &\frac{1}{n}\|x_i\|_2^2
\end{aligned} \tag{51}$$

$$\begin{aligned}
&\mathbb{E}[\Psi_s^2] \\
\overset{(i)}{\leq} &\frac{1}{n}\sum (c + k_i(\theta))^2 \\
= &\frac{1}{n}\sum (c + \log(1 + \exp(-y_i \widehat{\theta}^\top x_i - y_i \Delta x_i)))^2 \\
\overset{(ii)}{\lesssim} &\frac{1}{n}\sum \|x_i\|^2 \|\Delta\|_2^2 + \frac{1}{n}\sum (c + \log(1 + \exp(-y_i \widehat{\theta}^\top x_i)))^2,
\end{aligned} \tag{52}$$

where (i) follows from $\mathbb{E}[(\frac{\sum_{j=1}^S X_j}{S})^2] \leq \mathbb{E}[\frac{\sum_{j=1}^S X_j^2}{S}]$ and (ii) follows from $\log(1 + \exp(a+b)) \leq \log(1+e^b) + |a|$.

Thus we have

$$\begin{aligned}
&\mathbb{E}[\|g_s\|_2^2(\theta) \mid \theta] \\
= &\mathbb{E}[\Psi^2 \mid \theta]\mathbb{E}[\|\Upsilon\|_2^2 \mid \theta] \\
\lesssim &A\|\Delta\|_2^2 + B
\end{aligned} \tag{53}$$

for some data dependent constants $A$ and $B$.

$$\begin{aligned}
&\mathbb{E}[\|\Upsilon\|_2^4 \mid \theta] \\
= &\mathbb{E}[\|\frac{1}{S_\Upsilon}\sum_{i \in I_t^\Upsilon} \nabla \log(1 + \exp(-y_i \theta^\top x_i))\|_2^4] \\
\leq &\mathbb{E}[(\frac{1}{S_\Upsilon}\sum_{i \in I_t^\Upsilon} \|\nabla \log(1 + \exp(-y_i \theta^\top x_i))\|_2)^4] \\
\leq &\mathbb{E}[(\frac{1}{S_\Upsilon}\sum_{i \in I_t^\Upsilon} \|x_i\|_2)^4] \\
\leq &\frac{1}{n}\sum \|x_i\|_2^4.
\end{aligned} \tag{54}$$

$$\begin{aligned}
&\mathbb{E}[\Psi_s^4] \\
\overset{(i)}{\leq} &\frac{1}{n}\sum (c + k_i(\theta))^4 \\
= &\frac{1}{n}\sum (c + \log(1 + \exp(-y_i \widehat{\theta}^\top x_i - y_i \Delta x_i)))^4 \\
\overset{(ii)}{\lesssim} &\frac{1}{n}\sum \|x_i\|^4 \|\Delta\|_2^4 + \frac{1}{n}\sum (c + \log(1 + \exp(-y_i \widehat{\theta}^\top x_i)))^4,
\end{aligned} \tag{55}$$

where (i) follows from $\mathbb{E}[(\frac{\sum_{j=1}^{S} X_j}{S})^4] \leq \mathbb{E}[\frac{\sum_{j=1}^{S} X_j^4}{S}]$ and (ii) follows from $\log(1+\exp(a+b)) \leq \log(1+e^b)+|a|$.

Thus we have

$$\begin{aligned}&\mathbb{E}[\|g_s\|_2^4(\theta) \mid \theta]\\ =&\mathbb{E}[\Psi^4 \mid \theta]\mathbb{E}[\|\Upsilon\|_2^4 \mid \theta]\\ \lesssim& C\|\Delta\|_2^4 + D,\end{aligned} \tag{56}$$

for some data dependent constants $C$ and $D$.

