THE UNIVERSITY OF CHICAGO

MODERN STATISTICAL INFERENCE: PARAMETER ESTIMATION IN ONLINE SETTINGS AND GOODNESS-OF-FIT TESTING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
WANRONG ZHU

CHICAGO, ILLINOIS
JUNE 2024

To my family.

TABLE OF CONTENTS

LIST OF FIGURES viii

LIST OF TABLES x

ACKNOWLEDGMENTS xi

ABSTRACT xiii

1 INTRODUCTION 1
- 1.1 Statistical inference using stochastic gradient descent. 1
- 1.2 Goodness-of-fit testing. 3
- 1.3 Notation 4

2 ASYMPTOTIC NORMALITY FOR WEIGHTED AVERAGED STOCHASTIC GRADIENT DESCENT 6
- 2.1 Introduction 6
- 2.2 Main results: asymptotic normality 7
- 2.3 Examples 10
  - 2.3.1 Polynomial-decay averaging 11
  - 2.3.2 Suffix averaging 12
- 2.4 Non-asymptotic mean squared error 14
  - 2.4.1 Linear model 15
  - 2.4.2 A new averaging scheme: adaptive weighted averaging 16
- 2.5 Numerical experiment 18
  - 2.5.1 Asymptotic normality for different averaging schemes 18
  - 2.5.2 Non-asymptotic performance of different averaging schemes 20
- 2.6 Summary 23

3 SHARP CONCENTRATION ANALYSIS FOR STOCHASTIC GRADIENT DESCENT 24
- 3.1 Introduction 24
- 3.2 Upper bound 28
  - 3.2.1 Linear model setting 28
  - 3.2.2 Assumptions 30
  - 3.2.3 Nagaev type upper bound 31
  - 3.2.4 Technical overview and proof sketch for main results 33
- 3.3 Tightness of the upper bound 36
  - 3.3.1 Upper bound from Nagaev inequality 38
  - 3.3.2 Exact deviation 39
  - 3.3.3 A numerical study 40
- 3.4 Summary 41

4 ONLINE COVARIANCE MATRIX ESTIMATION IN STOCHASTIC GRADIENT DESCENT . . . 42
4.1 Introduction . . . 42
4.2 Online approach . . . 45
4.2.1 Online covariance matrix estimator based on batch means . . . 46
4.2.2 Statistical inference . . . 53
4.3 Theoretical results . . . 54
4.3.1 Preamble: mean estimation model . . . 54
4.3.2 Assumptions and existing convergence results . . . 55
4.3.3 Convergence properties for the online estimator . . . 57
4.3.4 Asymptotically accurate confidence intervals/regions . . . 58
4.4 Simulation studies . . . 61
4.4.1 Empirical performance of the proposed online approach . . . 61
4.4.2 Comparison with other methods . . . 63
4.5 Summary . . . 65

5 HIGH CONFIDENCE LEVEL INFERENCE IS ALMOST FREE USING PARALLEL STOCHASTIC OPTIMIZATION . . . 68
5.1 Introduction . . . 68
5.1.1 Background: existing confidence interval construction . . . 70
5.2 Inference with parallel runs of stochastic algorithms . . . 74
5.2.1 Parallel computing . . . 74
5.2.2 Asymptotic t-distribution . . . 75
5.3 Theoretical guarantee . . . 78
5.3.1 Convergence characterization for ASGD . . . 79
5.3.2 Main results . . . 82
5.4 Experiment . . . 84
5.4.1 Simulation . . . 84
5.4.2 Hand-written digit dataset . . . 90
5.5 Summary . . . 91

6 APPROXIMATE CO-SUFFICIENT SAMPLING WITH REGULARIZATION . . . 92
6.1 Introduction . . . 92
6.2 Background: goodness-of-fit testing via CSS and aCSS . . . 94
6.2.1 Co-sufficient sampling (CSS) . . . 95
6.2.2 Approximate co-sufficient sampling (aCSS) . . . 96
6.2.3 Additional related work . . . 99
6.3 The aCSS method with linear constraints . . . 100
6.3.1 Examples of constraints . . . 103
6.3.2 Formally defining the method . . . 104
6.4 Theoretical results . . . 110
6.4.1 Main result: Type I error control . . . 110
6.4.2 Special case: sparse structure . . . 113
6.4.3 Special case: Gaussian linear model . . . 117

6.5 Generalization of linear constraint: ℓ_1 penalty 120
6.5.1 The conditional density in the penalized case 121
6.5.2 The aCSS method in the penalized case 123
6.6 Numerical experiments 125
6.6.1 Necessary constraints: the Gaussian mixture model 125
6.6.2 High dimensional setting: structured Gaussian linear model 128
6.7 Summary 137

REFERENCES 138

A APPENDIX FOR CHAPTER 2 149
A.1 Technical Lemmas and Proofs 149
A.1.1 Technical Overview and Proof Sketch of the Main Theorem 150
A.1.2 Proof of Theorem 2.2.4 151
A.1.3 Proof of Lemma A.1.2 154
A.1.4 Proof of Lemma A.1.3 156
A.1.5 Proof of Corollary 2.3.1 159
A.1.6 Proof of Corollary 2.3.2 162
A.1.7 Proof of Proposition 2.4.1 163
A.1.8 Proof of Corollary 2.4.2 168
A.1.9 Proof of Lemma A.1.4 171
A.1.10 Proof of Lemma A.1.6 173

B APPENDIX FOR CHAPTER 3 175
B.1 Proofs 175
B.1.1 Some useful lemmas 175
B.1.2 Proof of Lemma 3.2.8 176
B.1.3 Proof of Lemma 3.2.9 177
B.1.4 Proof of Theorems 3.2.4, 3.2.6 180
B.1.5 Proof of Proposition 3.3.1 180
B.1.6 Proof of Proposition 3.3.2 181

C APPENDIX FOR CHAPTER 4 183
C.1 Technical Lemmas 183
C.2 The Linear Case 186
C.3 Proof of Main Theorems 207
C.3.1 Proof of Theorem 4.3.5 207
C.3.2 Proof of Theorem 4.3.8 212
C.4 Proof of Proposition 4.3.1 218
C.5 Simulation for stopping rule 220

D APPENDIX FOR CHAPTER 5 222
D.1 Proof 222
D.1.1 Proof of Theorem 5.3.4 222

D.1.2 Proof of Theorem 5.3.8 225
D.2 Additional numerical results 228

E APPENDIX FOR CHAPTER 6 232
E.1 Proofs of main results 232
E.1.1 Proof of Theorems 6.4.3, 6.4.5: error control for constrained aCSS . . 232
E.1.2 Proof of Theorem 6.4.7: constrained aCSS for the Gaussian linear model242
E.1.3 Proof of Lemma 6.3.3: conditional density 244
E.1.4 Proof of Theorem 6.5.3: error control for aCSS with an ℓ_1 penalty . . 250
E.1.5 Proof of Lemma 6.5.2: conditional density for aCSS with an ℓ_1 penalty 251
E.2 Additional proofs 256
E.2.1 Verifying that the plug-in version of $p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g})$ defines a density . . . 256
E.2.2 Proof of Lemma E.1.1 258
E.2.3 Proof of Lemma 6.4.4 259
E.3 Checking assumptions for examples 260
E.3.1 Verifying assumptions for Examples 2 (isotonic regression) and 3 (sparse regression) 260
E.3.2 Verifying assumptions for Example 1 (Gaussian mixture model) . . . 262
E.4 Sampling details 267
E.4.1 Implementation details for Example 1 (Gaussian mixture model . . . 268
E.4.2 Implementation details for Examples 2 (isotonic regression) and 3 (sparse regression) 271

LIST OF FIGURES

2.1 Realizations of online suffix averaging. Here $a_m, m \geq 0$, is the index of the staring point of the m-th block. 13
2.2 Density plot for the standardized error with and without prefactor w. The red line denotes a standard normal distribution. 20
2.3 Left: log-log plots for MSE. Right: the curves stand for the ratio of MSE between different averaging schemes and adaptive weighted averaging at each step. The baseline (black line) is for the adaptive weighted averaging. 21
2.4 Comparison of different weights under expectile regression model with $\rho = 0.8$. The oracle weights are numerically computed via Monte-Carlo simulation with 50000 repetitions. 22

3.1 Compare the two terms in sample complexity (3.2). Here X axis represents failure probability δ; the solid line denotes $\epsilon^{-1}\log(1/\delta)$, the dashed line denotes $(\delta^{-2/q}\epsilon^{-1})^{q/(2(q-1))}$. We choose $\epsilon = 0.01$ and $q = 2.5$. 28
3.2 Ratio of approximated and true tail probability. Here X axis represents deviation x. Red curves represent Gaussian approximation: $\left(1 - \Phi(x/\sqrt{\mu_{T,2}})\right)/\mathbb{P}(S_T \geq x)$; blue curves represent tail approximation: $R(T,x)/\mathbb{P}(S_T \geq x)$; black curves represent their sum: $\left(1 - \Phi(x/\sqrt{\mu_{T,2}}) + R(T,x)\right)/\mathbb{P}(S_T \geq x)$. 40

4.1 Linear regression: Log loss (operator norm) of the estimated covariance matrix against the log of total number of steps. Here F denotes the full overlapping version (4.3), NOL denotes the non-overlapping version (4.8), and C denotes the constant in $a_m = \left\lfloor Cm^{2/(1-\alpha)} \right\rfloor$. 62
4.2 Relative efficiency (ratio of MSE) of the full overlapping version (4.3) and non-overlapping version (4.8). We set $d = 5$ in linear regression. Here C denotes the constant in $a_m = \left\lfloor Cm^{2/(1-\alpha)} \right\rfloor$. 63
4.3 Linear regression with $d = 5$: (a): Empirical coverage rate against the number of steps. Red dashed line denotes the nominal coverage rate of 0.95. (b): Length of confidence intervals. (c): Density plot for the standardized error. Red curve denotes the standard normal density. 64
4.4 Comparison of online-BM and Plug-in estimators. First/Middle row: Empirical coverage rate against the number of steps in linear/logistic model. Red dashed line denotes the nominal coverage rate of 0.95. Third row: total computation time for updating covariance estimate and confidence intervals in SGD. 66

5.1 Effect of K. Plot (a): relative error of coverage; plot (b): the length of confidence interval. The nominal coverage probability is 0.99. The total sample size N is 60000 for linear models and 200000 for logistic models. 85
5.2 Linear Regression $d = 20$: Left: relative error of coverage; Middle: empirical coverage; Right: length of confidence intervals. 87

5.3 Logistic Regression $d = 20$: Left: relative error of coverage; Middle: empirical coverage; Right: length of confidence intervals. 88
5.4 Computation time: d = 20 89
5.5 Mean image before and after denoising. (a) shows the original estimated mean before denoising; (b) uses a uniform threshold $t = 0$; (c) uses a uniform threshold $t = -0.5$; (d) applies an adaptive threshold based on the upper bound of the confidence interval. 91

6.1 Power of the regularized (i.e., constrained) aCSS method, denoted as reg-aCSS in the plot, as compared to the oracle method. The oracle method knows the true parameter and samples (unconditionally) from the simple null. The constrained aCSS method controls the Type I error at the nominal 5% level (red dotted line) under the null. All tests are repeated for 500 independent trials. 129
6.2 Power of aCSS, regularized (i.e., constrained or penalized) aCSS (denoted as reg-aCSS in the plot), and the oracle method, for isotonic regression (left) and sparse regression (right), with different values of the parameter σ, over 5000 independent trials. The dotted red line denotes the nominal 10% level (i.e., $\alpha = 0.1$). For both settings, $\beta_0 = 0$ corresponds to the null hypothesis being true. 135
6.3 Type I error rate of aCSS, regularized (i.e., constrained or penalized) aCSS (denoted as reg-aCSS in the plot), and the oracle method, for isotonic regression (left) and sparse regression (right), with different values of the parameter σ, over 5000 independent trials. The dotted red line denotes the nominal 10% level (i.e., $\alpha = 0.1$). The shaded bands denote standard error for each method. 136
6.4 Histogram of p-values under the null, for aCSS and for regularized (i.e., constrained or penalized) aCSS, for isotonic regression (left) and sparse regression (right), over 5000 independent trials. The parameter σ is chosen as $\sigma = 7$ for both examples. 136

B.1 Left: Check the order of $\mu_{T,2}$. The X axis represents $\log(t)$; the Y axis represents $\log(\mu_{t,2})$. The slop of the log-log curve is about -1, which implies that $\mu_{T,2} \asymp T^{-1}$. Right: Check the order of V_t. The X axis represents $\log(t)$; the Y axis represents $\log(V_t)$. The slop of the log-log curve is much less than -1 when t is large, which means V_t is summable and $\sum_{i=1}^{T} V_i = O(1)$. 181
B.2 Check the uniform asymptotic negligibility of the variance of individual summands. The X axis represents t; the Y axis represents the ratio of the largest individual variance and variance of individual summands. 182

D.1 Linear Regression $d = 5$: Left: relative error of coverage; Middle: empirical coverage; Right: length of confidence intervals. 229
D.2 Logistic Regression $d = 5$: Left: relative error of coverage; Middle: empirical coverage; Right: length of confidence intervals. 230
D.3 Computation time: d = 5 231

LIST OF TABLES

4.1 Empirical coverage rates: the average coverage rate for the nominal coverage probability 95%. Standard errors are reported in the brackets. 65

5.1 Comparison of methods for online statistical inference: This table compares various methods including Plug-In [Chen et al., 2020], Online BM [Zhu et al., 2023], and Random Scale [Lee et al., 2022]. The symbol '$\checkmark$' indicates that the method can achieve the goal, while '/' signifies that it cannot. The rating symbols '$\star\star\star$', '$\star\star$', and '$\star$' denote the best, moderate, and lowest advantage, respectively. . . . 73

C.1 Apply fixed-width sequential stopping rule with the tolerance 0.01 (discussed in Section 4.3.4). We present termination iterations and coverage probabilities at termination. Standard errors are reported in the brackets. 221

D.1 Asymptotic one-sided critical values for asymptotic pivotal statistic in the random scaling method (5.4) via Monte Carlo simulation with 1,000,000 samples. 228

ACKNOWLEDGMENTS

First and foremost, I extend my heartfelt gratitude to my wonderful advisors, Profs. Rina Foygel Barber and Wei Biao Wu, for their unwavering guidance and support throughout the past years. My PhD journey would not have been as enriching without their presence. I am deeply indebted to Wei Biao for introducing me to the realm of research. His encouragement and sense of optimism have been pivotal in helping me navigate my PhD journey. I am also incredibly fortunate to have Rina alongside me in this endeavor, serving as my role model. Her dedication to research, meticulousness with students, and exemplary work ethic will continue to inspire and guide me in my future endeavors.

I also want to thank my committee member, Prof. Chao Gao, for valuable feedback on my thesis and for providing the best lectures, which have been instrumental in my research. I am also grateful to Prof. Xi Chen for valuable instructions in our collaboration and for support during my job search. My appreciation extends to all faculty and staff in the department.

I am immensely grateful for the intellectual stimulation provided by my fellow academic peers, especially my talented academic siblings, Yu Gui, Yuefeng Han, Qijia Li, Ruiting Liang, Yonghoon Lee, Yuetian Luo , Zhipeng Lou, Jake Soloff. Their insightful discussions and support during my job search have been invaluable. I am equally thankful to my other fellow doctoral students, particularly Hai Tran Bach, Pinhan Chen, Zihao Wang, Chih-Hsuan Wu, Daniel Xiang, Zhishen Xiao (and his wife, Peijun Xiao), Qing Yan, for making my PhD journey vibrant. I will always cherish the memories of our challenging hikes, late-night board game sessions, and gym workouts. I also want to thank Rui Da, Haoyang Liu, Sen Na, Yao Tong and Yuexi Wang for being most reliable friends, offering invaluable advice, encouragement and a lot of joy during my difficult times. Additionally, I am grateful to Dongyue for his companionship over the years, for exploring countless adventures together, and for infusing everyday life with happiness.

Lastly, but certainly not least, I want to thank my family. To my mother Weijun Xu and father Huiliang Zhu, your unconditional support, trust and love make me who I am today. I am also thankful to my grandparents and other big family members for fostering a nurturing environment and reminding me of the profound love that surrounds me. I hope I can always stay brave and keep thinking throughout the rest of my life, and I believe I can do so with all of your love.

ABSTRACT

Statistical inference plays a crucial role in realizing large-scale intelligent systems that can learn safely and efficiently. This thesis presents two interesting yet challenging problems of modern statistical inference.

In the first part, we consider statistical inference for estimation in online settings. Model parameter estimation through optimization is a classical problem in statistics and machine learning. Algorithms based on stochastic approximation, particularly stochastic gradient descent (SGD) and its variants, have emerged as the workhorses for solving such problems in modern statistical and machine learning. Despite SGD's tremendous success in practical applications, one cost of the SGD algorithm is the uncertainty of solutions. A crucial aspect of this thesis is to understand the variability inherent in these solutions and perform practical statistical inference. We will discuss both theoretical aspects of inference as well as methods to conduct practical inference. From the theoretical perspective, topics include studying the limiting distribution, where we extend the classic asymptotic normality results for averaged SGD to a general case of weighted averaged SGD. Beyond asymptotic distribution, we also study the concentration properties of SGD solutions under heavy-tailed noise settings. To provide a practical methodology, we introduce an approach to estimate the limiting covariance matrix of SGD estimates in an online fashion and construct confidence intervals as a byproduct. When only confidence intervals are of interest, we further introduce a more computationally efficient way to construct confidence intervals directly without estimating the covariance matrix and enable testing related to high-level confidence.

In the second part, we consider the problem of goodness-of-fit (GoF) testing for parametric models. This testing problem involves a composite null hypothesis, due to the unknown values of the model parameters. In some special cases, co-sufficient sampling (CSS) can remove the influence of these unknown parameters via conditioning on a sufficient statistic—often, the maximum likelihood estimator (MLE) of the unknown parameters. And the recent approximate co-sufficient sampling (aCSS) framework replacing sufficiency with an approximately sufficient statistic (namely, a noisy version of the MLE) to recovers power in a range of settings where CSS leads to a powerless test, but can only be applied in settings where the unconstrained MLE is well-defined and well-behaved, which implicitly assumes a low-dimensional regime. We extend aCSS to the setting of constrained and penalized maximum likelihood estimation, so that more complex estimation problems can now be handled within the aCSS framework, including those in high-dimensional settings.

CHAPTER 1

INTRODUCTION

In this thesis, we present two problems of modern statistical inference: statistical inference in online settings using stochastic gradient descent and goodness-of-fit testing in a high-dimensional setting.

1.1 Statistical inference using stochastic gradient descent.

In the first part, we consider statistical inference for estimation in online settings. Model parameter estimation through optimization of an objective function is a fundamental problem in statistics and machine learning. Here we consider the classic setting where the true model parameter $x^* \in \mathbb{R}^d$ can be characterized as the minimizer of a convex objective function $F : \mathbb{R}^d \to \mathbb{R}$, i.e.,

$$x^* = \underset{x \in \mathbb{R}^d}{\arg\min}\ F(x). \tag{1.1}$$

The objective function $F(x)$ is defined as $F(x) = \mathbb{E}_{\xi \sim \Pi} f(x, \xi)$, where $f(x, \xi)$ is a noisy measurement of $F(x)$ and ξ is a random variable following the distribution Π. For example, in linear regression, the coefficient can be modeled as the minimizer of the expected squared loss. In logistic regression, a linear classifier is derived by minimizing the expected log loss.

In recent years, huge data sets and streaming data arise frequently. Classic deterministic optimization methods that require storing all the data are not appealing due to expensive memory cost and computational inefficiency. To resolve these issues, one can apply the Robbins-Monro algorithm [Robbins and Monro, 1951], also known as Stochastic Gradient Descent (SGD), especially for online learning [Bottou, 1998, Mairal et al., 2010, Hoffman et al., 2010]. Setting x_0 as the initial point, the i-th iteration of the SGD algorithm takes the following form

$$x_i = x_{i-1} - \eta_i \nabla f(x_{i-1}, \xi_i),\ i \geq 1, \tag{1.2}$$

where $\{\xi_i\}_{i\geq 1}$ is a sequence of *i.i.d* samples from the distribution Π, ∇f is the gradient of $f(x, \xi)$ with respect to the first argument x, and η_i is the step size at the i-th step. This recursive adaptive algorithm performs one update at a time and does not need to remember outcomes in previous iterations. Therefore, it is computationally efficient, memory friendly, and able to process data on the fly. Despite SGD's tremendous success in practical applications, one cost of the SGD algorithm is the uncertainty of solutions. A crucial aspect of the research in this thesis is to understand the variability inherent in these solutions and perform uncertainty quantification.

The first natural question pertains to characterizing the distribution of SGD solutions. In Chapter 2, we establish the asymptotic normality of generalized weighted averaged SGD solutions under a set of mild assumptions. This extends and refines the results in the seminal work associated with Polyak-Ruppert averaging.[1] In addition to asymptotic distributional guarantees, practitioners often seek assurances regarding the performance stability of a single trial of an algorithm. This underpins the research in Chapter 3, where we explore the concentration properties of SGD solutions.[2] Traditional concentration analyses often impose restrictive conditions on the gradient noise, such as boundedness or sub-Gaussian traits. We consider a broader class of noise where only finitely many moments are required, thus accommodating heavy-tailed noise.

The next question is how to perform practical inference leveraging theoretical distribution characteristics. In Chapter 4, we discuss the estimation of the asymptotic covariance.[3] This task is particularly challenging due to the dependencies between SGD iterates and the goal of maintaining computational and memory efficiency in a fully online context. We introduce a fully online estimator for the asymptotic covariance of the averaged SGD (ASGD), which

[1] The paper corresponding to the work discussed in this chapter is available on arXiv:2307.06915

[2] The paper corresponding to the work discussed in this chapter was published in Journal of Machine Learning Research 23 (46), 1-22

[3] The paper corresponding to the work discussed in this chapter was published in Journal of the American Statistical Association 118 (541), 393-404

is also extendable to other SGD variants. This method possesses two key attributes. First, it utilizes solely the SGD iterates without the need for additional information. Second, it updates recursively with the arrival of new data, thereby aligning with the online nature of SGD while maintaining desired computational and memory efficiencies. With the estimated covariance matrix and asymptotic normality results, one can construct asymptotically exact confidence intervals. In tasks where only a confidence interval is needed and not the covariance matrix, we introduce a more computationally efficient method in Chapter 5.[4] The approach involves dividing the original SGD path into K independent and identically distributed runs. Based on the final estimates from these sequences, a t-statistic and t-based confidence interval is constructed. This method is free from covariance matrix estimation, requires minimal extra computation and memory beyond SGD updates, and provides valid coverage at exceedingly high confidence levels.

1.2 Goodness-of-fit testing.

In the second part, we will study Goodness-of-fit (GoF) testing. GoF testing is an essential statistical method, widely used in various fields such as biology, economics, engineering, and finance, to assess whether the observed data follows a certain pattern or distribution that is expected based on theoretical assumptions. Given data X belonging to some sample space $\mathcal{X}$, the fundamental problem addressed by GoF is the question of testing the null hypothesis

$$H_0 : X \sim P_\theta \text{ for some } \theta \in \Theta, \tag{1.3}$$

where $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ is a parametric family, versus a more complex (usually higher-dimensional) model. For example, we may be interested in testing whether a logistic regression model is appropriate for our binary data $X = (X_1, \ldots, X_n)$ (in the presence of some

[4] The paper corresponding to the work discussed in this chapter is available on arXiv:2401.09346

covariates), or whether a more complex—perhaps even nonparametric—model is needed.

In Chapter 6, we will reduce the testing problem to a sampling problem *How can we generate copies $\tilde{X}^{(1)}, ..., \tilde{X}^{(M)}$ of the observed data X such that, if H_0 is true, then $X, \tilde{X}^{(1)}, ..., \tilde{X}^{(M)}$ are (approximately) exchangeable?*[5] The difficulty of the problem lies in the composite null or unknown true parameter. We will provide an overview of related sampling techniques including co-sufficient sampling (CSS) and approximate co-sufficient sampling (aCSS). These methods avoid this issue by conditioning on a sufficient (or approximately sufficient) statistic for the unknown θ, but are not suited for addressing challenges such as high dimensionality. We will then introduce an extended version of aCSS that can accommodate more complex problems where robust and accurate parameter estimation is needed, particularly in high-dimensional settings.

1.3 Notation

Here we list notations that we will use throughout the thesis. For notations that will be used only in a specific chapter, we will introduce them therein. For a vector v, $\|v\|_0$ denotes the ℓ_0 norm (the number of nonzero entries), and $\|v\|_q$ denotes the usual ℓ_q norm for $1 \leq q \leq \infty$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{d\times d}$, $\|\mathbf{A}\|_F$ denotes its Frobenius norm $\|\mathbf{A}\|_F = \left(\sum_{i=1}^d \sum_{j=1}^d a_{ij}^2\right)^{1/2}$, $\|\mathbf{A}\|_2$ denotes its operator norm $\|\mathbf{A}\|_2 = \max_{\|x\|_2\leq 1} \|\mathbf{A}x\|_2$, $tr(\mathbf{A})$ denotes its trace, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denotes its largest and smallest eigenvalues. We use $\mathbf{I}_d$ to denote a $d \times d$ identity matrix, and $\mathbf{1}_d$ to denote the vector in $\mathbb{R}^d$ with all entries 1. We use $\mathbb{1}\{\mathcal{E}\}$ to denote the indicator variable for event $\mathcal{E}$. For $t \in \mathbb{R}$, $\lfloor t \rfloor$ is the largest integer less than or equal to t, and $\lceil t \rceil$ is the smallest integer greater than or equal to t. For positive sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, $a_n \lesssim b_n$ means there exists some constant C such that $a_n \leq Cb_n$ for all large n, and $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold. For notational simplicity, we use notation C for constants which can take different values

[5] The paper corresponding to the work discussed in this chapter is available on arXiv:2309.08063

in different equations. For a sequence of *i.i.d.* sample $\{\xi_i\}_{i\geq 1}$ from some distribution Π, we define conditional expectation $\mathbb{E}_n(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_n)$ and conditional probability $\mathbb{P}_n(\cdot) = \mathbb{P}(\cdot|\mathcal{F}_n)$. , where $\mathcal{F}_n$ is σ-algebra generated by $\{\xi_i\}_{i\leq n}$. Moreover, we use $\Rightarrow$ to denote convergence in distribution. Lastly, we use O to denote the Big-O notation.

CHAPTER 2

ASYMPTOTIC NORMALITY FOR WEIGHTED AVERAGED STOCHASTIC GRADIENT DESCENT

2.1 Introduction

The asymptotic convergence of SGD iterates has been studied extensively in the early years [Blum, 1954, Dvoretzky, 1956, Sacks, 1958, Fabian, 1968, Robbins and Siegmund, 1971, Ljung, 1977, Lai, 2003]. To further investigate the asymptotic distribution of SGD, Polyak and Juditsky [1992] and Ruppert [1988] introduced the averaged SGD (ASGD), a simple modification where iterates are averaged, and established the asymptotic normality of the obtained estimate. It is known that ASGD estimates achieve the optimal central limit theorem rate $O(1/\sqrt{n})$ by running SGD for n iterations under certain regularity conditions. However, it is not optimal from a non-asymptotic perspective [Moulines and Bach, 2011, Needell et al., 2014]. Moreover, for non-smooth objective functions, neither the final iterate nor ASGD can achieve the optimal convergence rate. To address these issues, various modified versions of ASGD have been proposed, such as suffix averaging [Rakhlin et al., 2012] and polynomial-decay averaging [Shamir and Zhang, 2013] for non-smooth problems, exponential weighted moving average (EWMA) for capturing time variation, elastic averaging in parallel computing environments [Zhang et al., 2015] and a simple weight proportional to $O(n)$ for the projected stochastic subgradient method [Lacoste-Julien et al., 2012].

As mentioned earlier, employing a suitable averaging scheme in specific settings can help accelerate convergence and necessitates only a straightforward modification to the original SGD algorithm. In this chapter, we will delve deeper into the characteristics of averaging by examining a comprehensive averaging scheme. Our main objectives include understanding the variability and statistical efficiency of a general weighted averaged SGD. Under certain mild assumptions, we establish the asymptotic normality of these general weighted averaged

SGD solutions. This result is applicable to a wide range of existing algorithms, including the polynomial-decay and suffix averaged SGD. Beyond asymptotic normality, we also investigate finite sample convergence. When considering finite sample MSE, it is challenging to identify a single averaging scheme that is optimal for all objective functions. To gain insights, we examine the linear model to derive adaptively weighted SGD iterates that minimize the finite sample MSE. Our findings indicate that the optimal weight derived from the linear model not only achieves the optimal statistical rate but also exhibits favorable non-asymptotic convergence on other models.

2.2 Main results: asymptotic normality

Recall the problem in (1.1), i.e.,

$$x^* = \underset{x \in \mathbb{R}^d}{\arg\min}\ F(x),$$

where $F(x)$ is defined as $F(x) = \mathbb{E}_{\xi \sim \Pi} f(x, \xi)$, $f(x, \xi)$ is a noisy measurement of $F(x)$ and ξ is a random variable following the distribution Π. The SGD sequence $\{x_i\}_{i \geq 1}$ is defined in (1.2). Denote $\bar{x}_n$ as the uniform average, i.e., $\bar{x}_n = (1/n) \sum_{i=1}^n x_i$. Under certain assumptions and conditions, Polyak and Juditsky [1992] shows that

$$\sqrt{n}(\bar{x}_n - x^*) \Rightarrow \mathcal{N}(0, A^{-1} S A^{-1}), \tag{2.1}$$

where $A = \nabla^2 F(x^*), S = \mathbb{E}([\nabla f(x^*, \xi)][\nabla f(x^*, \xi)]^T)$. In this chapter, one of our goals is to establish the asymptotic normality for the general weighted average

$$\tilde{x}_n = \sum_{i=1}^n w_{n,i} x_i,$$

where $w_{n,i}$ denotes the weight of x_i after the n-th update, $1 \leq i \leq n$. Define gradient noise $\epsilon_i = \nabla F(x_{i-1}) - \nabla f(x_{i-1}, \xi_i)$. We leverage the martingale CLT on $\sqrt{n}\sum_{i=1}^{n} w_{n,i}A^{-1}\epsilon_i$ to obtain the conclusion: under certain assumptions, we have

$$\sqrt{n}(\tilde{x}_n - x^*) \Rightarrow \mathcal{N}(0, wA^{-1}SA^{-T}),$$

where $w = \lim_{n\to\infty} n\sum_{i=1}^{n}(w_{n,i})^2$. This result holds for many existing averaging schemes, as well as the adaptive averaging in Section 2.4.2.

We begin by introducing several assumptions.

Assumption 2.2.1. *The objective function $F(x)$ is continuously differentiable and strongly convex with parameter $\mu > 0$. That is, for any x_1 and x_2,*

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

Further assume that $\nabla^2 F(x^*)$ exists and $\nabla F(x)$ is Lipschitz continuous with parameter L, i.e., for any x_1 and x_2 we have,

$$||\nabla F(x_1) - \nabla F(x_2)||_2 \leq L||x_1 - x_2||_2.$$

Assumption 2.2.2. *The function $f(x, \xi)$ is continuous differentiable with respect to x for any ξ and $||\nabla f(x, \xi)||_2$ is uniformly integrable for any x.*

Assumption 2.2.1 requires strong convexity and Lipschitz continuity, which are needed to derive the asymptotic normality of ASGD solutions. These properties are also important for obtaining the desired error bounds on SGD iterates and the asymptotic properties of weighted averaged SGD. Assumption 2.2.2 ensures that Leibniz's integration rule holds. Consequently, $\epsilon_n = \nabla F(x_{n-1}) - \nabla f(x_{n-1}, \xi_n)$ is a martingale difference, i.e., $\mathbb{E}_{n-1}(\epsilon_n) = 0$.

Assumption 2.2.3. *Recall the gradient noise $\epsilon_i = \nabla F(x_{i-1}) - \nabla f(x_{i-1}, \xi_i)$. There exists a constant C_1 such that the fourth conditional moment of ϵ_n is bounded as:*

$$\mathbb{E}_{n-1}(\|\epsilon_n\|_2^4) \leq C_1(1 + ||\delta_{n-1}||_2^4),$$

where $\delta_n = x_n - x^*$ is the error sequence. In addition, the conditional covariance of ϵ_n satisfies

$$||\mathbb{E}_{n-1}(\epsilon_n \epsilon_n^T) - S||_2 \leq C_2(||\delta_{n-1}||_2 + ||\delta_{n-1}||_2^2),$$

for some constant C_2.

Assumption 2.2.3 is a mild condition concerning the boundedness of the loss function. It holds when the (matrix norm of the) Hessian of $f(x, \xi)$ is bounded by some random function $H(\xi)$ with a bounded fourth moment. Easily verified examples include linear and logistic regression. Similar assumptions are proposed in Polyak and Juditsky [1992] for the asymptotic normality of ASGD. We will also use these assumptions in later chapters.

With the aforementioned assumptions in place, we will now present our main results regarding the asymptotic normality for a general weighted averaging scheme.

Theorem 2.2.4. *Given SGD iterates in* (1.2) *with step size $\eta_i = \eta i^{-\alpha}$ for some $\eta > 0$ and $0.5 < \alpha < 1$, we consider a general averaging scheme:*

$$\tilde{x}_n = \sum_{i=1}^{n} w_{n,i} x_i, \tag{2.2}$$

where the weight $w_{n,i}$ satisfies the following conditions:

1. *$\sum_{i=1}^{n} w_{n,i} = 1$, $|w_{n,i}| \leq Cn^{-1}$ for some constant C.*

2. *$w = \lim_{n\to\infty} n \sum_{i=1}^{n} (w_{n,i})^2$ exists.*

3. smoothness condition: for $\lambda = \min(\lambda_{\min}(A), \frac{1}{2\eta})$,

$$\lim_{n\to\infty} \sum_{i=1}^{n} \sum_{k=i+1}^{n} |w_{n,k} - w_{n,i}| \eta_i \exp(-\lambda \sum_{t=i+1}^{k} \eta_t) = 0.$$

Then under Assumptions 2.2.1-2.2.3, the weighted averaged SGD is asymptotically normal

$$\sqrt{n}(\tilde{x}_n - x^*) \Rightarrow \mathcal{N}(0, wA^{-1}SA^{-1}),$$

where $A = \nabla^2 F(x^*)$*, and* $S = \mathbb{E}([\nabla f(x^*, \xi)][\nabla f(x^*, \xi)]^T)$.

The asymptotic covariance of the weighted average $\tilde{x}_n$ here is composed of a prefactor w and the sandwich form $A^{-1}SA^{-1}$ ($A^{-1}SA^{-1}$ is the asymptotic covariance matrix of ASGD). In the context of ASGD, where $w_{n,i} = 1/n$, the prefactor $w = 1$, which aligns with our results. The smoothness condition requires that the majority of weights should not undergo drastic changes. A slightly stronger yet simpler condition is $|w_{n,i+1} - w_{n,i}| \leq \tilde{C} n^{-2}$ for some constant $\tilde{C} > 0$, as shown in the appendix.

2.3 Examples

In this section, we apply our results to two specific examples of averaging schemes: polynomial-decay averaging [Shamir and Zhang, 2013] and suffix averaging [Rakhlin et al., 2012]. We also modify suffix averaging to an online fashion. While these schemes are known for achieving $O(1/n)$ convergence rates in non-smooth settings, we will show that they are not statistically optimal, with a constant prefactor that is strictly greater than unity.

2.3.1 Polynomial-decay averaging

With a small number $\gamma \geq 1$, the polynomial-decay averaging [Shamir and Zhang, 2013] is defined as follows: given iterates $\{x_i\}_{i=1}^{\infty}$, define $\tilde{x}_1 = x_1$, and for any $n \geq 1$,

$$\tilde{x}_n = (1 - \frac{\gamma+1}{\gamma+n})\tilde{x}_{n-1} + \frac{\gamma+1}{\gamma+n}x_n. \tag{2.3}$$

The recursion form in (2.3) can be rewritten as the weighted average $\tilde{x}_n = \sum_{i=1}^{n} w_{n,i}x_i$ with weight $w_{n,i} = \theta_{n,i}$ and

$$\theta_{n,i} = \frac{\gamma+1}{\gamma+i}\prod_{j=i+1}^{n}\frac{j-1}{j+\gamma} = \frac{\gamma+1}{n}\frac{\Gamma(\gamma+i+1)\Gamma(n+1)}{\Gamma(\gamma+n+1)\Gamma(i+1)},$$

where $\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt, x > 0$, is the Gamma function. The weight $w_{n,i} = \theta_{n,i}$ satisfies conditions in Theorem 2.2.4. Moreover,

$$\lim_{n\to\infty} n\sum_{i=1}^{n}(w_{n,i})^2 = \frac{(\gamma+1)^2}{2\gamma+1}.$$

Therefore we have the following asymptotic normality:

Corollary 2.3.1. *Consider SGD iterates in* (1.2) *with step size* $\eta_i = \eta i^{-\alpha}$ *for* $\eta > 0$, $0.5 < \alpha < 1$, *and polynomial-decay averaging* $\tilde{x}_n$ *defined in* (2.3). *Under Assumptions 2.2.1-2.2.3, we have*

$$\sqrt{n}(\tilde{x}_n - x^*) \Rightarrow \mathcal{N}\left(0, \frac{(\gamma+1)^2}{2\gamma+1}A^{-1}SA^{-1}\right).$$

Since $(\gamma+1)^2/(2\gamma+1) > 1$, the covariance of polynomial-decay averaged SGD is larger than that of ASGD.

2.3.2 Suffix averaging

The κ-suffix averaging in Rakhlin et al. [2012] is defined as the average of the last $\lceil \kappa n \rceil$ iterates of $\{x_i\}_{i=1}^{\infty}$ for $0 < \kappa < 1$,

$$\tilde{x}_n = \frac{1}{\lceil \kappa n \rceil} \sum_{i=\lceil (1-\kappa)n \rceil}^{n} x_i. \tag{2.4}$$

For κ-suffix averaging, the weight $w_{n,i} = 1/\lceil \kappa n \rceil$ for $i > (1-\kappa)n$ otherwise 0. The weight satisfies conditions in Theorem 2.2.4 with

$$\lim_{n\to\infty} n \sum_{i=1}^{n} (w_{n,i})^2 = \frac{1}{\kappa}.$$

Therefore the κ-suffix averaged SGD is also asymptotically normal.

Corollary 2.3.2. *Consider SGD iterates in* (1.2) *with step size* $\eta_i = \eta i^{-\alpha}$ *for* $\eta > 0$, $0.5 < \alpha < 1$, *and* $\tilde{x}_n$ *defined in* (2.4). *Under Assumptions 2.2.1-2.2.3, we have*

$$\sqrt{n}(\tilde{x}_n - x^*) \Rightarrow \mathcal{N}(0, \frac{1}{\kappa} A^{-1} S A^{-1}).$$

Since $1/\kappa > 1$, the covariance of κ-suffix averaged SGD is larger than that of ASGD.

Remark 2.3.3 (Online algorithm for suffix averaging)**.** *Since* $(1-\kappa)n$ *depends on* n, *the* κ-*suffix averaging cannot be computed on-the-fly. To enable online update, we modify the suffix averaging procedure to an online method. We employ the concept of online batch scheme: divide the rounds into blocks and track iterations within the current block (or the most recent blocks). The block sizes are pre-defined based on various objectives and training parameters. For a pre-defined sequence* $(a_m)_{m\geq 0}$, *we treat* x_{a_m} *as the start of the* m*-th block. Let* m_t *denote the block index for the* t*-th iteration, satisfying* $a_{m_t} \leq t < a_{m_t+1}$.

In the online suffix averaging procedure, we partition the rounds into exponentially in-

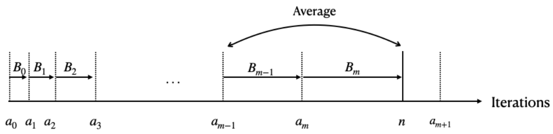


Figure 2.1: Realizations of online suffix averaging. Here $a_m, m \geq 0$, is the index of the staring point of the m-th block.

creasing blocks, and maintain the average of the last two blocks; see Figure 2.1 for an example of possible realizations. In particular, we set $a_m = \lfloor 2^{m-1} \rfloor + 1, m \geq 0$. *Then* $B_0 = \{x_1\}, B_1 = \{x_2\}, B_2 = \{x_3, x_4\}, B_3 = \{x_5, x_6, x_7, x_8\}, ...,$ *and the end index of the m-th block is* 2^m. *We have* $m_t = \lceil \log_2 t \rceil$, *i.e.,* $\lfloor 2^{m_t - 1} \rfloor < t \leq 2^{m_t}$. *Given the sequence of SGD iterates* $x_1, x_2 ...$, *the online suffix averaging procedure is defined as follows:* $\hat{x}_1 = x_1, \hat{x}_2 = (x_1 + x_2)/2$, *and for any* $t > 2$

$$\hat{x}_t = \frac{1}{t - 2^{\lceil \log_2 t \rceil - 2}} \left(\sum_{k=2^{\lceil \log_2 t \rceil - 2}+1}^{2^{\lceil \log_2 t \rceil - 1}} x_k + \sum_{k=2^{\lceil \log_2 t \rceil - 1}+1}^{t} x_k \right). \tag{2.5}$$

Note that $1/2 < (t - 2^{\lceil \log_2 t \rceil - 2})/t \leq 3/4$ *for* $t \geq 3$. *Therefore, the online suffix averaging is a form of robust suffix averaging with* $1/2 < \kappa \leq 3/4$, *i.e., the average would always correspond to a constant-portion suffix of all iterates. The online suffix average* $\hat{x}_t$ *in* (2.5) *can be updated recursively; see Algorithm 1.*

Algorithm 1: Online suffix averaging

```
Input: step sizes $\{\eta_t\}_{t\geq 1}$, initialization $x_0, m = 0, S_0 = 0, S_1 = 0$;
for $t = 1, 2, ...,$ do
    $x_t = x_{t-1} - \eta_t \nabla f(x_{t-1}, \xi_t)$;
    if $t > 2^m$ then
        $m = m + 1,\ S_0 = S_1,\ S_1 = x_t$;
    else
        $S_1 = S_1 + x_t$;
    end
    Output (if necessary): $\hat{x}_t = (S_0 + S_1)/(t - \lfloor 2^{m-2} \rfloor)$
end
```

2.4 Non-asymptotic mean squared error

In addition to the asymptotic distribution and statistical convergence rates, it is also important to consider finite sample performance when dealing with finite data problems or when early stopping is desired. In this section, we will examine the optimal weight for a linear model in terms of finite sample mean squared error (MSE), building upon the concept of *best linear unbiased estimation* (BLUE). Furthermore, we will introduce a novel adaptive averaging scheme based on insights from the mean estimation model. This particular scheme is both statistically efficient with optimal variance, and has a fast finite sample convergence rate, outperforming existing averaging schemes in the mean estimation model.

Given an SGD estimate $\tilde{x}_n$, we can evaluate its non-asymptotic performance through its MSE, i.e.,

$$\mathrm{MSE}(\tilde{x}_n) = \mathbb{E}\|\tilde{x}_n - x^*\|_2^2.$$

Consider the general weighted average $\tilde{x}_n = \sum_{i=1}^n w_{n,i} x_i$. To find the optimal weight with respect to the finite sample MSE, we solve the following problem:

$$\min_{c=(c_1,\,\cdots,\,c_n):c^T\mathbf{1}_d=1} \mathbb{E}\|\sum_{i=1}^{n} c_i x_i - x^*\|_2^2. \tag{2.6}$$

Given the "covariance" matrix Σ of the SGD sequence with $\Sigma_{i,j} = \mathbb{E}((x_i - x^*)^T(x_j - x^*))$, the solution to the above constrained optimization problem is

$$c = \frac{\Sigma^{-1}\mathbf{1}_d}{\mathbf{1}_d^T\Sigma^{-1}\mathbf{1}_d}. \tag{2.7}$$

The solution depends on the correlation between SGD iterates and can vary across different models. In this section, we examine the linear regression model, which provides insights into the properties of an optimal weight in a generalized form.

2.4.1 Linear model

Consider the following linear regression model:

$$b_i = a_i x^* + \epsilon_i \tag{2.8}$$

where x^* denotes the unknown parameter of interest, ϵ_i across $i = 1, 2, ...$ are *i.i.d.* from standard normal distribution, and $\{\xi_i = (a_i, b_i)\}$ denote the observed streaming data. To solve the above linear regression problem, we consider the squared loss function

$$F(x) = \mathbb{E}(f(x, \xi_i)) = \mathbb{E}\frac{1}{2}(a_i x - b_i)^2,$$

and SGD sequence with step size η_i at the i-th iteration:

$$x_i = x_{i-1} - \eta_i a_i(a_i x_{i-1} - b_i). \tag{2.9}$$

Proposition 2.4.1. *Consider the linear model in* (2.8) *and SGD sequence* x_i *defined in* (2.9) *with step size* $\eta_1 = a_1^{-2}$ *and general* $\eta_i, i \geq 1$. *The unique solution to the optimization problem* (2.6) *is given by*

$$c = \frac{\Theta^T D^{-1} \Theta \mathbf{1}_d}{\mathbf{1}_d^T \Theta^T D^{-1} \Theta \mathbf{1}_d},$$

where D *is a diagonal matrix with* $D_{i,i} = (\sigma^2 a_i^2 \eta_i^2)$ *and*

$$\Theta = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \eta_2 a_2^2 - 1 & 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \eta_n a_n^2 - 1 & 1 \end{pmatrix}.$$

More explicitly,

$$c_{n,i} = \frac{a_{i+1}^2 + \eta_i^{-1} - \eta_{i+1}^{-1}}{S_n}, 1 \leq i \leq n-1,$$

$$c_{n,n} = \frac{1}{\eta_n S_n},$$

where $S_n = \sum_{i=1}^n a_i^2$. *And*

$$\mathrm{MSE}(\sum_{i=1}^n c_{n,i} x_i) = \frac{1}{n}.$$

The weights in Proposition 2.4.1 are adjusted by learning rate η_i and data a_i. However, one characteristic of these weights is that the last weight is significantly larger than the preceding ones.

2.4.2 A new averaging scheme: adaptive weighted averaging

In the special case of the mean estimation model, where $a_i = 1, \forall i \geq 1$, and if we choose a polynomially decaying step size $\eta_i = i^{-\alpha}$, then the weights are given by:

$$c_{n,i} = \frac{1 + i^\alpha - (i+1)^\alpha}{n}, 1 \leq i \leq n-1, c_{n,n} = n^{\alpha - 1}. \tag{2.10}$$

The weighted averaged SGD $\tilde{x}_n$ with above optimal weights can also be easily computed on-the-fly. When the number of iteration increases from n to $n+1$, we have

$$c_{n+1,i} = \frac{n}{n+1} c_{n,i}, 1 \le i \le n-1.$$

Therefore we can obtain $\tilde{x}_{n+1}$ through

$$\tilde{x}_{n+1} = \frac{n}{n+1}\tilde{x}_n + \frac{1-(n+1)^\alpha}{n+1} x_n + (n+1)^{\alpha-1} x_{n+1}. \tag{2.11}$$

The newly introduced averaging scheme, referred to as *adaptive weighted averaging*, effectively decreases the weight of earlier iterates in comparison to the later ones.

From the definition of the optimal weight, the weight in (2.10) minimizes the finite sample MSE among all possible weights for the mean estimation model. On the other hand, this optimal weight (2.10) also satisfies the conditions in Theorem 2.2.4, except for the last weight $n^{\alpha-1}$ which is much greater than $O(1/n)$. However, we have $x_n - x^* = O(n^{-\alpha/2})$, so the last weighted error term $\sqrt{n} c_{n,n}(x_n - x^*) = O(n^{(\alpha-1)/2})$ will vanish as $n \to \infty$. As a result, the corresponding weighted averaged SGD still exhibits ideal asymptotic normality, with the asymptotic covariance matrix being the same as that of ASGD estimates; see the following Corollary 2.4.2. Thus, the proposed adaptive weighted average achieves both fast finite sample convergence rates and the optimal statistical rate.

Corollary 2.4.2. *Under the settings in Theorem 2.2.4, let* $\tilde{x}_n = \sum_{i=1}^n c_{n,i} x_i$ *with* $c_{n,i}$ *defined in* (2.10)*. Then we have*

$$\sqrt{n}(\tilde{x}_n - x^*) \Rightarrow \mathcal{N}(0, A^{-1} S A^{-1}),$$

where $A = \nabla^2 F(x^*), S = \mathbb{E}([\nabla f(x^*, \xi)][\nabla f(x^*, \xi)]^T)$*.*

Connection with uniform averaging. It is interesting to compare the adaptive weighted average in (2.11) with uniform average (ASGD). The recursion of ASGD $\bar{x}_n$ takes the following form

$$\bar{x}_{n+1} = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}x_{n+1}. \tag{2.12}$$

To build the connection between (2.11) and (2.12), we rewrite (2.11) as

$$\tilde{x}_{n+1} = \frac{n}{n+1}\tilde{x}_n + \frac{1}{n+1}x_{n+1} + \frac{(n+1)^\alpha - 1}{n+1}(x_{n+1} - x_n). \tag{2.13}$$

Thus we can consider the proposed adaptive weighted average as a modified ASGD with a correction term, where the correction term reduces the weight of earlier iterates and increases the weight of the latest iteration. This modification bears a certain similarity with other existing variance-reduced modifications on SGD where a correction term is applied on stochastic gradients, such as SGD with momentum.

2.5 Numerical experiment

In this section, we check the asymptotic normality property of the general weighted averaged SGD and investigate the non-asymptotic performance of various averaging schemes in different settings.

2.5.1 Asymptotic normality for different averaging schemes

To verify the asymptotic normality and the limiting covariance matrix derived in Theorem 2.2.4, we consider three averaging schemes: polynomial-decay, suffix averaging (as described in Section 2.3), and the adaptive averaging scheme proposed in (2.11).

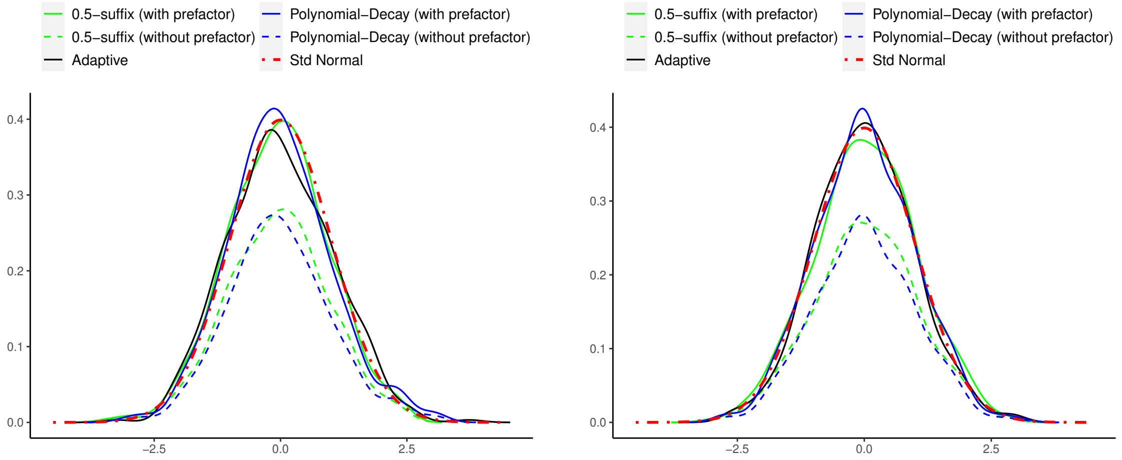
We focus on two classes of loss functions: squared loss $f(x,(a,b)) = (a^T x - b)^2/2$ for the linear regression model, and logit loss: $f(x,(a,b)) = \log(1+\exp(-ba^T x))$ for the logistic regression model. In both models, we assume that the data $\xi_i = (a_i, b_i)$, $i = 1, 2, ..., n$, are independent, where a_i represents the explanatory variable generated from $\mathcal{N}(0, \mathbf{I}_d)$, and b_i

represents the response variable generated from two different distributions correspondingly. For linear regression, we assume $b_i \sim \mathcal{N}(a_i^T x^*, 1)$, while for logistic regression, $b_i \in \{1, -1\}$ is generated from a Bernoulli distribution, where $\mathbb{P}(b_i|a_i) = 1/(1 + \exp(-b_i a_i^T x^*))$. Recall the asymptotic normality we are going to verify in Theorem 2.2.4:

$$\sqrt{n}(\tilde{x}_n - x^*) \Rightarrow \mathcal{N}(0, wV)$$

where $V = A^{-1}SA^{-T}$ is the sandwich form matrix. For squared loss, it is easy to derive that $A = S = \mathbf{I}_d$ and therefore $V = \mathbf{I}_d$. For logit loss, since the explicit forms for A and S are difficult to obtain, we use Monte-Carlo simulation to numerically compute the sandwich form matrix V.

In simulations, we set $d = 5$ and the true parameter $x^* = (1, -2, 0, 0, 4)^T$ for both models. We generate SGD sequences with $\eta_i = i^{-\alpha}, \alpha = 0.505$, and apply different averaging schemes. The number of iterations $n = 100000$, and all the measurements are averaged over 450 independent runs. For the polynomial-decay averaging, we choose $\gamma = 3$ [Shamir and Zhang, 2013], and for the suffix averaging we choose $\kappa = 0.5$ [Rakhlin et al., 2012]. Then the prefactors w for polynomial-decay and suffix averaging schemes are 16/7 and 2. For polynomial decay and suffix averaged SGD, we plot the density of the standardized error with and without prefactor w, i.e., $w^{-1}V^{-1}\sqrt{n}(\tilde{x}_n - x^*)$ and $V^{-1}\sqrt{n}(\tilde{x}_n - x^*)$. For adaptive weighted SGD, the prefactor is 1 according to Theorem 2.4.2, so we only plot the density of the standardized error $V^{-1}\sqrt{n}(\tilde{x}_n - x^*)$. As shown in Figure 2.2, the standardized error (scaled with the prefactor w) exhibits an approximate standard normal distribution for all three averaging schemes. However, for the polynomial decay and suffix averaging schemes, the standardized error without the prefactor w has a significantly different density compared to the standard normal distribution. These findings support the conclusion stated in Theorem 2.2.4, affirming the validity of the asymptotic normality of weighted SGD solutions and the correctness of the limiting covariance matrix.



(a) Squared Loss

(b) Logit Loss

Figure 2.2: Density plot for the standardized error with and without prefactor w. The red line denotes a standard normal distribution.

