

Copyright
by
Tianyang Li
2019

The Dissertation Committee for Tianyang Li
certifies that this is the approved version of the following dissertation:

Stochastic gradients methods for statistical inference

Committee:

Constantine Caramanis, Supervisor

Alexandros Dimakis

Qixing Huang

Anastasios Kyrillidis

Stochastic gradients methods for statistical inference

by

Tianyang Li

DISSERTATION

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2019

Dedicated to my parents Zihui and Bo.

Acknowledgments

First, I would like to thank my advisor, Constantine Caramanis, for his support and encouragements. Constantine has always been a resource for providing me with advice and feedback about my research, and served as a guide throughout my Ph.D. studies. I would also like to thank Anastasios Kyrillidis, for insightful discussions. Apart from Constantine and Anastasios, I would like to thank Pradeep Ravikumar for his mentorship.

In addition, many thanks to my collaborators: Dhruv Balwada, Liu Liu, Harsh Pareek, Adarsh Prasad, Yanyao Shen, Kevin Speer, and Xinyang Yi.

Last, but the most important of all, I would like to thank my parents for their support.

Stochastic gradients methods for statistical inference

Tianyang Li, Ph.D.

The University of Texas at Austin, 2019

Supervisor: Constantine Caramanis

Statistical inference, such as hypothesis testing and calculating a confidence interval, is an important tool for accessing uncertainty in machine learning and statistical problems. Stochastic gradient methods, such as stochastic gradient descent (SGD), have recently been successfully applied to point estimation in large scale machine learning problems. In this work, we present novel stochastic gradient methods for statistical inference in large scale machine learning problems.

Unregularized M -estimation using SGD. Using 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.

Approximate Newton-based statistical inference using only stochastic gradients for unregularized M -estimation. 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 for unregularized convex learning problems, 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.

High dimensional linear regression statistical inference using only stochastic gradients. As an extension of the approximate Newton-based statistical inference algorithm for unregularized problems, we present a similar algorithm, using only stochastic gradients, for statistical inference in high dimensional linear regression, where the number of features is much larger than the number of samples.

Stochastic gradient methods for time series analysis. We present a novel stochastic gradient descent algorithm for time series analysis, which correctly captures correlation structures in a time series dataset during optimization. Instead of uniformly sampling indices in vanilla SGD, we uniformly sample contiguous blocks of indices, where the block length depends on the dataset.

Table of Contents

Acknowledgments	v
Abstract	vi
List of Tables	xii
List of Figures	xiii
List of Algorithms	xv
Chapter 1. Introduction	1
1.1 Related work	3
1.1.1 Connection with Bootstrap methods	3
1.1.2 Other stochastic gradient methods for frequentist inference	4
1.1.3 Stochastic gradient methods for Bayesian inference	4
1.1.4 Related optimization algorithms	6
1.1.4.1 Connection with stochastic approximation methods	6
1.1.4.2 Connections to stochastic Newton methods	7
1.1.5 Statistical inference in high dimensional linear regression	7
Chapter 2. Statistical inference in M-estimation	9
Chapter 3. Statistical inference using SGD	11
3.1 Statistical inference using SGD	11
3.1.1 Theoretical guarantees	13
3.1.2 Intuitive interpretation via the Ornstein-Uhlenbeck process approximation	15
3.1.3 Logistic regression	16
3.2 Experiments	19
3.2.1 Synthetic data	19

3.2.1.1	Univariate models	19
3.2.1.2	Multivariate models	20
3.2.2	Real data	21
3.2.2.1	Splice data set	21
3.2.2.2	MNIST	23
3.2.3	Discussion	23
3.3	Linear Regression	24
3.4	Exact analysis of mean estimation	25
Chapter 4. Approximate Newton-based statistical inference using only stochastic gradients		26
4.1	Statistical inference with approximate Newton steps using only stochastic gradients	27
4.2	Experiments	31
4.2.1	Synthetic data	31
4.2.2	Real data	32
4.3	Statistical inference via approximate stochastic Newton steps using first order information with increasing inner loop counts	34
4.3.1	Unregularized M-estimation	34
4.4	SVRG based statistical inference algorithm in unregularized M-estimation	34
4.4.1	Lack of asymptotic normality in Algorithm 1 for mean estimation	36
Chapter 5. High dimensional linear regression		38
5.1	Experiments	41
5.1.1	Synthetic data	41
5.1.2	Real data	42
5.2	Statistical inference using approximate proximal Newton steps with stochastic gradients	44
5.3	Computing the de-biased estimator (5.4) via SVRG	46
5.4	Solving the high dimensional linear regression optimization objective (5.1) using proximal SVRG	48
5.5	Non-asymptotic covariance estimate bound and asymptotic normality in Algorithm 3	48
5.6	Plug-in statistical error covariance estimate	50

Chapter 6. Time series analysis	51
6.1 Time series analysis	51
6.2 Time series statistical inference with approximate Newton steps using only stochastic gradients (Section 6.1)	53
6.3 Experiments	55
6.3.1 Synthetic data	55
6.3.2 Real data	55
Chapter 7. An intuitive view of SVRG as approximate stochastic Newton descent	57
Chapter 8. Chapter 3 proofs	59
8.1 Proof of Theorem 1	59
8.2 Proof of Corollary 2	69
8.3 Proof of Corollary 1	70
Chapter 9. Chapter 4 proofs	76
9.1 Proof of Theorem 2	76
9.1.1 Proof of (4.8)	77
9.1.2 Proof of (4.9)	85
9.1.3 Proof of (4.6)	88
9.1.4 Proof of (4.7)	91
9.1.5 Proof of (4.10)	92
9.2 Proof of Corollary 3	95
9.3 Proof of Corollary 5	98
Chapter 10. Chapter 5 proofs	100
10.1 Proof of Theorem 3	100
10.1.1 Proof of Theorem 4	104
10.1.2 Proof of Theorem 5	107
10.1.3 Proof of Lemma 1	111
10.2 Technical lemmas	113
10.2.1 Lemma 10	113
10.2.2 Lemma 11	114

Appendix	117
Appendix 1. Chapter 3 appendix	118
1.1 Experiments	118
1.1.1 Synthetic data	118
1.1.1.1 Multivariate models	118
1.1.2 Real data	125
1.1.2.1 Higgs data set	125
1.1.2.2 Splice data set	127
1.1.2.3 MNIST	128
Appendix 2. Chapter 4 appendix	131
2.1 Experiments	131
2.1.1 Synthetic data	131
2.1.1.1 Low dimensional problems	131
2.1.2 Real data	132
2.1.2.1 Neural network adversarial attack detection	132
Appendix 3. Chapter 5 appendix	134
3.1 Experiments	134
3.1.1 Synthetic data	134
3.1.1.1 High dimensional linear regression	134
3.1.2 Real data	135
3.1.2.1 High dimensional linear regression	135
Appendix 4. Chapter 6 appendix	137
4.1 Experiments	137
4.1.1 Synthetic data	137
4.1.2 Real data	138
Bibliography	139

List of Tables

- 3.1 Linear regression. *Left: Experiment 1, Right: Experiment 2.* 18
- 3.2 Logistic regression. *Left: Experiment 1, Right: Experiment 2.* 19

- 4.1 Synthetic data average coverage & confidence interval length for low dimensional problems. 31
- 4.2 Average 95% confidence interval (coverage, length) after calibration 32

- 5.1 HIV drug resistance related mutations detected by our high dimensional inference procedure 44

List of Figures

3.1	Our SGD inference procedure	12
3.2	Estimation in univariate models.	18
3.3	Splice data set	22
3.4	MNIST	22
4.1	Distribution of scores for original, randomly perturbed, and adversarially perturbed images	33
5.1	95% confidence intervals	42
5.2	Distribution of two-sided Z-test p-values under the null hypothesis (high dimensional)	43
6.1	Exposure of US equities market to equities markets of other countries	56
1.1	Estimation in univariate models: Q-Q plots for samples shown in Figure 3.2	118
1.2	Linear regression experiment 1: Q-Q plots per coordinate	119
1.3	Linear regression experiment 2: Q-Q plots per coordinate	119
1.4	Logistic regression experiment 1: Q-Q plots per coordinate	120
1.5	Logistic regression experiment 2: Q-Q plots per coordinate	120
1.6	2-dimensional linear regression	121

1.7	11-dimensional linear regression: covariance matrix diagonal terms of SGD inference and sandwich estimator	122
1.8	2-dimensional logistic regression	123
1.9	11-dimensional logistic regression: $\rho = 0$	124
1.10	11-dimensional logistic regression: $\rho = 0.6$	124
1.11	Higgs data set with $n = 200$	126
1.12	Higgs data set with $n = 50000$	126
1.13	Higgs data set with $n = 90000$	127
1.14	Splice data set	127
1.15	MNIST	129
1.16	MNIST adversarial perturbation (scaled for display)	130
2.1	“Shirt” example	133
2.2	“T-shirt/top” example	133
2.3	Adversarial perturbation generated using the fast gradient sign method [GSS14]	133
3.1	Comparison of our de-biased estimator and oracle de-biased LASSO estimator	135
3.2	Comparison of our high dimensional linear regression point estimate with the vanilla LASSO estimate	136

List of Algorithms

1	Unregularized M-estimation statistical inference	28
2	SVRG based statistical inference algorithm in unregularized M-estimation . .	35
3	High dimensional linear regression statistical inference	47
4	Computing the de-biased estimator (5.4) via SVRG	48
5	Solving the high dimensional linear regression optimization objective (5.1) using proximal SVRG	49
6	Unregularized M-estimation statistical inference	53

Chapter 1

Introduction

Statistical inference, such as hypothesis testing and calculating a confidence interval, is an important tool for accessing uncertainty in machine learning and statistical problems, both for estimation and prediction purposes [FHT01, EH16]. *E.g.*, in unregularized linear regression and high-dimensional LASSO settings [vdGBRD14, JM15, TWH15], 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 [Was13]. Traditionally, the inverse Fisher information matrix [Edg08] 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 [TRVB06]. Alternatively, the Bootstrap [Efr82, ET94] method is a popular statistical inference algorithm, where we solve an optimization problem per dataset replicate, but can be expensive for large data sets [KTSJ14].

Stochastic gradient methods, such as stochastic gradient descent (SGD) [RM85, Bub15a, Bot10], have been recently been successfully applied to point estimation in large scale machine learning problems. For example, in deep learning [GBC16], stochastic gradient methods such as SGD and Adam [KB14] are widely used to train neural nets.

In this context, we follow a different path: we show that inference can also be accomplished by directly using stochastic gradient methods, such as SGD, both for point estimates and inference. While optimization is mostly used for point estimates, recently it

This chapter also appears in [LKLC18, LLKC18]. It was written by Tianyang Li, and edited by Anastasios Kyrillidis and Constantine Caramanis.

is also used as a means for statistical inference in large scale machine learning [LKLC18, LLKC18, CLTZ16, SZ18, FXY17].

In Chapter 3, we present a statistical inference procedure using SGD with fixed step size [LLKC18]. 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 empirical risk minimization (ERM), 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.

In Chapter 4, we present a framework for approximate Newton-based statistical inference using only stochastic gradients [LKLC18]. 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 ∇f , $\nabla^2 f$ denote the gradient and Hessian of f . Our method can be interpreted as a generalization of stochastic variance reduced gradient (SVRG) [JZ13] in optimization [JZ13] (Chapter 7); further, it is related to other stochastic Newton methods (e.g. [ABH17]) when $\delta \rightarrow 0$.

As an extension of the approximate Newton-based statistical inference procedure using stochastic gradients, in Chapter 5 we present a novel statistical inference procedure for high dimensional linear regression using stochastic gradients, where the number of features is much larger than the number of samples. The intuition behind our algorithm is that each proximal Newton descent step [LSS14] can be solved using proximal SVRG [XZ14].

In Chapter 6, we present a novel stochastic gradient framework for time series analysis, which correctly captures dependence relationships in a time series dataset. Unlike vanilla

SGD where we sample indices uniformly over the entire dataset, we sample contiguous blocks of indices, where the data-dependent block length is the *lag*. This enables our stochastic gradient procedure to compute a covariance estimate similar to the Driscoll-Kraay method [DK98, Hoe07]. The sampling scheme in our procedure is similar to that of moving block bootstrap [Lah13], and similar sampling schemes in conformal prediction for time series analysis [BHV14], which also use contiguous blocks chosen from the dataset.

1.1 Related work

1.1.1 Connection with Bootstrap methods

The classical approach for statistical inference is to use the bootstrap [ET94, ST12]. 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.

In bootstrap, given a dataset with n samples, each time we resample n times with replacement from the dataset, and compute an estimate (*replicate*) on this resampled dataset. We then perform statistical inference, such as hypothesis testing or computing confidence intervals, using the empirical distribution of bootstrap replicates.

In jackknife, we generate n datasets, where each dataset has $n - 1$ elements, by leaving out one element each time. We then use the variance of jackknife replicates in asymptotic normality to perform statistical inference, such as hypothesis testing or computing confidence intervals.

1.1.2 Other stochastic gradient methods for frequentist inference

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 [LLKC18, CLTZ16].

Compared to [CLTZ16] (and similar works [SZ18, FXY17] 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, [CLTZ16] guarantees $O(L^{-\frac{1-d_o}{2}} + T^{-\frac{1}{2}} + T^{\max\{\frac{1}{2}-\frac{d_o}{4(1-d_o)}, 0\}} \cdot L^{-\frac{d_o}{4}} + T^{\max\{\frac{1-2d_o}{2(1-d_o)}, 0\}} \cdot L^{\frac{1-2d_o}{2}})$, whereas our method guarantees $O(T^{-\frac{d_o}{2}})$. Further, [CLTZ16] is inconsistent, whereas our scheme guarantees consistency of computing the statistical error covariance.

Chapter 3 [LLKC18] 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 [LLKC18] states empirically constant step SGD performs well in statistical inference, it has been empirically shown [DDB17] 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 ([PJ92, Rup88] and our work) will converge to the optimal, and averaging consecutive iterates guarantees “fast” rates.

1.1.3 Stochastic gradient methods for 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

[WT11, BEL15, MHB16, MHB17]. We note that, here as well, the vast majority of the results rely on using a decreasing step size. Very recently, [MHB17] 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 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. Another key difference between our methods and SGLD methods, is that we use averages of consecutive iterates, whereas SGLD does not use averaging.

SGLD can be viewed as a discretization of the following stochastic differential equation in d -dimensional space

$$d\mathbf{z} = \mathbf{f}(\mathbf{z})dt + \sqrt{2\mathbf{D}(\mathbf{z})}d\mathbf{W}(t),$$

where $\mathbf{f}(z)$ is a deterministic drift, $\mathbf{W}(t)$ is a standard Brownian motion process, and $\mathbf{D}(\mathbf{z})$ is a positive semidefinite diffusion matrix. [MCF15] shows that its stationary distribution $\exp(-H(\mathbf{z}))$, when

$$\mathbf{f}(\mathbf{z}) = -[\mathbf{D}(\mathbf{z}) + \mathbf{Q}(\mathbf{z})]\nabla H(\mathbf{z}) + \Gamma(\mathbf{z}), \quad \Gamma_i(\mathbf{z}) = \sum_{j=1}^d \frac{\partial}{\partial \mathbf{z}_j} [\mathbf{D}_{ij}(\mathbf{z}) + \mathbf{Q}_{ij}(\mathbf{z})],$$

where $\mathbf{Q}(\mathbf{z})$ is a skew symmetric curl matrix.

1.1.4 Related optimization algorithms

1.1.4.1 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., [LPW12, PJ92], and more recently, [TA14, CLTZ16]; and also for *fixed step size*, e.g., [BPM90], 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, [BPM90] provides results only when the (constant) step-size approaches 0 (see Section 4.4 and 4.6, and in particular Theorem 7 in [BPM90]). Similarly, while [KY03] 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 [TA14] and [CLTZ16], which follow up on [PJ92]. Both of these rely on decreasing step size, for reasons mentioned above. The work in [CLTZ16] uses SGD with decreasing step size for estimating an M -estimate’s covariance. Work in [TA14] studies implicit SGD with decreasing step size and proves results similar to [PJ92], 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).

1.1.4.2 Connections to stochastic Newton methods

Our method is similar to stochastic Newton methods (e.g. [ABH17]); 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 [Ama98]. Also, we demonstrate an intuitive view of SVRG [JZ13] as a special case of approximate stochastic Newton steps using first order information (Chapter 7).

1.1.5 Statistical inference in high dimensional linear regression

[CLTZ16]’s high dimensional inference algorithm is based on [ANW12], 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 [XZ14, LSS14] towards the optimum, which guarantees the optimization error to be much smaller than the statistical error.

Our method of de-biasing the LASSO Chapter 5 is similar to [ZZ14, vdGBRD14, JM14, JM15]. Our method uses a new ℓ_1 regularized objective for high dimensional linear regression, and we have different de-biasing terms, because we also need to de-bias the covariance estimation. In Algorithm 3, our covariance estimate is similar to the classic *sandwich estimator* [Hub67, Whi80]. Previous methods require $O(p^2)$ space which is unsuitable for large scale problems, whereas our method only requires $O(p)$ space. Similar to our ℓ_1 -norm regularized objective, [YLR14, JD11] shows similar point estimate statistical guarantees for related estimators; however there are no confidence interval results. Further, although [YLR14] 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 be 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. [FGLS18, Wai09, BvdG11].

Time series analysis. Our approach of sampling contiguous blocks of indices to compute stochastic gradients for statistical inference in time series analysis is similar to resampling procedures in *moving block* or *circular* bootstrap [Car86, Kun89, Böh02, DH97, ET94, Lah13, PR92, PR94, KL12], and *conformal prediction* [BHV14, SV08, VGS05]. Also, our procedure is similar to Driscoll-Kraay standard errors [DK98, KD99, Hoe07], but does not waste computational resources to explicitly store entire matrices, and is suited for large scale time series analysis.

Chapter 2

Statistical inference in M -estimation

Here, we give a brief overview of statistical inference in unregularized M -estimation. Consider the problem of estimating a set of parameters $\theta^* \in \mathbb{R}^p$ 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 θ^* of the population risk:

$$\theta^* = \operatorname{argmin}_{\theta \in \mathbb{R}^p} \mathbb{E}_P[f(\theta; X)] = \operatorname{argmin}_{\theta \in \mathbb{R}^p} \int_x f(\theta; x) dP(x), \quad (2.1)$$

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

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n f(\theta; X_i). \quad (2.2)$$

Statistical inference consists of techniques for obtaining information beyond point estimates $\hat{\theta}$, such as confidence intervals. These can be performed if there is an asymptotic limiting distribution associated with $\hat{\theta}$ [Was13]. Indeed, under standard and well-understood regularity conditions, the solution to M -estimation problems satisfies asymptotic normality. That is, the distribution $\sqrt{n}(\hat{\theta} - \theta^*)$ converges weakly to a normal distribution:

$$\sqrt{n}(\hat{\theta} - \theta^*) \longrightarrow \mathcal{N}(0, H^{*-1}G^*H^{*-1}), \quad (2.3)$$

where

$$H^* = \mathbb{E}[\nabla^2 f(\theta^*; X)],$$

This chapter also appears in [LKLC18, LLKC18]. It was written by Tianyang Li, and edited by Anastasios Kyrillidis and Constantine Caramanis.

and

$$G^* = \mathbb{E}[\nabla f(\theta^*; X) \cdot \nabla f(\theta^*; X)^\top];$$

see also Theorem 5.21 in [vdV00]. We can therefore use this result, as long as we have a good estimate of the covariance matrix: $H^{*-1}G^*H^{*-1}$. The central goal of this paper is obtaining accurate estimates for $H^{*-1}G^*H^{*-1}$.

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

$$\begin{aligned} \widehat{H} &= \frac{1}{n} \sum_{i=1}^n \nabla^2 f(\widehat{\theta}; X_i) & \text{and} \\ \widehat{G} &= \frac{1}{n} \sum_{i=1}^n \nabla f(\widehat{\theta}; X_i) \nabla f(\widehat{\theta}; X_i)^\top. \end{aligned} \tag{2.4}$$

Beyond calculating¹ \widehat{H} and \widehat{G} , 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 $\widehat{H}^{-1}\widehat{G}\widehat{H}^{-1}$.

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

Chapter 3

Statistical inference using SGD

Here, we describe our procedure for statistical inference in unregularized M -estimation using SGD.

3.1 Statistical inference using SGD

Consider the optimization problem in (2.2). For instance, in maximum likelihood estimation (MLE), $f(\theta; X_i)$ is a negative log-likelihood function. For simplicity of notation, we use $f_i(\theta)$ and $f(\theta)$ for $f(\theta; X_i)$ and $\frac{1}{n} \sum_{i=1}^n f(\theta; X_i)$, respectively, for the rest of the paper.

The SGD algorithm with a fixed step size η , is given by the iteration

$$\theta_{t+1} = \theta_t - \eta g_s(\theta_t), \tag{3.1}$$

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 $g_s(\cdot) \equiv \nabla f_i(\cdot)$, where i is a uniformly random index over the samples X_i .

Our inference procedure uses the average of t consecutive SGD iterations. In particular, the algorithm proceeds as follows: Given a sequence of SGD iterates, we use the first SGD

This chapter also appears in [LLKC18]. The theoretical analysis was written by Tianyang Li, and the experiments were conducted in collaboration with Liu Liu. It was edited by Anastasios Kyrillidis and Constantine Caramanis.

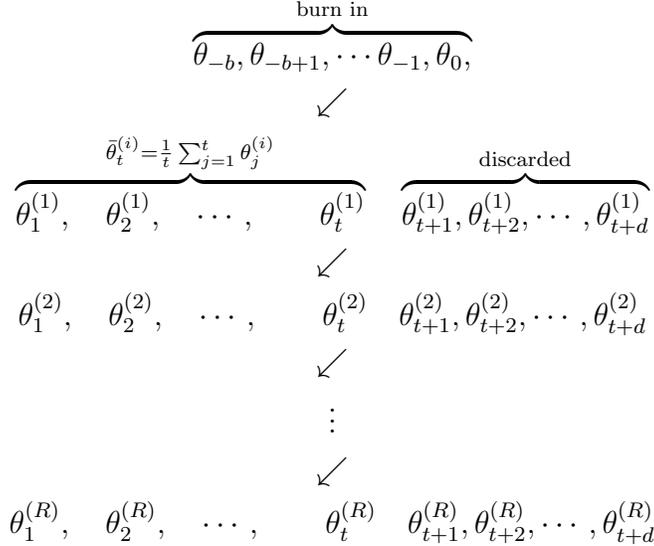


Figure 3.1: Our SGD inference procedure

iterates $\theta_{-b}, \theta_{-b+1}, \dots, \theta_0$ 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 3.1. As the final empirical minimum $\hat{\theta}$, we use in practice $\hat{\theta} \approx \frac{1}{R} \sum_{i=1}^R \bar{\theta}_t^{(i)}$ [Bub15b].

Some practical aspects of our scheme are discussed below.

Step size η 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) η gives equally accurate results with significantly reduced running time. We conjecture that a better understanding of t ’s and η ’s influence requires stronger bounds for SGD with constant step size. Heuristically, calibration methods for parameter tuning in subsampling methods ([ET94], Ch.18; [PRW12], 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 in the appendix, if we discard d SGD iterates in every segment, the correlation between consecutive $\theta^{(i)}$ and $\theta^{(i+1)}$ is of 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 [JZ13]) to find the optimum, and use a relatively small b for SGD to reach stationarity, similar to Markov Chain Monte Carlo burn-in.

Statistical inference using $\bar{\theta}_t^{(i)}$ and $\hat{\theta}$: Similar to ensemble learning [OM99], we use $i = 1, 2, \dots, R$ estimators for statistical inference:

$$\theta^{(i)} = \hat{\theta} + \sqrt{\frac{K_s \cdot t}{n}} \left(\bar{\theta}_t^{(i)} - \hat{\theta} \right). \quad (3.2)$$

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. Similar to other resampling methods such as bootstrap and subsampling, we use quantiles or variance of $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(R)}$ for statistical inference.

3.1.1 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 $\hat{\theta} \in \mathbb{R}^p$ be its minimizer, according to (2.2), and denote its Hessian at $\hat{\theta}$ by $H := \nabla^2 f(\hat{\theta}) = \frac{1}{n} \cdot \sum_{i=1}^n \nabla^2 f_i(\hat{\theta})$. Assume that $\forall \theta \in \mathbb{R}^p$, f satisfies:*

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

(F₂) *Lipschitz gradient continuity*: $\|\nabla f(\theta)\|_2 \leq L\|\theta - \hat{\theta}\|_2$, for constant $L > 0$,

(F₃) *Bounded Taylor remainder*: $\|\nabla f(\theta) - H(\theta - \hat{\theta})\|_2 \leq E\|\theta - \hat{\theta}\|_2^2$, for constant $E > 0$,

(F₄) *Bounded Hessian spectrum at $\hat{\theta}$* : $0 < \lambda_L \leq \lambda_i(H) \leq \lambda_U < \infty, \forall i$.

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

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

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

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

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

where $G = \mathbb{E}[g_s(\hat{\theta})g_s(\hat{\theta})^\top \mid \hat{\theta}]$ and, for positive, data dependent constants A, B, C, D, A_i , for $i = 1, \dots, 4$.

Assume that $\|\theta_1 - \hat{\theta}\|_2^2 = O(\eta)$; then for sufficiently small step size $\eta > 0$, the average SGD sequence, $\bar{\theta}_t$, satisfies:

$$\begin{aligned} & \left\| t\mathbb{E}[(\bar{\theta}_t - \hat{\theta})(\bar{\theta}_t - \hat{\theta})^\top] - H^{-1}GH^{-1} \right\|_2 \\ & \lesssim \sqrt{\eta} + \sqrt{\frac{1}{t\eta} + t\eta^2}. \end{aligned} \tag{3.3}$$

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. Further, in the next section, we relate a continuous approximation of SGD to Ornstein-Uhlenbeck process [RM51] to give an intuitive explanation of our results.

Discussion. For linear regression, assumptions (F₁), (F₂), (F₃), and (F₄) 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 2 in the appendix.

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. (3.7) below. When the empirical risk function is not degenerate, (3.7) satisfies assumptions (F_1) , (F_2) , (F_3) , and (F_4) . We cannot directly use vanilla SGD to minimize (3.7), instead we describe a modified SGD procedure for minimizing (3.7) in Section 3.1.3, 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 1.

Note that Theorem 1 proves consistency for SGD with fixed step size, requiring $\eta \rightarrow 0$ when $t \rightarrow \infty$. However, we empirically observe in our experiments that a sufficiently large constant η 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.

3.1.2 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 [RM51], to give an intuitive explanation of our results. In particular, under regularity conditions, the stochastic process $\Delta_t = \theta_t - \hat{\theta}$ asymptotically converges to an Ornstein-Uhlenbeck process $\Delta(t)$, [KH81, Pf86, BPM90, KY03, MHB16] that satisfies:

$$d\Delta(T) = -H\Delta(T) dT + \sqrt{\eta}G^{\frac{1}{2}} dB(T), \quad (3.4)$$

where $B(T)$ is a standard Brownian motion. Given (3.4), $\sqrt{t}(\bar{\theta}_t - \hat{\theta})$ can be approximated as

$$\begin{aligned}
\sqrt{t}(\bar{\theta}_t - \hat{\theta}) &= \frac{1}{\sqrt{t}} \sum_{i=1}^t (\theta_i - \hat{\theta}) \\
&= \frac{1}{\eta\sqrt{t}} \sum_{i=1}^t (\theta_i - \hat{\theta})\eta \approx \frac{1}{\eta\sqrt{t}} \int_0^{t\eta} \Delta(T) dT,
\end{aligned} \tag{3.5}$$

where we use the approximation that $\eta \approx dT$. By rearranging terms in (3.4) and multiplying both sides by H^{-1} , we can rewrite the stochastic differential equation (3.4) as $\Delta(T) dT = -H^{-1} d\Delta(T) + \sqrt{\eta}H^{-1}G^{\frac{1}{2}} dB(T)$. Thus, we have

$$\begin{aligned}
\int_0^{t\eta} \Delta(T) dT &= \\
&= -H^{-1}(\Delta(t\eta) - \Delta(0)) + \sqrt{\eta}H^{-1}G^{\frac{1}{2}}B(t\eta).
\end{aligned} \tag{3.6}$$

After plugging (3.6) into (3.5) we have

$$\begin{aligned}
\sqrt{t}(\bar{\theta}_t - \hat{\theta}) &\approx \\
&= -\frac{1}{\eta\sqrt{t}}H^{-1}(\Delta(t\eta) - \Delta(0)) + \frac{1}{\sqrt{t\eta}}H^{-1}G^{\frac{1}{2}}B(t\eta).
\end{aligned}$$

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

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

3.1.3 Logistic regression

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

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

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

$$f(\theta) = \frac{1}{2} \left(c + \frac{1}{n} \sum_{i=1}^n \log \left(1 + \exp(-y_i \theta^\top X_i) \right) \right)^2, \quad (3.7)$$

where $c > 0$ (e.g., $c = 1$).