$$\begin{aligned}&\|\mathbb{E}[\nabla g_s(\theta)\nabla g_s(\theta)^\top] - G\|_2\\ \leq&\|K_G(\theta)\frac{1}{n}\sum \nabla k_i(\theta)\nabla k_i(\theta)^\top - K_G(\widehat{\theta})\frac{1}{n}\sum \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top\|_2\\ \leq&\|K_G(\theta)\frac{1}{n}\sum \nabla k_i(\theta)\nabla k_i(\theta)^\top - K_G(\theta)\frac{1}{n}\sum \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top + K_G(\theta)\frac{1}{n}\sum \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top - K_G(\widehat{\theta})\frac{1}{n}\sum \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top\|_2\\ \leq&K_G(\theta)\frac{1}{n}\|\sum(\nabla k_i(\theta)\nabla k_i(\theta)^\top - \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top)\|_2 + |K_G(\theta) - K_G(\widehat{\theta})|\|\frac{1}{n}\sum \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top\|_2.\end{aligned} \tag{57}$$

Because

$$K_G(\theta) = O(1+\|\Delta\|_2+\|\Delta\|_2^2), \tag{58}$$

$$\frac{1}{n}\|\sum(\nabla k_i(\theta)\nabla k_i(\theta)^\top - \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top)\|_2 = O(\|\Delta\|_2+\|\Delta\|_2^2), \tag{59}$$

$$|K_G(\theta) - K_G(\widehat{\theta})| = O(\|\Delta\|_2+\|\Delta\|_2^2), \tag{60}$$

where we have data dependent constants.

Then, we have

$$\|\mathbb{E}[g_s(\theta)g_s(\theta)^\top \mid \theta] - G\|_2 \leq A_1\|\theta-\widehat{\theta}\|_2 + A_2\|\theta-\widehat{\theta}\|_2^2 + A_3\|\theta-\widehat{\theta}\|_2^3 + A_4\|\theta-\widehat{\theta}\|_2^4, \tag{61}$$

for some data dependent constants $A_1$, $A_2$, $A_3$, and $A_4$.

Combining above results and using Theorem 1, we have

$$\begin{aligned}&\|t\mathbb{E}[(\bar{\theta}_t-\widehat{\theta})(\bar{\theta}_t-\widehat{\theta})^\top] - H^{-1}GH^{-1}\|_2\\ \lesssim&\sqrt{\eta} + \sqrt{\frac{1}{t\eta}+t\eta^2}.\end{aligned} \tag{62}$$

□

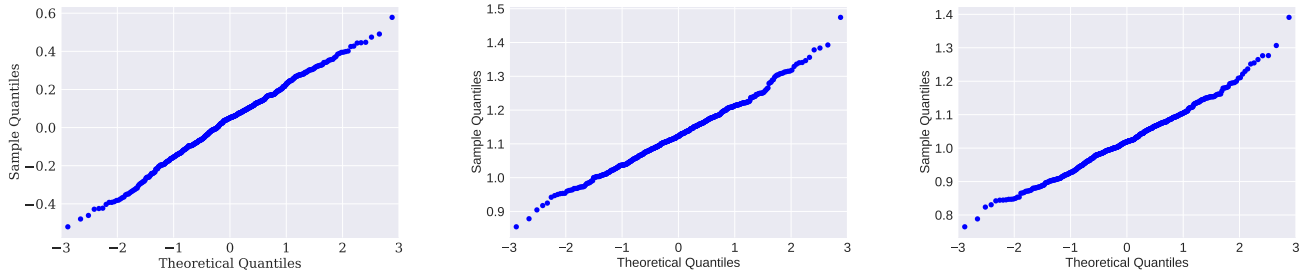
# B Experiments

Here we present additional experiments on our SGD inference procedure.

## B.1 Synthetic data

### B.1.1 Univariate models

Figure 8 shows Q-Q plots for samples shown in Figure 2.




(a) Normal. (b) Exponential. (c) Poisson.

Figure 8: Estimation in univariate models: Q-Q plots for samples shown in Figure 2

### B.1.2 Multivariate models

Here we show Q-Q plots per coordinate for samples from our SGD inference procedure.