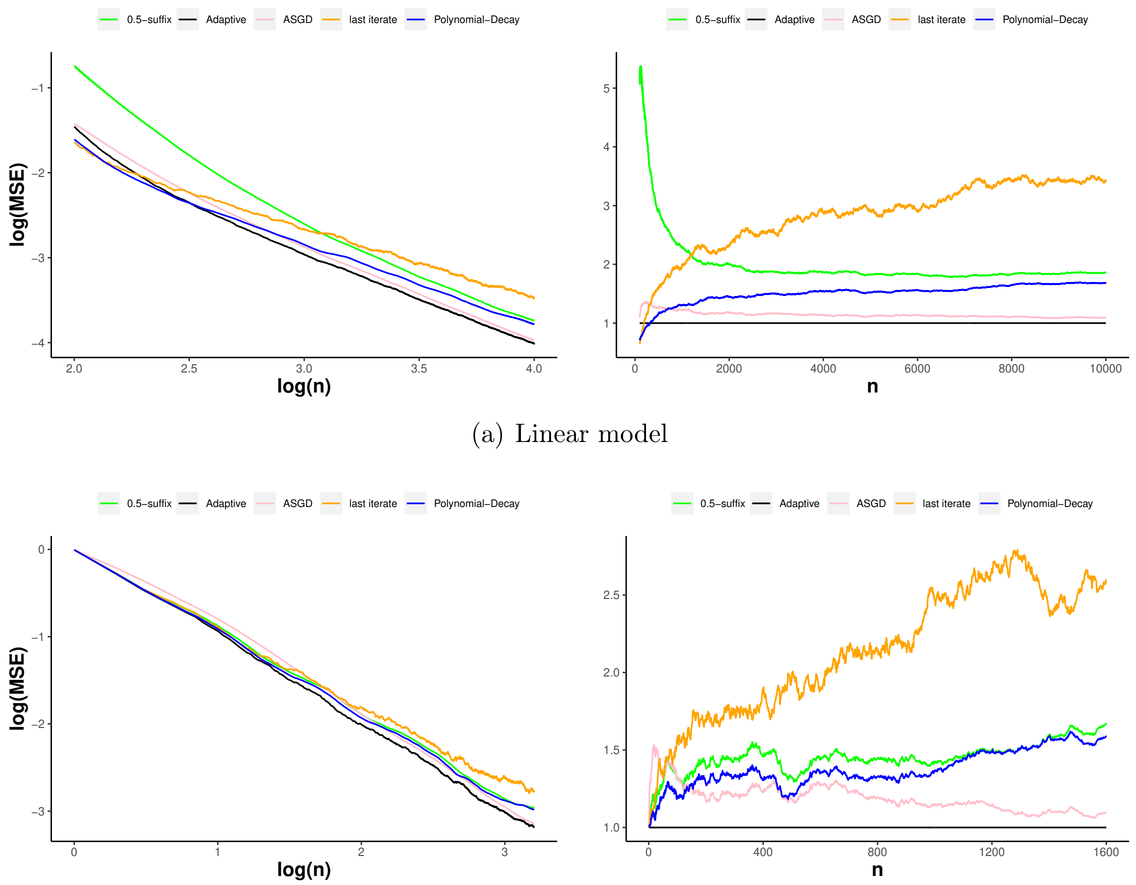
2.5.2 Non-asymptotic performance of different averaging schemes

In this section, we will show that the adaptive weighted averaging scheme in (2.11) has good non-asymptotic performances in different cases.

Linear model: optimal MSE

We first validate the optimality in terms of finite sample MSE in the linear regression model, which is a generalization of the mean estimation model. The linear regression model we examine employs identical simulation settings as described in Section 2.5.1. Additionally, we incorporate a specific scenario of the mean estimation model with $a_i = 1$ and $x^* = 0$.

We present coordinate-averaged MSE at certain steps in Figure 2.3. Here the step size $\eta_i = i^{-0.8}$, and all the measurements are averaged over 400 independent runs. We can see that the adaptive weighted averaging outperforms other averaging schemes. When n is large enough ASGD has a similar performance with adaptive weighted SGD. It is also consistent with our conclusion that adaptive weighted SGD and ASGD have the same limiting



(a) Linear model

(b) Mean estimation model

Figure 2.3: Left: log-log plots for MSE. Right: the curves stand for the ratio of MSE between different averaging schemes and adaptive weighted averaging at each step. The baseline (black line) is for the adaptive weighted averaging.

covariance. For the linear regression model, MSE of the optimal weighted SGD is always smaller than that of ASGD. It indicates that adaptive weighted SGD has the potential to beat ASGD for not only the mean estimation model but also a more general optimization problem.

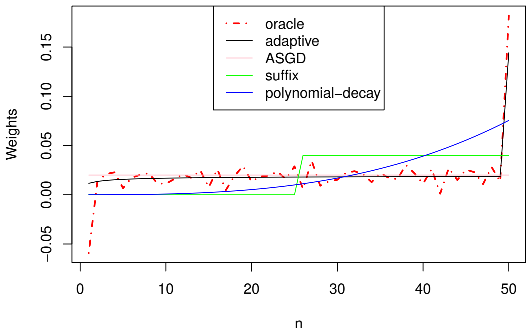


Figure 2.4: Comparison of different weights under expectile regression model with $\rho = 0.8$. The oracle weights are numerically computed via Monte-Carlo simulation with 50000 repetitions.

Expectile regression: trend of optimal weight

We next explore the example of Expectile regression, which has a non-smooth objective function,

$$F(x) = \mathbb{E}_{y\sim\Pi}(|\rho - 1_{\{y<x\}}|(y-x)^2), 0 < \rho < 1.$$

Expectiles have important applications in finance and risk management. They are closely associated with two commonly adopted measures: Value at Risk (VaR) and Conditional Expected Shortfall (CES). Expectile regression was proposed by Newey and Powell [1987], and has been widely used and researched in statistical and economic literature [Efron, 1991, Taylor, 2008]. For the expectile estimation problem, the optimal weights in Section 2.4 do not have a closed-form solution. Therefore we use Monte-Carlo simulation to numerically compute the inverse of the covariance matrix and obtain the oracle weights based on equation (2.7). We then compare these oracle weights with all the weighting schemes that we studied previously.

The weights for different averaging schemes are plotted in Figure 2.4 with the total iteration $n = 50$ and step size $\eta = i^{-0.505}$. The most remarkable feature of the oracle weight is its *highest* weight assigned to the last iterate. The adaptive weight we proposed in Section

2.4.2 is able to capture this characteristic, while the other averaging schemes fail to recover it. This observation shows that our adaptive weighted averaging scheme is also promising for the non-smooth optimization problem as it aligns the closest with the trend of the oracle weights.

2.6 Summary

In this paper, we present the asymptotic normality results for a broad range of weighted averaged SGD solutions, demonstrating that the limiting covariance matrix adopts a sandwich form—that of ASGD's limiting covariance matrix with an additional prefactor. This marks the first asymptotic distribution result for general weighted averaged SGD and holds significant importance for statistical inference. We note that although certain existing weighted averaged SGD methods exhibit faster convergence than ASGD in non-asymptotic views or specific settings without strong assumptions, they may also incur larger variance, indicating a trade-off. Additionally, we explore the non-asymptotic MSE of weighted averaged SGD in the linear regression model and propose a novel averaging scheme—adaptive averaged SGD. This scheme exhibits asymptotic normality, achieves optimal limiting covariance, and offers favorable finite sample MSE.

CHAPTER 3

SHARP CONCENTRATION ANALYSIS FOR STOCHASTIC GRADIENT DESCENT

As mentioned in the previous chapter, there have been extensive studies on the theoretical properties of SGD since 1951, from consistency to distributions/inference and from asymptotic to non-asymptotic investigations [Blum, 1954, Dvoretzky, 1956, Moulines and Bach, 2011, Rakhlin et al., 2012, Bach and Moulines, 2013, Toulis and Airoldi, 2017, Anastasiou et al., 2019a]. However, there are still gaps between the theory of SGD and applications, especially with heavy-tailed stochastic gradient noise which commonly arises in practice. In this chapter, we will focus on the concentration property of the SGD estimates. We obtain a nearly sharp high-probability error bound for SGD estimates with heavy-tailed noise in the linear model. We show that the tail behaviors of SGD estimates are quite different in heavy-tailed noise cases compared to those in sub-Gaussian noise cases.

For consistency with the notation used in the paper Lou et al. [2022], which is relevant to this chapter, we will use θ to denote the parameter of interest instead of x, and X, Y to represent data instead of ξ. To be more specific, we consider the convex optimization problem $\min_{\theta \in \mathbb{R}^p} F(\theta)$, where $F : \mathbb{R}^p \to \mathbb{R}$, and SGD updates the estimate of the minimum $\theta^\star$ based on the stochastic gradient $\hat{g}(\theta)$ at some θ, which is a noisy measurement of the gradient/subgradient $g(\theta) = \nabla F(\theta)$. It is important to note that this change in notation is specific to this chapter.

3.1 Introduction

Most of the literature on the quality of SGD estimates focuses on the *expected* error rate. Polyak and Juditsky [1992] and Ruppert [1988] introduced the averaged SGD (ASGD), a simple modification where iterates are averaged, and established the asymptotic normality of the obtained estimate. It is known that ASGD estimates achieve the optimal rate $O(1/\sqrt{T})$ due to the central limit theorem (CLT), after T steps of SGD, under certain regularity conditions. Further analyses on the error rate show that the *expected* squared error of the SGD estimate (with average if necessary) is $O(1/T)$ for strongly convex objective functions, and $O(1/\sqrt{T})$ for smooth convex and non-smooth Lipschitz objective functions [Nemirovski et al., 2009, Rakhlin et al., 2012, Shamir and Zhang, 2013, Lacoste-Julien et al., 2012].

Besides the guarantees in expectation, practitioners usually want to ensure that the output of a single trial of the algorithm is well behaved and may ask: how many iterations are needed in a single trial of the algorithm to achieve the desired accuracy? In other words, they would prefer high confidence guarantees, i.e., high-probability error bounds in the form of

$$\mathbb{P}(\|\hat{\theta} - \theta^\star\|_2^2 \geq \epsilon) \leq \delta,$$

where $\epsilon > 0$, $\delta \in (0,1)$ can be arbitrarily small, and $\hat{\theta}$ is the estimate of $\theta^\star$. These high-probability guarantees are usually adopted in statistical learning theory [Valiant, 1984], where a tight sample complexity bound is of great interest. Note that bounds in expectation are generally too conservative to derive high-probability guarantees. Specifically, if one has $\mathbb{E}\|\hat{\theta}_T - \theta^\star\|_2^q = O(T^{-2/q})$ [Chung, 1954], by Markov's inequality, one can only guarantee with probability at least $1 - \delta$,

$$\|\hat{\theta}_T - \theta^\star\|_2^2 \leq O(\delta^{-2/q} T^{-1}).$$

Then, the resulting sample complexity

$$T(\epsilon, \delta) = O\left(\frac{\delta^{-2/q}}{\epsilon}\right) \tag{3.1}$$

can be very high for a small δ. Also, the confidence intervals obtained from the CLT only hold asymptotically when the number of samples goes to infinity and cannot be used to rigorously compute sample complexity when $\delta \to 0$. Thus, additional non-asymptotic tail probability results are needed.

High-probability bounds on SGD are much less explored than the bounds in expectation. Some known high-probability results under light-tailed noise assumptions include Rakhlin et al. [2012], who showed that for the strongly convex setting and suffix averaging $\hat{\theta}_T$, with probability at least $1 - \delta$,

$$\|\hat{\theta}_T - \theta^\star\|_2^2 \le O\left(\log(\log(T)/\delta)/T\right).$$

Recently, Harvey et al. [2019a] improved the above bound to $O\left(\log(1/\delta)/T\right)$. Other similar results can be found in Hazan and Kale [2014], Cardot et al. [2017], Jain et al. [2019], Harvey et al. [2019b], Feldman and Vondrak [2019], Mou et al. [2020]. These high-probability bounds depend logarithmically on $1/\delta$, and the resulting sample complexity is

$$T(\epsilon, \delta) = O\left(\frac{\log(1/\delta)}{\epsilon}\right),$$

substantially improving $T(\epsilon, \delta) = O(\delta^{-2/q}\epsilon^{-1})$ in (3.1) when δ is small. Such bounds with a dependence on $\log(1/\delta)$ are often called *sub-Gaussian bounds* or with *sub-Gaussian performance*. Harvey et al. [2019a] also remark that a dependence on $\log(1/\delta)$ is necessary, which indicates that SGD can not achieve a better performance than this one under sub-Gaussianity.

The aforementioned high-probability results all rely on the *light-tailed* assumption on the gradient noise $z = \hat{g} - g$, such as boundedness or sub-Gaussianity. However, such assumptions can be violated in practice. The heavy-tailed phenomenon is not uncommon in applications [Simsekli et al., 2019]. It is also more likely to get a bad output in a single trail of SGD due to the more frequent outliers with heavy-tailed stochastic gradients. Thus, a high-probability guarantee is especially needed. Then a natural question is: *Can SGD achieve the* *sub-Gaussian performance with* $\log(1/\delta)$ *tail behavior in the case of heavy-tailed stochastic noise?* This paper answers this question by delivering a nearly tight high-probability bound in a linear model with heavy-tailed stochastic noise. In particular, with probability at least $1-\delta$, for any $\delta \in (0,1)$,

$$\|\bar{\theta}_T - \theta^\star\|_2^2 \le O\left(\frac{\log(1/\delta)}{T} + \frac{(1/\delta)^{2/q}}{T^{2-2/q}}\right),$$

where $\theta^\star$ is the true parameter and $\bar{\theta}_T$ is the ASGD estimate ($q > 2$ controls the tail of the stochastic noise). As a result, the sample complexity bound, with tolerance error $\epsilon > 0$ and failure probability $\delta \in (0,1)$, is

$$T(\epsilon, \delta) = O\left(\frac{\log(1/\delta)}{\epsilon} + \left(\frac{\delta^{-2/q}}{\epsilon}\right)^{q/(2(q-1))}\right). \tag{3.2}$$

It is better than the $T(\epsilon, \delta) = O(\delta^{-2/q}\epsilon^{-1})$ in (3.1). Besides the advantage of the logarithmical term, the polynomial term $O((\delta^{-2/q}\epsilon^{-1})^{q/(2(q-1))})$ is sharper than $O(\delta^{-2/q}\epsilon^{-1})$ since $q > 2$. We also compare the logarithmical term and the polynomial term in (3.2) numerically. Figure 3.1 shows that, when δ is big, the logarithmical dependence dominates, and therefore the sample complexity is the same as that in the sub-Gaussian case. While when δ is small, which is more of interest in most cases, the polynomial dependence term dominates, showing that the polynomial dependence on δ is unavoidable. Thus, one cannot achieve the sub-Gaussian performance when the gradient noise exhibits heavy-tailed distribution.

There has recently been renewed interest in obtaining robust guarantees for SGD without the light-tailed assumption. Robust modifications of SGD (or GD), such as gradient clipping and using the geometric median of stochastic gradients, are studied to accommodate heavy-tailed noise [Nazin et al., 2019, Holland and Ikeda, 2019, Davis and Drusvyatskiy, 2020, Gorbunov et al., 2020]. The question of whether these robust modifications are necessary

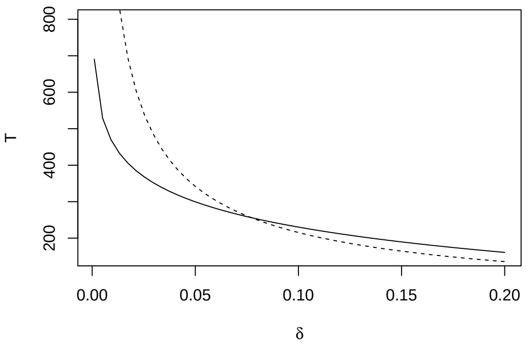


Figure 3.1: Compare the two terms in sample complexity (3.2). Here X axis represents failure probability δ; the solid line denotes $\epsilon^{-1}\log(1/\delta)$, the dashed line denotes $(\delta^{-2/q}\epsilon^{-1})^{q/(2(q-1))}$. We choose $\epsilon = 0.01$ and $q = 2.5$.

is vital since using SGD is a common heuristic in modern learning tasks and is easier to implement and more widely used than its modified versions. Our lower bound answers this question and indicates that such modifications are necessary when heavy-tailed noise exists.

3.2 Upper bound

3.2.1 Linear model setting

Assume that we observe data $(X_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i \geq 1$, from the following linear regression model:

$$y_i = X_i^\top \theta^\star + \epsilon_i, \quad i \geq 1,$$

where $\theta^\star \in \mathbb{R}^p$ is the unknown true parameter. The random draws (X_i, ϵ_i) across $i = 1, 2, ...$ are *i.i.d.* from $P_X \times P_\epsilon$. Here we assume that P_X is a distribution on $\mathbb{R}^p$ such that $\mathbb{E}(X_i X_i^T) = \Sigma$, while P_ϵ is a distribution on $\mathbb{R}$ such that $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. Note that, we consider a much richer class of gradient noise beyond sub-Gaussian, where only finitely-many moments are required allowing heavy-tailed noise. More detailed assumptions are included in Section 3.2.2.

To solve the above linear regression problem, we consider the optimization problem

$$\min_{\theta \in \mathbb{R}^p} F(\theta) = \mathbb{E}_{X,y} \frac{1}{2}(y - X^\top \theta)^2.$$

We apply the mini-batch SGD, which is a popular parallelization technique reducing the communication costs [Li et al., 2014, Reddi et al., 2016, Jain et al., 2017]. Mini-batching is efficient in practice and brings convenience in later proofs. Initialized at θ_0, the t-th iteration with step size η_t is given by:

$$\begin{aligned} &\theta_t = \theta_{t-1} - \eta_t \hat{g}_t(\theta_{t-1}), \;\; t \geq 1, \\ &\hat{g}_t(\theta_{t-1}) = \frac{1}{B} \sum_{i=(t-1)B+1}^{tB} X_i \left(X_i^\top \theta_{t-1} - y_i \right), \end{aligned} \tag{3.3}$$

where B is the mini-batch size and step sizes η_t will be discussed in later analysis. In this chapter, we are interested in the high-probability bound of the averaged iterate $\bar{\theta}_T = T^{-1} \sum_{t=1}^T \theta_t$ with T iterations ($n = TB$ samples) in total.

For $(X_i, y_i)_{i \geq 1}$ in above linear regression model, let

$$A_t = \frac{1}{B} \sum_{i=(t-1)B+1}^{tB} X_i X_i^\top \;\text{ and }\; b_t = \frac{1}{B} \sum_{i=(t-1)B+1}^{tB} X_i y_i.$$

We can rewrite the t-th iteration from the mini-batch SGD as:

$$\theta_t = \theta_{t-1} - \eta_t (A_t \theta_{t-1} - b_t), \;\; t \geq 1.$$

Note that $\mathbb{E}(A_t) = \mathbb{E}(XX^\top) = \Sigma$, and $b = \mathbb{E}(b_t) = \Sigma \theta^\star$. We can see that solving the linear regression problem through mini-batch SGD (3.3) is equivalent to solving the linear system of the form:

$$\Sigma \theta^\star = b,$$

through stochastic approximation [Mou et al., 2020].

3.2.2 Assumptions

Assumption 3.2.1. *For distribution P_ϵ, assume that for some constant $q > 2$,*

$$\mu_q = \mathbb{E}|\epsilon|^q < \infty.$$

Assumption 3.2.2. *Assume that $M_\psi := \max_{1\leq \ell \leq p} \left(\mathbb{E}|X_{i\ell}|^{2\psi}\right)^{1/2\psi} < \infty$, for some constant $\psi > \max\{4, q\}$. Let $\lambda_{\min}(\Sigma) > 0$, assume that the mini-batch size satisfy*

$$B \geq 16(\psi - 1)M_\psi^4 p^2/\lambda_{\min}(\Sigma)^2.$$

Remark 3.2.3. *In existing works, light-tailed assumptions of the gradient noise are required, i.e., finite exponential moments (e.g. bounded, sub-Gaussian, sub-exponential). While in our assumptions, the noise conditions are more general. We only require finite polynomial moments in Assumption 3.2.1, in which case heavy-tailed noise is allowed. Assumption 3.2.2 is a fairly mild condition on P_X and the mini-batch size B. It ensures that*

$$\left(\mathbb{E}\|A_t - \Sigma\|_2^\psi\right)^{1/\psi} \leq \lambda_{\min}(\Sigma)/2, \tag{3.4}$$

which is shown in Lemma B.1.3 and is a useful condition for controlling the correlation between SGD iterates in later proofs. On the other hand, if X_i is Gaussian, Corollary 2 in Koltchinskii and Lounici [2017] implies that a weaker assumption for (3.4) *is, for come constant C_ψ,*

$$B \geq C_\psi r(\Sigma)\mathcal{K}(\Sigma)^2,$$

where $r(\Sigma) = \mathrm{tr}(\Sigma)/\lambda_{\max}(\Sigma)$ is the effective rank of Σ, and $\mathcal{K}(\Sigma) = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ is the condition number. It is worth mentioning that the linear system $\Sigma\theta^\star = b$ becomes more *unstable when the condition number of Σ grows. Therefore, it is reasonable to require a larger mini-batch size B when the condition number is larger. For more discussion about concentration inequality and expectation inequality of the operator norm $\|A_t - \Sigma\|_2$, we refer to Vershynin [2010], Koltchinskii and Lounici [2017], Tropp [2012] and the references therein.*

3.2.3 Nagaev type upper bound

The step size sequence $(\eta_t)_{t\geq 1}$ controls the convergence of the SGD algorithm. In this section, we focus on two commonly used step size regimes: polynomial decay step size $\eta_t = \eta_0 t^{-\alpha}$ with $\alpha \in (0,1)$ and constant step size with $\eta_t = \eta_0$ for any $t \geq 1$.

We analyze the tail probability of the error $\bar{\theta}_T - \theta^\star$, after T steps of SGD, in the linear model setting. In what follows, we denote $\lambda_0 = \lambda_{\min}(\Sigma)/2$, $\lambda^* = \lambda_{\max}(\Sigma)$,

$$\mathcal{K}_q = \sup_{\nu \in \mathbb{S}^{p-1}} \mathbb{E}|\nu^\top X_t|^q,$$

and

$$\Upsilon_{\varpi,\alpha} = \int_1^\infty \exp\left(-\varpi \int_1^z x^{-\alpha} dx\right) dz.$$

Theorem 3.2.4 (polynomial decay step size)**.** *Let Assumptions 3.2.1 and 3.2.2 hold. Assume that $\eta_0 \leq 1/\lambda^*$ and*

$$\psi > \frac{2q - 4\alpha}{2 - \alpha}.$$

Then, for any $\omega \in \mathbb{S}^{p-1}$ and $x > 0$, we have

$$\mathbb{P}\left(|\omega^\top(\bar{\theta}_T - \theta^\star)| > x\right) \leq \frac{C_0\|\theta_0 - \theta^\star\|_2^\psi}{(Tx)^\psi} + \frac{C_1 W_q}{T^{q-1}x^q} + C\exp\left(-\frac{C_2 T x^2}{W_2}\right), \tag{3.5}$$

where $W_q = \mu_q \mathcal{K}_q \lambda_0^{-q} B^{1-q}$, $W_2 = \sigma^2 \lambda^* \lambda_0^{-2} B^{-1}$, $C_0 = (2\Upsilon_{\lambda_0 \eta_0, \alpha})^{\psi}$, C, C_1 *and* C_2 *are constants depending only on* q, ψ *and* α.

Remark 3.2.5. *The first term on the RHS of* (3.5) *characterizes the effect of the initial point* θ_0 *on the tail probability of* $\bar{\theta}_T - \theta^\star$. *The influence of* θ_0 *decays quickly, note that for any* $x \gtrsim T^{-1/2}$, *we have*

$$\frac{C_0 \|\theta_0 - \theta^\star\|_2^{\psi}}{(Tx)^{\psi}} \le \frac{C_1 W_q}{T^{q-1} x^q},$$

as long as $\|\theta_0 - \theta^\star\|_2^{\psi} \lesssim C_1 C_0^{-1} W_q T^{1+(\psi-q)/2}$, *which is a fairly mild condition on* θ_0 *as* $\psi > q$. *Consequently, in this case, Theorem 3.2.4 implies that*

$$\mathbb{P}\left(|\omega^\top(\bar{\theta}_T - \theta^\star)| > x\right) \le \frac{2C_1 W_q}{T^{q-1} x^q} + C \exp\left(-\frac{C_2 T x^2}{W_2}\right),$$

which, together with $\mathbb{P}(\|\bar{\theta}_T - \theta^\star\|_2 > x) \le \sum_{j=1}^{p} \mathbb{P}(|\bar{\theta}_{T,j} - \theta_j^\star| > x/\sqrt{p})$, *imply that*

$$\mathbb{P}\left(\|\bar{\theta}_T - \theta^\star\|_2 > x\right) \le \frac{2p^{1+q/2} C_1 W_q}{T^{q-1} x^q} + pC \exp\left(-\frac{C_2 T x^2}{p W_2}\right). \tag{3.6}$$

Theorem 3.2.6 (Constant step size)**.** *Let Assumptions 3.2.1 and 3.2.2 hold. Assume that* $\eta_0 \le 1/\lambda^*$ *and*

$$\eta_0 \gtrsim \frac{(\log T)^{(3\psi-4)/(\psi-4)}}{T^{2(\psi-q)/(\psi-4)}}.$$

Then, for any vector $\omega \in \mathbb{S}^{p-1}$ *and* $x > 0$, *we have*

$$\mathbb{P}\left(|\omega^\top(\bar{\theta}_T - \theta^\star)| > x\right) \le \frac{C_0 \|\theta_0 - \theta^\star\|_2^{\psi}}{(Tx)^{\psi}} + \frac{C_1 W_q}{T^{q-1} x^q} + C \exp\left(-\frac{C_2 T x^2}{W_2}\right),$$

where $W_q = \mu_q \mathcal{K}_q \lambda_0^{-q} B^{1-q}$, $W_2 = \sigma^2 \lambda^* \lambda_0^{-2} B^{-1}$,$C_0 = (2/\lambda_0 \eta_0)^{\psi}$, C, C_1 *and* C_2 *are constants depending only on* q *and* ψ.

Remark 3.2.7. *Both inequalities with different step size regimes imply two types of bounds:*

Gaussian type tail and polynomial type tail. When x is small, i.e., for small deviations, the Gaussian type tail is the dominating term for the tail of the estimation error. While for large x, the polynomial type tail dominates. A combination of these two types of tail approximation calibrates the tail behavior of SGD solutions more accurately in the case of heavy-tailed noise.

After elementary calculations, we can translate the tail probability results as following. For any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\bar{\theta}_T - \theta^\star\|_2 \leq O\left(\sqrt{\frac{\log(1/\delta)}{T}} + \frac{(1/\delta)^{1/q}}{T^{1-1/q}}\right).$$

For large or moderate failure probability $\delta > \delta^*$,

$$\|\bar{\theta}_T - \theta^\star\|_2 \leq O\left(\sqrt{\frac{\log(1/\delta)}{T}}\right),$$

where δ^* is the solution of the equation $\sqrt{T^{-1}\log(1/\delta)} = T^{-1+1/q}(1/\delta)^{1/q}$ and has the asymptotic form $\delta^* \asymp T^{1-q/2}(\log T)^{-q/2}$. This high probability error rate matches existing results considering sub-Gaussian/bounded gradient noise [Mou et al., 2020]. While for small failure probability $\delta < \delta^*$, which is more of interest in most applications, we have

$$\|\bar{\theta}_T - \theta^\star\|_2 \leq O\left(\frac{(1/\delta)^{1/q}}{T^{1-1/q}}\right).$$

3.2.4 Technical overview and proof sketch for main results

Let $\Delta_t = \theta_t - \theta^\star$ and $\mathcal{E}_t = B^{-1}\sum_{i=(t-1)B+1}^{tB} X_i\epsilon_i$. At the t-th step, the gradient g and the stochastic gradient $\hat{g}_t$ can be written with $\Sigma, A_t, \mathcal{E}_t, \Delta_t$ notations as follows:

$$g(\theta_{t-1}) = \Sigma\Delta_{t-1}, \quad \hat{g}_t(\theta_{t-1}) = A_t\Delta_{t-1} - \mathcal{E}_t.$$

Let $z_t(\theta_{t-1}) = \hat{g}_t(\theta_{t-1}) - g(\theta_{t-1})$ denote the gradient noise. Note that it is a martingale difference sequence since $\mathbb{E}_{t-1}(z_t(\theta_{t-1})) = 0$. The recursion of Δ_t is usually represented using martingales as follows

$$\Delta_t = (\mathbf{I}_p - \eta_t \Sigma)\Delta_{t-1} - \eta_t z_t(\theta_{t-1}), \;\; t \geq 1. \tag{3.7}$$

Then, the classic analysis uses properties of martingales, such as Freedman and Azuma inequalities. The high-probability bounds obtained from those general martingale inequalities are sharp only when finite exponential moments of the noise $z_t|\mathcal{F}_{t-1}$ exists. Therefore, existing studies require the gradient noise z_t (or equivalently A_t and b_t in linear stochastic approximation) to be sub-Gaussian or to be bounded. In our work, we extend the noise condition to a more general case, where heavy-tailed noise is allowed. To obtain sharp high-probability bounds for heavy-tailed noise, we study the detailed structure of the martingale differences and use inequalities which are nearly sharp under polynomial moment conditions.

We can see that the martingale difference z_t at θ_{t-1} can be decomposed as

$$z_t(\theta_{t-1}) = (A_t - \Sigma)\Delta_{t-1} - \mathcal{E}_t, \;\; t \geq 1,$$

which is the sum of two parts, one related to the noise from A_t and the other part $\mathcal{E}_t$. Note that the dependence between $\{z_t(\theta_{t-1})\}_{t\geq 1}$ comes from the dependence between $(\theta_t)_{t\geq 1}$, and $(\mathcal{E}_t)_{t\geq 1}$ are independent. Then leveraging the structure of z_t, we study the recursion with a different representation:

$$\Delta_t = (\mathbf{I}_p - \eta_t A_t)\Delta_{t-1} + \eta_t \mathcal{E}_t, \;\; t \geq 1. \tag{3.8}$$

Compared with the form in (3.7), although more considerations are needed for the correlation term $(\mathbf{I}_p - \eta_t A_t)$ as variability is introduced (we now have $(\mathbf{I}_p - \eta_t A_t)$ instead of $(\mathbf{I}_p - \eta_t \Sigma)$), the remaining independent structure makes it possible to obtain a tight tail bound under heavy-tailed noise assumptions.

In the following, we sketch the proof of our main results under the step size regime $\eta_t = \eta_0 t^{-\alpha}$ with $\alpha \in (0, 1)$. Proof for the constant step size regime shares the same spirit with minor modifications. We defer the complete proof to Section B.1. From (3.8) we can see that $(\Delta_t)_{t \geq 1}$ has a closed form expression

$$\Delta_t = \prod_{\ell=1}^{t} (\mathbf{I}_p - \eta_\ell A_\ell)\Delta_0 + \sum_{m=1}^{t} \prod_{\ell=m+1}^{t} (\mathbf{I}_p - \eta_\ell A_\ell)\eta_m \mathcal{E}_m.$$

Let $S_T = T(\bar{\theta}_T - \theta) = \sum_{t=1}^{T} \Delta_t$ which is further decomposed as $S_T = S_T^\diamond + S_T^\star$, where

$$S_T^\diamond = \sum_{t=1}^{T} \prod_{\ell=1}^{t} (\mathbf{I}_p - \eta_\ell A_\ell)\Delta_0 \text{ and } S_T^\star = \sum_{t=1}^{T} \sum_{m=1}^{t} \prod_{\ell=m+1}^{t} (\mathbf{I}_p - \eta_\ell A_\ell)\eta_m \mathcal{E}_m.$$

To bound the target $\omega^\top S_T / T$ for any $\omega \in \mathbb{S}^{p-1}$ in Section 3.2.3, we deal with $S_T^\diamond$ and $S_T^\star$ separately.

Lemma 3.2.8. *Under Assumption 3.2.2, for any vector $\omega \in \mathbb{S}^{p-1}$ and $x > 0$, we have*

$$\mathbb{P}\left(|\omega^\top S_T^\diamond| > x \right) \leq \frac{\|\theta_0 - \theta^\star\|_2^\psi \Upsilon_{\lambda_0 \eta_0, \alpha}^\psi}{x^\psi}.$$

Next, we observe that for any $\omega \in \mathbb{S}^{p-1}$,

$$\omega^\top S_T^\star = \frac{1}{B} \sum_{m=1}^{T} \sum_{i=(m-1)B+1}^{mB} \eta_m \omega^\top H_m X_i \epsilon_i, \text{ where } H_m = \sum_{t=m}^{T} \prod_{\ell=m+1}^{t} (\mathbf{I}_p - \eta_\ell A_\ell),$$

which means $\omega^\top S_T^\star$ is a sum of independent zero-mean random variables conditional on

$\mathcal{F}_{X,n} = \sigma\{X_1, X_2, \ldots, X_n\}$. Hence, for $x > 0$, by Lemma B.1.2 (Nagaev inequality),

$$\mathbb{P}\left(|\omega^\top S_T^\star| > x|\mathcal{F}_{X,n}\right) \leq \frac{C_{q,1} D_{T,q}}{(Bx)^q} + 2\exp\left(-\frac{C_{q,2}B^2x^2}{D_{n,2}}\right),$$

where $C_{q,1}$ and $C_{q,2}$ are constants depending only on q and

$$D_{T,q} = \mu_q \sum_{m=1}^{T} \eta_m^q \sum_{i=(m-1)B+1}^{mB} |\omega^\top H_m X_i|^q.$$

We bound the conditional variance $D_{T,2}$ in Lemma B.1.4, which is a main technical step, and obtain the following results for $S_T^\star$.

Lemma 3.2.9. *Under the conditions of Theorem 3.2.4, we have*

$$\mathbb{P}\left(|\omega^\top S_T^\star| > x\right) \leq \frac{C_1 W_q T}{x^q} + C\exp\left(-\frac{C_2 x^2}{TW_2}\right),$$

Consequently, Theorem 3.2.4 directly follows from Lemma 3.2.8 and Lemma 3.2.9.

3.3 Tightness of the upper bound

This section shows that the Nagaev type upper bound obtained in the above section is tight through the example of the mean estimation model. Therefore, the polynomial term in the upper bounds in Section 3.2.3 is unavoidable, and the sub-Gaussian performance with $\log(1/\delta)$ tail behavior cannot be achieved through SGD with heavy-tailed gradient noise. In particular, we consider the model

$$y_i = \theta^\star + \epsilon_i, \ \ i \geq 1, \tag{3.9}$$

where $\theta^\star \in R$ is the mean we want to estimate and $\{\epsilon_i\}_{i\geq 1}$ are i.i.d. generated from a t-distribution with degree of freedom $\nu > 2$. For initial value θ_0, the t-th iterate θ_t from SGD algorithm, with mini-batch size $B = 1$, takes the following form:

$$\theta_t = \theta_{t-1} + \eta_t(y_t - \theta_{t-1}), \;\; t \geq 1, \tag{3.10}$$

where η_t is the step size at the t-th iteration.

The gradient noise $z_t = \epsilon_t$ is *heavy-tailed.* The mean estimation model (3.9) is a special case of the linear regression model. Assumptions 3.2.1 and 3.2.2 can be easily verified since $A_t = 1$ with no randomness here. Then we can apply theorems in Section 3.2.3 and get the upper bounds for the estimation error $\bar{\theta}_T - \theta^\star$ when there are T iterations in total. We focus on the polynomial decay step size regime, i.e., $\eta_t = \eta_0 t^{-\alpha}$, with $\eta_0 = 0.1, \alpha = 0.55$ in the rest of this section. We modify the upper bound (3.6) in Section 3.2.3 as follows.

Nagaev type upper bound: We have for all $x > 0$,

$$\mathbb{P}\left(|\bar{\theta}_T - \theta^\star| > x\right) \leq \frac{C_1}{x^q T^{q-1}} + \exp\left(-C_2 T x^2\right), \tag{3.11}$$

for some constant C_1, C_2. Then, with probability at least $1 - \delta$,

$$|\bar{\theta}_T - \theta^\star| \leq O\left(\frac{1}{(\delta T^{q-1})^{1/q}} + \sqrt{\frac{\log(1/\delta)}{T}}\right).$$

Next, we will show that the Nagaev type upper bound for the estimation error $\bar{\theta}_T - \theta^\star$ is tight by taking advantage of the simple structure of the mean estimation model. First, we introduce the following notation

$$\begin{aligned} V_i &= \prod_{k=1}^{i}(1 - \eta_k), \;\; i \geq 1, \;\; V_0 = 1; \\ V_i^j &= \frac{V_j}{V_i}, \;\; j \geq i. \end{aligned} \tag{3.12}$$

Then, $\bar{\theta}_T - \theta^\star$ has the closed form as follows:

$$\bar{\theta}_T - \theta^\star = \frac{1}{T}\sum_{i=1}^{T} V_i \Delta_0 + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=t}^{T} V_t^i \eta_t \epsilon_t,$$

where Δ_0 is the initialization error $\theta_0 - \theta^\star$. Since $\{\epsilon_t\}_{t\geq 1}$ is a sequence of i.i.d. random errors, the estimation error above (deducted by the initialization error) can be view as the weighted sum of T i.i.d. random variables with mean 0. We can then further analyze the estimation error based on existing studies about deviations and tail probabilities of linear processes.

3.3.1 Upper bound from Nagaev inequality

The Nagaev inequality [Nagaev, 1979] for tail probability is a useful result in probability theory. It is known that the performance bounds obtained from Nagaev inequality are nearly sharp under polynomial moment conditions.

Proposition 3.3.1. *Consider the mean estimation model in* (3.9) *and the SGD iterates* $\{\theta_t\}_{t=1,\dots,T}$ *defined in* (3.10)*. For any* $x > 0$ *and* $2 < q < \nu$*, we have*

$$\mathbb{P}\left(\left|\bar{\theta}_T - \theta^\star\right| \geq \frac{C|\Delta_0|}{T} + x\right) \leq \frac{(1+2/q)^q \mathbb{E}|\epsilon|^q}{x^q T^{q-1}} + 2\exp\left(-c_q x^2 T\right), \qquad (3.13)$$

where $c_q = 2e^{-q}(q+1)^{-2}/\mathbb{E}|\epsilon|^2$*, and* $C = \sum_{i=1}^{\infty} V_i, V_i$ *is defined in* (3.12)*.*

While the Nagaev inequality gives more precise constants, the upper bound in (3.13) is of the same order as that in (3.11). Thus, the tightness of Nagaev inequality implies that our proposed Nagaev type upper bound is also tight.

3.3.2 Exact deviation

Furthermore, instead of an upper bound, we give the exact asymptotic tail probability of the estimation error in the mean estimation model. Inspired by Peligrad et al. [2014], which studied the exact moderate and large deviation of linear processes, we obtain Proposition 3.3.2.

Proposition 3.3.2. *Consider the mean estimation model* (3.9) *and the SGD iterates* $\{\theta_t\}_{t=1,\ldots,T}$ *defined in* (3.10). *Define*

$$\sigma_T^2 = \mathbb{E}|\epsilon|^2 \sum_{t=1}^{T} \left(\sum_{i=t}^{T} V_t^i \eta_t / T \right)^2 .$$

For $x \geq \sigma_T$,

$$\mathbb{P}\left(\left| \bar{\theta}_T - \theta^\star - \frac{\gamma_T \Delta_0}{T} \right| \geq x \right) = (2 + o(1)) \left(1 - \Phi(x/\sigma_T) + R(T,x) \right), \tag{3.14}$$

where $\gamma_T = \sum_{i=1}^{T} V_i, V_i$ *is defined in* (3.12), *and*

$$R(T,x) = \sum_{t=1}^{T} \mathbb{P}\left(\epsilon_t \geq Tx / \sum_{i=t}^{T} V_t^i \eta_t \right) .$$

The right-hand-side (RHS) of (3.14) comprises two parts: Gaussian approximation $1 - \Phi(x/\sigma_T)$ and tail approximation $R(T,x)$. Note that $\sigma_T^2 \asymp 1/T$ as discussed in Section B.1.6 . Then the Gaussian approximation refines the term $\exp(-C_2 T x^2)$ in (3.11). Also,

$$\mathbb{P}\left(\epsilon_t \geq y\right) \sim c_\nu / y^\nu, y \to \infty,$$

where $c_\nu = \nu^{-3/2} \pi^{-1/2} \Gamma((\nu+1)/2) / \Gamma(\nu/2)$ according to the property of t_ν distribution, and $\sum_{i=t}^{\infty} V_t^i \eta_t = O(1)$ as discussed in Section B.1.6. Then the tail approximation

$$R(T,x) \asymp 1/(x^\nu T^{\nu-1}),$$

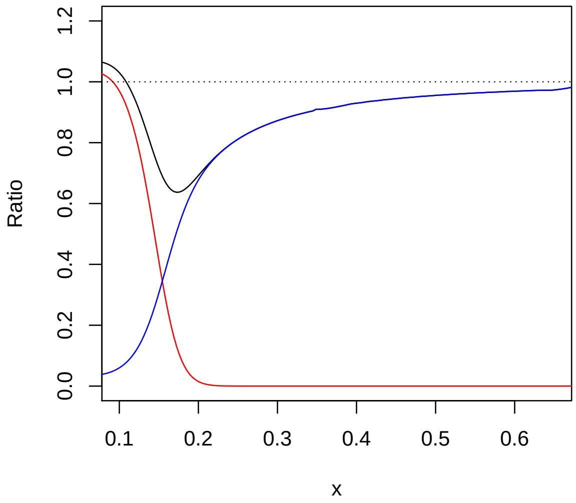


Figure 3.2: Ratio of approximated and true tail probability. Here X axis represents deviation x. Red curves represent Gaussian approximation: $\left(1-\Phi(x/\sqrt{\mu_{T,2}})\right)/\mathbb{P}(S_T \geq x)$; blue curves represent tail approximation: $R(T,x)/\mathbb{P}(S_T \geq x)$; black curves represent their sum: $\left(1-\Phi(x/\sqrt{\mu_{T,2}})+R(T,x)\right)/\mathbb{P}(S_T \geq x)$.

matching the polynomial term in our proposed Nagaev type upper bound (3.11). Therefore, we can see that the tail probability polynomial dependence on $1/\delta$ is necessary in the tail bound of SGD and sub-Gaussian tails cannot be achieved under heavy-tailed assumptions.

3.3.3 A numerical study

We conduct a numerical study of the accuracy of the exact tail probability in (3.14) for $\nu = 3$. The true tail probability of the estimation error (LHS of (3.14)) can be calculated through the inversion formula. Let

$$S_T = \bar{\theta}_T - \theta^* - \frac{1}{T}\sum_{i=1}^{T} V_i \Delta_0 = \sum_{t=1}^{T}\left(\sum_{i=t}^{T} V_t^i \eta_t / T\right)\epsilon_t.$$

Then the characteristic function of S_T is

$$\phi_{S_T}(x) = \prod_{t=1}^{T} \phi\left(\left(\sum_{i=t}^{T} V_t^i \eta_t / T\right) x\right),$$

where ϕ is the characteristic function of a t_3-distribution. By the inversion formula,

$$\mathbb{P}\left(S_T \leq x\right) - \mathbb{P}\left(S_T \leq 0\right) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{\sqrt{-1}yx} - 1}{\sqrt{-1}y} \phi_{S_T}(y) dy.$$

Since S_T is symmetric, $\mathbb{P}\left(S_T \leq 0\right) = \frac{1}{2}$. In our numerical study, we use the above formula to compute the probability $\mathbb{P}(S_T \geq x)$. In figure 3.2, we report the ratios $R(T, x)/\mathbb{P}(S_T \geq x)$, $\left(1 - \Phi(x/\sqrt{\mu_{T,2}})\right)/\mathbb{P}(S_T \geq x)$ and $\left(1 - \Phi(x/\sqrt{\mu_{T,2}}) + R(T, x)\right)/\mathbb{P}(S_T \geq x)$. We can see that the Gaussian approximation is good for small deviations, while the tail approximation is better when the deviation is moderate or large. The numerical study confirms that the polynomial term in the upper bound (3.11) is necessary in the case of heavy-tailed gradient noise, especially for moderate and large deviations.

3.4 Summary

In this paper, we established nearly tight tail probabilities for SGD errors with heavy-tailed noises in linear models. The resulting high probability error bounds and sample complexity are quite different from those obtained in light-tailed noise cases. In particular, with probability at least $1 - \delta$, we have $\|\bar{\theta}_T - \theta^\star\|_2^2 \leq O\left(T^{-1} \log(1/\delta) + (\delta T^{q-1})^{-2/q}\right)$, where the polynomial dependence on the failure probability δ is generally unavoidable. For future directions, it is interesting to extend our concentration analysis under heavy-tailed noise assumptions to other examples of SA. Also, the robust modification of SGD can be a promising topic to accommodate the heavy-tailed noise.

CHAPTER 4

ONLINE COVARIANCE MATRIX ESTIMATION IN STOCHASTIC GRADIENT DESCENT

From this chapter, we consider the problem of practical inference in the online setting where data can arrive sequentially. In this section, we focus on the estimation of the limiting covariance matrix, which appears in the asymptotic normality results discussed in the previous chapter. In particular, we introduce a fully online approach to estimate the covariance matrix using only the iterates from SGD. We can then construct confidence intervals for model parameters based on the estimated covariance matrix and asymptotic normality results.

4.1 Introduction

Recall that applying vanilla SGD algorithm (1.2) to solve the problem (1.1), we obtain iterates $\{x_i\}_{i\geq 1}$. In this chapter, we will use the ASGD iterate

$$\bar{x}_n = n^{-1}\sum_{i=1}^{n} x_i$$

as the final estimate for the model parameter at the n-th step, and set step size $\eta_i = \eta i^{-\alpha}(i \geq 1)$ with $\eta > 0$ and $\alpha \in (0.5, 1)$ as suggested by Polyak and Juditsky [1992]. Recall the definition

$$A = \nabla^2 F(x^*),\ S = \mathbb{E}\left([\nabla f(x^*, \xi)][\nabla f(x^*, \xi)]^T\right). \tag{4.1}$$

From Polyak and Juditsky [1992], under suitable conditions, $\bar{x}_n$ has the asymptotic normality:

$$\sqrt{n}(\bar{x}_n - x^*) \Rightarrow N(0, \Sigma), \tag{4.2}$$

where $\Sigma = A^{-1}SA^{-1}$, which is known as the “sandwich” form of the covariance matrix. To leverage the asymptotic normality result for inference, it is critical to estimate the limiting covariance matrix Σ. Intuitively, one can estimate S with a simple sample average $\hat{S}_n = n^{-1}\sum_{i=1}^{n}[\nabla f(x_{i-1},\xi_i)][\nabla f(x_{i-1},\xi_i)]^T$, and similarly estimate A with $\hat{A}_n = n^{-1}\sum_{i=1}^{n}\nabla^2 f(x_{i-1},\xi_i)$. Then the limiting covariance matrix Σ can be estimated by the consistent plug-in estimator $\hat{A}_n^{-1}\hat{S}_n\hat{A}_n^{-1}$ (see Chen et al. [2020]). However, computation of the Hessian matrix of the loss function is not always available, e.g., certain computations are not available in many existing codebases that only adopt SGD for optimization and in cases such as quantile regression, the Hessian matrix does not even exist. Also, the plug-in estimator may be computationally costly when d is large since it involves matrix inversion with $O(d^3)$ time complexity in general.

Our goal is to obtain an online estimate of the covariance matrix of $\sqrt{n}\bar{x}_n$, only through the SGD iterates $\{x_1, x_2, ..., x_n\}$. Our approach is attractive in situations where the computation for A^{-1} and S are difficult, which is quite typical in practice. Also, the approach is efficient in both computation and memory due to its recursive property, i.e., the estimate at n-th step $\hat{\Sigma}_n$ can be updated from $\hat{\Sigma}_{n-1}$ within $O(d^2)$ computation. With the estimate, we can perform uncertainty quantification and statistical inference with desirable computation and memory efficiency. The approach is useful for online learning, where the data is constantly arriving over time, such as streaming data.

For the time-homogeneous Markov chain, $\{x_i\}_{i\in\mathbb{Z}}$ is a stationary process. Under certain short-range dependence conditions, we have

$$\sqrt{n}\,(\bar{x}_n - \mathbb{E}x_i) \Rightarrow N(0,\sigma^2),$$

where

$$\sigma^2 = \lim_{n\to\infty}\mathrm{Var}(\sqrt{n}\bar{x}_n) = \sum_{i=-\infty}^{\infty}\mathrm{cov}(x_0, x_i)$$

is the long-run variance, and it plays a fundamental role in the statistical inference of stationary processes. To estimate the long-run variance, one can apply the batch-means method

[Glynn and Whitt, 1991, Flegal and Jones, 2010, Politis et al., 1999, Lahiri, 2003, Kitamura et al., 1997]. Given $x_1, ..., x_n$, let $1 \leq l_n \leq n$ be the batch size. Based on batch-means $\sum_{k=i}^{i+l_n} x_k/l_n - \bar{x}_n$ for $1 \leq i \leq n - l_n + 1$, one can estimate σ^2 by

$$\sigma_n^2 = \frac{l_n}{n - l_n + 1} \sum_{i=1}^{n-l_n+1} \left(\sum_{k=i}^{i+l_n-1} x_k/l_n - \bar{x}_n \right)^2 .$$

As an alternative, one can use the non-overlapping batch-means $\sum_{k=i}^{i+l_n} x_k/l_n - \bar{x}_n$ for $i = 1, 1 + l_n, 1 + 2l_n, ...,$ to construct a similar estimate. Properties of overlapping and non-overlapping batch-means estimators are discussed in Politis et al. [1999] and Lahiri [2003]. In our problem, estimation of Σ in (4.2) becomes more complicated since SGD iterates form a non-stationary Markov Chain.

To apply to SGD, Chen et al. [2020] modified the classic non-overlapping batch-means by allowing increasing batch sizes and showed that the modified batch-means estimator is consistent. However, their approach is not in line along with the spirit of SGD, the fully online fashion. Their construction of covariance estimator $\hat{\Sigma}_n$ requires the information on the total number of iterations n a priori. There is no simple algebraic relation between $\hat{\Sigma}_n$ and $\hat{\Sigma}_{n+1}$. In other words, when a new data point x_{n+1} arrives later, their algorithm needs to recompute their estimate from the beginning and cannot perform efficient sequential updating. So the approach is computationally expensive for online learning, where the dynamic training data is arriving over time, and the goal is to make sequential predictions; see Remark 4.2.1 for a detailed discussion of Chen et al. [2020].

To address the above problems, we develop a fully *online approach* for asymptotic covariance matrix estimation, which we refer to as online batch means method. The construction does not require prior knowledge of the total sample size. Immediate updates from $\hat{\Sigma}_n$ to $\hat{\Sigma}_{n+1}$ can be performed recursively as new data is coming in, which fits our online setting. To achieve this goal, we design a novel construction of batches with time-varying size, which substantially extends the one in Chen et al. [2020]. Similar to the recursive nature of SGD, our algorithm is also recursive and it updates the covariance matrix estimate once at a time only through the stochastic gradient within $O(d^2)$ computation. Note that since we are learning a $d \times d$ covariance matrix, it requires at least $O(d^2)$ computation to update the covariance matrix estimates. In the important special case of marginal inference of each coordinate of the parameter vector, our online batch means estimator only needs to compute and store diagonals of the covariance matrix estimate, which only require $O(d)$ computation and $O(d)$ memory. The idea of online estimation is motivated by Wu [2009], who studied the estimation of long-run variances of stationary and ergodic processes. As mentioned above, the SGD iterates in (1.2) form a non-homogeneous (non-stationary) Markov Chain since the step size η_k decays as k increases, for example $\eta_k = \eta k^{-\alpha}$ for $\alpha \in (1/2, 1)$ as suggested by Polyak and Juditsky [1992]. Hence, the asymptotic behaviors of SGD and stationary processes are fundamentally different. The construction, which is associated with batch sizes, is novel and different for SGD iterates and stationary sequences. This non-stationarity also brings substantial difficulties in technical analysis. The convergence of our estimator is far from being trivial. We formally establish the consistency result and obtain the convergence rate of our online estimator in Section 6.4.

4.2 Online approach

We first introduce a time varying batch scheme used in our online approach. Consider infinite sequentially arriving SGD iterates $\{x_i\}_{i=1,2,...}$ in (1.2). Let $\{a_m\}_{m \in \mathbb{N}}$ be a strictly increasing integer-valued sequence with $a_1 = 1$. For the i-th iterate x_i, we consider a data block B_i including iterates from past iterations t_i to i, i.e.,

$$B_i = \{x_{t_i}, ..., x_i\},$$

where t_i is the index of iterate we trace back to at the i-th step. The value of t_i is determined by the sequence $\{a_m\}_{m\in\mathbb{N}}$ through $t_i = a_m$ when $i \in [a_m, a_{m+1})$. For example, $t_i = \lfloor\sqrt{i}\rfloor^2$ if $a_m = m^2$. In this case we have:

$B_1 = \{x_1\}$, $B_2 = \{x_1, x_2\}$, $B_3 = \{x_1, x_2, x_3\}$,
$B_4 = \{x_4\}$, $B_5 = \{x_4, x_5\}$, $B_6 = \{x_4, x_5, x_6\}$, $B_7 = \{x_4, x_5, x_6, x_7\}$, $B_8 = \{x_4, x_5, x_6, x_7, x_8\}$,
$B_9 = \{x_9\}$, $B_{10} = \{x_9, x_{10}\}$, $B_{11} = \{x_9, x_{10}, x_{11}\}, \ldots$.

We can see that the batch sizes are time-varying. The blocks $\{B_i : a_m \leq i < a_{m+1}\}$ can also be viewed as the so-called "forward scans" in block subsampling [McElroy et al., 2007, Nordman et al., 2013]. That is, given non-overlapping blocks $\{x_{a_m}, ..., x_{a_{m+1}-1}\}$, the forward scans are overlapping blocks of sequentially increasing length starting from x_{a_m}.

4.2.1 Online covariance matrix estimator based on batch means

Based on blocks $\{B_i\}_{i\in\mathbb{N}}$, the covariance matrix estimator is defined as the sum of squared block sums (centered) divided by the sum of block lengths, i.e., at the n-th step

$$\hat{\Sigma}_n = \frac{\sum_{i=1}^{n}\left(\sum_{k=t_i}^{i} x_k - l_i\bar{x}_n\right)\left(\sum_{k=t_i}^{i} x_k - l_i\bar{x}_n\right)^T}{\sum_{i=1}^{n} l_i}, \tag{4.3}$$

where $l_i = |B_i| = i - t_i + 1$ denotes the length of B_i. The novel idea of constructing data block B_i, which only includes past iterates, is the key to make the algorithm fully online. Next, we will show that the estimate $\hat{\Sigma}_n$ can be computed recursively. Let W_i denote the sum of the block $B_i = \{x_{t_i}, ..., x_i\}$, i.e.,

$$W_i = \sum_{k=t_i}^{i} x_k. \tag{4.4}$$

When $t_{i+1} = t_i = a_m$ for some m, $B_{i+1} = B_i \cup \{x_{i+1}\}$ and

$$W_{i+1} = W_i + x_{i+1},\ l_{i+1} = l_i + 1.$$

When $t_{i+1} = a_{m+1}$ for some m, we start a new block $B_{i+1} = \{x_{i+1}\}$ and

$$W_{i+1} = x_{i+1},\ l_{i+1} = 1.$$

We can see that both the batch sum W_i and the batch length l_i can be updated recursively. With the notation of W_i, the estimator in (4.3) can be expressed as

$$\hat{\Sigma}_n = \frac{\sum_{i=1}^n W_i W_i^T + \sum_{i=1}^n l_i^2 \bar{x}_n \bar{x}_n^T - \left(\sum_{i=1}^n l_i W_i\right) \bar{x}_n^T - \bar{x}_n \left(\sum_{i=1}^n l_i W_i\right)_n^T}{\sum_{i=1}^n l_i}. \tag{4.5}$$

To further simplify the form, we introduce

$$\begin{aligned} V_n &= \sum_{i=1}^n W_i W_i^T\ ,\ P_n = \sum_{i=1}^n l_i W_i. \\ v_n &= \sum_{i=1}^n l_i, \quad \text{and} \quad q_n = \sum_{i=1}^n l_i^2. \end{aligned} \tag{4.6}$$

They can be computed recursively since both W_i and l_i can be updated recursively. Now, $\hat{\Sigma}_n$ in (4.3) can be finally rewritten as

$$\hat{\Sigma}_n = \frac{V_n + q_n \bar{x}_n \bar{x}_n^T - P_n \bar{x}_n^T - \bar{x}_n P_n^T}{v_n}. \tag{4.7}$$

All five components in (4.7): $V_n, q_n, P_n, v_n, \bar{x}_n$ can be updated recursively. Thus, $\hat{\Sigma}_n$ can be updated through results in the $(n-1)$-th step and the new iterate x_n within $O(d^2)$ computation.