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 $\mathbb{E}[\Psi_s | \theta] = \Psi$ and $\mathbb{E}[\Upsilon_s | \theta] = \Upsilon$. For Υ_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^Υ are S_Υ indices sampled from $[n]$ uniformly at random with replacement. For Ψ_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^Ψ are S_Ψ indices uniformly sampled from $[n]$ with or without replacement. Given the above, we have $\nabla f(\theta)^\top (\theta - \hat{\theta}) \geq \alpha \|\theta - \hat{\theta}\|_2^2$ for some constant α by the generalized self-concordance of logistic regression [Bac10, Bac14], 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 $f(\theta) = (k(\theta) + c)^2$, $\mathbb{E}[\Psi_s | \theta] = k(\theta) + c$, and $\mathbb{E}[\Upsilon_s | \theta] = \nabla k(\theta)$.

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

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

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

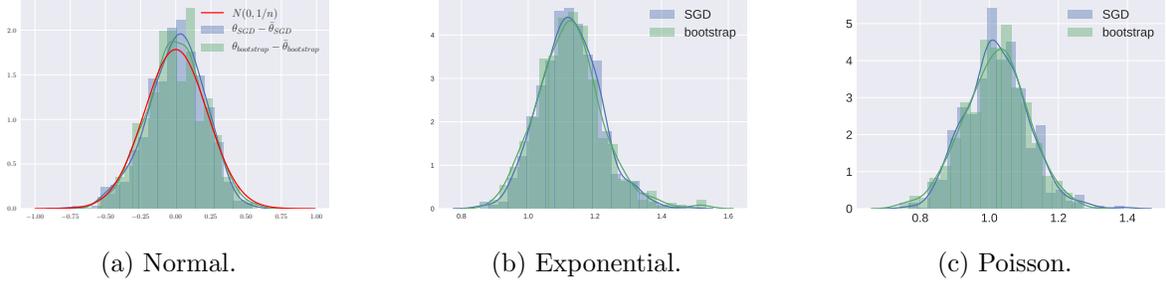


Figure 3.2: Estimation in univariate models.

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

Quantities other than t and η 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 $\frac{(\nabla^2 k(\hat{\theta}))^{-1}}{n} \left(\frac{\sum_{i=1}^n \nabla k_i(\hat{\theta}) \nabla k_i(\hat{\theta})^\top}{n} \right) \left(\nabla^2 k(\hat{\theta}) \right)^{-1}$, we set the scaling factor $K_s = \frac{(c + k(\hat{\theta}))^2}{K_G(\hat{\theta})}$ in (3.2) for statistical inference. Note that $K_s \approx 1$ for sufficiently large S_Ψ .

η	$t = 100$	$t = 500$	$t = 2500$	η	$t = 100$	$t = 500$	$t = 2500$
0.1	(0.957, 4.41)	(0.955, 4.51)	(0.960, 4.53)	0.1	(0.949, 4.74)	(0.962, 4.91)	(0.963, 4.94)
0.02	(0.869, 3.30)	(0.923, 3.77)	(0.918, 3.87)	0.02	(0.845, 3.37)	(0.916, 4.01)	(0.927, 4.17)
0.004	(0.634, 2.01)	(0.862, 3.20)	(0.916, 3.70)	0.004	(0.616, 2.00)	(0.832, 3.30)	(0.897, 3.93)

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

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

Table 3.1: Linear regression. *Left*: Experiment 1, *Right*: Experiment 2.

η	$t = 100$	$t = 500$	$t = 2500$
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)

η	$t = 100$	$t = 500$	$t = 2500$
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 3.2: Logistic regression. *Left*: Experiment 1, *Right*: Experiment 2.

3.2 Experiments

3.2.1 Synthetic data

The coverage probability is defined as $\frac{1}{p} \sum_{i=1}^p \mathbb{P}[\theta_i^* \in \hat{C}_i]$ where $\theta^* = \operatorname{argmin}_{\theta} \mathbb{E}[f(\theta, X)] \in \mathbb{R}^p$, and \hat{C}_i is the estimated confidence interval for the i^{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^{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. For results given as a pair (α, β) , it usually indicates (coverage probability, confidence interval length).

3.2.1.1 Univariate models

In Figure 3.2, 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 1.1 in the appendix shows Q-Q plots of samples from our SGD inference procedure.

Normal distribution mean estimation: Figure 3.2a 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 Section 3.4. 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 3.2b 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 3.2c 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).

3.2.1.2 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 $1_p = [1 \ 1 \ \dots \ 1]^\top \in \mathbb{R}^p$. Additional simulations comparing covariance matrix computed with different methods are given in Sec. 1.1.1.1.

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 3.1a. 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 3.1b. 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. 3.1.3) are given in Sec. 1.1.1.1. *Experiment 1:* Results for the case where $\mathbb{P}[Y = +1] = \mathbb{P}[Y = -1] = 1/2$, $X | Y \sim \mathcal{N}(0.01Y1_p/\sqrt{p}, I) \in \mathbb{R}^{10}$ with $n = 1000$ samples is given in Table 3.2a. 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 $\mathbb{P}[Y = +1] = \mathbb{P}[Y = -1] = 1/2$, $X | Y \sim \mathcal{N}(0.01Y1_p/\sqrt{p}, \Sigma) \in \mathbb{R}^{10}$, $\Sigma_{ij} = 0.2^{|i-j|}$ with $n = 1000$ samples is given in Table 3.2b. 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).

3.2.2 Real data

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

3.2.2.1 Splice data set

The Splice data set ¹ contains 60 distinct features with 1000 data samples. This is a classification problem between two classes of splice junctions in a DNA sequence. We use a logistic regression model trained using vanilla SGD.

In Figure 3.3, we compare the covariance matrix computed using our SGD inference

¹<https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html>

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.

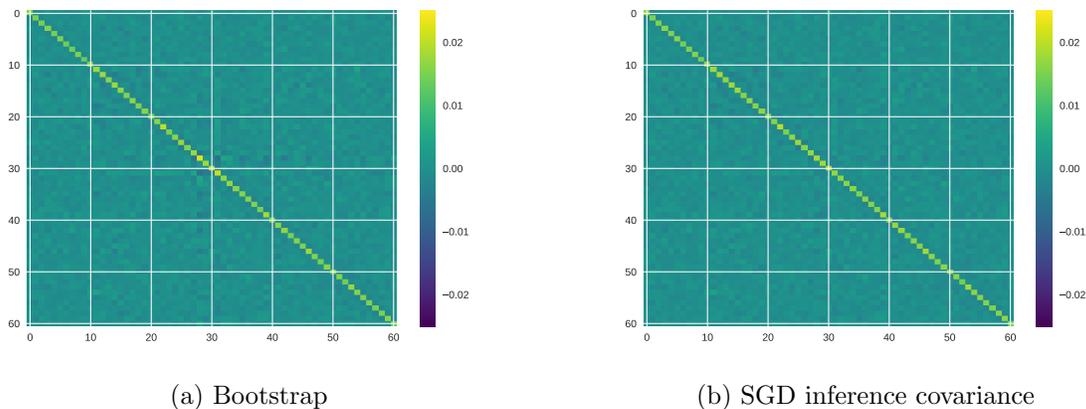


Figure 3.3: Splice data set

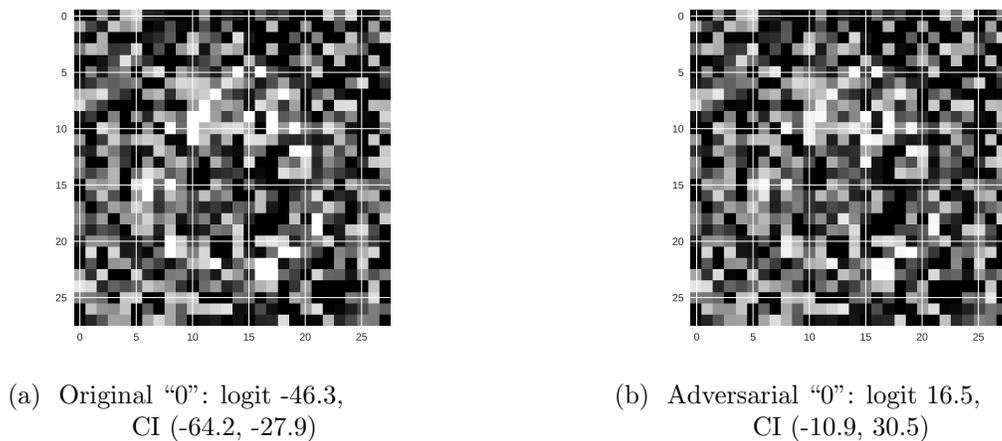


Figure 3.4: MNIST

3.2.2.2 MNIST

Here, we train a binary logistic regression classifier to classify 0/1 using a noisy MNIST data set, and demonstrate that adversarial examples produced by gradient attack [GSS14] (perturbing an image in the direction of loss function’s gradient with respect to data) can be detected using prediction intervals. We flatten each 28×28 image into a 784 dimensional vector, and train a linear classifier using pixel values as features. To add noise to each image, where each original pixel is either 0 or 1, we randomly changed 70% pixels to random numbers uniformly on $[0, 0.9]$. Next we train the classifier on the noisy MNIST data set, and generate adversarial examples using this noisy MNIST data set. Figure 3.4 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 quantiles from our SGD inference procedure.

3.2.3 Discussion

In our experiments, we observed that using a larger step size η produces accurate results with significantly accelerated convergence time. This might imply that the η 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 $\eta t \rightarrow \infty$ and $\eta^2 t \rightarrow 0$ 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 η and only require $\eta t \rightarrow \infty$.

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

3.3 Linear Regression

In linear regression, the empirical risk function satisfies:

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

where y_i denotes the observations of the linear model and x_i are the regressors. To find an estimate to θ^* , one can use SGD with stochastic gradient give 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 $\nabla f(\theta) - H(\theta - \hat{\theta}) = 0$, linear regression has a stronger result than general M -estimation problems.

Corollary 2. *Assume that $\|\theta_1 - \hat{\theta}\|_2^2 = O(\eta)$, we have*

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

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

We assume that $S = O(1)$ is bounded, and quantities other than t and η 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 $\frac{1}{n} (\frac{1}{n} \sum_{i=1}^n x_i x_i^\top)^{-1} (\frac{1}{n} (x_i^\top \hat{\theta} - y_i) (x_i^\top \hat{\theta} - y_i)^\top) (\frac{1}{n} \sum_{i=1}^n x_i x_i^\top)^{-1}$, we set the scaling factor $K_s = S$ in (3.2) for statistical inference.

3.4 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, \dots, X_n , the mean is estimated by solving the following optimization problem

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

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^{th} mini batch SGD step, the update step is

$$\theta_{t+1} = \theta_t - \eta(\theta_t - Y_t) = (1 - \eta)\theta_t + \eta Y_t, \quad (3.8)$$

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

$$\sqrt{t}\hat{\theta}_t = -\frac{1}{\eta\sqrt{t}}(\theta_{t+1} - \theta_1) + \frac{1}{\sqrt{t}} \sum_{i=1}^n Y_i. \quad (3.9)$$

Assume that $\|\theta_1 - \hat{\theta}\|_2^2 = O(\eta)$, then from Chebyshev's inequality $-\frac{1}{\eta\sqrt{t}}(\theta_{t+1} - \theta_1) \rightarrow 0$ almost surely when $t\eta \rightarrow \infty$. By the central limit theorem, $\frac{1}{\sqrt{t}} \sum_{i=1}^n Y_i$ converges weakly to $\mathcal{N}(\hat{\theta}, \frac{1}{S}\hat{\Sigma})$ with $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\theta})(X_i - \hat{\theta})^\top$. From (3.8), 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 $\|\operatorname{Cov}(\theta^{(i)}, \theta^{(j)})\|_2 = O(\eta(1 - \eta)^{d+|i-j|})$. 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 (3.2).

Chapter 4

Approximate Newton-based statistical inference using only stochastic gradients

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

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

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

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

Statistical inference, such as computing one-dimensional confidence intervals, gives us information beyond the point estimate $\hat{\theta}$, when $\hat{\theta}$ has an asymptotic limit distribution [Was13]. *E.g.*, under regularity conditions, the M -estimator satisfies asymptotic normality [vdV98, Theorem 5.21]. *I.e.*, $\sqrt{n}(\hat{\theta} - \theta^*)$ weakly converges to a normal distribution:

$$\sqrt{n}(\hat{\theta} - \theta^*) \rightarrow \mathcal{N}(0, H^{*-1}G^*H^{*-1}),$$

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

This chapter also appears in [LKLC18]. The theoretical analysis was written by Tianyang Li, and the experiments were conducted in collaboration with Liu Liu. It was edited by Anastasios Kyrillidis and Constantine Caramanis.

plug-in covariance estimator $\widehat{H}^{-1}\widehat{G}\widehat{H}^{-1}$ for $H^{\star-1}G^{\star}H^{\star-1}$, where:

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

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 $\widehat{H}^{-1}\widehat{G}\widehat{H}^{-1}$.

4.1 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:

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

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 θ_0 which is sufficiently close to $\widehat{\theta}$, which can be effectively achieved using SVRG [JZ13]; a description of the SVRG algorithm can be found in Chapter 7. 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,

$$\min_{g \in \mathbb{R}^p} \left\langle \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\theta_t), g \right\rangle + \frac{1}{2\rho_t} \langle g, \nabla^2 f(\theta_t) g \rangle, \quad (4.1)$$

which can be seen as a generalization of SVRG; this relationship is described in more detail in Chapter 7. In particular, these lines correspond to solving (4.1) using SGD by uniformly

Algorithm 1 Unregularized M-estimation statistical inference

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

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

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

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

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

6: **for** $j = 0$ to $L - 1$ **do** // solving (4.1) approximately using SGD

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

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

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

10: **end for**

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

13: **end for**

sampling a random f_i , and approximating:

$$\nabla^2 f(\theta)g \approx \frac{\nabla f(\theta + \delta_t^j g) - \nabla f(\theta)}{\delta_t^j} = \mathbb{E} \left[\frac{\nabla f_i(\theta + \delta_t^j g) - \nabla f_i(\theta)}{\delta_t^j} \mid \theta \right]. \quad (4.2)$$

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 [Ama98].

In terms of parameters, similar to [PJ92, Rup88], 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 2. *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

$$\rho_t = \rho_0 \cdot (t+1)^{-d_o}, \quad \text{where } d_o \in \left(\frac{1}{2}, 1\right) \text{ is the decaying rate.} \quad (4.3)$$

In each outer loop, the inner loop step size is

$$\tau_j = \tau_0 \cdot (j+1)^{-d_i}, \quad \text{where } d_i \in \left(\frac{1}{2}, 1\right) \text{ is the decaying rate.} \quad (4.4)$$

The scaling constant for Hessian vector product approximation is

$$\delta_t^j = \delta_0 \cdot \rho_t^4 \cdot \tau_j^4 = o\left(\frac{1}{(t+1)^2(j+1)^2}\right). \quad (4.5)$$

Then, for the outer iterate θ_t we have

$$\mathbb{E} \left[\|\theta_t - \hat{\theta}\|_2^2 \right] \lesssim t^{-d_o}, \quad (4.6) \quad \text{and} \quad \mathbb{E} \left[\|\theta_t - \hat{\theta}\|_2^4 \right] \lesssim t^{-2d_o}. \quad (4.7)$$

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

$$\mathbb{E} \left[\left\| \frac{\bar{g}_t}{\rho_t} - [\nabla^2 f(\theta_t)]^{-1} g_t^0 \right\|_2^2 \mid \theta_t \right] \lesssim \frac{1}{L} \|g_t^0\|_2^2, \quad (4.8)$$

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

$$\mathbb{E} \left[\left\| g_t^{j+1} - [\nabla^2 f(\theta_t)]^{-1} g_t^0 \right\|_2^4 \mid \theta_t \right] \lesssim (j+1)^{-2d_i} \|g_t^0\|_2^4. \quad (4.9)$$

After T steps of the outer loop, we have a non-asymptotic bound on the “covariance”:

$$\mathbb{E} \left[\left\| H^{-1} G H^{-1} - \frac{S_o}{T} \sum_{t=1}^T \frac{\bar{g}_t \bar{g}_t^\top}{\rho_t^2} \right\|_2 \right] \lesssim T^{-\frac{d_o}{2}} + L^{-\frac{1}{2}}, \quad (4.10)$$

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

Some comments on the results in Theorem 2. The main outcome is that (4.10) 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 (4.8), and the optimization bound in (4.6). As a side note, the rates in Theorem 2 are very similar to classic results in stochastic approximation [PJ92, Rup88]; 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 ([ET94], Ch.18; [PRW12], 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 $\mathcal{N}\left(0, \frac{1}{S_o} H^{-1} G H^{-1}\right)$; we give an example using mean estimation in Section 4.4.1. For a similar algorithm based on SVRG (Algorithm 2 in Section 4.4), 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 Section 4.3, 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 $H^{-1} G H^{-1}$), 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 [TA17, HAV⁺15, LH15, DLH16] 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 [PJ92, Rup88].

Corollary 3. *In Algorithm 1’s outer loop, the average of consecutive iterates satisfies*

$$\mathbb{E} \left[\left\| \frac{\sum_{t=1}^T \theta_t}{T} - \hat{\theta} \right\|_2^2 \right] \lesssim \frac{1}{T}, \quad (4.11) \quad \text{and} \quad \frac{1}{\sqrt{T}} \left(\frac{\sum_{t=1}^T \theta_t}{T} - \hat{\theta} \right) = W + \Delta, \quad (4.12)$$

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

Corollary 3 uses 2nd, 4th moment bounds on individual iterates (eqs. (4.6), (4.7) in the above theorem), and the approximation of inverse Hessian conditioned stochastic gradient in (4.9).

4.2 Experiments

4.2.1 Synthetic data

The coverage probability is defined as $\frac{1}{p} \sum_{i=1}^p \mathbb{P}[\theta_i^* \in \hat{C}_i]$, where \hat{C}_i is the estimated confidence interval for the i^{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^{th} coordinate. In our experiments, coverage probability and average confidence interval length are estimated through simulation. Result given as a pair (α, β) 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 4.1: Synthetic data average coverage & confidence interval length for low dimensional problems.

Table 4.1 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 2.1.1.1. Compared with Bootstrap and Jackknife [ET94], Algorithm 1 uses less numerical operations, while achieving similar results. Compared with the

Approximate Newton	Bootstrap	Inverse Fisher information
(0.951, 0.224)	(0.946 0.205)	(0.966, 0.212)

Table 4.2: Average 95% confidence interval (coverage, length) after calibration

averaged SGD method [LLKC18, CLTZ16], our algorithm performs much better, while using the same amount of computation, and is much less sensitive to the choice hyper-parameters. And we observe that calibrated approximate Newton confidence intervals [ET94, PRW12] are better than bootstrap and inverse Fisher information (Table 4.2).

4.2.2 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 [BJ16, CM16], and this causes the risk function’s Hessian to be degenerate [SEG⁺17]. 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 Chapter 4). When we make a prediction $\Psi(x; \hat{\theta})$ using a fixed data point x as input (i.e., conditioned on x), using the delta method [vdV98], the confidence interval of the prediction can be derived from the asymptotic normality of $\Psi(x; \hat{\theta})$

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

To detect adversarial attacks, we use the score

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

to measure how much $\nabla_{\theta} \Psi(x; \hat{\theta})$ 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 [XRV17]. Figure 4.1 shows distributions of scores of original images, adversarial images generated using the fast gradient sign method [GSS14], and randomly perturbed images. Adversarial and random perturbations have the same ℓ_∞ norm. The adversarial perturbations and example images are shown in Appendix 2.1.2.1. 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.

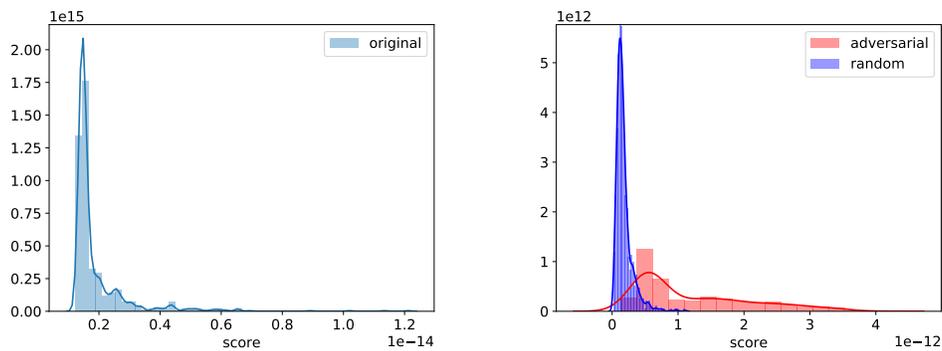


Figure 4.1: Distribution of scores for original, randomly perturbed, and adversarially perturbed images

4.3 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.

4.3.1 Unregularized M-estimation

Similar to Theorem 2’s proof, we have the following result when the number of inner loop increases in the outer loops.

Corollary 4. *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 $(L)_t = L(t+1)^{d_L}$ for some $d_L > 0$, then $\sqrt{\frac{1}{T} \sum_{i=1}^T \frac{1}{(L)_t}} = O(\frac{1}{\sqrt{L}} T^{-\frac{d_L}{2}})$.

4.4 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 2 has stronger guarantees than Algorithm 1, Algorithm 2 requires a full gradient evaluation in each outer loop.

Corollary 5. *In Algorithm 2, 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*

Algorithm 2 SVRG based statistical inference algorithm in unregularized M-estimation

- 1: **for** $t \leftarrow 0; t < T; ++t$ **do**
 - 2: $d_t^0 \leftarrow -\eta \nabla f(\theta_t) = -\eta \left(\frac{1}{n} \sum_{i=1}^n \nabla f_i(\theta) \right)$ // point estimation via SVRG
 - 3: $I_o \leftarrow$ uniformly sample S_o indices with replacement from $[n]$
 - 4: $g_t^0 \leftarrow -\rho_t \left(\frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\theta_t) \right)$ // statistical inference
 - 5: **for** $j \leftarrow 0; j < L; ++j$ **do** // solving (4.1) approximately using SGD
 - 6: $I_i \leftarrow$ uniformly sample S_i indices without replacement from $[n]$
 - 7: $d_t^{j+1} \leftarrow d_t^j - \eta \left(\frac{1}{S_i} \sum_{k \in I_i} (\nabla f_k(\theta_t + d_t^j) - \nabla f_k(\theta_t)) \right) + d_t^0$ // point estimation via SVRG
 - 8: $g_t^{j+1} \leftarrow g_t^j - \tau_j \left(\frac{1}{S_i} \sum_{k \in I_i} \frac{1}{\delta_t^j} [\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)] \right) + \tau_j g_t^0$ // statistical inference
 - 9: **end for**
 - 10: Use $\sqrt{S_o} \cdot \frac{\bar{g}_t}{\rho_t}$ for statistical inference // $\bar{g}_t = \frac{1}{L+1} \sum_{j=0}^L g_t^j$
 - 11: $\theta_{t+1} \leftarrow \theta_t + \bar{d}_t$ // $\bar{d}_t = \frac{1}{L+1} \sum_{j=0}^L d_t^j$
 - 12: **end for**
-

the outer loop, we have a non-asymptotic bound on the “covariance”

$$\mathbb{E} \left[\left\| \left\| H^{-1}GH^{-1} - \frac{S_o}{T} \sum_{t=1}^T \frac{\bar{g}_t \bar{g}_t^\top}{\rho_t^2} \right\| \right\|_2 \right] \lesssim L^{-\frac{1}{2}}, \quad (4.13)$$

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 $T \rightarrow \infty$ and $L \rightarrow \infty$ ($\mathbb{E}[\|\Delta\|_2] \lesssim \frac{1}{\sqrt{T}} + \frac{1}{L}$).

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 4.

A better understanding of concentration, and Edgeworth expansion of the average consecutive iterates averaged (beyond [Dip08a, Dip08b]) in stochastic approximation, would

give stronger guarantees for our algorithms, and better compare and understand different algorithms.

4.4.1 Lack of asymptotic normality in Algorithm 1 for mean estimation

In mean estimation, we solve the following optimization problem

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

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

$$\begin{aligned} & \frac{1}{\sqrt{T}} \left(\sum_{i=1}^T \frac{\bar{g}_i}{\rho_i} \right) \\ &= \frac{1}{\sqrt{T}} \left(\sum_{t=1}^T X_t - \sum_{t=1}^T \sum_{i=0}^{t-1} \rho(1-\rho)^{t-1-i} X_i \right) \\ &= \frac{1}{\sqrt{T}} \left(\sum_{t=1}^T X_t - \sum_{i=0}^{T-1} \left(\sum_{t=i+1}^T \rho(1-\rho)^{t-1-i} \right) X_i \right) \\ &= \frac{1}{\sqrt{T}} \left(\sum_{t=1}^T X_t - \sum_{i=0}^{T-1} (1 - (1-\rho)^{T-i}) X_i \right) \end{aligned}$$

$$= \frac{1}{\sqrt{T}} (X_T - X_0 + \sum_{i=1}^{T-1} (1 - \rho)^{T-i} X_i),$$

whose ℓ_2 norm's expectation converges to 0 when $T \rightarrow \infty$, 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 $\mathcal{N} \left(0, \frac{1}{s_o} H^{-1} G H^{-1} \right)$.

Chapter 5

High dimensional linear regression

In this chapter, 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 [ZZ14, vdGBRD14, JM15] propose statistical inference via de-biased LASSO estimators. Here, we present a new ℓ_1 -norm regularized objective and propose an approximate stochastic *proximal* Newton algorithm, using only first order information.

We consider the linear model $y_i = \langle \theta^*, x_i \rangle + \epsilon_i$, for some sparse $\theta^* \in \mathbb{R}^p$. 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 θ* : **(i)** θ^* is s -sparse; **(ii)** $\|\theta^*\|_2 = O(1)$, which implies that $\|\theta^*\|_1 \lesssim \sqrt{s}$.
- *Assumptions on Σ* : **(i)** Σ is sparse, where each column (and row) has at most b non-zero entries;¹ **(ii)** Σ is well conditioned: all of Σ ’s eigenvalues are $\Theta(1)$; **(iii)** Σ 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 ℓ_∞ norm of \widehat{S}^{-1} [Var75]. A commonly used design covariance that satisfies all of our assumptions is I .

This chapter also appears in [LKLC18]. The theoretical analysis was written by Tianyang Li, and the experiments were conducted in collaboration with Liu Liu. It was edited by Anastasios Kyrillidis and Constantine Caramanis.

¹This is satisfied when Σ is block diagonal or banded. Covariance estimation under this sparsity assumption has been extensively studied [BL08, BRT09, CZ12], and soft thresholding is an effective yet simple estimation method [RLZ09].

We estimate θ^* using:

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{2} \left\langle \theta, \left(\hat{S} - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \theta \right\rangle + \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (x_i^\top \theta - y_i)^2 + \lambda \|\theta\|_1, \quad (5.1)$$

where $\hat{S}_{jk} = \text{sign} \left(\left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)_{jk} \right) \left(\left| \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)_{jk} \right| - \omega \right)_+$ is an estimate of Σ by soft-thresholding each element of $\frac{1}{n} \sum_{i=1}^n x_i x_i^\top$ with $\omega = \Theta \left(\sqrt{\frac{\log p}{n}} \right)$ [RLZ09]. Under our assumptions, \hat{S} is positive definite with high probability when $n \gg b^2 \log p$ (Lemma 9), and this guarantees that the optimization problem (5.1) 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

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

which is similar to LASSO regression [BvdG11, NRWY12] and related estimators using thresholded covariance [YLR14, JD11].

Point estimate. Theorem 3 provides guarantees for our proposed point estimate (5.1).

Theorem 3. *When $n \gg b^2 \log p$, the solution $\hat{\theta}$ in (5.1) satisfies*

$$\left\| \hat{\theta} - \theta^* \right\|_1 \lesssim s (\sigma + \|\theta^*\|_1) \sqrt{\frac{\log p}{n}} \lesssim s (\sigma + \sqrt{s}) \sqrt{\frac{\log p}{n}}, \quad (5.2)$$

$$\left\| \hat{\theta} - \theta^* \right\|_2 \lesssim \sqrt{s} (\sigma + \|\theta^*\|_1) \sqrt{\frac{\log p}{n}} \lesssim \sqrt{s} (\sigma + \sqrt{s}) \sqrt{\frac{\log p}{n}}, \quad (5.3)$$

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

Confidence intervals. We next present a de-biased estimator $\hat{\theta}^{\text{d}}$ (5.4), based on our proposed estimator. $\hat{\theta}^{\text{d}}$ can be used to compute confidence intervals and p-values for each

coordinate of $\hat{\theta}^d$, which can be used for false discovery rate control [JJ18]. The estimator satisfies:

$$\hat{\theta}^d = \hat{\theta} + \hat{S}^{-1} \left[\frac{1}{n} \sum_{i=1}^n (y_i - x_i^\top \hat{\theta}) x_i \right]. \quad (5.4)$$

Our de-biased estimator is similar to [ZZ14, vdGBRD14, JM14, JM15]. However, we have different terms, since we need to de-bias covariance estimation. Our estimator assumes $n \gg b^2 \log p$, since then \hat{S} is positive definite with high probability (Lemma 9). The assumption that Σ is diagonally dominant guarantees that the ℓ_∞ norm $\|\hat{S}^{-1}\|_\infty$ is bounded by $O\left(\frac{1}{D_\Sigma}\right)$ with high probability when $n \gg \frac{1}{D_\Sigma^2} \log p$.