Q-Q plots per coordinate for samples in linear regression experiment 1 is shown in Figure 9. Q-Q plots per coordinate for samples in linear regression experiment 2 is shown in Figure 10.

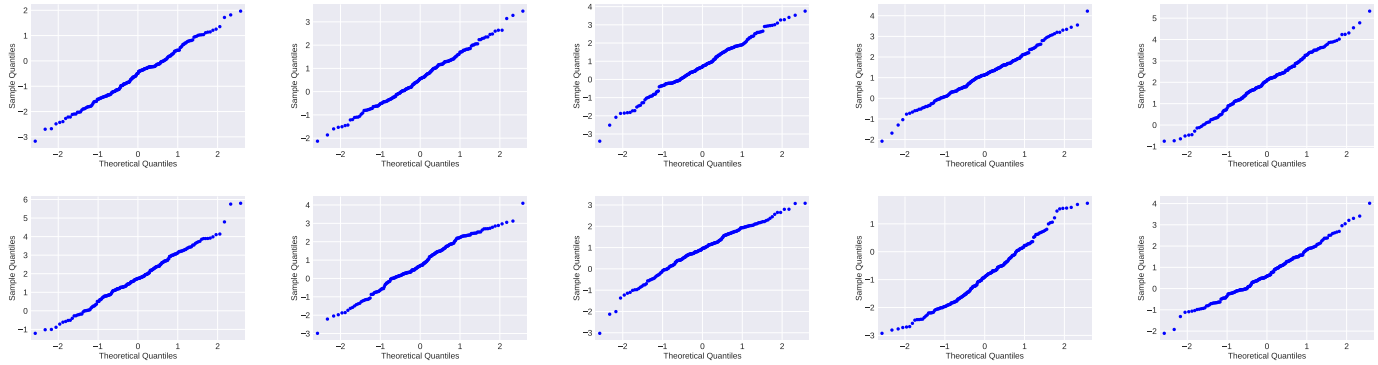



Figure 9: Linear regression experiment 1: Q-Q plots per coordinate

Q-Q plots per coordinate for samples in logistic regression experiment 1 is shown in Figure 11. Q-Q plots per coordinate for samples in logistic regression experiment 2 is shown in Figure 12.

**Additional experiments**

*2-Dimensional Linear Regression.* Consider:

$$y = x_1 + x_2 + \epsilon, \qquad \text{where } \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right) \text{ and } \epsilon \sim \mathcal{N}(0, \sigma^2 = 10^2).$$

Each sample consists of $Y = y$ and $X = [x_1, \ x_2]^\top$. We use linear regression to estimate $w_1, w_2$ in $y = w_1 x_1 + w_2 x_2$. In this case, the minimizer of the population least square risk is $w_1^\star = 1, w_2^\star = 1$.