To summarize, we propose Algorithm 2. As shown in Algorithm 2, the five components

Algorithm 2: Update ASGD iterate and covariance matrix estimate recursively

```
Input: function $f(\cdot)$, parameter $(\alpha, \eta)$, step size $\eta_i = \eta i^{-\alpha}$ for $i \geq 1$, pre-defined
 sequence $\{a_m\}_{m \in N}$.
Initialize: $m_0 = l_0 = 0, v_0 = P_0 = q_0 = V_0 = W_0 = \bar{x}_0 = 0, x_0$;
For $n$ = 0, 1, 2, 3, ...
    Receive: new data $\xi_{n+1}$
    Do the following update:
        1. $x_{n+1} = x_n - \eta_{n+1} \nabla f(x_n, \xi_{n+1})$;
        2. $\bar{x}_{n+1} = (n\bar{x}_n + x_{n+1})/(n+1)$;
        3. if $n + 1 = a_{m_n+1}$, then:
            $m_{n+1} = m_n + 1$; $l_{n+1} = 1$; $W_{n+1} = x_{n+1}$;
         else:
            $m_{n+1} = m_n$; $l_{n+1} = l_n + 1$; $W_{n+1} = W_n + x_{n+1}$;
        4. $q_{n+1} = q_n + l_{n+1}^2$;
        5. $v_{n+1} = v_n + l_{n+1}$;
        6. $V_{n+1} = V_n + W_{n+1} W_{n+1}^T$;
        7. $P_{n+1} = P_n + l_{n+1} W_{n+1}$;
        8. $S = V_{n+1} + q_{n+1} \bar{x}_{n+1} \bar{x}_{n+1}^T - P_{n+1} \bar{x}_{n+1}^T - \bar{x}_{n+1} P_{n+1}^T$;
    Output: ASGD estimator $\bar{x}_{n+1}$, estimated covariance $\hat{\Sigma}_{n+1} = S/v_{n+1}$
```

of $\hat{\Sigma}_{n+1}$ can be easily updated from their values in the n-th step. There is no need to store all the outcomes in the previous steps. The memory complexity is $O(d^2)$, independent of the sample size n. In the update step, the computational complexity is also $O(d^2)$. The total computational cost scales linearly in n. The algorithm is much more efficient compared to non-recursive methods and naturally fits online learning scenarios.

An alternative version

The estimate $\hat{\Sigma}_n$ in (4.3) includes squared block sums from all n blocks $\{B_i\}_{i=1,2,\ldots,n}$. Block B_i and B_j are overlapped when $a_m \leq i < j < a_{m+1}$ for some m. So $\hat{\Sigma}_n$ in (4.3) is a full overlapping version of the online batch means estimator. We also introduce an alternative non-overlapping version with a slightly simpler form which has a comparable performance. As data arriving sequentially, we follow the same batch scheme above to construct $\{B_i\}_{i=1,2,\ldots,}$ while only include a few squared block sums. At the n-th step, define set

Algorithm 3: Update ASGD estimator and covariance matrix estimate (non-overlapping version) recursively

```
Input: function $f(\cdot)$, parameter $(\alpha, \eta)$, step size $\eta_i = \eta i^{\alpha}$ for $i \geq 1$, pre-defined
  sequence $\{a_m\}_{m\in\mathbb{N}}$.
Initialize: $m_0 = l_0 = 0, v_0 = P_0 = q_0 = V_0 = W_0 = \bar{x}_0 = 0, x_0$;
For $n$ = 0, 1, 2, 3, ...
      Receive: new data $\xi_{n+1}$
      Do the following update:
         1. $x_{n+1} = x_n - \eta_{n+1}\nabla f(x_n, \xi_{n+1})$;
         2. $\bar{x}_{n+1} = (n\bar{x}_n + x_{n+1})/(n+1)$;
         4. if $n+1 = a_{m_n+1}$, then:
            $m_{n+1} = m_n + 1$; $l_{n+1} = 1$; $W_{n+1} = x_{n+1}$;
            $q_{n+1} = q_n + l_n^2$; $V_{n+1} = V_n + W_n W_n^T$; $P_{n+1} = P_n + l_n W_n$
          else:
            $m_{n+1} = m_n$; $l_{n+1} = l_n + 1$; $W_{n+1} = W_n + x_{n+1}$;
            $q_{n+1} = q_n$; $V_{n+1} = V_n$; $P_{n+1} = P_n$
         5. $S' = W_{n+1}W_{n+1}^T + l_{n+1}^2\bar{x}_{n+1}\bar{x}_{n+1}^T - l_{n+1}W_{n+1}\bar{x}_{n+1}^T - l_{n+1}\bar{x}_{n+1}W_{n+1}^T$;
         6. $S = V_{n+1} + q_{n+1}\bar{x}_{n+1}\bar{x}_{n+1}^T - P_{n+1}\bar{x}_{n+1}^T - \bar{x}_{n+1}P_{n+1}^T + S'$;
      Output: ASGD estimator $\bar{x}_{n+1}$, estimated covariance $\hat{\Sigma}_{n+1,NOL} = S/(n+1)$
```

$S_n = \{n\}\bigcup\{a_i - 1 : i > 1, a_i \leq n\}$. Consider a set of non-overlapping blocks $\{B_i\}_{i\in S_n}$, i.e.,

$$\{\underbrace{\{x_{a_1},..., x_{a_2-1}\}}_{B_{a_2-1}}, ..., \underbrace{\{x_{a_{m-1}},..., x_{a_m-1}\}}_{B_{a_m-1}}, \underbrace{\{x_{a_m},..., x_n\}}_{B_n}\}.$$

The alternative non-overlapping estimate at the n-th step includes squared block sums of $\{B_i\}_{i\in S_n}$. It is then defined as

$$\hat{\Sigma}_{n,NOL} = \frac{1}{n}\sum_{i\in S_n}\left(\sum_{k=t_i}^{i} x_k - l_i\bar{x}_n\right)\left(\sum_{k=t_i}^{i} x_k - l_i\bar{x}_n\right)^T. \tag{4.8}$$

The non-overlapping version estimator is also recursive and can perform a real-time update. The algorithm is almost the same as the overlapping one with same computational and memory complexity. One can follow the derivation of Algorithm 2 to get Algorithm 3.

In the stationary process case, Lahiri [2003, 1999] showed that the mean squared error of the classic (non-recursive) non-overlapping batch-means estimate is 33% larger than that of its overlapping version, while the convergence rates are the same. The comparison between the full overlapping version and the non-overlapping version of our online estimators is more complicated in the non-stationary case. In Section 4.3.3, we provide upper bounds for estimation errors for both overlapping and non-overlapping estimators. The two upper bounds are of the same order. The non-overlapping version is easier to analyze theoretically, given its simpler structure. In the mean estimation model, we can obtain the precise order of the mean squared error for the non-overlapping one; see Section 4.3.1. We also compare the empirical performance of the two versions in Section 4.4.1. However, it is hard to tell which one is more efficient based on the simulation results. We leave the rigorous comparison as a future research problem by extending Lahiri [2003] to non-stationary processes.

Remark 4.2.1 (Comparison with the non-recursive batch-means covariance matrix estimator)**.** *The non-overlapping version* (4.8) *appears similar to the batch-means estimator [Chen et al., 2020]. However, the batch schemes of the two methods are fundamentally different. Chen et al. [2020] split n iterates of SGD into $M+1$ non-overlapping blocks, where M and batch sizes $b_{m,n}$ $(m = 0, ..., M)$ are chosen based on n for desired convergence. With $e_{m,n}$ denoting the ending index of the k-th block, the covariance matrix estimator at n-th iteration in Chen et al. [2020] is defined as*

$$\hat{\Sigma}_{n,BM} = \frac{1}{M}\sum_{m=1}^{M} b_{m,n}\left(\sum_{k=e_{m-1,n}+1}^{e_{m,n}} x_k/b_{m,n} - \bar{x}_n\right)\left(\sum_{k=e_{m-1,n}+1}^{e_{m,n}} x_k/b_{m,n} - \bar{x}_n\right)^T, \quad (4.9)$$

where $e_{M,n} = n$. The optimal batch size setting as suggested in Chen et al. [2020] is $e_{m,n} = ((m+1)/(M+1))^{1/(1-\alpha)}n$ with the number of batches $M = n^{(1-\alpha)/2}$. Since $e_{m,n}$ must depend on n to ensure the desired convergence rate at the n-th iteration, there is no simple algebraic relation between $\hat{\Sigma}_{n,BM}$ and $\hat{\Sigma}_{n+1,BM}$. So the batch-means estimator [Chen et al., *2020] is only suitable for offline tasks requiring final prediction/inference given the* prespecified *total sample size n. In contrast, our fully online estimator can sequentially improve over each iteration. Also, n does not need to be specified beforehand.*

Remark 4.2.2 (Choice of batch-sizes when n is unknown)**.** *Chen et al. [2020] also propose an approach based on a target error tolerance to apply the batch-means estimator when n is unknown. In particular, given the pre-specified error ϵ, Chen et al. [2020] propose to set the ending index of the k-th batch by $e_k = \left((k+1)C\epsilon^{-2}\right)^{1/(1-\alpha)}$, where C is a constant. The approach indeed enables an online updating, thus achieve the goal of recursive processing. However, choosing the constant C can be difficult or arbitrary in online settings. Moreover, there is a fundamental difference. The approach in Chen et al. [2020] only ensures that the expected spectrum norm loss of the covariance matrix is smaller than ϵ (up to a constant) for large n, rather than goes to* 0*. In other words, the covariance matrix estimator is not necessarily* consistent*. While our online method constantly improves the covariance matrix estimate as $n \to \infty$, and the estimation error goes to* 0*.*

Choice of batch sizes

The remaining question is to specify the sequence $\{a_m\}_{m\in\mathbb{N}}$. This pre-defined sequence does not depend on n. This ensures that we can construct batches even if the total number of data is unknown (which is a typical situation), and the incoming data will not affect the recursive estimation process. In Section 4.3.3, we show that a_m is required to take a polynomial form so that the estimator is consistent. Next, we shall give some intuitive explanation and one example of choice.

The formula in (4.3) bears a certain similarity to the sample covariance matrix $S_n = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)^T$. On the other hand, in contrast to the standard sample covariance matrix where $\{x_i\}_{i\geq 1}$ are independent, our SGD iterates in (4.3) are highly correlated. In other words, we cannot ignore the covariance between data as in the construction of the sample covariance matrix. According to (1.2), the correlation between x_i and x_j diminishes as the distance $|j-i|$ becomes larger, while the correlation between x_i and x_{i+1} becomes stronger as i goes to infinity. The idea of online estimation is to choose sequence $(a_m)_{m\in\mathbb{N}}$ and form non-overlapping blocks $\{B_{a_m-1}\}_{m>1}$ as mentioned above such that the correlation between x_i and x_j is sufficiently small when they are in different non-overlapping blocks. So when considering the effect of x_i, we trace back to the starting point of the non-overlapping block x_i belongs to, i.e., construct data block $B_i = \{x_{t_i}, ..., x_i\}$. Recall that the i-th iterate x_i through SGD takes the form

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

Let $\delta_i = x_i - x^*$ be the error sequence, where x^* is the minimizer in (1.1). Then

$$\delta_i = \delta_{i-1} - \eta_i \nabla F(x_{i-1}) + \eta_i \epsilon_i, \tag{4.10}$$

where $\epsilon_i = \nabla F(x_{i-1}) - \nabla f(x_{i-1}, \xi_i)$. Note that $\nabla F(x^*) = 0$ since x^* is the minimizer of $F(x)$. By Taylor's expansion of $\nabla F(x_{i-1})$ around x^*, we have $\nabla F(x_{i-1}) \approx \nabla A\delta_{i-1}$, where $A = \nabla^2 F(x^*)$. Thus, by modifying equation (4.10) with $\nabla F(x_{i-1})$ approximated by $A\delta_{i-1}$, for large i

$$\delta_i \approx (\mathbf{I}_d - \eta_i A)\delta_{i-1} + \eta_i \epsilon_i. \tag{4.11}$$

Then for the i-th iterate x_i and the j-th iterate x_j (assume $j < i$), the strength of correlation between them is roughly

$$\Pi_{k=j+1}^{i} \|\mathbf{I}_d - \eta_k A\|_2 \leq (1 - \eta\lambda_A i^{-\alpha})^{i-j}, \tag{4.12}$$

when $\eta_k = \eta k^{-\alpha}$ and λ_A is the smallest eigenvalue of A. To make the correlation small, one can choose $i - j \approx Ki^{(\alpha+1)/2}$, where K is a constant. Then the correlation is less than $(1 - \eta\lambda_A i^{-\alpha})^{Ki^{\alpha} i^{(1-\alpha)/2}}$, which goes to zero as i goes to infinity. Combining the correlation between x_i, x_j and the form of $i - j$, a reasonable setting is that the sequence $\{a_m\}_{m\in\mathbb{N}}$ satisfies

$$a_m - a_{m-1} = Ka_m^{(\alpha+1)/2}. \tag{4.13}$$

Let a_m increase polynomially, i.e., $a_m = Cm^{\beta}$ for some constant C. We obtain $\beta = 2/(1-\alpha)$ by solving equation (4.13). Thus a natural choice of a_m is

$$a_m = \left\lfloor Cm^{2/(1-\alpha)} \right\rfloor. \tag{4.14}$$

This is also the best choice in the general setting, as discussed in Section 4.3.3. However, the best choice of β may change considering specific objective functions.

4.2.2 Statistical inference

Now the limiting covariance matrix Σ can be approximated through the online estimation proposed above. Let $0 < q < 1$. Based on the asymptotic normality of ASGD in (4.2), the $(1 - q)100\%$ confidence interval for x_i^*, the i-th coordinate of x^*, can be constructed as

$$\left[\bar{x}_{n,i} - z_{1-q/2}\sqrt{\hat{\sigma}_{ii}/n},\ \bar{x}_{n,i} + z_{1-q/2}\sqrt{\hat{\sigma}_{ii}/n}\right], \tag{4.15}$$

where $\bar{x}_{n,i}$ is the i-th coordinate of $\bar{x}_n$, $z_{1-q/2}$ is the $(1 - q/2)$-th percentile of the standard Gaussian distribution and $\hat{\sigma}_{ii}$ is the i-th diagonal of the covariance matrix estimate. The confidence interval is constructed in a fully online fashion since both $\bar{x}_{n,i}$ and $\hat{\sigma}_{ii}$ can be computed recursively. Joint confidence regions and general form of confidence intervals are discussed in Section 4.3.4.

Relation to empirical likelihood

As pointed out by a reviewer, the construction of the non-overlapping version estimator shares a similar spirit with the blocking scheme and covariance estimator by Kim et al. [2013], who developed a progressive block empirical likelihood (PBEL) method. They consider a stationary, weakly dependent sequence $(X_1, ..., X_n)$ with mean μ such that the CLT $\sqrt{n}(\bar{X}_n - \mu) \Rightarrow N(0, \sigma^2)$ holds. The variance estimator $\hat{\sigma}^2_{n,NOL}$ in Kim et al. [2013] matches our scheme in Section 4.2.1 with $a_m = (m-1)m/2 + 1$ (or the i-th block has length i) and is shown to be a consistent variance estimator. The chi-squared limit of the log-likelihood ratio based on PBEL is established following the consistency of $\hat{\sigma}^2_{n,NOL}$. It would be interesting to see if one can obtain similar results as the PBEL ratio and establish a limiting distribution that can be used to calibrate confidence regions in the SGD case here.

4.3 Theoretical results

4.3.1 Preamble: mean estimation model

Before investigating the convergence property of the online batch means estimators in the general setting, we shall look at the simple mean estimation example. Taking advantage of the simpler structure of the non-overlapping version, we can obtain the exact order of convergence. Consider the mean estimation model:

$$y = x^* + e,$$

where $x^* \in \mathbb{R}$ is the mean we want to estimate, e is the random error with mean 0. Let $\{y_i\}_{i \in \mathbb{N}}$ be a sequence of $i.i.d$ sample from the model. Consider the squared loss function at x, $F(x) = (y - x)^2/2$. The i-th SGD iterate takes the form

$$x_i = x_{i-1} + \eta_i(y_i - x_{i-1}), i \geq 1, \tag{4.16}$$

where we choose the step size $\eta_i = \eta i^{-\alpha}$, $\alpha \in (1/2, 1)$. Then the error $\delta_i = x_i - x^*$ takes the form

$$\delta_i = (1 - \eta_i)\delta_{i-1} + \eta_i e_i.$$

In this case, one can have an explicit form of $\mathrm{var}(\sqrt{n}\bar{x}_n)$ and $\hat{\Sigma}_{n,NOL}$. Additionally, we can have an explicit form for the order of magnitude of the mean squared error of $\hat{\Sigma}_{n,NOL}$. Let the variance $\mathrm{var}(\sqrt{n}\bar{x}_n) = \sigma_n^2$. We have the following proposition.

Proposition 4.3.1. *For $m \geq 2$, let $a_m = \lfloor cm^\beta \rfloor$, where $\beta > 1$ and $c > 0$ are constants. Given the SGD iterates defined in* (4.16)*, we have*

$$\mathbb{E}(\hat{\Sigma}_{n,NOL} - \sigma_n^2)^2 \asymp n^{-1/\beta} + n^{2\alpha + 2/\beta - 2}. \tag{4.17}$$

Choose $\beta = 3/(2(1-\alpha))$. In the mean estimation model, the above proposition asserts that the convergence rate of the mean squared error of our recursive non-overlapping variance estimate is $n^{-2(1-\alpha)/3}$. For α close to 1/2, the latter rate approaches $n^{-1/3}$. This rate is faster than that of the batch-means estimator in Chen et al. [2020], which approaches $n^{-1/4}$. So, besides the advantage of the recursive property, our estimator may improve the convergence rate.

In the general setting, the analysis is much more complicated due to the nonlinearity. Upper bounds for the convergence rates of online estimators for both overlapping and non-overlapping versions are given in Section 4.3.3.

4.3.2 Assumptions and existing convergence results

In the work of Polyak and Juditsky [1992], assumptions on the objective function $F(x)$ and the gradient difference are proposed to prove the asymptotic normality of ASGD estimate. Those assumptions are necessary for our problem since we adopt the ASGD as the point estimator and require the asymptotic normality for statistical inference. Those assumptions, as well as some error bounds, are also proposed in other literature. We impose similar assumptions and review some existing results in this section.

Assumption 4.3.2. *Assume that the objective function $F(x)$ is continuously differentiable and strongly convex with parameter $\mu > 0$. That is, for any x_1 and x_2,*

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

Furthermore, assume that $\nabla^2 F(x^*)$ exists and $\nabla F(x)$ is Lipschitz continuous in the sense that there exist $L > 0$ such that,

$$\|\nabla F(x_1) - \nabla F(x_2)\|_2 \leq L\|x_1 - x_2\|_2.$$

Assumption 4.3.3. *For the n-th iteration, define error $\delta_n = x_n - x^*$ and gradient difference $\epsilon_n = \nabla F(x_{n-1}) - \nabla f(x_{n-1}, \xi_n)$. Recall that $\mathbb{E}_n(\cdot) = \mathbb{E}(\cdot|\xi_n, \xi_{n-1}, ...)$. The following hold:*

1). *The function $f(x, \xi)$ is continuously differentiable with respect to x for any ξ and $\|\nabla f(x, \xi)\|_2$ is uniformly integrable for any x. So $\mathbb{E}_{n-1}\left[\nabla f(x_{n-1}, \xi_n)\right] = \nabla F(x_{n-1})$, which implies that $\mathbb{E}_{n-1}(\epsilon_n) = 0$.*

2). *The conditional covariance of ϵ_n has an expansion around S which satisfies*

$$\left\|\mathbb{E}_{n-1}\left(\epsilon_n \epsilon_n^T\right) - S\right\|_2 \leq C\left(\|\delta_{n-1}\|_2 + \|\delta_{n-1}\|_2^2\right), \tag{4.18}$$

where $C > 0$ is some constant. Here S is defined in (4.1).

3). *There exists a constant $C > 0$ such that the fourth conditional moment of ϵ_n is bounded by*

$$\mathbb{E}_{n-1}\left(\|\epsilon_n\|_2^4\right) \leq C\left(1 + \|\delta_{n-1}\|_2^4\right).$$

Assumption 4.3.2 imposes strong convexity of the objective function $F(x)$ and Lipschitz continuity of its gradient. Assumption 4.3.3 asserts the regularity and the bound of the noisy gradient. These assumptions are widely used in SGD literature [Ruppert, 1988, Polyak and Juditsky, 1992, Moulines and Bach, 2011, Rakhlin et al., 2012]. With these assumptions, we have the asymptotic normality for averaged SGD iterates by Polyak and Juditsky [1992] and Ruppert [1988]. We also review the error bound for SGD iterates in Lemma 4.3.4.

Lemma 4.3.4. *Under Assumptions 4.3.2 and 4.3.3, for some constant $C > 0$ and $n_0 \in \mathbb{N}$, we have for any $n > n_0$, the sequence of error $\delta_n = x_n - x^*$ satisfies*

$$\begin{aligned} \mathbb{E}(\|\delta_n\|_2) &\leq Cn^{-\alpha/2}(1 + \|\delta_0\|_2), \\ \mathbb{E}(\|\delta_n\|_2^2) &\leq Cn^{-\alpha}(1 + \|\delta_0\|_2^2), \\ \mathbb{E}(\|\delta_n\|_2^4) &\leq Cn^{-2\alpha}(1 + \|\delta_0\|_2^4), \end{aligned}$$

when the step size is chosen to be $\eta_n = \eta n^{-\alpha}$ with $1/2 < \alpha < 1$.

4.3.3 Convergence properties for the online estimator

Theorem 4.3.5. *Under Assumptions 4.3.2 and 4.3.3, let $a_m = \left\lfloor Cm^{\beta} \right\rfloor$, where $C > 0$ is a constant, $\beta > (1 - \alpha)^{-1}$. Set step size at the i-th iteration as $\eta_i = \eta i^{-\alpha}$ with $1/2 < \alpha < 1$. Then for $\hat{\Sigma}_n$ defined in (4.3)*

$$\mathbb{E}\left\|\hat{\Sigma}_n - \Sigma\right\|_2 \lesssim n^{-1/(2\beta)} + n^{(\alpha-1)/2+1/(2\beta)}. \tag{4.19}$$

Theorem 4.3.5 shows that as $n \to \infty$, the estimator $\hat{\Sigma}_n$ converges to the limiting covariance matrix of the averaged SGD iterates in terms of operator norm loss. The convergence rate is associated with the parameters α and β. We state the following Corollary 4.3.6 to suggest the best choice of β.

Corollary 4.3.6. *Under conditions in Theorem 4.3.5 and let $\beta = 2/(1-\alpha)$, we have*

$$\mathbb{E}\left\|\hat{\Sigma}_n - \Sigma\right\|_2 \lesssim n^{-(1-\alpha)/4}. \tag{4.20}$$

Remark 4.3.7. *This convergence rate is the same as that of the non-recursive batch-means estimator in Chen et al. [2020]. According to Corollary 4.5 in Chen et al. [2020], the upper bound of the batch means estimator is also $O(n^{-(1-\alpha)/4})$ with the prior knowledge of the sample size n. So we make it possible that online estimation of covariance matrix achieves the same efficiency as offline methods. The plug-in approach in Chen et al. [2020] achieves the rate of $O(n^{-\alpha/2})$ when the i-th step size is chosen to be $i^{-\alpha}$. As a tradeoff, the online estimator enjoys efficient computation without the necessity of accessing Hessian information but pays the price in terms of the slower convergence rate.*

Next, we will show in Theorem 4.3.8 that the alternative version $\hat{\Sigma}_{n,NOL}$ shares the same upper bound.

Theorem 4.3.8. *Under conditions in Theorem 4.3.5, the alternative version $\hat{\Sigma}_{n,NOL}$ defined in (4.8) satisfies*

$$\mathbb{E}\left\|\hat{\Sigma}_{n,NOL} - \Sigma\right\|_2 \lesssim n^{-1/(2\beta)} + n^{(\alpha-1)/2+1/(2\beta)}. \tag{4.21}$$

4.3.4 Asymptotically accurate confidence intervals/regions

The next corollary shows that the confidence interval/region based on the online estimator achieves asymptotically correct coverage level $1-q$ for a pre-specified q with $0 < q < 1$.

Corollary 4.3.9. *Under conditions in Theorem 4.3.5, as n goes to infinity*

$$\mathbb{P}(x_i^* \in CI_{q,n,i}) \to 1 - q, \tag{4.22}$$

where

$$CI_{q,n,i} = \left[\bar{x}_{n,i} - z_{1-q/2}\sqrt{\hat{\sigma}_{ii}/n},\ \bar{x}_{n,i} + z_{1-q/2}\sqrt{\hat{\sigma}_{ii}/n}\right]$$

and $\hat{\sigma}_{ii}$ is the i-th diagonal of the online batch-means estimator $\hat{\Sigma}_n$ (or $\hat{\Sigma}_{n,NOL}$). We can also construct joint confidence regions as follows:

$$\mathbb{P}\left(x^* \in C_{q,n}\right) \to 1 - q, \tag{4.23}$$

where

$$C_{q,n} = \left\{x \in \mathbb{R}^d : n\left(\bar{x}_n - x\right)^T \hat{\Sigma}_n^{-1}\left(\bar{x}_n - x\right) \leq \chi^2_{d,1-2/q}\right\}.$$

Corollary 4.3.9 constructs asymptotic valid confidence intervals for each coordinate of x^* and joint confidence regions for $x^* \in \mathbb{R}^d$. More generally, for any unit length vector $w \in \mathbb{R}^d$ (i.e., $\|w\|_2 = 1$), we have by Theorem 4.3.5 and Slutsky's theorem,

$$\frac{\sqrt{n}w^T(\bar{x}_n - x^*)}{\sqrt{w^T\hat{\Sigma}_n w}} \Rightarrow N(0,1). \tag{4.24}$$

Therefore, the $(1-q)100\%$ confidence interval for $w^T x^*$ can be constructed as

$$\left[w^T\bar{x}_n - z_{1-q/2}\sqrt{w^T\hat{\Sigma}_n w/n}, w^T\bar{x}_n + z_{1-q/2}\sqrt{w^T\hat{\Sigma}_n w/n}\right]. \tag{4.25}$$

Stopping rule

In principle, SGD constantly improves the quality of $\bar{x}_n$, and our method constantly improves the covariance estimate $\hat{\Sigma}_n$ as n grows. A natural questions is when can we stop updating $\bar{x}_n$ and $\hat{\Sigma}_n$? There are several heuristics of stopping rules widely used in machine learning. For example, an online algorithm can stop when the neighboring estimates become sufficiently close. Or a more widely used approach in stopping SGD is to evaluate the error on a separate validation dataset and stops the SGD when the error becomes stable.

We can better answer this question and assess the SGD error based on the inference results, inspired by stopping rules for Markov Chain Monte Carlo (MCMC) that rely on a Markov chain central limit theorem. Especially, one can apply the fixed-width sequential stopping rule in Jones et al. [2006], where the updating is terminated the first time when the width of the confidence interval for each component is small enough. More formally, for a desired tolerance of ϵ_i for the i-th coordinate, the rule terminates updating the first time after the n-th iteration when the following condition is satisfied for all the coordinates $i = 1, \ldots, d$,

$$t_* \frac{\hat{\sigma}_{n,i}}{\sqrt{n}} + n^{-1} \leq \epsilon_i,$$

where $\hat{\sigma}_{n,i}$ is the i-th diagonal of the online estimator $\hat{\Sigma}_n$ (or $\hat{\Sigma}_{n,NOL}$), and t_* is an appropriate t-distribution quantile. For the joint inference, one may consider simplifying the relative standard deviation fixed-volume sequential stopping rule in Vats et al. [2019], where updating is terminated the first time when the volume of the confidence region C_n (4.23) is small enough. For a desired tolerance of ϵ, the rule terminates updating the first time after the n-th iteration when

$$\mathrm{Vol}(C_n)^{1/d} + n^{-1} \leq \epsilon,$$

where $\mathrm{Vol}(C_n) = 2\left(\pi\chi_*^2/n\right)^{d/2} |\hat{\Sigma}_n|^{1/2}/(d\Gamma(d/2))$, $|\cdot|$ denotes determinant, χ_*^2 is an appropriate chi-square distribution quantile, and $\hat{\Sigma}_n$ is our online estimator. We also include a simple simulation study of the stopping rule in the last section of the Supplement.

Remark 4.3.10. *The original stopping rule in Vats et al. [2019] avoids the practical issue of choosing ϵ with the idea of effective sample size (ESS). They consider an F-invariant Harris recurrent Markov chain and define a multivariate approach to ESS. The stopping rule in Vats et al. [2019] terminates the MCMC simulation the first time the estimated ESS is larger than a pre-specified lower bound. However, we need to re-define ESS in the non-stationary case, which requires more careful considerations. We will leave it as a future research direction.*

4.4 Simulation studies

In this section, we evaluate the empirical performance of the proposed online approach. We focus on two classes of examples: linear regression and logistic regression. Let $\{\xi_i \equiv (a_i, b_i)\}_{i=1,2,...}$ denotes an *i.i.d* sequence of pairs, and x^* denote the true parameter in the models. In both linear regression and logistic regression cases, $a_i \in \mathbb{R}^d$ is generated from $N(0, \mathbf{I}_d)$. In the former case, $b_i = a_i^T x^* + \epsilon_i$, where ϵ_i is independently generated from $N(0, 1)$. In the latter case, $b_i | a_i \sim Bernoulli((1 + \exp(-a_i^T x^*))^{-1})$. The loss function $f(\cdot)$ is defined as the negative log likelihood function, so we have

$$f(x, a_i, b_i) = \begin{cases} \frac{1}{2}(a_i^T x - b_i)^2 & \text{linear regression} \\ (1 - b_i)a_i^T x + \log(1 + \exp(-a_i^T x)) & \text{logistic regression.} \end{cases}$$

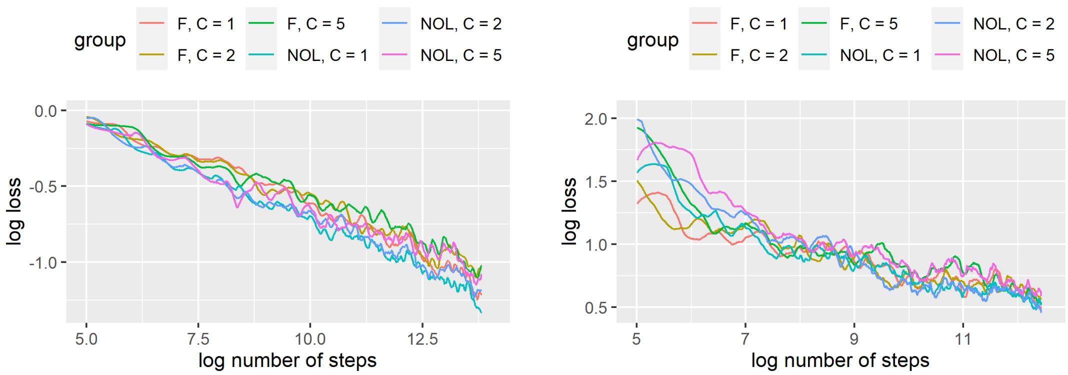
The true coefficient x^* is a d-dimensional vector linearly spaced between 0 and 1. In the SGD procedure, the step size η_j is set to be $0.5j^{-\alpha}$ and the parameter α is chosen to be 0.505. The sequence $\{a_k\}_{k\geq 1}$ in our online approach is chosen in the form of $a_m = \left\lfloor Cm^{2/(1-\alpha)} \right\rfloor$, for some constant C. All the measurements in the following discussions are averaged over 200 independent runs.

4.4.1 Empirical performance of the proposed online approach

Convergence of the recursive estimator. We focus on linear regression here since the true limiting covariance matrix is easy to compute. In the linear regression model described above,

$$A = \mathbb{E}\left[\nabla^2 f(x^*)\right] = \mathbb{E}\left(aa^T\right) = \mathbf{I}_d,$$

$$S = \mathbb{E}\left([\nabla f(x^*, \xi)][\nabla f(x^*, \xi)]^T\right) = \mathbb{E}(\epsilon^2)\mathbb{E}\left(aa^T\right) = \mathbf{I}_d.$$



(a) d=1

(b) d=5

Figure 4.1: Linear regression: Log loss (operator norm) of the estimated covariance matrix against the log of total number of steps. Here F denotes the full overlapping version (4.3), NOL denotes the non-overlapping version (4.8), and C denotes the constant in $a_m = \left\lfloor Cm^{2/(1-\alpha)} \right\rfloor$.

Then the limiting covariance matrix

$$\Sigma = A^{-1} S A^{-1} = \mathbf{I}_d.$$

We check the convergence of our proposed online estimators, both the full overlapping and the non-overlapping versions, by computing the operator norm loss of the covariance matrix estimate, i.e., $\|\hat{\Sigma}_n - \Sigma\|_2$. Figure 4.1 shows that the log loss of the online estimators are approximately linear with the log number of steps and the slopes are about $-1/8$ for the large total number of steps. It suggests that both the full overlapping and the non-overlapping versions converge to the limiting covariance matrix with the same convergence rate, about $O(n^{-1/8})$. We also compute the relative efficiency (MSE of the full overlapping version (4.3) divided by MSE of the non-overlapping version (4.8)); see Figure 4.2. Their performances are comparable. Also, the performance is relatively insensitive to the choice of C in $a_m = \left\lfloor Cm^{2/(1-\alpha)} \right\rfloor$. Therefore, we will implement the non-overlapping version and set $C = 1$ in the subsequent simulations without any specification.

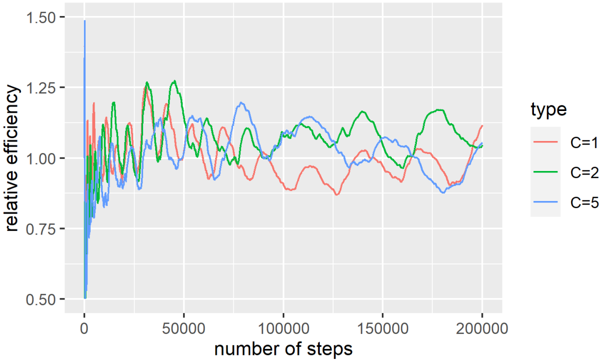


Figure 4.2: Relative efficiency (ratio of MSE) of the full overlapping version (4.3) and non-overlapping version (4.8). We set $d = 5$ in linear regression. Here C denotes the constant in $a_m = \left\lfloor Cm^{2/(1-\alpha)} \right\rfloor$.

Asymptotic normality and CI coverage. With the covariance matrix estimates, we construct 95% confidence intervals for the averaged coefficient $\mu = \mathbf{1}_d^T x^*$ according to (4.25), i.e.,

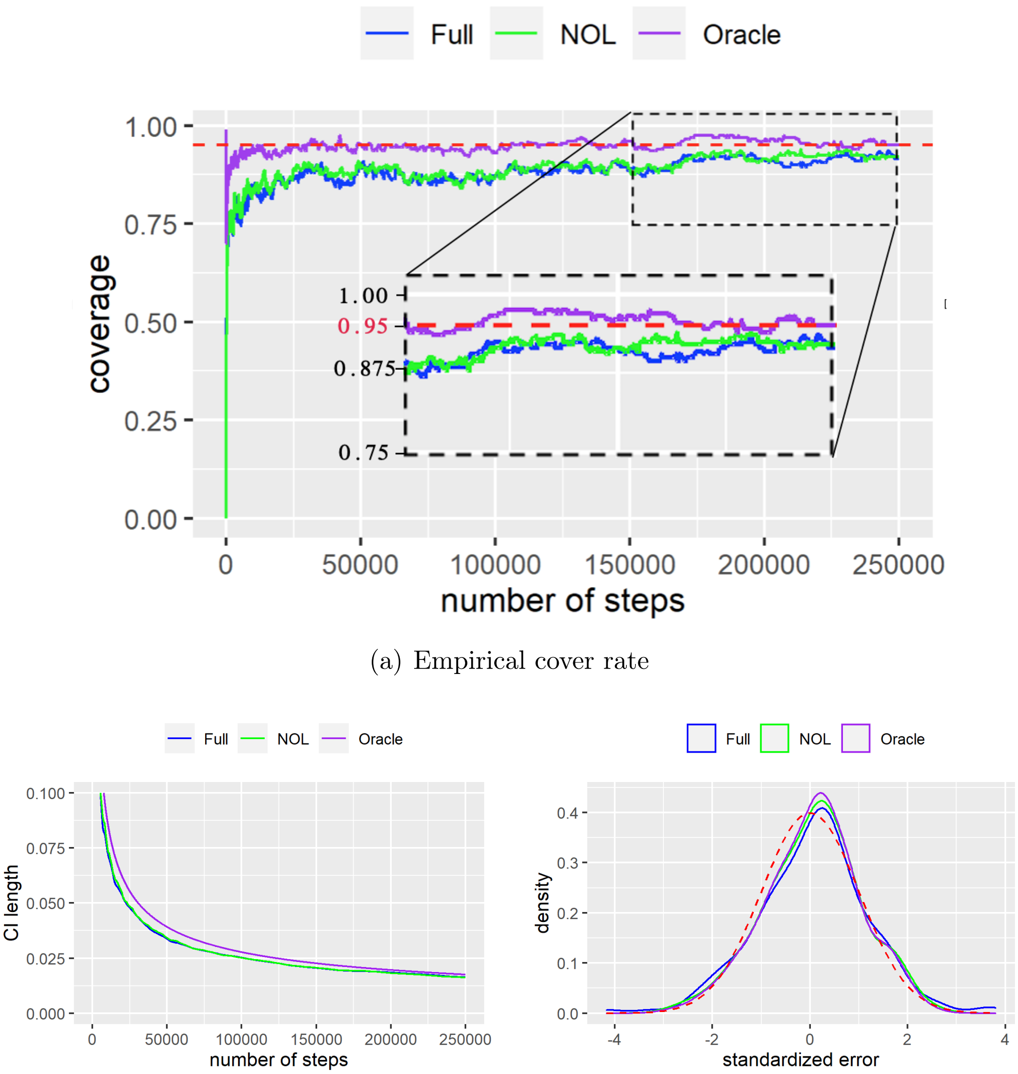
$$\left[1^T \bar{x}_n - z_{1-q/2}\sqrt{\mathbf{1}_d^T \hat{\Sigma}_n \mathbf{1}_d / n}, 1^T \bar{x}_n + z_{1-q/2}\sqrt{\mathbf{1}_d^T \hat{\Sigma}_n \mathbf{1}_d / n} \right].$$

We also compute the oracle 95% confidence intervals based on the true limiting covariance matrix. Figure 4.3 shows that for both overlapping and non-overlapping versions, the empirical coverage rate converges to 95%, and the standardized error $\sqrt{n}\mathbf{1}_d^T(\hat{x} - x^*)/\sqrt{\mathbf{1}_d^T \hat{\Sigma}_n \mathbf{1}_d}$ is approximately standard normal. Also, the estimated CI length converges to the oracle length.

4.4.2 Comparison with other methods

In this section, we compare the performance of the proposed online estimator, which we refer to as online-BM in the subsequent numerical experiments, with other estimators for marginal inference of each individual regression coefficient. We consider both linear and logistic regression examples. The nominal coverage probability is set to 95%.

We first compare the empirical coverage rates of the proposed estimator with the plug-in



(a) Empirical cover rate

(b) CI length

(c) Normality

Figure 4.3: Linear regression with $d = 5$: (a): Empirical coverage rate against the number of steps. Red dashed line denotes the nominal coverage rate of 0.95. (b): Length of confidence intervals. (c): Density plot for the standardized error. Red curve denotes the standard normal density.

estimator in Chen et al. [2020]. As we mentioned in the introduction, the plug-in estimator requires the computation of the Hessian matrix (of the loss function) and its inverse. Figure 4.4 shows that our online estimator (online-BM) has a comparable performance as the plug-in estimator when the number of iterations is large enough. Although the online-BM has a

Table 4.1: Empirical coverage rates: the average coverage rate for the nominal coverage probability 95%. Standard errors are reported in the brackets.

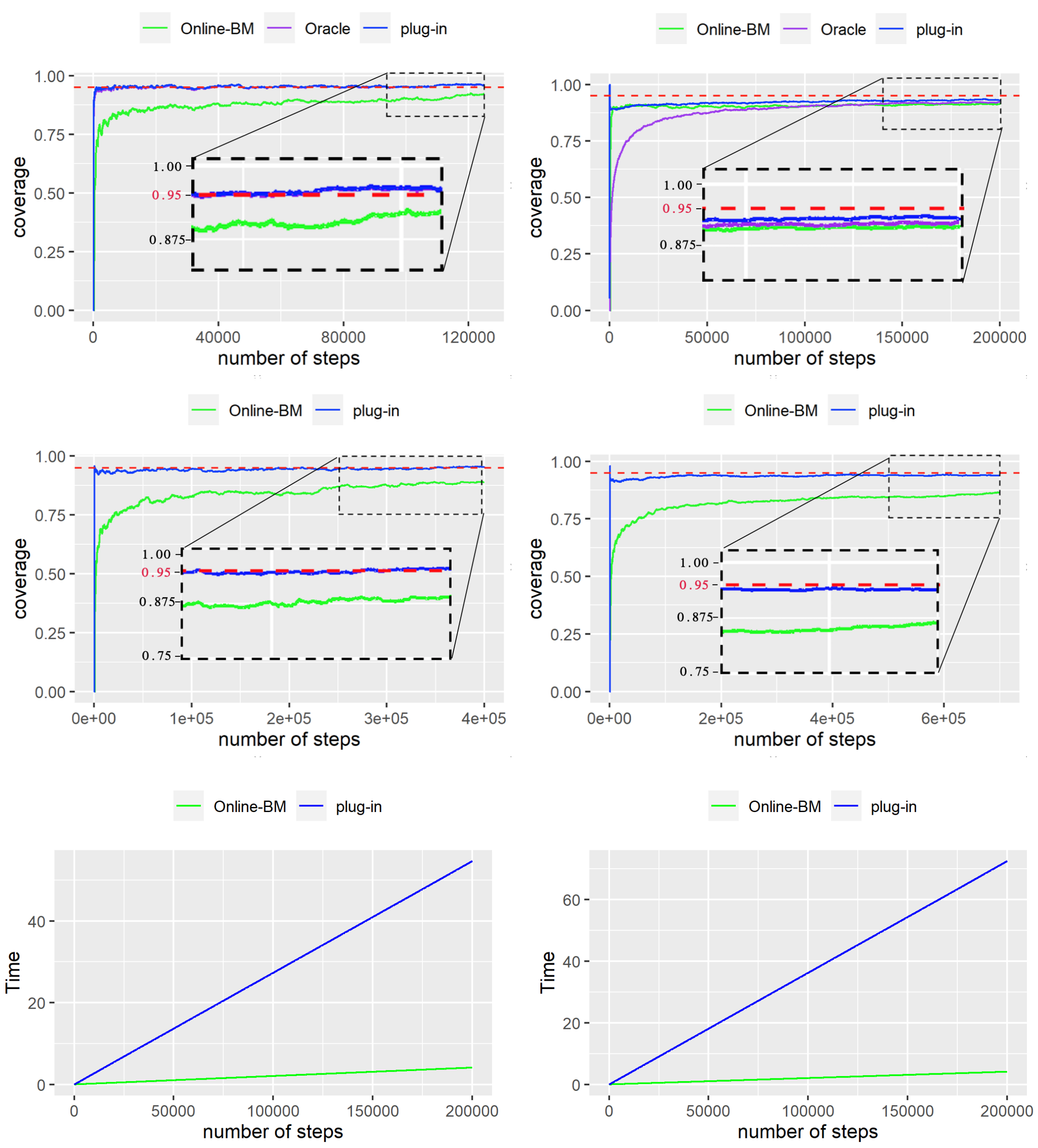
linear model				
$(d = 5)$	$n = 50000$	$n = 80000$	$n = 100000$	n= 125000
online-BM	0.894 (0.02177)	0.901 (0.02114)	0.917 (0.01951)	0.935 (0.01746)
BM	0.894 (0.02177)	0.904 (0.02085)	0.910 (0.02022)	0.928 (0.01831)
$(d = 20)$	n= 50000	n= 100000	n= 150000	n= 200000
online-BM	0.904 (0.02078)	0.907 (0.02050)	0.910 (0.02022)	0.914 (0.01986)
BM	0.878 (0.02312)	0.901 (0.02121)	0.908 (0.02043)	0.910 (0.02029)
logistic model				
$(d = 5)$	$n = 100000$	$n = 200000$	$n = 300000$	$n = 400000$
online-BM	0.828 (0.01011)	0.844 (0.00933)	0.875 (0.00770)	0.889 (0.00700)
BM	0.822 (0.01032)	0.847 (0.00919)	0.875 (0.00771)	0.885 (0.00721)
$(d = 20)$	$n = 100000$	$n = 300000$	$n = 500000$	$n = 700000$
online-BM	0.791 (0.01167)	0.829 (0.01004)	0.845 (0.00926)	0.864 (0.00834)
BM	0.787 (0.01188)	0.827 (0.01011)	0.839 (0.00955)	0.859 (0.00856)

slower convergence rate, it has an advantage in computational efficiency since it only uses the iterates from SGD. The online-BM is more desirable for practitioners when the computation is limited or only stochastic gradient information is available.

Next, we compare the finite sample coverage rate of the proposed online-BM estimator and the batch means covariance matrix estimator from Chen et al. [2020], which we refer to as BM. Table 4.1 shows that the finite sample coverage rates of the two estimators are close to each other in all cases, and the finite sample performance of our method slightly outperforms Chen et al. [2020] when n is large. In fact, this is not a totally fair comparison for us since we implement the method in Chen et al. [2020] based on the prior knowledge of the exact sample size.

4.5 Summary

We propose a fully online approach to estimate the asymptotic covariance matrix of the ASGD solution and conduct statistical inference. The fully online fashion allows efficient sequentially updating. It is important for online learning, where data comes in a stream and



(a) $d = 5$

(b) $d = 20$

Figure 4.4: Comparison of online-BM and Plug-in estimators. First/Middle row: Empirical coverage rate against the number of steps in linear/logistic model. Red dashed line denotes the nominal coverage rate of 0.95. Third row: total computation time for updating covariance estimate and confidence intervals in SGD.

real-time update of predictions is needed before seeing future data. Our method is efficient in both computation and memory. In particular, the computational and memory complexity at the update step is $O(d^2)$, and the total computational cost only scales linearly in n. In terms of theoretical merits, the proposed estimator is the first fully online fashion estimator with rigorous convergence property for asymptotic covariance of ASGD. We show that the convergence rate of our online estimator is comparable to the offline counterparts.

CHAPTER 5

HIGH CONFIDENCE LEVEL INFERENCE IS ALMOST FREE USING PARALLEL STOCHASTIC OPTIMIZATION

This chapter will continue with practical inference. As observed in the previous chapter, the convergence of the covariance matrix estimator is slow, as is the coverage of the confidence interval. Given the complexities inherent in estimating the covariance matrix, one may wonder if, for certain tasks, we can bypass the need for the covariance matrix and focus solely on the confidence interval to achieve better results. By “better”, we mean improved coverage of the confidence interval and enhanced computational efficiency, without the need for matrix updating at each iteration. This motivates our work in this chapter.

5.1 Introduction

We still consider the problem in (1.1). With streaming data $\{\xi_i\}_{i\geq 1}$, and assuming we obtain iterates/outputs of a stochastic approximation (SA) algorithm, the primary goal of this chapter is to enhance statistical inference by constructing confidence intervals based on these iterates in an online setting. Specifically, for a given vector $v \in \mathbb{R}^d$, we aim to construct a valid $(1-\alpha)\times 100\%$ confidence interval $\hat{\text{CI}}$ for the linear functional $v^\top x^*$, that is

$$\mathbb{P}(v^\top x^* \in \hat{\text{CI}}) - (1-\alpha) = \alpha - \mathbb{P}(v^\top x^* \notin \hat{\text{CI}}) \approx 0, \tag{5.1}$$

where $\alpha \in (0,1)$. To fit in an online setting, the proposed confidence interval can be updated recursively as new data becomes available. It utilize only previous SA iterates, requiring minimal extra computation for the inference purpose beyond the original computation, thus allowing for easy integration into existing codebases.

We consider a *high level of confidence*, i.e., $\alpha \approx 0$, as uncertainty quantification is particularly important in applications involving high-stakes decisions, where a nearly 100% confidence interval is required. Moreover, with datasets growing increasingly large, the demand for higher-level confidence intervals becomes more prevalent. Additionally, in applications involving multiple simultaneous tests, such as high-dimensional parameter analysis, correction techniques like the Bonferroni method are employed. This leads to each individual test maintaining a sufficiently high confidence level (related to dimension). In such cases, the guarantee in (5.1) may not be sufficient. In particular, we shall construct confidence intervals $\hat{\mathrm{CI}}$ such that the relative error

$$\Delta_\alpha := \left| \frac{\mathbb{P}(v^\top x^* \in \hat{\mathrm{CI}}) - (1-\alpha)}{\alpha} \right| = \left| \frac{\mathbb{P}(v^\top x^* \notin \hat{\mathrm{CI}})}{\alpha} - 1 \right| \tag{5.2}$$

is small. Note that (5.2) offers a much more refined assessment than (5.1). For example, if $\alpha = 10^{-4}$ and $\mathbb{P}(v^\top x^* \notin \hat{\mathrm{CI}}) = 10^{-3}$, then (5.1) is not severely violated, while Δ_α in (5.2) is very different from 0. In this context, it is crucial to recognize that even a slight undercoverage can be significant due to low tolerance for error. Conversely, an extremely wide confidence interval that nearly always covers can become uninformative, underscoring the importance of precision in interval construction. Hence, employing a method that provides confidence intervals with rapid convergence to the desired coverage level is essential. We derive the upper bound (with explicit rate) of the relative error of coverage for the constructed confidence intervals and explicitly detail the dependence on α in the upper bound. The results indicate that our method remains valid even when α is potentially very small or decreases with the total sample size or the number of hypotheses.

5.1.1 Background: existing confidence interval construction

Practical inference methods are based on the limiting distribution of SA solutions. Recall that the vanilla SGD iterates with the recursion form:

$$x_i = x_{i-1} - \eta_i \nabla f(x_{i-1}, \xi_i), \quad i = 1, 2, \dots,$$

where $\nabla f(x, \xi)$ is the gradient vector of $f(x, \xi)$ with respect to the first variable x, and η_i is the step size at the i-th step. Consider the average of all past iterations $\bar{x}_n = n^{-1} \sum_{i=1}^{n} x_i$. In the celebrated work of Polyak and Juditsky [1992], it is shown that, under suitable conditions, the averaged SGD (ASGD) exhibits asymptotic normality, that is,

$$\sqrt{n}(\bar{x}_n - x^*) \Rightarrow \mathcal{N}(0, \Sigma), \tag{5.3}$$

where $\Sigma = A^{-1} S A^{-1}$ is the sandwich form covariance matrix with $A = \nabla^2 F(x^*)$ and $S = \mathbb{E}\left([\nabla f(x^*, \xi)][\nabla f(x^*, \xi)]^T\right)$. Similar asymptotic normality results have been established for other variants of SGD with adjusted asymptotic covariance matrices [Li et al., 2022a, Wei et al., 2023, Na and Mahoney, 2022]. These asymptotic normality results form the foundation of statistical inference in an online setting. As the limiting covariance matrix is unknown in practice, to perform practical inference, there are three primary methods for constructing confidence intervals.

- The first method relies on recursively estimating the asymptotic covariance matrix Σ. Chen et al. [2020] proposes the plug-in method to estimate A and S separately using sample averages and then applying them in the sandwich form. Zhu et al. [2023] proposed the online batch-means method, which only utilizes SGD iterates and is more computationally efficient. Both methods provide consistent estimators for the asymptotic covariance matrix Σ of ASGD solutions. With a consistent covariance estimate $\hat{\Sigma}_n$, one can construct confidence intervals for $\upsilon^\top x^*$ as

$$\hat{\mathrm{CI}}_{n,\mathrm{cov}} = \left[\upsilon^\top \bar{x}_n - z_{1-\alpha/2}\sqrt{\frac{\upsilon^\top \hat{\Sigma}_n \upsilon}{n}},\ \upsilon^\top \bar{x}_n + z_{1-\alpha/2}\sqrt{\frac{\upsilon^\top \hat{\Sigma}_n \upsilon}{n}}\right],$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2) \times 100\%$ percentile of the standard normal distribution.

- The second method takes advantage of statistical pivotal statistics. One example is the random scaling method. Instead of consistently estimating the asymptotic covariance matrix, Lee et al. [2022] leverages the asymptotic normality result by constructing asymptotic pivotal statistics after self-normalization. Specifically, they studentize $\sqrt{n}(\bar{x}_n - x^*)$ via the random scaling matrix

$$\hat{V}_{rs,n} = \frac{1}{n}\sum_{s=1}^{n}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{s}(x_i - \bar{x}_n)\right\}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{s}(x_i - \bar{x}_n)\right\}^T.$$

The resulting statistic is asymptotically pivotal and the confidence interval for $\upsilon^\top x^*$ is then constructed as

$$\hat{\mathrm{CI}}_{n,\mathrm{rs}} = \left[\upsilon^\top \bar{x}_n - q_{rs,1-\alpha/2}\sqrt{\frac{\upsilon^\top \hat{V}_{rs,n} \upsilon}{n}},\ \upsilon^\top \bar{x}_n + q_{rs,1-\alpha/2}\sqrt{\frac{\upsilon^\top \hat{V}_{rs,n} \upsilon}{n}}\right], \tag{5.4}$$

where $q_{rs,1-\alpha/2}$ is the $(1-\alpha/2)\times 100\%$ percentile for $W_1(1)/[\int_0^1\{W_1(r)-rW_1(1)\}^2 dr]^{1/2}$ with $W_1(r)$ stands for a standard Brownian motion.

- An alternative method for inference is via bootstrap. One can apply bootstrap perturbations and modify the original SGD path. Then, the asymptotic distribution of the online estimate as well as other quantities such as variance or quantiles can be estimated using a large number of bootstrapped sequences [Fang et al., 2018, Li et al., 2018, Su and Zhu, 2023].

Same ideas have also been applied for inference when using different algorithms or dealing with online decision-making problems [Luo et al., 2022, Li et al., 2022b, Chen et al., 2021, Ramprasad et al., 2022, Su and Zhu, 2023]. Note that all the three methods above have their advantages and applicable use cases. The first and third methods can provide consistent estimators of the limiting covariance matrix. However, the cost of using bootstrap (the third method) involves heavy computation or complicated modification to existing code base, and we will not consider this method. The online covariance matrix estimation (the second method) is a difficult task in SGD settings. The plug-in estimator requires Hessian information which is typically unavailable, and involves matrix computation that requires an $O(d^3)$ computational cost, which is not desirable for large dimensions. The online batch-means methods do not require extra information such as the Hessian and are computationally efficient, but they come at the cost of slow convergence. The random scaling method does not provide a consistent covariance matrix estimator, yet in terms of confidence interval construction, it is computationally comparable to the online batch-means method while offering better coverage. However, the critical values of the self-normalized statistics are not easy to obtain for arbitrary α, we simulate the value via MCMC in Section 5.4.