Theorem 4 shows that we can compute valid confidence intervals for each coordinate when $n \gg \left(\frac{1}{D_\Sigma} s (\sigma + \|\theta^*\|_1) \log p\right)^2$. This is satisfied when $n \gg \left(\frac{1}{D_\Sigma} s (\sigma + \sqrt{s}) \log p\right)^2$. And the covariance is similar to the sandwich estimator [Hub67, Whi80].

Theorem 4. *Under our assumptions, when $n \gg \max\{b^2, \frac{1}{D_\Sigma^2}\} \log p$, we have:*

$$\sqrt{n}(\hat{\theta}^d - \theta^*) = Z + R, \quad (5.5)$$

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

Our estimate in (5.1) has similar error rates to the estimator in [YLR14]; 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 [RWY11], and the sample complexity in Theorem 4 is slightly higher than other methods [vdGBRD14, JM14, JM15], our criterion is strongly convex with high probability: this allows us to use linearly convergent proximal algorithms [XZ14, LSS14], whereas provable linearly convergent optimization bounds for LASSO only guarantees convergence to a neighborhood of the LASSO solution within statistical error [ANW10]. This

is crucial for computing the de-biased estimator, as we need the optimization error to be much less than the statistical error.

We present our algorithm for statistical inference in high dimensional linear regression using stochastic gradients below. 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 [Hub67, Whi80]. Algorithm 3 computes the statistical error covariance. Similar to Algorithm 1, Algorithm 3 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 [XZ14]. 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 [MB11] and proximal SGD [AFM17] 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 $O(T \cdot n \cdot p^2 \cdot \log(p) \cdot \log(T))$ numerical operations. And Algorithm 4 calculates the de-biased estimator $\widehat{\theta}^{\text{d}}$ (5.4) via SVRG.

5.1 Experiments

5.1.1 Synthetic data

We use 600 i.i.d. samples from a model with $\Sigma = I$, $\sigma = 0.7$, $\theta^* = [1/\sqrt{8}, \dots, 1/\sqrt{8}, 0, \dots, 0]^\top \in \mathbb{R}^{1000}$ which is 8-sparse. Figure 5.1 shows 95% confidence intervals for the first 20 coordinates. The average confidence interval length is 0.14 and average coverage is 0.83. Additional experimental results, including p-value distribution under the null hypothesis, are presented in Appendix 3.1.1.1.

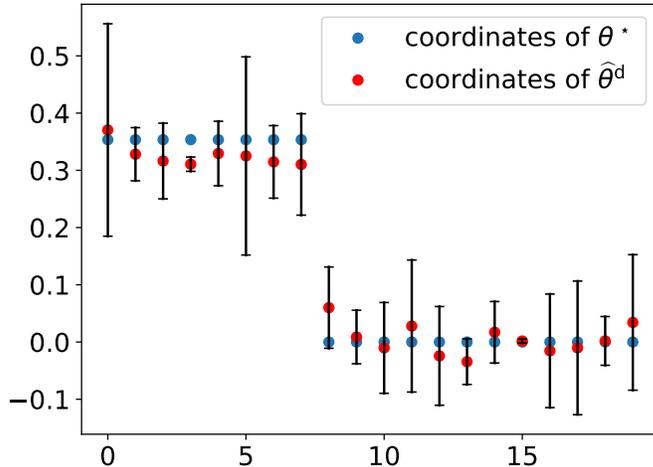


Figure 5.1: 95% confidence intervals

Figure 5.2 shows p-value distribution under the null hypothesis for our method and the de-biased LASSO estimator with known covariance, using 600 i.i.d. samples generated from a model with $\Sigma = I$, $\sigma = 0.7$, and we can see that it is close to a uniform distribution, similar results are observed for other high dimensional statistical inference procedures such as [CFJL18].

5.1.2 Real data

HIV drug resistance mutations dataset. We apply our high dimensional inference procedure to the dataset in [RTW+06] to detect mutations related to HIV drug resistance, where we randomly sub-sample the dataset so that the number of features is larger than the number of samples. When we control the family-wise error rate (FWER) at 0.05 using the Bonferroni correction [Bon36], our procedure is able to detect verified mutations in an expert dataset [JBVC+05] (Table 5.1), and the details are given in Appendix 3.1.2.1.

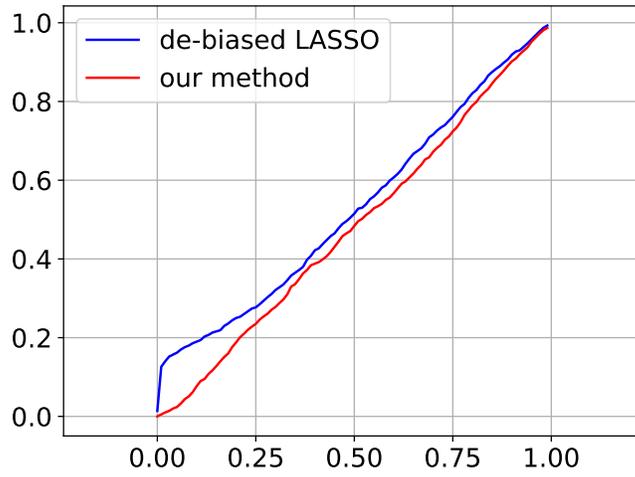
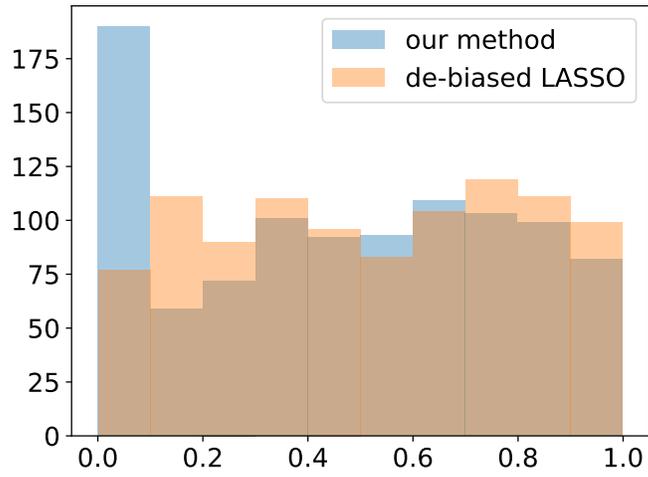


Figure 5.2: Distribution of two-sided Z-test p-values under the null hypothesis (high dimensional)

Drug		Mutations
PI	APV	10F
	ATV	33F, 43T, 84V
	IDV	48V, 84A
	LPV	46I
	NFV	46L
	RTV	10I, 54V
	SQV	20R, 84V
NRTI	3TC	184V
	ABC	41L
	AZT	41L, 210W
	D4T	41L, 215Y
	DDI	62V, 151M
	TDF	41L, 75M
NNRTI	DLV	228R
	EFV	74V, 103N
	NVP	103N, 181C

Table 5.1: HIV drug resistance related mutations detected by our high dimensional inference procedure

Riboflavin (vitamin B2) production rate data set. We apply our high dimensional linear regression statistical inference procedure to a high-throughput genomic data set concerning riboflavin (vitamin B2) production rate [BKM14], which contains $n = 71$ samples of $p = 4088$ genes. We set $\lambda = 4.260$ and $\omega = 0.5$. In Appendix 3.1.2.1, we show that our point estimate is similar to the vanilla LASSO estimate, and compare our statistical inference results with those of [JM14, BKM14, Böh13, MMB09].

5.2 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:

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

which is related to the empirical sandwich estimator [Hub67, Whi80]. Lemma 1 shows this is a good estimate of the covariance when $n \gg \frac{1}{D_\Sigma^4} \max\{1, \sigma^2\} s^2 (\sigma + \|\theta^*\|_1)^2$.

Algorithm 3 performs statistical inference in high dimensional linear regression (5.1), by computing the statistical error covariance in Theorem 4, based on the plug-in estimate in Lemma 1. We denote the soft thresholding of A by ω as an element-wise procedure $(\mathbf{S}_\omega(A))_e = \text{sign}(A_e)(|A_e| - \omega)_+$. For a vector v , we write v 's i^{th} coordinate as $v(i)$. The optimization objective (5.1) is denoted as:

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

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

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

where $\mathbf{e}_i \in \mathbb{R}^p$ is the basis vector where the i^{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 [LSS14] at θ solves the optimization problem

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

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

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

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 [XZ14], because in the high dimensional setting, the variance using SGD [MB11] and proximal SGD [AFM17] for solving Newton steps is too large. However because $p \gg n$, instead of sampling *by sample*, we sample *by feature*. We start from θ_0 sufficiently close to $\hat{\theta}$ (see Theorem 5 for details), which can be effectively achieved using proximal SVRG (Section 5.4). 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 Section 5.5. 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 $O(T \cdot n \cdot p^2 \cdot \log(p) \cdot \log(T))$ numerical operations. Calculation of the de-biased estimator $\hat{\theta}^d$ (5.4) via SVRG is described in Section 5.3.

5.3 Computing the de-biased estimator (5.4) 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

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

And we compute $\hat{S}^{-1} \frac{1}{n} \sum_{i=1}^n \nabla f_i(\hat{\theta})$ using SVRG [JZ13] by solving the following optimization

Algorithm 3 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: $d_t^0 \leftarrow -\left(\mathbf{g}_{\widehat{S}}(\theta_t) - \frac{1}{n} \sum_{i=1}^n [\nabla f_i(\theta_t + \theta_t) - \nabla f_i(\theta_t)] + \frac{1}{n} \sum_{i=1}^n \nabla f_i(\theta_t)\right)$
 - 6: **for** $j = 1$ **to** L_o^t **do** // solving Newton steps using SVRG
 - 7: $u_t^j \leftarrow \mathbf{g}_{\widehat{S}}(g_t^{j-1}) - g_t^0$
 - 8: $v_t^j \leftarrow \mathbf{g}_{\widehat{S}}(d_t^{j-1}) - d_t^0$
 - 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: $g_t^j \leftarrow g_t^j - \tau \left[u_t^j + \frac{p}{S_i} \sum_{k \in S_i} \left[g_t^j(k) - g_t^{j-1}(k) \right] \cdot \mathbf{S}_\omega (\nabla f_k(\theta_t + \mathbf{e}_k) - \nabla f_k(\theta_t)) \right]$
 - 13: $d_t^j \leftarrow \mathbf{S}_{\eta\lambda} \left(d_t^j - \eta \left[v_t^j + \frac{p}{S_i} \sum_{k \in S_i} \left[d_t^j(k) - d_t^{j-1}(k) \right] \cdot \mathbf{S}_\omega (\nabla f_k(\theta_t + \mathbf{e}_k) - \nabla f_k(\theta_t)) \right] \right)$
 - 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 $\bar{d}_t = \frac{1}{L_o^t+1} \sum_{j=0}^{L_o^t} d_t^j$ // point estimation (optimization)
 - 18: **end for**
-

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 4 Computing the de-biased estimator (5.4) via SVRG

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

Similar to Algorithm 3, we choose $\eta = \Theta\left(\frac{1}{p}\right)$ and $L_i = \Theta(p)$.

5.4 Solving the high dimensional linear regression optimization objective (5.1) using proximal SVRG

We solve our high dimensional linear regression optimization problem using proximal SVRG [XZ14]

$$\widehat{\theta} = \arg \min_{\theta} \frac{1}{2} \theta^\top \left(\widehat{S} - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \theta + \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (x_i^\top \theta - y_i)^2 + \lambda \|\theta\|_1. \quad (5.6)$$

Similar to Algorithm 3, we choose $\eta = \Theta\left(\frac{1}{p}\right)$ and $L_i = \Theta(p)$.

5.5 Non-asymptotic covariance estimate bound and asymptotic normality in Algorithm 3

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

Algorithm 5 Solving the high dimensional linear regression optimization objective (5.1) using proximal SVRG

```

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

```

Theorem 5. 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]: $\max_{1 \leq i \leq n} |\epsilon_i| \lesssim \sigma \sqrt{\log n}$,
- [B]: $\max_{1 \leq i \leq n} \|x_i\|_\infty \lesssim \sqrt{\log p + \log n}$,
- [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 3.

Here, we denote the objective function as

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

Then, we have a non-asymptotic covariance estimate bound

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

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

And we have asymptotic normality

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

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

Note that when we choose $L_o^t = \Theta(\log(p) \cdot \log(t))$, and start from θ_0 satisfying $P(\theta_0) - P(\hat{\theta}) \lesssim \frac{1}{p(\log p + \log n)^2}$ which can be effectively achieved using proximal SVRG (Section 5.4), 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 $O(T \cdot n \cdot p^2 \cdot \log(p) \cdot \log(T))$ numerical operations.

5.6 Plug-in statistical error covariance estimate

Algorithm 3 is similar to using plug-in estimator $\frac{1}{n} \sum_{i=1}^n (x_i^\top \hat{\theta} - y_i)^2 x_i x_i^\top$ for $\sigma^2 (\frac{1}{n} \sum_{i=1}^n x_i x_i^\top)$ in Theorem 4, similar to the sandwich estimator [Hub67, Whi80]. Lemma 1 gives a bound on using this plug-in estimator in the statistical error covariance (Theorem 4) 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*

$$\begin{aligned} & \left\| \hat{S}^{-1} \left(\frac{1}{n} \sum_{i=1}^n (x_i^\top \hat{\theta} - y_i)^2 x_i x_i^\top \right) \hat{S}^{-1} - \sigma^2 \hat{S}^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \hat{S}^{-1} \right\|_{\max} \\ & \lesssim \frac{1}{D_\Sigma^2} \left(\sigma \sqrt{\log n} + s (\sigma + \|\theta^*\|_1) \sqrt{\log p + \log n} \sqrt{\frac{\log p}{n}} \right) s (\sigma + \|\theta^*\|_1) (\log p + \log n)^{\frac{3}{2}} \sqrt{\frac{\log p}{n}}, \end{aligned}$$

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

Chapter 6

Time series analysis

6.1 Time series analysis

In this section, we present a sampling scheme for statistical inference in time series analysis using M -estimation, where we sample contiguous blocks of indices, instead of uniformly.

We consider a linear model $y_i = \langle x_i, \theta^* \rangle + \epsilon_i$, where $\mathbb{E}[\epsilon_i x_i] = 0$, but $\{x_i, y_i\}_{i=1}^n$ may not be i.i.d. as this is a time series. And we use ordinary least squares (OLS) $\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^n \frac{1}{2} (\langle x_i, \theta \rangle - y_i)^2$ to estimate θ^* . Applications include multifactor financial models for explaining returns [BBMS13, RM73]. For non-i.i.d. time series data, OLS may not be the optimal estimator, as opposed to the maximum likelihood estimator [SS11], but OLS is simple yet often robust, compared to more sophisticated models that take into account time series dynamics. And it is widely used in econometrics for time series analysis [Ber91]. To perform statistical inference, we use the asymptotic normality

$$\sqrt{n} (\hat{\theta} - \theta^*) \rightarrow \mathcal{N}(0, H^{*-1} G^* H^{*-1}), \quad (6.1)$$

where $H^* = \lim_{n \rightarrow \infty} \frac{1}{n} (\sum_{i=1}^n \nabla^2 f_i(\theta^*))$ and $G^* = \lim_{n \rightarrow \infty} \frac{1}{n} (\sum_{i=1}^n \sum_{j=1}^n \nabla f_i(\theta^*) \nabla f_j(\theta^*)^\top)$, with $f_i(\theta) = \frac{1}{2} (\langle x_i, \theta \rangle - y_i)^2$. The difference compared with the i.i.d. case (Chapter 4) is that G^* now includes autocovariance terms. We use the plug-in estimate $\hat{H} = \frac{1}{n} \sum_{i=1}^n \nabla^2 f_i(\hat{\theta})$ as

This chapter also appears in [LKLC18]. The theoretical analysis was written by Tianyang Li, and the experiments were conducted in collaboration with Liu Liu. It was edited by Anastasios Kyrillidis and Constantine Caramanis.

before, and we estimate G^* using the Newey-West covariance estimator [NW86] for HAC (heteroskedasticity and autocorrelation consistent) covariance estimation

$$\widehat{G} = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\widehat{\theta}) f_i(\widehat{\theta})^\top + \sum_{j=1}^{\mathbf{l}} w(j, \mathbf{l}) \sum_{i=j+1}^n \left(\nabla f_i(\widehat{\theta}) \nabla f_{i-j}(\widehat{\theta})^\top + \nabla f_{i-j}(\widehat{\theta}) \nabla f_i(\widehat{\theta})^\top \right), \quad (6.2)$$

where $w(j, \mathbf{l})$ is sample autocovariance weight, such as Bartlett weight $w(j, \mathbf{l}) = 1 - j/(\mathbf{l} + 1)$ [Bar46], and \mathbf{l} is the *lag* parameter, which captures data dependence across time. Note that this is an essential building block in time series statistical inference procedures, such as Driscoll-Kraay standard errors [DK98, KD99], moving block bootstrap [Kun89], and circular bootstrap [PR92, PR94].

In our framework, we solve OLS using our approximate Newton procedure with a slight modification to Algorithm 1. Instead of uniformly sampling indices as in line 4 of Algorithm 1, we uniformly select some $i_o \in [n]$, and set the outer mini-batch indexes I_o to the random contiguous block $\{i_o, i_o + 1, \dots, i_o + \mathbf{l} - 1\} \bmod n$, where we let the indexes circularly wrap around, as in line 4 of Algorithm 6, and this sampling scheme is similar to *circular bootstrap*. Here \mathbf{l} is the lag parameter, similar to the Newey-West estimator. And the stochastic gradient's expectation is still the full gradient. The complete algorithm is in Algorithm 6, and its guarantees are given in Corollary 6. Our approximate Newton statistical inference procedure is equivalent to using weight $w(j, \mathbf{l}) = 1 - j/\mathbf{l}$ in the Newey-West covariance estimator (6.2), with negligible terms for blocks that wrap around, and this is the same as circular bootstrap. Note that the connection between sampling scheme and Newey-West estimator was also observed in [Kun89]. Following [PR92], we can set the lag parameter such that $\mathbf{l} \cdot n^{-1/3} \rightarrow 0$, and run at least n outer loops. In practice, other methods for tuning the lag parameter can be used, such as [NW94]. For more details, we refer the reader to Section 6.2.

6.2 Time series statistical inference with approximate Newton steps using only stochastic gradients (Section 6.1)

Here, we give the complete approximate Newton-based time series statistical inference algorithm using only stochastic gradients.

Algorithm 6 Unregularized M-estimation statistical inference

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

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

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

4: Uniformly select some $i_o \in [n]$, then set I_o to the random contiguous block $\{i_o, i_o + 1, \dots, i_o + \mathbf{l} - 1\} \bmod n$, which circularly wraps around

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

6: **for** $j = 0$ to $L - 1$ **do** // solving (4.1) approximately using SGD

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

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

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

10: **end for**

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

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

13: **end for**

Corollary 6 gives guarantees for Algorithm 6, and is similar to the i.i.d. case (Theorem 2).

Corollary 6. *Under the same assumptions as Theorem 2, in Algorithm 6, for the outer iterate θ_t we have*

$$\mathbb{E} \left[\|\theta_t - \hat{\theta}\|_2^2 \right] \lesssim t^{-d_o}, \quad (6.3)$$

$$\mathbb{E} \left[\|\theta_t - \hat{\theta}\|_2^4 \right] \lesssim t^{-2d_o}. \quad (6.4)$$

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

$$\mathbb{E} \left[\left\| \frac{\bar{g}_t}{\rho_t} - [\nabla^2 f(\theta_t)]^{-1} g_t^0 \right\|_2^2 \mid \theta_t \right] \lesssim \frac{1}{L} \|g_t^0\|_2^2, \quad (6.5)$$

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

$$\mathbb{E} \left[\left\| g_t^{j+1} - [\nabla^2 f(\theta_t)]^{-1} g_t^0 \right\|_2^4 \mid \theta_t \right] \lesssim (j+1)^{-2d_i} \|g_t^0\|_2^4. \quad (6.6)$$

After T steps of the outer loop, we have a non-asymptotic bound on the ‘‘covariance’’:

$$\mathbb{E} \left[\left\| H^{-1}GH^{-1} - \frac{S_o}{T} \sum_{t=1}^T \frac{\bar{g}_t \bar{g}_t^\top}{\rho_t^2} \right\|_2 \right] \lesssim T^{-\frac{d_o}{2}} + L^{-\frac{1}{2}}, \quad (6.7)$$

where $H = \nabla^2 f(\hat{\theta})$, and

$$G = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\hat{\theta}) f_i(\hat{\theta})^\top + \sum_{j=1}^{\mathbf{1}} w(j, \mathbf{1}) \sum_{i=j+1}^n \left(\nabla f_i(\hat{\theta}) \nabla f_{i-j}(\hat{\theta})^\top + \nabla f_{i-j}(\hat{\theta}) \nabla f_i(\hat{\theta})^\top \right), \quad (6.8)$$

with $w(j, \mathbf{1}) = 1 - \frac{j}{\mathbf{1}}$.

Also, in Algorithm 6’s outer loop, the average of consecutive iterates satisfies

$$\mathbb{E} \left[\left\| \frac{\sum_{t=1}^T \theta_t}{T} - \hat{\theta} \right\|_2^2 \right] \lesssim \frac{1}{T}, \quad (6.9)$$

$$\frac{1}{\sqrt{T}} \left(\frac{\sum_{t=1}^T \theta_t}{T} - \hat{\theta} \right) = W + \Delta, \quad (6.10)$$

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

Our approximate Newton time series statistical inference procedure estimates $H^{-1}GH^{-1}$, where G is the Newey-West covariance estimator (6.2) with weight

$$w(j, \mathbf{1}) = 1 - \frac{j}{\mathbf{1}}, \quad (6.11)$$

which is because when we estimate the variance in Algorithm 6, for $j > 0$, terms $\nabla f_i \nabla f_{i+j}^\top$ and $\nabla f_{i+j} \nabla f_i^\top$ appear $\mathbf{1} - j$ times, and the term $\nabla f_i \nabla f_i^\top$ appears $\mathbf{1}$ times. Note that the connection between sampling scheme and Newey-West estimator was also observed in

[Kun89]. Thus, our stochastic approximate Newton statistical inference procedure for time series analysis has similar statistical properties compared circular bootstrap [PR92, PR94].

Because expectation of the stochastic gradient in line 5 of Algorithm 6 is the full gradient $\frac{1}{n} \sum_{i=1}^n f_i(\hat{\theta})$, we have the same optimization guarantees as the i.i.d. case (Corollary 3).

6.3 Experiments

6.3.1 Synthetic data

In our linear regression simulation, we generate i.i.d. random explanatory variables, and the observation noise is a 0-mean moving average (MA) process independent of the explanatory variables. Results on average 95% confidence interval coverage and length are given in Appendix 4.1.1, and they validate our theory.

6.3.2 Real data

Using monthly equities returns data from [FP14], we use our approximate Newton statistical inference procedure to show that the correlation between US equities market returns and non-US global equities market returns is statistically significant, which validates the capital asset pricing model (CAPM) [Sha64, Lin65, FF04]. The details are given in Appendix 4.1.2.

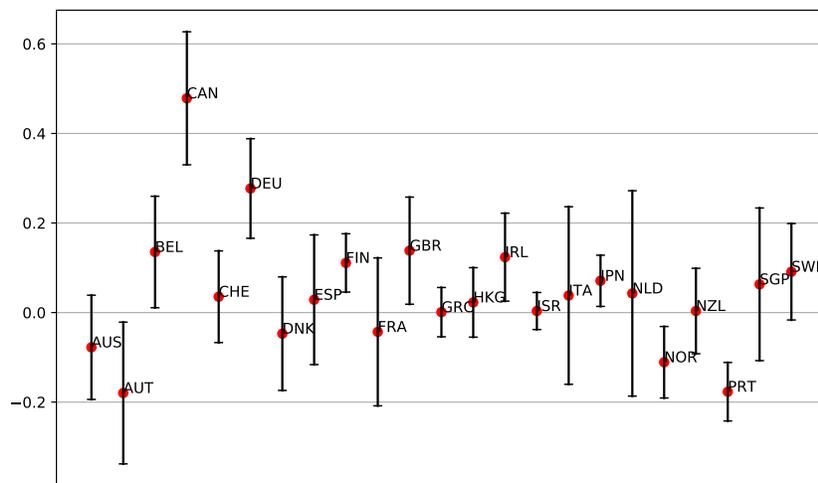


Figure 6.1: Exposure of US equities market to equities markets of other countries

Chapter 7

An intuitive view of SVRG as approximate stochastic Newton descent

Here we present an intuitive view of stochastic variance reduced gradient (SVRG) [JZ13] as approximate stochastic Newton descent, which is the inspiration behind our work.

Gradient descent solves the optimization problem $\hat{\theta} = \arg \min_{\theta} f(\theta)$, 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
 - $d \leftarrow d - \eta(\nabla f_k(\theta_t + d) - \nabla f_k(\theta_t) + g)$
- $\theta_{t+1} = \theta_t + d$

This chapter also appears in [LKLC18]. It was written by Tianyang Li, and edited by Anastasios Kyrillidis and Constantine Caramanis.

SVRG [JZ13] 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 [Bub15a] that the Newton direction is exactly the solution of

$$\min_d \langle \nabla f(\theta), d \rangle + \frac{1}{2}d^\top [\nabla^2 f(\theta)]d. \quad (7.1)$$

Next, let's consider solving (7.1) 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 $H^{-1} = \sum_{i=0}^{\infty} (I - \eta H)^i$, which corresponds to unrolling the gradient descent recursion for the optimization problem $H^{-1} = \arg \min_{\Omega} \text{Tr} \left(\frac{1}{2} \Omega^\top H \Omega - \Omega \right)$.

Chapter 8

Chapter 3 proofs

8.1 Proof of Theorem 1

We first assume that $\theta_1 = \hat{\theta}$. For ease of notation, we denote

$$\Delta_t = \theta_t - \hat{\theta}, \tag{8.1}$$

and, without loss of generality, we assume that $\hat{\theta} = 0$. 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, \dots is a martingale difference sequence. We use

$$g_i = \nabla f_i(\hat{\theta}) \quad \text{and} \quad H_i = \nabla^2 f_i(\hat{\theta}) \tag{8.2}$$

to denote the gradient component at index i , and the Hessian component at index i , at optimum $\hat{\theta}$, respectively. Note that $\sum g_i = 0$ and $\frac{1}{n} \sum H_i = H$.

For each f_i , its Taylor expansion around $\hat{\theta}$ is

$$f_i(\theta) = f_i(\hat{\theta}) + g_i^\top (\theta - \hat{\theta}) + \frac{1}{2} (\theta - \hat{\theta})^\top H_i (\theta - \hat{\theta}) + R_i(\theta, \hat{\theta}), \tag{8.3}$$

This chapter also appears in [\[LLKC18\]](#). It was written by Tianyang Li, and edited by Anastasios Kyrillidis and Constantine Caramanis.

where $R_i(\theta, \hat{\theta})$ 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 $t \rightarrow \infty$ and $\eta \rightarrow 0$.

Lemma 2. *For data dependent, positive constants α, A, B according to assumptions (F_1) and (G_2) in Theorem 1, and given assumption (G_1) , we have*

$$\mathbb{E} [\|\Delta_t\|_2^2] \leq (1 - 2\alpha\eta + A\eta^2)^{t-1} \|\Delta_1\|_2^2 + \frac{B\eta}{2\alpha - A\eta}, \quad (8.4)$$

under the assumption $\eta < \frac{2\alpha}{A}$.

Proof. As already stated, we assume without loss of generality that $\hat{\theta} = 0$. This further implies that: $g_s(\theta_t) = g_s(\theta_t - \hat{\theta}) = g_s(\Delta_t)$, 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:

$$\begin{aligned} \mathbb{E} [\|\Delta_{t+1}\|_2^2 \mid \Delta_t] &= \mathbb{E} [\|\Delta_t - \eta g_s(\Delta_t)\|_2^2 \mid \Delta_t] \\ &= \mathbb{E} [\|\Delta_t\|_2^2 \mid \Delta_t] + \eta^2 \cdot \mathbb{E} [\|g_s(\Delta_t)\|_2^2 \mid \Delta_t] - 2\eta \cdot \mathbb{E} [g_s(\Delta_t)^\top \Delta_t \mid \Delta_t] \\ &= \|\Delta_t\|_2^2 + \eta^2 \cdot \mathbb{E} [\|g_s(\Delta_t)\|_2^2 \mid \Delta_t] - 2\eta \cdot \nabla f(\Delta_t)^\top \Delta_t \\ &\stackrel{(i)}{\leq} \|\Delta_t\|_2^2 + \eta^2 \cdot (A \cdot \|\Delta_t\|_2^2 + B) - 2\eta \cdot \alpha \|\Delta_t\|_2^2 \\ &= (1 - 2\alpha\eta + A\eta^2) \|\Delta_t\|_2^2 + \eta^2 B. \end{aligned} \quad (8.5)$$

where (i) is due to assumptions (F_1) and (G_2) of Theorem 1. Taking expectations for every step $t = 1, \dots$ over the whole history, we obtain the recursion:

$$\mathbb{E} [\|\Delta_{t+1}\|_2^2] \leq (1 - 2\alpha\eta + A\eta^2)^{t-1} \|\Delta_1\|_2^2 + \eta^2 B \cdot \sum_{i=0}^{t-1} (1 - 2\alpha\eta + A\eta^2)^i$$

$$\begin{aligned}
&= (1 - 2\alpha\eta + A\eta^2)^{t-1} \|\Delta_1\|_2^2 + \eta^2 B \cdot \frac{1 - (1 - 2\alpha\eta + A\eta^2)^t}{2\alpha\eta - A\eta^2} \\
&\leq (1 - 2\alpha\eta + A\eta^2)^{t-1} \|\Delta_1\|_2^2 + \frac{\eta B}{2\alpha - A\eta}.
\end{aligned}$$

□

The following lemma states that $\mathbb{E}[\|\Delta_t\|_2^4] = O(\eta^2)$ as $t \rightarrow \infty$ and $\eta \rightarrow 0$.