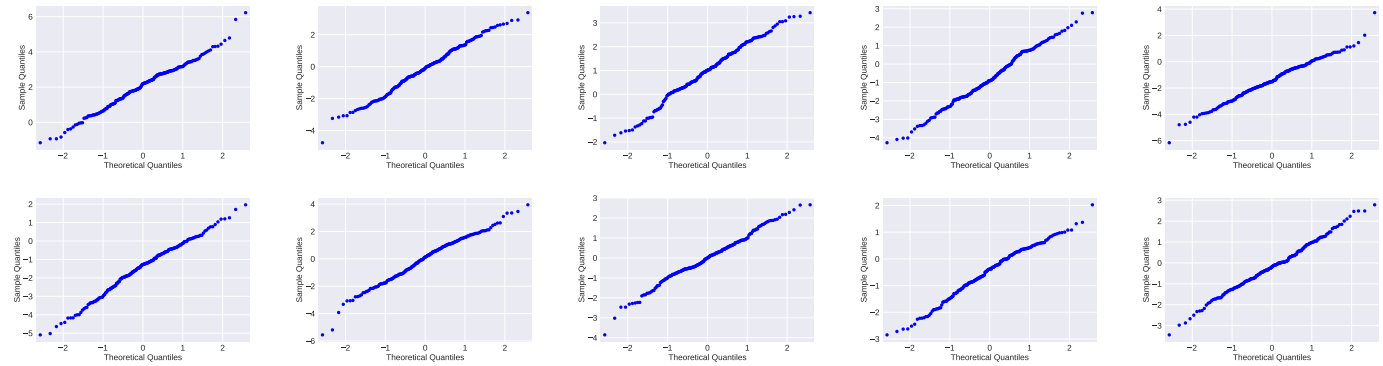



Figure 10: Linear regression experiment 2: Q-Q plots per coordinate

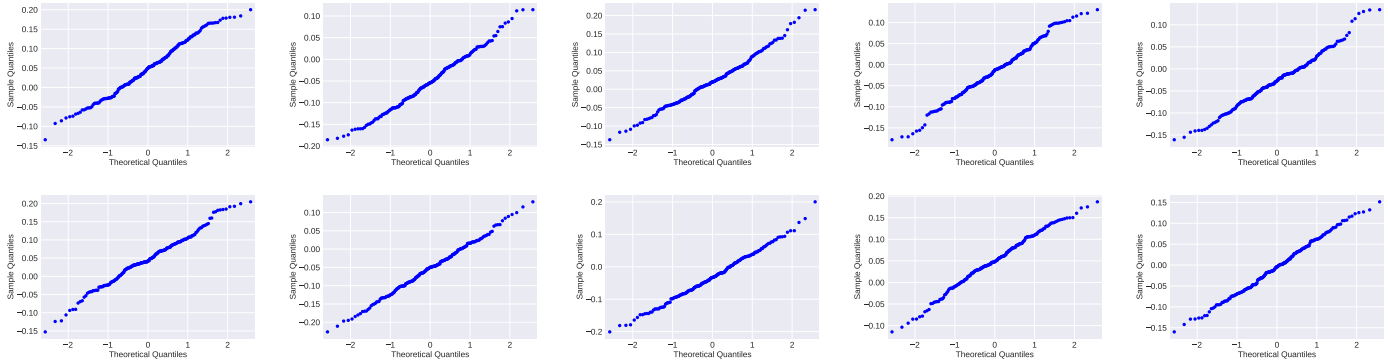



Figure 11: Logistic regression experiment 1: Q-Q plots per coordinate

For 300 i.i.d. samples, we plotted 100 samples from SGD inference in Figure 13. We compare our SGD inference procedure against bootstrap in Figure 13a. Figure 13b and Figure 13c show samples from our SGD inference procedure with different parameters.

*10-Dimensional Linear Regression.*

Here we consider the following model

$$y = x^\top w^\star + \epsilon,$$

where $w^\star = \frac{1}{\sqrt{10}}[1, \ 1, \ \cdots, \ 1]^\top \in \mathbb{R}^{10}$, $x \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{ij} = 0.8^{|i-j|}$, and $\epsilon \sim \mathcal{N}(0, \sigma^2 = 20^2)$, and use $n = 1000$ samples. We estimate the parameter using

$$\widehat{w} = \operatorname*{argmin}_w \frac{1}{n} \sum_{i=1}^{n} \tfrac{1}{2}(x_i^\top w - y_i)^2.$$

Figure 14 shows the diagonal terms of of the covariance matrix computed using the sandwich estimator and our SGD inference procedure with different parameters. 100000 samples from our SGD inference procedure are used to reduce the effect of randomness.