In terms of theoretical guarantees, although all three methods demonstrate asymptotically valid coverage of confidence intervals, the convergence guarantee without a specific rate and without explicit dependence on α is not sufficient, as demonstrated above and in Section 5.3. This limitation is not significant at moderate confidence levels but becomes substantial at higher levels, leading to unstable coverage. This may result in either undercoverage (failing to meet the standard) or overcoverage (producing an excessively wide confidence interval, thereby diminishing the interval's meaningfulness). A detailed comparison and discussion of these methods can be found in Lee et al. [2022]. We also make a brief summary in Table 5.1 comparing the above online inference methods.

We propose a new method that utilizes a small number of parallel runs to effectively

Method	Plug-In	Online BM	Random Scale	This paper
consistent covariance estimator?	✓	✓	/	/
to avoid Hessian?	/	✓	✓	✓
CI coverage convergence rate?	/	/	/	✓
empirical CI coverage	⋆⋆⋆	⋆	⋆⋆	⋆⋆⋆
computation time	⋆	⋆⋆	⋆⋆	⋆⋆⋆

Table 5.1: Comparison of methods for online statistical inference: This table compares various methods including Plug-In [Chen et al., 2020], Online BM [Zhu et al., 2023], and Random Scale [Lee et al., 2022]. The symbol '✓' indicates that the method can achieve the goal, while '/' signifies that it cannot. The rating symbols '⋆⋆⋆', '⋆⋆', and '⋆' denote the best, moderate, and lowest advantage, respectively.

acquire information about the distribution, while maintaining a fully online status. We demonstrate that the confidence intervals constructed via the parallel inference method provide asymptotically exact coverage with more rigorous theoretical guarantees than existing methods, featuring an explicit convergence rate of the relative error of coverage. Additionally, this method offers better coverage compared to other methods, as demonstrated in Section 5.4. It is also the most computationally efficient among all considered inference methods. Our approach avoids the heavy cost of resampling. Unlike methods based on covariance matrix estimation or the random scaling method, our method does not require updating a $d \times d$ matrix at each iteration. Additional computation or memory for inference beyond running SGD is required only when necessary at specific steps and is minimal, making the inference almost free. Another advantage of our method is its suitability for settings where parallel computing is needed, which can further accelerate computation. This is particularly relevant in scenarios such as processing extremely large and high-frequency datasets, or in federated learning scenarios where data are distributed across different clients [Zinkevich et al., 2010, Dean et al., 2012, Li et al., 2020b, Karimireddy et al., 2020, McMahan et al., 2017, Ghosh et al., 2020]. In our work, the requirement for parallel processing is seen not as a burden but as a beneficial tool.

5.2 Inference with parallel runs of stochastic algorithms

In this section, we introduce the parallel run inference method for constructing confidence intervals. The method involves K parallel runs of a predetermined stochastic algorithm, calculating the sample variance of the linear functional of interest from K parallel runs, and self-normalizing to obtain asymptotic pivotal t-statistics and the corresponding confidence interval.

5.2.1 Parallel computing

Consider a general stochastic algorithm characterized by the update rule h_i at the i-th step and K parallel run sequences. For each of the k-th sequence where $k = 1, \ldots, K$, we begin with a random initialization $\hat{x}_0^{(k)}$. The estimate for the k-th machine at the i-th iterate is denoted by $\hat{x}_i^{(k)}$. The recursive update is given by

$$\hat{x}_i^{(k)} = h_i(\xi_i^{(k)}, \mathcal{F}_{i-1}^{(k)}), \quad i = 1, 2, \ldots, \tag{5.5}$$

where $\mathcal{F}_{i-1}^{(k)} = \sigma(\xi_{i-1}^{(k)}, \xi_{i-2}^{(k)}, \ldots)$ encapsulates information from the previous step, such as $\hat{x}_{i-1}^{(k)}$ or other intermediate estimates according to the algorithm. For example, in the case of ASGD, we have

$$\begin{cases} x_i^{(k)} = x_{i-1}^{(k)} - \eta_i \nabla f(x_{i-1}^{(k)}, \xi_i^{(k)}), \\ \hat{x}_i^{(k)} = \{(i-1)\hat{x}_{i-1}^{(k)} + \hat{x}_i^{(k)}\}/i, \end{cases} \tag{5.6}$$

where $\nabla f(x_{i-1}^{(k)}, \xi_i^{(k)})$ is the derivative of the objective function f with respect to the first variable, and step size is usually chosen as $\eta_i = \eta \times i^{-\beta}$ for some $\beta \in (0.5, 1]$. If we seek output for estimation or inference at the n-th step (with $N = nK$ total samples), we can aggregate the results from the K machines by averaging the estimates or predictions. Specifically, we define the parallel average of the K sequences as

$$\bar{x}_{K,n} = \frac{1}{K} \sum_{k=1}^{K} \hat{x}_n^{(k)}. \tag{5.7}$$

In a practical online setting, sequential data $\{\xi_i\}_{i=1,2,...}$ can be distributed across K different machines, with $\xi_i^{(k)} = \xi_{k+K(i-1)}$. Alternatively, in an offline setting, the dataset can be randomly divided into K batches. Note that when the initialization $\hat{x}_0^{(k)}$ is the same for all $k = 1, \ldots, K$, the output from K sequences $\hat{x}_i^{(k)}$, $k = 1, \ldots, K$, will be independent and identically distributed (i.i.d.), given that the data components $\xi_i^{(k)}$ are i.i.d..

Note that, unlike local/federated SGD as discussed in Yu et al. [2019], Li et al. [2022b], Woodworth et al. [2020], we do not communicate local solutions/gradients to obtain a common averaged iterate for parallel runs at intermediate iterations. Our method is more akin to model averaging, often referred to as one-shot averaging [Zinkevich et al., 2010]. On one hand, model averaging without communication costs can still achieve good convergence when K is small or moderate. On the other hand, it ensures the K sequences are i.i.d. and enables us to construct a asymptotically pivotal t-statistic and demonstrate strong convergence in a later section. Additionally, the straightforward parallel running and model averaging make it easier to apply to different stochastic algorithms. In some cases, local updates with more frequent periodic averaging would improve statistical efficiency of the algorithm and communication costs may not be a problem. The inference procedure may still hold with refined proof. However, discussing the difference between vanilla SGD, parallel SGD and local SGD is beyond the scope of this thesis.

5.2.2 Asymptotic t-distribution

In this context, various stochastic approximation algorithms can be employed to run parallel sequences. To derive a valid t-distribution, it is essential to consider cases where asymptotic normality is applicable for the estimate $\hat{x}_n^{(k)}$ in each sequence. Specifically, for each $k = 1, \ldots, K$,

$$\sqrt{n}(\hat{x}_n^{(k)} - x^*) \Rightarrow \mathcal{N}(0, \Sigma), \text{ as } n \to \infty.$$

In the case of ASGD as denoted in (5.6), the celebrate work of Polyak and Juditsky Polyak and Juditsky [1992] demonstrated the asymptotic normality with the sandwich form Σ as mentioned before. Other algorithms, such as various versions of weighted-averaged SGD [Wei et al., 2023], Root-SGD [Li et al., 2022a], and StoSQP [Na and Mahoney, 2022], have also been shown to possess this asymptotic normality property, albeit with adjusted limiting covariance matrices.

For any vector $v \in \mathbb{R}^d$, considering inference for the linear functional $v^\top x^*$ at the n-th iteration (with $N = nK$ total samples), define the sample variance $\hat{\sigma}_v^2$ as

$$\hat{\sigma}_v^2 = \frac{1}{K-1} \sum_{k=1}^{K} (v^\top \hat{x}_n^{(k)} - v^\top \bar{x}_{K,n})^2,$$

where $\bar{x}_{K,n}$ is the sample average defined in (5.7). It is worth noting that $\hat{\sigma}_v^2$ is not a consistent estimator for the variance of $v^\top x^*$. However, we can studentize $\sqrt{K}(v^\top \bar{x}_{K,n} - v^\top x^*)$ with $\hat{\sigma}_v$ to obtain a t-statistic which is asymptotically pivotal. Assuming the validity of the asymptotic normality result, together with the i.i.d. property of $\{\hat{x}_n^{(k)}\}_{k=1,\ldots,K}$, we can infer a t-type distribution, that is,

$$\hat{t}_v := \frac{\sqrt{K}(v^\top \bar{x}_{K,n} - v^\top x^*)}{\hat{\sigma}_v} \Rightarrow t_{K-1}. \tag{5.8}$$

Based on (5.8), we can construct a $(1-\alpha) \times 100\%$ confidence interval for $v^\top x^*$ as follows,

$$\hat{\text{CI}}_v = \left[v^\top \bar{x}_{K,n} - \frac{t_{1-\alpha/2,K-1} \hat{\sigma}_v}{\sqrt{K}},\ v^\top \bar{x}_{K,n} + \frac{t_{1-\alpha/2,K-1} \hat{\sigma}_v}{\sqrt{K}} \right], \tag{5.9}$$

where $t_{1-\alpha/2,K-1}$ is the $(1-\alpha/2)\times 100\%$ percentile for the t_{K-1} distribution. The proposed confidence interval in (5.9) is fairly easy and efficient to construct. In particular, when $K = 2$, $\hat{t}_v$ is asymptotically distributed as the standard Cauchy distribution with upper quantile $t_{1-\alpha/2,1} = \tan((1-\alpha)\pi/2)$. The entire procedure is summarized in Algorithm 4.

Remark 5.2.1 (Almost cost-free)**.** *We observe that the inference step in our method can be performed whenever necessary with minimal calculation and memory requirements, without needing any modifications to existing stochastic algorithms. This makes it almost cost-free and can be easily integrated into existing codebases. In contrast, all other methods typically demand considerable extra effort for inference. This may involve complex modifications, as seen in Su and Zhu [2023], or entail storing and updating a matrix at each iteration, as required by covariance-matrix-estimation-based methods or the random scaling method. In these cases, the computing and memory costs for inference purposes usually far exceed those involved in the SGD update itself.*

Remark 5.2.2 (Choice of K)**.** *A larger K brings more stable and usually shorter (in expectation) confidence intervals, as indicated in* (5.9)*. And a smaller K will decrease the effect of sample splitting and lead to better convergence of the estimate in a single run. As demonstrated in Section 5.4, the performance of inference is not that sensitive to the choice of K when K is in a reasonable range, and $K = 6$ is a good choice in practice. Users can opt for a smaller K when dealing with a moderate dataset to ensure sufficient sample size and faster convergence in a single trial. If parallel computing resources are available and preferred, especially when dealing with very large datasets or high data acquisition rates, users are encouraged to utilize more machines.*

Algorithm 4: Online Parallel Inference

```
Input: stochastic algorithm $h$, number of parallel runs $K$
for $i = 1, 2, \dots$ do
    for $k = 1, \dots, K$ do
        Update $\hat{x}_i^{(k)} = h_i(\xi_i^{(k)}, \mathcal{F}_{i-1}^{(k)})$ ($\xi_i^{(k)}$ is data received);
    end
    Output if necessary:
    $\bar{x}_{K,i} \leftarrow \frac{1}{K}\sum_{k=1}^{K} \hat{x}_{k,i}$;
    $\hat{\sigma}_\upsilon^2 \leftarrow \frac{1}{K-1}\sum_{k=1}^{K} \left(\upsilon^\top \hat{x}_{k,i} - \upsilon^\top \bar{x}_{K,i}\right)^2$;
    $\hat{\mathrm{CI}}_\upsilon \leftarrow \left[\upsilon^\top \bar{x}_{K,i} - \hat{\sigma}_\upsilon t_{1-\alpha/2,K-1}/\sqrt{K},\ \upsilon^\top \bar{x}_{K,i} + \hat{\sigma}_\upsilon t_{1-\alpha/2,K-1}/\sqrt{K}\right]$
end
```

5.3 Theoretical guarantee

In this section, we provide a theoretical foundation for the confidence interval (5.9) constructed using the t-distribution. Recall that we consider a *high level of confidence* where the noncoverage level α can be potentially very small or decrease with the total sample size (or dimension). This level of validation requires a more stringent guarantee than just showing that

$$\mathbb{P}(\upsilon^\top x^* \in \hat{\mathrm{CI}}) - (1-\alpha) \to 0, \tag{5.10}$$

which can be derived from the convergence of relevant statistics in distribution as shown in other works. Our focus is to establish the bound of the relative error of coverage

$$\Delta_N := \sup_{\alpha(N) \le \alpha < 1} \left| \frac{\mathbb{P}(\upsilon^\top x^* \in \hat{\mathrm{CI}}) - (1-\alpha)}{\alpha} \right|,$$

where $\alpha(N)$ goes to zero at an appropriate rate. Compared with (5.10), this bound offers a more rigorous assessment. It is critical in cases where we require high precision in our confidence assessments, ensuring that the constructed interval genuinely reflects the desired confidence level. For example, suppose we use Bonferroni method to construct simultaneous confidence intervals for m parameters at overall level 0.95 with large m, then the CI for each individual parameter should be at level $1 - 0.05/m$. In this case $\alpha = 0.05/m$ and a small Δ_N is needed, while (5.10) is not sufficient. Also, it is important to make the dependence on level α explicit since we may consider a decreasing α.

To derive the upper bound of the relative error, it is important to obtain the rate of convergence of the t-statistic. In the rest of this section, we will first explore the application of ASGD in each parallel run and then extend our results to a broader class of stochastic algorithms that meet certain mild assumptions.

5.3.1 Convergence characterization for ASGD

Among various stochastic approximation algorithms, SGD is notably convenient and popular. Its variant, ASGD, is also widely used and has been the subject of extensive study. Beyond the well-known asymptotic normality results related to convergence in distribution, the rate of convergence to normality is of growing interest and has been studied in the literature. Notably, Anastasiou et al. [2019b] derived the non-asymptotic rate of convergence to normal using non-asymptotic rates of the martingale Central Limit Theorem (CLT), and Shao and Zhang [2022] established a Berry–Esseen type bound for the Kolmogorov distance between the cumulative distribution functions of the ASGD estimator and its Gaussian analogue.

In this subsection, to better characterize the distributional approximation of the ASGD estimator, we develop a new Gaussian approximation of which the asymptotic normality is a direct consequence. Before presenting the main approximation result, we first introduce some regularity assumptions on the objective function and basic definitions.

Assumption 5.3.1. *There exist positive constants τ and L such that*

$$(x - x')^\top (\nabla F(x) - \nabla F(x')) \geq \tau \|x - x'\|_2^2,$$
$$\|\nabla F(x) - \nabla F(x')\|_2 \leq L \|x - x'\|_2.$$

Assumption 5.3.2. *Denote $\Delta(x, \xi) = \nabla F(x) - \nabla f(x, \xi)$ for $x \in \mathbb{R}^d$ and $\xi \sim \Pi$. Given $q > 4$, we have $\mathbb{E}_\xi \|\Delta(x^*, \xi)\|_2^q < \infty$ and there exists some positive constant γ such that for any $x, x' \in \mathbb{R}^d$,*

$$\big(\mathbb{E}_\xi \|\Delta(x, \xi) - \Delta(x', \xi)\|_2^q\big)^{1/q} \leq \gamma \|x - x'\|_2.$$

Assumption 5.3.3. *There exists some positive constant $\mathcal{L}$ such that for $x \in \mathbb{R}^d$,*

$$\|\nabla F(x) - \nabla^2 F(x^*)(x - x^*)\|_2 \leq \mathcal{L} \|x - x^*\|_2^2.$$

Assumptions 5.3.1–5.3.3 are common and fairly mild in the context of convex optimization based on the SGD algorithm and its variants [Chen et al., 2020, Zhu et al., 2023]. For $n \geq 1$, we define

$$\Gamma_n = \frac{1}{n} \sum_{k=1}^n U_k S U_k^\top, \text{ where } U_k = \sum_{i=k}^n Y_k^i \eta_k \tag{5.11}$$

with $Y_k^k = \mathbf{I}_d, Y_k^i = \prod_{l=k+1}^i (\mathbf{I}_d - \eta_l \nabla^2 F(x^*)), i > k$. In the following theorem, we establish a Gaussian approximation result for the ASGD estimator.

Theorem 5.3.4. *Assume that $\{x_i\}_{i=1}^n$ is a SGD sequence defined by:*

$$x_i = x_{i-1} - \eta_i \nabla f(x_{i-1}, \xi_i), \quad i = 1, 2, \ldots,$$

where $\eta_i = \eta \times i^{-\beta}$ for some constant $\beta \in (1/2, 1)$. Let $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$. Under Assumptions 5.3.1–5.3.3, on a sufficiently rich probability space, there exist a random vector

$W_n \overset{\mathcal{D}}{=} \sqrt{n}(\bar{x}_n - x^*)$ *and a centered Gaussian random vector* $Z_n \sim \mathcal{N}(0, \Gamma_n)$*, where* Γ_n *is defined in* (5.11)*, such that*

$$\mathbb{E}\|W_n - Z_n\|_2^2 \lesssim \max\left(n^{1-2\beta}, \frac{\log n}{n^{1-2/q}}, \frac{\|x_0 - x^*\|_2^2}{n}\right). \tag{5.12}$$

Remark 5.3.5. *Theorem 5.3.4 reveals that the ASGD estimator can be approximated by a centered Gaussian random vector with approximately the same covariance matrix without imposing additional structural and moment assumptions and the approximation error is asymptotically negligible as long as* $\beta > 1/2$ *and* $\|x_0 - x^*\|_2 \ll n^{1/2}$*. To the best of our knowledge, this is the first Gaussian approximation result for online estimators based on the SGD algorithm. It is worth noting the SGD iterates* $x_i \in \mathbb{R}^d, i = 1, 2, \ldots,$ *are neither independent nor stationary. Hence the existing strong invariance principle results for the partial sums of independent random elements [Komlós et al., 1975, 1976, Csörgő and Révész, 1975, Einmahl, 1987] or general stationary sequences [Wu, 2007, Liu and Lin, 2009, Berkes et al., 2014] are not applicable here. To handle the nonstationary property of the sequence* $\{x_i\}_{i\geq 1}$*, we shall invoke the recently established strong approximation result for non-stationary time series [Mies and Steland, 2023]. The detailed proof of Theorem 5.3.4 is given in the Appendix.*

Remark 5.3.6. *It is worth noting the covariance matrix* Γ_n *defined in* (5.11) *and hence the distribution of the coupled Gaussian vector* Z_n *do not depend on the initial estimate* x_0*. Therefore, our result in* (5.12) *can also be viewed as a quenched Gaussian approximation in the sense that the impact of the initial point* x_0 *diminishes, as asserted by the third term* $\|x_0 - x^*\|_2^2/n$ *in the upper bound (5.12). As a direct consequence, the distribution of the random vector* $\sqrt{n}(\bar{x}_n - x^*)$ *can be approximated by that of* $\mathcal{N}(0, \Gamma_n)$*. Moreover, the multivariate central limit theorem* (5.3) *can be easily derived as* Γ_n *converges to the sandwich form covariance matrix* $\Sigma = A^{-1}SA^{-1}$*; see, for example Polyak and Juditsky [1992]. It is* *important to mention that our procedure does not rely on the convergence of Γ_n to Σ, which can be slow and introduce additional approximation error in practical implementations. Our constructed t-statistic is asymptotically pivotal as long as* (5.12) *holds. Particularly, the simulation studies in Section 5.4 demonstrate that our procedure has better finite-sample performance than the oracle procedure based on the multivariate central limit theorem* (5.3) *with the population covariance matrix Σ given.*

The rate of convergence to normality plays a crucial role in assessing the approximation of the t-distribution. As we will discuss in a later section through a general theorem, the convergence of the t-statistic and the upper bound of the relative error relies on the convergence rate (of a single parallel run sequence) to normality.

5.3.2 Main results

As discussed above, there are many variants of SGD exhibiting asymptotic normality results, and the specific convergence rate to normality may vary across different algorithms. We will not study those rates for other algorithms rigorously since they are beyond the scope of this paper. To derive general results with the potential to be applied to different algorithms, we propose the following assumption.

Assumption 5.3.7 (Convergence rate to normality)**.** *For a chosen stochastic algorithm and number of parallel runs K, let $\hat{x}_n^{(k)} (k = 1, ..., K)$ denote the result at the n-th iteration of the k-th parallel run used in calculating the parallel average $\bar{x}_{K,n}$ in* (5.7)*. There exists a centered Gaussian random vector $Z_n \sim \mathcal{N}(0, \Sigma_n)$ (for some $\Sigma_n \in \mathbb{R}^{d\times d}$) such that*

$$\left(\mathbb{E}\|\sqrt{n}(\hat{x}_n^{(k)} - x^*) - Z_n)\|_2^2\right)^{1/2} \lesssim \delta(n),$$

where the approximation rate $\delta(n) \to 0$.

Theorem 5.3.4 demonstrates that if we employ ASGD as defined in (5.6), Assumption

5.3.7 is satisfied with $\Sigma_n = \Gamma_n$ defined in (5.11) and

$$\delta(n) = \max\left(n^{1/2-\beta}, \frac{\sqrt{\log n}}{n^{1/2-1/q}}\right)$$

With this assumption, we are ready to show that the statistic in (5.8) is asymptotically pivotal with a specific convergence rate.

Theorem 5.3.8. *Suppose we run Algorithm 4 and Assumption 5.3.7 holds. For any v and $\hat{t}_v$ defined in* (5.8) *we have*

$$\sup_{z\in\mathbb{R}} \left|\mathbb{P}\left(\hat{t}_v \geq z\right) - \mathbb{P}(T_{K-1} \geq z)\right| \lesssim (\delta(N/K))^{1/4},$$

where T_{K-1} is a random variable following t distribution with degree of freedom $K-1$, N is the total sample size and K is the number of parallel runs. Consequently, for any confidence level $\alpha \in (0,1)$,

$$\left|\frac{\mathbb{P}\left(|\hat{t}_\nu| \geq t_{1-\alpha/2,K-1}\right)}{\alpha} - 1\right| \lesssim \alpha^{-1}\delta(N/K)^{1/4},$$

where $t_{1-\alpha/2,K-1}$ is the $(1-\alpha/2)\times 100\%$ percentile for the t_{K-1} distribution and the constant in $\lesssim$ does not depend on α. For $\alpha(N)$ goes to zero with $\delta(N/K)^{1/4} \ll \alpha(N)$, the relative error of coverage goes to zero when $\alpha \geq \alpha(N)$, i.e.,

$$\Delta_N = \sup_{\alpha(N)\leq\alpha<1} \left|\frac{\mathbb{P}(v^\top x^* \in \hat{\mathrm{CI}}) - (1-\alpha)}{\alpha}\right| \to 0.$$

The results suggest that any stochastic algorithm demonstrating appropriate convergence in certain scenarios can be selected, provided its single sequence exhibits convergence towards normality. The convergence of the t-statistic, as well as the relative error in the coverage of the confidence interval, can be bounded based on the rate of convergence to normality.

Furthermore, this study comprehensively examines the reliance on the value of α. The uniform convergence of Δ_N indicates that an extremely small α, or decreasing α, is feasible.

5.4 Experiment

5.4.1 Simulation

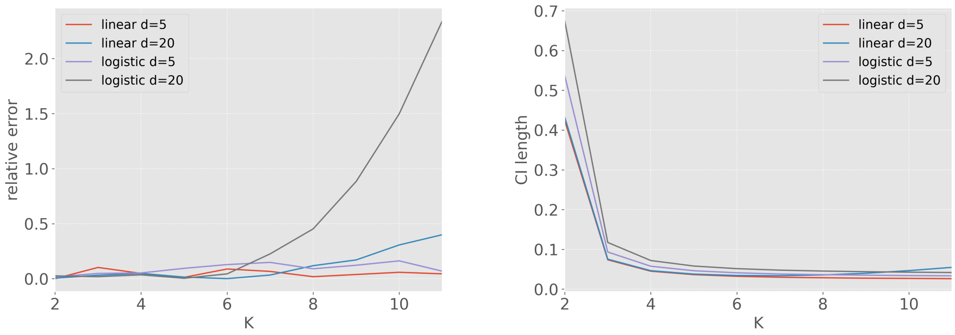
In this section, we investigate the empirical performance of our proposed parallel inference under the linear regression model and logistic regression model. The true coefficient of interest x^* is a d-dimensional vector with $x^* = (0, 1/d, 2/d, ..., (d-1)/d)$. We generate a sequence of i.i.d. random samples $\{(a_i, b_i)\}_{i=1}^n$, where a_i stands for the explanatory variable generated from $\mathcal{N}(0, \mathbf{I}_d)$, and b_i stands for the response variable. In the linear regression model, we have

$$b_i = a_i^T x^* + \epsilon_i,$$

where ϵ_i follows $\mathcal{N}(0,1)$ independently. The corresponding loss function is $f(x, \xi_i) = (a_i^T x - b_i)^2/2$. In the logistic regression model, $b_i \in \{0,1\}$ is generated from a Bernoulli distribution, where

$$\mathbb{P}(b_i = 1|a_i) = \frac{1}{1+\exp(-a_i^T x^*)},$$

and the loss function is logit loss as $f(x, \xi_i) = (1-b_i)a_i^T x + \log(1+\exp(-a_i^T x))$. We employ ASGD for our parallel method, with $\beta = 0.505$, and $\eta = 0.5$, consistent with the settings used in Lee et al. [2022], Zhu et al. [2023]. We consider the case of marginal inference of coordinates, that is the vector v in linear functional is chosen as the canonical basis. To analyze the empirical performance, we record the coverage of the constructed confidence intervals, the relative error of coverage Δ_α as defined in (5.2), the length of the confidence intervals, and the running time. All reported results are the average of 10000 independent trials.



(a) (b)

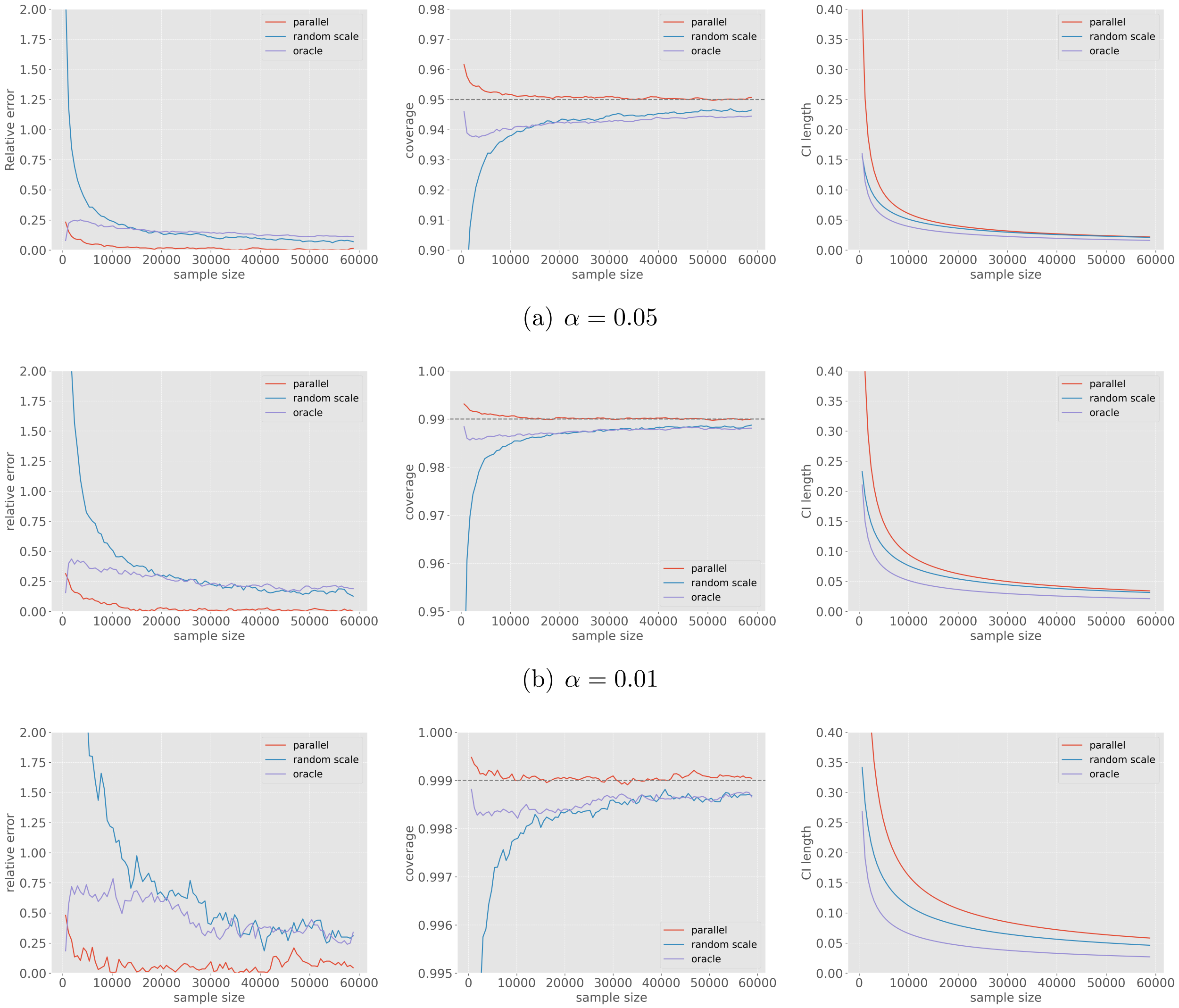
Figure 5.1: Effect of K. Plot (a): relative error of coverage; plot (b): the length of confidence interval. The nominal coverage probability is 0.99. The total sample size N is 60000 for linear models and 200000 for logistic models.

Choice of K. We first examine the effect of K. We construct 99% confidence intervals ($\alpha = 0.01$) with a total sample size of 60000 ($N = nK = 60000$) for linear regression, and with a total sample size of 200000 ($N = nK = 200000$) for logistic regression. From Figure 5.1, we observe that a small K may decrease the bias of coverage in some challenging scenarios, such as logistic regression with $d = 20$. On the other side, it will result in longer confidence intervals. However, when K falls within a reasonable range, say between 2 to 11, the results appear satisfactory and are not overly sensitive to the choice of K. In the following simulation results, we will use $K = 6$.

Compare to another method. We compare the finite sample performance of our proposed inference method, referred to as the parallel method, with that of the state-of-the-art method: the random scaling method [Lee et al., 2022], which also leverages an asymptotic pivotal statistic. The confidence interval constructed by the random scaling method is given in (5.4), and we obtain critical values through Monte Carlo simulation as tabulated in Table D.1. We did not include comparisons with other methods such as the Plug-in [Chen et al., 2020] or Online Batch-means [Zhu et al., 2023], as the random scaling method has already demonstrated comparable coverage to the Plug-in method, superior coverage compared to

Online Batch-means, and faster computing times. For both methods, we apply the ASGD algorithm with $\beta = 0.505$ and $\eta = 0.5$. The number of parallel runs, K, is set to 6 for the parallel method. We consider constructing confidence intervals every 600 samples. Overall, the performance of the parallel method is satisfactory and better than that of the random scaling method, with faster convergence, comparable confidence interval lengths, and less computation.

In Figures 5.2 and 5.3, we present results for confidence intervals where the nominal coverage probability is set at 0.95, 0.99, and 0.999, i.e., $\alpha = 0.05, 0.01, 0.001$ for both linear regression and logistic regression. We plot the relative error of coverage, the empirical coverage rate, and the length of the confidence intervals. We also compare the running time of a single trial in Figure 5.4. More results are summarized in Table **??** and Appendix. The relative error of our parallel method converges to zero faster than that of the random scaling method in all cases. The advantage becomes more obvious as confidence levels increase. Note that in logistic regression with $d = 20$, both methods exhibit relatively large errors when the sample size is small. In this case, the parallel method converges more slowly, which can be attributed to the fact that data splitting in the parallel run exacerbates the issue of a small sample size. However, as the sample size increases, the convergence rate of the parallel method improves and eventually surpasses that of the random scaling method. The lengths of the confidence intervals are comparable between the two methods, with those derived from the parallel method being slightly larger. We also observe that our parallel method has a distinct advantage in terms of computing time, as it does not necessitate additional computations at each iteration, such as updating a d by d matrix, which is required by the random scaling method. Apart from the SGD update, the only additional computation needed for inference is calculating a sample covariance matrix or a sample variance (for the linear functional). This computation is minimal, making the inference process almost cost-free. The advantage in computing becomes even more significant when utilizing parallel

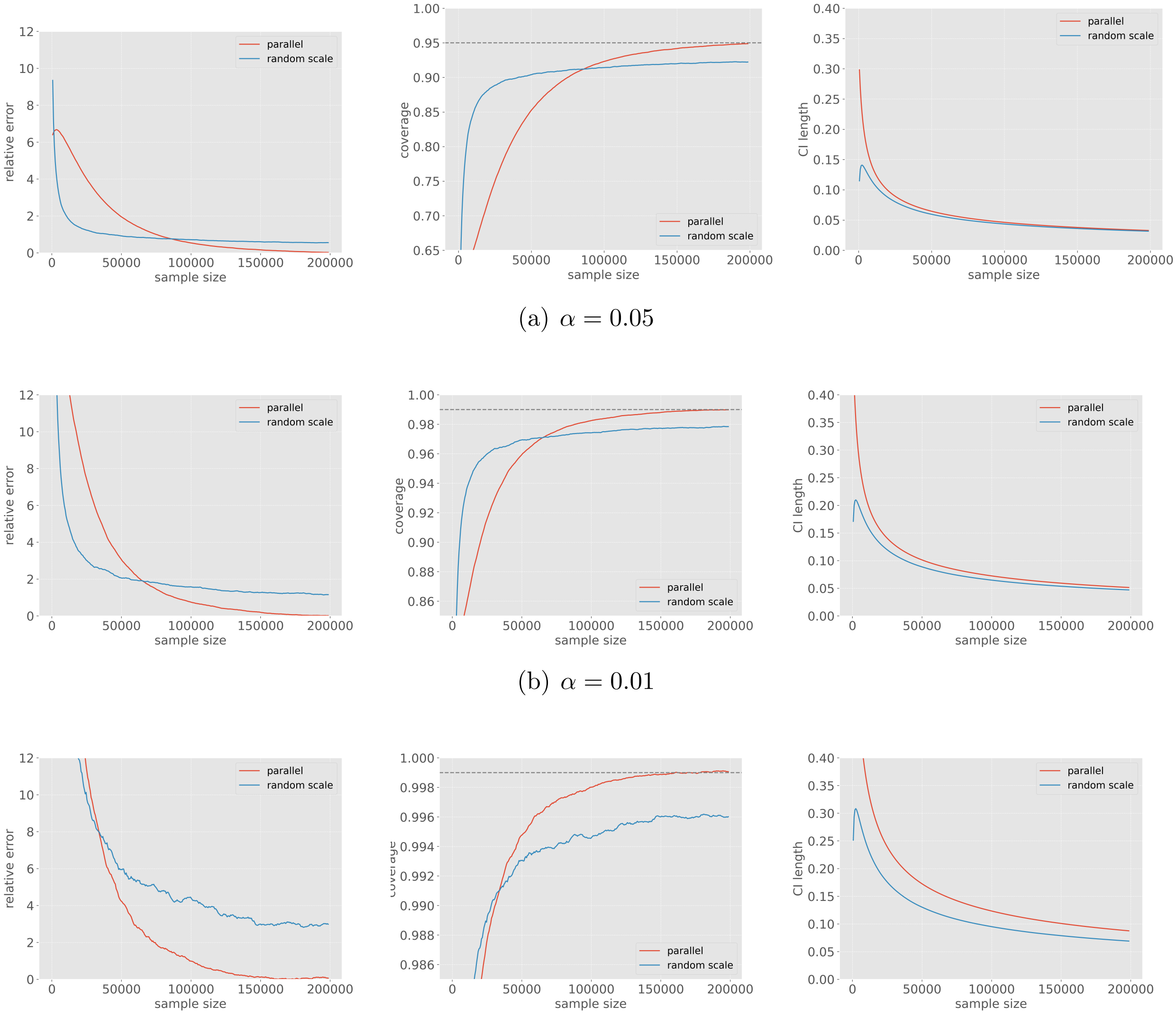


(a) $\alpha = 0.05$

(b) $\alpha = 0.01$

(c) $\alpha = 0.001$

Figure 5.2: Linear Regression $d = 20$: Left: relative error of coverage; Middle: empirical coverage; Right: length of confidence intervals.

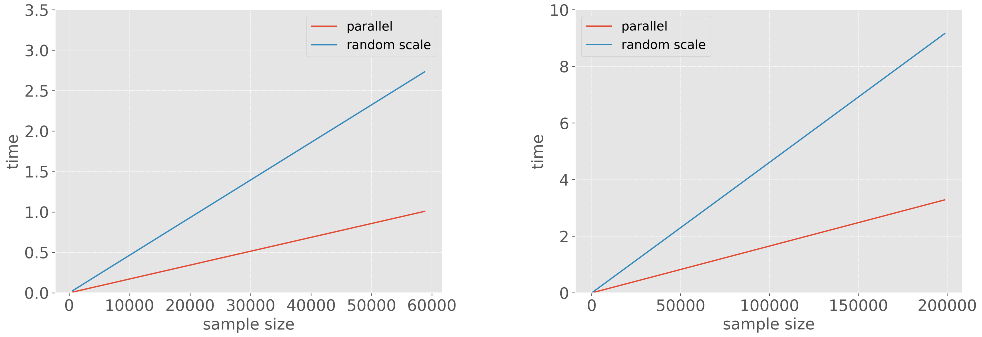


(a) $\alpha = 0.05$

(b) $\alpha = 0.01$

(c) $\alpha = 0.001$

Figure 5.3: Logistic Regression $d = 20$: Left: relative error of coverage; Middle: empirical coverage; Right: length of confidence intervals.



(a) Linear regression

(b) Logistic regression

Figure 5.4: Computation time: d = 20

computing across different cores.

Compare to oracle. We also compare our method to the oracle approach, which constructs confidence intervals using the true limiting covariance matrix Σ and the principles of asymptotic normality. The oracle method is given by:

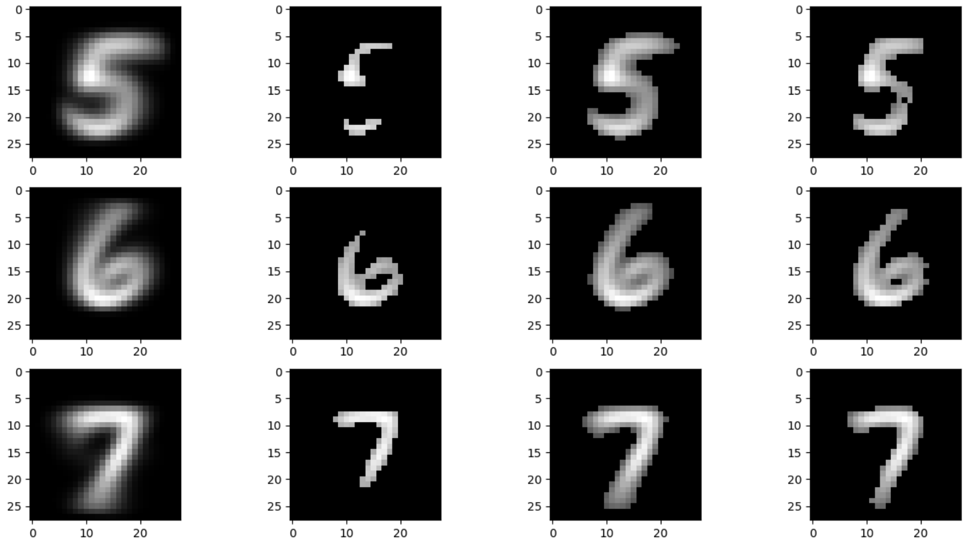
$$\hat{\text{CI}}_{n,\text{oracle}} = \left[\upsilon^\top \bar{x}_n - z_{1-\alpha/2}\sqrt{\frac{\upsilon^\top \Sigma \upsilon}{n}},\ \upsilon^\top \bar{x}_n + z_{1-\alpha/2}\sqrt{\frac{\upsilon^\top \Sigma \upsilon}{n}} \right].$$

In the linear regression model described above, the limiting covariance matrix $\Sigma = \mathbf{I}_d$. We focus on linear regression here since the true limiting covariance matrix is straightforward to compute. As illustrated in Figure 5.2 the coverage achieved by the parallel method surpasses that of the oracle method. This could be attributed to a discrepancy between the finite sample covariance matrix of ASGD and the limiting covariance Σ. Employing an asymptotic pivotal statistic helps to mitigate the impact of this difference. For further details and discussion on this topic, refer to Section 5.3.1.

5.4.2 Hand-written digit dataset

To further explore the application of confidence intervals, we consider the task of estimating the mean image for each digit in the MNIST handwritten digit dataset.

The MNIST dataset comprises 60000 training images, each each measuring 28×28 pixels in dimension and labeled with digits ranging from 0 and 9. For each label $m = 0, 1, \ldots, 9$, we hypothesize the existence of a mean image $x^{*,m} \in \mathbb{R}^d, d = 28 \times 28$, and individual image instances $\{\xi_i^{(m)}\}$ sampled from a normal distribution with mean $x^{*,m}$ and an unknown variance. Our goal is to estimate these mean images. The objective function is defined as $f(x, \xi) = \frac{1}{2}\|\xi - x\|_2^2$. We employ the online parallel algorithm (4) with $K = 6$ and ASGD with a step size at the n-th iteration $\eta_n = n^{-0.505}$. The parallel means computed from this process serve as our final mean estimates. We construct coordinate-wise confidence intervals based on Algorithm 4, choosing vectors $v = e_j, j = 1, ..., 784$. We visualize the mean images in Figure 5.5 (a), noting that in the grayscale representation, 0 denotes gray, -1 denotes black, and 1 denotes white. Our approach further includes denoising; that is, truncating (or 'shrinking') values below a certain threshold to -1 to make the mean image sharper. Traditionally, this threshold lacks formal guidance. We first try a one-size-fits-all threshold. In Figure 5.5 (b) , the uniform threshold is set at 0. The results are not satisfactory; significant portions of the digit 5 were missing, and it did not make sense to shrink the upper part of 6. Changing the uniform threshold to a smaller number, -0.5, as shown in Figure 5.5 (c), the denoising step has no effect. It remains unclear which threshold is effective, and it is uncertain if a single, uniform threshold is appropriate. Then, we refined the denoising step by leveraging confidence intervals constructed by the parallel method: a mean value is set to 0 if its coordinate-wise upper confidence bound is below 0. Given the desire to preserve sufficient pixel detail, we opt for a high confidence level, adjusted for the number of parameters. Thus, we set our confidence level to $1 - \alpha/d$ with $\alpha = 0.001$, effectively achieving a 99.9% confidence interval across all coordinates simultaneously. The



(a) Original mean (b) $t = 0$ (c) $t = -0.5$ (d) Adaptive

Figure 5.5: Mean image before and after denoising. (a) shows the original estimated mean before denoising; (b) uses a uniform threshold $t = 0$; (c) uses a uniform threshold $t = -0.5$; (d) applies an adaptive threshold based on the upper bound of the confidence interval.

results in Figure 5.5 (d) demonstrate that using confidence intervals yields more accurate and visually coherent mean images.

5.5 Summary

In this chapter, we introduce a novel inference framework designed to construct confidence intervals for model parameters by employing stochastic algorithms in an online environment. This method stands out for its simplicity and ease of implementation, offering flexibility across various algorithms. The construction of confidence intervals is the most computationally efficient among all existing online methods and incurs almost no cost post-SGD update. Furthermore, we bolster our approach with rigorous theoretical guarantees, demonstrating its capability to facilitate inference at a high confidence level.

CHAPTER 6

APPROXIMATE CO-SUFFICIENT SAMPLING WITH REGULARIZATION

In previous chapters, we discussed inference for parameter estimation in a model. In this chapter, we turn to the second part of this thesis: goodness-of-fit (GoF) testing, which assesses whether observed data follows a specific pattern or distribution. GoF testing is an essential statistical method, widely used across various fields such as biology, economics, engineering, and finance. Here, we will focus on scenarios where the null distribution is only partially known—limited to a parameterized family of distributions—rather than known exactly. We will approach this question through resampling and address challenges in a high-dimensional setting.

6.1 Introduction

Consider the GoF testing

$$H_0 : X \sim P_\theta \text{ for some } \theta \in \Theta, \tag{6.1}$$

where $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ is a parametric family, versus a more complex (usually higher-dimensional) model. As for any standard hypothesis testing problem, our approach to GoF testing involves two core ingredients: finding a test statistic that captures the important trends in the data (with the convention that large values of $T = T(X)$ indicate evidence against H_0), and deriving the null distribution of this test statistic $T(X)$ so that we can appropriately calibrate our test to make sure we do not exceed the allowable Type I error level. In many settings, this second component often poses the larger challenge; it is often the case that the null distribution of $T(X)$ cannot be computed exactly or even estimated accurately. An alternative approach, common in many statistical problems, is to mimic this null distribution with some form of resampling—e.g., methods based on permutations, on bootstrapping, or on knockoffs [Barber and Candès, 2015, Barber et al., 2020, Beran, 1988, Berrett et al., 2020, Candès et al., 2018, Davidson and MacKinnon, 2007, Efron, 1979, Ernst, 2004, Lehmann et al., 1986, Welch, 1990, Wu, 1986] all have this flavor. At a high level, we can consider sampling *copies* of the observed data, $\tilde{X}^{(1)}, ..., \tilde{X}^{(M)}$, and using the empirical distribution of the statistic, given by the corresponding values $T(\tilde{X}^{(1)}), ..., T(\tilde{X}^{(M)})$, as a null distribution against which we compare the evidence $T(X)$. More concretely, given these sampled copies, we can define a p-value corresponding to the observed evidence $T(X)$ as

$$\text{pval} = \text{pval}_T(X, \tilde{X}^{(1)}, ..., \tilde{X}^{(M)}) = \frac{1}{M+1}\left(1 + \sum_{m=1}^{M} \mathbb{1}\left\{T(\tilde{X}^{(m)}) \geq T(X)\right\}\right). \tag{6.2}$$

If it holds that the real data and its copies $X, \tilde{X}^{(1)}, ..., \tilde{X}^{(M)}$ are exchangeable under the null, then it follows immediately that this p-value is valid under the null, $\mathbb{P}_{H_0}(\text{pval} \leq \alpha) \leq \alpha$ (for any rejection threshold α). The core challenge for this type of approach is therefore reduced to the following question:

> How can we generate copies $\tilde{X}^{(1)}, ..., \tilde{X}^{(M)}$ of the observed data X such that, if H_0 is true, then $X, \tilde{X}^{(1)}, ..., \tilde{X}^{(M)}$ are (approximately) exchangeable?

Now we consider this question specifically for the GoF testing problem. Of course, in the case that $\Theta = \{\theta_0\}$ is a singleton set, the problem is trivial—we can simply draw the $\tilde{X}^{(m)}$'s from the known null distribution P_{θ_0}, so that $X, \tilde{X}^{(1)}, ..., \tilde{X}^{(M)}$ are i.i.d. (and thus, exchangeable). Beyond this trivial case, however, this simple strategy can no longer be used. For example, drawing $\tilde{X}^{(m)}$'s from $P_{\hat{\theta}}$ for a plug-in estimate $\hat{\theta}$, which is often called the *parametric bootstrap* [Efron, 2012, Efron and Tibshirani, 1994, Hall and Maiti, 2006, Singh, 1981], may work well in some settings but has the potential to substantially inflate the Type I error rate [Barber and Janson, 2022, Section 1]. The co-sufficient sampling (CSS) and approximate co-sufficient sampling (aCSS) approaches, which we will describe in detail below, avoid this issue by conditioning on a sufficient (or approximately sufficient) statistic for the unknown θ. aCSS in particular can be applied to a range of models, but is not suited for addressing challenges such as high dimensionality.

In this chapter, our aim is to extend the aCSS approach to the setting where θ cannot be estimated via unconstrained maximum likelihood estimation—for example, a high-dimensional sparse linear regression problem, where unconstrained estimation is not consistent but adding ℓ_1 regularization restores consistency. We develop a form of aCSS that is able to handle constrained maximum likelihood estimation (and will also extend to the penalized case). Consequently, this new approach allows for aCSS to accommodate more robust and accurate parameter estimation in complex problems, particularly in high-dimensional settings.

Notation of the Chapter. For an integer $n \geq 1$, $[n]$ denotes the set $\{1, \ldots, n\}$. We write $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ to denote expectation or probability taken with respect to the distribution P_θ.

6.2 Background: goodness-of-fit testing via CSS and aCSS

First, we recall the general framework for GoF testing. Our goal is to test the null hypothesis H_0 (6.1) that the data X is drawn from P_θ, for some (unknown) $\theta \in \Theta$. We begin by designing a statistic $T : \mathcal{X} \to \mathbb{R}$ that tests this null, with the convention that a larger value $T(X)$ will indicate more evidence against this null. We then need to choose a rejection threshold: can we find a value t_* such that, under the null, $T(X) > t_*$ occurs with probability at most α, while under the alternative, $T(X) > t_*$ is much more likely?

Since the null hypothesis H_0 is composite (aside from the trivial case that Θ is a singleton set), we cannot compute an exact null distribution for $T(X)$, and thus instead we aim to sample copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ that are approximately exchangeable with the observed data X under the null H_0, so that we can then assess $T(X)$ via the p-value defined in (6.2) above. Of course, we can trivially achieve exchangeability by simply taking $\tilde{X}^{(m)} = X$ for each copy m—but this would lead to zero power for testing any alternative, since the p-value defined in (6.2) would be equal to 1 regardless of the choice of test statistic.

In the remainder of this section, we will give background on the CSS and aCSS methods for producing these copies, the $\tilde{X}^{(m)}$'s, along with some examples to illustrate the types of settings where these methods may be applied. From this point on, we will write $\theta_0 \in \Theta$ to denote the unknown true value of the parameter.

6.2.1 Co-sufficient sampling (CSS)

We cannot sample the copies $\tilde{X}^{(m)}$ from the distribution P_{θ_0} of the data X, because of its dependence on the unknown θ_0. To remove this dependence we can condition on a *sufficient statistic* $S(X)$. To be precise, $S(X)$ is a sufficient statistic if the conditional distribution of X no longer depends on θ—that is, we can construct a conditional distribution $P(X \mid S)$ such that, for any $\theta \in \Theta$,

$$\text{If } X \sim P_\theta, \text{ then } X \mid S(X) \text{ has distribution } P(\cdot \mid S(X)).$$

Co-sufficient sampling (see, e.g., Agresti [1992], Engen and Lillegård [1997], Stephens [2012]) leverages this property to sample the copies:

$$\text{CSS method: after observing } X, \text{ sample } \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)} \text{ i.i.d. from } P(\cdot \mid S(X)).$$

By construction, $X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ are exchangeable when $X \sim P_\theta$, for *any* θ—and thus, the p-value constructed in (6.2) is valid under the null H_0 (6.1).

As a concrete example, suppose that $X = (X_1, \ldots, X_n)$ follows a Gaussian linear model,

$$X \sim \mathcal{N}(Z\theta, \nu^2 \mathbf{I}_n),$$

for known covariates $Z \in \mathbb{R}^{n \times d}$ (assumed to have full column rank), known variance $\nu^2 > 0$, and unknown coefficients $\theta \in \Theta = \mathbb{R}^d$. Then $S(X) = Z^\top X$ is a sufficient statistic for this parametric family, and we can calculate the conditional distribution

$$X \mid S(X) \sim \mathcal{N}(Z(Z^\top Z)^{-1}S(X), \nu^2 \mathcal{P}_Z^\perp),$$

where $\mathcal{P}_Z^\perp \in \mathbb{R}^{d\times d}$ is the projection matrix for the subspace orthogonal to the column span of Z. As long as $d < n$, then, the copies $\tilde{X}^{(m)}$ are distinct from X (and from each other), and we may be able to achieve high power under a suitable alternative hypothesis. Additional background and discussion of CSS can be found in [Barber and Janson, 2022, Section 1].

6.2.2 Approximate co-sufficient sampling (aCSS)

While the CSS method performs well for certain goodness-of-fit problems, there are many settings where CSS leads to a degenerate method and consequently zero power. Barber and Janson [2022] consider the example of logistic regression: suppose $X = (X_1, \ldots, X_n)$ follows a logistic regression model, where

$$X_i \sim \text{Bernoulli}(1/(1 + e^{-Z_i^\top \theta}))$$

independently for each $i \in [n]$, where again $Z_1, \ldots, Z_n \in \mathbb{R}^d$ are known covariate vectors, while $\theta \in \Theta = \mathbb{R}^d$ is unknown. In this case, for generic values of the Z_i's (for instance, if these covariates are drawn from some continuous distribution), the minimal sufficient statistic $S(X) = Z^\top X$ uniquely determines X ($Z \in \mathbb{R}^{n\times d}$ is the matrix with rows Z_i)—that is, the conditional distribution of $X \mid S(X)$ is simply a point mass. Consequently, applying CSS to this problem would lead to zero power since we would have $X = \tilde{X}^{(1)} = \cdots = \tilde{X}^{(M)}$.

To address this type of degenerate scenario, Barber and Janson [2022] propose *approximate co-sufficient sampling* (aCSS). The idea of aCSS is to condition on less information (to restore power), while ensuring that the sampled copies are approximately exchangeable (to retain Type I error control). (We refer the reader to [Barber and Janson, 2022, Section 1] for a more comprehensive discussion on the comparison between bootstrap, CSS, and aCSS methods.)

Concretely, consider an approximate maximum likelihood estimator,

$$\hat{\theta} = \hat{\theta}(X, W) = \text{argmin}_{\theta \in \Theta} \left\{ -\log f(X;\theta) + R(\theta) + \sigma W^\top \theta \right\},$$

where $f(\cdot;\theta)$ is the density for distribution P_θ (with respect to some base measure), $R(\theta)$ is an optional twice-differentiable regularizer (e.g., a ridge penalty), $W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ is Gaussian noise that adds a perturbation to the maximum likelihood estimation problem, and $\sigma > 0$ is a parameter that controls the magnitude of this perturbation. For each $\theta \in \Theta$, define $P_\theta(\cdot \mid \hat{\theta})$ as the conditional distribution of $X \mid \hat{\theta}$, when $X \sim P_\theta$ and $\hat{\theta} = \hat{\theta}(X, W)$ is defined as above.

Now we return to the GoF problem, where $X \sim P_{\theta_0}$ for an unknown θ_0. Note that, even if the unperturbed MLE were a sufficient statistic (as would be the case for a Gaussian linear model, for example), the perturbed MLE $\hat{\theta}$ is no longer a sufficient statistic in the exact sense, and so the conditional distribution $P_{\theta_0}(\cdot \mid \hat{\theta})$ does depend on the unknown parameter θ_0. However, it turns out that $\hat{\theta}$ is approximately sufficient, meaning that $P_{\theta_0}(\cdot \mid \hat{\theta})$ depends only weakly on θ_0. In particular, Barber and Janson [2022]'s method proposes replacing θ_0 with $\hat{\theta}$ as a plug-in estimate:

aCSS method: after observing X, draw $W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$, compute $\hat{\theta} = \hat{\theta}(X, W)$, then sample $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ i.i.d. from $P_{\hat{\theta}}(\cdot \mid \hat{\theta})$.