Lemma 3. *For data dependent, positive constants α, A, B, C, D according to assumptions $(F_1), (G_1), (G_2)$ in Theorem 1, we have:*

$$\begin{aligned}
\mathbb{E}[\|\Delta_t\|_2^4] &\leq (1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + B(3\eta + \eta^2) + C(2\eta^3 + \eta^4))^{t-1} \|\Delta_1\|_2^4 \\
&\quad + \frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - B(3 + \eta) - C(2\eta^2 + \eta^3)}. \tag{8.6}
\end{aligned}$$

Proof. Given Δ_t , we have the following sets of (in)equalities:

$$\begin{aligned}
&\mathbb{E}[\|\Delta_{t+1}\|_2^4 \mid \Delta_t] \\
&= \mathbb{E}[\|\Delta_t - \eta g_s(\Delta_t)\|_2^4 \mid \Delta_t] \\
&= \mathbb{E}[(\|\Delta_t\|_2^2 - 2\eta \cdot g_s(\Delta_t)^\top \Delta_t + \eta^2 \|g_s(\Delta_t)\|_2^2)^2 \mid \Delta_t] \\
&= \mathbb{E}[\|\Delta_t\|_2^4 + 4\eta^2 (g_s(\Delta_t)^\top \Delta_t)^2 + \eta^4 \|g_s(\Delta_t)\|_2^4 - 4\eta \cdot g_s(\Delta_t)^\top \Delta_t \|\Delta_t\|_2^2 \\
&\quad + 2\eta^2 \cdot \|g_s(\Delta_t)\|_2^2 \|\Delta_t\|_2^2 - 4\eta^3 \cdot g_s(\Delta_t)^\top \Delta_t \|g_s(\Delta_t)\|_2^2 \mid \Delta_t] \\
&\stackrel{(i)}{\leq} \mathbb{E}[\|\Delta_t\|_2^4 + 4\eta^2 \cdot \|g_s(\Delta_t)\|_2^2 \cdot \|\Delta_t\|_2^2 + \eta^4 \|g_s(\Delta_t)\|_2^4 - 4\eta \cdot g_s(\Delta_t)^\top \Delta_t \|\Delta_t\|_2^2 \\
&\quad + 2\eta^2 \cdot \|g_s(\Delta_t)\|_2^2 \cdot \|\Delta_t\|_2^2 + 2\eta^3 \cdot (\|g_s(\Delta_t)\|_2^2 + \|\Delta_t\|_2^2) \cdot \|g_s(\Delta_t)\|_2^2 \mid \Delta_t] \\
&\stackrel{(ii)}{\leq} \mathbb{E}[\|\Delta_t\|_2^4 + (2\eta^3 + \eta^4) \|g_s(\Delta_t)\|_2^4 + (6\eta^2 + 2\eta^3) \|g_s(\Delta_t)\|_2^2 \|\Delta_t\|_2^2 \mid \Delta_t] - 4\alpha\eta \|\Delta_t\|_2^4 \\
&\stackrel{(iii)}{\leq} (1 - 4\alpha\eta) \|\Delta_t\|_2^4 + (6\eta^2 + 2\eta^3)(A\|\Delta_t\|_2^2 + B)\|\Delta_t\|_2^2 + (2\eta^3 + \eta^4)(C\|\Delta_t\|_2^4 + D) \\
&= (1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + C(2\eta^3 + \eta^4)) \|\Delta_t\|_2^4 + B(6\eta^2 + 2\eta^3) \|\Delta_t\|_2^2 + D(2\eta^3 + \eta^4) \\
&\stackrel{(iv)}{\leq} (1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + C(2\eta^3 + \eta^4)) \cdot \|\Delta_t\|_2^4 + B(3\eta + \eta^2)(\eta^2 + \|\Delta_t\|_2^4) + D(2\eta^3 + \eta^4) \\
&= (1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + B(3\eta + \eta^2)
\end{aligned}$$

$$+ C(2\eta^3 + \eta^4)) \cdot \|\Delta_t\|_2^4 + B\eta^2(3\eta + \eta^2) + D(2\eta^3 + \eta^4), \quad (8.7)$$

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:

$$\begin{aligned} & \mathbb{E} [\|\Delta_{t+1}\|_2^4] \\ & \leq (1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + B(3\eta + \eta^2) + C(2\eta^3 + \eta^4))^{t-1} \|\Delta_1\|_2^4 \\ & \quad + (B\eta^2(3\eta + \eta^2) + D(2\eta^3 + \eta^4)) \cdot \sum_{i=0}^{t-1} (1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + B(3\eta + \eta^2) + C(2\eta^3 + \eta^4))^i \\ & \leq (1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + B(3\eta + \eta^2) + C(2\eta^3 + \eta^4))^{t-1} \|\Delta_1\|_2^4 \\ & \quad + \frac{B\eta^2(3\eta + \eta^2) + D(2\eta^3 + \eta^4)}{4\alpha\eta - A(6\eta^2 + 2\eta^3) - B(3\eta + \eta^2) - C(2\eta^3 + \eta^4)}, \end{aligned}$$

which is the target inequality, after simple transformations. \square

We know that:

$$\Delta_t = \Delta_{t-1} - \eta g_s(\Delta_{t-1})$$

Using the Taylor expansion formula around the point Δ_{t-1} and using the assumption that $\hat{\theta} = 0$, we have:

$$f(\Delta_{t-1}) = f(\hat{\theta}) + \nabla f(\hat{\theta})^\top \Delta_{t-1} + \frac{1}{2} \Delta_{t-1}^\top H \Delta_{t-1} + R(\Delta_{t-1})$$

Taking further the gradient w.r.t. Δ_{t-1} in the above expression, we have:

$$\nabla f(\Delta_{t-1}) = H \Delta_{t-1} + \nabla R(\Delta_{t-1})$$

Using the identity $g_s(\Delta_{t-1}) = \nabla f(\Delta_{t-1}) + e_{t-1}$, our SGD recursion can be re-written as:

$$\begin{aligned}\Delta_t &= (I - \eta H) \Delta_{t-1} - \eta (\nabla R(\Delta_{t-1}) + e_{t-1}) \\ &= (I - \eta H)^{t-1} \Delta_1 - \eta \sum_{i=1}^{t-1} (I - \eta H)^{t-1-i} (e_i + \nabla R(\Delta_i)).\end{aligned}\quad (8.8)$$

For $t \geq 2$ and since: $\bar{\theta} = \bar{\theta} - \hat{\theta} = \bar{\Delta}_t = \frac{1}{t} \sum_{i=1}^t (\theta_i - \hat{\theta}) = \frac{1}{t} \sum_{i=1}^t \Delta_i$, we get:

$$\begin{aligned}t(\bar{\theta} - \hat{\theta}) &= \sum_{i=1}^t \Delta_i = \sum_{i=1}^t (I - \eta H)^{i-1} \Delta_1 - \eta \sum_{j=1}^t \sum_{i=1}^j (I - \eta H)^{j-1-i} (e_i + \nabla R(\Delta_i)) \\ &\stackrel{(i)}{=} (I - (I - \eta H)^t) \frac{H^{-1}}{\eta} \Delta_1 - \eta \sum_{j=1}^t \sum_{i=1}^j (I - \eta H)^{j-1-i} (e_i + \nabla R(\Delta_i)).\end{aligned}\quad (8.9)$$

where (i) holds due to the assumption that the eigenvalues of $I - \eta H$ satisfy $|\lambda_i(I - \eta H)| < 1$, and thus, the geometric series of matrices: $\sum_{k=0}^{n-1} T^k = (I - T)^{-1}(I - T^n)$, is utilized above. In our case, $T = (I - \eta H)$.

For the latter term in (8.9), using a variant of Abel's sum formula, we have:

$$\begin{aligned}&\eta \sum_{j=1}^t \sum_{i=1}^j (I - \eta H)^{j-1-i} (e_i + \nabla R(\Delta_i)) \\ &= \eta \sum_{j=0}^{t-1} \sum_{i=1}^j (I - \eta H)^{j-i} (e_i + \nabla R(\Delta_i)) \\ &= \eta \sum_{i=1}^{t-1} \left(\sum_{j=0}^{t-i-1} (I - \eta H)^j \right) (e_i + \nabla R(\Delta_i)) \\ &= \sum_{i=1}^{t-1} (I - (I - \eta H)^{t-i}) H^{-1} (e_i + \nabla R(\Delta_i)) \\ &= H^{-1} \sum_{i=1}^{t-1} e_i + H^{-1} \sum_{i=1}^{t-1} \nabla R(\Delta_i) - H^{-1} \sum_{i=1}^{t-1} (I - \eta H)^{t-i} (e_i + \nabla R(\Delta_i)) \\ &\stackrel{(i)}{=} H^{-1} \sum_{i=1}^{t-1} e_i + H^{-1} \sum_{i=1}^{t-1} \nabla R(\Delta_i) + \frac{H^{-1}}{\eta} (I - \eta H) (\Delta_t - (I - \eta H)^{t-1} \Delta_1),\end{aligned}\quad (8.10)$$

where (i) follows from the fact $\sum_{i=1}^{t-1} (I - \eta H)^{t-i} (e_i + \nabla R(\Delta_i)) = (I - \eta H)^{\frac{1}{\eta}} (\Delta_t - (I - \eta H)^{t-1} \Delta_1)$, based on the expression (8.8).

The above combined lead to:

$$\begin{aligned}
& \sqrt{t} \bar{\Delta}_t \\
&= \underbrace{\frac{1}{\sqrt{t}} (I - (I - \eta H)^t) \frac{H^{-1}}{\eta} \Delta_1}_{\varphi_1} \\
&\quad - \underbrace{\frac{1}{\sqrt{t}} H^{-1} \sum_{i=1}^{t-1} e_i}_{\varphi_2} \\
&\quad - \underbrace{\frac{1}{\sqrt{t}} H^{-1} \sum_{i=1}^{t-1} \nabla R(\Delta_i)}_{\varphi_3} \\
&\quad - \underbrace{\frac{1}{\sqrt{t}} \frac{H^{-1}}{\eta} (I - \eta H) (\Delta_t - (I - \eta H)^{t-1} \Delta_1)}_{\varphi_4}. \tag{8.12}
\end{aligned}$$

For the main result of the theorem, we are interested in the following quantity:

$$\left\| t \mathbb{E}[(\bar{\theta}_t - \hat{\theta})(\bar{\theta}_t - \hat{\theta})^\top] - H^{-1} G H^{-1} \right\|_2 = \left\| t \mathbb{E}[\bar{\Delta}_t \bar{\Delta}_t^\top] - H^{-1} G H^{-1} \right\|_2$$

Using the φ_i notation, we have $\mathbb{E}[t \bar{\Delta}_t \bar{\Delta}_t^\top] = \mathbb{E}[(\varphi_1 + \varphi_2 + \varphi_3 + \varphi_4)(\varphi_1 + \varphi_2 + \varphi_3 + \varphi_4)^\top]$.

Thus, we need to bound:

$$\begin{aligned}
& \left\| t \mathbb{E}[(\bar{\theta}_t - \hat{\theta})(\bar{\theta}_t - \hat{\theta})^\top] - H^{-1} G H^{-1} \right\|_2 \\
&= \left\| \mathbb{E}[(\varphi_1 + \varphi_2 + \varphi_3 + \varphi_4)(\varphi_1 + \varphi_2 + \varphi_3 + \varphi_4)^\top] - H^{-1} G H^{-1} \right\|_2 \\
&= \left\| \mathbb{E}[\varphi_2 \varphi_2^\top] - H^{-1} G H^{-1} \right. \\
&\quad \left. + \mathbb{E}[\varphi_2(\varphi_1 + \varphi_4 + \varphi_3)^\top + (\varphi_1 + \varphi_4 + \varphi_3) \varphi_2^\top + (\varphi_1 + \varphi_4 + \varphi_3)(\varphi_1 + \varphi_4 + \varphi_3)^\top] \right\|_2 \\
&\leq \left\| \mathbb{E}[\varphi_2 \varphi_2^\top] - H^{-1} G H^{-1} \right\|_2 + \left\| \mathbb{E}[\varphi_2(\varphi_1 + \varphi_4 + \varphi_3)^\top] \right\|_2 + \left\| \mathbb{E}[(\varphi_1 + \varphi_4 + \varphi_3) \varphi_2^\top] \right\|_2 \\
&\quad + \left\| \mathbb{E}[(\varphi_1 + \varphi_4 + \varphi_3)(\varphi_1 + \varphi_4 + \varphi_3)^\top] \right\|_2
\end{aligned}$$

$$\begin{aligned}
& + \left\| \mathbb{E}[(\varphi_1 + \varphi_4 + \varphi_3)(\varphi_1 + \varphi_4 + \varphi_3)^\top] \right\|_2 \\
& \stackrel{(i)}{\lesssim} \left\| \mathbb{E}[\varphi_2 \varphi_2^\top] - H^{-1} G H^{-1} \right\|_2 + \sqrt{\mathbb{E}[\|\varphi_2\|_2^2] (\mathbb{E}[\|\varphi_1\|_2^2] + \mathbb{E}[\|\varphi_4\|_2^2] + \mathbb{E}[\|\varphi_3\|_2^2])} \\
& \quad + \mathbb{E}[\|\varphi_1\|_2^2] + \mathbb{E}[\|\varphi_4\|_2^2] + \mathbb{E}[\|\varphi_3\|_2^2]
\end{aligned} \tag{8.13}$$

where (i) is due to the successive use of the AM-GM rule:

$$\left\| \mathbb{E}[ab^\top] \right\|_2 \leq \sqrt{\mathbb{E}[\|a\|_2^2] \mathbb{E}[\|b\|_2^2]} \leq \frac{1}{2} \mathbb{E}[\|a\|_2^2] + \mathbb{E}[\|b\|_2^2]. \tag{8.14}$$

for two p -dimensional random vectors a and b . Indeed, for any fixed unit vector u we have $\|\mathbb{E}[ab^\top]u\|_2 = \|\mathbb{E}[a(b^\top u)]\|_2 \leq \mathbb{E}[\|a\|_2 |b^\top u|] \leq \mathbb{E}[\|a\|_2 \|b\|_2] \leq \sqrt{\mathbb{E}[\|a\|_2^2] \mathbb{E}[\|b\|_2^2]}$. We used the fact $\|\mathbb{E}[x]\|_2 \leq \mathbb{E}[\|x\|_2]$ because $\|x\|_2$ is convex. Here also, the \lesssim hides any constants appearing from applying successively the above rule.

Therefore, to proceed bounding the quantity of interest, we need to bound the terms $\mathbb{E}[\|\varphi_i\|_2^2]$. 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 each of the above φ_i terms we have the following.

$$\varphi_1 := \frac{1}{\sqrt{t}} (I - (I - \eta H)^t) \frac{H^{-1}}{\eta} \Delta_1 = 0, \quad (\text{due to } \Delta_1 = 0) \tag{8.15}$$

$$\begin{aligned}
\mathbb{E}[\|\varphi_4\|_2^2] & := \mathbb{E} \left[\left\| -\frac{1}{\sqrt{t}} \frac{H^{-1}}{\eta} (I - \eta H) (\Delta_t - (I - \eta H)^{t-1} \Delta_1) \right\|_2^2 \right] \\
& \leq \mathbb{E} \left[\|H^{-1}\|_2^2 \cdot \|I - \eta H\|_2^2 \cdot \frac{1}{\eta^2 t} \|\Delta_t\|_2^2 \right] \stackrel{(i)}{\leq} \frac{1 - \eta \lambda_U}{\lambda_L} \cdot \frac{1}{\eta^2 t} \cdot \mathbb{E}[\|\Delta_t\|_2^2] \\
& \stackrel{(ii)}{\leq} \frac{1 - \eta \lambda_U}{\lambda_L} \frac{1}{\eta^2 t} \left((1 - 2\alpha\eta + A\eta^2)^{t-1} \|\Delta_1\|_2^2 + \frac{B\eta}{2\alpha - A\eta} \right) \\
& = \frac{1 - \eta \lambda_U}{\lambda_L} \frac{B}{t\eta(2\alpha - A\eta)} \\
& = O\left(\frac{1}{t\eta}\right)
\end{aligned} \tag{8.16}$$

where (i) is due to Assumption (F₄), (ii) is due to Lemma 2, and we used in several places the fact that $\Delta_1 = 0$.

$$\begin{aligned}
\mathbb{E}[\|\varphi_3\|_2^2] &:= \mathbb{E} \left[\left\| -\frac{1}{\sqrt{t}} H^{-1} \sum_{i=1}^{t-1} \nabla R(\Delta_i) \right\|_2^2 \right] \\
&\leq \mathbb{E} \left[\frac{1}{t} \cdot \|H^{-1}\|_2^2 \cdot \left\| \sum_{i=1}^{t-1} \nabla R(\Delta_i) \right\|_2^2 \right] \stackrel{(i)}{\leq} \mathbb{E} \left[\frac{1}{\lambda_L} \frac{1}{t} \left(\sum_{i=1}^{t-1} \|\nabla R(\Delta_i)\|_2 \right)^2 \right] \\
&\stackrel{(ii)}{\leq} \mathbb{E} \left[\frac{E^2}{\lambda_L \cdot t} \left(\sum_{i=1}^{t-1} \|\Delta_i\|_2^2 \right)^2 \right] \stackrel{(iii)}{\leq} \frac{E^2}{\lambda_L \cdot t} (t-1) \cdot \mathbb{E} \left[\sum_{i=1}^{t-1} \|\Delta_i\|_2^4 \right] \\
&\leq \frac{E^2}{\lambda_L \cdot t} (t-1) \sum_{i=1}^{t-1} \left((1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + C(2\eta^3 + \eta^4))^{t-1} \|\Delta_1\|_2^4 \right. \\
&\quad \left. + \frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - C(2\eta^2 + \eta^3)} \right) \\
&\stackrel{(iv)}{=} \frac{E^2}{\lambda_L} \frac{(t-1)^2}{t} \frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - C(2\eta^2 + \eta^3)} \stackrel{(v)}{=} O(t\eta^2). \tag{8.17}
\end{aligned}$$

where (i) is due to Assumption (F₄) and due to $|\sum_i \chi_i|^2 \leq \sum_i |\chi_i|^2$, (ii) is due to Assumption (F₃) on bounded remainder, (iii) is due to the inequality $(\sum_{i=1}^n \chi_i^2)^2 \leq n \cdot \sum_{i=1}^n \chi_i^2$, (iv) is due to $\Delta_1 = 0$, (v) is due to η being an small constant compared to α and thus $\frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - C(2\eta^2 + \eta^3)} = \frac{O(\eta^2)}{O(1)}$.

$$\begin{aligned}
\mathbb{E}[\|\varphi_2\|_2^2] &:= \mathbb{E} \left[\left\| -\frac{1}{\sqrt{t}} H^{-1} \sum_{i=1}^{t-1} e_i \right\|_2^2 \right] \stackrel{(i)}{=} \frac{1}{t} \sum_{i=1}^{t-1} \mathbb{E}[\|H^{-1} e_i\|_2^2] \stackrel{(ii)}{\leq} \frac{\lambda_U}{t} \sum_{i=1}^{t-1} \mathbb{E}[\|e_i\|_2^2] \\
&= \frac{\lambda_U}{t} \sum_{i=1}^{t-1} \mathbb{E}[\|g_s(\Delta_i) - \nabla f(\Delta_i)\|_2^2] \leq \frac{2\lambda_U}{t} \left(\sum_{i=1}^{t-1} \mathbb{E}[\|g_s(\Delta_i)\|_2^2] + \sum_{i=1}^{t-1} \mathbb{E}[\|\nabla f(\Delta_i)\|_2^2] \right) \\
&\stackrel{(iii)}{\leq} \frac{2\lambda_U}{t} \left((t-1)B + (A + L^2) \sum_{i=1}^{t-1} \mathbb{E}[\|\Delta_i\|_2^2] \right)
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(iv)}{\leq} \frac{2\lambda_U}{t} \left((t-1)B + (A+L^2) \sum_{i=1}^{t-1} \left((1-2\alpha\eta + A\eta^2)^{t-1} \|\Delta_1\|_2^2 + \frac{B\eta}{2\alpha - A\eta} \right) \right) \\
&= \frac{2\lambda_U(t-1)}{t} \left(B + (A+L^2) \frac{B\eta}{2\alpha - A\eta} \right) = O(1), \tag{8.18}
\end{aligned}$$

where (i) is due to $\mathbb{E}[(H^{-1}e_i)^\top H^{-1}e_j] = 0$ for $i \neq j$, (ii) is due to Assumption (F₄), (iii) is due to Assumptions (F₂) and (G₂), (iv) is due to Lemma 2.

Finally, for the term $\mathbb{E}[\varphi_2\varphi_2^\top]$, we have

$$\mathbb{E}[\varphi_2\varphi_2^\top] = \mathbb{E} \left[\left(-\frac{1}{\sqrt{t}} H^{-1} \sum_{i=1}^{t-1} e_i \right) \left(-\frac{1}{\sqrt{t}} H^{-1} \sum_{i=1}^{t-1} e_i \right)^\top \right] = \frac{1}{t} H^{-1} \left(\sum_{i=1}^{t-1} \mathbb{E}[e_i e_i^\top] \right) H^{-1}. \tag{8.19}$$

and thus:

$$\begin{aligned}
\|\mathbb{E}[\varphi_2\varphi_2^\top] - H^{-1}GH^{-1}\|_2 &= \left\| \frac{1}{t} H^{-1} \left(\sum_{i=1}^{t-1} \mathbb{E}[e_i e_i^\top] \right) H^{-1} - H^{-1}GH^{-1} \right\|_2 \\
&= \left\| \frac{1}{t} H^{-1} \left(\sum_{i=1}^{t-1} \mathbb{E}[e_i e_i^\top] - G + G \right) H^{-1} - H^{-1}GH^{-1} \right\|_2 \\
&= \left\| \frac{1}{t} H^{-1} \left(\sum_{i=1}^{t-1} \mathbb{E}[e_i e_i^\top] - G \right) H^{-1} - \frac{t-1}{t} \cdot H^{-1}GH^{-1} \right\|_2 \\
&\leq \frac{1}{t} H^{-1} \left(\sum_{i=1}^{t-1} \|\mathbb{E}[e_i e_i^\top] - G\|_2 \right) H^{-1} + \frac{t-1}{t} \|H^{-1}GH^{-1}\|_2
\end{aligned}$$

For each term $\|\mathbb{E}[e_i e_i^\top] - G\|_2, \forall i$, we have

$$\begin{aligned}
&\|\mathbb{E}[e_i e_i^\top] - G\|_2 \\
&= \left\| \mathbb{E}[g_s(\Delta_i)g_s(\Delta_i)^\top] - \mathbb{E}[\nabla f(\Delta_i)\nabla f(\Delta_i)^\top] - G \right\|_2 \\
&= \left\| \mathbb{E}[(g_s(\Delta_i) - \nabla f(\Delta_i))(g_s(\Delta_i) - \nabla f(\Delta_i))^\top] - G \right\|_2 \\
&= \left\| \mathbb{E}[g_s(\Delta_i)g_s(\Delta_i)^\top] - \mathbb{E}[g_s(\Delta_i)\nabla f(\Delta_i)^\top] - \mathbb{E}[\nabla f(\Delta_i)g_s(\Delta_i)^\top] + \mathbb{E}[\nabla f(\Delta_i)\nabla f(\Delta_i)^\top] - G \right\|_2
\end{aligned}$$

$$\begin{aligned}
& \stackrel{(i)}{=} \left\| \mathbb{E}[g_s(\Delta_i)g_s(\Delta_i)^\top] - 2\mathbb{E}[\nabla f(\Delta_i)\nabla f(\Delta_i)^\top] + \mathbb{E}[\nabla f(\Delta_i)\nabla f(\Delta_i)^\top] - G \right\|_2 \\
& \stackrel{(ii)}{\leq} \mathbb{E}[\|\nabla f(\Delta_i)\|_2^2] + \mathbb{E} \left[A_1\|\Delta_i\|_2 + A_2\|\Delta_i\|_2^2 + A_3\|\Delta_i\|_2^3 + A_4\|\Delta_i\|_2^4 \right] \\
& \stackrel{(iii)}{\leq} L^2\mathbb{E}[\|\Delta_i\|_2^2] + A_1\sqrt{\mathbb{E}[\|\Delta_i\|_2^2]} + A_2\mathbb{E}[\|\Delta_i\|_2^2] + \frac{A_3}{2}\mathbb{E}[\|\Delta_i\|_2^2 + \|\Delta_i\|_2^4] + A_4\mathbb{E}[\|\Delta_i\|_2^4] \\
& = A_1\sqrt{\mathbb{E}[\|\Delta_i\|_2^2]} + (L^2 + A_2 + \frac{A_3}{2})\mathbb{E}[\|\Delta_i\|_2^2] + (\frac{A_3}{2} + A_4)\mathbb{E}[\|\Delta_i\|_2^4] \\
& \stackrel{(iv)}{\leq} A_1\sqrt{(1 - 2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \frac{B\eta}{2\alpha - A\eta}} \\
& + \left(L^2 + A_2 + \frac{A_3}{2} \right) \left((1 - 2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \frac{B\eta}{2\alpha - A\eta} \right) \\
& + \left(\frac{A_3}{2} + A_4 \right) \left((1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + C(2\eta^3 + \eta^4))^{t-1}\|\Delta_1\|_2^4 + \frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - C(2\eta^2 + \eta^3)} \right) \\
& = A_1\sqrt{\frac{B\eta}{2\alpha - A\eta}} + \left(L^2 + A_2 + \frac{A_3}{2} \right) \frac{B\eta}{2\alpha - A\eta} + \left(\frac{A_3}{2} + A_4 \right) \frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - C(2\eta^2 + \eta^3)}. \tag{8.20}
\end{aligned}$$

where (i) is due to Assumption (G_1) , (ii) is due to Cauchy-Schwartz inequality and Assumption (G_4) , (iii) is due to Assumption (F_2) , (iv) is due to Lemmas 2-3.

Then, we have:

$$\begin{aligned}
& \left\| \mathbb{E}[\varphi_2\varphi_2^\top] - H^{-1}GH^{-1} \right\|_2 \\
& \stackrel{(i)}{\leq} \frac{t-1}{t} \left\| H^{-1}GH^{-1} \right\|_2 \\
& + \frac{t-1}{\lambda_L^2 t} \left(A_1\sqrt{\frac{B\eta}{2\alpha - A\eta}} + \left(L^2 + A_2 + \frac{A_3}{2} \right) \frac{B\eta}{2\alpha - A\eta} + \left(\frac{A_3}{2} + A_4 \right) \frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - C(2\eta^2 + \eta^3)} \right) \\
& = O(\sqrt{\eta}). \tag{8.21}
\end{aligned}$$

where (i) is due to Assumption, and (ii) is after removing constants and observing that the dominant term in the second part is $O(\sqrt{\eta})$. Combining all the above in (8.13), we obtain:

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

■

8.2 Proof of Corollary 2

Proof of Corollary 2. Here we use the same notations as the proof of Theorem 1. Because linear regression satisfies $\nabla f(\theta) - H(\theta - \hat{\theta}) = 0$, we do not have to consider the Taylor remainder term in our analysis. And we do not need 4-th order bound for SGD. Due to the fact that the quadratic function is strongly convex, we have $\Delta^\top \nabla f(\Delta + \hat{\theta}) \geq \lambda_L \|\Delta\|_2^2$.