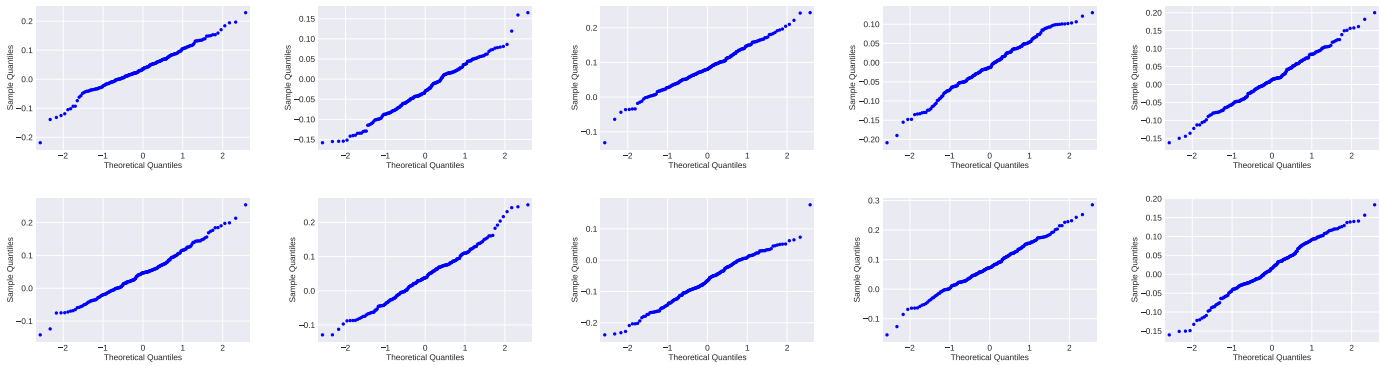



Figure 12: Logistic regression experiment 2: Q-Q plots per coordinate

*2-Dimensional Logistic Regression.*
Here we consider the following model

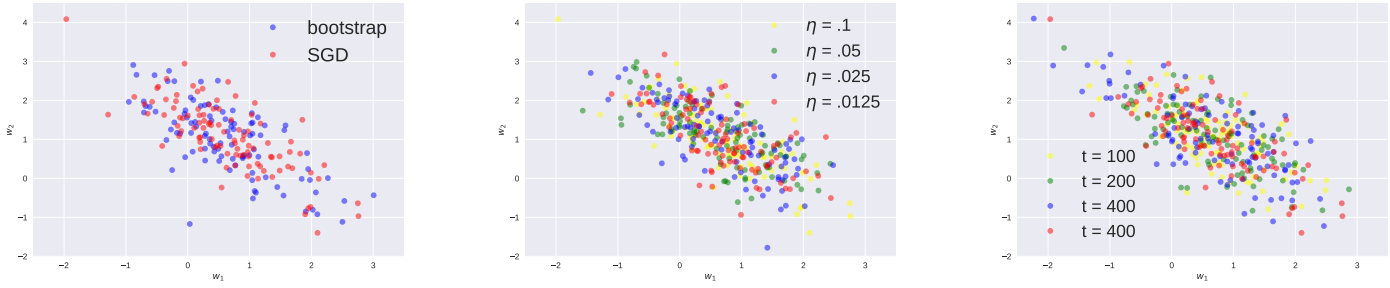
$$\mathbb{P}[Y=+1]=\mathbb{P}[Y=-1]=\frac{1}{2}, \qquad X \mid Y \sim \mathcal{N}\left(\mu=1.1+0.1Y,\ \sigma^2=1\right). \tag{63}$$

We use logistic regression to estimate $w, b$ in the classifier $\operatorname{sign}(wx+b)$ where the minimizer of the population logistic risk is $w^\star = 0.2, b^\star = -0.22$.

For 100 i.i.d. samples, we plot 1000 samples from SGD in Figure 15. In our simulations, we notice that our modified SGD for logistic regression behaves similar to vanilla logistic regression. T his suggests that an assumption weaker than $(\theta-\widehat{\theta})^\top \nabla f(\theta) \geq \alpha\|\theta-\widehat{\theta}\|_2^2$ (assumption $(F_1)$ in Theorem 1) is sufficient for SGD analysis. Figure 15b and Figure 15d suggest that the $t\eta^2$ term in Corollary 2 is an artifact of our analysis, and can be improved.