Of course, these copies are no longer exactly exchangeable with X under the null, since in general we will have $P_{\hat{\theta}}(\cdot \mid \hat{\theta}) \neq P_{\theta_0}(\cdot \mid \hat{\theta})$. To quantify this issue, Barber and Janson [2022] define the "distance to exchangeability",

$$d_{\text{exch}}(A_1, \ldots, A_k) = \inf \left\{ d_{\text{TV}}((A_1, \ldots, A_k), (B_1, \ldots, B_k)) : B_1, \ldots, B_k \text{ are exchangeable} \right\},$$

where d_{TV} denotes the total variation distance. The p-value defined in (6.2) is then approximately valid with

$$\mathbb{P}(\text{pval}_T(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \leq \alpha) \leq \alpha + d_{\text{exch}}(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}),$$

where $d_{\text{exch}}(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)})$ can be bounded under certain conditions on the parametric family $\{P_\theta : \theta \in \Theta\}$.

While aCSS is able to handle a far broader range of models and problems than the CSS framework, there are nonetheless limitations to this method that motivate our present work. In particular, Barber and Janson [2022]'s work assumes a bound on $\|\hat{\theta} - \theta_0\|_2$, i.e., consistency of the perturbed MLE $\hat{\theta}$, which may not be possible to achieve in high dimensional settings unless we regularize using constraints or non-smooth penalization. Moreover, computing $P_\theta(\cdot \mid \hat{\theta})$, which is a key step in the aCSS procedure, relies heavily on the fact that $\hat{\theta}$ is the solution to an *unconstrained, differentiable* optimization problem over a *convex, open* parameter space $\Theta \subseteq \mathbb{R}^d$ (as these assumptions allow for using first-order optimality conditions on $\hat{\theta}$ to derive this conditional distribution), and consequently, aCSS is not able to handle optimization under constraints or under a non-differentiable penalty.

The role of σ

Here we pause to discuss the role of the noise parameter σ in the aCSS method, and the tradeoffs inherent in choosing the value of σ. The aCSS method requires choosing a parameter $\sigma > 0$ that controls the amount by which the MLE is perturbed. As discussed by Barber and Janson [2022], the choice of σ represents a tradeoff between Type I error control, and the statistical and computational efficiency of the method. A smaller σ leads to a lower inflation of the Type I error (that is, Barber and Janson [2022]'s bound on $d_{\text{exch}}(X, \tilde{X}^{(1)}, \dots, \tilde{X}^{(M)})$ increases with σ). On the other hand, choosing σ to be too small can lead to low power—if the perturbed MLE $\hat{\theta}$ reveals too much information about X, the copies $\tilde{X}^{(m)}$ may be extremely similar to X and therefore, our power to reject the null is low. Moreover, a small value of σ makes it more challenging to sample the $\tilde{X}^{(m)}$'s from the conditional distribution of $X \mid \hat{\theta}$, since this distribution becomes more concentrated as σ tends to zero.

As we will see later on, these considerations will play an important role in our constrained version of aCSS, as well. We will return to a discussion of this parameter in Section 6.4.1 below, after defining our new methods and presenting theoretical results.

6.2.3 Additional related work

The literature on GoF testing is extensive, particularly in low-dimensional settings, and giving an overview of this broad field is beyond the scope of the present work. Here we discuss some challenges faced in the high-dimensional regime. A crucial prerequisite for valid testing is the reliable estimation of underlying parameters. In high-dimensional settings, achieving consistent parameter estimation is impossible without additional structural assumptions. Constraints serve as an effective tool for incorporating prior knowledge about the structure into the estimation process. The most common illustration of this is the application of LASSO [Tibshirani, 1996] and the Dantzig selector [Candès and Tao, 2007] under specific sparsity assumptions. These techniques, linked with ℓ_1-regularization, have been demonstrated to be consistent [Bickel et al., 2009, Zhang and Huang, 2008, Zhao and Yu, 2006]. When applied to GoF testing, there has been much work on inference and testing in high dimensional generalized linear models (GLM) considering lasso and sparse models (see, e.g., Janková et al. [2020], Van de Geer et al. [2014], Zhang and Zhang [2014]). For two-sample test in high dimensions, Srivastava et al. [2016] focus on projecting the high-dimensional data onto a lower-dimensional subspace and Li et al. [2020a] propose a test based on a ridge-regularized Hotelling's T^2. While much of the aforementioned work centers on simple null settings, the methods used to manage high-dimensional data provide valuable insights for our scenarios.

6.3 The aCSS method with linear constraints

Our constrained aCSS method will address the problem of goodness-of-fit testing for the hypothesis

$$H_0 : X \sim P_\theta \text{ for some } \theta \in \Theta,$$

where as before, $\{P_\theta : \theta \in \Theta\}$ is a parametric family, indexed by a convex and open subset $\Theta \subseteq \mathbb{R}^d$. For Barber and Janson [2022]'s aCSS method to provide approximate Type I error control, we need consistency of the (perturbed) MLE, i.e., a bound on $\|\hat{\theta} - \theta_0\|_2$. Many important problems are therefore excluded from this framework. In particular, consistency of the MLE cannot be assumed for problems where the unconstrained MLE is not well-defined—for example, a mixture of two Gaussians with unknown means and variances, due to the degenerate behavior of the likelihood as we take one component's variance to zero. In addition, consistency of the MLE will not hold for high-dimensional problems, such as Gaussian linear regression with dimension d larger than the sample size n—even if we add a ridge regularizer $R(\theta)$ so that the solution $\hat{\theta}$ is unique, in general $\hat{\theta}$ will not be a consistent estimator of θ. In contrast to aCSS, however, where we need to be able to estimate the true parameter θ_0 accurately with the *unconstrained* MLE solution $\hat{\theta}$, here we are interested in settings where θ_0 can only be accurately estimated with a *constrained* optimization problem.

To this end, we now introduce constraints,

$$A\theta \leq b,$$

for a fixed and known matrix $A \in \mathbb{R}^{r\times d}$ and vector $b \in \mathbb{R}^r$. The inequality should be interpreted elementwise, i.e., we are requiring $(A\theta)_i \leq b_i$ for each $i = 1, \dots, r$. (Of course, in the special case $r = 0$, this reduces to the earlier, unconstrained setting.) At a high level, to run aCSS in this setting, we first need to compute a constrained MLE (with a random perturbation),

$$\hat{\theta} = \hat{\theta}(X, W) = \mathrm{argmin}_{\theta\in\Theta} \left\{(\theta; X, W) \,:\, A\theta \leq b\right\}, \tag{6.3}$$

where

$$(\theta; X, W) = -\log f(X; \theta) + R(\theta) + \sigma W^\top \theta.$$

As before, $f(\cdot;\theta)$ is the density for distribution P_θ, $R(\theta)$ is an optional twice-differentiable regularizer, $W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ is independent Gaussian noise, and $\sigma > 0$ is a parameter that controls the magnitude of this perturbation. We then compute the conditional distribution of X given $\hat{\theta}$, and sample the copies $\tilde{X}^{(1)}, \dots, \tilde{X}^{(M)}$ from this conditional distribution (or rather, sample from an approximation, since θ_0 is unknown). Defining

$$\hat{g} = \hat{g}(X, W) = \nabla_\theta(\hat{\theta}(X, W); X, W), \tag{6.4}$$

we can see that we would trivially have $\hat{g} \equiv 0$ in the unconstrained setting but may in general have $\hat{g} \neq 0$ now that constraints have been introduced. We will see that, in the constrained optimization setting, while $\hat{\theta}$ on its own does not carry enough information to serve as an approximately sufficient statistic, instead the pair $(\hat{\theta}, \hat{g})$ now plays this role.

For each $\theta \in \Theta$, we will define $P_\theta(\cdot \mid \hat{\theta}, \hat{g})$ as the conditional distribution of $X \mid (\hat{\theta}, \hat{g})$ if we assume that X was drawn as $X \sim P_\theta$. Using $\hat{\theta}$ as a plug-in for the true parameter θ_0, we will use $P_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ as the distribution from which the copies $\tilde{X}^{(m)}$ are drawn. The constrained aCSS algorithm is then defined via the following steps:

Constrained aCSS algorithm (informal version):

1. Observe data $X \sim P_{\theta_0}$.
2. Draw noise $W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$.
3. Solve for a constrained perturbed MLE $\hat{\theta} = \hat{\theta}(X, W)$ as in (6.3), and compute the corresponding gradient $\hat{g} = \hat{g}(X, W)$ as in (6.4).
4. Sample the copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ from the approximate conditional distribution $P_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$.
5. Compute the p-value defined in (6.2) for our choice of test statistic T.

As compared to (unconstrained) aCSS, the difference lies in the fact that $\hat{\theta}$ is computed via a constrained optimization problem, and as a result, the conditional distribution $P_\theta(\cdot \mid \hat{\theta}, \hat{g})$ is now more challenging to compute; we will return to this question shortly.

When running constrained aCSS, we note that we are not assuming explicitly that the true parameter θ_0 itself satisfies the constraints—that is, we do not assume $A\theta_0 \leq b$ must hold. However, in order for the method to retain approximate Type I error control, $\hat{\theta} = \hat{\theta}(X, W)$ will need to be an accurate estimator of θ_0; this implicitly requires that $A\theta_0 \leq b$ must at least approximately hold.

The choice of σ controls the amount of perturbation in the constrained MLE $\hat{\theta}$. This choice represents a tradeoff between Type I error, which is better for small σ, versus statistical power and computational efficiency, which tend to improve with larger σ—this tradeoff occurs for unconstrained aCSS as well (see Section 6.2.2). For constrained aCSS, additional challenges can arise since we may now be working in a high-dimensional setting—we will discuss these questions more in Section 6.4 below, when presenting our theoretical results, and will explore the role of σ empirically in our simulations in Section 6.6.

6.3.1 Examples of constraints

Before defining the method more formally, we present several key examples of constraints $A\theta \le b$ to motivate this method.

- Nonnegativity constraint: if we believe θ_0 has only nonnegative entries, we can choose

  $$A = -\mathbf{I}_d, \quad b = \mathbf{0}_d$$

  to enforce $\theta_i \ge 0$ for all i.

- Bounding away from zero: if we believe the entries of θ_0 cannot be too close to zero, we can choose

  $$A = -\mathbf{I}_d, \quad b = -c \cdot \mathbf{1}_d,$$

  for a small constant $c > 0$ (or we can take a submatrix of the identity, if we want to place a lower bound on only certain entries of θ), to enforce $\theta_i \ge c$ for all i (or for certain entries). For example, for a Gaussian mixture model, we need to place a positive lower bound on the variance of each component in order for the MLE to be well-defined.

- Monotonicity constraint: if we believe θ_0 has entries that appear in nondecreasing order, i.e., $(\theta_0)_1 \le \cdots \le (\theta_0)_d$, we can choose

  $$A = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}, \quad b = \mathbf{0}_d,$$

  to enforce the monoticity constraint $\theta_1 \le \cdots \le \theta_d$.

- ℓ_∞ constraint: if we believe θ_0 has bounded entries, we can choose

$$A = \begin{pmatrix} \mathbf{I}_d \\ -\mathbf{I}_d \end{pmatrix}, \quad b = C \cdot \mathbf{1}_{2d},$$

  to enforce the constraint $\|\theta\|_\infty \leq C$.

- ℓ_1 constraint: if we believe that θ_0 is sparse or approximately sparse, such as in a high-dimensional regression problem, we can choose

$$A \in \{\pm 1\}^{2^d \times d} \text{ (with rows given by the set of sign vectors of length } d\text{)}, \quad b = C \cdot \mathbf{1}_{2^d}$$

  in order to enforce the constraint $\|\theta\|_1 \leq C$. (Note that, in high-dimensional statistics, it is more common to use an ℓ_1 penalty—i.e., the lasso—rather than an ℓ_1 constraint, when defining the regularized MLE. We will define a penalized version of our method later on, in Section 6.5.)

- Fused ℓ_1 norm constraint: if we believe θ_0 is locally constant (or is smooth and therefore can be well approximated by a locally constant vector), we can choose to constrain $\|D\theta\|_1 \leq C$, where $D \in \{-1, 0, +1\}^{(d-1)\times d}$ is defined with first row $(+1, -1, 0, \dots, 0)$, second row $(0, +1, -1, 0, \dots, 0)$, etc, so that $\|D\theta\|_1 = \sum_{i=1}^{d-1} |\theta_i - \theta_{i+1}|$. This corresponds to choosing $A \in \mathbb{R}^{2^{d-1}\times d}$ given by $A = A' \cdot D$, where $A' \in \{\pm 1\}^{2^{d-1}\times(d-1)}$ has rows given by all possible sign vectors of length $d-1$, and $b = C \cdot \mathbf{1}_{2^{d-1}}$.

6.3.2 Formally defining the method

We now turn to the details of the method and its implementation, including questions of optimization and sampling, then combine all these ingredients to formally define the constrained aCSS method.

The second-order stationary condition

First we consider the question of optimization. In certain settings, it may be the case that we cannot reliably solve for the global minimizer of $(\theta; X, W)$, or, that this global minimizer may not be well-defined or may not be unique—for example, the negative log-likelihood might be nonconvex. Formally, we define

$$\hat{\theta} : \mathcal{X} \times \mathbb{R}^d \to \Theta$$

to be *any* measurable function, which represents the output of our solver when we input the constrained optimization problem (6.3). For each subset $\mathcal{I} \subseteq [r]$ of constraints, define a matrix $U_{\mathcal{I}}$ that forms an orthonormal basis for subspace orthogonal to $\mathrm{span}\{A_i : i \in \mathcal{I}\}$ (where $A_i \in \mathbb{R}^d$ is the vector given by the ith row of A), that is,

$$U_{\mathcal{I}} \in \mathbb{R}^{d \times (d - \mathrm{rank}(\mathrm{span}\{A_i\}_{i \in \mathcal{I}}))} \text{ satisfies } U_{\mathcal{I}} U_{\mathcal{I}}^\top = \mathcal{P}^{\perp}_{\mathrm{span}\{A_i\}_{i \in \mathcal{I}}}, \tag{6.5}$$

so that $U_{\mathcal{I}} U_{\mathcal{I}}^\top$ projects to the subspace orthogonal to the span of constraints indexed by $\mathcal{I}$.

Definition 6.3.1 (SSOSP). *A parameter $\theta \in \Theta$ is a strict second-order stationary point (SSOSP) of the optimization problem* (6.3) *if it satisfies all of the following:*

1. *Feasibility:*

$$A\theta \leq b.$$

2. *First-order necessary conditions, i.e., Karush–Kuhn–Tucker (KKT) conditions:*

$$\nabla_\theta(\theta; X, W) + \sum_{i=1}^{r} \lambda_i A_i = 0,$$

where $\lambda_i \geq 0$ for all i, and $\lambda_i = 0$ for all $i \in [r] \backslash \mathcal{I}(\theta)$, where $\mathcal{I}(\theta) = \{i \in [r] : A_i^\top \theta = b_i\}$ is the set of active constraints.

3. *Second-order sufficient condition:*

$$U_{\mathcal{I}(\theta)}^\top \nabla_\theta^2(\theta; X, W) U_{\mathcal{I}(\theta)} \succ 0,$$

that is, the Hessian $\nabla_\theta^2(\theta; X, W)$ is strictly positive definite when restricted to the subspace orthogonal to the active constraints.

As in the unconstrained aCSS algorithm [Barber and Janson, 2022], to allow for the possibility that our solver might fail to find a valid solution, if $\hat{\theta}(X, W)$ fails the SSOSP condition then we will set $\tilde{X}^{(1)} = \cdots = \tilde{X}^{(M)} = X$ to trivially obtain a p-value of 1 (i.e., to avoid the possibility of a rejection in this scenario where our estimate $\hat{\theta}$ of θ_0 is unreliable).

The conditional distribution

With the SSOSP condition in place, we are now ready to define the conditional distribution $P_\theta(\cdot \mid \hat{\theta}, \hat{g})$. We first need some regularity conditions.

Assumption 6.3.2. *Assume the family $\{P_\theta : \theta \in \Theta\}$ and regularization function $R(\theta)$ satisfy:*

- $\Theta \subseteq \mathbb{R}^d$ *is a convex and open set;*
- *For each $\theta \in \Theta$, P_θ has density $f(x; \theta) > 0$ with respect to a common base measure $\nu_{\mathcal{X}}$;*
- *for each $x \in \mathcal{X}$, the function $\theta \to (\theta; x) = -\log f(x; \theta) + R(\theta)$ is continuously twice differentiable.*

This first assumption is the same as Assumption 1 of Barber and Janson [2022], for the unconstrained aCSS setting. The following result, however, is a strict generalization of [Barber and Janson, 2022, Lemma 1], computing the conditional density of X after solving for $\hat{\theta}$ under linear constraints (with the unconstrained setting as a special case).

Lemma 6.3.3 (Conditional density)**.** *Suppose Assumption 6.3.2 holds. For* $A \in \mathbb{R}^{r\times d}$*,* $b \in \mathbb{R}^r$*, fix any* $\theta_0 \in \Theta$ *and let* $(X, W, \hat{\theta}, \hat{g})$ *be drawn from the joint model*

$$\begin{cases} X \sim P_{\theta_0}, \\ W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d), \\ \hat{\theta} = \hat{\theta}(X, W), \\ \hat{g} = \hat{g}(X, W) = \nabla_\theta(\hat{\theta}; X, W). \end{cases} \tag{6.6}$$

Fix any $\mathcal{I} \subseteq [r]$*, and assume that the event that* $\hat{\theta}(X, W)$ *is a SSOSP of* (6.3) *with active set* $\mathcal{I}(\hat{\theta}(X, W)) = \mathcal{I}$ *has positive probability. Then, conditional on this event, the conditional distribution of* $X|\hat{\theta}, \hat{g}$ *has density*

$$p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}) \propto f(x; \theta_0) \cdot \exp\left\{-\frac{\|\hat{g} - \nabla_\theta(\hat{\theta}; x)\|_2^2}{2\sigma^2/d}\right\} \cdot \det\left(U_{\mathcal{I}}^\top \nabla_\theta^2(\hat{\theta}; x) U_{\mathcal{I}}\right) \cdot \mathbb{1}_{x \in \mathcal{X}_{\hat{\theta}, \hat{g}}} \tag{6.7}$$

with respect to the base measure $\nu_{\mathcal{X}}$*, where* $U_{\mathcal{I}}$ *is defined in* (6.5) *and*

$$\mathcal{X}_{\theta,g} = \left\{x \in \mathcal{X} : \text{ for some } w \in \mathbb{R}^d,\, \theta = \hat{\theta}(x, w) \text{ is a SSOSP of (6.3), and } g = \nabla(\theta; x, w)\right\}.$$

The four terms of the conditional density reflect, respectively, the original distribution of X in the first term; the Gaussian distribution of the noise W in the second term; the determinant term, which captures a change-of-variables type calculation relating (X, W) with $(X, \hat{\theta}, \hat{g})$; and the final indicator term, which accounts for possible failure to find a SSOSP. In the case where $\mathcal{I} = \emptyset$, i.e., no active constraints, we have $\hat{g} \equiv 0$ (by first-order optimality) and the conditional density then coincides with the calculations in Barber and Janson [2022] for the unconstrained case.

With this calculation in place, we can now specify the estimated conditional distribution $P_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$, from which we would like to sample the copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ for the constrained aCSS algorithm: it is the distribution obtained by plugging in $\hat{\theta}$ in place of the unknown θ_0, in the conditional distribution computed in Lemma 6.3.3, namely,[1]

$$p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}) \propto f(x;\hat{\theta}) \cdot \exp\left\{-\frac{\|\hat{g} - \nabla_\theta(\hat{\theta};x)\|_2^2}{2\sigma^2/d}\right\} \cdot \det\left(U_{\mathcal{I}(\hat{\theta})}^\top \nabla_\theta^2(\hat{\theta};x) U_{\mathcal{I}(\hat{\theta})}\right) \cdot \mathbb{1}_{x\in\mathcal{X}_{\hat{\theta},\hat{g}}}. \quad (6.8)$$

Sampling strategies

In the informal version of the algorithm defined above, we require that the copies $\tilde{X}^{(m)}$ are drawn i.i.d. from the conditional density $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$, as calculated in (6.8). In other words, conditional on $X, \hat{\theta}, \hat{g}$, the collection of copies is drawn from a product distribution,

$$(\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \mid (X, \hat{\theta}, \hat{g}) \sim p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}) \times \cdots \times p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}). \quad (6.9)$$

In some settings, this may be computationally very easy—we will see some examples of this type below when the parametric family $\{P_\theta\}$ is Gaussian. In more complex settings, however, sampling directly from $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ may be infeasible, and we will instead turn to approximations, such as MCMC-based strategies. Of course, without analyzing complex conditions such as the mixing time of the Markov chain, we cannot ensure that theoretical guarantees enjoyed by the algorithm would be preserved when sampling directly from $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ is replaced with an approximation—particularly as this approximation might induce additional dependence among the copies.

In the unconstrained aCSS setting, Barber and Janson [2022] describe several exchangeable MCMC strategies, based on the work of Besag and Clifford [1989], that avoid these difficulties. For completeness, we will describe these schemes in more detail in Appendix E.4.1. In general, following Barber and Janson [2022], we can generalize the sampling strategy (6.9),

[1] For this to result in a well defined density, we need to verify that the right-hand side integrates to a positive and finite value; in fact, this holds almost surely on the event that $\hat{\theta} = \hat{\theta}(X, W)$ is a SSOSP, as we will verify in Appendix E.2.1.

drawing the copies as

$$(\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \mid (X, \hat{\theta}, \hat{g}) \sim \tilde{P}_M(\cdot; X, \hat{\theta}, \hat{g})$$

where the family of conditional distributions $\{\tilde{P}_M(\cdot; x, \theta, g)\}$ is required to satisfy the following condition:

$$\begin{gathered}\text{If } X \sim p_\theta(\cdot \mid \theta, g) \text{ and } (\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \mid X \sim \tilde{P}_M(\cdot; X, \theta, g), \text{ then} \\ \text{the random vector } (X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \text{ is exchangeable.}\end{gathered} \tag{6.10}$$

In particular, we note that choosing

$$\tilde{P}_M(\cdot; x, \theta, g) = p_\theta(\cdot \mid \theta, g) \times \cdots \times p_\theta(\cdot \mid \theta, g),$$

i.e., sampling the copies i.i.d. from $p_\theta(\cdot \mid \theta, g)$, will trivially always satisfy the exchangeability condition (6.10). More generally, however, if sampling the copies directly from $p_\theta(\cdot \mid \theta, g)$ is computationally infeasible, the MCMC based strategy described in Appendix E.4.1 will also satisfy (6.10) while allowing for more complex problems where direct sampling is not achievable.

Combining everything

With all our formal calculations and definitions in place, we can now state the full version of the constrained aCSS algorithm.

Constrained aCSS algorithm:

1. Observe data $X \sim P_{\theta_0}$.
2. Draw noise $W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$.

3. Solve for a constrained perturbed MLE $\hat{\theta} = \hat{\theta}(X, W)$ as in (6.3), and compute the corresponding gradient $\hat{g} = \hat{g}(X, W)$ as in (6.4).
4. If $\hat{\theta}$ is not a SSOSP of (6.3), then set $\tilde{X}^{(1)} = \cdots = \tilde{X}^{(M)} = X$. Otherwise, sample copies $(\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \mid (X, \hat{\theta}, \hat{g}) \sim \tilde{P}_M(\cdot; X, \hat{\theta}, \hat{g})$, where $\tilde{P}_M$ is chosen to satisfy property (6.10) relative to the conditional density $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ as computed in (6.8).
5. Compute the p-value defined in (6.2) for our choice of test statistic T.

This more general form of the constrained aCSS algorithm is more flexible than our original informal definition: it allows us to handle settings where solving for the (perturbed, constrained) MLE is more challenging (e.g., convergence may not be guaranteed), as well as settings where sampling directly from the estimated conditional density (6.8) may be computationally infeasible.

6.4 Theoretical results

In this section, we provide theoretical guarantees for the constrained aCSS procedures, establishing an upper bound on the Type I error level of the test. First, in Section 6.4.1, we give a general result that holds for any problem where constrained aCSS can be applied. We will then refine the result to provide a stronger bound for two special cases: Section 6.4.2 addresses the setting where $\hat{\theta}$ is sparse in some basis, and Section 6.4.3 considers the setting of (potentially high-dimensional) Gaussian data.

6.4.1 Main result: Type I error control

In order to establish a bound on the Type I error level of the constrained aCSS procedure, we first need several assumptions (in addition to the regularity conditions of Assumption 6.3.2). The following assumption ensures that, with high probability, we successfully find a strict second-order stationary point (SSOSP) $\hat{\theta}$ of the optimization problem (6.3), and this solution $\hat{\theta}$ is a good approximation to the true parameter θ_0.

Assumption 6.4.1. *For any $\theta_0 \in \Theta$ in Assumption 6.3.2, the estimator $\hat{\theta} : \mathcal{X} \times \mathbb{R}^d \to \Theta$ satisfies*

$$\begin{cases} \hat{\theta}(X, W) \text{ is a SSOSP of the constrained optimization problem (6.3),} \\ \|\hat{\theta}(X, W) - \theta_0\|_2 \leq r(\theta_0), \end{cases}$$

with probability at least $1-\delta(\theta_0)$, where the probability is taken with respect to the distribution $(X, W) \sim P_{\theta_0} \times N(0, \frac{1}{d}\mathbf{I}_d)$.

Next, we need an assumption on the Hessian of the log-likelihood. Define $H(\theta; x) = -\nabla^2_\theta \log f(x; \theta)$, and let $H(\theta) = \mathbb{E}_{\theta_0}\left[H(\theta; x)\right]$.

Assumption 6.4.2. *For any $\theta_0 \in \Theta$, the expectation $H(\theta)$ exists for all $\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta$, and furthermore*

$$\mathbb{E}_{\theta_0}\left[\sup_{\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta} r(\theta_0)^2 \left(\lambda_{\max}\left(H(\theta) - H(\theta; X)\right)\right)_+\right] \leq \epsilon(\theta_0), \tag{6.11}$$

$$\log \mathbb{E}_{\theta_0}\left[\exp\left\{\sup_{\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta} r(\theta_0)^2 \cdot (\lambda_{\max}(H(\theta; X) - H(\theta)))_+\right\}\right] \leq \epsilon(\theta_0). \tag{6.12}$$

Here $r(\theta_0)$ is the same constant as that appears in Assumption 6.4.1.

These two assumptions are analogous to Assumptions 2 and 3 in Barber and Janson [2022]'s theoretical results for unconstrained aCSS. However, in the present work $\hat{\theta}$ is defined as the solution to the constrained, rather than unconstrained, perturbed maximum likelihood estimation problem. Since constraints allow for more accurate estimation in many settings, we can expect that the error $\|\hat{\theta} - \theta_0\|_2$ might be substantially smaller in this constrained setting, making these assumptions more realistic for a broader range of problems.

Theorem 6.4.3. *Suppose Assumptions 6.3.2, 6.4.1, 6.4.2 hold, and the data is generated as* $X \sim P_{\theta_0}$*. Then the copies* $\tilde{X}^{(1)}, \dots, \tilde{X}^{(M)}$ *generated by the constrained aCSS procedure are approximately exchangeable with* X*, satisfying*

$$d_{\text{exch}}(X, \tilde{X}^{(1)}, \dots, \tilde{X}^{(M)}) \leq 3\sigma r(\theta_0) + \epsilon(\theta_0) + \delta(\theta_0),$$

where $r(\theta_0), \epsilon(\theta_0), \delta(\theta_0)$ *are defined in Assumptions 6.4.1 and 6.4.2. In particular, this implies that for any predefined test statistic* $T : \mathcal{X} \to \mathbb{R}$ *and rejection threshold* $\alpha \in [0,1]$*, the p-value defined in* (6.2) *satisfies*

$$\mathbb{P}\left(\text{pval}_T(X, \tilde{X}^{(1)}, \dots, \tilde{X}^{(M)}) \leq \alpha\right) \leq \alpha + 3\sigma r(\theta_0) + \epsilon(\theta_0) + \delta(\theta_0).$$

The above upper bound on the Type I error appears identical to the result of [Barber and Janson, 2022, Theorem 1], but in fact this new result offers important contributions. Firstly, this new result holds for the more complex setting of a constrained optimization problem, which requires a more technical analysis. Moreover, as mentioned above, the estimation error $\|\hat{\theta} - \theta_0\|_2$ may be much smaller for the constrained optimization problem, since constraints can reduce the effective dimensionality of the statistical problem; consequently, the value of $r(\theta_0)$ can be much smaller in the constrained setting, leading to a tighter bound on Type I error control. (We will see that our empirical results, shown in Section 6.6, support this intuition.)

Revisiting the role of σ

As discussed earlier in Section 6.2.2, the choice of σ plays an important role in the performance of the method, typically with better Type I error control when σ is smaller versus better power when σ is larger. Now we return to this question in the context of constrained aCSS. The upper bound on Type I error shown in Theorem 6.4.3 suggests that σ should not be too large—in particular, for most statistical settings with sample size n, we can expect $r(\theta_0) \asymp n^{-1/2}$ at best, suggesting that we need to choose $\sigma \ll n^{1/2}$ to ensure a meaningful bound on Type I error. On the other hand, recalling that the noise W in the perturbed maximum likelihood estimation problem (6.3) is generated as $W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$, in a high-dimensional setting where $d \gg n$ the perturbation term $\sigma W^\top \theta$ in (6.3) may therefore be negligible. This might lead to extremely low power and/or to computational challenges in sampling the copies $\tilde{X}^{(m)}$. This issue leads us to our next question: are there any settings where we can improve the result of Theorem 6.4.3, and allow for a larger value of σ?

6.4.2 Special case: sparse structure

We next turn to the special case where, due to the constraints imposed on the estimate $\hat{\theta}$, we can assume that the error $\hat{\theta} - \theta_0$ is likely to be sparse, relative to some basis. We will see that, in this setting, the upper bound on Type I error given in Theorem 6.4.3 can be improved to account for the lower effective dimension of $\hat{\theta}$, and that we are therefore free to use a substantially larger value of σ in the constrained aCSS procedure—leading downstream to higher power and easier computation.

To formalize this idea, consider a fixed set of vectors $v_1, \dots, v_p \in \mathbb{R}^d$. We are interested in settings where the solution $\hat{\theta}$ to the perturbed constrained maximum likelihood estimation problem (6.3) is likely to lie in the span of a small subset of v_i's. To motivate this setting, we can revisit several examples that we considered in Section 6.3.1:

- Sparsity: in a setting where we believe θ_0 is sparse, we might use an ℓ_1 constraint for the optimization problem, requiring $\|\theta\|_1 \leq C$, which is likely to lead to a solution $\hat{\theta}$ that is sparse as well. In this setting, we can take $p = d$ and choose the set of vectors to be the canonical basis, i.e., $v_i = \mathbf{e}_i$ for $i \in [d]$, reflecting our belief that the error $\hat{\theta} - \theta_0$ will itself be sparse.

- Locally constant signal: if we believe θ_0 is locally constant, we might choose the constraint $\sum_{i=1}^{d-1} |\theta_i - \theta_{i+1}| \leq C$. This constraint often leads to solutions $\hat{\theta}$ that are piecewise constant, with $\hat{\theta}_i = \hat{\theta}_{i+1}$ for many indices $i \in [d-1]$, and therefore the error $\hat{\theta} - \theta_0$ will also be piecewise constant. Consequently, we can take $p = d$, and choose $v_i = \mathbf{e}_1 + ... + \mathbf{e}_i$ for $i \in [d]$. (This choice of vectors $\{v_i\}$ means that, for any $w \in \mathbb{R}^d$, if w has ℓ many changepoints—that is, $w_i \neq w_{i+1}$ for ℓ many indices i—then w can be written as a linear combination of at most $\ell + 1$ many v_i's.)

- Monotonicity: in a setting where we believe θ_0 is monotone nondecreasing, we might use the isotonic constraint, choosing A and b to constrain $\theta_1 \leq \cdots \leq \theta_d$. This constraint often leads to solutions $\hat{\theta}$ that are piecewise constant, with $\hat{\theta}_i = \hat{\theta}_{i+1}$ for many indices $i \in [d-1]$. If the true parameter θ_0 is also piecewise constant, we therefore again have an error $\hat{\theta} - \theta_0$ that is likely to be piecewise constant, and we can then choose the same v_i's as for the preceding example.

Notation and definitions

For a given choice of vectors $\{v_i\}_{i \in [p]}$, we define

$$\|w\|_{v,0} = \begin{cases} \min \{|S| : S \subseteq [p], w \in \mathrm{span}(\{v_i\}_{i \in S})\}, & w \in \mathrm{span}(\{v_i\}_{i \in [p]}), \\ +\infty, & \text{otherwise.} \end{cases}$$

for any $w \in \mathbb{R}^d$. In other words, $\|w\|_{v,0}$ is the minimum number of vectors v_i needed so that w lies in their span. Note that, despite the notation, the function $w \mapsto \|w\|_{v,0}$ is not a norm. We choose this notation to agree with the commonly used "ℓ_0 norm", $\|w\|_0$, the number of nonzero elements of the vector w; in particular, in the first example where $v_i = \mathbf{e}_i$, $i \in [d]$, we have $\|w\|_{v,0} = \|w\|_0$.

Next, for each $k = 0, \ldots, d$, we define

$$h_v(k) = \mathbb{E}_{Z \sim \mathcal{N}(0,\mathbf{I}_d)} \left[\max_{S \subseteq [p], |S| \leq k} \|\mathcal{P}_{v_S}(Z)\|_2^2 \right],$$

where $\mathcal{P}_{v_S}$ denotes projection to $\text{span}(\{v_i\}_{i \in S})$. This quantity will play an important role in our theory below. We can think of $h_v(k)$ as describing the "effective dimension" of vectors that can be written as a k-sparse combination of the vectors $v_1, \ldots, v_p$. In particular, we can see that for any k, we have $h_v(k) \leq \mathbb{E}_{Z \sim \mathcal{N}(0,\mathbf{I}_d)}[\|Z\|_2^2] = d$. On the other hand, if $k \ll d$, the following result shows that $h_v(k)$ can be substantially smaller:

Lemma 6.4.4. *For each k it holds that $h_v(k) \leq \min\{4k \log(4p/k), d\}$.*

Theoretical result

For this setting, our main result given in Theorem 6.4.3 can be strengthened to the following tighter bound.

Theorem 6.4.5. *Under the notation and assumptions of Theorem 6.4.3, suppose it also holds that*

$$\mathbb{P}\{\|\hat{\theta}(X,W) - \theta_0\|_{v,0} \leq k(\theta_0)\} \geq 1 - \tilde{\delta}(\theta_0),$$

for a fixed set of vectors $v_1, \ldots, v_p \in \mathbb{R}^d$. Then the copies $\tilde{X}^{(1)}, ..., \tilde{X}^{(M)}$ generated by the constrained aCSS procedure are approximately exchangeable with X, satisfying

$$d_{\text{exch}}(X, \tilde{X}^{(1)}, ..., \tilde{X}^{(M)}) \leq 3\sigma r(\theta_0) \cdot \sqrt{\frac{h_v(k(\theta_0))}{d}} + \epsilon(\theta_0) + \delta(\theta_0) + \tilde{\delta}(\theta_0).$$

In particular, this implies that for any predefined test statistic $T : \mathcal{X} \to \mathbb{R}$ and rejection threshold $\alpha \in [0,1]$, the p-value defined in (6.2) *satisfies*

$$\mathbb{P}\left(\text{pval}_T(X, \tilde{X}^{(1)}, ..., \tilde{X}^{(M)}) \leq \alpha\right) \leq \alpha + 3\sigma r(\theta_0) \cdot \sqrt{\frac{h_v(k(\theta_0))}{d}} + \epsilon(\theta_0) + \delta(\theta_0) + \tilde{\delta}(\theta_0).$$

As discussed above, a small value of $k(\theta_0)$ indicates that the error vector, $\hat{\theta}-\theta_0$, typically lies in a region of $\mathbb{R}^d$ that is characterized by a lower effective dimension. As another interpretation, we can think of $k(\theta_0)$ as capturing the effective degrees of freedom in our estimation problem.

The result of Theorem 6.4.5 is strictly stronger than that of Theorem 6.4.3. In particular, Theorem 6.4.3 can be derived as a special case, by taking $v_1 = \mathbf{e}_1, \dots, v_d = \mathbf{e}_d$ and $k(\theta_0) \equiv d$—then the additional condition of Theorem 6.4.5 holds trivially with $\tilde{\delta}(\theta_0) = 0$, and so the two theorems give the same bound (since $h_v(d) = d$). On the other hand, if the constrained estimation problem exhibits sparsity relative to the chosen set of vectors $\{v_i\}$, we may be able to choose a value $k(\theta_0) \ll d$ that allows for a low value of $\tilde{\delta}(\theta_0)$; in this setting, $h_v(k(\theta_0)) \ll d$ by Lemma 6.4.4, and consequently, we see that we can afford to choose a much larger value of the perturbation noise parameter σ while still retaining approximate Type I error control. Of course, to have $k(\theta_0) \ll d$ (or equivalently, $h_v(k(\theta_0)) \ll d$), we need to choose a suitable set $\{v_i\}$ that corresponds well to the structure induced by the constraints $A\theta \leq b$, as in the examples given above.

Remark 6.4.6. *As we will see in the proof, the result of Theorem 6.4.5 holds even if we replace Assumption 6.4.2 with a weaker condition: defining*

$$\Theta_0 = \{\theta \in \Theta : \|\theta - \theta_0\|_2 \leq r(\theta_0), \|\theta - \theta_0\|_{v,0} \leq k(\theta_0)\},$$

and writing $\theta_t = (1-t)\theta_0 + t\theta$ for any θ, it suffices to assume

$$\mathbb{E}_{\theta_0}\left[\sup_{\theta \in \Theta_0, t \in [0,1]} \Big((\theta - \theta_0)^\top \left(H(\theta_t) - H(\theta_t; X)\right)(\theta - \theta_0)\Big)_+\right] \leq \epsilon(\theta_0),$$

and

$$\log \mathbb{E}_{\theta_0}\left[\exp\left\{\sup_{\theta \in \Theta_0, t \in [0,1]} \Big((\theta - \theta_0)^\top \left(H(\theta_t; X) - H(\theta_t)\right)(\theta - \theta_0)\Big)_+\right\}\right] \leq \epsilon(\theta_0).$$

in place of conditions (6.11) *and* (6.12), *respectively. That is, we only need to establish concentration of the error in the Hessian along directions $\theta - \theta_0$ that have sparse structure with respect to the chosen vectors $\{v_i\}$, which may be a much more realistic condition in high-dimensional settings.*

6.4.3 Special case: Gaussian linear model

In this section, we turn to another setting where the scaling of our result has a much more favorable dependence on dimension d, for the special case of a Gaussian linear model. Unlike the result in Theorem 6.4.5 above, here we do not need to assume an underlying sparse structure.

For this special case, we assume that the parametric family $\{P_\theta\}$ is given by

$$P_\theta : \ X \sim \mathcal{N}(Z\theta, \nu^2 \mathbf{I}_n) \tag{6.13}$$

where both the covariate matrix $Z \in \mathbb{R}^{n\times d}$ and the variance $\nu^2 > 0$ are fixed and known. This model is parametrized by the coefficient vector, $\theta \in \Theta = \mathbb{R}^d$. In this setting, as described earlier in Section 6.2.1, co-sufficient sampling (CSS) can be directly applied to sample copies $\tilde{X}^{(m)}$ that are *exactly* exchangeable with X. Concretely, we can consider the sufficient statistic $\mathcal{P}_Z X$, where $\mathcal{P}_Z \in \mathbb{R}^{n\times n}$ denotes the projection matrix to the column span of Z, and sample the copies as

$$\tilde{X}^{(m)} \mid \mathcal{P}_Z X \overset{\text{iid}}{\sim} \mathcal{N}(\mathcal{P}_Z X, \nu^2 \mathcal{P}_Z^{\perp}).$$

Then, under the null, $(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)})$ is exchangeable, and so the p-value defined in (6.2) is *exactly* valid for any test statistic T.

In a low-dimensional regime where $n > d$, the copies $\tilde{X}^{(m)}$ are distinct from X, and the resulting test can have high power against the alternative for a suitably chosen statistic

T. However, in the high-dimensional setting with $d \geq n$, we will have $\mathcal{P}_Z = \mathbf{I}_n$, leading to copies $\tilde{X}^{(m)}$ that are identical to X and, therefore, a powerless test. In the high-dimensional setting, therefore, we turn to aCSS as a practical alternative that can offer nontrivial power, while sacrificing some Type I error control.

The challenge for applying aCSS is that, as we are in a high-dimensional setting, the estimator $\hat{\theta}$ may have low accuracy—but we need a tight bound $r(\theta_0)$ on its error in order to achieve approximate Type I error control. In many settings, the accuracy of the estimator $\hat{\theta}$ will be greatly improved by adding constraints that reflect structure in the problem (e.g., an ℓ_1 constraint if we believe θ_0 is sparse), and so we would expect that constrained aCSS can offer a strong advantage in this setting.

However, the power of the method will rely on being able to choose a sufficiently large value of σ in the implementation. We are therefore motivated to develop a theoretical guarantee that is stronger than the general result of Theorem 6.4.3, so that we can choose a higher value of σ and, consequently, achieve higher power. We will now see that the Gaussian case offers both computational and theoretical advantages.

First, we will assume that R is chosen to ensure that the loss has strongly positive definite Hessian, i.e.,

$$\frac{1}{\nu^2} Z^\top Z + \nabla^2_\theta R(\theta) \succ c\mathbf{I}_d \text{ for all } \theta \in \mathbb{R}^d, \text{ for some } c > 0. \tag{6.14}$$

For example, if $n \geq d$ and Z has full rank d, then this holds with $R(\theta) \equiv 0$. More generally, for any d, n and any Z, a ridge penalty $R(\theta) = \frac{\lambda_{\text{ridge}}}{2}\|\theta\|_2^2$ (for some positive penalty parameter $\lambda_{\text{ridge}} > 0$) will ensure that this condition holds.

Then $\hat{\theta}$ is defined by the optimization problem

$$\hat{\theta} = \hat{\theta}(X, W) = \text{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2\nu^2}\|X - Z\theta\|_2^2 + R(\theta) + \sigma W^\top \theta \ : \ A\theta \leq b \right\},$$

and we compute the gradient as

$$\hat{g} = \frac{1}{\nu^2} Z^\top (Z\hat{\theta} - X) + \nabla_\theta R(\hat{\theta}) + \sigma W.$$

Note that, by our assumptions on R, this optimization problem is guaranteed to have a unique minimizer, and moreover, this minimizer is guaranteed to be a SSOSP. In other words, we can assume that the event $X \in \mathcal{X}_{\hat{\theta},\hat{g}}$ holds almost surely. Then, applying Lemma 6.3.3, we can compute the distribution $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ as

$$\mathcal{N}\left(Z\hat{\theta} + \frac{d}{\sigma^2} \left(\mathbf{I}_n + \frac{d}{\sigma^2 \nu^2} ZZ^\top \right)^{-1} Z(\nabla_\theta R(\hat{\theta}) - \hat{g}), \nu^2 \left(\mathbf{I}_n + \frac{d}{\sigma^2 \nu^2} ZZ^\top \right)^{-1} \right). \tag{6.15}$$

This means that it is possible to draw the copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ directly as i.i.d. draws from $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$.

Next we turn to our theoretical guarantee, which shows an $O(\sqrt{d})$ improvement in the excess Type I error for the Gaussian case.

Theorem 6.4.7. *Consider the Gaussian linear model* (6.13)*, and assume that $R(\theta)$ is chosen so that condition* (6.14) *is satisfied. Assume also that $\mathbb{P}\{\|\hat{\theta}(X, W) - \theta_0\|_2 \leq r(\theta_0)\} \geq 1 - \delta(\theta_0)$. Then the copies $\tilde{X}^{(1)}, ..., \tilde{X}^{(M)}$ generated by the constrained aCSS procedure are approximately exchangeable with X, satisfying*

$$d_{\text{exch}}(X, \tilde{X}^{(1)}, ..., \tilde{X}^{(M)}) \leq \frac{\sigma}{2\sqrt{d}} r(\theta_0) + \delta(\theta_0).$$

In particular, this implies that for any predefined test statistic $T : \mathcal{X} \to \mathbb{R}$ and rejection threshold $\alpha \in [0, 1]$, the p-value defined in (6.2) *satisfies*

$$\mathbb{P}\left(\text{pval}_T(X, \tilde{X}^{(1)}, ..., \tilde{X}^{(M)}) \leq \alpha \right) \leq \alpha + \frac{\sigma}{2\sqrt{d}} r(\theta_0) + \delta(\theta_0).$$

The Type I error inflation described above offers an improvement by a factor of $O(\sqrt{d})$ in terms of dependence on σ, when compared to Theorem 6.4.3. In other words, we see that we are free to choose a substantially larger σ in this Gaussian setting to increase power without losing the guarantee of approximate Type I error control.

6.5 Generalization of linear constraint: ℓ_1 penalty

Thus far, we have considered settings where the estimator $\hat{\theta}$ is obtained via a constrained optimization problem. Section 6.4 shows that the constraints introduced can improve the estimation of unknown parameters, thereby leading to a tighter bound on Type I error control. One important example is placing a bound on $\|\theta\|_1$ to encourage sparsity, a technique that is popular in high-dimensional settings. However, in many statistical applications, it is more common—and more effective—to use a ℓ_1 penalty rather than a constraint. Therefore, in this section, we will consider a ℓ_1-penalized, rather than constrained, form of aCSS.

We consider replacing the constrained optimization problem

$$\hat{\theta}_C = \text{argmin}_{\theta\in\Theta}\{(\theta; X, W) : \|\theta\|_1 \leq C\}$$

with its penalized version,

$$\hat{\theta}_\lambda = \text{argmin}_{\theta\in\Theta}\{(\theta; X, W) + \lambda\|\theta\|_1\}, \tag{6.16}$$

(i.e., the lasso [Tibshirani, 1996], but with an added perturbation term due to W). The penalized and constrained forms of the optimization problem have a natural correspondence—for ℓ_1 regularization, each constrained solution $\hat{\theta}_C$ corresponds to some penalized solution $\hat{\theta}_\lambda$ for some data-dependent λ, and vice versa. However, in a statistical analysis, these two versions of the problem often behave very differently: for ℓ_1 regularization, the fact that the correspondence between C and λ is data-dependent means that theoretical results obtained for $\hat{\theta}_\lambda$ at a fixed λ do not transfer over to a theoretical guarantee for $\hat{\theta}_C$ for a fixed C, and vice versa. Therefore, proper modification is needed for the ℓ_1-penalized aCSS.

Before state the modified method, we first define SSOSP for the penalized problem. For $\theta \in \mathbb{R}^d$, we will write $S(\theta) = \{j \in [d] : \theta_j \neq 0\}$ to denote the support of θ.

Definition 6.5.1 (SSOSP for the ℓ_1-penalized problem)**.** *A parameter $\theta \in \Theta$ is a strict second-order stationary point (SSOSP) of the optimization problem* (6.16) *if it satisfies all of the following:*

1. *First-order necessary conditions, i.e., Karush–Kuhn–Tucker (KKT) conditions:*

$$\nabla(\theta; X, W) + \lambda s = 0, \quad \text{where } \begin{cases} s_j = \text{sign}(\theta_j), & j \in S(\theta), \\ s_j \in [-1, 1], & j \notin S(\theta). \end{cases}$$

2. *Second-order sufficient condition:*

$$\nabla^2_\theta(\theta; X, W)_{S(\theta)} \succ 0,$$

where for a matrix $M \in \mathbb{R}^{d\times d}$ and a nonempty subset $J \subseteq [d]$, $M_J \in \mathbb{R}^{|J|\times|J|}$ denotes the submatrix of M restricted to row and column subsets J. That is, the Hessian $\nabla^2_\theta(\theta; X, W)$ is strictly positive definite when restricted to the support of θ.

6.5.1 The conditional density in the penalized case

Next we compute the conditional density of X given $(\hat{\theta}, \hat{g})$. We will see that this calculation looks quite similar to the constrained case (which was addressed in Lemma 6.3.3).

Lemma 6.5.2 (Conditional density for the ℓ_1-penalized case)**.** *Suppose Assumption 6.3.2* *holds. Fix any $\theta_0 \in \Theta$ and let $(X, W, \hat{\theta}, \hat{g})$ be drawn from the joint model*

$$\begin{cases} X \sim P_{\theta_0}, \\ W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d), \\ \hat{\theta} = \hat{\theta}(X, W) \\ \hat{g} = \hat{g}(X, W) = \nabla_\theta(\hat{\theta}; X, W). \end{cases}$$

Let $S \subseteq [d]$. Assume that the event that $\hat{\theta}(X, W)$ is a SSOSP of (6.16) *with support $S(\hat{\theta}(X, W)) = S$ has positive probability. Then, conditional on this event, the conditional distribution of $X|\hat{\theta}, \hat{g}$ has density*

$$p_{\theta_0}(\cdot|\hat{\theta}, \hat{g}) \propto f(x; \theta_0) \exp\left\{ -\frac{\|\hat{g} - \nabla_\theta(\hat{\theta}; x)\|_2^2}{2\sigma^2/d} \right\} \det\left(\nabla_\theta^2(\hat{\theta}; x)_S\right) \mathbb{1}_{x \in \tilde{\mathcal{X}}_{\hat{\theta},\hat{g}}} \tag{6.17}$$

with respect to the base measure $\nu_{\mathcal{X}} \times$Leb, and

$$\tilde{\mathcal{X}}_{\theta,g} = \left\{ x \in \mathcal{X} : \text{ for some } w \in \mathbb{R}^d,\, \theta = \hat{\theta}(x, w) \text{ is a SSOSP of (6.16), and } g = \nabla(\theta; x, w) \right\}.$$

Comparing to the analogous result given in Lemma 6.3.3 for the constrained case, we see that the only difference is in the det$(\cdot)$ term: the density involves the determinant of a different matrix (namely, $U_{\mathcal{I}}^\top \nabla_\theta^2(\hat{\theta}; x) U_{\mathcal{I}}$ in the constrained case, and $\nabla_\theta^2(\hat{\theta}; x)_S$ in the penalized case). This is not merely a difference in notation: the matrices will actually have different dimension in the ℓ_1-constrained and ℓ_1-penalized settings, because under the constrained setting, if we know the support is S, the solution $\hat{\theta}$ effectively has $|S|-1$ degrees of freedom (due to the ℓ_1 constraint which specifies the sum of the terms), in contrast to $|S|$ for the ℓ_1-penalized setting.

6.5.2 The aCSS method in the penalized case

To implement an ℓ_1-penalized version of aCSS, we can modify the constrained aCSS method in a straightforward way: we simply replace the constrained optimization problem (6.3) with the ℓ_1-penalized optimization problem (6.16), and then proceed as before, using our new calculation for the conditional density as given in Lemma 6.5.2. In particular, the copies $\tilde{X}^{(m)}$ will be sampled as

$$(\tilde{X}^{(1)}, \dots, \tilde{X}^{(M)}) \mid (X, \hat{\theta}, \hat{g}) \sim \tilde{P}_M(\cdot; X, \hat{\theta}, \hat{g})$$

where $\{\tilde{P}_M(\cdot; x, \theta, g)\}$ is required to satisfy (6.10), the same property as before, but now relative to the conditional density $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ calculated as

$$p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}) \propto f(x; \hat{\theta}) \cdot \exp\left\{ -\frac{\|\hat{g} - \nabla_\theta(\hat{\theta}; x)\|_2^2}{2\sigma^2/d} \right\} \cdot \det\left(\nabla^2_\theta(\hat{\theta}; x)_{S(\hat{\theta})} \right) \cdot \mathbb{1}_{x \in \mathcal{X}_{\hat{\theta}, \hat{g}}}. \tag{6.18}$$

As a special case, if computationally feasible, we can choose

$$\tilde{P}_M(\cdot; x, \hat{\theta}, \hat{g}) = p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}) \times \cdots \times p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}),$$

i.e., sampling the copies i.i.d. from the conditional density $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ defined in (6.18).

Formally, the algorithm is defined as follows. The bold text highlights the only modifications in the algorithm, relative to constrained aCSS.

ℓ_1-penalized aCSS algorithm:

1. Observe data $X \sim P_{\theta_0}$.
2. Draw noise $W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$.
3. **Solve for an ℓ_1-penalized perturbed MLE $\hat{\theta} = \hat{\theta}(X, W)$ as in** (6.16)**.**

   Compute the corresponding gradient $\hat{g} = \hat{g}(X, W)$ as in (6.4).

4. If $\hat{\theta}$ is not a SSOSP of (6.3), then set $\tilde{X}^{(1)} = \cdots = \tilde{X}^{(M)} = X$. Otherwise, sample copies $(\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \mid (X, \hat{\theta}, \hat{g}) \sim \tilde{P}_M(\cdot; X, \hat{\theta}, \hat{g})$, where $\tilde{P}_M$ is chosen to satisfy property (6.10) **relative to the conditional density** $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ **as computed in** (6.18)**.**
5. Compute the p-value defined in (6.2) for our choice of test statistic T.

In contrast to the typical challenges for translating results between the constrained and penalized form of a regularized estimation problem, in the context of aCSS, both the conditional density in Lemma 6.5.2 and our next result establish that the exact same results can be obtained for the ℓ_1-penalized case. This unusually favorable behavior is due to the fact that aCSS operates conditionally on the solution $\hat{\theta}$—effectively, once we condition on $\hat{\theta}$, we no longer face the challenge of the data-dependent correspondence between the penalty parameter λ versus the constraint parameter C, since both values are revealed by $\hat{\theta}$ itself.

Theorem 6.5.3. *The results of Theorems 6.4.3, 6.4.5, and 6.4.7 all hold for the ℓ_1-penalized form of aCSS in place of constrained aCSS.*

In the context of utilizing the ℓ_1 penalty, it is commonly the case that the parameter is high-dimensional and sparse. This naturally directs our attention towards Theorem 6.4.5, which offers the most relevant insights for this scenario. Specifically, we can select the set of vectors $\{v_i\}$ as the canonical basis $\{\mathbf{e}_i\}_{i=1,\ldots,d}$. Then we have $\|w\|_{v,0} = \|w\|_0$ (i.e., the cardinality of the support of w). The result of Theorem 6.4.5 then gives a much stronger bound on the excess Type I error rate, as long as we can assume that

$$\|\hat{\theta} - \theta_0\|_0 \leq k(\theta_0)$$

holds with high probability. This is very favorable for the ℓ_1 penalized setting: if θ_0 itself is sparse, then the sparsity of $\hat{\theta}$ (which is ensured by the ℓ_1 penalty) means that the difference $\hat{\theta} - \theta_0$ will also be sparse.

6.6 Numerical experiments

In this section, we will study the performance of aCSS with regularization on three simulated examples.[2] The first, Example 1, is a Gaussian mixture model, which showcases a scenario where constraints on the parameters being estimated are essential to ensure the existence of a well-defined MLE. In the remaining examples, Example 2 (isotonic regression) and Example 3 (sparse regression), we shift our focus to a high-dimensional Gaussian linear model, where the imposition of suitable constraints or penalties can allow for accurate estimation despite high dimensionality.

6.6.1 Necessary constraints: the Gaussian mixture model

In this section, we will examine the Gaussian mixture model example, where constraints are needed for ensuring the existence of a well-defined MLE.