By sampling with replacement, we have

$$\begin{aligned}
\mathbb{E}[\|g_s(\theta_t)\|_2^2 \mid \theta_t] &= \|\nabla f(\theta_t)\|_2^2 + \mathbb{E}[\|e_t\|_2^2 \mid \theta_t] \\
&= \|\nabla f(\theta_t)\|_2^2 + \frac{1}{S} \left(\frac{1}{n} \sum \|\nabla f_i(\theta_t)\|_2^2 - \|\nabla f(\theta_t)\|_2^2 \right) \\
&\leq L^2 \left(1 - \frac{1}{S}\right) \|\Delta_t\|_2^2 + \frac{1}{S} \frac{1}{n} \sum \|x_i(x_i^\top \theta_t - y_i)\|_2^2 \\
&= L^2 \left(1 - \frac{1}{S}\right) \|\Delta_t\|_2^2 + \frac{1}{S} \frac{1}{n} \sum \|x_i x_i^\top \Delta_t + x_i x_i^\top \hat{\theta} - y_i x_i\|_2^2 \\
&\leq L^2 \left(1 - \frac{1}{S}\right) \|\Delta_t\|_2^2 + 2 \frac{1}{S} \frac{1}{n} \sum (\|x_i x_i^\top \Delta_t\|_2^2 + \|x_i x_i^\top \hat{\theta} - y_i x_i\|_2^2) \\
&\leq \left(L^2 \left(1 - \frac{1}{S}\right) + 2 \frac{1}{S} \frac{1}{n} \sum \|x_i\|_2^4 \right) \|\Delta_t\|_2^2 + 2 \frac{1}{S} \frac{1}{n} \sum \|x_i x_i^\top \hat{\theta} - y_i x_i\|_2^2. \quad (8.22)
\end{aligned}$$

We also have

$$\begin{aligned}
\|\mathbb{E}[g_s(\theta)g_s(\theta)^\top \mid \theta] - G\|_2 &= \left\| \frac{1}{S} \frac{1}{n} \sum \nabla f_i(\theta) f_i(\theta)^\top - \nabla f(\theta) \nabla f(\theta)^\top - G \right\|_2 \\
&\leq \|\nabla f(\theta)\|_2^2 + \frac{1}{S} \left\| \frac{1}{n} \sum \nabla f_i(\theta) f_i(\theta)^\top - G \right\|_2 \\
&\leq \|\nabla f(\theta)\|_2^2 + \frac{1}{S} \left\| \frac{1}{n} \sum (g_i + H_i \Delta)(g_i + H_i \Delta)^\top - G \right\|_2 \\
&\leq \|\nabla f(\theta)\|_2^2 + \frac{1}{S} \left\| \frac{1}{n} \sum H_i \Delta g_i^\top + g_i \Delta^\top H_i + H_i \Delta \Delta^\top H_i \right\|_2 \\
&\leq \|\nabla f(\theta)\|_2^2 + \frac{1}{S} \left(\frac{2}{n} \|H_i\|_2 \|g_i\|_2 \right) \|\Delta\|_2 + \frac{1}{S} \left(\frac{1}{n} \sum \|H_i\|_2^2 \right) \|\Delta\|_2^2 \\
&\leq \frac{1}{S} \left(\frac{2}{n} \|H_i\|_2 \|g_i\|_2 \right) \|\Delta\|_2 + \left(L^2 + \frac{1}{S} \frac{1}{n} \sum \|H_i\|_2^2 \right) \|\Delta\|_2^2, \quad (8.23)
\end{aligned}$$

where $g_i = x_i(x_i^\top \hat{\theta} - y_i)$ and $H_i = x_i x_i^\top$.

Following Theorem 1's proof, we have

$$\left\| t \mathbb{E}[(\bar{\theta}_t - \hat{\theta})(\bar{\theta}_t - \hat{\theta})^\top] - H^{-1} G H^{-1} \right\|_2 \lesssim \sqrt{\eta} + \frac{1}{\sqrt{t\eta}}. \quad (8.24)$$

□

8.3 Proof of Corollary 1

Proof of Corollary 1. 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 4.

$$\|\nabla f(\theta)\|_2 \leq L\|\Delta\|_2 \quad (8.25)$$

for some data dependent constant L .

Proof. First, because

$$\nabla k(\theta) = \frac{1}{n} \sum -\frac{-y_i x_i}{1 + \exp(y_i \theta^\top x_i)}, \quad (8.26)$$

we have

$$\|\nabla k(\theta)\|_2 \leq \frac{1}{n} \sum \|x_i\|_2. \quad (8.27)$$

Also, we have

$$\|\nabla^2 k(\theta)\|_2 = \left\| \frac{1}{n} \sum \frac{\exp(y_i \theta^\top x_i)}{(1 + \exp(y_i \theta^\top x_i))^2} x_i x_i^\top \right\|_2 \leq \frac{1}{4n} \sum \|x_i\|_2^2, \quad (8.28)$$

which implies

$$\|\nabla k(\theta)\|_2 \leq \frac{1}{4n} \sum \|x_i\|_2^2 \|\Delta\|_2. \quad (8.29)$$

Further:

$$k(\theta) = \frac{1}{n} \sum \log(1 + \exp(-y_i \Delta^\top x_i - y_i \widehat{\theta}^\top x_i))$$

$$\begin{aligned}
&\leq \frac{1}{n} \sum \log(1 + \exp(\|x_i\|_2 \|\Delta\|_2 - y_i \widehat{\theta}^\top x_i)) \\
&\stackrel{(i)}{\leq} \frac{1}{n} \sum (\log(1 + \exp(-y_i \widehat{\theta}^\top x_i)) + \|x_i\|_2 \|\Delta\|_2)
\end{aligned} \tag{8.30}$$

where step (i) follows from $\log(1 + \exp(a + b)) \leq \log(1 + e^b) + |a|$. Thus, we have

$$\begin{aligned}
\|\nabla f(\theta)\|_2 &= \|(k(\theta) + c)\nabla k(\theta)\|_2 \leq k(\theta)\|\nabla k(\theta)\|_2 + c\|\nabla k(\theta)\|_2 \\
&\leq \left(c + \frac{1}{n} \sum \log(1 + \exp(-y_i \widehat{\theta}^\top x_i))\right) \|\nabla k(\theta)\|_2 + \left(\frac{1}{n} \sum \|x_i\|_2\right)^2 \|\Delta\|_2,
\end{aligned} \tag{8.31}$$

and we can conclude that $\|\nabla f(\theta)\|_2 \leq L\|\Delta\|_2$ for some data dependent constant L . \square

Next, we show that $f(\theta)$ has a bounded Taylor remainder.

Lemma 5.

$$\|\nabla f(\theta) - H(\theta - \widehat{\theta})\|_2 \leq E\|\theta - \widehat{\theta}\|_2^2, \tag{8.32}$$

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 . \square

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 α .

Lemma 6.

$$\nabla f(\theta)^\top (\theta - \widehat{\theta}) \geq \alpha\|\theta - \widehat{\theta}\|_2^2, \tag{8.33}$$

for some data dependent constant α .

Proof. We know that

$$\nabla f(\theta)^\top \Delta = (k(\theta) + c)\nabla k(\theta)^\top \Delta. \quad (8.34)$$

First, notice that locally (when $\|\Delta\|_2 = O(\frac{\lambda_L}{E})$) we have

$$\nabla k(\theta)^\top \Delta \gtrsim \Delta^\top H \Delta \gtrsim \lambda_L \|\Delta\|_2^2, \quad (8.35)$$

because of the optimality condition. This lower bounds $\nabla f(\theta)^\top (\theta - \hat{\theta})$ 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

$$\begin{aligned} g(t) &= \nabla f(\hat{\theta} + ut)^\top ut \\ &= (k(\hat{\theta} + ut) + c)\nabla k(\hat{\theta} + ut)^\top ut \\ &= k(\hat{\theta} + ut)\nabla k(\hat{\theta} + ut)^\top ut + c\nabla k(\hat{\theta} + ut)^\top ut, \end{aligned} \quad (8.36)$$

where $u = \frac{\Delta}{\|\Delta\|_2}$. Because $k(\theta)$ is convex, $\nabla k(\hat{\theta} + ut)^\top u$ is an increasing function in t , thus we have $\nabla k(\hat{\theta} + ut)^\top u = \Omega(\frac{\lambda_L^2}{E})$ when $t = \Omega(\frac{\lambda_L}{E})$. And we can deduce $\nabla k(\hat{\theta} + ut)^\top ut = \Omega(\frac{\lambda_L^2}{E}t)$ when $t = \Omega(\frac{\lambda_L}{E})$.

Similarly, because $k(\theta)$ is convex, $k(\hat{\theta} + ut)$ is an increasing function in t . Its derivative $\nabla k(\hat{\theta} + ut)^\top u = \Omega(\frac{\lambda_L^2}{E})$ when $t = \Omega(\frac{\lambda_L}{E})$. So we have $k(\hat{\theta} + ut) = \Omega(\frac{\lambda_L^2}{E}t)$ when $t = \Omega(\frac{\lambda_L}{E})$.

Thus, we have

$$k(\hat{\theta} + ut)\nabla k(\hat{\theta} + ut)^\top ut = \Omega\left(\frac{\lambda_L^4}{E^2}t^2\right), \quad (8.37)$$

when $t = \Omega(\frac{E}{\lambda_L})$.

And we can conclude that $\nabla f(\theta)^\top (\theta - \hat{\theta}) \geq \alpha \|\theta - \hat{\theta}\|_2^2$ for some data dependent constant $\alpha = \Omega(\min\{\lambda_L, \frac{\lambda_L^4}{E^2}\})$. \square

Next, we will prove properties about $g_s = \Psi_s \Upsilon_s$.

$$\mathbb{E}[\|\Upsilon\|_2^2 \mid \theta] = \frac{1}{S_\Upsilon} \left(\frac{1}{n} \sum \|\nabla k_i(\theta)\|_2^2 - \|\nabla k(\theta)\|_2^2 \right) + \|\nabla k(\theta)\|_2^2 \lesssim \frac{1}{n} \|x_i\|_2^2 \quad (8.38)$$

$$\begin{aligned} \mathbb{E}[\Psi_s^2] &\stackrel{(i)}{\leq} \frac{1}{n} \sum (c + k_i(\theta))^2 \\ &= \frac{1}{n} \sum \left(c + \log(1 + \exp(-y_i \widehat{\theta}^\top x_i - y_i \Delta x_i)) \right)^2 \\ &\stackrel{(ii)}{\lesssim} \frac{1}{n} \sum \|x_i\|_2^2 \|\Delta\|_2^2 + \frac{1}{n} \sum (c + \log(1 + \exp(-y_i \widehat{\theta}^\top x_i)))^2, \end{aligned} \quad (8.39)$$

where (i) follows from $\mathbb{E} \left[\left(\frac{\sum_{j=1}^S X_j}{S} \right)^2 \right] \leq \mathbb{E} \left[\frac{\sum_{j=1}^S X_j^2}{S} \right]$ and (ii) follows from $\log(1 + \exp(a + b)) \leq \log(1 + e^b) + |a|$.

Thus, we have

$$\mathbb{E}[\|g_s\|_2^2(\theta) \mid \theta] = \mathbb{E}[\Psi^2 \mid \theta] \cdot \mathbb{E}[\|\Upsilon\|_2^2 \mid \theta] \lesssim A \|\Delta\|_2^2 + B \quad (8.40)$$

for some data dependent constants A and B .

For the fourth-moment quantities, we have:

$$\begin{aligned} \mathbb{E}[\|\Upsilon\|_2^4 \mid \theta] &= \mathbb{E} \left[\left\| \frac{1}{S_\Upsilon} \sum_{i \in I_\Upsilon^r} \nabla \log(1 + \exp(-y_i \theta^\top x_i)) \right\|_2^4 \right] \\ &\leq \mathbb{E} \left[\left(\frac{1}{S_\Upsilon} \sum_{i \in I_\Upsilon^r} \|\nabla \log(1 + \exp(-y_i \theta^\top x_i))\|_2 \right)^4 \right] \\ &\leq \mathbb{E} \left[\left(\frac{1}{S_\Upsilon} \sum_{i \in I_\Upsilon^r} \|x_i\|_2 \right)^4 \right] \leq \frac{1}{n} \sum \|x_i\|_2^4. \end{aligned} \quad (8.41)$$

$$\begin{aligned}
\mathbb{E}[\Psi_s^4] &\stackrel{(i)}{\leq} \frac{1}{n} \sum (c + k_i(\theta))^4 = \frac{1}{n} \sum \left(c + \log(1 + \exp(-y_i \widehat{\theta}^\top x_i - y_i \Delta x_i)) \right)^4 \\
&\stackrel{(ii)}{\lesssim} \frac{1}{n} \sum \|x_i\|^4 \|\Delta\|_2^4 + \frac{1}{n} \sum \left(c + \log(1 + \exp(-y_i \widehat{\theta}^\top x_i)) \right)^4, \tag{8.42}
\end{aligned}$$

where (i) follows from $\mathbb{E} \left[\left(\frac{\sum_{j=1}^S X_j}{S} \right)^4 \right] \leq \mathbb{E} \left[\frac{\sum_{j=1}^S X_j^4}{S} \right]$ and (ii) follows from $\log(1 + \exp(a + b)) \leq \log(1 + e^b) + |a|$.

Combining the above, we get:

$$\mathbb{E}[\|g_s\|_2^4(\theta) \mid \theta] = \mathbb{E}[\Psi^4 \mid \theta] \cdot \mathbb{E}[\|\Upsilon\|_2^4 \mid \theta] \lesssim C \|\Delta\|_2^4 + D, \tag{8.43}$$

for some data dependent constants C and D .

Finally, we need a bound for the quantity $\|\mathbb{E}[\nabla g_s(\theta) \nabla g_s(\theta)^\top] - G\|_2$. We observe:

$$\begin{aligned}
&\|\mathbb{E}[\nabla g_s(\theta) \nabla g_s(\theta)^\top] - G\|_2 \\
&\leq \left\| \frac{K_G(\theta)}{n} \sum \nabla k_i(\theta) \nabla k_i(\theta)^\top - \frac{K_G(\widehat{\theta})}{n} \sum \nabla k_i(\widehat{\theta}) \nabla k_i(\widehat{\theta})^\top \right\|_2 \\
&\leq \left\| \frac{K_G(\theta)}{n} \sum \nabla k_i(\theta) \nabla k_i(\theta)^\top - \frac{K_G(\theta)}{n} \sum \nabla k_i(\widehat{\theta}) \nabla k_i(\widehat{\theta})^\top \right. \\
&\quad \left. + \frac{K_G(\theta)}{n} \sum \nabla k_i(\widehat{\theta}) \nabla k_i(\widehat{\theta})^\top - \frac{K_G(\widehat{\theta})}{n} \sum \nabla k_i(\widehat{\theta}) \nabla k_i(\widehat{\theta})^\top \right\|_2 \\
&\leq \frac{K_G(\theta)}{n} \left\| \sum \left(\nabla k_i(\theta) \nabla k_i(\theta)^\top - \nabla k_i(\widehat{\theta}) \nabla k_i(\widehat{\theta})^\top \right) \right\|_2 \\
&\quad + |K_G(\theta) - K_G(\widehat{\theta})| \cdot \left\| \frac{1}{n} \sum \nabla k_i(\widehat{\theta}) \nabla k_i(\widehat{\theta})^\top \right\|_2. \tag{8.44}
\end{aligned}$$

Because

$$K_G(\theta) = O(1 + \|\Delta\|_2 + \|\Delta\|_2^2), \tag{8.45}$$

$$\frac{1}{n} \left\| \sum (\nabla k_i(\theta) \nabla k_i(\theta)^\top - \nabla k_i(\widehat{\theta}) \nabla k_i(\widehat{\theta})^\top) \right\|_2 = O(\|\Delta\|_2 + \|\Delta\|_2^2), \tag{8.46}$$

$$|K_G(\theta) - K_G(\widehat{\theta})| = O(\|\Delta\|_2 + \|\Delta\|_2^2), \tag{8.47}$$

we may conclude that

$$\|\mathbb{E}[g_s(\theta)g_s(\theta)^\top | \theta] - G\|_2 \leq A_1\|\theta - \widehat{\theta}\|_2 + A_2\|\theta - \widehat{\theta}\|_2^2 + A_3\|\theta - \widehat{\theta}\|_2^3 + A_4\|\theta - \widehat{\theta}\|_2^4, \quad (8.48)$$

for some data dependent constants A_1 , A_2 , A_3 , and A_4 .

Combining above results and using Theorem 1, we have

$$\left\| t\mathbb{E}[(\bar{\theta}_t - \widehat{\theta})(\bar{\theta}_t - \widehat{\theta})^\top] - H^{-1}GH^{-1} \right\|_2 \lesssim \sqrt{\eta} + \sqrt{\frac{1}{t\eta} + t\eta^2}. \quad (8.49)$$

□

Chapter 9

Chapter 4 proofs

9.1 Proof of Theorem 2

Given assumptions about strong convexity, Lipschitz gradient continuity and Hessian Lipschitz continuity in Theorem 2, 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:

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

This chapter also appears in [LKLC18]. It was written by Tianyang Li, and edited by Anastasios Kyrillidis and Constantine Caramanis.

9.1.1 Proof of (4.8)

We first prove (4.8); the proof is similar to standard SGD convergence proofs (e.g. [LLKC18, CLTZ16, PJ92]). For the rest of our discussion, we assume that

$$\delta_t^j \cdot \bar{h} \leq \delta_t^j \cdot \sqrt{h_2} \ll 1, \quad \forall t, j.$$

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

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

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

$$g_t^{j+1} - H_t^{-1} g_t^0 = g_t^j - H_t^{-1} g_t^0 - \tau_j \cdot \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0 - \tau_j e_t^j. \quad (9.1)$$

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

$$\begin{aligned} \mathbb{E} \left[\|g_t^{j+1} - H_t^{-1} g_t^0\|_2^2 \mid \theta_t \right] &= \mathbb{E} \left[\|g_t^j - H_t^{-1} g_t^0\|_2^2 - \tau_j \underbrace{\left\langle g_t^j - H_t^{-1} g_t^0, \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - g_t^0 \right\rangle}_{[1]} \right. \\ &\quad \left. + \tau_j^2 \underbrace{\left\| \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - g_t^0 \right\|_2^2}_{[2]} + \tau_j^2 \underbrace{\|e_t^j\|_2^2}_{[3]} \mid \theta_t \right]. \end{aligned} \quad (9.2)$$

For term [1], we have

$$\begin{aligned}
& \left\langle g_t^j - H_t^{-1}g_t^0, \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - g_t^0 \right\rangle \\
&= (g_t^j - H_t^{-1}g_t^0)^\top H_t (g_t^j - H_t^{-1}g_t^0) + \left\langle g_t^j - H_t^{-1}g_t^0, \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - H_t \right\rangle \\
&\geq (g_t^j - H_t^{-1}g_t^0)^\top H_t (g_t^j - H_t^{-1}g_t^0) - \left| \left\langle g_t^j - H_t^{-1}g_t^0, \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - H_t \right\rangle \right|
\end{aligned}$$

by Hessian approximation

$$\geq (g_t^j - H_t^{-1}g_t^0)^\top H_t (g_t^j - H_t^{-1}g_t^0) - \delta_t^j \cdot \bar{h} \cdot \|g_t^j - H_t^{-1}g_t^0\|_2 \cdot \|g_t^j\|_2$$

by AM-GM inequality

$$\begin{aligned}
&\geq (g_t^j - H_t^{-1}g_t^0)^\top H_t (g_t^j - H_t^{-1}g_t^0) - \frac{\delta_t^j \cdot \bar{h}}{2} \cdot \|g_t^j - H_t^{-1}g_t^0\|_2^2 - \frac{\delta_t^j \cdot \bar{h}}{2} \cdot \|g_t^j\|_2^2 \\
&= (g_t^j - H_t^{-1}g_t^0)^\top H_t (g_t^j - H_t^{-1}g_t^0) - \frac{\delta_t^j \cdot \bar{h}}{2} \cdot \|g_t^j - H_t^{-1}g_t^0\|_2^2 - \frac{\delta_t^j \cdot \bar{h}}{2} \cdot \|g_t^j - H_t^{-1}g_t^0 + H_t^{-1}g_t^0\|_2^2
\end{aligned}$$

$$\|x + u\|_2^2 \leq 2\|x\|_2^2 + 2\|u\|_2^2$$

$$\geq (g_t^j - H_t^{-1}g_t^0)^\top H_t (g_t^j - H_t^{-1}g_t^0) - \frac{3\delta_t^j \cdot \bar{h}}{2} \cdot \|g_t^j - H_t^{-1}g_t^0\|_2^2 - \delta_t^j \bar{h} \cdot \|H_t^{-1}g_t^0\|_2^2$$

by strong convexity

$$\geq (g_t^j - H_t^{-1}g_t^0)^\top H_t (g_t^j - H_t^{-1}g_t^0) - \frac{3\delta_t^j \cdot \bar{h}}{2} \cdot \|g_t^j - H_t^{-1}g_t^0\|_2^2 - \frac{\delta_t^j \bar{h}}{\alpha^2} \cdot \|g_t^0\|_2^2. \quad (9.3)$$

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:

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

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

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

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

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

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

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

$$\begin{aligned}
&\mathbb{E}_{I_i} \left[\|e_t^j\|_2^2 \mid g_t^j, \theta_t \right] \\
&\leq 6 \left(\frac{1}{S_i} \cdot \left(1 - \frac{S_i-1}{n-1} \right) \right) \left((\delta_t^j \bar{h})^2 + 2\bar{\beta}^2 + (\delta_t^j)^2 \bar{h}_2 + 2\bar{\beta}_2 \right) \cdot \left(\|g_t^j - H_t^{-1} g_t^0\|_2^2 + \frac{1}{\alpha^2} \cdot \|g_t^0\|_2^2 \right) \\
&\leq 18 \left(\frac{1}{S_i} \left(1 - \frac{S_i-1}{n-1} \right) \right) \cdot ((\delta_t^j)^2 \bar{h}_2 + \bar{\beta}_2) \cdot \left(\|g_t^j - H_t^{-1} g_t^0\|_2^2 + \frac{1}{\alpha^2} \|g_t^0\|_2^2 \right). \tag{9.4}
\end{aligned}$$

Combining all above, we have

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

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

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

we have

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

For term [4], let us consider the α strongly convex and β smooth quadratic function

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

who attains its minimum at $g = H_t^{-1}g_t^0$. Using a well known property of α strongly convex and β smooth functions (Lemma 10), we have

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

Thus, when we choose

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

we have

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

and we have

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

Next, we set

$$\tau_0 = \min \left\{ \frac{0.476}{\beta}, \frac{0.025 \cdot \alpha}{\frac{1}{S_i} \left(1 - \frac{S_i-1}{n-1} \right) \bar{\beta}_2} \right\}, \quad D_j = (j+1)^{-d_i}, \quad \tau_j = \tau_0 D_j, \quad (9.5)$$

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

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

$$+ 18.5D_j^2\tau_0^2 \left(\frac{1}{S_i} \left(1 - \frac{S_{i-1}}{n-1} \right) \right) \frac{\bar{\beta}_2}{\alpha^2} \cdot \|g_t^0\|_2^2.$$

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

$$\begin{aligned} \delta_t^j &= o \left(\min \left\{ 1, \frac{1}{h} \right\} \cdot \min \left\{ 1, \alpha, \min \left\{ 1, \tau_0^4 \left(\frac{\tau_j}{\tau_0} \right)^4 \right\} \frac{1}{S_i} \left(1 - \frac{S_{i-1}}{n-1} \right) \right\} \right) \cdot \delta_t^0 = o((j+1)^{-2}) \cdot \delta_t^0, \\ \delta_t^0 &= O(\rho_t^4) = o((t+1)^{-2}) = o(1). \end{aligned} \quad (9.6)$$

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

Note that:

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

Applying Lemma 11, we have:

$$\mathbb{E} \left[\|g_t^j - H_t^{-1}g_t^0\|_2^2 \mid \theta_t \right] = O(t^{-d_i} \cdot \|g_t^0\|_2^2), \quad (9.7)$$

where we have assumed that α, β, S_i , etc. are (data dependent) constants. Further, (9.7) implies:

$$\mathbb{E} \left[\|g_t^j\|_2^2 \right] \leq 2\mathbb{E} \left[\|g_t^j - H_t^{-1}g_t^0\|_2^2 + \|H_t^{-1}g_t^0\|_2^2 \mid \theta_t \right] \lesssim \|g_t^0\|_2^2, \quad \text{for all } j. \quad (9.8)$$

In Algorithm 1, we have

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

By unrolling the recursion we have:

$$g_t^{j+1} - H_t^{-1}g_t^0 = \sum_{k=0}^j \left(\prod_{l=k+1}^j (I - \tau_l H_t) \right) \cdot \tau_k \cdot \left(-e_t^k - \frac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k \right). \quad (9.9)$$

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

$$\begin{aligned}
\bar{g}_t - H_t^{-1}g_t^0 &= \frac{1}{L+1} \sum_{j=0}^L (g_t^j - H_t^{-1}g_t^0) \\
&= \frac{1}{L+1} \sum_{j=0}^L \sum_{k=0}^{j-1} \left(\prod_{l=k+1}^{j-1} (I - \tau_l H_t) \right) \cdot \tau_k \left(-e_t^k - \frac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k \right) \\
&= \frac{1}{L+1} \sum_{k=0}^{L-1} \underbrace{\tau_k \sum_{j=k+1}^L \prod_{l=k+1}^{j-1} (I - \tau_l H_t)}_{[5]} \left(-e_t^k - \frac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k \right) \\
&= \frac{1}{L+1} \underbrace{\sum_{k=0}^{L-1} \tau_k \sum_{j=k+1}^L \prod_{l=k+1}^{j-1} (I - \tau_l H_t)}_{[6]} (-e_t^k) \\
&\quad + \frac{1}{L+1} \underbrace{\sum_{k=0}^{L-1} \tau_k \sum_{j=k+1}^L \prod_{l=k+1}^{j-1} (I - \tau_l H_t)}_{[7]} \left(-\frac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k \right). \tag{9.10}
\end{aligned}$$

For the term [5], we have:

$$\begin{aligned}
&\left\| \tau_k \sum_{j=k+1}^L \prod_{l=k+1}^{j-1} (I - \tau_l H_t) \right\|_2 \\
&\leq \tau_k \sum_{j=k+1}^L \prod_{l=k+1}^{j-1} \|I - \tau_l H_t\|_2 \\
&I - \tau_l H_t \text{ is positive definite by our choice of } \tau_l \text{ (9.5) and } \|I - \tau_l H_t\|_2 \leq 1 - \tau_l \alpha \\
&\leq \tau_k \sum_{j=k+1}^L \prod_{l=k+1}^{j-1} (1 - \tau_l \alpha) \\
&\leq \tau_k \sum_{j=k+1}^L \prod_{l=k+1}^{j-1} \left(1 - \frac{1}{2} \tau_l \alpha \right)^2 \\
&\tau_k \prod_{l=k+1}^{j-1} \left(1 - \frac{1}{2} \tau_l \alpha \right) \leq \tau_k \exp\left(-\frac{1}{2} \alpha \sum_{l=k+1}^{j-1} \tau_l\right) \lesssim k^{-d_i} \exp(\Theta(-j^{1-d_i} + k^{1-d_i})) \lesssim j^{-d_i} \lesssim \tau_j
\end{aligned}$$

because for a fixed d_i $x^{-d_i} e^{\Theta(x^{1-d_i})}$ is an increasing function when x is sufficiently large

$$\begin{aligned}
&\lesssim \sum_{j=k+1}^L \tau_j \prod_{l=k+1}^{j-1} \left(1 - \frac{\tau_l \alpha}{2}\right) = \frac{2}{\alpha} \sum_{j=k+1}^L \frac{1}{2} \tau_j \alpha \prod_{l=k+1}^{j-1} \left(1 - \frac{\tau_l \alpha}{2}\right) \\
&= \frac{2}{\alpha} \left(1 - \prod_{j=k+1}^L \left(1 - \frac{\tau_j \alpha}{2}\right)\right) = O(1),
\end{aligned} \tag{9.11}$$

where we have assumed that α, β, S_i , etc. are (data-dependent) constants.