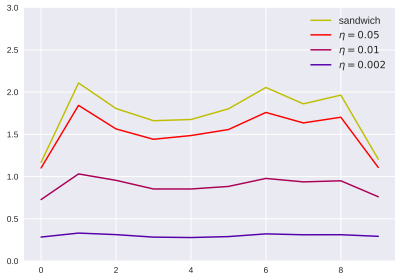
*11-Dimensional Logistic Regression.*
Here we consider the following model

$$\mathbb{P}[Y=+1]=\mathbb{P}[Y=-1]=\frac{1}{2}, \qquad X \mid Y \sim \mathcal{N}\left(0.01Y\mu, \Sigma\right),$$

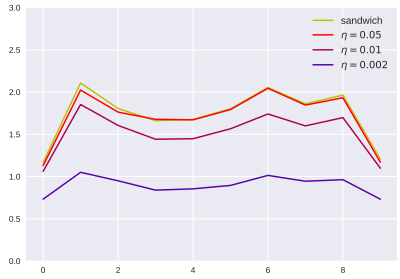



(a) SGD inference vs. bootstrap (b) $t = 800$ (c) $\eta = 0.1$

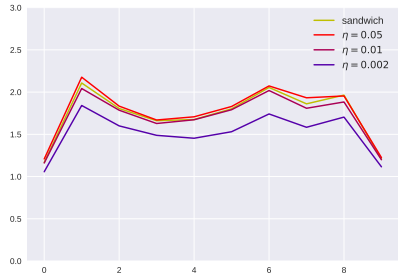
Figure 13: 2-dimensional linear regression



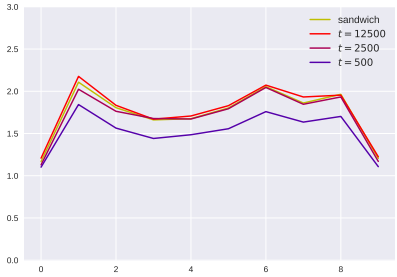

(a) $t = 500$



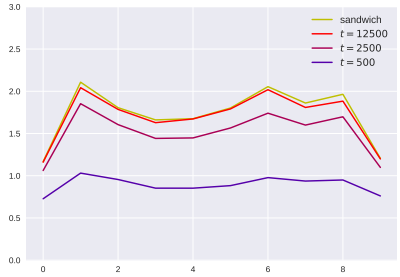

(b) $t = 2500$



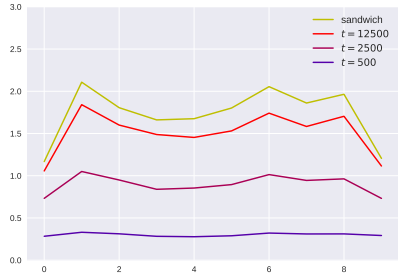

(c) $t = 12500$




(d) $\eta = 0.05$

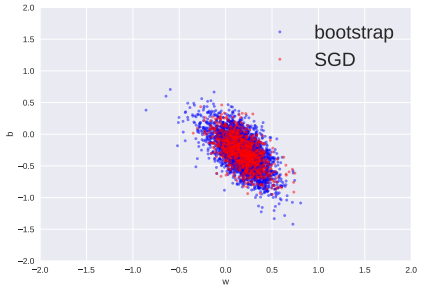



(e) $\eta = 0.01$

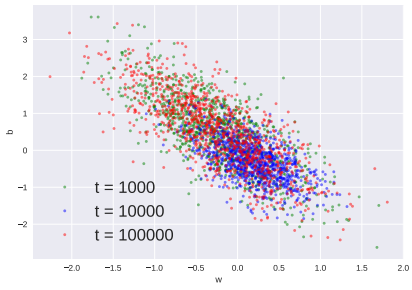



(f) $\eta = 0.002$

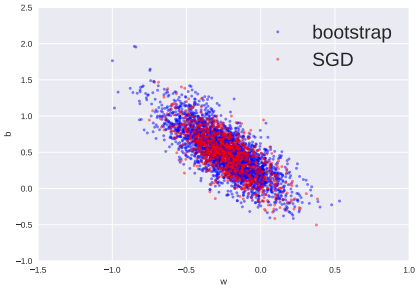
Figure 14: 11-dimensional linear regression: covariance matrix diagonal terms of SGD inference and sandwich estimator