Example 1 (Gaussian mixture model)**.** *Suppose we observe data from the Gaussian mixture model with a known number of components J,*

$$X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} \sum_{j=1}^{J} \pi_j \mathcal{N}(\mu_j, \sigma_j^2),$$

where $\{\pi_j\}_{j\in[J]}$ are the weights on the components, with $\pi_j > 0$ and $\sum_j \pi_j = 1$. The family of distributions $\{P_\theta\}_{\theta\in\Theta}$ is parameterized by $\theta = (\pi_1, ..., \pi_{J-1}, \mu_1, \sigma_1, ..., \mu_J, \sigma_J) \in \Theta$ where

$$\Theta = \{t \in \mathbb{R}_+^{J-1} : \sum_i t_i < 1\} \times (\mathbb{R} \times \mathbb{R}_+)^J.$$

Consequently we have $\Theta \subseteq \mathbb{R}^d$ with $d = 3J - 1$. The density of P_θ, the distribution on the

[2] Code for reproducing all experiments is available at `http://rinafb.github.io/code/reg_acss.zip`.

data $X = (X_1, \dots, X_n)$, is thus given by

$$f(x;\theta) = \prod_{i=1}^{n} \sum_{j=1}^{J} \pi_j \phi(x_i; \mu_j, \sigma_j^2), \tag{6.19}$$

where $\phi(\cdot; \mu, \sigma^2)$ is the density of the normal distribution with mean μ and variance σ^2.

Why is constrained aCSS useful for this example? The Gaussian mixture model does not possess straightforward, compact sufficient statistics due to the presence of unobserved latent variables (i.e., identifying which of the J components corresponds to the draw of each data point X_i). Any sufficient statistic would reveal essentially all the information about the data X. However, if we attempt to apply aCSS (without constraints), we are faced with a fundamental challenge: the MLE does not exist for this model, because the likelihood approaches infinity if, for any component j, we take $\mu_j = X_i$ for some observation $i \in [n]$ and take $\sigma_j \to 0$. To prevent this divergence of the likelihood, one can impose a lower bound on the component variances, requiring $\sigma_j \geq c$ for each $j \in [J]$, where $c > 0$ is some small constant. Under this restriction, it can be shown that MLE is strongly consistent if the true parameter lies within the restricted parameter space Tanaka and Takemura [2006]. Then the constrained aCSS framework is indeed suitable when generating sampling copies in the context of this example. As we will show in Appendix E.3, for an appropriately-chosen initial estimator this example satisfies Assumptions 6.3.2, 6.4.1, and 6.4.2 with $r(\theta_0) = O(\sqrt{\log n/n})$, $\delta(\theta_0) = O(n^{-1})$, and $\epsilon(\theta_0) = O(\sqrt{\frac{\log^3 n}{n}})$, as long as we assume $(\mu_1)_0 \neq (\mu_2)_0$, i.e., the two components have distinct means under the true parameter θ_0. Therefore, Theorem 6.4.3 implies that constrained aCSS will have approximate Type I error control for this example.

Simulation: setting

We next examine the empirical performance of constrained aCSS for the Gaussian mixture model (Example 1). For this setting, we will compare the null hypothesis of a Gaussian mixture model with $J = 2$ components, against an alternative where there are more (specifically, 3) components. The setup of the simulation is summarized as follows:

- To generate data, we take $n = 200$, and draw the data points $X_1, \dots, X_n$ from a mixture of Gaussians

$$\pi_0 \mathcal{N}(0, 0.01) + \frac{1 - \pi_0}{2} \mathcal{N}(0.4, 0.01) + \frac{1 - \pi_0}{2} \mathcal{N}(-0.4, 0.01).$$

- Our null hypothesis is a mixture of *two* Gaussians (i.e., a density of the form (6.19) with $J = 2$). The data generating distribution above therefore corresponds to the null hypothesis (6.19) with parameter

$$\theta_0 = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2) = (0.5,\ 0.4,\ 0.1,\ -0.4,\ 0.1)$$

  in the case that $\pi_0 = 0$, while if $\pi_0 > 0$ then the null hypothesis is not true.

- The test statistic T (used both for aCSS and for the oracle) is chosen as the decrease in total within-cluster sum of squares of the k-means algorithm, when the number of estimated clusters is increased from 2 to 3.

- We enforce $r = 2$ constraints, given by $\sigma_j \geq 0.098$, $j = 1, 2$. (We choose the lower bound to be only slightly smaller than the true value $\sigma_j = 0.1$, so that there will be a reasonable proportion of constraints being active—this way, running our constrained aCSS procedure is meaningfully different than running unconstrained aCSS.) Constrained aCSS is then run with perturbation noise level $\sigma = 8$, and with $M = 300$

copies $\tilde{X}^{(m)}$. The copies are sampled using an MCMC sampler (additional details are provided in Appendix E.4.1).

- We compare constrained aCSS to the oracle method, which uses the same test statistic T but is given full knowledge of the distribution of X under null hypothesis, i.e., $P_{\theta_0} = 0.5\mathcal{N}(0.4, 0.01) + 0.5\mathcal{N}(-0.4, 0.01)$, and can therefore sample the copies $\tilde{X}^{(m)}$ i.i.d. from the known null distribution. (Note that, for this experiment, we cannot compare to unconstrained aCSS, because the unconstrained MLE problem is degenerate, as described above.)

Simulation: results

The results of the simulation are shown in Figure 6.1. We see that the constrained aCSS method is empirically valid as a test of H_0, since the rejection probability when $\pi_0 = 0$ (i.e., when H_0 is true) closely matches the nominal level $\alpha = 0.05$. Of course, the power of constrained aCSS is lower than that of the oracle method, as is expected since the oracle is given knowledge of the true null parameter θ_0; nonetheless, constrained aCSS shows a good increase in power as the signal strength π_0 grows.

6.6.2 High dimensional setting: structured Gaussian linear model

We will now turn to the high-dimensional setting, where the data is distributed according to a Gaussian linear model with dimension $d \geq n$,

$$X \sim \mathcal{N}(Z\theta, \nu^2 \mathbf{I}_n), \quad \text{with } Z \in \mathbb{R}^{n \times d},\ \nu^2 > 0 \text{ known},$$

as in (6.13). The family of distributions $\{P_\theta\}_{\theta \in \Theta}$ is parameterized by $\theta \in \Theta = \mathbb{R}^d$ and has density

$$f(x; \theta) = \frac{1}{(2\pi\nu^2)^{n/2}} e^{-\frac{\|x - Z\theta\|_2^2}{2\nu^2}}.$$

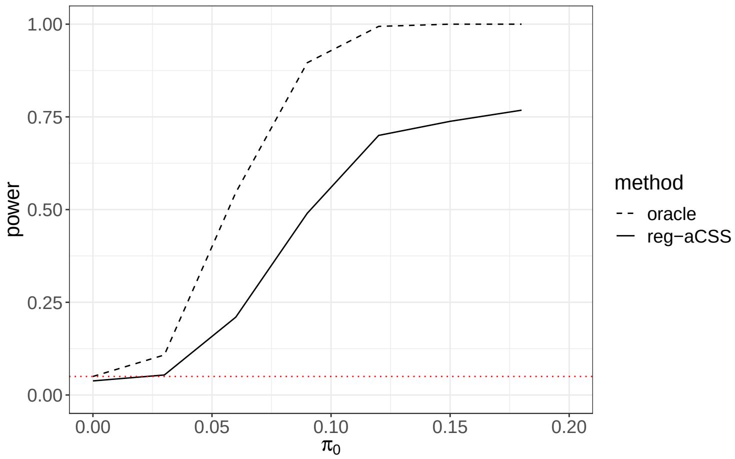


Figure 6.1: Power of the regularized (i.e., constrained) aCSS method, denoted as reg-aCSS in the plot, as compared to the oracle method. The oracle method knows the true parameter and samples (unconditionally) from the simple null. The constrained aCSS method controls the Type I error at the nominal 5% level (red dotted line) under the null. All tests are repeated for 500 independent trials.

In Section 6.4.3, we examined the limitations of CSS testing, which will be powerless for this problem when $d \geq n$, as the copies $\tilde{X}^{(m)}$ will be identically equal to X. We can instead run the aCSS method; however, the results of Barber and Janson [2022] indicate that the inflation in Type I error will scale with our estimation error $\|\hat{\theta} - \theta_0\|_2$, which will in general be large when $d \geq n$, since the estimator $\hat{\theta}$ is computed with an unregularized maximum likelihood estimation problem. (More precisely, aCSS does allow for a *smooth* regularizer $R(\theta)$, such as a ridge penalty; however, it is challenging to achieve accurate estimation in a high-dimensional setting unless we use *nonsmooth* regularization, e.g., the ℓ_1 norm).

In contrast, our proposed version of aCSS allows for constraints (or penalties) that allow us to achieve an accurate estimator $\hat{\theta}$, and consequently low Type I error, in the high-dimensional setting. We now consider two specific examples where the application of appropriate regularization assists in the estimation process.

Example 2 (Isotonic regression)**.** *In the isotonic regression model, we are given a noisy* *observation $X \in \mathbb{R}^n$ of some monotone increasing signal $\theta_0 \in \mathbb{R}^n$ with*

$$(\theta_0)_1 \leq \cdots \leq (\theta_0)_n.$$

If the noise is Gaussian, with $X \sim \mathcal{N}(\theta_0, \nu^2 \mathbf{I}_n)$, then this model is a special case of the Gaussian linear model with $d = n$ and $Z = \mathbf{I}_n$.

To run constrained aCSS, the perturbed isotonic (least squares) regression is given by

$$\hat{\theta}_{\text{iso}} = \arg\min_{\theta \in \mathbb{R}^n} \{(\theta; X, W) \ : \ \theta_1 \leq \cdots \leq \theta_n\}$$

to estimate the underlying signal. Zhang [2002] demonstrated that the isotonic least squares estimator (LSE), which is given by minimizing $\|\theta - X\|_2$ subject to the constraints $\theta_1 \leq \cdots \leq \theta_n$, has an error rate scaling as $\|\hat{\theta} - \theta_0\|_2 = O(n^{1/6})$ (and choosing a sufficiently small σ means that the perturbation will not substantially inflate this rate). This rate matches the minimax rate over the class of monotone and Lipschitz signals [Chatterjee et al., 2015]. Thus, adding the monotonicity constraint will substantially reduce the error $\|\hat{\theta} - \theta_0\|_2$, which can help control the excess Type I error for our setting. In Appendix E.3, we will see that this example satisfies Assumptions 6.3.2, 6.4.1, and 6.4.2 with $r(\theta_0) = O\left(n^{1/6} (\log n)^{1/3}\right)$, $\delta(\theta_0) = 1/n$, and $\epsilon(\theta_0) = 0$, if we choose $\sigma = O(1)$. Therefore, Theorem 6.4.7 implies that constrained aCSS will have approximate Type I error control for this example.

Next, we examine a high-dimensional setting with a sparse parameter.

Example 3 (Sparse regression)**.** *Let $d > n$, and let $Z \in \mathbb{R}^{d \times n}$ be a fixed covariate matrix. We assume the model*

$$X \sim \mathcal{N}(Z\theta, \nu^2 \mathbf{I}_n),$$

for a known noise level ν^2. This model is unidentifiable without further assumptions, but becomes identifiable once we assume θ_0 is sparse—specifically, as long as Z satisfies some *standard conditions (e.g., a restricted eigenvalue assumption). We will assume that the underlying parameter θ_0 is sparse, with*

$$\|\theta_0\|_0 \leq k$$

for some sparsity bound k.

To address the problem of estimating a sparse θ_0 in a linear model, the Lasso estimator [Tibshirani, 1996], which combines the least squares loss with an ℓ_1 penalty, is frequently employed. Under certain conditions, the error rate of the Lasso estimator can be on the order of $O(\sqrt{k\log(d)/n})$ [Bickel et al., 2009, Hastie et al., 2015]. Thus the perturbed Lasso is a suitable candidate for the estimator in this context: for a given penalty level $\lambda > 0$, we define

$$\hat{\theta}_{\text{lasso}} = \arg\min_{\theta\in\mathbb{R}^d}\{(\theta; X, W) + \lambda\|\theta\|_1\}.$$

In Appendix E.3, we will see that this example satisfies Assumptions 6.3.2, 6.4.1, and 6.4.2 with $r(\theta_0) = O(\sqrt{k\log d/n})$, $\delta(\theta_0) = 1/n$, and $\epsilon(\theta_0) = 0$, under suitable conditions. Therefore, Theorem 6.5.3 implies that constrained aCSS will have approximate Type I error control for this example.

Simulation: setting

In this section, we demonstrate the advantage of regularized aCSS in high-dimensional settings. Specifically, we will compare against the (unconstrained) aCSS method of Barber and Janson [2022], to see how adding regularization allows for better estimation—consequently, we can allow a high value of σ without losing (approximate) Type I error control, which in turn leads to higher power.

For the isotonic regression setting (Example 2), we will compare the null hypothesis that X is given by an isotonic signal θ_0 plus Gaussian noise, against the alternative where X

also has dependence on an additional random variable Y. (Equivalently, we can take our covariate matrix Z to be the identity, $Z = \mathbf{I}_d$, with $d = n$.) The setup of the simulation for isotonic regression is as follows:

- To generate data, we take $n = 100$, $\nu = 1$, and set the signal θ_0 as

  $$\theta_0 = (0.1, \ldots, 0.1, 0.2, \ldots, 0.2, \ldots, 1, \ldots, 1),$$

  with each value $0.1, 0.2, 0.3, \ldots, 1$ appearing 10 times. We then generate $X \sim \mathcal{N}(\theta_0, \nu^2 \mathbf{I}_n)$. The additional random vector Y is then drawn as

  $$Y \mid X \sim \mathcal{N}(\beta_0 X, \mathbf{I}_n),$$

  where $\beta_0 \in \{0, 0.05, 0.1, \ldots, 0.5\}$, with $\beta_0 = 0$ corresponding to the null hypothesis. Formally, our null hypothesis is given by assuming that $X \mid Y \sim \mathcal{N}(\theta, \nu^2 \mathbf{I}_n)$ for some $\theta \in \Theta = \mathbb{R}^n$, i.e., that the Gaussian model for X is true even after conditioning on Y. If $\beta_0 \neq 0$, then this null hypothesis does not hold.

- For Barber and Janson [2022]'s aCSS method, $\hat{\theta}$ is computed via perturbed and unconstrained maximum likelihood estimation,

  $$\hat{\theta} = \hat{\theta}_{\mathrm{OLS}} = \mathrm{argmin}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|X - \theta\|_2^2 + \sigma W^\top \theta \right\}.$$

  For our proposed constrained aCSS method, $\hat{\theta}$ is computed with the isotonic constraint,

  $$\hat{\theta} = \hat{\theta}_{\mathrm{iso}} = \mathrm{argmin}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|X - \theta\|_2^2 + \sigma W^\top \theta \ : \ \theta_1 \leq \cdots \leq \theta_n \right\}.$$

  For both methods, we sample the copies $\tilde{X}^{(m)}$ directly from the conditional distribution (6.15) (additional details provided in Appendix E.4.2).

- For the oracle method, we assume oracle knowledge of the parameter θ_0 that defines the null distribution, and sample the copies $\tilde{X}^{(m)}$ i.i.d. from $P_{\theta_0} = \mathcal{N}(\theta_0, \mathbf{I}_n)$.

- For all methods, the test statistic T is given by the absolute value of the sample correlation between X and Y.

For the sparse regression setting (Example 3), we will compare the null hypothesis that $X \mid Z$ follows a (sparse) Gaussian linear model, against the alternative where X also has dependence on an additional random variable Y. The setup of the simulation for sparse regression is as follows:

- To generate data, we set $n = 50$, $d = 100$, $\nu = 1$, and $\theta_0 = (5, 5, 5, 5, 5, 0, ..., 0)$. The covariate matrix $Z \in \mathbb{R}^{n \times d}$ is generated with i.i.d. $\mathcal{N}(0, 1/d)$ entries, and we draw $X \mid Z \sim \mathcal{N}(Z\theta_0, \nu^2 \mathbf{I}_n)$. The random vector $Y \in \mathbb{R}^n$ is then generated with each entry Y_i drawn as
$$Y_i \mid X_i, Z_i \sim \mathcal{N}(\beta_0 X_i + \sum_{j=1}^{5} Z_{i,j}, 1).$$
We consider $\beta_0 \in \{0, 0.1, 0.2, ..., 1\}$ with β_0 corresponding to the setting where $Y \perp\!\!\!\perp X \mid Z$. Formally, our null hypothesis is given by assuming that $X \mid Y, Z \sim \mathcal{N}(Z\theta, \nu^2 \mathbf{I}_n)$ for some $\theta \in \Theta = \mathbb{R}^d$. If $\beta_0 \neq 0$, then this null does not hold.

- For Barber and Janson [2022]'s aCSS method, we will use a ridge regularizer, $R(\theta) = \frac{\lambda_{\text{ridge}}}{2} \|\theta\|_2^2$, for parameter estimation. We define
$$\hat{\theta} = \hat{\theta}_{\text{ridge}} = \text{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2} \|X - Z\theta\|_2^2 + \frac{\lambda_{\text{ridge}}}{2} \|\theta\|_2^2 + \sigma W^\top \theta \right\}.$$
Adding ridge regularization allows for a unique solution $\hat{\theta}$, achieving strict second-order stationarity conditions, to avoid a trivial result where the method achieves zero power (as would be the case if the SSOSP conditions are never satisfied). For our proposed ℓ_1-penalized aCSS method, in order to be more comparable to aCSS, we also add the

regularizer $R(\theta)$. This means that our estimator is given by the elastic net [Zou and Hastie, 2005], incorporating both ℓ_1 and ℓ_2 penalization:

$$\hat{\theta} = \hat{\theta}_{\text{elastic-net}} = \text{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2} \|X - Z\theta\|_2^2 + \frac{\lambda_{\text{ridge}}}{2} \|\theta\|_2^2 + \lambda \|\theta\|_1 + \sigma W^\top \theta \right\}.$$

For both methods, we sample the copies $\tilde{X}^{(m)}$ directly from the conditional distribution (6.15) (additional details provided in Appendix E.4.2).

- For the oracle method, we assume oracle knowledge of the parameter θ_0 that defines the null distribution, and sample the copies $\tilde{X}^{(m)}$ i.i.d. from $P_{\theta_0} = \mathcal{N}(Z\theta_0, \mathbf{I}_n)$.
- For all methods, the test statistic T is given by the absolute value of the estimate of the coefficient on X, when Y is regressed on X, Z with elastic net for penalization on the coefficients on Z—specifically, the fitted coefficient $\hat{\beta}_X$ in the optimization problem

$$(\hat{\beta}_X, \hat{\beta}) = \text{argmin}_{\beta_X, \beta} \left\{ \frac{1}{2} \|Y - X\beta_X - Z\beta\|_2^2 + \frac{3}{2} \|\beta\|_2^2 + 7\|\beta\|_1 \right\}.$$

Simulation: results

Next, we turn to the results of this simulation. In Figure 6.2, we show the power of the methods for isotonic regression (left) and sparse regression (right). We see that aCSS (in its original unconstrained form as proposed by Barber and Janson [2022]) quickly loses Type I error control as σ increases—this is exactly as expected from the theory, since the excess Type I error rate is characterized by a term scaling as $\sigma r(\theta_0)$, where $r(\theta_0)$ bounds the estimation error $\|\hat{\theta} - \theta_0\|_2$ and therefore is high in the unconstrained setting. This means that, to maintain (approximate) Type I error control with aCSS, we would need to use a small value of σ, which in turn leads to low power under the alternative. On the other hand, for our proposed methods—constrained aCSS in the isotonic example, and ℓ_1-penalized aCSS in the sparse example—we see that approximate Type I error control is well maintained even for

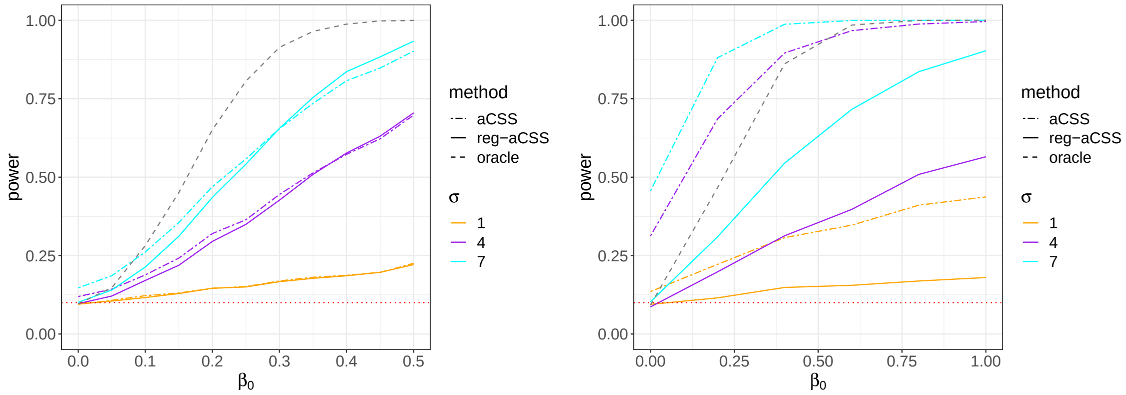


Figure 6.2: Power of aCSS, regularized (i.e., constrained or penalized) aCSS (denoted as reg-aCSS in the plot), and the oracle method, for isotonic regression (left) and sparse regression (right), with different values of the parameter σ, over 5000 independent trials. The dotted red line denotes the nominal 10% level (i.e., $\alpha = 0.1$). For both settings, $\beta_0 = 0$ corresponds to the null hypothesis being true.

larger values of σ, which allows for fairly high power without losing validity. Of course, in each case, the power of the oracle method is higher, as the oracle is given access to the true parameter θ_0 for the null distribution.

To better understand the difference in performance in terms of Type I error rate, in Figure 6.3 we show the Type I error as a function of the parameter σ. For both settings, we see that aCSS suffers a rapid increase in Type I error rate, thus necessitating a very small value of σ to maintain validity, while constrained or penalized aCSS maintains Type I error control across a broad range of values of σ. Finally, Figure 6.4 illustrates the issue of Type I error in more detail for the specific choice $\sigma = 7$ for both examples (chosen to be large enough so that the methods can achieve substantial power). This figure shows a highly nonuniform distribution of the p-values for aCSS, in contrast to the approximately uniform distribution for constrained or penalized aCSS.

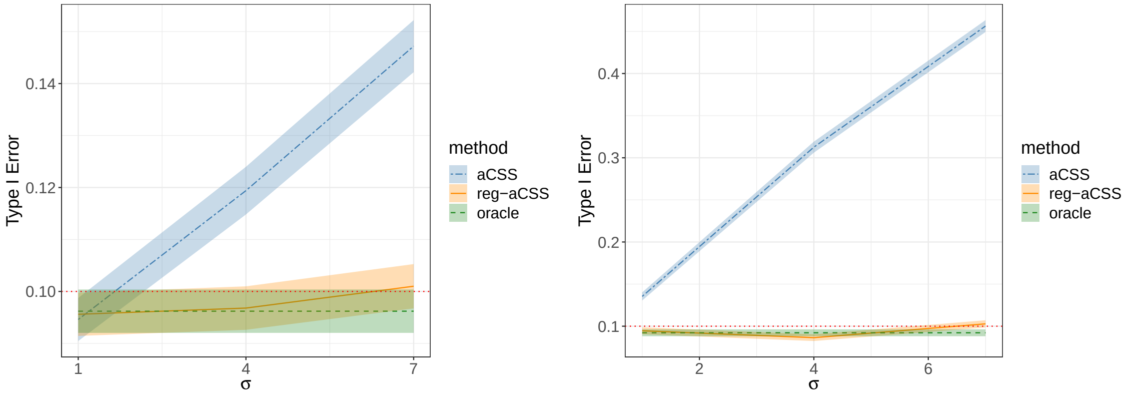


Figure 6.3: Type I error rate of aCSS, regularized (i.e., constrained or penalized) aCSS (denoted as reg-aCSS in the plot), and the oracle method, for isotonic regression (left) and sparse regression (right), with different values of the parameter σ, over 5000 independent trials. The dotted red line denotes the nominal 10% level (i.e., $\alpha = 0.1$). The shaded bands denote standard error for each method.

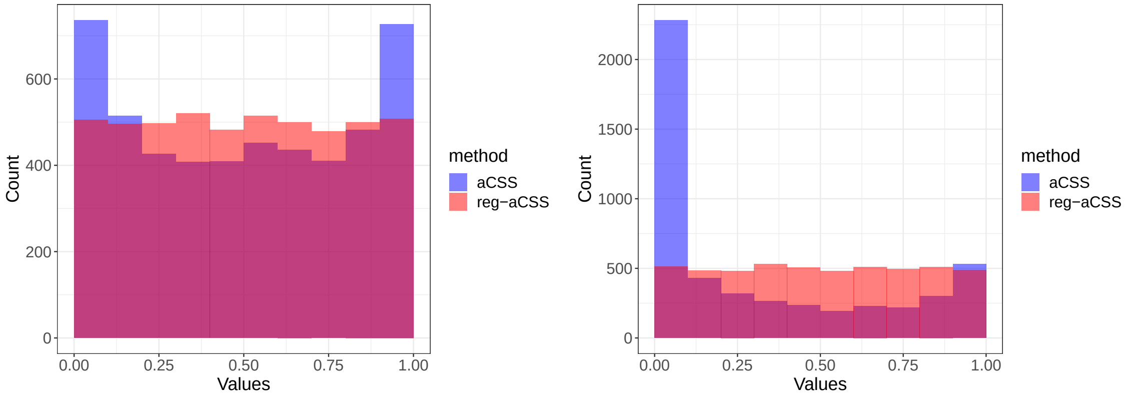


Figure 6.4: Histogram of p-values under the null, for aCSS and for regularized (i.e., constrained or penalized) aCSS, for isotonic regression (left) and sparse regression (right), over 5000 independent trials. The parameter σ is chosen as $\sigma = 7$ for both examples.

6.7 Summary

In this chapter, we discuss how to extend the aCSS algorithm to cases where linear constraints, such as an ℓ_1 constraint or an isotonicity constraint, are applied to enable better accuracy in the estimator $\hat{\theta}$. We also extend to the case of an ℓ_1 penalty (e.g., the lasso). This methodology addresses one of the primary open questions proposed in Barber and Janson [2022], who pose the problem of “Relaxing regularity conditions and extending to high dimensions". We demonstrate that this extension of the aCSS algorithm can accommodate complex estimators $\hat{\theta}$, which may be more stable and accurate in high-dimensional settings. Moreover, we show that the regularized aCSS testing has theoretical guarantees for high dimensions when the estimator exhibits a low-dimensional structure. A remaining challenge is the problem of efficient sampling for aCSS: as for Barber and Janson [2022]'s earlier work in the unconstrained setting, aside from special cases such as a Gaussian linear model, overcoming computational challenges for sampling the copies $\tilde{X}^{(m)}$ will greatly increase the practical utility of this methodology, and remains an important issue to address in future work.

REFERENCES

Alan Agresti. A survey of exact inference for contingency tables. *Stat. Sci.*, 7(1):131–153, 1992.

Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A. Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale clt. In *Conference on Learning Theory*, 2019a.

Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale clt. In *Conference on Learning Theory*, 2019b.

Francis R. Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $\mathcal{O}(1/n)$. In *Advances in Neural Information Processing Systems*, 2013.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Ann. Statist.*, 45(1):77–120, 2017.

Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43:2055–2085, 2015.

Rina Foygel Barber and Lucas Janson. Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling. *Ann. Statist.*, 50(5):2514–2544, 2022.

Rina Foygel Barber, Emmanuel J Candès, and Richard J Samworth. Robust inference with knockoffs. *Ann. Statist.*, 48(3):1409–1431, 2020.

Javad Behboodian. Information matrix for a mixture of two normal distributions. *J. Stat. Comput. Simul.*, 1(4):295–314, 1972.

Rudolf Beran. Prepivoting test statistics: a bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.*, 83(403):687–697, 1988.

István Berkes, Weidong Liu, and Wei Biao Wu. Komlós–major–tusnády approximation under dependence. *The Annals of Probability*, 42(2):794–817, 2014.

Thomas B Berrett, Yi Wang, Rina Foygel Barber, Richard J Samworth, et al. The conditional permutation test for independence while controlling for confounders. *J. R. Stat. Soc. Ser. B*, 82(1):175–197, 2020.

Julian Besag and Peter Clifford. Generalized Monte Carlo significance tests. *Biometrika*, 76 (4):633–642, 1989.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37:1705–1732, 2009.

Julius R. Blum. Approximation methods which converge with probability one. *Ann. Math. Statist.*, 25(2):382–386, 1954.

Léon Bottou. Online learning and stochastic approximations. In *On-line learning and stochastic approximations*, 1998.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

Donald L Burkholder. Sharp inequalities for martingales and stochastic integrals. *Astérisque*, 157(158):75–94, 1988.

Emmanuel J Candès and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

Emmanuel J Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B*, 80(3):551–577, 2018.

Hervé Cardot, Peggy Cénac, Antoine Godichon-Baggioni, et al. Online estimation of the geometric median in hilbert spaces: Nonasymptotic confidence balls. *The Annals of Statistics*, 45(2):591–614, 2017.

Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, 43(4):1774–1800, 2015.

Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534): 708–719, 2021.

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

K. L. Chung. On a Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 25(3):463 – 483, 1954.

M Csörgő and Pal Révész. A new method to prove strassen type laws of invariance principle. 1. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 31(4):255–259, 1975.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, 2019.

Russell Davidson and James G MacKinnon. Improving the reliability of bootstrap tests with the fast double bootstrap. *Comput. Stat. Data Anal.*, 51(7):3259–3281, 2007.

Damek Davis and Dmitriy Drusvyatskiy. High probability guarantees for stochastic convex optimization. In *Conference on Learning Theory*, 2020.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. *Advances in Neural Information Processing Systems*, 2012.

Aryeh Dvoretzky. On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1956.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

Bradley Efron. Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, pages 93–125, 1991.

Bradley Efron. Bayesian inference and the parametric bootstrap. *Ann. Appl. Stat.*, 6(4): 1971–1997, 2012.

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

Uwe Einmahl. Strong invariance principles for partial sums of independent random vectors. *The Annals of Probability*, 15(4):1419–1440, 1987.

Steinar Engen and Magnar Lillegård. Stochastic simulations conditioned on sufficient statistics. *Biometrika*, 84(1):235–240, 1997.

Michael D Ernst. Permutation methods: a basis for exact inference. *Stat. Sci.*, 19:676–685, 2004.

Vaclav Fabian. On asymptotic normality in stochastic approximation. *Ann. Math. Statist.*, 39(4):1327–1332, 1968.

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

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, 2019.

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

Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic analysis of the Ruppert-Polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156: 312–348, 2023.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 2020.

Peter W. Glynn and Ward Whitt. Estimating the asymptotic variance with batch means. *Oper. Res. Lett.*, 10(8):431–435, 1991.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, 2020.

Peter Hall. Theoretical comparison of bootstrap confidence intervals. *Ann. Statist.*, 16: 927–953, 1988.

Peter Hall and Christopher C Heyde. *Martingale limit theory and its application.* Academic press, 2014.

Peter Hall and Tapabrata Maiti. On parametric bootstrap methods for small area prediction. *J. R. Stat. Soc. Ser. B*, 68(2):221–238, 2006.

Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the Forty-Seventh Annual ACM symposium on Theory of computing*, pages 753–760, 2015.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, 2019a.

Nicholas JA Harvey, Christopher Liaw, and Sikander Randhawa. Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. Preprint. Available at arXiv:1909.00843, 2019b.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations.* CRC press, 2015.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(71):2489–2512, 2014.

Matthew Hoffman, Francis R. Bach, and David M. Blei. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, 2010.

Matthew Holland and Kazushi Ikeda. Better generalization with less data using robust gradient descent. In *International Conference on Machine Learning*, 2019.

Dongming Huang and Lucas Janson. Relaxing the assumptions of knockoffs by conditioning. *Ann. Statist.*, 48(5):3021–3042, 2020.

Prateek Jain, Praneeth Netrapalli, Sham M Kakade, Rahul Kidambi, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *The Journal of Machine Learning Research*, 18(1): 8258–8299, 2017.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, 2019.

Jana Janková, Rajen D Shah, Peter Bühlmann, and Richard J Samworth. Goodness-of-fit testing in high dimensional generalized linear models. *J. R. Stat. Soc. Ser. B*, 82(3): 773–795, 2020.

Galin L. Jones, Murali Haran, Brian S. Caffo, and Ronald Neath. Fixed-width output analysis for markov chain monte carlo. *J. Amer. Statist. Assoc.*, 101(476):1537–1547, 2006.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, 2020.

Young Min Kim, Soumendra N Lahiri, and Daniel J Nordman. A progressive block empirical likelihood method for time series. *J. Amer. Statist. Assoc.*, 108(504):1506–1516, 2013.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Yuichi Kitamura et al. Empirical likelihood methods with weakly dependent processes. *Ann. Statist.*, 25(5):2084–2102, 1997.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

János Komlós, Péter Major, and Gábor Tusnády. An approximation of partial sums of independent rv'-s, and the sample df. i. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32:111–131, 1975.

János Komlós, Péter Major, and Gábor Tusnády. An approximation of partial sums of independent rv's, and the sample df. ii. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 34:33–58, 1976.

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $\mathcal{O}(1/t)$ convergence rate for the projected stochastic subgradient method. Preprint. Available at arXiv:1212.2002, 2012.

S. N. Lahiri. Theoretical comparisons of block bootstrap methods. *Ann. Statist.*, 27(1): 386–404, 1999.

S. N. Lahiri. *Resampling methods for dependent data.* Springer Series in Statistics. Springer-Verlag, New York, 2003.

Tze Leung Lai. Stochastic approximation. *Ann. Statist.*, 31(2):391–406, 2003.

Lucien Le Cam. Sufficiency and approximate sufficiency. *Ann. Math. Statist.*, 35:1419–1455, 1964.

Lucien Le Cam. On the information contained in additional observations. *Ann. Statist.*, 2 (4):630–649, 1974.

Lucien Le Cam. *Asymptotic methods in statistical decision theory.* Springer Science & Business Media, 2012.

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 3. Springer, 1986.

Chris Junchi Li, Wenlong Mou, Martin Wainwright, and Michael Jordan. Root-sgd: Sharp nonasymptotics and asymptotic efficiency in a single algorithm. In *Conference on Learning Theory*, 2022a.

Haoran Li, Alexander Aue, Debashis Paul, Jie Peng, and Pei Wang. An adaptable generalization of hotelling's t^2 test in high dimension. *Ann. Statist.*, 48(3):1815–1847, 2020a.

Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3): 50–60, 2020b.

Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis. Statistical inference using sgd. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Xiang Li, Jiadong Liang, Xiangyu Chang, and Zhihua Zhang. Statistical estimation and online inference via local sgd. In *Conference on Learning Theory*, 2022b.

Weidong Liu and Zhengyan Lin. Strong approximation for a class of stationary processes. *Stochastic Processes and their Applications*, 119(1):249–280, 2009.

Lennart Ljung. Analysis of recursive stochastic algorithms. *IEEE. T. Automat. Contr.*, 22 (4):551–575, 1977.

Zhipeng Lou, Wanrong Zhu, and Wei Biao Wu. Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent. *Journal of Machine Learning Research*, 23:1–22, 2022.

Yiling Luo, Xiaoming Huo, and Yajun Mei. Covariance estimators for the root-sgd algorithm in online learning. *arXiv preprint arXiv:2212.01259*, 2022.

Julien Mairal, Francis R. Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11(1):19–60, 2010.

Tucker McElroy, Dimitris N Politis, et al. Computer-intensive rate estimation, diverging statistics and scanning. *Ann. Statist.*, 35(4):1827–1848, 2007.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 2017.

Fabian Mies and Ansgar Steland. Sequential gaussian approximation for nonstationary time series in high dimensions. *Bernoulli*, 29(4):3114–3140, 2023.

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, 2020.

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

Sen Na and Michael W Mahoney. Statistical inference of constrained stochastic optimization via sketched sequential quadratic programming. *arXiv preprint arXiv:2205.13687*, 2022.

S. V. Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, 7(5):745–789, 1979.

Alexander V Nazin, Arkadi S Nemirovsky, Alexandre B Tsybakov, and Anatoli B Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.

Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in Neural Information Processing Systems*, 2014.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, 27(4):538 – 557, 2012.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

Whitney K Newey and James L Powell. Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847, 1987.

Daniel J. Nordman, Helle Bunzel, and Soumendra N. Lahiri. A nonstandard empirical likelihood for time series. *Ann. Statist.*, 41(6):3050–3073, 12 2013.

Magda Peligrad, Hailin Sang, Yunda Zhong, and Wei Biao Wu. Exact moderate and large deviations for linear processes. *Statistica Sinica*, 24:957–969, 2014.

Mark Semenovich Pinsker. The information content of observations, and asymptotically sufficient statistics. *Probl. Peredachi Inf.*, 8(1):45–61, 1972.

Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York, 1999.

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

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*, 2012.

Pratik Ramprasad, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 118(544):2901–2914, 2022.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, 2016.

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Rev.*, 26(2):195–239, 1984.

Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the International Conference on World Wide Web*, 2007.

Emmanuel Rio. Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability*, 22(1):146–163, 2009.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951. doi:10.1214/aoms/1177729586. URL `https://doi.org/10.1214/aoms/1177729586`.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*. Academic Press, 1971.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Jerome Sacks. Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.*, 29(2):373–405, 06 1958.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, 2013.

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

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, 2019.

Kesar Singh. On the asymptotic accuracy of efron's bootstrap. *Ann. Statist.*, 9:1187–1195, 1981.

Radhendushka Srivastava, Ping Li, and David Ruppert. Raptt: An exact two-sample test in high dimensions using random projections. *J. Comput. Graph. Stat.*, 25(3):954–970, 2016.

Michael A Stephens. Goodness-of-fit and sufficiency: Exact and approximate tests. *Meth. Comput. Appl. Probab.*, 14:785–791, 2012.

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

Kentaro Tanaka and Akimichi Takemura. Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when the scale parameters are exponentially small. *Bernoulli*, 12(6):1003 – 1017, 2006.

James W Taylor. Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics*, 6(2):231–252, 2008.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, 58(1):267–288, 1996.

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.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3): 1166–1202, 2014.

Dootika Vats, James M Flegal, and Galin L Jones. Multivariate output analysis for markov chain monte carlo. *Biometrika*, 106(2):321–337, 2019.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. Preprint. Available at arXiv:1011.3027, 2010.

Ziyang Wei, Wanrong Zhu, and Wei Biao Wu. Weighted averaged stochastic gradient descent: Asymptotic normality and optimality. *arXiv preprint arXiv:2307.06915*, 2023.

William J Welch. Construction of permutation tests. *J. Amer. Statist. Assoc.*, 85(411): 693–698, 1990.

Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, 2020.

Chien-Fu Jeff Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1261–1295, 1986.

Wei Biao Wu. Strong invariance principles for dependent random variables. *The Annals of Probability*, 35(6):2294–2320, 2007.

Wei Biao Wu. Recursive estimation of time-average variance constants. *Ann. Appl. Probab.*, 19(4):1529–1552, 2009.

Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. *Advances in Neural Information Processing Systems*, 29, 2016.

Lei Xu and Michael I Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural Comput.*, 8(1):129–151, 1996.

Fan Yang and Rina Foygel Barber. Contraction and uniform convergence of isotonic regression. *Electronic Journal of Statistics*, 13(1):646 – 677, 2019.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Cun-Hui Zhang. Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555, 2002.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B*, 76:217–242, 2014.

Jin-Ting Zhang, Jia Guo, Bu Zhou, and Ming-Yen Cheng. A simple two-sample test in high dimensions based on l_2-norm. *J. Amer. Statist. Assoc.*, 115(530):1011–1027, 2020.

Sixin Zhang, Anna E Choromanska, and Yann LeCun. Deep learning with elastic averaging sgd. *Advances in Neural Information Processing Systems*, 2015.

Weinan Zhang, Tianxiong Zhou, Jun Wang, and Jian Xu. Bid-aware gradient descent for unbiased learning with censored data in display advertising. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7: 2541–2563, 2006.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.

Wanrong Zhu, Zhipeng Lou, Ziyang Wei, and Wei Biao Wu. High confidence level inference is almost free using parallel stochastic optimization. *arXiv preprint arXiv:2401.09346*, 2024.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in Neural Information Processing Systems*, 2010.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

APPENDIX A

APPENDIX FOR CHAPTER 2

A.1 Technical Lemmas and Proofs

We will use $\mathcal{F}_{n,i} = \mathcal{F}_i,\ 1 \leq i < n$ to denote the nested σ-algebra generated by $\{\xi_1, ..., \xi_i\}$. Before we start, we also present a lemma in Polyak and Juditsky [1992] as follows:

Lemma A.1.1. *Choose the step size as $\eta_i = \eta i^{-\alpha}$ with $\eta > 0$ and $0.5 < \alpha < 1$. For a real symmetric positive definite matrix A, define a matrices sequence Y_i^k: $Y_i^i = \mathbf{I}_d$ and for any $k > i$:*

$$Y_i^k = \prod_{j=i+1}^{k} (\mathbf{I}_d - \eta_j A).$$

We also define $\bar{Y}_i^n$ and ϕ_i^n as follows,

$$\bar{Y}_i^n = \eta_i \sum_{k=i}^{n} Y_i^k,\ n \geq i,$$

$$\phi_i^n = A^{-1} - \bar{Y}_i^n.$$

Then $\exists\ 0 < K < \infty$ such that $\forall\ j$ and $i \geq j$

$$||\phi_i^n||_2 \leq K,$$

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} ||\phi_i^n||_2 = 0.$$

Lemma A.1.1 is a simple reduction of Lemma 1 in Polyak and Juditsky [1992]. The term Y_i^k appears frequently in the explicit form of weighted SGD solutions.

A.1.1 Technical Overview and Proof Sketch of the Main Theorem

The error sequence $\delta_i = x_i - x^*$ takes the following form

$$\delta_i = \delta_{i-1} - \eta_i \nabla F(x_{i-1}) + \eta_i \epsilon_i, \ i \geq 1, \tag{A.1}$$

where $\epsilon_i = \nabla F(x_{i-1}) - \nabla f(x_{i-1}, \xi_i)$. Since $\nabla F(x^*) = 0$, by Taylor's expansion of F around x^* we have $\nabla F(x_n) \approx A\delta_n$, which inspires the idea to approximate the general SGD sequence with a corresponding linear sequence.

Consider a linear sequence:

$$\delta_i' = (I - \eta_i A)\delta_{i-1}' + \eta_i \epsilon_i, \ \delta_0' = \delta_0. \tag{A.2}$$

The following lemma shows the asymptotic normality for the weighted average of the linear sequence δ_i'.

Lemma A.1.2. *Let δ_i' be defined in* (C.8)*. Then under the settings in Theorem 2.2.4, for $\tilde{\delta}_n' = \sum_{i=1}^n w_{n,i}\delta_i'$ we have*

$$\sqrt{n}\tilde{\delta}_n' \Rightarrow \mathcal{N}(0, wA^{-1}SA^{-1}),$$

where $w = \lim_{n\to\infty} n\sum_{i=1}^n (w_{n,i})^2$, $A = \nabla^2 F(x^*)$, and $S = \mathbb{E}([\nabla f(x^*, \xi)][\nabla f(x^*, \xi)]^T)$. Furthermore, we can weaken the constraints so that we do not require $\sum_{i=1}^n w_{n,i} = 1$ in this lemma, and the conclusion still holds.

To prove Lemma A.1.2, we decompose $\sqrt{n}\tilde{\delta}_n'$ into four terms:

$$\begin{aligned}\sqrt{n}\tilde{\delta}_n' = \sqrt{n}\sum_{i=1}^n w_{n,i}A^{-1}\epsilon_i + \sqrt{n}\sum_{i=1}^n w_{n,i}Y_0^i\delta_0 \\ + \sqrt{n}\sum_{i=1}^n w_{n,i}a_i^n\epsilon_i + \sqrt{n}\sum_{i=1}^n b_i^n\epsilon_i,\end{aligned} \tag{A.3}$$

where $a_i^n = \sum_{k=i}^n (Y_i^k \eta_i - A^{-1})$, $b_i^n = \sum_{k=i+1}^n (w_{n,k} - w_{n,i}) Y_i^k \eta_i$ and

$$Y_i^k = \prod_{j=i+1}^k (\mathbf{I}_d - \eta_j A), k > i, Y_i^i = \mathbf{I}_d.$$

The last three terms in (A.3) vanish as n goes to infinity. The first term $\sqrt{n}\sum_{i=1}^n w_{n,i} A^{-1}\epsilon_i$ is a linear combination of martingale differences and the following lemma shows that it is asymptotic normal.

Lemma A.1.3 (Martingale difference asymptotic normality)**.** *Under the settings in Theorem 2.2.4,*

$$\sqrt{n}\sum_{i=1}^n w_{n,i} A^{-1}\epsilon_i \Rightarrow \mathcal{N}(0, wA^{-1}SA^{-T}).$$

Once Lemma A.1.2 is established, we can prove Theorem 2.2.4 using the well-known linear approximation technique.

A.1.2 Proof of Theorem 2.2.4

Proof. Recall the error sequence of SGD iterates $\delta_n = x_n - x^*$. It also takes the form $\delta_n = \delta_{n-1} - \eta_n \nabla F(x_{n-1}) + \eta_n \epsilon_n$. The weighted averaged error sequence is $\tilde{\delta}_n = \sum_{i=1}^n w_{n,i}\delta_i$. Since $\sum_{i=1}^n w_{n,i} = 1$, we have $\tilde{\delta}_n = \tilde{x}_n - x^*$. We have also defined the linear error sequence

$$\delta_n' = \delta_{n-1}' - \eta_n A \delta_{n-1}' + \eta_n \epsilon_n,\ \delta_0' = x_0 - x^*,$$

$$\tilde{\delta}_n' = \sum_{i=1}^n w_{n,i}\delta_i'.$$

We claim that Lemma A.1.2 is true, i.e.,

$$\sqrt{n}\tilde{\delta}_n' \Rightarrow \mathcal{N}(0, wA^{-1}SA^{-T}),$$

then it suffices to prove that $\sqrt{n}\tilde{\delta}'_n$ and $\sqrt{n}\tilde{\delta}_n$ are asymptotically equally distributed. Let s_n be the difference between the nonlinear and linear sequence. It also takes the following recursion form:

$$\begin{aligned} s_n = \delta_n - \delta'_n &= \delta_{n-1} - \eta_n \nabla F(x_{n-1}) - (\mathrm{I} - \eta_n A)\delta'_{n-1} \\ &= (\mathrm{I} - \eta_n A)(\delta_{n-1} - \delta'_{n-1}) - \eta_n(\nabla F(x_{n-1}) - A\delta_{n-1}) \\ &= (\mathrm{I} - \eta_n A)s_{n-1} - \eta_n(\nabla F(x_{n-1}) - A\delta_{n-1}). \end{aligned} \tag{A.4}$$

Recall the definition of Y_i^n:

$$Y_i^k = \prod_{j=i+1}^{k} (\mathbf{I}_d - \eta_j A), k > i, Y_i^i = \mathbf{I}_d.$$

We can use Y_i^n to rewrite s_n as

$$s_n = \sum_{i=1}^{n} Y_i^n \eta_i [A\delta_{i-1} - \nabla F(x_{i-1})].$$

Define the weighted average difference between the nonlinear and linear sequence:

$$\tilde{s}_n = \sum_{i=1}^{n} w_{n,i} s_i = \sum_{i=1}^{n} w_{n,i} \sum_{j=1}^{i} Y_j^i \eta_j [A\delta_{j-1} - \nabla F(x_{j-1})].$$

Note that $\sqrt{n}\tilde{\delta}_n = \sqrt{n}\tilde{s}_n + \sqrt{n}\tilde{\delta}'_n$ and $\sqrt{n}\tilde{\delta}'_n \xrightarrow{D} \mathcal{N}(0, wA^{-1}SA^{-T})$. To prove

$$\sqrt{n}\tilde{\delta}_n = \sqrt{n}(\tilde{x}_n - x^*) \Rightarrow \mathcal{N}(0, wA^{-1}SA^{-T}),$$

it is suffice to prove $\sqrt{n}\tilde{s}_n$ converges to 0 in probability.

$$
\begin{aligned}
||\tilde{s}_n||_2 &\leq \sum_{i=1}^{n} w_{n,i} \sum_{j=1}^{i} ||Y_j^i||_2 \eta_j ||A\delta_{j-1} - \nabla F(x_{j-1})||_2 \\
&\lesssim \frac{1}{n} \sum_{j=1}^{n} ||A\delta_{j-1} - \nabla F(x_{j-1})||_2 \eta_j \Big(\sum_{i=j}^{n} ||Y_j^i||_2 \Big) \\
&\lesssim \frac{1}{n} \sum_{j=1}^{n} ||A\delta_{j-1} - \nabla F(x_{j-1})||_2 j^{-\alpha} \big(1 + (j+1)^{\alpha}\big) \qquad \text{(A.5)} \\
&\lesssim \frac{1}{n} \sum_{j=1}^{n} ||A\delta_{j-1} - \nabla F(x_{j-1})||_2 \\
&\lesssim \frac{1}{n} \sum_{j=1}^{n} ||\delta_{j-1}||_2^2.
\end{aligned}
$$

The second inequality is obtained by upper bounding $w_{n,i}$ and exchange the order of summations. The third inequality comes from Lemma A.2 in Zhu et al. [2023]. The last inequality is from Taylor's expansion around x^*. From Lemma 3.2 in Zhu et al. [2023] we know that

$$\mathbb{E}||\delta_{j-1}||_2^2 \lesssim (j-1)^{-\alpha}.$$

So there exists a constant $C > 0$ such that

$$\sum_{j=1}^{n} \frac{1}{\sqrt{j}} \mathbb{E}||\delta_{j-1}||_2^2 \lesssim \sum_{j=1}^{n} \frac{1}{\sqrt{j}} (j-1)^{-\alpha} \lesssim \sum_{j=1}^{n} j^{-0.5-\alpha} \leq C.$$

By Kronecker's lemma,

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathbb{E}||\delta_{j-1}||_2^2 \to 0.$$

As a result, for any fixed $h > 0$,

$$\mathbb{P}(\sqrt{n}||\tilde{s}_n||_2 > h) \leq \mathbb{P}\Big(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} ||\delta_{j-1}||_2^2 > h \Big) \leq \frac{1}{\sqrt{n}h} \mathbb{E} \sum_{j=1}^{n} ||\delta_{j-1}||_2^2 \to 0.$$

Thus we proved that $\sqrt{n}\tilde{s}_n$ converges to 0 in probability, and the theorem is proved.