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

$$\begin{aligned}
&\mathbb{E} \left[\left\| \frac{1}{L+1} \sum_{k=0}^{L-1} \tau_k \sum_{j=k+1}^L \prod_{l=k+1}^{j-1} (I - \tau_l H_t)(-e_t^k) \right\|_2^2 \mid \theta_t \right] \\
&= \frac{1}{(L+1)^2} \mathbb{E} \left[\sum_{k=0}^{L-1} \left\| \tau_k \sum_{j=k+1}^L \prod_{l=k+1}^{j-1} (I - \tau_l H_t)(-e_t^k) \right\|_2^2 \mid \theta_t \right] \\
&\text{using (9.11)} \\
&\lesssim \frac{1}{(L+1)^2} \mathbb{E} \left[\sum_{k=0}^{L-1} \|e_t^k\|_2^2 \mid \theta_t \right] \\
&\text{using (9.4) and (9.7)} \\
&\lesssim \frac{1}{L} \|g_t^0\|_2^2.
\end{aligned} \tag{9.12}$$

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:

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

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

using (9.11) and Lemma 7

$$\lesssim \frac{1}{(L+1)^2} \mathbb{E} \left[\sum_{0 \leq a, b, \leq L-1} \delta_t^a \bar{h} \|g_t^a\|_2 \delta_t^b \bar{h} \|g_t^b\|_2 \mid \theta_t \right] \\ \leq \frac{2\bar{h}^2}{(L+1)^2} \sum_{0 \leq a, b, \leq L-1} \delta_t^a \delta_t^b \cdot \mathbb{E} [\|g_t^a\|_2^2 + \|g_t^b\|_2^2 \mid \theta_t] \\ \lesssim \frac{\|g_t^0\|_2^2}{(L+1)^2} \sum_{0 \leq a, b, \leq L-1} \delta_t^a \delta_t^b \lesssim \frac{\|g_t^0\|_2^2}{L^2} \left(\sum_{k=0}^L \delta_t^k \right)^2 \quad (9.13)$$

using (9.8) and our choice of δ_t^k (9.6)

$$\lesssim \frac{1}{L^2} \delta_t^{0^2} \left(\sum_{k=0}^L \tau_k \right)^2 \cdot \|g_t^0\|_2^2 \lesssim \frac{1}{L^2} \delta_t^{0^2} \left(\sum_{k=0}^L (k+1)^{-d_i} \right)^2 \cdot \|g_t^0\|_2^2 \\ \text{because } \left(\sum_{k=0}^L (k+1)^{-d_i} \right)^2 = O(L^{1-d_i}) \text{ and } d_i \in \left(\frac{1}{2}, 1 \right) \\ \ll \frac{1}{L} \|g_t^0\|_2^2. \quad (9.14)$$

Combining (9.12) and (9.14), we have

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

9.1.2 Proof of (4.9)

Using (9.1), we have

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

$$\begin{aligned}
&= \mathbb{E}[\|\mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0 - \tau_j \frac{\nabla f(\theta_t + \delta_t^j \mathbf{g}_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j \mathbf{g}_t^0\|_2^2 \\
&\quad - 2\langle \tau_j \mathbf{e}_t^j, \mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0 - \tau_j \frac{\nabla f(\theta_t + \delta_t^j \mathbf{g}_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j \mathbf{g}_t^0 \rangle + \tau_j^2 \|\mathbf{e}_t^j\|_2^2 \mid \mathbf{g}_t^j] \\
&= \mathbb{E}[\|\mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0 - \tau_j \frac{\nabla f(\theta_t + \delta_t^j \mathbf{g}_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j \mathbf{g}_t^0\|_2^4 \\
&\quad + 4(\langle \tau_j \mathbf{e}_t^j, \mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0 - \tau_j \frac{\nabla f(\theta_t + \delta_t^j \mathbf{g}_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j \mathbf{g}_t^0 \rangle)^2 + \tau_j^4 \|\mathbf{e}_t^j\|_2^4 \\
&\quad + 2\|\mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0 - \tau_j \frac{\nabla f(\theta_t + \delta_t^j \mathbf{g}_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j \mathbf{g}_t^0\|_2^2 \tau_j^2 \|\mathbf{e}_t^j\|_2^2 \\
&\quad - 4\langle \tau_j \mathbf{e}_t^j, \mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0 - \tau_j \frac{\nabla f(\theta_t + \delta_t^j \mathbf{g}_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j \mathbf{g}_t^0 \rangle \|\mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0 - \tau_j \frac{\nabla f(\theta_t + \delta_t^j \mathbf{g}_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j \mathbf{g}_t^0\|_2^2 \\
&\quad - 4\langle \tau_j \mathbf{e}_t^j, \mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0 - \tau_j \frac{\nabla f(\theta_t + \delta_t^j \mathbf{g}_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j \mathbf{g}_t^0 \rangle \tau_j^2 \|\mathbf{e}_t^j\|_2^2 \mid \mathbf{g}_t^j]. \tag{9.15}
\end{aligned}$$

Because we have

$$\mathbb{E}[\mathbf{e}_t^j \mid \mathbf{g}_t^j] = 0,$$

$$\begin{aligned}
&\|\mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0 - \tau_j \frac{\nabla f(\theta_t + \delta_t^j \mathbf{g}_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j \mathbf{g}_t^0\|_2^4 \\
&= \|(I - \tau_j H_t)(\mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0) + \tau_j \left(-\frac{\nabla f(\theta_t + \delta_t^j \mathbf{g}_t^j) - \nabla f(\theta_t)}{\delta_t^j} + H_t \mathbf{g}_t^j\right)\|_2^4 \\
&= (\|(I - \tau_j H_t)(\mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0)\|_2^2 \\
&\quad + 2\tau_j \langle (I - \tau_j H_t)(\mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0), -\frac{\nabla f(\theta_t + \delta_t^j \mathbf{g}_t^j) - \nabla f(\theta_t)}{\delta_t^j} + H_t \mathbf{g}_t^j \rangle \\
&\quad + \tau_j^2 \left\| -\frac{\nabla f(\theta_t + \delta_t^j \mathbf{g}_t^j) - \nabla f(\theta_t)}{\delta_t^j} + H_t \mathbf{g}_t^j \right\|_2^2)^2
\end{aligned}$$

using Lemma 7

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

by our choice of $\tau_j = \Theta((j+1)^{-d_i}) = o(1)$ (9.5)

and using $\|\mathbf{g}_t^j\|_2 \leq \|\mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0\|_2 + \|H_t^{-1}\mathbf{g}_t^0\|_2 \lesssim \|\mathbf{g}_t^j - H_t^{-1}\mathbf{g}_t^0\|_2 + \|\mathbf{g}_t^0\|_2$

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

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

using Lemma 7 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 $\tau_j = \Theta((j+1)^{-d_i}) = o(1)$ (9.5) and $\delta_t^j = O(\tau_j^4)$ (9.6), after repeatedly applying the AM-GM inequality, Lemma 7, triangle inequality, and (9.7), we can bound (9.15) by

$$\begin{aligned}
&\mathbb{E}[\|g_t^{j+1} - H_t^{-1} g_t^0\|_2^4 \mid g_t^j] \\
&\leq (1 - \Theta(\tau_j)) \|g_t^j - H_t^{-1} g_t^0\|_2^4 + O(\tau_j^3 \|g_t^0\|_2^4).
\end{aligned} \tag{9.16}$$

Applying Lemma 11, we have

$$\mathbb{E}[\|g_t^{j+1} - H_t^{-1} g_t^0\|_2^4 \mid \theta_t] = O((j+1)^{-2d_i} \|g_t^0\|_2^4), \tag{9.17}$$

and using the AM-GM in equality we have

$$\mathbb{E}[\|g_t^{j+1}\|_2^4 \mid \theta_t] = O(\|g_t^0\|_2^4). \tag{9.18}$$

9.1.3 Proof of (4.6)

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

Lemma 8.

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

Proof. Using (9.9), and because $\mathbb{E}[e_t^j \mid \theta_t = 0]$, we have

$$\begin{aligned} &\mathbb{E}[\langle \nabla f(\theta_t), -g_t^L \rangle \mid \theta_t] \\ &= \rho_t \nabla f(\theta_t)^\top H_t^{-1} \nabla f(\theta_t) - \mathbb{E} \left[\left\langle \nabla f(\theta_t), \sum_{k=0}^{L-1} (\prod_{l=k+1}^{L-1} (I - \tau_l H_t)) \tau_k \left(\frac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} - H_t g_t^k \right) \right\rangle \mid \theta_t \right] \\ &\text{using strong convexity and Lemma 7} \\ &\geq \frac{1}{\beta} \rho_t \|\nabla f(\theta_t)\|_2^2 - \|\nabla f(\theta_t)\|_2 \underbrace{\mathbb{E} \left[\sum_{k=0}^{L-1} \prod_{l=k+1}^{L-1} \|I - \tau_l H_t\|_2 \tau_k \delta_t^k \|g_t^k\|_2 \mid \theta_t \right]}_{[8]}. \end{aligned}$$

By our choice of $\tau_j = \Theta((j+1)^{-d_i}) = o(1)$ (9.5) and $\delta_t^j = O(\delta_t^0 \tau_j^4)$ (9.6), and using (9.8), 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

$$\mathbb{E}[\langle \nabla f(\theta_t), -g_t^L \rangle \mid \theta_t]$$

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

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

□

Now, we continue our proof of (4.6).

In Algorithm 1, because f is β smooth, we have

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1}) - f(\hat{\theta}) \mid \theta_t] \\
&= \mathbb{E}[f(\theta_t + g_t^L) - f(\hat{\theta}) \mid \theta_t] \\
&\leq f(\theta_t) - f(\hat{\theta}) + \mathbb{E} \left[\langle \nabla f(\theta_t), g_t^L \rangle + \frac{\beta}{2} \|g_t^L\|_2^2 \mid \theta_t \right] \\
&\quad \text{using Lemma 8 and (9.8)} \\
&\leq f(\theta_t) - f(\hat{\theta}) - \Omega(\rho_t \|\nabla f(\theta_t)\|_2^2) + \mathbb{E}[O(\|g_t^0\|_2^2 + \delta_t^0 \|g_t^0\|_2 \|\nabla f(\theta_t)\|_2) \mid \theta_t]. \tag{9.19}
\end{aligned}$$

For g_t^0 , we have

$$\begin{aligned}
\frac{g_t^0}{\rho_t} &= \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\theta_t) \\
&= \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\hat{\theta}) + \frac{1}{S_o} \sum_{i \in I_o} (\nabla f_i(\theta_t) - \nabla f_i(\hat{\theta})), \tag{9.20}
\end{aligned}$$

which implies that

$$\mathbb{E} \left[\left\| \frac{g_t^0}{\rho_t} \right\|_2^2 \mid \theta_t \right]$$

$$\begin{aligned}
&\leq 2\mathbb{E} \left[\left\| \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\hat{\theta}) \right\|_2^2 \mid \theta_t \right] + 2\mathbb{E} \left[\left\| \frac{1}{S_o} \sum_{i \in I_o} (\nabla f_i(\theta_t) - \nabla f_i(\hat{\theta})) \right\|_2^2 \mid \theta_t \right] \\
&\quad \text{because we sample uniformly with replacement and } \nabla f(\hat{\theta}) = 0 \\
&\leq \frac{2}{S_o} \sum_{i=1}^n \|\nabla f_i(\hat{\theta})\|_2^2 + \mathbb{E}[\|\nabla f_i(\theta_t) - \nabla f_i(\hat{\theta})\|_2^2 \mid \theta_t] \\
&\leq \frac{2}{S_o} \sum_{i=1}^n \|\nabla f_i(\hat{\theta})\|_2^2 + \|\theta_t - \hat{\theta}\|_2^2 \mathbb{E}[\beta_i^2 \mid \theta_t] \\
&\lesssim 1 + \|\theta_t - \hat{\theta}\|_2^2. \tag{9.21}
\end{aligned}$$

Thus, continuing (9.19), using (9.21) and strong convexity $\alpha^2 \|\theta_t - \hat{\theta}\|_2^2 \leq \|\nabla f(\theta_t)\|_2^2$, we have

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1}) - f(\hat{\theta}) \mid \theta_t] \\
&\leq f(\theta_t) - f(\hat{\theta}) - C_1 \rho_t \|\nabla f(\theta_t)\|_2^2 + C_2 \rho_t \delta_t^0 (1 + \|\nabla f(\theta_t)\|_2) \|\nabla f(\theta_t)\|_2 + C_3 \rho_t^2 (1 + \|\nabla f(\theta_t)\|_2^2) \\
&\quad = f(\theta_t) - f(\hat{\theta}) - \rho_t (C_1 - C_2 \delta_t^0 - C_3 \rho_t) \|\nabla f(\theta_t)\|_2^2 + C_3 \rho_t^2 + C_2 \rho_t \delta_t^0 \|\nabla f(\theta_t)\|_2 \\
&\quad \text{because we have } C_2 \rho_t \delta_t^0 \|\nabla f(\theta_t)\|_2 = \frac{1}{2} C_1 \rho_t \delta_t^0 \frac{C_2^2 \|\nabla f(\theta_t)\|_2}{\delta_t^0} \leq \frac{1}{2} C_1 \rho_t \delta_t^0 \left(\left(\frac{C_2}{C_1} \right)^2 + \left(\frac{\|\nabla f(\theta_t)\|_2}{\delta_t^0} \right)^2 \right) \\
&\leq f(\theta_t) - f(\hat{\theta}) - \rho_t \left(\frac{1}{2} C_1 - C_2 \delta_t^0 - C_3 \rho_t \right) \|\nabla f(\theta_t)\|_2^2 + C_3 \rho_t^2 + \frac{C_2^2}{C_1} \rho_t \delta_t^0 \\
&\quad \text{using strong convexity } \frac{1}{2\alpha} \|\nabla f(\theta_t)\|_2^2 \geq f(\theta_t) - f(\hat{\theta}) \text{ and smoothness } \frac{1}{2\beta} \|\nabla f(\theta_t)\|_2^2 \leq f(\theta_t) - f(\hat{\theta}) \\
&\leq [f(\theta_t) - f(\hat{\theta})] - \rho_t \left(\frac{1}{2} C_1 - C_2 \delta_t^0 - C_3 \rho_t \right) \frac{1}{2\alpha} [f(\theta_t) - f(\hat{\theta})] + C_3 \rho_t^2 + \frac{C_2^2}{C_1} \rho_t \delta_t^0 \\
&\quad \text{when we set } \delta_t^0 = O(\rho_t) \text{ in (9.6)} \\
&\leq [f(\theta_t) - f(\hat{\theta})] - \rho_t \left(\frac{1}{2} C_1 - C_2 \delta_t^0 - C_3 \rho_t \right) \frac{1}{2\alpha} [f(\theta_t) - f(\hat{\theta})] + (C_3 + O(1)) \rho_t^2, \tag{9.22}
\end{aligned}$$

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

In (9.22) we choose $\rho_t = \Theta((t+1)^{-d_o})$ for some $d_o \in (\frac{1}{2}, 1)$, and after applying Lemma 11 we have

$$\mathbb{E}[\|\theta_t - \hat{\theta}\|_2^2]$$

$$\begin{aligned}
&\leq \mathbb{E}\left[\frac{2}{\alpha}(f(\theta_t) - f(\widehat{\theta}))\right] \\
&\lesssim t^{-d_o} + e^{-\Theta(t^{1-d_o})} \|\theta_0 - \widehat{\theta}\|_2^2,
\end{aligned} \tag{9.23}$$

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

9.1.4 Proof of (4.7)

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

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

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 $\delta_t^0 = O(\rho_t^4)$ (9.6), after repeatedly applying the AM-GM inequality and (9.23), 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 11, we have

$$\begin{aligned} & \mathbb{E}[\|\theta_t - \widehat{\theta}\|_2^4] \\ & \leq \mathbb{E}\left[\frac{4}{\alpha^2}(f(\theta_t) - f(\widehat{\theta}))^2\right] \\ & \lesssim t^{-2d_o}. \end{aligned} \tag{9.24}$$

9.1.5 Proof of (4.10)

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

$$\begin{aligned} \frac{\bar{g}_t}{\rho_t} &= \underbrace{-H^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\widehat{\theta})}_{[1]} \\ &+ \underbrace{H^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\widehat{\theta}) - H_t^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\widehat{\theta}) + H_t^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\widehat{\theta}) - H_t^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\theta_t)}_{[2]} \\ &\underbrace{-H_t^{-1} \frac{g_t^0}{\rho_t} + \frac{\bar{g}_t}{\rho_t}}_{[3]}. \end{aligned} \tag{9.25}$$

Thus, for the ‘‘covariance’’ of our replicates, we have

$$\begin{aligned} & \left\| H^{-1} G H^{-1} - \frac{S_o}{T} \sum_{t=1}^T \frac{\bar{g}_t \bar{g}_t^\top}{\rho_t^2} \right\|_2 \\ & \lesssim \left\| H^{-1} G H^{-1} - \frac{S_o}{T} \sum_{t=1}^T [1]_t [1]_t^\top \right\|_2 \end{aligned}$$

$$\begin{aligned}
& + \left\| \frac{S_o}{T} \sum_{t=1}^T [\mathbf{1}]_t ([\mathbf{2}]_t + [\mathbf{3}]_t)^\top \right\|_2 + \left\| \frac{S_o}{T} \sum_{t=1}^T ([\mathbf{2}]_t + [\mathbf{3}]_t) [\mathbf{1}]_t^\top \right\|_2 \\
& + \left\| \frac{S_o}{T} \sum_{t=1}^T ([\mathbf{2}]_t + [\mathbf{3}]_t) ([\mathbf{2}]_t + [\mathbf{3}]_t)^\top \right\|_2
\end{aligned}$$

because for two vectors a, b the operator norm $\|ab^\top\|_2 \leq \|a\|_2 \|b\|_2$

$$\begin{aligned}
& \lesssim \left\| H^{-1}GH^{-1} - \frac{S_o}{T} \sum_{t=1}^T [\mathbf{1}]_t [\mathbf{1}]_t^\top \right\|_2 \\
& + \frac{1}{T} \sum_{t=1}^T \|\mathbf{1}\|_2 (\|\mathbf{2}\|_2 + \|\mathbf{3}\|_2) \\
& + \frac{1}{T} \sum_{t=1}^T (\|\mathbf{2}\|_2^2 + \|\mathbf{3}\|_2^2).
\end{aligned}$$

Because $\sum_{t=1}^T [\mathbf{1}]_t$ consists of $S_o \cdot T$ i.i.d. samples from $\{H^{-1}\nabla f_i(\theta)\}_{i=1}^n$ and the mean $H^{-1}\nabla f(\hat{\theta}) = 0$, using matrix concentration [Tro15], we know that

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

For term [3], using (9.10), because we have

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

by using (9.11) and (9.13), we have

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

$$\begin{aligned}
&\lesssim \mathbb{E} \left[\frac{1}{T} \left(\sum_{t=1}^T \frac{1}{L} + \left(\sum_{t=1}^T \frac{\sum_{k=0}^L \delta_t^k}{L} \right)^2 \right) \left\| \frac{g_t^0}{\rho_t} \right\|_2^2 \right] \\
&\quad \text{using (9.21), and by our choice of } \delta_t^k = \delta_t^0 o((k+1)^{-2}) \text{ and } \delta_t^0 = o((t+1)^{-2}) \text{ (9.6)} \\
&\lesssim \mathbb{E} \left[\left(\frac{1}{L} + \frac{\sum_{t=1}^T \delta_t^{02}}{T} \right) \left(1 + \|\theta_t - \hat{\theta}\|_2^2 \right) \right] \\
&\lesssim \frac{1}{L} + \frac{1}{T}. \tag{9.26}
\end{aligned}$$

And because we have

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

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{[2]}_t\|_2^2 \mid \theta_t] \\
&\lesssim \mathbb{E} \left[\left\| (H^{-1} - H_t^{-1}) \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\hat{\theta}) \right\|_2^2 + \left\| H_t^{-1} \frac{1}{S_o} \sum_{i \in I_o} (\nabla f_i(\hat{\theta}) - \nabla f_i(\theta_t)) \right\|_2^2 \mid \theta_t \right] \\
&\quad \text{because } H^{-1} - H_t^{-1} = H^{-1}(H_t - H)H_t^{-1} \text{ and using Lemma 7} \tag{9.27}
\end{aligned}$$

$$\begin{aligned}
&\lesssim \mathbb{E}[\|\theta_t - \hat{\theta}\|_2^2 \mid \theta_t] \\
&\lesssim (t+1)^{-d_o}, \tag{9.28}
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E} \left[\left\| H^{-1}GH^{-1} - \frac{S_o}{T} \sum_{t=1}^T \frac{\bar{g}_t \bar{g}_t^\top}{\rho_t^2} \right\|_2 \right] \\
&\lesssim \frac{1}{\sqrt{T}} + \frac{1}{T} \sum_{t=1}^T (t+1)^{-\frac{d_o}{2}} + \frac{1}{T} \sum_{t=1}^T (t+1)^{-d_o} + \frac{1}{\sqrt{L}} + \frac{1}{L} \\
&\quad \text{because } \sum_{t=1}^T (t+1)^{-\frac{d_o}{2}} = T^{1-\frac{d_o}{2}} \text{ for } d_o \in \left(\frac{1}{2}, 1\right) \\
&\lesssim \frac{1}{T^{\frac{d_o}{2}}} + \frac{1}{\sqrt{L}}.
\end{aligned}$$

9.2 Proof of Corollary 3

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

$$\begin{aligned}
\frac{g_t^L}{\rho_t} &= \underbrace{-H^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\hat{\theta})}_{[1]} \\
&\quad + \underbrace{H^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\hat{\theta}) - H_t^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\hat{\theta}) + H_t^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\hat{\theta}) - H_t^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\theta_t) + H_t^{-1} \nabla f(\theta_t)}_{[2]} \\
&\quad - \underbrace{H_t^{-1} \nabla f(\theta_t) + (\theta_t - \hat{\theta})}_{[3]} - \underbrace{H_t^{-1} \frac{g_t^0}{\rho_t} + \frac{g_t^L}{\rho_t}}_{[4]} - (\theta_t - \hat{\theta}), \tag{9.29}
\end{aligned}$$

which gives

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

And we have

$$\begin{aligned}
&\sqrt{T} \left(\frac{\sum_{t=1}^T \theta_t}{T} - \hat{\theta} \right) \\
&= \frac{1}{\sqrt{T}} \left(\sum_{t=1}^T \prod_{i=0}^{t-1} (1 - \rho_i) \right) (\theta_0 - \hat{\theta}) + \frac{1}{\sqrt{T}} \sum_{t=1}^T \sum_{i=0}^{t-1} \left(\prod_{j=i+1}^{t-1} (1 - \rho_j) \right) \rho_i ([1]_i + [2]_i + [3]_i + [4]_i) \\
&= \frac{1}{\sqrt{T}} \left(\sum_{t=1}^T \prod_{i=0}^{t-1} (1 - \rho_i) \right) (\theta_0 - \hat{\theta}) + \frac{1}{\sqrt{T}} \sum_{i=0}^{T-1} \sum_{t=i+1}^T \left(\prod_{j=i+1}^{t-1} (1 - \rho_j) \right) \rho_i ([1]_i + [2]_i + [3]_i + [4]_i). \tag{9.30}
\end{aligned}$$

For the first term in (9.30), which is non-stochastic, we have

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

For the second term in (9.30), 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 (9.11)) and satisfies

$$\begin{aligned}
& \rho_i \sum_{t=i+1}^T \prod_{j=i+1}^{t-1} (1 - \rho_j) \\
&= \sum_{t=i+1}^T \frac{\rho_i}{\rho_t} \rho_t \prod_{j=i+1}^{t-1} (1 - \rho_j) \\
&\leq \frac{\rho_i}{\rho_s} \sum_{t=i+1}^s \rho_t \prod_{j=i+1}^{t-1} (1 - \rho_j) + \rho_i \left(\prod_{j=i+1}^s (1 - \rho_j) \right) \sum_{t=s+1}^T \prod_{j=s+1}^{t-1} (1 - \rho_j) \\
&= \left(1 + \frac{\rho_i - \rho_s}{\rho_s}\right) \left(1 - \prod_{t=i+1}^s (1 - \rho_t)\right) + \rho_i \left(\prod_{j=i+1}^s (1 - \rho_j) \right) \sum_{t=s+1}^T \prod_{j=s+1}^{t-1} (1 - \rho_j) \\
&\leq \left(1 + \frac{\rho_i - \rho_s}{\rho_s}\right) (1 - (1 - \rho_s)^{s-i}) + \rho_i (1 - \rho_s)^{s-i} \sum_{t=s+1}^T \prod_{j=s+1}^{t-1} (1 - \rho_j) \\
&\leq 1 + \left(\left(1 + \frac{s-i}{i+1}\right)^{d_o} - 1 \right) + \rho_i e^{-(s-i)\rho_s} \sum_{t=s+1}^{\infty} \prod_{j=s+1}^{t-1} (1 - \rho_j) \\
&\leq 1 + \frac{s-i}{i} + \rho_i e^{-(s-i)\rho_s} \sum_{t=s+1}^{\infty} \prod_{j=s+1}^{t-1} (1 - \rho_j),
\end{aligned}$$

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

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

$$\geq 1 - \exp\left(-\frac{1}{1-d_o}((T+2)^{1-d_o} - (i+2)^{1-d_o})\right)$$

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

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

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

For term [2], we have

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

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

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

For term [3], we have

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

By using (4.7) and Cauchy-Schwarz inequality, we have

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

For term [4], similar to similar to (9.26), we have

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

9.3 Proof of Corollary 5

Using Theorem 6.5 of [Bub15a], we have

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

Similar to (4.8) in Theorem 2 (Section 9.1.1), we have

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

Similar to the proof of (4.10) in Theorem 2 (Section 9.1.5), using (9.25), we have

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

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

$$\begin{aligned} & \frac{\bar{g}_t}{\rho_t} \\ &= \underbrace{-H^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\hat{\theta})}_{[1]} \\ & \quad + \underbrace{H^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\hat{\theta}) - H_t^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\hat{\theta}) + H_t^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\hat{\theta}) - H_t^{-1} \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\theta_t) + H_t^{-1} \nabla f(\theta_t)}_{[2]} \\ & \quad - \underbrace{H_t^{-1} \nabla f(\theta_t)}_{[3]} - \underbrace{H_t^{-1} \frac{g_t^0}{\rho_t} + \frac{g_t^L}{\rho_t}}_{[4]}. \end{aligned} \tag{9.31}$$

For term [1], we have

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

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

For term [2], similar to (9.27), we have

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

For term [3], we have

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

For term [4], similar to (9.26), we have

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

Chapter 10

Chapter 5 proofs

10.1 Proof of Theorem 3

The error bound proof is similar to standard LASSO proofs [BvdG11, NRWY12].

We will use Lemma 9 for the covariance estimate using soft thresholding.

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

Lemma 9. *Under our assumptions in Chapter 5, 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

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

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

Under this event, ℓ_2 operator norm of $\widehat{S} - \Sigma$ satisfies

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

This chapter also appears in [LKLC18]. It was written by Tianyang Li, and edited by Anastasios Kyrillidis and Constantine Caramanis.

ℓ_1 and ℓ_∞ 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, [BL08].

Our assumption that Σ 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^{th} sample, for each i.i.d. sample $x = [x(1), x(2), \dots, x(p)]^\top \sim \mathcal{N}(0, \Sigma)$, 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 Σ 's eigenvalues are upper and lower bounded. Thus, $x(j) \pm x(k)$ are χ_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 $\Theta(1)$ [Wai17]. 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 [Wai17], 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 9 guarantees that the optimization problem (5.1) is well defined with high probability when $n \gg b\sqrt{\frac{\log p}{n}}$. Because the ℓ_2 operator norm $\|\widehat{S} - \Sigma\|_2 \lesssim b\sqrt{\frac{\log p}{n}} \ll 1$, and the positive definite matrix Σ '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

$$0 \leq v^\top \widehat{S} v = \Theta(\|v\|_2^2). \quad (10.1)$$

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

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

which, after rearranging terms, is equivalent to

$$\frac{1}{2} (\widehat{\theta} - \theta^*)^\top \widehat{S} (\widehat{\theta} - \theta^*) + \left\langle \left(\widehat{S} - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \theta^* + \frac{1}{n} \sum_{i=1}^n \epsilon_i x_i, \widehat{\theta} - \theta^* \right\rangle \leq \lambda (\|\theta^*\|_1 - \|\widehat{\theta}\|_1). \quad (10.2)$$

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

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

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

For the j^{th} coordinate of $\epsilon_i x_i$, because ϵ_i and x_i are independent Gaussian random variables, we know that it is sub-exponential [Wai17]

$$\mathbb{P}[|\epsilon_i x_i(j)| > t] \lesssim \exp\left(-\Theta\left(\min\left\{\frac{t^2}{\sigma^2}, \frac{t}{\sigma}\right\}\right)\right), \quad (10.3)$$

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

Using Bernstein inequality, we have

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

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

$$\left\|\frac{1}{n} \sum_{i=1}^n \epsilon_i x_i\right\|_\infty \lesssim \sigma \sqrt{\frac{\log p}{n}}, \quad (10.4)$$

when $n \gg \log p$.

Thus, we set the regularization parameter

$$\begin{aligned} \lambda &= \Theta\left(\left(\sigma + \|\theta^*\|_1\right) \sqrt{\frac{\log p}{n}}\right) \\ &\geq 2 \left\| \left(\widehat{S} - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top\right) \theta^* + \frac{1}{n} \sum_{i=1}^n \epsilon_i x_i \right\|_\infty, \end{aligned} \quad (10.5)$$

which holds under the events in Lemma 9 and (10.4).

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

(10.2) and (10.5) implies that

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

which is equivalent to

$$\|\theta^{\bar{S}}\|_1 \leq 3\|(\theta - \theta^*)^S\|_1, \quad (10.6)$$

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 (10.2), we have

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

which implies that

$$\|(\theta - \theta^*)^S\|_1 \lesssim s(\sigma + \|\theta^*\|_1)\sqrt{\frac{\log p}{n}}. \quad (10.7)$$

Combining (10.7) and (10.6), we have proven (5.2)

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

In (10.2) because $\langle (\widehat{S} - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top) \theta^* + \frac{1}{n} \sum_{i=1}^n \epsilon_i x_i, \widehat{\theta} - \theta^* \rangle \geq 0$ by convexity, and using (10.1), we have proven (5.3)

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

10.1.1 Proof of Theorem 4

At the solution $\widehat{\theta}$ of the optimization problem (5.1), using the KKT condition, we have

$$\left(\widehat{S} - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \widehat{\theta} + \frac{1}{n} \sum_{i=1}^n x_i (x_i^\top \widehat{\theta} - y_i) + \lambda \widehat{g} = 0, \quad (10.8)$$

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

$$\widehat{S}\widehat{\theta} - \frac{1}{n} \sum_{i=1}^n x_i(x_i^\top \theta^* + \epsilon_i) + \lambda \widehat{g} = 0, . \quad (10.9)$$

By Lemma 9, we know that \widehat{S} is invertible when $n \gg b^2 \log p$.