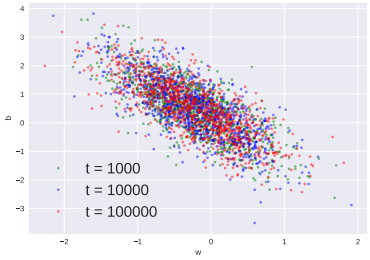

(a) Modified SGD with $t = 1000$ and $\eta = 0.1$



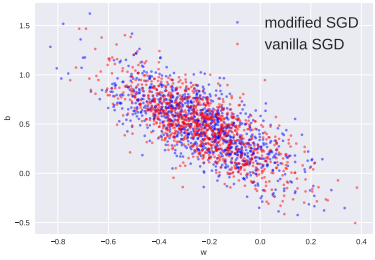

(b) Modified SGD with $\eta = 0.1$




(c) Vanilla SGD with $t = 1000$ and $\eta = 0.1$




(d) Vanilla SGD with $\eta = 0.1$




(e) $t = 1000$ and $\eta = 0.1$

Figure 15: 2-dimensional logistic regression

where $\Sigma_{ii} = 1$ and when $i \neq j$ $\Sigma_{ij} = \rho^{|i-j|}$ for some $\rho \in [0, 1)$, and $\mu = \frac{1}{\sqrt{10}}[1, \ 1, \ \cdots, \ 1]^\top \in \mathbb{R}^{10}$. We estimate a classifier $\text{sign}(w^\top x + b)$ using

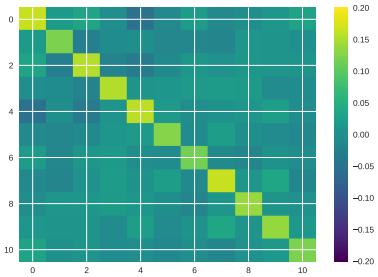
$$\widehat{w}, \widehat{b} = \underset{w,b}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp(-Y_i(w^\top X_i + b))\right). \tag{64}$$

Figure 16 shows results for $\rho = 0$ with $n = 80$ samples. We use $t = 100$, $d = 70$, $\eta = 0.8$, and mini batch of size 4 in vanilla SGD. Bootstrap and our SGD inference procedure each generated 2000 samples. In bootstrap, we used Newton method to perform optimization over each replicate, and 6-7 iterations were used. In figure 17, we follow the same procedure for $\rho = 0.6$ with $n = 80$ samples. Here, we use $t = 200$, $d = 70$, $\eta = 0.85$; the rest of the setting is the same.

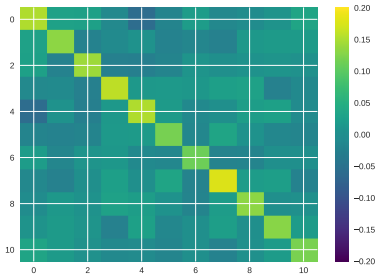
## B.2 Real Data

### B.2.1 MNIST

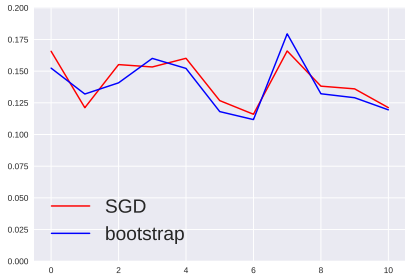
Here, we train a binary logistic regression classifier to classify $0/1$ using perturbed MNIST data set, and demonstrate that certain adversarial examples (e.g. [10]) can be detected using prediction confidence intervals. For each image, where each original pixel is either 0 or 1, we randomly changed 70% pixels to random numbers uniformly on $[0, 0.9]$. Figure 18 shows