□

A.1.3 Proof of Lemma A.1.2

Proof. By definition of the linear error term, we have

$$\delta_n' = \prod_{i=1}^{n}(\mathbf{I}_d - \eta_i A)\delta_0 + \sum_{i=1}^{n}\prod_{j=i+1}^{n}(\mathbf{I}_d - \eta_j A)\eta_i \epsilon_i,$$

where the matrix sequence $Y_i^k,\ k \geq i$ is defined as

$$Y_i^k = \prod_{j=i+1}^{k}(\mathbf{I}_d - \eta_j A), k > i, Y_i^i = \mathbf{I}_d.$$

Here we also use the convention that $\prod_{j=n+1}^{n}(\mathbf{I}_d - \eta_j A) = I$. Then the weighted averaged error sequence $\tilde{\delta}_n'$ takes the form:

$$\begin{aligned}
\tilde{\delta}_n' &= \sum_{i=1}^{n} w_{n,i}\prod_{j=1}^{i}(\mathbf{I}_d - \eta_j A)\delta_0 + \sum_{k=1}^{n} w_{n,k}\sum_{i=1}^{k}\prod_{j=i+1}^{k}(\mathbf{I}_d - \eta_j A)\eta_i\epsilon_i \\
&= \sum_{i=1}^{n} w_{n,i}Y_0^i\delta_0 + \sum_{i=1}^{n}\sum_{k=i}^{n} w_{n,k}Y_i^k\eta_i\epsilon_i \\
&= \sum_{i=1}^{n} w_{n,i}Y_0^i\delta_0 + \sum_{i=1}^{n} w_{n,i}\sum_{k=i}^{n} Y_i^k\eta_i\epsilon_i + \sum_{i=1}^{n}\sum_{k=i+1}^{n}(w_{n,k} - w_{n,i})Y_i^k\eta_i\epsilon_i \\
&= \sum_{i=1}^{n} w_{n,i}A^{-1}\epsilon_i + \sum_{i=1}^{n} w_{n,i}Y_0^i\delta_0 + \sum_{i=1}^{n} w_{n,i}(\sum_{k=i}^{n} Y_i^k\eta_i - A^{-1})\epsilon_i + \\
&\quad \sum_{i=1}^{n}\sum_{k=i+1}^{n}(w_{n,k} - w_{n,i})Y_i^k\eta_i\epsilon_i \\
&\triangleq I + II + III + IV
\end{aligned} \tag{A.6}$$

By Lemma A.2 in Zhu et al. [2023],

$$\sum_{k=i+1}^{n} ||Y_i^k||_2 \lesssim (i+1)^{\alpha}.$$

So we have,

$$\lim_{n\to\infty} ||\sqrt{n}\sum_{i=1}^{n} w_{n,i}Y_0^i\delta_0||_2 \lesssim \lim_{n\to\infty}\frac{1}{n}||\sqrt{n}\sum_{i=1}^{n} Y_0^i||_2 \lesssim \lim_{n\to\infty}\frac{1}{\sqrt{n}} = 0.$$

Recall $\phi_i^n = \sum_{k=i}^{n} Y_i^k\eta_i - A^{-1}$. Then by Lemma A.1.1,

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}||\sqrt{n}\sum_{i=1}^{n} w_{n,i}(\sum_{k=i}^{n} Y_i^k\eta_i - A^{-1})\epsilon_i||_2^2 &= \lim_{n\to\infty}\mathbb{E}||\sqrt{n}\sum_{i=1}^{n} w_{n,i}\phi_i^n\epsilon_i||_2^2 \\ &\lesssim \frac{1}{n}\sum_{i=1}^{n}||\phi_i^n||_2^2 \\ &\lesssim \frac{1}{n}\sum_{i=1}^{n}||\phi_i^n||_2 = 0.\end{aligned}$$

The last inequality is because $||\phi_i^n||_2 \leq K$. The result shows that $\sqrt{n}II$ and $\sqrt{n}III$ converge to 0 in L^2 norm. For $\sqrt{n}IV$, let $a_i^n = \sum_{k=i+1}^{n}(w_{n,k} - w_{n,i})Y_i^k\eta_i$. By Lemma A.2 in Zhu et al. [2023], we have

$$||Y_i^k||_2 \leq \exp(-\lambda\sum_{t=i+1}^{k}\eta_t),$$

and

$$||a_i^n||_2 \lesssim \frac{1}{n}\sum_{k=i+1}^{n}||Y_i^k||_2\eta_i \lesssim \frac{1}{n}$$

Under the smoothness condition,

$$\lim_{n\to\infty}\sum_{i=1}^{n}||a_i||_2 \leq \lim_{n\to\infty}\sum_{i=1}^{n}\sum_{k=i+1}^{n}|w_{n,k} - w_{n,i}|\eta_i\exp(-\lambda\sum_{t=i+1}^{k}\eta_t) = 0.$$

As a result,

$$\begin{aligned}\mathbb{E}||\sqrt{n}IV||_2^2 &= \mathbb{E}||\sqrt{n}\sum_{i=1}^n a_i^n\epsilon_i||_2^2 \\ &\lesssim \frac{1}{n}\sum_{i=1}^n ||na_i^n||_2^2 \\ &\lesssim \frac{1}{n}\sum_{i=1}^n ||na_i^n||_2 = \sum_{i=1}^n ||a_i^n||_2 \to 0.\end{aligned}$$

So we only need to show the asymptotic normality of the first term. By Lemma A.1.3 we have

$$\sqrt{n}\sum_{i=1}^n w_{n,i}A^{-1}\epsilon_i \Rightarrow \mathcal{N}(0, wA^{-1}SA^{-1}).$$

Thus we have proved Lemma A.1.2. □

A.1.4 Proof of Lemma A.1.3

Proof. Let $X_{ni} = \sqrt{n}w_{n,i}A^{-1}\epsilon_i$ $(1 \le i \le n)$ denote a martingale difference array. Then we need to prove

$$\sum_{i=1}^n X_{ni} \Rightarrow \mathcal{N}(0, wA^{-1}SA^{-1}).$$

We first check the conditional Lindeberg condition: $\forall r > 0$

$$\begin{aligned}\mathbb{E}_{i-1}[||X_{ni}||^2\mathbb{1}(||X_{ni}|| > r)] &\le \sqrt{\mathbb{E}_{i-1}[||X_{ni}||^4]}\sqrt{\mathbb{E}_{i-1}[\mathbb{1}(||X_{ni}|| > r)^2]} \\ &\le \sqrt{Cn^2(w_i^n)^4\mathbb{E}_{i-1}||\epsilon_i||^4}\sqrt{\mathbb{P}_{i-1}(||X_{ni}|| > r)} \\ &\le \sqrt{\frac{K_1}{n^2}(1+||\delta_{i-1}||^4)}\sqrt{\mathbb{P}_{i-1}(||\epsilon_i||^4 > K_2^4n^2r^4)}\end{aligned} \tag{A.7}$$

The first inequality is Cauchy-Schwarz. The third inequality is from Assumption 2.2.3. By Markov's inequality,

$$\mathbb{P}_{i-1}(||\epsilon_i||^4 > K_2^4n^2r^4) \le \frac{K_3(1+||\delta_{i-1}||^4)}{n^2},$$

We get the following bound

$$\mathbb{E}_{i-1}[||X_{ni}||^2 \mathbb{1}(||X_{ni}|| > r)] \leq \frac{K_4}{n^2}(1 + ||\delta_{i-1}||^4).$$

As a result

$$\sum_{i=1}^n \mathbb{E}_{i-1}[||X_{ni}||^2 \mathbb{1}(||X_{ni}|| > r)] \leq \frac{K_4}{n} + \frac{K_4}{n^2} \sum_{i=1}^n ||\delta_{i-1}||^4. \tag{A.8}$$

By Lemma 3.2 in Zhu et al. [2023], we have

$$\lim_{n\to\infty} \mathbb{E}[\frac{K_4}{n} + \frac{K_4}{n^2} \sum_{i=1}^n ||\delta_{i-1}||^4] = 0.$$

So both sides of equation (A.8) also L^1 converges to 0, which implies the conditional Lindeberg condition: $\forall r > 0$,

$$\sum_{i=1}^n \mathbb{E}[||X_{ni}||^2 \mathbb{1}(||X_{ni}|| > r) | \mathcal{F}_{n,i-1}] \xrightarrow{P} 0.$$

The next step is to show that:

$$\sum_{i=1}^n \mathbb{E}[n(w_{n,i})^2 A^{-1} \epsilon_i \epsilon_i^T A^{-T} | \mathcal{F}_{n,i-1}] \Rightarrow w A^{-1} S A^{-T}.$$

Since $n \lim_{n\to\infty} \sum_{i=1}^n (w_{n,i})^2 = w$,

$$w A^{-1} S A^{-T} = \lim_{n\to\infty} \sum_{i=1}^n n(w_{n,i})^2 A^{-1} S A^{-T}$$

We estimate the difference

$$\Delta = \sum_{i=1}^n n(w_{n,i})^2 A^{-1} S A^{-T} - \sum_{i=1}^n \mathbb{E}[n(w_{n,i})^2 A^{-1} \epsilon_i \epsilon_i^T A^{-T} | \mathcal{F}_{n,i-1}]$$

using Assumption 2.2.3 :

$$
\begin{aligned}
||\Delta||_F &= ||\sum_{i=1}^{n} \mathbb{E}_{i-1}[n(w_{n,i})^2 A^{-1}\epsilon_i\epsilon_i^T A^{-T}] - \sum_{i=1}^{n} n(w_{n,i})^2 A^{-1} S A^{-T}||_F \\
&\leq \sum_{i=1}^{n} n(w_{n,i})^2 ||A^{-1}(\mathbb{E}_{i-1}\epsilon_i\epsilon_i^T - S)A^{-T}||_F \\
&\leq \sum_{i=1}^{n} n(w_{n,i})^2 ||A^{-1}||_F^2 ||\mathbb{E}_{i-1}\epsilon_i\epsilon_i^T - S||_2 \\
&\leq \frac{C}{n}\sum_{i=1}^{n}(||\delta_{i-1}||_2 + ||\delta_{i-1}||_2^2)
\end{aligned}
\tag{A.9}
$$

By Lemma 3.2 in Zhu et al. [2023], we have

$$
\lim_{n\to\infty} \mathbb{E}[\frac{C}{n}\sum_{i=1}^{n}(||\delta_{i-1}||_2 + ||\delta_{i-1}||_2^2)] = 0.
$$

So the Frobenius norm of Δ also L^1 converges to 0. With triangle inequality it implies that

$$
\begin{aligned}
&||\sum_{i=1}^{n} \mathbb{E}_{i-1}[n(w_{n,i})^2 A^{-1}\epsilon_i\epsilon_i^T A^{-T}] - wA^{-1}SA^{-T}||_F \\
\leq &||\Delta||_F + ||wA^{-1}SA^{-T} - \sum_{i=1}^{n} n(w_{n,i})^2 A^{-1}SA^{-T}||_F.
\end{aligned}
$$

And both terms on the right hand side L^1 converge to 0. As a result,

$$
\lim_{n\to\infty} \mathbb{E}\Big| ||\sum_{i=1}^{n} \mathbb{E}_{i-1}[n(w_{n,i})^2 A^{-1}\epsilon_i\epsilon_i^T A^{-T}] - wA^{-1}SA^{-T}||_F \Big| = 0.
$$

Since L^1 convergence implies convergence in probability, we have

$$
||\sum_{i=1}^{n} \mathbb{E}_{i-1}[n(w_{n,i})^2 A^{-1}\epsilon_i\epsilon_i^T A^{-T}] - wA^{-1}SA^{-T}||_F \xrightarrow{P} 0,
$$

which implies

$$\sum_{i=1}^{n} \mathbb{E}_{i-1}[n(w_{n,i})^2 A^{-1}\epsilon_i \epsilon_i^T A^{-T}] \xrightarrow{P} wA^{-1}SA^{-T}$$

at every entries. Then all conditions of Corollary 3.1 in Hall and Heyde [2014] hold. By Theorem 3.3 in Hall and Heyde [2014],

$$\sqrt{n}\sum_{i=1}^{n} X_n^i = \sqrt{n}\sum_{i=1}^{n} w_{n,i}A^{-1}\epsilon_i \Rightarrow \mathcal{N}(0, wA^{-1}SA^{-1}).$$

□

A.1.5 Proof of Corollary 2.3.1

Proof. Recall the definition of $\theta_{n,i}$:

$$\begin{aligned}\theta_{n,i} &= \frac{\gamma+1}{\gamma+i}\prod_{j=i+1}^{n}\frac{j-1}{j+\gamma} \\ &= \frac{\gamma+1}{n}\frac{\Gamma(\gamma+i+1)\Gamma(n+1)}{\Gamma(\gamma+n+1)\Gamma(i+1)}.\end{aligned}$$

Lemma A.1.4. *The weight $w_{n,i} = \theta_{n,i}$ satisfies $\sum_{i=1}^{n} w_{n,i} = 1$, $w_{n,i} \leq (\gamma+1)/n$ and*

$$\lim_{n\to\infty} n\sum_{i=1}^{n}(w_{n,i})^2 = \frac{(\gamma+1)^2}{2\gamma+1}.$$

Now we only have to verify the smoothness condition. We first show that there exists a constant $\tilde{C} = \gamma(\gamma+1)$ such that for all $1 \leq i < n$,

$$|w_{n,i+1} - w_{n,i}| \leq \tilde{C}n^{-2}.$$

Notice for any $n \in \mathbb{N}^+$,

$$|\theta_{n+1,n+1} - \theta_{n,n}| = \frac{(\gamma+1)(\gamma+n) - n(\gamma+1)}{(\gamma+n+1)(\gamma+n)} = \frac{\gamma(\gamma+1)}{(\gamma+n+1)(\gamma+n)} \leq \frac{\gamma(\gamma+1)}{(n+1)^2},$$

and

$$\begin{aligned} |\theta_{n+1,n} - \theta_{n+1,n-1}| &= (1 - \frac{\gamma+1}{\gamma+n})(\theta_{n,n} - \theta_{n,n-1}) \\ &= \frac{\gamma(\gamma+1)}{(\gamma+n)(\gamma+n-1)} \frac{n}{\gamma+n+1}. \end{aligned} \tag{A.10}$$

Since $n \geq 1$, we have $|\theta_{n+1,n} - \theta_{n+1,n-1}| \leq |\theta_{n+1,n+1} - \theta_{n,n}|$. Similarly we can prove that $|\theta_{n+1,i+1} - \theta_{n+1,i}| \leq |\theta_{n+1,n+1} - \theta_{n,n}|$ for any $1 \leq i \leq n$. So $|\theta_{n+1,i+1} - \theta_{n+1,i}| \leq \gamma(\gamma+1)/(n+1)^2$ for any $1 \leq i \leq n$, or equivalently, $|w_{n,i+1} - w_{n,i}| \leq \tilde{C} n^{-2}$ for all $1 \leq i < n$.

Then we claim the following lemma holds, and the conclusion follows.

Lemma A.1.5. *If $|w_{n,i+1} - w_{n,i}| \leq \tilde{C} n^{-2}$ for some constant $\tilde{C} > 0$ and all $1 \leq i < n$, then the smoothness condition in Theorem 2.2.4 holds.*

To prove the lemma, we need the following 3 steps:

Step 1: Define $m_i^i = 0$ and for any $k > 1$,

$$m_i^k = \sum_{t=i+1}^{k} \eta_t.$$

Then our goal is to prove $\sum_{i=1}^n \sum_{k=i+1}^n |w_{n,k} - w_{n,i}| \eta_i \exp(-\lambda m_i^k) \to 0$. Recall that $\lambda = \min(\lambda_{\min}(A), 1/(2\eta))$. Let $\mu = \lfloor 1/\lambda \rfloor + 1$. Choose an $N \in \mathbb{N}^+$ such that $\forall k > i \geq N$,

$$m_i^k \geq \mu \log \frac{k}{i}.$$

Since $m_i^k \geq \eta(k^{1-\alpha} - i^{1-\alpha})/(1-\alpha)$, we can always find such an N.

Step 2: Let $b_i^n = \sum_{k=i+1}^n |w_{n,k} - w_{n,i}| \eta_i \exp(-\lambda m_i^k)$. We decompose $\sum_{i=1}^n b_i^n$ into two parts:

$$
\begin{aligned}
\sum_{i=1}^{n} b_i^n &= \sum_{i=1}^{n} \eta_i \sum_{k=i+1}^{n} (w_{n,k} - w_{n,i}) \exp(-\lambda m_i^k) \\
&\leq \sum_{i=1}^{N} \eta_i \sum_{k=i+1}^{n} |w_{n,k} - w_{n,i}| \exp(-\lambda m_i^k) + \sum_{i=N+1}^{n} \eta_i \sum_{k=i+1}^{n} |w_{n,k} - w_{n,i}| \exp(-\lambda m_i^k) \\
&\triangleq I_1 + I_2
\end{aligned}
\tag{A.11}
$$

Step 3: Show that each term goes to 0 when $n \to \infty$. For the first term we have

$$
I_1 \leq \sum_{i=1}^{N} \eta_i \frac{2C}{n} \sum_{k=i+1}^{n} \exp(-\lambda m_i^k) \lesssim \frac{1}{n} \sum_{i=1}^{N} (i+1)^{\alpha} i^{-\alpha} \lesssim \frac{1}{n}
$$

The second inequality is due to Lemma A.1 and Lemma A.2 in Zhu et al. [2023]. Now there exists a constant $\tilde{C}$ such that $|w_{n,t+1} - w_{n,t}| \leq \tilde{C}/n^2$. So for the second term we have

$$
\begin{aligned}
I_2 &= \sum_{i=N+1}^{n} \eta_i \sum_{k=i+1}^{n} \{\sum_{t=i}^{k-1} |w_{n,t+1} - w_{n,t}|\} e^{-\lambda m_i^k} \\
&\lesssim \sum_{i=N+1}^{n} \eta_i \sum_{k=i+1}^{n} \sum_{t=i}^{k-1} \frac{1}{n^2} e^{-\lambda m_i^k} \\
&\lesssim n^{\alpha-2} \sum_{i=N+1}^{n} \eta_i \sum_{k=i+1}^{n} \sum_{t=i}^{k-1} \frac{1}{t^{\alpha}} e^{-\lambda m_i^k} \\
&\lesssim n^{\alpha-2} \sum_{i=N+1}^{n} \eta_i \sum_{k=i+1}^{n} m_i^k e^{-\lambda m_i^k} \\
&= n^{\alpha-2} \sum_{i=N+1}^{n} \sum_{k=i+1}^{n} \frac{m_i^k \eta_i}{\eta_k} e^{-\lambda m_i^k} (m_i^k - m_i^{k-1})
\end{aligned}
\tag{A.12}
$$

The second inequality is because $t^\alpha \le n^\alpha$. Notice that $\frac{\eta_i}{\eta_k} \le \frac{k}{i} \le e^{\frac{m_i^k}{\mu}}$ by step 2,

$$\begin{aligned} I_2 &\lesssim n^{\alpha-2} \sum_{i=N+1}^{n} \sum_{k=i+1}^{n} m_i^k e^{(\frac{1}{\mu}-\lambda)m_i^k}(m_i^k - m_i^{k-1}) \\ &\lesssim n^{\alpha-2} \sum_{i=N+1}^{n} \int_0^{+\infty} m e^{(\frac{1}{\mu}-\lambda)m} dm \\ &\lesssim n^{\alpha-1} \to 0. \end{aligned} \tag{A.13}$$

We have showed that

$$\sum_{i=1}^{n} b_i^n \lesssim n^{\alpha-1} \to 0.$$

So Lemma A.1.5 and Corollary 2.3.1 is proved.

□

A.1.6 Proof of Corollary 2.3.2

We first propose a lemma which bounds some exponential series.

Lemma A.1.6. *Let $\alpha \in (0.5, 1)$ and $\nu > 0$. For all $i < j < n$,*

$$\sum_{i=1}^{j} i^{-\alpha} \exp(\nu(i+1)^{1-\alpha}) \lesssim \exp(\nu j^{1-\alpha}),$$

$$\sum_{k=j}^{n} \exp(-\nu k^{1-\alpha}) \lesssim \exp(-\nu j^{1-\alpha}) j^{\alpha}.$$

Proof. Let $w_{n,i} = 1/\lceil \kappa n \rceil$ for $i > (1-\kappa)n$ otherwise 0. It's clear that $\sum_{i=1}^{n} w_{n,i} = 1$, $|w_{n,i}| \le 1/\kappa n$, and $\lim_{n\to\infty} n \sum_{i=1}^{n} (w_{n,i})^2 = 1/\kappa$. So we only need to verify the smoothness condition. By Lemma A.1. in Zhu et al. [2023] we have

$$\exp(-\lambda \sum_{t=i+1}^{k} \eta_t) \le \exp\Big(-\frac{\lambda\eta}{1-\alpha}\big(k^{1-\alpha} - (i+1)^{1-\alpha}\big)\Big).$$

Therefore

$$\begin{aligned}
&\sum_{i=1}^{n} \eta_i \sum_{k=i+1}^{n} |w_{n,k} - w_{n,i}| \exp(-\lambda \sum_{t=i+1}^{k} \eta_t) \\
\leq &\sum_{i=1}^{\lfloor \kappa n \rfloor} \eta_i \sum_{k=\lceil \kappa n \rceil}^{n} \frac{1}{\kappa n} \exp\left(- \frac{\lambda \eta}{1-\alpha} \big(k^{1-\alpha} - (i+1)^{1-\alpha}\big) \right) \\
\lesssim &\frac{1}{n} \sum_{i=1}^{\lfloor \kappa n \rfloor} i^{-\alpha} \exp\left(\frac{\lambda \eta}{1-\alpha} (i+1)^{1-\alpha} \right) \sum_{k=\lceil \kappa n \rceil}^{n} \exp\left(- \frac{\lambda \eta}{1-\alpha} k^{1-\alpha} \right) \\
\lesssim &\frac{1}{n} \exp\left(\frac{\lambda \eta}{1-\alpha} \lfloor \kappa n \rfloor^{1-\alpha} \right) \exp\left(- \frac{\lambda \eta}{1-\alpha} \lceil \kappa n \rceil^{1-\alpha} \right) \lceil \kappa n \rceil^{\alpha} \\
\lesssim &\frac{\lceil \kappa n \rceil^{\alpha}}{n}
\end{aligned} \tag{A.14}$$

So the smoothness condition $\lim_{n\to\infty} \sum_{i=1}^{n} \eta_i \sum_{k=i+1}^{n} |w_{n,k} - w_{n,i}| \exp(-\lambda \sum_{t=i+1}^{k} \eta_t) = 0$ holds, and the Corollary is proved.

□

A.1.7 Proof of Proposition 2.4.1

Instead of requiring $\eta_0 = a_1^{-2}$ in Proposition 2.4.1, we first consider a general step size. Recall that we have the squared loss function $f(x, \xi_i = (a_i, b_i)) = \dfrac{(a_i x - b_i)^2}{2}$, and SGD iterates

$$x_i = x_{i-1} - \eta_i a_i (a_i x_{i-1} - b_i). \tag{A.15}$$

Here $\eta_i = \eta_0 i^{-\alpha}$ with $0.5 < \alpha < 1$.

Proposition A.1.7. *The unique solution to the optimization problem*

$$\min_{c=(c_0,\,\cdots,\,c_n):c^T\mathbf{1}=1} \mathbb{E}\| \sum_{i=0}^{n} c_i (x_i - x^*) \|^2$$

with x_i defined in (A.15) is given by

$$c = \frac{\Theta^T D^{-1} \Theta \mathbf{1}_d}{\mathbf{1}_d^T \Theta^T D^{-1} \Theta \mathbf{1}_d},$$

where

$$D = \begin{pmatrix} (x_0 - x^*)^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma^2 a_1^2 \eta_1^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 a_2^2 \eta_2^2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 a_n^2 \eta_n^2 \end{pmatrix},$$

$$\Theta = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \eta_1 a_1^2 - 1 & 1 & 0 & \cdots & 0 \\ 0 & \eta_2 a_2^2 - 1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \eta_n a_n^2 - 1 & 1 & . \end{pmatrix}$$

More explicitly,

$$c_{n,0} = \frac{(\frac{\sigma}{x_0 - x^*})^2 + a_1^2 - \eta_1^{-1}}{S_n},$$

$$c_{n,i} = \frac{\eta_i^{-1} + a_{i+1}^2 - \eta_{i+1}^{-1}}{S_n}, 1 \leq i \leq n-1,$$

$$c_{n,n} = \frac{1}{\eta_n S_n},$$

where $S_n = (\frac{\sigma}{x_0 - x^*})^2 + \sum_{i=1}^n a_i^2$.

Proof. The SGD error sequence $x_i - x^*$ takes the recursion form

$$x_i - x^* = (1 - \eta_i a_i^2)(x_{i-1} - x^*) + \eta_i a_i (b_i - a_i x^*)$$

Let

$$\Theta = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \eta_1 a_1^2 - 1 & 1 & 0 & \cdots & 0 \\ 0 & \eta_2 a_2^2 - 1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \eta_n a_n^2 - 1 & 1 & . \end{pmatrix}$$

Then we have

$$\Theta \begin{pmatrix} x_0 - x^* \\ x_1 - x^* \\ x_2 - x^* \\ \vdots \\ x_n - x^* \end{pmatrix} = \begin{pmatrix} x_0 - x^* \\ \eta_1(b_1 - a_1 x^*)a_1 \\ \eta_2(b_2 - a_2 x^*)a_2 \\ \vdots \\ \eta_n(b_n - a_n x^*)a_n \end{pmatrix}. \tag{A.16}$$

We further treat a_i as fixed and denote Θ as the matrix in the left hand side above. Similar as the mean estimation model, here the optimal weights solution is also determined by $\Sigma = (\mathbb{E}(x_i x_j))_{i,j\geq 0}$, the "covariance" matrix of $(x_0,\ x_1,\ x_2,\ \cdots,\ x_n)$.

We further define

$$\Phi = \begin{pmatrix} x_0 - x^* \\ \eta_1(b_1 - a_1 x^*)a_1 \\ \eta_2(b_2 - a_2 x^*)a_2 \\ \vdots \\ \eta_n(b_n - a_n x^*)a_n \end{pmatrix},\ X = \begin{pmatrix} x_0 - x^* \\ x_1 - x^* \\ x_2 - x^* \\ \vdots \\ x_n - x^* \end{pmatrix},$$

and

$$D = \begin{pmatrix} (x_0 - x^*)^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma^2 a_1^2 \eta_1^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 a_2^2 \eta_2^2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 a_n^2 \eta_n^2 \end{pmatrix}.$$

Then $\Theta X = \Phi$ implies that $\mathbb{E}\Theta X X^T \Theta^T = \Theta \mathbb{E} X X^T \Theta^T = \Theta \Sigma \Theta^T = \mathbb{E}\Phi\Phi^T$. By the fact that $\mathbb{E}(x_i - x^*)(x_j - x^*) = 0$ for all $i \neq j$, we have $\mathbb{E}\Phi\Phi^T = D$. Thus we have a diagonalization of Σ as $\Theta\Sigma\Theta^T = D$.

Using the Lagrangian multiplier method, we can obtain the closed-form solution as

$$\frac{\Sigma^{-1}\mathbf{1}_d}{\mathbf{1}_d^T \Sigma^{-1} \mathbf{1}_d},$$

and it remains to show that the solution

$$\frac{\Sigma^{-1}\mathbf{1}_d}{\mathbf{1}_d^T \Sigma^{-1} \mathbf{1}_d} = \frac{\Theta^T D^{-1} \Theta \mathbf{1}_d}{\mathbf{1}_d^T \Theta^T D^{-1} \Theta \mathbf{1}_d}$$

is the form of Proposition A.1.7. Here we give the closed form of the matrix $\Theta^T D^{-1} \Theta$,

$$\Theta^T D^{-1} \Theta =$$

$$\begin{pmatrix} \frac{1}{(x_0 - x^*)^2} + \frac{(\eta_1 a_1^2 - 1)^2}{\sigma^2 a_1^2 \eta_1^2} & \frac{\eta_1 a_1^2 - 1}{\sigma^2 a_1^2 \eta_1^2} & 0 & 0 & \cdots & 0 \\ \frac{\eta_1 a_1^2 - 1}{\sigma^2 a_1^2 \eta_1^2} & \frac{1}{\sigma^2 a_1^2 \eta_1^2} + \frac{(\eta_2 a_2^2 - 1)^2}{\sigma^2 a_2^2 \eta_2^2} & \frac{\eta_2 a_2^2 - 1}{\sigma^2 a_2^2 \eta_2^2} & 0 & \cdots & 0 \\ 0 & \frac{\eta_2 a_2^2 - 1}{\sigma^2 a_2^2 \eta_2^2} & \frac{1}{\sigma^2 a_2^2 \eta_2^2} + \frac{(\eta_3 a_3^2 - 1)^2}{\sigma^2 a_3^2 \eta_3^2} & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \frac{\eta_n a_n^2 - 1}{\sigma^2 a_n^2 \eta_n^2} \\ 0 & 0 & \cdots & 0 & \frac{\eta_n a_n^2 - 1}{\sigma^2 a_n^2 \eta_n^2} & \frac{1}{\sigma^2 a_n^2 \eta_n^2} \end{pmatrix},$$

and the conclusion can be easily verified. □

The optimal weight of the initialization term $c_{n,0}$ in Proposition A.1.7 depends on σ^2 and the initial error $x_0 - x^*$, both of which can not be observed. To solve this problem, we may consider the two-step estimation: first estimate σ and x^* using ASGD or other averaged schemes with a small batch of SGD iterates, then plug them in to obtain the optimal weights. Another approach is to modify the structure of Θ and equation A.16. Choosing $\eta_1 = a_1^{-2}$, we can exclude $x_0 - x^*$ from (A.16) and reduce it to

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ \eta_2 a_2^2 - 1 & 1 & \cdots & 0 \\ \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \eta_n a_n^2 - 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 - x^* \\ x_2 - x^* \\ \vdots \\ x_n - x^* \end{pmatrix} = \begin{pmatrix} \eta_1 (b_1 - a_1 x^*) a_1 \\ \eta_2 (b_2 - a_2 x^*) a_2 \\ \vdots \\ \eta_n (b_n - a_n x^*) a_n \end{pmatrix}.$$

In other words, if we plug $\eta_1 = a_1^{-2}$ in Proposition A.1.7, we have $x_1 = \eta_i b_i a_i$ and all SGD iterates will be free of $x_0 - x^*$.

Denote $\tilde{\Theta}$ as this reduced matrix in the left hand side above, and we can perform a diagonalization of $\Sigma_{-0} = (\mathbb{E} x_i x_j)_{i,j \geq 1}$, the "covariance" matrix of SGD sequence without x_0, as follows

$$\tilde{\Theta} \Sigma_{-0} \tilde{\Theta}^T = D_{-0} = \begin{pmatrix} \sigma^2 a_1^2 \eta_1^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 a_2^2 \eta_2^2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 a_n^2 \eta_n^2 \end{pmatrix}.$$

Instead of the optimization problem in proposition 2.4.1, the decomposition of Σ_{-0} enables us to solve a reduced problem excluding the weight on x_0

$$\min_{c=(c_1,\, \cdots,\, c_n): c^T \mathbf{1}_d = 1} \mathbb{E} \| \sum_{i=1}^{n} c_i (x_i - x^*) \|^2,$$

and get the minimizer in the form of

$$c = \frac{\Sigma_{-0}^{-1}\mathbf{1}_d}{\mathbf{1}_d^T\Sigma_{-0}^{-1}\mathbf{1}_d} = \frac{\tilde{\Theta}^T D_{-0}^{-1}\tilde{\Theta}\mathbf{1}}{\mathbf{1}_d^T\tilde{\Theta}^T D_{-0}^{-1}\tilde{\Theta}\mathbf{1}_d}.$$

The σ^2 terms in D_{-0} cancel out. Finally, the weighting scheme with $\eta_1 = a_1^{-2}$ is

$$c_{n,i} = \frac{\eta_i^{-1} + a_{i+1}^2 - \eta_{i+1}^{-1}}{S_n}, 1 \leq i \leq n-1,$$

$$c_{n,n} = \frac{1}{\eta_n S_n},$$

where $S_n = \sum_{i=1}^n a_i^2$.

A.1.8 Proof of Corollary 2.4.2

We start with a lemma describing the rate of the last error term of SGD, which can be found in Lemma 3.2 in Chen et al. [2020].

Lemma A.1.8. *Under the setting in Theorem 2.2.4, we have*

$$n^{\frac{\alpha}{2}}(x_n - x^*) = O_p(1).$$

Proof. First of all, define $\tilde{x}_n' = \sum_{i=1}^{n-1} c_{n,i}x_i$. Notice that

$$\sqrt{n}(\tilde{x}_n - x^*) - \sqrt{n}(\tilde{x}_n' - (1 - n^{\alpha-1})x^*) = n^{\alpha-\frac{1}{2}}(x_n - x^*).$$

By Lemma A.1.8,

$$n^{\alpha-\frac{1}{2}}(x_n - x^*) = n^{\frac{\alpha}{2}-\frac{1}{2}}n^{\frac{\alpha}{2}}(x_n - x^*) = o_p(1).$$

So $\sqrt{n}(\tilde{x}_n - x^*)$ and $\sqrt{n}(\tilde{x}_n' - (1 - n^{\alpha-1})x^*) = \sqrt{n}\sum_{i=1}^{n-1}(x_i - x^*)$ has the same asymptotic distribution. Define the linear error term

$$\delta'_n = \delta'_{n-1} - \eta_n A\delta'_{n-1} + \eta_n\epsilon_n,\ \ \delta'_0 = x_0 - x^*,$$

and $\tilde{\delta}'_n = \sum_{i=1}^{n-1} c_{n,i}\delta'_i$. Previously we have showed that the weighted averaged linear error term can well approximate the weighted averaged original error term (see Subsection A.1.2 for details). So it suffices to prove that

$$\sqrt{n}\tilde{\delta}'_n \Rightarrow \mathcal{N}(0, A^{-1}SA^{-1}).$$

Notice that $c_{n,i} \leq O(1/n)$ holds for $i = 1, ..., n-1$. By Lemma A.1.2, we only need to verify the smoothness and limitation condition. For the smoothness condition, Define

$$\tau_i^n = \eta_i \sum_{k=i+1}^{n-1} |c_{n,k} - c_{n,i}|e^{-\lambda m_k^i} = \eta_i \sum_{k=i+1}^{n-1} \frac{|k^\alpha - (k+1)^\alpha - i^\alpha + (i+1)^\alpha|}{n} e^{-\lambda m_k^i},$$

where $m_k^i = \sum_{t=i+1}^{k} \eta_t$ as we previously defined.

Since $|i^\alpha - (i+1)^\alpha| \asymp i^{\alpha-1}$, we have $||\tau_i^n||_2 \lesssim 1/n$. Let $N = \lfloor\sqrt{n}\rfloor$, then

$$\sum_{i=1}^{n-1} \tau_i^n = \sum_{i=1}^{N} \tau_i^n + \sum_{i=N+1}^{n-1} \tau_i^n.$$

We estimate the two terms respectively.

$$\begin{aligned}\sum_{i=1}^{N} \tau_i^n &\lesssim \frac{1}{n}\sum_{i=1}^{N}\sum_{k=i+1}^{n-1} e^{-\lambda m_k^i}\eta_i \\ &\lesssim \frac{1}{n}\sum_{i=1}^{N}(i+1)^\alpha i^{-\alpha} \\ &\lesssim \frac{N}{n} \leq n^{-\frac{1}{2}}.\end{aligned} \tag{A.17}$$

The second inequality is due to Lemma A.1. and Lemma A.2. in Zhu et al. [2023].

$$\begin{aligned}\sum_{i=N+1}^{n-1} \tau_i^n &\lesssim \sum_{i=N+1}^{n-1} \sum_{k=i+1}^{n-1} \frac{k^{\alpha-1} + i^{\alpha-1}}{n} e^{-\lambda m_k^i} \eta_i \\ &\lesssim \frac{2\sqrt{n}^{\alpha-1}}{n} \sum_{i=N+1}^{n-1} (i+1)^{\alpha} i^{-\alpha} \\ &\lesssim n^{\frac{\alpha-1}{2}}.\end{aligned} \tag{A.18}$$

As a result,

$$\lim_{n\to\infty} \sum_{i=1}^{n-1} \tau_i^n = 0$$

and the smoothness condition holds. Then we compute the limitation. Notice that

$$\begin{aligned}(n-1)\sum_{i=1}^{n-1}(c_{n,i}^2 - \frac{1}{n^2}) &= (n-1)\sum_{i=1}^{n-1}(c_{n,i} - \frac{1}{n})(c_{n,i} + \frac{1}{n}) \\ &= (n-1)\sum_{i=1}^{n-1}\left[\frac{(i+1)^{\alpha} - i^{\alpha}}{n}\right]\left[\frac{(i+1)^{\alpha} - i^{\alpha} + 2}{n}\right] \\ &\lesssim \sum_{i=1}^{n-1} \frac{(i+1)^{\alpha} - i^{\alpha}}{n} \le n^{\alpha-1},\end{aligned} \tag{A.19}$$

we have

$$\lim_{n\to\infty} (n-1)\sum_{i=1}^{n-1} c_{n,i}^2 = \lim_{n\to\infty}(n-1)\sum_{i=1}^{n-1}\frac{1}{n^2} = 1.$$

By Lemma A.1.2,

$$\sqrt{n-1}\tilde{\delta}_n' \Rightarrow \mathcal{N}(0, A^{-1}SA^{-1}).$$

Finally by Slutsky's theorem,

$$\sqrt{n}\tilde{\delta}_n' = \sqrt{\frac{n}{n-1}}\sqrt{n-1}\tilde{\delta}_n' \Rightarrow \mathcal{N}(0, A^{-1}SA^{-1}).$$

□

A.1.9 Proof of Lemma A.1.4

Proof. Since $\Gamma(\gamma+i+1)\Gamma(n+1) < \Gamma(\gamma+n+1)\Gamma(i+1)$, we have

$$|\theta_{n,i}| \leq \frac{\gamma+1}{n}.$$

From the recursive form of polynomial decay averaged SGD, it is easy to see $\sum_{i=1}^{n} \theta_{n,i} = 1$. The last step is to prove the limitation holds. Define

$$\Gamma_n(x) = \int_0^n t^{x-1}(1-\frac{t}{n})^n dt = \frac{n^x n!}{z(z+1)(z+2)\cdots(z+n)},$$

where the last equality is from integration by parts. It's well known that $(1-\frac{t}{n})^n \leq (1-\frac{t}{n+1})^{n+1}$ and $\lim_{n\to\infty}(1-\frac{t}{n})^n = e^{-t}$ for any t. So $\Gamma_n(x) \leq \Gamma(x)$. Meanwhile we have an equivalent definition of $\Gamma(x)$:

$$\Gamma(x) = \lim_{n\to\infty} \frac{n^x n!}{x(x+1)(x+2)\cdots(x+n)} = \lim_{n\to\infty} \Gamma_n(x).$$

So for any $\tau > 0$, there exists an $N > 0$ such that for all $n \geq N$,

$$0 \leq \Gamma(\gamma) - \frac{n^\gamma n!}{\gamma(\gamma+1)(\gamma+2)\cdots(\gamma+n)} \leq \tau.$$

As a result, for $n \geq i \geq N$, we have $0 \leq \frac{\Gamma(\gamma+i+1)}{\Gamma(i+1)i^\gamma} - 1 \leq \Gamma(\gamma)\tau$ and $0 \leq \frac{\Gamma(\gamma+n+1)}{\Gamma(n+1)i^\gamma} - 1 \leq \Gamma(\gamma)\tau$, which implies

$$\left|\frac{\Gamma(\gamma+n+1)}{\Gamma(n+1)i^\gamma} - \frac{\Gamma(\gamma+i+1)}{\Gamma(i+1)i^\gamma}\right| \leq 2\Gamma(\gamma)\tau.$$

Furthermore we have

$$\begin{aligned}\left|\frac{i^\gamma}{n^\gamma}-\frac{\Gamma(\gamma+i+1)\Gamma(n+1)}{\Gamma(\gamma+n+1)\Gamma(i+1)}\right| &= \left|\frac{t^\gamma\Gamma(n+1)}{\Gamma(\gamma+n+1)}\right|\left|\frac{\Gamma(\gamma+n+1)}{\Gamma(n+1)i^\gamma}-\frac{\Gamma(\gamma+i+1)}{\Gamma(i+1)i^\gamma}\right| \\ &\leq 2\Gamma(\gamma)\tau\left|\frac{n^\gamma\Gamma(n+1)}{\Gamma(\gamma+n+1)}\right| \\ &\leq 2\Gamma(\gamma)\tau.\end{aligned} \tag{A.20}$$

Now we estimate the following summation

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}[(\gamma+1)(\frac{i}{n})^\gamma-n(\theta_{n,i})] \\ \leq&\frac{1}{n}\sum_{i=1}^{N}|(\gamma+1)(\frac{i}{n})^\gamma|+\sum_{i=1}^{N}|(\theta_{n,i})|+\frac{\gamma+1}{n}\sum_{i=N+1}^{n}\left|(\frac{i}{n})^\gamma-\frac{\Gamma(\gamma+i+1)\Gamma(n+1)}{\Gamma(\gamma+n+1)\Gamma(i+1)}\right| \\ \leq&\frac{N(\gamma+1)}{n}+\frac{N}{n}\theta_{n,N}+\frac{n-N}{n}\tau(\gamma+1) \\ \leq&(\gamma+1)(\frac{N}{2n}+\tau).\end{aligned} \tag{A.21}$$

Let $\tau \to 0$ and $n \to \infty$,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}[(\gamma+1)(\frac{i}{n})^\gamma-n(\theta_{n,i})]=0. \tag{A.22}$$

Since

$$\begin{aligned}[(\gamma+1)^2(\frac{i}{n})^{2\gamma}-n^2(\theta_{n,i})^2] &= [(\gamma+1)(\frac{i}{n})^\gamma-n(\theta_{n,i})][(\gamma+1)(\frac{i}{n})^\gamma+n(\theta_{n,i})] \\ &\leq 2(\gamma+1)[(\gamma+1)(\frac{i}{n})^\gamma-n(\theta_{n,i})],\end{aligned} \tag{A.23}$$

we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}[(\gamma+1)^2(\frac{i}{n})^{2\gamma}-n^2(\theta_{n,i})^2]\leq 2(\gamma+1)\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}[(\gamma+1)(\frac{i}{n})^\gamma-n(\theta_{n,i})]=0. \tag{A.24}$$

Finally, notice that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\gamma+1)^2(\frac{i}{n})^{2\gamma}=\int_0^1(\gamma+1)^2x^{2\gamma}dx=\frac{(\gamma+1)^2}{2\gamma+1},$$

We have proved the limitation in Theorem 2.2.4 holds with

$$\lim_{n\to\infty}n\sum_{i=1}^{n}\theta_{n,i}^2=\frac{(\gamma+1)^2}{2\gamma+1}.$$

□

A.1.10 Proof of Lemma A.1.6

Proof. First we have $\exp(\nu(i+1)^{1-\alpha})\asymp\exp(\nu i^{1-\alpha})$, so

$$\sum_{i=1}^{j}i^{-\alpha}\exp(\nu(i+1)^{1-\alpha})\lesssim\sum_{i=1}^{j}i^{-\alpha}\exp(\nu i^{1-\alpha}).$$

Let $\psi(x)=x^{-\alpha}\exp(\nu x^{1-\alpha})$. It's an increasing function when $x>(\alpha/(1-\alpha)\nu)^{1/(1-\alpha)}$. So we can further bound our target term as,

$$\begin{aligned}\sum_{i=1}^{j}i^{-\alpha}\exp(\nu(i+1)^{1-\alpha})&\lesssim\sum_{i=1}^{j}i^{-\alpha}\exp(\nu i^{1-\alpha})\\&\lesssim\int_0^{j+1}\psi(x)dx\\&=\frac{1}{\nu(1-\alpha)}\exp(\nu(j+1)^{1-\alpha})\\&\lesssim\exp(\nu j^{1-\alpha}).\end{aligned}\tag{A.25}$$

For the second inequality, similarly we have

$$\begin{aligned}\sum_{k=j}^{n}\exp(-\nu k^{1-\alpha}) &\lesssim \int_{j-1}^{n}\exp(-\nu x^{1-\alpha}))dx\\ &= \int_{\nu(j-1)^{1-\alpha}}^{\infty}\frac{s^{\frac{\alpha}{1-\alpha}}}{\nu(1-\alpha)}e^{-s}ds\\ &\lesssim (j-1)^{\alpha}\exp(-\nu(j-1)^{1-\alpha})\\ &\asymp \exp(-\nu j^{1-\alpha})j^{\alpha}.\end{aligned} \tag{A.26}$$

□

APPENDIX B

APPENDIX FOR CHAPTER 3

B.1 Proofs

We first introduce some notations. For a random variable X and $q > 0$, we write $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}$ if $\mathbb{E}|X|^q < \infty$. Moreover, for any random matrix A, we write $\|A\|_q = (\mathbb{E}\|A\|_2^q)^{1/q}$ by convention. From this point on, abusing notation, depending on context we may write $\|\cdot\|_2$ to denote the matrix operator norm, or may also write $\|\cdot\|_2$ to denote the random matrix norm discussed here.

B.1.1 Some useful lemmas

In Lemma B.1.1, the case $1 < q \leq 2$ follows from Burkholder [1988] and the other case $q > 2$ is due to Rio [2009]. Lemma B.1.2 follows from Corollary 1.8 of Nagaev [1979].

Lemma B.1.1 (Burkholder)**.** *Let $q > 1$ and $q' = \min\{q, 2\}$. Let $(D_t)_{t\in\mathbb{Z}}$ be martingale differences with $\mathbb{E}|D_t|^q < \infty$ for every $t \in \mathbb{Z}$. Write $M_n = \sum_{t=1}^n D_t$. Then*

$$\|M_n\|_q^{q'} \leq C_q^{q'} \sum_{t=1}^n \|D_t\|_q^{q'}, \text{ where } C_q = \begin{cases} (q-1)^{-1}, & 1 < q \leq 2, \\ \sqrt{q-1}, & q > 2. \end{cases}$$

Lemma B.1.2 (Nagaev)**.** *Let $(e_t)_{t\in\mathbb{Z}}$ be independent zero-mean random variables with $\sup_{t\in\mathbb{Z}} \mathbb{E}|e_t|^q < \infty$ for some $q > 2$. Let $S_n = \sum_{t=1}^n e_t$ and $c_q = 2e^{-q}(q+2)^{-2}$. Then, for $x > 0$, we have*

$$\mathbb{P}(|S_n| \geq x) \leq (1 + 2/q)^q \frac{\sum_{t=1}^n \mathbb{E}|e_i|^q}{x^q} + 2\exp\left(-\frac{c_q x^2}{\sum_{t=1}^n \mathbb{E}|e_i|^2}\right).$$

Lemma B.1.3 (Moment bounds for sample covariance operators)**.** *Under Assumption 3.2.2, we have*

$$\left(\mathbb{E}\|A_t - \Sigma\|_2^{\psi}\right)^{1/\psi} \leq \lambda_0.$$

Proof. For simplicity, write

$$\Delta_{B,jk} = \sum_{i=1}^{B} (X_{ij}X_{ik} - \Sigma_{jk})$$

for $1 \leq j, k \leq p$. By Lemma B.1.1, it follows that

$$\|\Delta_{B,jk}\|_{\psi}^2 \leq (\psi - 1) \sum_{i=1}^{B} \|X_{ij}X_{ik} - \Sigma\|_{\psi}^2 \leq 4B(\psi - 1)M_{\psi}^4.$$

Consequently, under Assumption 3.2.2, we have

$$\left(\mathbb{E}\|A_t - \Sigma\|_2^{\psi}\right)^{1/\psi} \leq \frac{1}{B} \left(\sum_{j,k=1}^{p} \|\Delta_{B,jk}\|_{\psi}^2 \right)^{1/2} \leq \frac{2p(\psi - 1)^{1/2} M_{\psi}^2}{B^{1/2}} \leq \lambda_0.$$

□

B.1.2 Proof of Lemma 3.2.8

By Lemma B.1.3, triangle inequality and the fact that $\eta_0 \leq 1/\|\Sigma\|_2$, we have for each $\ell \geq 1$,

$$\|I_p - \eta_\ell A_\ell\|_{\psi} \leq \|I_p - \eta_\ell \Sigma\|_{\psi} + \eta_\ell \|A_\ell - \Sigma\|_{\psi} \leq 1 - 2\lambda_0 \eta_\ell + \lambda_0 \eta_\ell = 1 - \lambda_0 \eta_\ell.$$

Consequently, by the triangle inequality, it follows that

$$\|\omega^\top S_T^\diamond\|_\psi \le \|\theta_0 - \theta^\star\|_2 \sum_{t=1}^T \prod_{\ell=1}^t (1-\lambda_0\eta_\ell) \le \|\theta_0 - \theta^\star\|_2 \Upsilon_{\lambda_0\eta_0,\alpha}. \tag{B.1}$$

Then Lemma 3.2.8 is obtained through Markov's inequality.

B.1.3 Proof of Lemma 3.2.9

We first introduce the following lemma, providing a concentration inequality for $D_{T,2}$, where

$$D_{T,2} = B\sigma^2 \sum_{m=1}^T \eta_m^2 \omega^\top H_m A_m H_m^\top \omega =: B\sigma^2 \sum_{m=1}^T \eta_m^2 \xi_m.$$

Lemma B.1.4 (Main Technical Lemma)**.** *Under Assumption 3.2.2, for $z > 0$, we have*

$$\mathbb{P}(|D_{T,2} - \mathbb{E}(D_{T,2})| > z) \le \frac{C_{\psi,\alpha} T L^{\psi/4-1} \|\Sigma\|_2^{\psi/2}}{\lambda_0^\psi (z/B)^{\psi/2}} + C\exp\left\{-\frac{C'_{\psi,\alpha}(z/B)^2\lambda_0^4}{T\|\Sigma\|_2^2}\right\},$$

where $C_{\psi,\alpha}$ and $C'_{\psi,\alpha}$ are positive constants depending only on ψ and α, and

$$L \asymp \frac{T^\alpha}{\lambda_0\eta_0} \log\left(\frac{B\|\Sigma\|_2 T^{1+\alpha}}{\lambda_0^2}\right). \tag{B.2}$$

Proof. For any $k \ge 1$, define $\mathcal{F}_{A,k} = \sigma\{A_1, A_2, \dots, A_k\}$ and the projection operator

$$\mathcal{P}_{A,k}(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_{A,k}) - \mathbb{E}(\cdot|\mathcal{F}_{A,k-1}).$$

Denote $H_m = \mathcal{H}(A_{m+1}, A_{m+2}, \dots, A_T)$. For any $h \ge 1$, define

$$H_{m,\{m+h\}} = \mathcal{H}(A_{m+1}, A_{m+2}, \dots, A_{m+h-1}, A^\star_{m+h}, A_{m+h+1}, \dots, A_T).$$

where $(A_t^\star)_{t\in\mathbb{Z}}$ are i.i.d. random matrix with $A_t^\star \overset{\mathcal{D}}{=} A_t$. Note that

$$H_m - H_{m,\{m+h\}} = \sum_{k=m+h}^{T} \prod_{\ell=m+h+1}^{k} (\mathbf{I}_p - \eta_\ell A_\ell)\eta_{m+h}(A_{m+h} - A_{m+h}^\star) \prod_{\ell=m+1}^{m+h-1} (\mathbf{I}_p - \eta_\ell A_\ell).$$

Hence, by Assumption 3.2.2, we have $\|A_{m+h} - \Sigma\|_\psi \le \lambda_0$ and consequently

$$\begin{aligned}\|H_m - H_{m,\{m+h\}}\|_\psi &\lesssim \sum_{k=m+h}^{T} \eta_{m+h}\|A_{m+h} - \Sigma\|_\psi \prod_{\ell=m+1}^{k} (1 - \lambda_0\eta_\ell) \\ &\le \lambda_0\eta_{m+h} \int_{m+h}^{\infty} \exp\left(-\lambda_0\eta_0 \int_{m+1}^{z} x^{-\alpha} dx\right) dz.\end{aligned}$$

Therefore, together with the fact that $\|A_m\|_\psi \le \|A_m - \Sigma\|_\psi + \|\Sigma\|_2 \le 2\|\Sigma\|_2$, we have

$$\begin{aligned}\|\mathcal{P}_{A,m+h}(\xi_m)\|_{\psi/2} &\le 2\|A_m\|_\psi \|H_m - H_{m,\{m+h\}}\|_\psi \|H_m\|_\psi \\ &\lesssim \lambda_0\|\Sigma\|\eta_{m+h}\|H_m\|_\psi \int_{m+h}^{\infty} \exp\left(-\lambda_0\eta_0 \int_{m+1}^{z} x^{-\alpha} dx\right) dz.\end{aligned}$$

Define the L-approximation of $D_{T,2}$ as

$$D_{T,2,L} = B\sigma^2 \sum_{m=1}^{T} \eta_m^2 \mathbb{E}(\xi_m | \mathcal{P}_{A,m+L}) = D_{T,2} - B\sigma^2 \sum_{m=1}^{T} \eta_m^2 \sum_{h=L+1}^{T-h} \mathcal{P}_{A,m+h}(\xi_m).$$

Note that $\mathbb{E}(D_{T,2}) = \mathbb{E}(D_{T,2,L})$. Hence, by Lemma B.1.1 and (B.2),

$$\begin{aligned}\|D_{T,2} - D_{T,2,L}\|_{\psi/2} &\le C_\psi B\sigma^2 \sum_{h=L+1}^{T-1} \left\{ \sum_{m=1}^{T-h} \eta_m^4 \|\mathcal{P}_{A,m+h}(\xi_m)\|_{\psi/2}^2 \right\}^{1/2} \\ &\le \frac{C_{\psi,\alpha} B\sigma^2\|\Sigma\|T^{1+\alpha}}{\lambda_0^2 L^\alpha} \exp\left(-\frac{\lambda_0\eta_0 L}{2^\alpha T^\alpha}\right) \le C_{\psi,\alpha}T^{-1/2}.\end{aligned}$$

Now we bound $|D_{T,2,L} - \mathbb{E}(D_{T,2,L})|$. By Lemma B.1.2 and a similar argument as that of

(B.3),

$$\mathbb{P}(|D_{T,2,L} - \mathbb{E}(D_{T,2,L})| > z) \leq \frac{C_{\psi,\alpha} T L^{\psi/4-1} \|\Sigma\|_2^{\psi/2}}{\lambda_0^{\psi} (z/B)^{\psi/2}} + C \exp\left\{ -\frac{C'_{\psi,\alpha}(z/B)^2 \lambda_0^4}{T\|\Sigma\|_2^2} \right\}.$$

Consequently, Lemma B.1.4 follows in view of

$$\mathbb{P}(|D_{T,2} - \mathbb{E}(D_{T,2})| > z) \leq \mathbb{P}(|D_{T,2} - D_{T,2,L}| > z/2) + \mathbb{P}(|D_{T,2,L} - \mathbb{E}(D_{T,2,L})| > z/2).$$

□

Remaining proof: As discussed in Section 3.2.4, it suffices to bound $D_{T,q}$ and $D_{T,2}$. By Assumption 3.2.2 and a similar argument as (B.1),

$$\|H_m\|_q \leq 1 + \int_{m+1}^{\infty} \exp\left(-\lambda_0 \eta_0 \int_{m+1}^{z} x^{-\alpha} dx \right) dz,$$

which leads to

$$\mathbb{E}(D_{T,q}) = B\mu_q \sum_{m=1}^{T} \eta_m^q \mathbb{E}|\omega^\top H_m X_i|^q \leq B\mu_q \mathcal{K}_q \sum_{m=1}^{T} \eta_m^q \|H_m\|_q^q \leq \frac{C_{q,\alpha} n \mu_q \mathcal{K}_q}{\lambda_0^q}. \tag{B.3}$$

Hence, by Lemma B.1.4, we have

$$\mathbb{P}\left(D_{T,2} > \mathbb{E}(D_{T,2}) + \frac{x^2}{\log x} \right) \leq \frac{C_{\psi,\alpha} T L^{\psi/4-1} \|\Sigma\|_2^{\psi/2} B^{\psi/2}}{\lambda_0^{\psi} (x^2/\log x)^{\psi/2}} + C \exp\left\{ -\frac{C'_{\psi,\alpha}(x^2/\log x)^2 \lambda_0^4}{T\|\Sigma\|_2^2 B^2} \right\}.$$

As $\psi > (2q - 4\alpha)/(2-\alpha)$, for any $x \gtrsim \sqrt{T}$, we have

$$\frac{T L^{\psi/4-1} (\log x)^{\psi/2}}{x^{\psi}} = o\left(\frac{T}{x^q} \right).$$

Consequently, as $\mathbb{E}(D_{T,2}) \leq C_{q,\alpha} n W_2$, we have

$$\mathbb{P}\left(|\omega^\top S_T^\star| > x\right) \leq \frac{C_1 T W_q}{x^q} + C \exp\left(-\frac{C_2 x^2}{T W_2 + x^2/\log x}\right) \leq \frac{C_1 T W_q}{x^q} + C \exp\left(-\frac{C_2 x^2}{T W_2}\right),$$

where C_1 and C_2 are positive constants depending only on q, α and ψ.

B.1.4 Proof of Theorems 3.2.4, 3.2.6

As discussed in Section 3.2.4, Theorem 3.2.4 directly follows from Lemma 3.2.8 and Lemma 3.2.9. The proof of Theorem 3.2.6 is similar to that of Theorem 3.2.4 and thus omitted.

B.1.5 Proof of Proposition 3.3.1

Proof. Let

$$\mu_{T,q} = \sum_{t=1}^{T} \left(\sum_{i=t}^{T} V_t^i \eta_t / T \right)^q .$$

Note that

$$\bar{\theta}_T - \theta^* - \frac{1}{T} \sum_{i=1}^{T} V_i \Delta_0 = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=t}^{T} V_t^i \eta_t \epsilon_t.$$

Since $\{\epsilon_t\}_{t \geq 1}$ are i.i.d. , according to Corollary 1.8 in Nagaev [1979] we have

$$\mathbb{P}\left(\left| \bar{\theta}_T - \theta^* - \frac{1}{T} \sum_{i=1}^{T} V_i \Delta_0 \right| \geq x \right) \leq (1 + 2/q)^q \frac{\mu_{T,q} \mathbb{E}|\epsilon|^q}{x^q} + 2 \exp\left(-\frac{c_q x^2}{\mu_{T,2} \mathbb{E}|\epsilon|^2} \right).$$

Then all we need to show is that $\mu_{T,q} \asymp T^{1-q}$ for $2 \leq q < \nu$, and $\sum_{i=1}^{T} V_i = O(1)$. Since there's no randomness in $\mu_{T,q}$ and V_i, we can check the order through numerical computation; see figure B.1. Also, according to Lemma A.2 in Zhu et al. [2023], $\sum_{i=t}^{T} V_t^i = O(t^\alpha)$ for $\alpha \in (1/2, 1)$, which implies that $\mu_{T,q} = O(T^{1-q})$. □

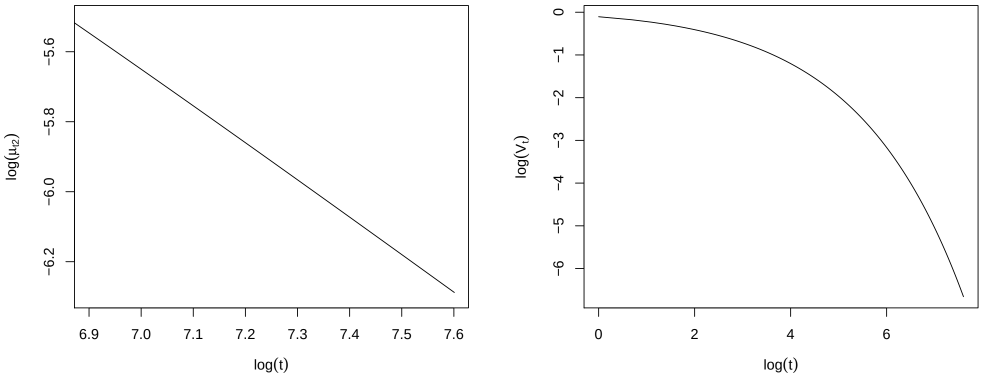


Figure B.1: Left: Check the order of $\mu_{T,2}$. The X axis represents $\log(t)$; the Y axis represents $\log(\mu_{t,2})$. The slop of the log-log curve is about -1, which implies that $\mu_{T,2} \asymp T^{-1}$. Right: Check the order of V_t. The X axis represents $\log(t)$; the Y axis represents $\log(V_t)$. The slop of the log-log curve is much less than -1 when t is large, which means V_t is summable and $\sum_{i=1}^{T} V_i = O(1)$.