Plugging (5.4) into (10.9), we have

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

which is equivalent to

$$\widehat{S}(\widehat{\theta}^d - \theta^*) - \frac{1}{n} \sum_{i=1}^n \epsilon_i x_i + \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top - \widehat{S} \right) (\widehat{\theta} - \theta^*) = 0, \quad (10.10)$$

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

Rewriting (10.10), we have

$$\widehat{\theta}^d - \theta^* = \widehat{S}^{-1} \frac{1}{n} \sum_{i=1}^n \epsilon_i x_i + \left(I - \widehat{S}^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \right) (\widehat{\theta} - \theta^*). \quad (10.11)$$

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

$$\begin{aligned} & \max_{1 \leq j, k \leq p} \left| \left(I - \widehat{S}^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \right)_{jk} \right| \\ &= \max_{1 \leq j, k \leq p} \left| \left(\widehat{S}^{-1} \left(S - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \right)_{jk} \right| \\ &\leq \|\widehat{S}^{-1}\|_\infty \max_{1 \leq j, k \leq p} \left| \left(S - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)_{jk} \right|. \end{aligned} \quad (10.12)$$

Under the event in Lemma 9, we have

$$\max_{1 \leq j, k \leq p} \left| \left(S - \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)_{jk} \right| \lesssim \sqrt{\frac{\log p}{n}}. \quad (10.13)$$

Also under the event in Lemma 9, we have

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

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, [Var75], when $n \gg \frac{1}{D_\Sigma^2} \log p$, we have

$$\|\widehat{S}\|_\infty \lesssim \frac{1}{D_\Sigma}, \quad (10.14)$$

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

And using (10.13) and (10.14) in (10.12), we have

$$\max_{1 \leq j, k \leq p} \left| \left(I - \widehat{S}^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \right)_{jk} \right| \lesssim \frac{1}{D_\Sigma} \sqrt{\frac{\log p}{n}}. \quad (10.15)$$

Using (10.15) and the bound on $\|\widehat{\theta} - \theta^*\|_1$ (5.2), in (10.11), we have

$$\left\| \left(I - \widehat{S}^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \right) (\widehat{\theta} - \theta^*) \right\|_\infty \lesssim \frac{1}{D_\Sigma} s (\sigma + \|\theta^*\|_1) \frac{\log p}{n} \lesssim \frac{1}{D_\Sigma} s (\sigma + \sqrt{s}) \frac{\log p}{n}. \quad (10.16)$$

Combining (10.16) and (10.11), we have proven Theorem 4, when $n \gg \max\{b^2, \frac{1}{D_\Sigma^2}\} \log p$, we have

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

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 $\|R\|_\infty \lesssim \frac{1}{D_\Sigma} s (\sigma + \|\theta^*\|_1) \frac{\log p}{\sqrt{n}} \lesssim \frac{1}{D_\Sigma} s (\sigma + \sqrt{s}) \frac{\log p}{\sqrt{n}}$ with probability at least $1 - p^{-\Theta(1)}$.

10.1.2 Proof of Theorem 5

We analyze the optimization problem conditioned on the data set $\{x_i\}_{i=1}^n$, which satisfies Lemma 9 with probability at least $1 - p^{\Theta(-1)}$ when $n \gg b^2 \log p$.

Here, we denote the objective function as

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

In Algorithm 3, lines 6 to 15 are using SVRG [JZ13] to solve the Newton step

$$\min_{\Delta} \frac{1}{2} \Delta^\top \widehat{S} \Delta + \left\langle \frac{1}{S_o} \sum_{k \in I_o} \nabla f_k(\theta_t), \Delta \right\rangle, \quad (10.17)$$

and using proximal SVRG [XZ14] to solve the proximal Newton step

$$\min_{\Delta} \frac{1}{2} \Delta^\top \widehat{S} \Delta + \left\langle \frac{1}{n} \sum_{k=1}^n \nabla f_k(\theta_t), \Delta \right\rangle + \lambda \|\theta + \Delta\|_1. \quad (10.18)$$

The gradient of (10.17) is

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

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

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 9, when $n \gg b^2 \log p$, the ℓ_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^{th} diagonal element of $\widehat{S}^\top \widehat{S}$, we have $\|\widehat{S}_k\|_2^2 = O(1)$ for all $1 \leq k \leq p$. Thus, each $[p\widehat{S}_k] \Delta(k)$ is a $O(p)$ -Lipschitz function.

By Theorem 6.5 of [Bub15a], when conditioned on θ_t , and choosing

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

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 $\frac{1}{2} \Delta^\top \widehat{S} \Delta + \langle \frac{1}{n} \sum_{k=1}^n \nabla f_k(\theta_t), \Delta \rangle$ in (10.18) is

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

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 [XZ14], when conditioned on θ_t , and choosing

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

after L_o^t proximal SVRG outer steps, we have

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

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

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

At each θ_t , we have

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

For the first term, we have

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

which implies that

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

For the second term, we have

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

Because when $n \gg \log p$, from (10.21) 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 (10.23) 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 θ_t (and the data set $\{x_i\}_{i=1}^n$) we have

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

under the events of (10.21), (10.14), and (10.23), where we used the fact (10.14) that the ℓ_∞ 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 (10.21), (10.23), and (10.14), we have we have an asymptotic normality result

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

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

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

which implies

$$\mathbb{E} [\|R\|_\infty \mid \{x_i\}_{i=1}^n, (10.21), (10.23), (10.14)]$$

$$\begin{aligned}
&\lesssim \mathbb{E} \left[\frac{1}{\sqrt{t}} \sum_{t=1}^T 0.95^{L_o^t} (1 + \|\theta_t - \widehat{\theta}\|_2) + \sqrt{p}(\log p + \log n) \|\theta_t - \widehat{\theta}\|_2 \mid \{x_i\}_{i=1}^n, (10.21), (10.23), (10.14) \right] \\
&\lesssim \frac{1}{\sqrt{T}} \sum_{t=1}^T 0.95^{L_o^t} (1 + \sqrt{P(\theta_0) - P(\widehat{\theta})} 0.95^{\sum_{i=0}^{t-1} L_o^i}) + \sqrt{p}(\log p + \log n) \sqrt{P(\theta_0) - P(\widehat{\theta})} 0.95^{\sum_{i=0}^{t-1} L_o^i}.
\end{aligned}$$

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 (10.21), (10.23), and (10.14), using a union bound over all matrix entries, and sub-Gaussian concentration inequalities [Wai17] similar to Lemma 1's proof, when $T \gg \left((\log p + \log n) \|\widehat{\theta} - \theta^*\|_1 + \sigma \sqrt{(\log p + \log n) \log n}\right) \log p$, we also have

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

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

10.1.3 Proof of Lemma 1

We analyze the optimization problem conditioned on the data set $\{x_i\}_{i=1}^n$, which satisfies Lemma 9 with probability at least $1 - p^{\Theta(-1)}$ when $n \gg b^2 \log p$.

Because we have

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

we can write

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

$$= \frac{1}{n} \sum_{i=1}^n (\sigma^2 - \epsilon_i^2) x_i x_i^\top + \frac{1}{n} \sum_{i=1}^n (2\epsilon_i x_i^\top (\hat{\theta} - \theta^*) - (x_i^\top (\hat{\theta} - \theta^*))^2) x_i x_i^\top. \quad (10.19)$$

Conditioned on $\{x_i\}_{i=1}^n$, because $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d., and ϵ_i^2 is sub-exponential, using Bernstein inequality [Wai17], we have

$$\begin{aligned} & \mathbb{P} \left[\left| \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{\epsilon_i^2}{\sigma^2} \right) x_i(j) x_i(k) \right| > t \mid \{x_i\}_{i=1}^n \right] \\ & \lesssim \exp \left(-n \min \left\{ \frac{t}{\max_{1 \leq i \leq n} |x_i(j) x_i(k)|}, \left(\frac{t}{\max_{1 \leq i \leq n} |x_i(j) x_i(k)|} \right)^2 \right\} \right), \end{aligned} \quad (10.20)$$

for $1 \leq j, k \leq p$, where $x_i(j)$ is the j^{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

$$\max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}} |x_i(j)| \lesssim \sqrt{\log p + \log n}, \quad (10.21)$$

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

Combining (10.20) and (10.21), and taking a union bound over all entries of the matrix $\frac{1}{n} \sum_{i=1}^n (\sigma^2 - \epsilon_i^2) x_i x_i^\top$, when $n \gg \log p$, we have

$$\max_{1 \leq j, k \leq p} \left| \left(\frac{1}{n} \sum_{i=1}^n (\sigma^2 - \epsilon_i^2) x_i x_i^\top \right)_{jk} \right| \lesssim \sigma^2 (\log p + \log n) \sqrt{\frac{\log p}{n}}, \quad (10.22)$$

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

$$\max_{1 \leq i \leq n} |\epsilon_i| \lesssim \sigma \sqrt{\log n}, \quad (10.23)$$

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

Using (10.21), we have

$$\max_{1 \leq i \leq n} |x_i^\top (\hat{\theta} - \theta^*)|$$

$$\begin{aligned}
&\leq \|\widehat{\theta} - \theta^*\|_1 \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |x_i(j)| \\
&\lesssim s (\sigma + \|\theta^*\|_1) \sqrt{\frac{\log p}{n} (\log p + \log n)} \lesssim s (\sigma + \sqrt{s}) \sqrt{\frac{\log p}{n} (\log p + \log n)}, \tag{10.24}
\end{aligned}$$

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

Combining (10.21), (10.22), (10.23), (10.24), and using (10.19), when $n \gg \log p$, we have

$$\begin{aligned}
&\max_{1 \leq j, k \leq p} \left| \left(\frac{1}{n} \sum_{i=1}^n (x_i^\top \widehat{\theta} - y_i)^2 x_i x_i^\top - \sigma^2 \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)_{jk} \right| \\
&\lesssim \sigma^2 (\log p + \log n) \sqrt{\frac{\log p}{n}} + \sigma s (\sigma + \|\theta^*\|_1) (\log p + \log n)^{\frac{3}{2}} \sqrt{\frac{\log p \cdot \log n}{n}} \\
&\quad + s^2 (\sigma + \|\theta^*\|_1)^2 (\log p + \log n)^2 \frac{\log p}{n}, \tag{10.25}
\end{aligned}$$

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

Combining (10.25) and (10.14), when $n \gg \max\{b^2, \frac{1}{D_\Sigma^2}\} \log p$, we have

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

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

10.2 Technical lemmas

10.2.1 Lemma 10

Next lemma is a well known property of convex functions (Lemma 3.11 of [Bub15a]).

Lemma 10. *For a α strongly convex and β smooth function $F(x)$, we have*

$$\langle \nabla F(x_1) - \nabla F(x_2), x_1 - x_2 \rangle \geq \frac{\alpha \beta}{\alpha + \beta} \|x_1 - x_2\|_2^2 + \frac{1}{\beta + \alpha} \|\nabla F(x_1) - \nabla F(x_2)\|_2^2$$

$$\geq \frac{1}{2}\alpha\|x_1 - x_2\|_2^2 + \frac{1}{2\beta}\|\nabla F(x_1) - \nabla F(x_2)\|_2^2.$$

10.2.2 Lemma 11

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

Lemma 11. *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

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

When we assume that a_1, C, κ, p, d are all constants, we have

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

Proof. Unrolling the recursion, we have

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

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

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

where we used

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

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

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

and using the fact that

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

we have

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

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

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

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

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

□

Appendix

Appendix 1

Chapter 3 appendix

1.1 Experiments

Here we present additional experiments on our SGD inference procedure.

1.1.1 Synthetic data

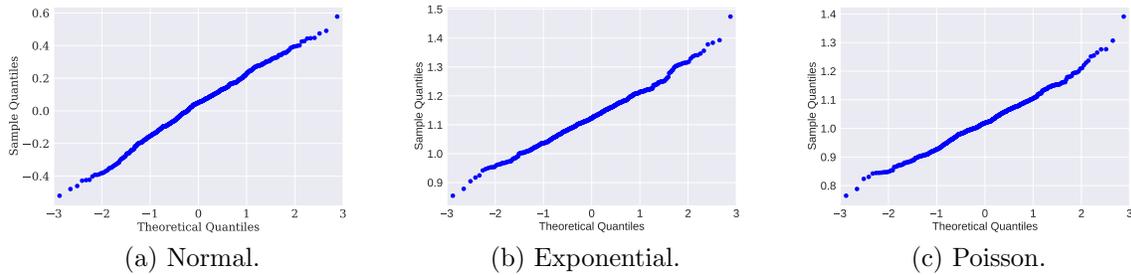


Figure 1.1: Estimation in univariate models: Q-Q plots for samples shown in Figure 3.2

Figure 1.1 shows Q-Q plots for samples shown in Figure 3.2.

1.1.1.1 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

This chapter also appears in [LLK18]. The experiments were conducted in collaboration with Liu Liu. It was edited by Anastasios Kyrillidis and Constantine Caramanis.

Figure 1.2. Q-Q plots per coordinate for samples in linear regression experiment 2 is shown in Figure 1.3.

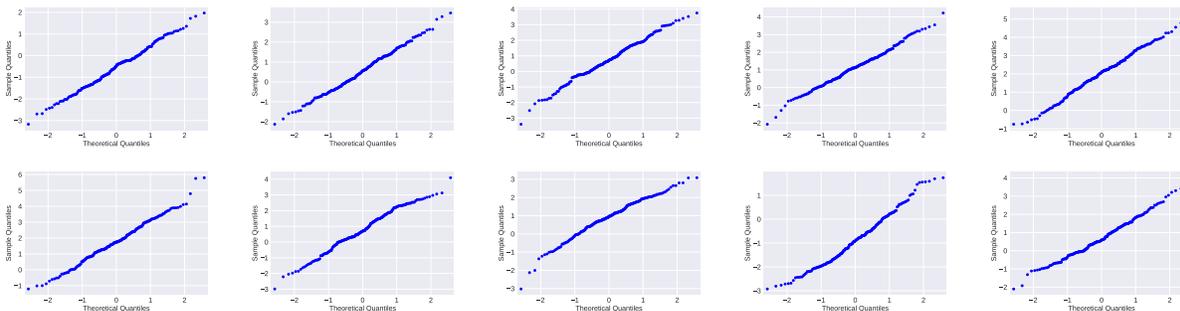


Figure 1.2: Linear regression experiment 1: Q-Q plots per coordinate

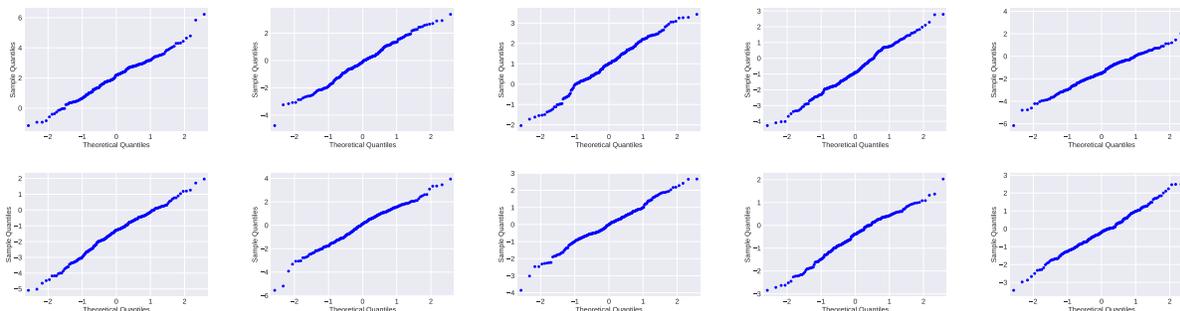


Figure 1.3: Linear regression experiment 2: Q-Q plots per coordinate

Q-Q plots per coordinate for samples in logistic regression experiment 1 is shown in Figure 1.4. Q-Q plots per coordinate for samples in logistic regression experiment 2 is shown in Figure 1.5.

Additional experiments

2-Dimensional Linear Regression. Consider:

$$y = x_1 + x_2 + \epsilon, \quad \text{where } \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right) \text{ and } \epsilon \sim \mathcal{N}(0, \sigma^2 = 10^2).$$

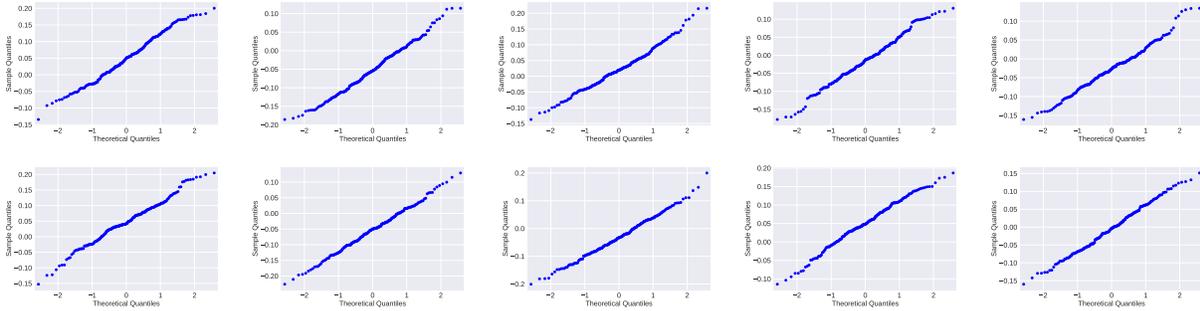


Figure 1.4: Logistic regression experiment 1: Q-Q plots per coordinate

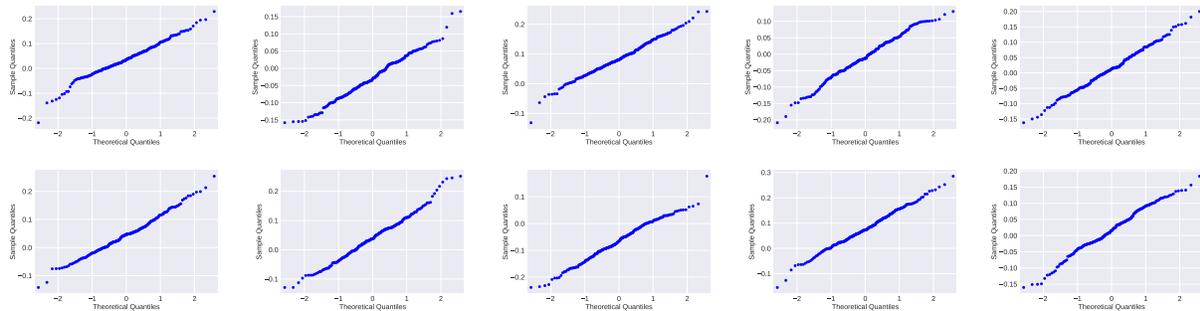


Figure 1.5: Logistic regression experiment 2: Q-Q plots per coordinate

Each sample consists of $Y = y$ and $X = [x_1, x_2]^\top$. We use linear regression to estimate w_1, w_2 in $y = w_1x_1 + w_2x_2$. In this case, the minimizer of the population least square risk is $w_1^* = 1, w_2^* = 1$.

For 300 i.i.d. samples, we plotted 100 samples from SGD inference in Figure 1.6. We compare our SGD inference procedure against bootstrap in Figure 1.6a. Figure 1.6b and Figure 1.6c show samples from our SGD inference procedure with different parameters.

10-Dimensional Linear Regression.

Here we consider the following model

$$y = x^\top w^* + \epsilon,$$

where $w^* = \frac{1}{\sqrt{10}}[1, -1, \dots, -1]^\top \in \mathbb{R}^{10}$, $x \sim \mathcal{N}(0, \Sigma)$ 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

$$\hat{w} = \operatorname{argmin}_w \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (x_i^\top w - y_i)^2.$$

Figure 1.7 shows the diagonal terms 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.

2-Dimensional Logistic Regression.

Here we consider the following model

$$\mathbb{P}[Y = +1] = \mathbb{P}[Y = -1] = \frac{1}{2}, \quad X | Y \sim \mathcal{N}(\mu = 1.1 + 0.1Y, \sigma^2 = 1). \quad (1.1)$$

We use logistic regression to estimate w, b in the classifier $\operatorname{sign}(wx + b)$ where the minimizer of the population logistic risk is $w^* = 0.2, b^* = -0.22$.

For 100 i.i.d. samples, we plot 1000 samples from SGD in Figure 1.8. In our simulations, we notice that our modified SGD for logistic regression behaves similar to vanilla logistic regression. This suggests that an assumption weaker than $(\theta - \hat{\theta})^\top \nabla f(\theta) \geq \alpha \|\theta - \hat{\theta}\|_2^2$

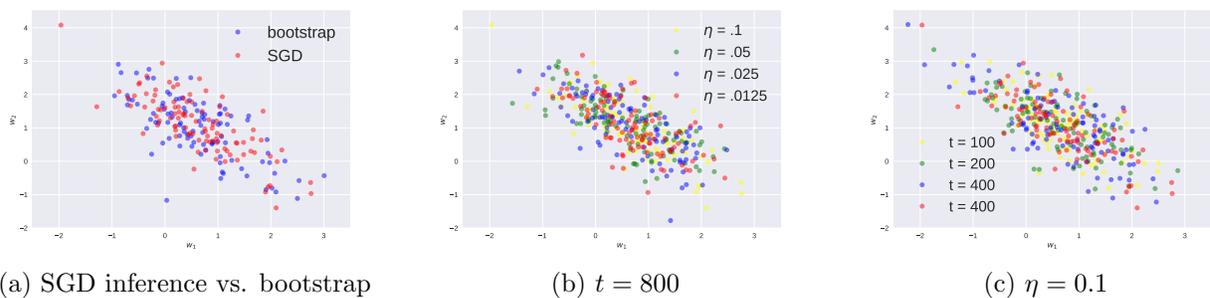
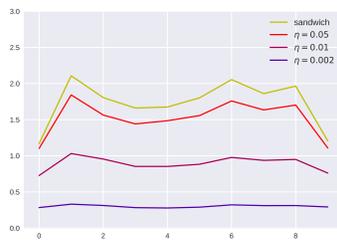
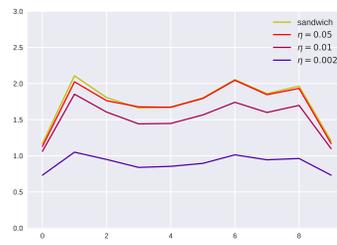


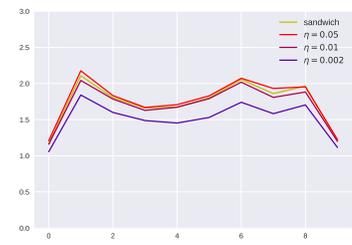
Figure 1.6: 2-dimensional linear regression



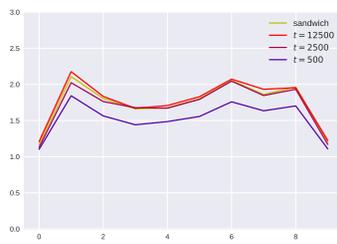
(a) $t = 500$



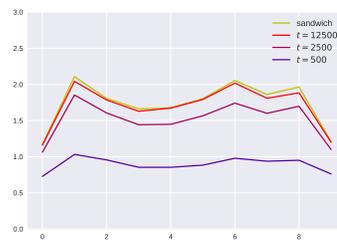
(b) $t = 2500$



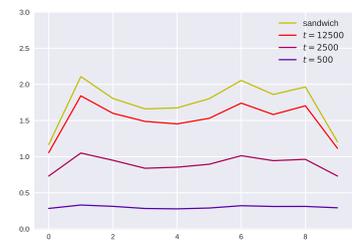
(c) $t = 12500$



(d) $\eta = 0.05$

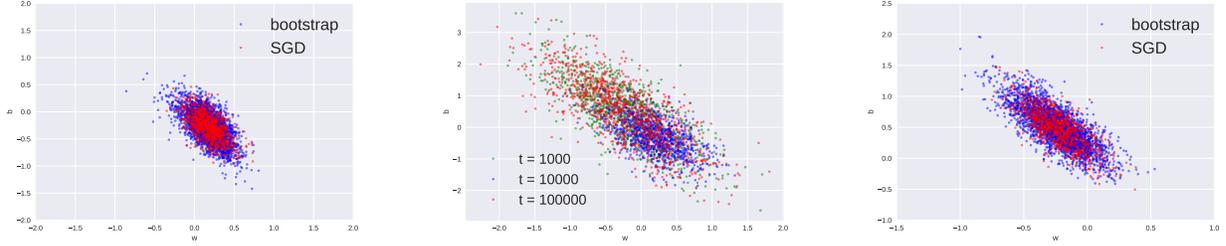


(e) $\eta = 0.01$

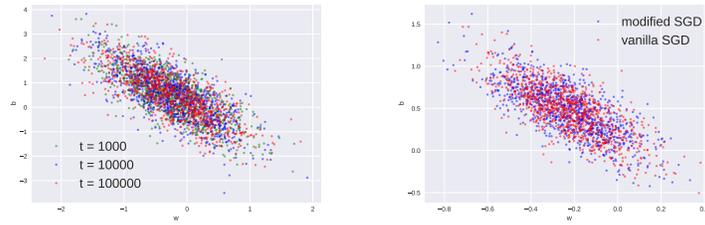


(f) $\eta = 0.002$

Figure 1.7: 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 1.8: 2-dimensional logistic regression

(assumption (F_1) in Theorem 1) is sufficient for SGD analysis. Figure 1.8b and Figure 1.8d suggest that the $t\eta^2$ term in Corollary 1 is an artifact of our analysis, and can be improved.

11-Dimensional Logistic Regression.

Here we consider the following model

$$\mathbb{P}[Y = +1] = \mathbb{P}[Y = -1] = \frac{1}{2}, \quad X | Y \sim \mathcal{N}(0.01Y\mu, \Sigma),$$

where $\Sigma_{ii} = 1$ and when $i \neq j$ $\Sigma_{ij} = \rho^{|i-j|}$ for some $\rho \in [0, 1)$, and $\mu = \frac{1}{\sqrt{10}}[1, 1, \dots, 1]^\top \in \mathbb{R}^{10}$. We estimate a classifier $\text{sign}(w^\top x + b)$ using

$$\widehat{w}, \widehat{b} = \underset{w, b}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-Y_i(w^\top X_i + b))). \quad (1.2)$$

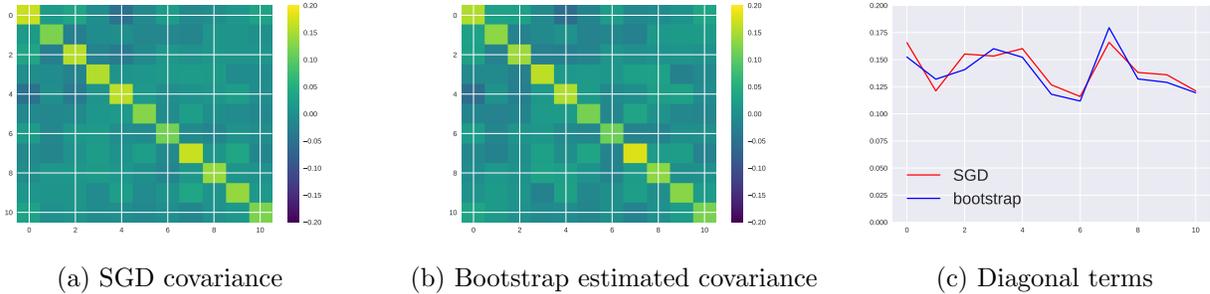


Figure 1.9: 11-dimensional logistic regression: $\rho = 0$

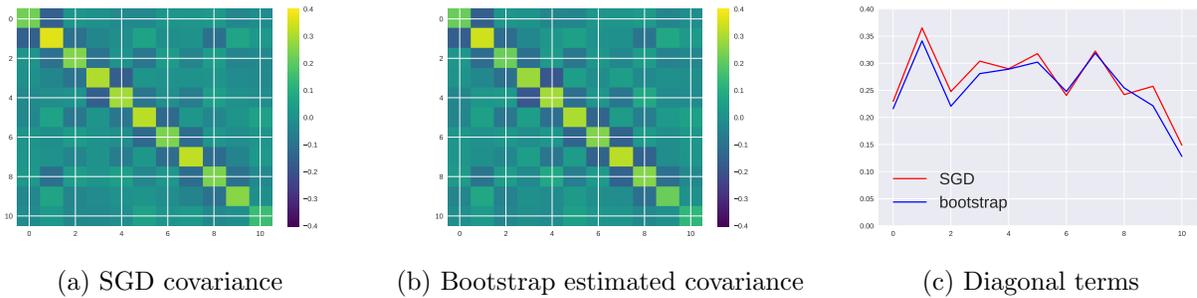


Figure 1.10: 11-dimensional logistic regression: $\rho = 0.6$

Figure 1.9 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 1.10, 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.

1.1.2 Real data

Here, we compare covariance matrix computed using our SGD inference procedure, bootstrap, and inverse Fisher information matrix on the Higgs data set [BSW14] and the LIBSVM Splice data set, and we observe that they have similar statistical properties.