(a) SGD covariance

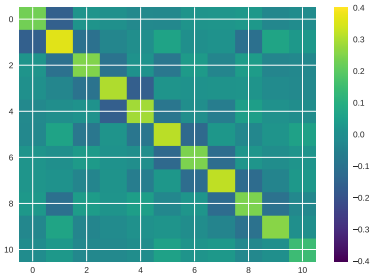


(b) Bootstrap estimated covariance

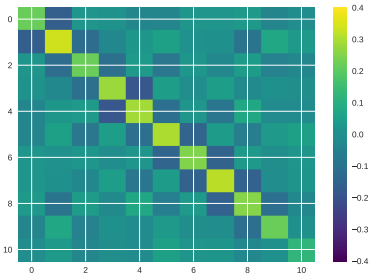



(c) Diagonal terms

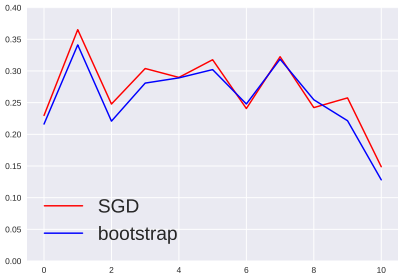
Figure 16: 11-dimensional logistic regression: $\rho = 0$



(a) SGD covariance



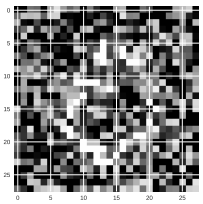
(b) Bootstrap estimated covariance



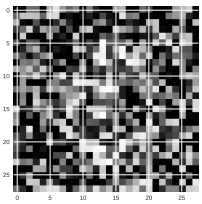

(c) Diagonal terms

Figure 17: 11-dimensional logistic regression: $\rho = 0.6$

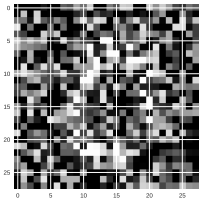
each image's logit value ($\log \frac{\mathbb{P}[1|\text{image}]}{\mathbb{P}[0|\text{image}]}$) and its 95% confidence interval computed using our SGD inference procedure. The adversarial perturbation used here is shown in Figure 19 (scaled for display).



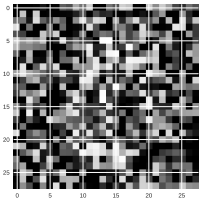
(a) Original "0": logit -72.6, CI (-106.4, -30.0)



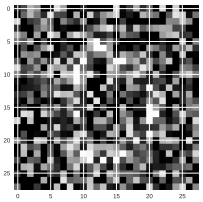
(b) Adversarial "0": logit 15.3, CI (-6.5, 26.2)



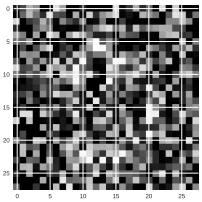
(c) Original "0": logit -62.1, CI (-101.6, -5.5)



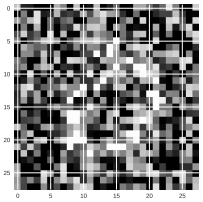
(d) Adversarial "0": logit 1.9, CI (-4.9, 11.6)



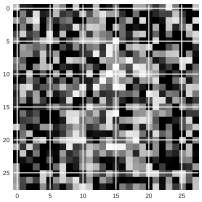
(e) Original "0": logit -42.9, CI (-75.4, 5.1)



(f) Adversarial "0": logit 4.8, CI (-3.4, 17.9)



(g) Original "0": logit -77.0, CI (-110.7, -32.2)



(h) Adversarial "0": logit 13.3, CI (-8.0, 25.7)

Figure 18: MNIST

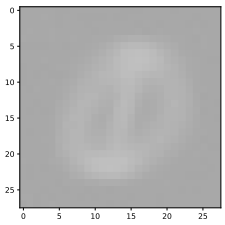



Figure 19: MNIST adversarial perturbation (scaled for display)