1.1.2.1 Higgs data set

The Higgs data set ¹ [BSW14] 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 3.1.3.

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 1.11.

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 3.1). For bootstrap, we use 2000 replicates, which is much larger than the sample size $n = 200$.

Figure 1.11 shows that the covariance matrix obtained by SGD inference is comparable to the estimation given by bootstrap and inverse Fisher information.

¹<https://archive.ics.uci.edu/ml/datasets/HIGGS>

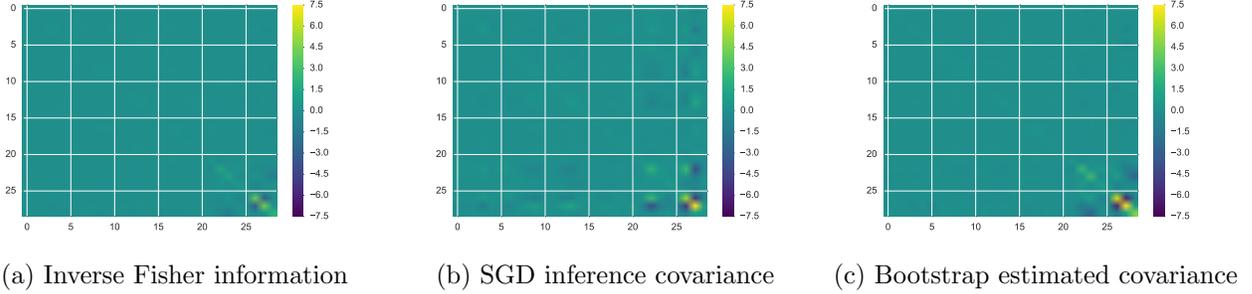


Figure 1.11: Higgs data set with $n = 200$

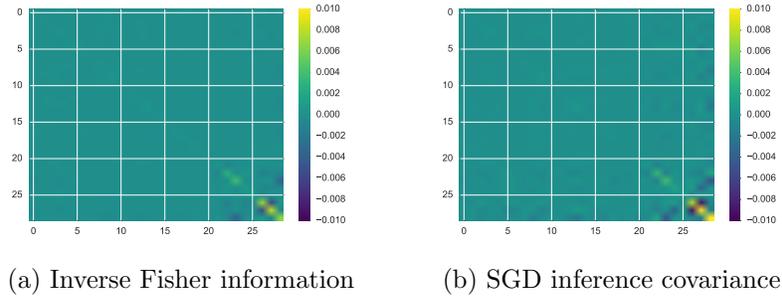


Figure 1.12: Higgs data set with $n = 50000$

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 3.1). For this large data set, we present the estimated covariance of SGD inference procedure and inverse Fisher information (the asymptotic covariance) in Figure 1.12 because bootstrap is computationally prohibitive. Similar to the small sample result in Figure 1.11, the covariance of our SGD inference procedure is comparable to the inverse Fisher information.

In Figure 1.13, 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.

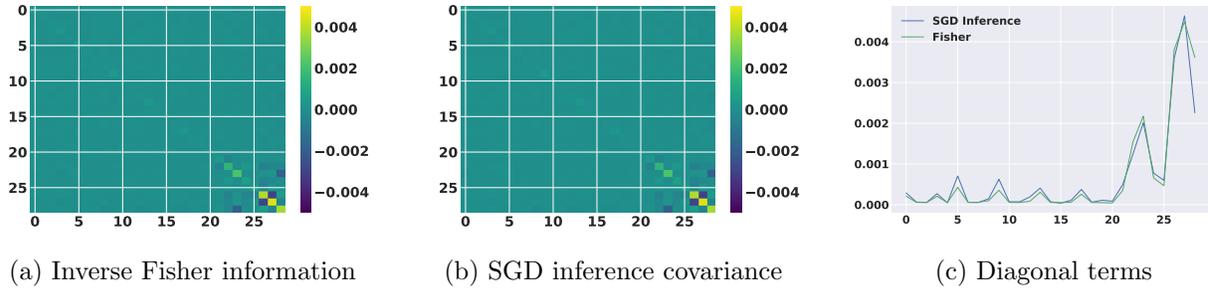


Figure 1.13: Higgs data set with $n = 90000$

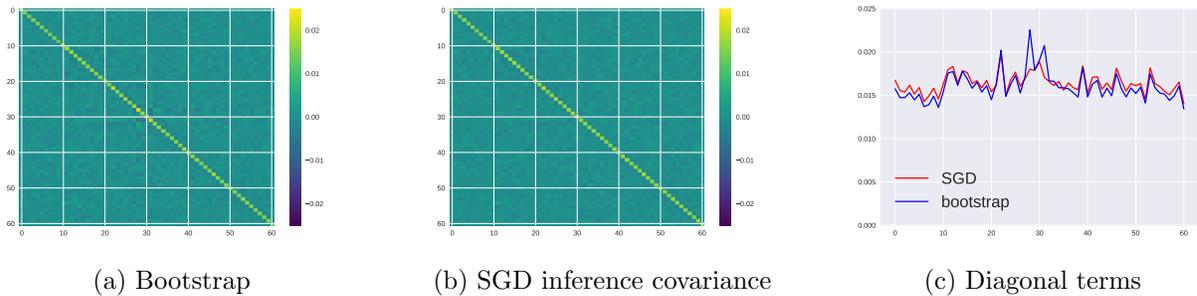


Figure 1.14: Splice data set

1.1.2.2 Splice data set

The Splice data set ² 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.

In Figure 1.14, 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.

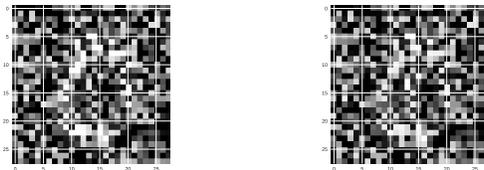
²<https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html>

1.1.2.3 MNIST

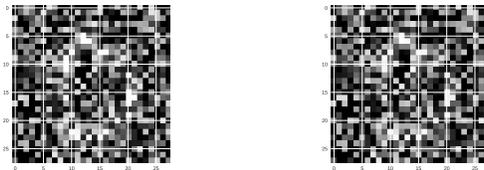
Here, we train a binary logistic regression classifier to classify 0/1 using perturbed MNIST data set, and demonstrate that certain adversarial examples (e.g. [GSS14]) can be detected using prediction confidence intervals. For each image, where each original pixel is either 0 or 1, we randomly changed 70% pixels to random numbers uniformly on $[0, 0.9]$. Figure 1.15 shows 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 1.16 (scaled for display).



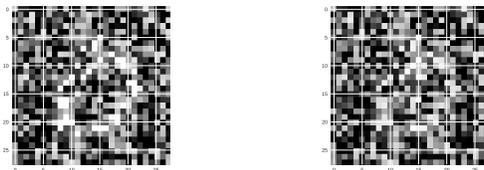
(a) Original “0”: logit -72.6, CI (-106.4, -30.0) (b) Adversarial “0”: logit 15.3, CI (-6.5, 26.2)



(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 1.15: MNIST

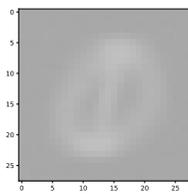


Figure 1.16: MNIST adversarial perturbation (scaled for display)

Appendix 2

Chapter 4 appendix

2.1 Experiments

2.1.1 Synthetic data

2.1.1.1 Low dimensional problems

Here, we provide the exact configurations for linear/logistic regression examples provided in Table 4.1.

Linear regression. We consider the model $y = \langle [1, \dots, 1]^T / \sqrt{10}, x \rangle + \epsilon$, 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 ([GP17] recently has shown that

This chapter also appears in [LKLC18]. The experiments were conducted in collaboration with Liu Liu. It was edited by Anastasios Kyrillidis and Constantine Caramanis.

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 $x \mid y \sim \mathcal{N}(0.1/\sqrt{10} \cdot [1, \dots, 1]^\top, \Sigma) \in \mathbb{R}^{10}$, 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.

Calibration. Here, we give empirical results on calibrating confidence intervals ([ET94], Ch.18; [PRW12], Ch. 9) produced by our approximate Newton procedure. We consider the model $y = \langle [1, \dots, 1]^\top / \sqrt{20}, x \rangle + \epsilon$, where $x \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^{20}$ and $\epsilon \sim \mathcal{N}(0, 0.7^2)$, with 200 i.i.d. data points. We ran 100 simulations. In each simulation, we bootstrapped the dataset 100 times, and computed confidence intervals on each bootstrap replicate using our approximate Newton procedure, bootstrap, and inverse Fisher information. For each method, we then used grid search to find a multiplier such that the empirical point estimate is covered by the bootstrap confidence intervals 95% of the time. Average 95% confidence interval coverage and length after calibration are given in Table 4.2.

2.1.2 Real data

2.1.2.1 Neural network adversarial attack detection

The adversarial perturbation used in our experiments is shown in Figure 2.3. It is generated using the fast gradient sign method [GSS14] Figure 2.1 shows images in a “Shirt”

example. Figure 2.2 shows images in a “T-shirt/top” example.

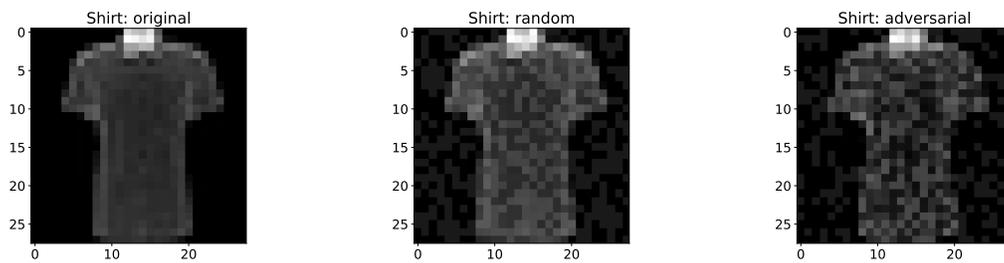


Figure 2.1: “Shirt” example

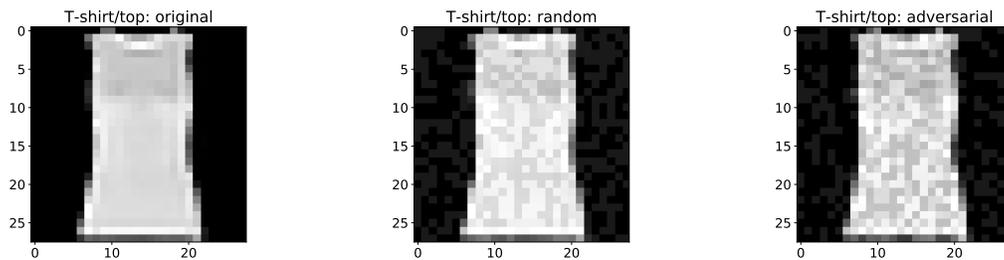


Figure 2.2: “T-shirt/top” example

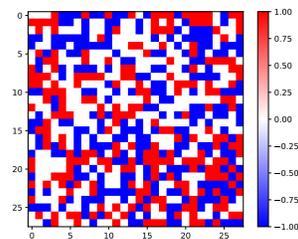


Figure 2.3: Adversarial perturbation generated using the fast gradient sign method [GSS14]

Appendix 3

Chapter 5 appendix

3.1 Experiments

3.1.1 Synthetic data

3.1.1.1 High dimensional linear regression

For comparison with de-biased LASSO [JM15, vdGBRD14], we use the *oracle* de-biased LASSO estimator

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

and its corresponding statistical error covariance estimate

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

which assumes that the true inverse covariance Σ^{-1} and observation noise variance σ^2 are known.

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

This chapter also appears in [LKLC18]. The experiments were conducted in collaboration with Liu Liu. It was edited by Anastasios Kyrillidis and Constantine Caramanis.

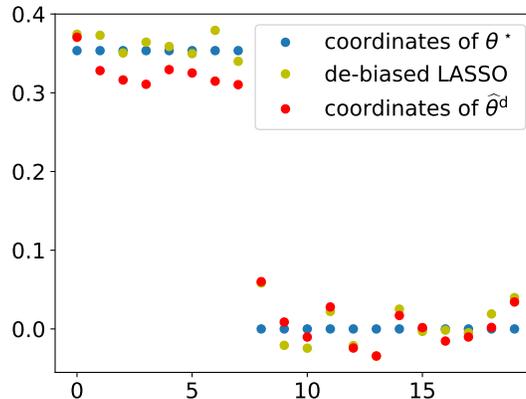


Figure 3.1: Comparison of our de-biased estimator and oracle de-biased LASSO estimator

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

3.1.2 Real data

3.1.2.1 High dimensional linear regression

HIV drug resistance mutations dataset. We apply our high dimensional inference procedure to the dataset in [RTW⁺06] to detect mutations related to HIV drug resistance. Our procedure is able to detect verified mutations in an expert dataset [JBVC⁺05], when we control the family-wise error rate (FWER) at 0.05.

Riboflavin (vitamin B2) production rate data set. For the vanilla LASSO estimate on the high-throughput genomic data set concerning riboflavin (vitamin B2) production rate [BKM14], we set $\lambda = 0.021864$. Figure 3.2, 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% significance level, our method does not find any significant gene. [JM14, BKM14] report that [Büh13] also does not find any significant gene, whereas [MMB09] finds one significant gene (YXLD-at), and [JM14] finds two significant genes (YXLD-at and YXLE-at). This indicates that our method is more conservative than [JM14, MMB09].

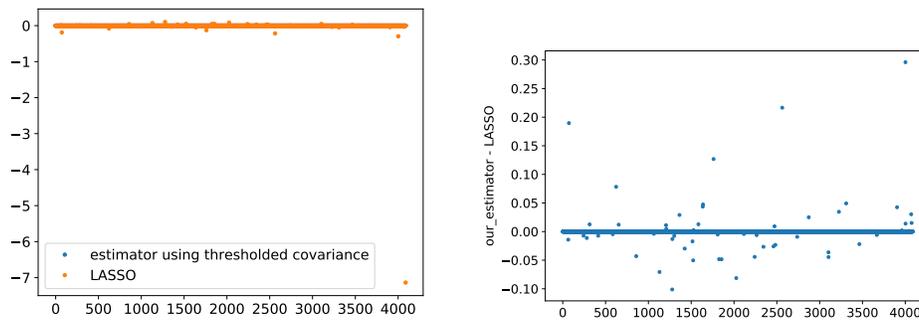


Figure 3.2: Comparison of our high dimensional linear regression point estimate with the vanilla LASSO estimate

Appendix 4

Chapter 6 appendix

4.1 Experiments

4.1.1 Synthetic data

In our linear regression simulation, we generate i.i.d. random explanatory variables, and the observation noise is a 0-mean moving average (MA) process independent of the explanatory variables.

For the linear model

$$y_i = \langle x_i, \theta^* \rangle + \epsilon_i,$$

$x_i \in \mathbb{R}^{20}$ are i.i.d. samples generated from $\mathcal{N}\left([1, 1, \dots, 1]^\top / \sqrt{k}, I\right)$, and ϵ_i is a 0-mean moving average process

$$\epsilon_i = 0.6 \cdot z_i + 0.8 \cdot z_{i-1},$$

where z_i are i.i.d. $\mathcal{N}(0, 0.7^2)$.

We ran 10,000 simulations, with each time series containing $n = 10,000$ samples, and set the lag $\mathbf{l} = 32$. For our approximate Newton statistical inference procedure (Algorithm 6), average 95% confidence interval (coverage, length) is (0.958, 0.0142), and it matches our theory. For circular bootstrap, where each replicate contains $n - \mathbf{l}$ samples, average 95% confidence interval (coverage, length) is (0.949, 0.0136).

This chapter also appears in [\[LKLC18\]](#). The experiments were conducted in collaboration with Liu Liu. It was edited by Anastasios Kyrillidis and Constantine Caramanis.

4.1.2 Real data

Using monthly equities returns data from [FP14], we use our approximate Newton statistical inference procedure to show that the correlation between US equities market returns and non-US global equities market returns is statistically significant, which validates the capital asset pricing model (CAPM) [Sha64, Lin65, FF04].

We regress monthly US equities market returns from 1995 to 2018 against other countries' equities market returns, and each country's coefficient and its 95% confidence interval is shown in Figure 6.1. And we observe that the US market is highly positively correlated with Canada and other advanced economies such as Germany and UK.

Bibliography

- [ABH17] 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. [2](#), [7](#)
- [AFM17] Yves F. Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *J. Mach. Learn. Res.*, 18(1):310–342, 2017. [41](#), [46](#)
- [Ama98] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998. [7](#), [28](#)
- [ANW10] 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. [40](#)
- [ANW12] 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. [7](#)
- [Bac10] Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010. [17](#)
- [Bac14] Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014. [17](#)

- [Bar46] Maurice Bartlett. On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8(1):27–41, 1946. [52](#)
- [BBMS13] Jennifer Bender, Remy Briand, Dimitris Melas, and Raman Subramanian. Foundations of factor investing. 2013. [51](#)
- [BEL15] Sebastien Bubeck, Ronen Eldan, and Joseph Lehec. Finite-time analysis of projected langevin monte carlo. In *Advances in Neural Information Processing Systems*, pages 1243–1251, 2015. [5](#)
- [Ber91] Ernst Berndt. *The practice of econometrics: classic and contemporary*. Addison Wesley Publishing Company, 1991. [51](#)
- [BHV14] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014. [3](#), [8](#)
- [BJ16] Ronen Basri and David Jacobs. Efficient representation of low-dimensional manifolds using deep networks. *arXiv preprint arXiv:1602.04723*, 2016. [32](#)
- [BKM14] 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. [44](#), [135](#), [136](#)
- [BL08] Peter Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604, 2008. [38](#), [101](#)
- [Bon36] C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936. [42](#)

- [Bot10] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. [1](#)
- [BPM90] Albert Benveniste, Pierre Priouret, and Michel Métivier. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag New York, Inc., New York, NY, USA, 1990. [6](#), [15](#)
- [BRT09] Peter Bickel, Ya'acov Ritov, and Alexandre Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009. [38](#)
- [BSW14] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5, 2014. [125](#)
- [Bub15a] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015. [1](#), [58](#), [98](#), [108](#), [113](#)
- [Bub15b] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, November 2015. [12](#)
- [Büh02] Peter Bühlmann. Bootstraps for time series. *Statistical science*, pages 52–72, 2002. [8](#)
- [Büh13] Peter Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, 2013. [44](#), [136](#)
- [BvdG11] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011. [8](#), [39](#), [100](#)

- [Car86] Edward Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, 14(3):1171–1179, 1986. [8](#)
- [CFJL18] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:model-xknockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018. [42](#)
- [CLTZ16] Xi Chen, Jason Lee, Xin Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*, 2016. [2](#), [4](#), [6](#), [7](#), [32](#), [77](#)
- [CM16] Charles K Chui and Hrushikesh Narhar Mhaskar. Deep nets for local manifold learning. *arXiv preprint arXiv:1607.07110*, 2016. [32](#)
- [CZ12] Tony Cai and Harrison Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, pages 2389–2420, 2012. [38](#)
- [DDB17] 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. [4](#)
- [DH97] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Cambridge University Press, 1997. [8](#)
- [Dip08a] Jürgen Dippon. Asymptotic expansions of the Robbins-Monro process. *Mathematical Methods of Statistics*, 17(2):138–145, 2008. [35](#)
- [Dip08b] Jürgen Dippon. Edgeworth expansions for stochastic approximation theory. *Mathematical Methods of Statistics*, 17(1):44–65, 2008. [35](#)

- [DK98] John Driscoll and Aart Kraay. Consistent covariance matrix estimation with spatially dependent panel data. *Review of economics and statistics*, 80(4):549–560, 1998. [3](#), [8](#), [52](#)
- [DLH16] Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Starting small-learning with adaptive sample sizes. In *International conference on machine learning*, pages 1463–1471, 2016. [30](#)
- [Edg08] Francis Ysidro Edgeworth. On the probable errors of frequency-constants. *Journal of the Royal Statistical Society*, 71(2):381–397, 1908. [1](#)
- [Efr82] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982. [1](#)
- [EH16] B. Efron and T. Hastie. *Computer age statistical inference*. Cambridge University Press, 2016. [1](#)
- [ET94] Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994. [1](#), [3](#), [8](#), [12](#), [30](#), [31](#), [32](#), [132](#)
- [FF04] Eugene F. Fama and Kenneth R. French. The capital asset pricing model: Theory and evidence. *Journal of economic perspectives*, 18(3):25–46, 2004. [55](#), [138](#)
- [FGLS18] 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. [8](#)
- [FHT01] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001. [1](#)

- [FP14] Andrea Frazzini and Lasse Heje Pedersen. Betting against beta. *Journal of Financial Economics*, 111(1):1–25, 2014. [55](#), [138](#)
- [FXY17] Yixin Fang, Jinfeng Xu, and Lei Yang. On Scalable Inference with Stochastic Gradient Descent. *arXiv preprint arXiv:1707.00192*, 2017. [2](#), [4](#)
- [GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning (adaptive computation and machine learning series). *Adaptive Computation and Machine Learning series*, page 800, 2016. [1](#)
- [GP17] Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. *arXiv preprint arXiv:1709.03342*, 2017. [131](#)
- [GSS14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. [xiv](#), [23](#), [33](#), [128](#), [132](#), [133](#)
- [HAV⁺15] 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. [30](#)
- [Hoe07] Daniel Hoehle. Robust standard errors for panel regressions with cross-sectional dependence. *Stata Journal*, 7(3):281–312, 2007. [3](#), [8](#)
- [Hub67] 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. [7](#), [40](#), [41](#), [45](#), [50](#)

- [JBVC⁺05] Victoria A Johnson, Francoise Brun-Vezinet, Bonaventura Clotet, Brian Conway, Daniel R. Kuritzkes, Deenan Pillay, Jonathan Schapiro, Amalio Telenti, and Douglas Richman. Update of the drug resistance mutations in HIV-1: 2005. *Topics in HIV medicine: a publication of the International AIDS Society, USA*, 13(1):51–57, 2005. [42](#), [135](#)
- [JD11] Jessie Jeng and John Daye. Sparse covariance thresholding for high-dimensional variable selection. *Statistica Sinica*, pages 625–657, 2011. [7](#), [39](#)
- [JJ18] Adel Javanmard and Hamid Javadi. False Discovery Rate Control via Debiased Lasso. *arXiv preprint arXiv:1803.04464*, 2018. [40](#)
- [JM14] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014. [7](#), [40](#), [44](#), [136](#)
- [JM15] Adel Javanmard and Andrea Montanari. De-biasing the Lasso: Optimal sample size for Gaussian designs. *arXiv preprint arXiv:1508.02757*, 2015. [1](#), [7](#), [38](#), [40](#), [134](#)
- [JZ13] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013. [2](#), [7](#), [13](#), [27](#), [46](#), [57](#), [58](#), [107](#)
- [KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [1](#)
- [KD99] Aart Kraay and John Driscoll. *Spatial Correlations in Panel Data*. The World Bank, 1999. [8](#), [52](#)

- [KH81] Harold Kushner and Hai Huang. Asymptotic properties of stochastic approximations with constant coefficients. *SIAM Journal on Control and Optimization*, 19(1):87–105, 1981. [15](#)
- [KL12] Jens-Peter Kreiss and Soumendra Nath Lahiri. Bootstrap methods for time series. In *Handbook of statistics*, volume 30, pages 3–26. Elsevier, 2012. [8](#)
- [KTSJ14] 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. [1](#)
- [Kun89] Hans Kunsch. The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, pages 1217–1241, 1989. [8](#), [52](#), [55](#)
- [KY03] H. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer New York, 2003. [6](#), [15](#)
- [Lah13] Soumendra Nath Lahiri. *Resampling methods for dependent data*. Springer Science & Business Media, 2013. [3](#), [8](#)
- [LH15] Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015. [30](#)
- [Lin65] John Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, pages 13–37, 1965. [55](#), [138](#)
- [LKLC18] Tianyang Li, Anastasios Kyrillidis, Liu Liu, and Constantine Caramanis. Approximate Newton-based statistical inference using only stochastic gradients.

- arXiv preprint arXiv:1805.08920*, 2018. [1](#), [2](#), [9](#), [26](#), [38](#), [51](#), [57](#), [76](#), [100](#), [131](#), [134](#), [137](#)
- [LLKC18] 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. [1](#), [2](#), [4](#), [9](#), [11](#), [32](#), [59](#), [77](#), [118](#)
- [LPW12] Lennart Ljung, Georg Ch Pflug, and Harro Walk. *Stochastic approximation and optimization of random systems*, volume 17. Birkhäuser, 2012. [6](#)
- [LSS14] Jason Lee, Yuekai Sun, and Michael Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014. [2](#), [7](#), [40](#), [45](#)
- [MB11] 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](#), [46](#)
- [MCF15] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015. [5](#)
- [MHB16] Stephan Mandt, Matthew Hoffman, and David Blei. A Variational Analysis of Stochastic Gradient Algorithms. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 354–363, 2016. [5](#), [15](#)
- [MHB17] Stephan Mandt, Matthew Hoffman, and David Blei. Stochastic Gradient Descent as Approximate Bayesian Inference. *arXiv preprint arXiv:1704.04289*, 2017. [5](#)

- [MMB09] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. p -values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009. [44](#), [136](#)
- [NRWY12] 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. [39](#), [100](#)
- [NW86] Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix. *National Bureau of Economic Research Cambridge, Mass., USA*, 1986. [52](#)
- [NW94] Whitney Newey and Kenneth West. Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):631–653, 1994. [52](#)
- [OM99] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999. [13](#)
- [Pfl86] Georg Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986. [15](#)
- [PJ92] Boris Polyak and Anatoli Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. [4](#), [6](#), [28](#), [30](#), [77](#)
- [PR92] Dimitris Politis and Joseph Romano. A circular block-resampling procedure for stationary data. *Exploring the limits of bootstrap*, pages 263–270, 1992. [8](#), [52](#), [55](#)
- [PR94] Dimitris Politis and Joseph Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313, 1994. [8](#), [52](#), [55](#)

- [PRW12] D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer Series in Statistics. Springer New York, 2012. [12](#), [30](#), [32](#), [132](#)
- [RLZ09] Adam Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009. [38](#), [39](#)
- [RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951. [14](#), [15](#)
- [RM73] Barr Rosenberg and Walt McKibben. The prediction of systematic and specific risk in common stocks. *Journal of Financial and Quantitative Analysis*, 8(2):317–333, 1973. [51](#)
- [RM85] Herbert Robbins and Sutton Monro. A stochastic approximation method. In *Herbert Robbins Selected Papers*, pages 102–109. Springer, 1985. [1](#)
- [RTW⁺06] Soo-Yon Rhee, Jonathan Taylor, Gauhar Wadhera, Asa Ben-Hur, Douglas L. Brutlag, and Robert W. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360, 2006. [42](#), [135](#)
- [Rup88] David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988. [4](#), [28](#), [30](#)
- [RWY11] Garvesh Raskutti, Martin Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011. [40](#)

- [SEG⁺17] 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. [32](#)
- [Sha64] William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964. [55](#), [138](#)
- [SS11] Robert Shumway and David Stoffer. Time series regression and exploratory data analysis. In *Time series analysis and its applications*, pages 47–82. Springer, 2011. [51](#)
- [ST12] Jun Shao and Dongsheng Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012. [3](#)
- [SV08] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008. [8](#)
- [SZ18] Weijie Su and Yuancheng Zhu. Statistical Inference for Online Learning and Stochastic Approximation via Hierarchical Incremental Gradient Descent. *arXiv preprint arXiv:1802.04876*, 2018. [2](#), [4](#)
- [TA14] Panos Toulis and Edoardo Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *arXiv preprint arXiv:1408.2923*, 2014. [6](#)
- [TA17] 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. [30](#)
- [Tro15] Joel Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015. [93](#)

- [TRVB06] 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. [1](#)
- [TWH15] R. Tibshirani, M. Wainwright, and T. Hastie. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015. [1](#)
- [Var75] James Varah. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11(1):3–5, 1975. [38](#), [106](#)
- [vdGBRD14] 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. [1](#), [7](#), [38](#), [40](#), [134](#)
- [vdV98] Aad W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998. [26](#), [32](#)
- [vdV00] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000. [10](#)
- [VGS05] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world: conformal prediction*. Springer, 2005. [8](#)
- [Wai09] Martin Wainwright. Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1 -Constrained Quadratic Programming (Lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009. [8](#)
- [Wai17] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. To appear, 2017. [101](#), [103](#), [111](#), [112](#)
- [Was13] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013. [1](#), [9](#), [26](#)

- [Whi80] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838, 1980. [7](#), [40](#), [41](#), [45](#), [50](#)
- [WT11] Max Welling and Yee Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011. [5](#)
- [XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017. [33](#)
- [XZ14] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014. [2](#), [7](#), [40](#), [41](#), [46](#), [48](#), [107](#), [108](#)
- [YLR14] Eunho Yang, Aurelie Lozano, and Pradeep Ravikumar. Elementary estimators for high-dimensional linear regression. In *International Conference on Machine Learning*, pages 388–396, 2014. [7](#), [39](#), [40](#)
- [ZZ14] 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. [7](#), [38](#), [40](#)