B.1.6 Proof of Proposition 3.3.2

Proof. Let

$$S_T = \bar{\theta}_T - \theta^* - \frac{1}{T}\sum_{i=1}^{T} V_i \Delta_0 = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=t}^{T} V_t^i \eta_t \epsilon_t.$$

To apply Theorem 1 in Peligrad et al. [2014], we need to verify the basic assumption, the uniform asymptotic negligibility of the variance of individual summands, that is

$$\max_t \left(\sum_{i=t}^{T} V_t^i \eta_t\right)^2 / \sum_{t=1}^{T}\left(\sum_{i=t}^{T} V_t^i \eta_t\right)^2 \to 0. \tag{B.4}$$

Since Lemma $A.2$ in Zhu et al. [2023] shows $\sum_{i=t}^{T} V_t^i \asymp t^{\alpha}$ as $T \to \infty$ and $\eta_t = \eta_0 t^{-\alpha}$, the above limit is of order T^{-1}. (Note that $\sigma_T^2 = \mathbb{E}|\epsilon|^2 \mu_{T,2} \asymp T^{-1}$.) We can also verify (B.4) from numerical computation; see Figure B.2. Then, according to Theorem 1 in Peligrad

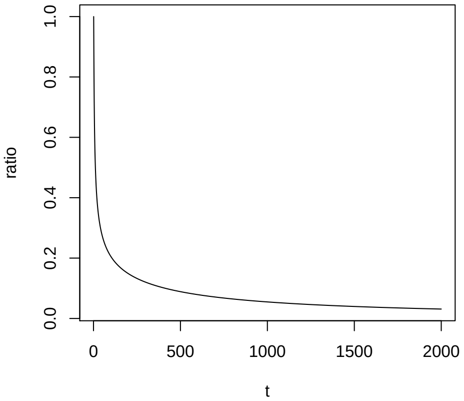


Figure B.2: Check the uniform asymptotic negligibility of the variance of individual summands. The X axis represents t; the Y axis represents the ratio of the largest individual variance and variance of individual summands.

et al. [2014], we have

$$\mathbb{P}\left(S_T \geq x\right) = (1 + o(1)) \left(1 - \Phi(x/\sigma_T) + \sum_{t=1}^{T} P\left(\sum_{i=t}^{T} V_t^i \eta_t \epsilon_t / T \geq x\right)\right),$$

which naturally yields (3.14).

□

APPENDIX C

APPENDIX FOR CHAPTER 4

This chapter is organized as follows: In section C.1, we introduce some technical lemmas, which are useful for our proofs later. In section C.2, we prove the convergence of our proposed online estimator in the special case of linear processes, i.e., Lemma C.2.1. We break down the proof of Lemma C.2.1 into several parts: C.2.2, C.2.3, C.2.4, and C.2.5 in the rest of this section. Based on the results for the special case, we prove in section C.3 the convergence in general cases, i.e., Theorems 4.3.5 and 4.3.8. We provide proof of Proposition 4.3.1 in section C.4. We also include a simple simulation study applying the fixed-width sequential stopping rule in Section C.5. We use $\mathbf{I}$ to denote a $d \times d$ identity matrix.

C.1 Technical Lemmas

Lemma C.1.1. *Assume that A is a positive definite matrix. For any $i \in \mathbb{N}$, define the matrix sequence $\{Y_i^j\}$ with $Y_i^i = \mathbf{I}$ and for any $j > i$*

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

where η_k is chosen to be $\eta k^{-\alpha}$ for $\alpha \in (1/2, 1)$. Then we have

$$\|Y_i^j\|_2 \le \exp\left(-\eta\gamma \sum_{k=i+1}^{j} k^{-\alpha}\right) \le \exp\left[-\frac{\gamma\eta}{1-\alpha}\left(j^{1-\alpha} - (i+1)^{1-\alpha}\right)\right],$$

where $\gamma = \min(\lambda_{\min}(A), 1/(2\eta))$.

Proof. Since A is positive definite, there exists an orthonormal matrix Q and a diagonal matrix Λ such that $A = Q\Lambda Q^T$. We have

$$\|Y_i^j\|_2 \leq \prod_{k=i+1}^{j} \|(\mathbf{I} - \eta_k A)\|_2 = \prod_{k=i+1}^{j} \|(\mathbf{I} - \eta_k \Lambda)\|_2 \leq \prod_{k=i+1}^{j} \left(1 - \gamma\eta k^{-\alpha}\right).$$

Note that $1 - x \leq \exp(-x)$ for any $x \in [0, 1]$. So $\|Y_i^j\|_2$ can be further bounded as

$$\|Y_i^j\|_2 \leq \exp\left(-\sum_{k=i+1}^{j} \gamma\eta k^{-\alpha}\right).$$

The lemma can be verified using the fact that

$$\sum_{k=i+1}^{j} k^{-\alpha} \geq \int_{i+1}^{j+1} k^{-\alpha} dk = \frac{1}{1-\alpha}\left((j+1)^{1-\alpha} - (i+1)^{1-\alpha}\right).$$

□

Lemma C.1.2. *With Y_i^j defined in Lemma C.1.1, let $S_i^j = \sum_{k=i+1}^{j} Y_i^k$ for any $j > i$ and $S_i^i = 0$. Then we have*

$$\|S_i^j\|_2 \lesssim (i+1)^{\alpha}.$$

Proof. Through triangle inequality and Lemma C.1.1,

$$\|S_i^j\|_2 \leq \sum_{k=i+1}^{j} \|Y_i^k\|_2 \leq \sum_{k=i+1}^{j} \exp\left[-\frac{\gamma\eta}{1-\alpha}\left(k^{1-\alpha} - (i+1)^{1-\alpha}\right)\right]. \tag{C.1}$$

Note that $\exp\left(-\frac{\gamma\eta}{1-\alpha}k^{1-\alpha}\right)$ is decreasing with k, so

$$\sum_{k=i+1}^{j} \exp\left(-\frac{\gamma\eta}{1-\alpha}k^{1-\alpha}\right) \leq \int_{i+1}^{j} \exp\left(-\frac{\gamma\eta}{1-\alpha}k^{1-\alpha}\right) dk \lesssim \int_{(i+1)^{1-\alpha}}^{k^{1-\alpha}} \exp\left(-\frac{\gamma\eta}{1-\alpha}t\right) t^{\frac{\alpha}{1-\alpha}} dt.$$

For any $1 \le a \le b$ and any $1 < \beta$, we have by elementary manipulation that

$$\int_a^b e^{-x} x^\beta dx \le \int_a^\infty e^{-x} x^\beta dx \lesssim a^\beta e^{-a} C_\beta,$$

where C_β is a constant depending only on β. Then we have

$$\sum_{k=i+1}^{j} \exp\left(-\frac{\gamma\eta}{1-\alpha} k^{1-\alpha}\right) \lesssim \exp\left(-\frac{\gamma\eta}{1-\alpha}(i+1)^{1-\alpha}\right)(i+1)^\alpha. \tag{C.2}$$

Combining (C.1) and (C.2),

$$\|S_i^j\|_2 \le \exp\left(\frac{\gamma\eta}{1-\alpha}(i+1)^{1-\alpha}\right) \sum_{k=i+1}^{j} \exp\left(-\frac{\gamma\eta}{1-\alpha} k^{1-\alpha}\right) \lesssim (i+1)^\alpha.$$

□

Lemma C.1.3. *With definition of Y_i^j in Lemma C.1.1, sequence U_n can be rewritten as*

$$U_k = (\mathbf{I} - \eta_k A) U_{k-1} + \eta_k \epsilon_k = Y_s^k U_s + \sum_{p=s+1}^{k} Y_p^k \eta_p \epsilon_p.$$

According to Lemma B.3 in Chen et al. [2020], we have

$$\mathbb{E}\|U_k\|_2^2 \lesssim k^{-\alpha}.$$

Lemma C.1.4. *Let $a_m = \left\lfloor C m^\beta \right\rfloor, m \ge 2$ ($a_1 = 1$), for some constant $C > 0$ and $\beta > 1/(1-\alpha)$. For $a_M \le n < a_{M+1}$, define $n_m = a_{m+1} - a_m, 1 \le m < M$, and $n_M = n - a_M + 1$. We have*

1.

$$\lim_{M\to\infty} \frac{\sum_{i=1}^n l_i}{\sum_{i=1}^{a_{M+1}-1} l_i} = 1. \tag{C.3}$$

2.

$$\frac{(a_{M+1}-a_M)^2}{\sum_{m=1}^{M}(a_{m+1}-a_m)^2} \lesssim M^{-1}, \text{ and } \frac{a_M^{\alpha}}{n_M} \to 0. \tag{C.4}$$

Proof. Since $n \geq a_M$, we have

$$\sum_{i=1}^{n} l_i \geq \sum_{i=1}^{a_M-1} l_i = \sum_{m=1}^{M-1}\sum_{i=a_m}^{a_{m+1}-1}(i-a_m+1) = \sum_{m=1}^{M-1}\frac{n_m(n_m+1)}{2}.$$

Also,

$$\sum_{i=1}^{a_{M+1}-1} l_i = \sum_{m=1}^{M}\frac{n_m(n_m+1)}{2}.$$

Then according to the choice of a_k, we have

$$\lim_{M\to\infty}\frac{\sum_{i=1}^{n} l_i}{\sum_{i=1}^{a_{M+1}-1} l_i} \geq 1 - \frac{n_M(n_M+1)}{\sum_{m=1}^{M} n_m(n_m+1)} = \lim_{M\to\infty}(1-M^{-1}) = 1. \tag{C.5}$$

Since $\sum_{i=1}^{n} l_i \leq \sum_{i=1}^{a_{M+1}-1} l_i$, the limit is 1. Equation (C.4) is easy to verify by using the form of a_k. □

C.2 The Linear Case

Recall that the error $\delta_n = x_n - x^*$ takes the form:

$$\delta_n = \delta_{n-1} - \eta_n \nabla F(x_{n-1}) + \eta_n \epsilon_n, \tag{C.6}$$

where $\epsilon_n = \nabla F(x_{n-1}) - \nabla f(x_{n-1}, \xi_n)$. The sequence $\{\epsilon_n\}$ is a martingale difference sequence since

$$\mathbb{E}_{n-1}\epsilon_n = \nabla F(x_{n-1}) - \mathbb{E}_{n-1}\nabla f(x_{n-1}, \xi_n) = 0. \tag{C.7}$$

Note that $\nabla F(x^*) = 0$ since x^* is the minimizer of $F(x)$. By Taylor's expansion of $\nabla F(x_{n-1})$ around x^*, we have $\nabla F(x_{n-1}) \approx \nabla A\delta_{n-1}$, where $A = \nabla^2 F(x^*)$. Thus, modifying equation

(C.6) with $\nabla F(x_{n-1})$ approximated by $A\delta_{n-1}$, we have for large n

$$\delta_n \approx (\mathbf{I} - \eta_n A)\delta_{n-1} + \eta_n \epsilon_n. \tag{C.8}$$

Inspired by (C.8), we define the linear sequence $(U_n)_{n\in\mathbb{N}}$ as follows:

$$U_n = (\mathbf{I} - \eta_n A)U_{n-1} + \eta_n \epsilon_n, \quad U_0 = \delta_0. \tag{C.9}$$

Now we define a new estimator $\tilde{\Sigma}_n$ based on U_n:

$$\tilde{\Sigma}_n = \frac{1}{\sum_{i=1}^n l_i} \sum_{i=1}^n \left(\sum_{k=t_i}^{i} U_k - l_i \bar{U}_n \right) \left(\sum_{k=t_i}^{i} U_k - l_i \bar{U}_n \right)^T . \tag{C.10}$$

In certain cases when $\nabla F(x_{n-1}) = \nabla A\delta_{n-1}$, such as mean estimation model and linear regression model, error δ_n exactly takes the form of U_n. Then we have $\hat{\Sigma}_n = \tilde{\Sigma}_n$. In general cases, we can use U_n to approximate δ_n since the difference between them is small. In other words, studying covariance matrix of $\bar{U}_n$ can give us insight into the covariance matrix of $\bar{x}_n$. Next lemma shows that the estimator $\tilde{\Sigma}_n$ is consistent. It can be viewed as a special case of linear processes.

Lemma C.2.1. *Let $a_m = \left\lfloor Cm^\beta \right\rfloor$, where $C > 0$ and $\beta > 1/(1-\alpha)$. Set step size at the i-th iteration $\eta_i = \eta i^{-\alpha}$ with $\frac{1}{2} < \alpha < 1$. Then under Assumptions 4.3.2 and 4.3.3,*

$$\mathbb{E}\left\| \tilde{\Sigma}_n - \Sigma \right\|_2 \lesssim M^{-\alpha\beta/2} + M^{-1/2} + M^{((\alpha-1)\beta+1)/2}, \tag{C.11}$$

where M is the number of batches such that $a_M \leq n < a_{M+1}$.

Proof. Recall that

$$\tilde{\Sigma}_n = \left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i U_k - l_i \bar{U}_n\right) \left(\sum_{k=t_i}^i U_k - l_i \bar{U}_n\right)^T.$$

Using triangle inequality we have

$$\begin{aligned}\mathbb{E}\left\|\tilde{\Sigma}_n - \Sigma\right\|_2 \leq \mathbb{E}&\left\|\left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i U_k\right)\left(\sum_{k=t_i}^i U_k\right)^T - \Sigma\right\|_2 \\ &+ \mathbb{E}\left\|\left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n l_i^2 \bar{U}_n \bar{U}_n^T\right\|_2 + 2\mathbb{E}\left\|\left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i U_k\right)(l_i \bar{U}_n)^T\right\|_2.\end{aligned} \tag{C.12}$$

By Lemmas C.2.3, C.2.4 and C.2.5 (proved in the rest of this section), all these three terms in (C.12) are bounded, which implies Lemma C.2.1. □

Let

$$\hat{S}_n = \left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i \epsilon_k\right)\left(\sum_{k=t_i}^i \epsilon_k\right)^T.$$

In Lemma C.2.2, we show that $\hat{S}_n$ converges to S, the covariance matrix of $\nabla f(x^*, \xi)$. Using this fact, we have Lemma C.2.3, which provides an upper bound for the first term in (C.12). The other two terms in (C.12) are bounded by Lemma C.2.4 and C.2.5 respectively.

Lemma C.2.2. *Let $a_M \leq n < a_{M+1}$. Under conditions in Lemma C.2.1, we have*

$$\mathbb{E}\left\|\hat{S}_n - S\right\|_2 \lesssim M^{-\alpha\beta/2} + M^{-1/2}. \tag{C.13}$$

Proof. Here we introduce sequence $\{\epsilon_n^*\}$ as follows

$$\epsilon_n^* = \nabla F(x^*) - \nabla f(x^*, \xi_n) = -\nabla f(x^*, \xi_n), n \geq 1.$$

Note that $\{\epsilon_n^*\}$ is a sequence of $i.i.d$ variables with mean 0, and therefore $\{\epsilon_n - \epsilon_n^*\}$ is still a martingale difference sequence. We further define

$$\hat{S}_n^* = \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \left(\sum_{k=t_i}^{i} \epsilon_k^*\right) \left(\sum_{k=t_i}^{i} \epsilon_k^*\right)^T .$$

Then we can bound $\mathbb{E}\|\hat{S}_n - S\|_2$ through triangle inequality

$$\mathbb{E}\|\hat{S}_n - S\|_2 \leq \mathbb{E}\|\hat{S}_n^* - S\|_2 + \mathbb{E}\|\hat{S}_n - \hat{S}_n^*\|_2. \tag{C.14}$$

Step 1: Bound $\mathbb{E}\|\hat{S}_n^* - S\|_2$.

Since $\hat{S}_n^* - S$ is symmetric,

$$\mathbb{E}\|\hat{S}_n^* - S\|_2 = \mathbb{E}|\lambda_{max}(\hat{S}_n^* - S)| = \mathbb{E}\sqrt{\lambda_{max}(\hat{S}_n^* - S)^2}. \tag{C.15}$$

Note that $(\hat{S}_n^* - S)^2$ is positive semidefinite. For any positive semidefinite matrix C we have $\lambda_{max}(C) \leq \mathrm{tr}(C) \leq d\|C\|_2$. So $\lambda_{max}(\hat{S}_n^* - S)^2 \leq \mathrm{tr}(\hat{S}_n^* - S)^2$. Further using Jensen's inequality, we have

$$\mathbb{E}\|\hat{S}_n^* - S\|_2 \leq \mathbb{E}\sqrt{\mathrm{tr}(\hat{S}_n^* - S)^2} \leq \sqrt{\mathrm{tr}\mathbb{E}(\hat{S}_n^* - S)^2} \leq \sqrt{d\|\mathbb{E}(\hat{S}_n^* - S)^2\|_2}. \tag{C.16}$$

Note that by definition of S,

$$\mathbb{E}(\hat{S}_n^*) = \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \sum_{k=t_i}^{i} \mathbb{E}\epsilon_k^* \epsilon_k^{*T} = S.$$

Then

$$\|\mathbb{E}(\hat{S}_n^* - S)^2\|_2 = \|\mathbb{E}\hat{S}_n^{*2} - S^2\|_2.$$

Note that $\mathbb{E}(\epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_3}\epsilon^{*T}_{p_4})$ is nonzero if and only if for any r there exist $r' \neq r$ such that $p_r = p_{r'}$, $r, r' \in \{1,2,3,4\}$. There are two cases we can consider. The first case is $p_1 = p_3 \neq p_2 = p_4$ or $p_1 = p_4 \neq p_2 = p_3$. This requires i and j in the same block. The second case is $p_1 = p_2$ and $p_3 = p_4$. We can expand $\mathbb{E}\hat{S}^{*2}_n$ and rewrite it into two parts,

$$\begin{aligned}\mathbb{E}\hat{S}^{*2}_n =&\mathbb{E}\left(\sum_{i=1}^n l_i\right)^{-2}\sum_{1\le i,j\le n}\left(\sum_{k=t_i}^{i}\epsilon^*_k\right)\left(\sum_{k=t_i}^{i}\epsilon^*_k\right)^T\left(\sum_{k=t_j}^{j}\epsilon^*_k\right)\left(\sum_{k=t_j}^{j}\epsilon^*_k\right)^T\\ =&\left(\sum_{i=1}^n l_i\right)^{-2} I + \left(\sum_{i=1}^n l_i\right)^{-2} II,\end{aligned} \tag{C.17}$$

where

$$\begin{aligned}I =&\mathbb{E}\sum_{m=1}^{M-1}\sum_{i=a_m}^{a_{m+1}-1}\left[2\sum_{j=a_m}^{i-1}\sum_{a_m\le p_1\neq p_2\le j}\left(\epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_1}\epsilon^{*T}_{p_2}+\epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_2}\epsilon^{*T}_{p_1}\right)\right.\\ &\left.+\sum_{a_m\le p_1\neq p_2\le i}\left(\epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_1}\epsilon^{*T}_{p_2}+\epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_2}\epsilon^{*T}_{p_1}\right)\right]\\ &+\mathbb{E}\sum_{i=a_M}^{n}\left[2\sum_{j=a_M}^{i-1}\sum_{a_M\le p_1\neq p_2\le j}\left(\epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_1}\epsilon^{*T}_{p_2}+\epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_2}\epsilon^{*T}_{p_1}\right)\right.\\ &\left.+\sum_{a_M\le p_1\neq p_2\le i}\left(\epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_1}\epsilon^{*T}_{p_2}+\epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_2}\epsilon^{*T}_{p_1}\right)\right],\end{aligned}$$

and

$$II = \sum_{i=1}^n\sum_{j=1}^n\sum_{p=t_i}^{i}\sum_{q=t_j}^{j}\mathbb{E}(\epsilon^*_p\epsilon^{*T}_p\epsilon^*_q\epsilon^{*T}_q).$$

Let $\|\mathbb{E}(\epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_3}\epsilon^{*T}_{p_4})\|_2$ be bounded by constant C for any p_r,$r \in \{1,2,3,4\}$. Then we can bound I as follows,

$$\begin{aligned}
\|I\|_2 &\leq \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \left[2 \sum_{j=a_m}^{i-1} \sum_{a_m \leq p_1 \neq p_2 \leq j} (C + C) + \sum_{a_m \leq p_1 \neq p_2 \leq i} (C + C) \right] \\
&\lesssim \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} (1 \times 2 + 2 \times 3 + ... + (l_i - 1) \times l_i) \\
&\lesssim \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} l_i^3 \lesssim \sum_{m=1}^{M} n_m^4.
\end{aligned} \tag{C.18}$$

Since $\sum_{i=1}^{n} l_i \asymp \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} l_i \asymp \sum_{m=1}^{M} n_m^2$ and $n_M^2 / \sum_{m=1}^{M} n_m^2 \lesssim M^{-1}$, we have

$$\left(\sum_{i=1}^{n} l_i \right)^{-2} \|I\|_2 \lesssim \frac{\sum_{m=1}^{M} n_m^4}{(\sum_{m=1}^{M} n_m^2)^2} \lesssim \frac{\max_{1 \leq m \leq M} n_m^2}{\sum_{m=1}^{M} n_m^2} \lesssim M^{-1}. \tag{C.19}$$

Next, note that $\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{p=t_i}^{i} \sum_{q=t_j}^{j} 1 = (\sum_{i=1}^{n} l_i)^2$. Then,

$$\begin{aligned}
&\left\| \left(\sum_{i=1}^{n} l_i \right)^{-2} II - S^2 \right\|_2 \leq \left(\sum_{i=1}^{n} l_i \right)^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{p=t_i}^{i} \sum_{q=t_j}^{j} \left\| \mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T}) - S^2 \right\|_2 \\
&\lesssim \left(\sum_{i=1}^{a_{M+1}-1} l_i \right)^{-2} \sum_{m=1}^{M} \sum_{k=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \sum_{j=a_k}^{a_{k+1}-1} \sum_{p=a_m}^{i} \sum_{q=a_k}^{j} \left\| \mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T}) - S^2 \right\|_2 .
\end{aligned} \tag{C.20}$$

We consider two cases here. One is when p and q are in the same block. Let

$$III = \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \sum_{j=a_m}^{a_{m+1}-1} \sum_{p=a_m}^{i} \sum_{q=a_m}^{j} \left\| \mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T}) - S^2 \right\|_2 .$$

Here $\|\mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T})\|_2$ is still bounded by constant C. Then we have

$$\begin{aligned}
\left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-2} III \leq & \left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-2} \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \sum_{j=a_m}^{a_{m+1}-1} \sum_{p=a_m}^{i} \sum_{q=a_m}^{j} \left(C + \left\|S^2\right\|_2\right) \\
\lesssim & \left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-2} \sum_{m=1}^{M} \left(\sum_{i=a_m}^{a_{m+1}-1} l_i\right)^2 \\
\lesssim & \frac{\sum_{m=1}^{M} n_m^4}{(\sum_{m=1}^{M} n_m^2)^2} \lesssim \frac{\max_{1\leq m\leq M} n_m^2}{\sum_{m=1}^{M} n_m^2} \lesssim M^{-1}.
\end{aligned} \tag{C.21}$$

The other case is when p and q are in different blocks. Let

$$IV = \sum_{m\neq k} \sum_{j=a_k}^{a_{k+1}-1} \sum_{i=a_m}^{a_{m+1}-1} \sum_{q=a_k}^{j} \sum_{p=a_m}^{i} \left\|\mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T}) - S^2\right\|_2.$$

Note that $\mathbb{E}(\epsilon_n^* \epsilon_n^{*T}) = S$ by definition of S and $\epsilon_p^* \epsilon_p^{*T}$ is independent of $\epsilon_q^* \epsilon_q^{*T}$, $\forall p > q$. Then for $p > q$,

$$\left\|\mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T}) - S^2\right\|_2 = \left\|\mathbb{E}(\epsilon_p^* \epsilon_p^{*T})\mathbb{E}(\epsilon_q^* \epsilon_q^{*T}) - S^2\right\|_2 = 0. \tag{C.22}$$

Then we have

$$\left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-2} IV = 0 \tag{C.23}$$

Combining (C.20), (C.21) and (C.23), we have

$$\left\|\left(\sum_{i=1}^{n} l_i\right)^{-2} II - S^2\right\|_2 \lesssim \left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-2} III + \left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-2} IV \lesssim M^{-1}. \tag{C.24}$$

Further combining (C.17), (C.19) and (C.24), we have

$$\|\mathbb{E}\hat{S}_n^{*2} - S^2\| \leq \left\| \left(\sum_{i=1}^n l_i\right)^{-2} II - S^2 \right\|_2 + \left(\sum_{i=1}^n l_i\right)^{-2} \|I\|_2 \lesssim M^{-1}. \tag{C.25}$$

Therefore

$$\mathbb{E}\|\hat{S}_n^* - S\|_2 \leq \sqrt{d\|\mathbb{E}(\hat{S}_n^* - S)^2\|_2} = \sqrt{d\|\mathbb{E}\hat{S}_n^{*2} - S^2\|_2} \lesssim M^{-1/2}.$$

Step 2: Bound $\mathbb{E}\|\hat{S}_n - \hat{S}_n^*\|_2$.

Let $v_k = \epsilon_k - \epsilon_k^*, k \geq 1$. We can expand $\mathbb{E}\|\hat{S}_n - \hat{S}_n^*\|_2$ as

$$\begin{aligned}\mathbb{E}\|\hat{S}_n - \hat{S}_n^*\|_2 =& \mathbb{E}\left\| \left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left[\left(\sum_{k=t_i}^i \epsilon_k\right)\left(\sum_{k=t_i}^i \epsilon_k\right)^T - \left(\sum_{k=t_i}^i \epsilon_k^*\right)\left(\sum_{k=t_i}^i \epsilon_k^*\right)^T \right] \right\| \\ \leq& 2\mathbb{E}\left\| \left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i v_k\right)\left(\sum_{k=t_i}^i \epsilon_k^*\right)^T \right\|_2 + \mathbb{E}\left\| \left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i v_k\right)\left(\sum_{k=t_i}^i v_k\right)^T \right\|_2.\end{aligned} \tag{C.26}$$

Apply Cauchy's inequality

$$\begin{aligned}&\mathbb{E}\left\| \left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i v_k\right)\left(\sum_{k=t_i}^i \epsilon_k^*\right)^T \right\|_2 \\ \leq& \sqrt{\mathbb{E}\|\hat{S}_n^*\|_2} \sqrt{\mathbb{E}\left\| \left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i v_k\right)\left(\sum_{k=t_i}^i v_k\right)^T \right\|_2}.\end{aligned} \tag{C.27}$$

Then we only need to bound $\mathbb{E}\left\|\left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i v_k\right)\left(\sum_{k=t_i}^i v_k\right)^T\right\|_2$. By triangle inequality and the fact $\|C\|_2 \leq \mathrm{tr}(C)$ for any positive semi-definite matrix C,

$$\begin{aligned}\mathbb{E}\left\|\left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n}\left(\sum_{k=t_i}^{i} v_k\right)\left(\sum_{k=t_i}^{i} v_k\right)^T\right\|_2 &\leq \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \mathbb{E}\mathrm{tr}\left(\left(\sum_{k=t_i}^{i} v_k\right)\left(\sum_{k=t_i}^{i} v_k\right)^T\right) \\ &= \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \mathbb{E}\left\|\sum_{k=t_i}^{i} v_k\right\|_2^2 .\end{aligned} \tag{C.28}$$

Note that the sequence $\{v_k\}$ is still a martingale difference sequence since

$$\mathbb{E}_{k-1} v_k = \mathbb{E}_{k-1}\epsilon_k - \mathbb{E}_{k-1}\epsilon_k^* = 0.$$

Then we have

$$\mathbb{E}\| \sum_{k=t_i}^{i} v_k\|_2^2 = \sum_{k=t_i}^{i} \mathbb{E}\|v_k\|_2^2.$$

We also have

$$\begin{aligned}\mathbb{E}\|v_k\|_2^2 = \mathbb{E}\|\epsilon_k - \epsilon_k^*\|_2^2 &= \mathbb{E}\|\nabla F(x_{k-1}) - \nabla F(x^*) - (\nabla f(x_{k-1}, \xi_k) - \nabla f(x^*, \xi_k))\|_2^2 \\ &\leq 2\mathbb{E}\|\nabla F(x_{k-1}) - \nabla F(x^*)\|_2^2 + 2\mathbb{E}\|\nabla f(x_{k-1}, \xi_k) - \nabla f(x^*, \xi_k)\|_2^2 \\ &\lesssim \mathbb{E}\|x_{k-1} - x^*\|_2^2 \lesssim (k-1)^{-\alpha}.\end{aligned} \tag{C.29}$$

The second last inequality comes from Lipschitz continuity of objective function (here we also assume $f(x, \xi)$ is Lipschitz continuous with respect to the first argument x). Last inequality comes from Lemma 4.3.4. Then we have

$$\begin{aligned}
&\mathbb{E}\left\|\left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\left(\sum_{k=t_i}^{i} v_k\right)\left(\sum_{k=t_i}^{i} v_k\right)^T\right\|_2 \le \left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{k=t_i}^{i} v_k\right\|_2^2 \\
&\le \left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\sum_{k=t_i}^{i}\mathbb{E}\|v_k\|_2^2 \lesssim \left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\sum_{k=t_i}^{i}(k-1)^{-\alpha} \qquad \text{(C.30)} \\
&\le \left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{m=1}^{M}\sum_{i=a_m}^{a_{m+1}-1} l_i(a_m-1)^{-\alpha}.
\end{aligned}$$

Since

$$\sum_{i=1}^{n} l_i \asymp \sum_{m=1}^{M} n_m^2, \sum_{m=1}^{M}\sum_{i=a_m}^{a_{m+1}-1} l_i(a_m-1)^{-\alpha} \asymp \sum_{m=1}^{M} n_m^2 a_m^{-\alpha},$$

we have

$$\mathbb{E}\left\|\left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\left(\sum_{k=t_i}^{i} v_k\right)\left(\sum_{k=t_i}^{i} v_k\right)^T\right\|_2 \lesssim M^{-\alpha\beta}.$$

Then

$$\mathbb{E}\|\hat{S}_n - \hat{S}_n^*\|_2 \lesssim M^{-\alpha\beta/2}.$$

Finally, we reach the result

$$\mathbb{E}\|\hat{S}_n - S\|_2 \lesssim \mathbb{E}\|\hat{S}_n^* - S\|_2 + \mathbb{E}\|\hat{S}_n - \hat{S}_n^*\|_2 \lesssim M^{-\alpha\beta/2} + M^{-1/2}. \qquad \text{(C.31)}$$

□

Lemma C.2.3. *Under conditions in Lemma C.2.1, we have*

$$\mathbb{E}\left\|\left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\left(\sum_{k=t_i}^{i} U_k\right)\left(\sum_{k=t_i}^{i} U_k\right)^T - \Sigma\right\|_2 \le M^{-\alpha\beta/2} + M^{-1/2} + M^{((\alpha-1)\beta+1)/2}, \qquad \text{(C.32)}$$

where $a_M \le n < a_{M+1}$.

Proof. With the formula of U_k in Lemma C.1.3, for $k \in [t_i, i]$ we have

$$U_k = Y_{t_i-1}^k U_{t_i-1} + \sum_{p=t_i}^{k} Y_p^k \eta_p \epsilon_p.$$

With definition of S_j^k, we have

$$\sum_{k=t_i}^{i} U_k = \sum_{k=t_i}^{i} \left(Y_{t_i-1}^k U_{t_i-1} + \sum_{p=t_i}^{k} Y_p^k \eta_p \epsilon_p \right) = S_{t_i-1}^i U_{t_i-1} + \sum_{p=t_i}^{i} (\mathbf{I} + S_p^i) \eta_p \epsilon_p.$$

Then we have the following expansion:

$$\begin{aligned}
&\left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \left(\sum_{k=t_i}^{i} U_k\right) \left(\sum_{k=t_i}^{i} U_k\right)^T \\
=& \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \left(S_{t_i-1}^i U_{t_i-1} + \sum_{p=t_i}^{i} (\mathbf{I} + S_p^i) \eta_p \epsilon_p \right) \left(S_{t_i-1}^i U_{t_i-1} + \sum_{p=t_i}^{i} (\mathbf{I} + S_p^i) \eta_p \epsilon_p \right)^T \\
=& \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \left(A^{-1} \left(\sum_{p=t_i}^{i} \epsilon_p\right) \left(\sum_{p=t_i}^{i} \epsilon_p\right)^T A^{-1} + B_i A_i^T + A_i B_i^T + B_i B_i^T \right),
\end{aligned} \tag{C.33}$$

where $A_i = \sum_{p=t_i}^{i} A^{-1}\epsilon_p$ and $B_i = S_{t_i-1}^{i} U_{t_i-1} + \sum_{p=t_i}^{i}(\eta_p S_p^i + \eta_p \mathbf{I} - A^{-1})\epsilon_p$. We then have

$$\begin{aligned}
&\mathbb{E}\left\| \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \left(\sum_{k=t_i}^{i} U_k\right) \left(\sum_{k=t_i}^{i} U_k\right)^T - \Sigma \right\|_2 \\
\lesssim & E\left\| \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \left(A^{-1} \left(\sum_{p=t_i}^{i} \epsilon_p\right) \left(\sum_{p=t_i}^{i} \epsilon_p\right)^T A^{-1} - \Sigma \right) \right\|_2 \\
&+ \mathbb{E}\left\| \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} B_i A_i^T \right\|_2 + \mathbb{E}\left\| \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} B_i B_i^T \right\|_2 \\
= & I + II + III.
\end{aligned} \tag{C.34}$$

It is suffices to show that all three parts above can be bounded. Recall that

$$\hat{S}_n = \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \left(\sum_{k=t_i}^{i} \epsilon_k\right) \left(\sum_{k=t_i}^{i} \epsilon_k\right)^T,$$

and $\Sigma = A^{-1} S A^{-1}$. We can bound I using Lemma C.2.2.

$$I \leq \|A^{-1}\|_2^2 \mathbb{E}\|\hat{S}_n - S\|_2 \lesssim M^{-\alpha\beta/2} + M^{-1/2}. \tag{C.35}$$

For the third part III, since $B_i B_i^T$ is positive semi-definite, we have

$$\mathbb{E}\|B_i B_i^T\|_2 \leq \mathbb{E}\mathrm{tr}(B_i B_i^T) = \mathrm{tr}(\mathbb{E}(B_i B_i^T)) \leq d\|\mathbb{E}(B_i B_i^T)\|_2.$$

Since ϵ_p are martingale differences, we have $\mathbb{E}(U_{a_m-1}\epsilon_p^T) = 0$ for any $p \geq a_m$ and $\mathbb{E}(\epsilon_{p_1}\epsilon_{p_2}^T) =$

0 for any $p_1 \neq p_2$. So,

$$\begin{aligned}\left\|\mathbb{E}(B_i B_i^T)\right\|_2 &= \Bigg\|S_{a_m-1}^i \mathbb{E}(U_{a_m-1} {U_{a_m-1}}^T) {S_{a_m-1}^i}^T \\ &\quad + \sum_{p=a_m}^{i} (\eta_p S_p^i + \eta_p \mathbf{I} - A^{-1}) \mathbb{E}(\epsilon_p \epsilon_p^T)(\eta_p S_p^i + \eta_p \mathbf{I} - A^{-1})^T \Bigg\|_2 \\ &\leq \left\|S_{a_m-1}^i\right\|_2^2 \left\|\mathbb{E}(U_{a_m-1} U_{a_m-1}^T)\right\|_2 + \sum_{p=a_m}^{i} \left\|\eta_p S_p^i + \eta_p \mathbf{I} - A^{-1}\right\|_2^2 \left\|\mathbb{E}(\epsilon_p \epsilon_p^T)\right\|_2 .\end{aligned} \tag{C.36}$$

From Lemmas C.1.2 and C.1.3, we can see that $\left\|S_{a_m-1}^i\right\|_2^2 \lesssim a_m^{2\alpha}$ and

$$\|\mathbb{E}(U_{a_m-1} U_{a_m-1}^T)\|_2 \lesssim \mathrm{tr}\mathbb{E}(U_{a_m-1} U_{a_m-1}^T) \lesssim \mathbb{E}\mathrm{tr}(U_{a_m-1} U_{a_m-1}^T) \lesssim E\|U_{a_m-1}\|_2^2 \lesssim (a_m - 1)^{-\alpha}.$$

So we have

$$\left\|S_{a_m-1}^i\right\|_2^2 \left\|\mathbb{E}(U_{a_m-1} U_{a_m-1}^T)\right\|_2 \lesssim a_m^{\alpha}.$$

For the remaining part in (C.36), $\left\|\mathbb{E}(\epsilon_p \epsilon_p^T)\right\|_2$ is bounded and

$$\sum_{p=a_m}^{i} \|\eta_p S_p^i + \eta_p \mathbf{I} - A^{-1}\|_2^2 \lesssim \sum_{p=a_m}^{i} \left(\|\eta_p S_p^i - A^{-1}\|_2^2 + \|\eta_p \mathbf{I}\|_2^2 \right). \tag{C.37}$$

Next, we need to bound $\|\eta_p S_p^i - A^{-1}\|_2^2$. When $\eta_j = \eta j^{-\alpha}$ and $a_m \leq p \leq i < a_{m+1}$, based on Lemma D.2 (3) in Chen et al. [2020], we have

$$\|\eta_p S_p^i - A^{-1}\|_2^2 \lesssim p^{2\alpha-2} + \exp\left(-2\gamma \sum_{j=p}^{i} \eta_j\right).$$

Also,

$$\sum_{p=a_m}^{i} \exp\left(-2\gamma \sum_{j=p}^{i} \eta_j\right) \leq \sum_{p=a_m}^{i} \exp\left(-2\gamma\eta(i-p)i^{-\alpha}\right) \leq \sum_{k=0}^{\infty} \exp\left(-2\gamma\eta i^{-\alpha} k\right).$$

Note that $\int_0^\infty e^{-ax}dx = a^{-1}$. Then we can use integration to bound the summation above as

$$\sum_{k=0}^{\infty} \exp\left(-2\gamma\eta i^{-\alpha} k\right) \leq \int_0^\infty \exp\left(-2\gamma\eta i^{-\alpha} k\right) \lesssim i^{\alpha}.$$

Furthermore, $p^{2\alpha-2} \geq p^{-2\alpha}$ since $\alpha > 1/2$. So

$$\sum_{p=a_m}^{i} \left\| \eta_p S_p^i + \eta_p \mathbf{I} - A^{-1} \right\|_2^2 \lesssim l_i a_m^{2\alpha-2} + i^{\alpha}.$$

Recall the definition of B_i, when $t_i = a_m$

$$\|\mathbb{E}(B_i B_i^T)\|_2 \lesssim i^{\alpha} + l_i a_m^{2\alpha-2}. \tag{C.38}$$

Now since $\sum_{i=1}^{n} l_i \asymp \sum_{m=1}^{M} n_m^2$, we can bound III as follows:

$$\begin{aligned} III &\lesssim \left(\sum_{m=1}^{M} n_m^2\right)^{-1} \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \mathbb{E}\left\|B_i B_i^T\right\|_2 \\ &\leq \left(\sum_{m=1}^{M} n_m^2\right)^{-1} \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \left(i^{\alpha} + l_i a_m^{2\alpha-2}\right) \\ &\lesssim \left(\sum_{m=1}^{M} n_m^2\right)^{-1} \sum_{m=1}^{M} \left(n_m^2 a_m^{2\alpha-2} + n_m a_m^{\alpha}\right). \end{aligned} \tag{C.39}$$

Recall that $a_m \asymp m^{\beta}$ and $n_m \asymp m^{\beta-1}$, we then have

$$III \lesssim a_M^{2\alpha-2} + \frac{a_M^{\alpha}}{n_M} \lesssim M^{(\alpha-1)\beta+1}. \tag{C.40}$$

For the second part, using Cauchy's inequality we have

$$II \leq {}^{1}\sqrt{\frac{\sum_{i=1}^{n}\mathbb{E}\left\|A_iA_i^T\right\|_2}{\sum_{i=1}^{n} l_i}\frac{\sum_{i=1}^{n}\mathbb{E}\left\|B_iB_i^T\right\|_2}{\sum_{i=1}^{n} l_i}}. \tag{C.41}$$

We already have the bound for $(\sum_{i=1}^{n} l_i)^{-1}\sum_{i=1}^{n}\mathbb{E}\left\|B_iB_i^T\right\|_2$. To finish the proof, the only term remained to bound is $(\sum_{i=1}^{n} l_i)^{-1}\sum_{i=1}^{n}\mathbb{E}\left\|A_iA_i^T\right\|_2$. Recall the definition of $A_i = \sum_{p=a_m}^{i} A^{-1}\epsilon_p$ when $a_m \leq i < a_{m+1}$. Since $A_iA_i^T$ is positive semi-definite, we have

$$\mathbb{E}\|A_iA_i^T\|_2 \leq \mathbb{E}\mathrm{tr}(A_iA_i^T) = \mathrm{tr}\left(\mathbb{E}(A_iA_i^T)\right) = \mathrm{tr}\left(A^{-1}\mathbb{E}\left(\left(\sum_{p=a_m}^{i}\epsilon_p\right)\left(\sum_{p=a_m}^{i}\epsilon_p^T\right)\right)A^{-T}\right). \tag{C.42}$$

When $q \neq q$, we have $\mathbb{E}(\epsilon_p\epsilon_q^T) = 0$. Furthermore, Let $\mathbb{E}_{n-1}(\epsilon_n\epsilon_n^T) - S = \Sigma_1(\delta_{n-1})$. Then,

$$\begin{aligned}
\mathbb{E}\|A_iA_i^T\|_2 &\leq \mathrm{tr}\left(A^{-1}\left(\sum_{p=a_m}^{i} S + \mathbb{E}\Sigma_1(\delta_{p-1})\right)A^{-T}\right)\\
&= \mathbb{E}\mathrm{tr}\left(A^{-1}\left(l_iS + \sum_{p=a_m}^{i}\Sigma_1(\delta_{p-1})\right)A^{-T}\right)\\
&\lesssim \mathbb{E}\left\|A^{-1}\left(l_iS + \sum_{p=a_m}^{i}\Sigma_1(\delta_{p-1})\right)A^{-T}\right\|_2\\
&\lesssim l_i\|S\|_2 + \sum_{p=a_m}^{i}\mathbb{E}\left\|\Sigma_1(\delta_{p-1})\right\|_2.
\end{aligned} \tag{C.43}$$

In Assumption 4.3.3, we have $\|\Sigma_1(\delta)\|_2 \leq C(\|\delta\|_2 + \|\delta\|_2^2)$ for any δ. Also Lemma 4.3.4 shows that $\mathbb{E}\|\delta_n\|_2 \leq n^{-\alpha/2}(1 + \|\delta_0\|_2)$ and $\mathbb{E}\|\delta_n\|_2^2 \leq n^{-\alpha}(1 + \|\delta_0\|_2^2)$. Then we can further bound

[1] Apply Cauchy's inequality twice: $\mathbb{E}|\sum_{i=1}^{n} x_iy_i| \leq \mathbb{E}\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2)} \leq \sqrt{\sum_{i=1}^{n}\mathbb{E}x_i^2\sum_{i=1}^{n}\mathbb{E}y_i^2}$.

$\mathbb{E}\|A_iA_i^T\|_2$ as

$$\begin{aligned}\mathbb{E}\|A_iA_i^T\|_2 &\lesssim l_i\|S\|_2 + \sum_{p=a_m}^{i}\left(\mathbb{E}\left\|\delta_{p-1}\right\|_2 + \mathbb{E}\left\|\delta_{p-1}\right\|_2^2\right)\\ &\lesssim l_i + \sum_{p=a_m}^{i}(p-1)^{-\alpha/2}\\ &\lesssim l_i + l_i(a_m-1)^{-\alpha/2} \lesssim l_i.\end{aligned} \tag{C.44}$$

Then we can bound the remaining term as

$$\left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\mathbb{E}\left\|A_iA_i^T\right\|_2 \lesssim O(1). \tag{C.45}$$

Combining (C.45) and the bound of III, we have $II \lesssim M^{((\alpha-1)\beta+1)/2}$.

Now, all three parts I, II, III are bounded by $M^{-\alpha\beta/2} + M^{-1/2} + M^{((\alpha-1)\beta+1)/2}$. □

Lemma C.2.4. *Under the same conditions in Lemma C.2.1, we have*

$$\mathbb{E}\left\|\left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n} l_i^2\bar{U}_n\bar{U}_n^T\right\|_2 \lesssim M^{-1}, \tag{C.46}$$

where $a_M \leq n < a_{M+1}$.

Proof. Since $\bar{U}_n\bar{U}_n^T$ is positive semi-definite, we have

$$\mathbb{E}\left\|\bar{U}_n\bar{U}_n^T\right\|_2 \leq \mathbb{E}\mathrm{tr}\left(\bar{U}_n\bar{U}_n^T\right) = n^{-2}\mathrm{tr}\left(\mathbb{E}\left(\sum_{i=1}^{n} U_i\right)\left(\sum_{i=1}^{n} U_i\right)^T\right). \tag{C.47}$$

Recall that $U_i = Y_0^iU_0 + \sum_{p=1}^{i} Y_p^i\eta_p\epsilon_p$, then

$$\sum_{i=1}^{n} U_i = \sum_{i=1}^{n}\left(Y_0^iU_0 + \sum_{p=1}^{i} Y_p^i\eta_p\epsilon_p\right) = S_0^nU_0 + \sum_{p=1}^{n}\left(\mathbf{I} + S_p^n\right)\eta_p\epsilon_p.$$

Note that ϵ_p are martingale differences. We have the following inequality after plugging in the expansion above:

$$\begin{aligned}
\mathbb{E}\left\|\bar{U}_n\bar{U}_n^T\right\|_2 &\leq n^{-2}\mathrm{tr}\left(\mathbb{E}\left(\left(S_0^n U_0 + \sum_{p=1}^{n}\left(\mathbf{I}+S_p^n\right)\eta_p\epsilon_p\right)\left(S_0^n U_0 + \sum_{p=1}^{n}\left(\mathbf{I}+S_p^n\right)\eta_p\epsilon_p\right)^T\right)\right)\\
&= n^{-2}\mathrm{tr}\left(S_0^n\mathbb{E}\left(U_0U_0^T\right)S_0^{nT} + \sum_{p=1}^{n}\left(\mathbf{I}+S_p^n\right)\eta_p^2\mathbb{E}(\epsilon_p\epsilon_p^T)\left(\mathbf{I}+S_p^n\right)^T\right)\\
&= n^{-2}\left(\|S_0^n\|_2^2\mathbb{E}\|U_0\|_2^2 + \sum_{p=1}^{n}\left\|\left(\mathbf{I}+S_p^n\right)\right\|_2^2\eta_p^2\mathbb{E}\|\epsilon_p\|_2^2\right).
\end{aligned} \tag{C.48}$$

In Lemma C.1.2 we show that $\|S_i^j\|_2 \lesssim (i+1)^\alpha$. So here we have $\|S_0^n\|_2^2 = O(1)$ and

$$\sum_{p=1}^{n}\left\|\left(\mathbf{I}+S_p^n\right)\right\|_2^2\eta_p^2 \lesssim O(n).$$

Since $\mathbb{E}\|U_0\|_2^2$ and $\mathbb{E}\|\epsilon_p\|_2^2$ are bounded, we have

$$\mathbb{E}\left\|\bar{U}_n\bar{U}_n^T\right\|_2 \lesssim O(n^{-1}). \tag{C.49}$$

Note that

$$\frac{\sum_{i=1}^n l_i^2}{\sum_{i=1}^n l_i} \leq \frac{\sum_{i=1}^n l_i \max_{k\leq M}(a_{k+1}-a_k)}{\sum_{i=1}^n l_i} \leq n_M. \tag{C.50}$$

Since $n_M = M^{\beta-1}$ and $n \asymp M^{1/\beta}$, we have

$$\mathbb{E}\left\|\left(\sum_{i=1}^n l_i\right)^{-1}\sum_{i=1}^n l_i^2\bar{U}_n\bar{U}_n^T\right\|_2 \leq \frac{\sum_{i=1}^n l_i^2}{\sum_{i=1}^n l_i}\mathbb{E}\left\|\bar{U}_n\bar{U}_n^T\right\|_2 \lesssim n_M n^{-1} \asymp M^{-1}.$$

□

Lemma C.2.5. *Under conditions in Lemma C.2.1, for* $a_M \leq n < a_{M+1}$*, we have*

$$\mathbb{E}\left\|\left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n}\left(\sum_{k=t_i}^{i} U_k\right)\left(l_i \bar{U}_n\right)^T\right\|_2 \lesssim M^{-1/2}. \tag{C.51}$$

Proof. Apply Cauchy's inequality twice we have

$$\begin{aligned} &\mathbb{E}\left\|\left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n}\left(\sum_{k=t_i}^{i} U_k\right)\left(l_i \bar{U}_n\right)^T\right\|_2 \\ \leq^2 &\sqrt{\frac{\mathbb{E}\left\|\sum_{i=1}^{n}\left(\sum_{k=t_i}^{i} U_k\right)\left(\sum_{k=t_i}^{i} U_k\right)^T\right\|_2}{\sum_{i=1}^{n} l_i} \frac{\mathbb{E}\left\|\sum_{i=1}^{n} l_i^2 \bar{U}_n \bar{U}_n^T\right\|_2}{\sum_{i=1}^{n} l_i}}. \end{aligned} \tag{C.52}$$

In Lemma C.2.4, we already have $\mathbb{E}\| \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} l_i^2 \bar{U}_n \bar{U}_n^T \|_2 \lesssim M^{-1}$. Moreover, the L_2 norm of $(\sum_{k=a_m}^{i} U_k)(\sum_{k=a_m}^{i} U_k)^T$ is less than or equal to its trace since it is positive semi-definite. Then we have LHS of the above equation bounded by

$$O(M^{-\frac{1}{2}})\sqrt{\left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \mathbb{E}\mathrm{tr}\left(\left(\sum_{k=t_i}^{i} U_k\right)\left(\sum_{k=t_i}^{i} U_k\right)^T\right)}.$$

Let

$$I = \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \mathbb{E}\mathrm{tr}\left(\left(\sum_{k=t_i}^{i} U_k\right)\left(\sum_{k=t_i}^{i} U_k\right)^T\right).$$

To show Lemma C.2.5, it is suffices to show $I \lesssim O(1)$. Note that

$$\lim_{M\to\infty} \sum_{i=1}^{n} l_i / \sum_{i=1}^{a_M - 1} l_i = 1$$

[2] Apply Cauchy's inequality twice: $\mathbb{E}|\sum_{i=1}^{n} x_i y_i| \leq \mathbb{E}\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2)} \leq \sqrt{\sum_{i=1}^{n} \mathbb{E}x_i^2 \sum_{i=1}^{n} \mathbb{E}y_i^2}$.

and $\operatorname{tr}((\sum_{k=t_i}^{i} U_k)(\sum_{k=t_i}^{i} U_k)^T) \geq 0$. Plug $U_k = Y_0^k U_0 + \sum_{p=1}^{k} Y_p^k \eta_p \epsilon_p$ into I, we have

$$
\begin{aligned}
I \lesssim & \left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-1} \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \operatorname{tr}\left(\left(\sum_{k=a_m}^{i} Y_0^k\right) \mathbb{E}(U_0 U_0^T) \left(\sum_{k=a_m}^{i} Y_0^k\right)^T\right) \\
& + \left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-1} \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \sum_{p=1}^{i} \operatorname{tr}\left(\left(\sum_{k=(a_m \vee p)}^{i} Y_p^k\right) \mathbb{E}(\epsilon_p \epsilon_p^T) \left(\sum_{k=(a_m \vee p)}^{i} Y_p^k\right)^T\right) \eta_p^2 \\
= & II + III,
\end{aligned}
\tag{C.53}
$$

where $(a_m \vee p) = \max(a_m, p)$. Next we shall show that both II and III are bounded by $O(1)$. The first term

$$
II = \left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-1} \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \operatorname{tr}\left(\left(\sum_{k=a_m}^{i} Y_0^k\right) \mathbb{E}(U_0 U_0^T) \left(\sum_{k=a_m}^{i} Y_0^k\right)^T\right).
$$

It can be bounded using $\operatorname{tr}(C) \leq d\|C\|_2$ as follows

$$
II \lesssim \left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-1} \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \left\|\left(\sum_{k=a_m}^{i} Y_0^k\right)\right\|_2^2 \left\|\mathbb{E}(U_0 U_0^T)\right\|_2. \tag{C.54}
$$

From Lemma C.1.2,

$$
\left\|\left(\sum_{k=a_m}^{i} Y_0^k\right)\right\|_2^2 = \left\|S_0^i - S_0^{a_m}\right\|_2^2 \lesssim \left\|S_0^i\right\|_2^2 + \left\|S_0^{a_m}\right\|_2^2 \lesssim O(1).
$$

Also note that $\left\|\mathbb{E}(U_0 U_0^T)\right\|_2 \lesssim O(1)$. Then

$$
II \lesssim \left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-1} \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} O(1) = O(1). \tag{C.55}
$$

The term III can be bounded as:

$$\begin{aligned} III &= \left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-1} \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \sum_{p=1}^{i} \operatorname{tr}\left(\left(\sum_{k=(a_m \vee p)}^{i} Y_p^k\right) \mathbb{E}(\epsilon_p \epsilon_p^T) \left(\sum_{k=(a_m \vee p)}^{i} Y_p^k\right)^T\right) \eta_p^2 \\ &\lesssim \left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-1} \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \sum_{p=1}^{i} \left\| \sum_{k=(a_m \vee p)}^{i} Y_p^k \right\|_2^2 \eta_p^2 \\ &\leq \left(\sum_{i=1}^{a_{M+1}-1} l_i\right)^{-1} \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \sum_{p=1}^{i} \left(\sum_{k=(a_m \vee p)}^{i} \|Y_p^k\|_2\right)^2 \eta_p^2. \end{aligned} \tag{C.56}$$

Let $IV = \sum_{p=1}^{i} \left(\sum_{k=\max(a_m,p)}^{i} \|Y_p^k\|_2\right)^2 \eta_p^2$. From Lemma C.1.1,

$$\|Y_i^j\|_2 \leq \exp\left[-\frac{\gamma\eta}{1-\alpha}\left(j^{1-\alpha} - (i+1)^{1-\alpha}\right)\right].$$

Then for $a_m \leq i < a_{m+1}$, we have

$$IV \leq \sum_{p=1}^{i} \left(\sum_{k=\max(a_m,p)}^{i} \exp\left(-\eta\gamma \frac{k^{1-\alpha}}{1-\alpha}\right)\right)^2 \eta_p^2 e^{\frac{2\eta\gamma}{1-\alpha} p^{1-\alpha}}. \tag{C.57}$$

Using the integration, we can further bound IV as

$$\begin{aligned} IV &\lesssim \sum_{p=1}^{i} \left(\int_{\max(a_m,p)}^{i} \exp\left(-\eta\gamma \frac{k^{1-\alpha}}{1-\alpha} \right) dk \right)^2 p^{-2\alpha} e^{\frac{2\eta\gamma}{1-\alpha} p^{1-\alpha}} \\ &\lesssim \sum_{p=1}^{i} \left(\int_{\max(a_m,p)^{1-\alpha}}^{i^{1-\alpha}} e^{-\frac{\eta\gamma}{1-\alpha} t} t^{\frac{\alpha}{1-\alpha}} dt \right)^2 p^{-2\alpha} e^{\frac{2\eta\gamma}{1-\alpha} p^{1-\alpha}} \\ &\lesssim \sum_{p=1}^{i} e^{-\frac{2\eta\gamma}{1-\alpha} \max(a_m,p)^{1-\alpha}} \max(a_m,p)^{2\alpha} p^{-2\alpha} e^{\frac{2\eta\gamma}{1-\alpha} p^{1-\alpha}} \\ &\lesssim \sum_{p=1}^{a_m-1} e^{-\frac{2\eta\gamma}{1-\alpha} (a_m^{1-\alpha} - p^{1-\alpha})} \left(\frac{a_m}{p} \right)^{2\alpha} + l_i. \end{aligned} \tag{C.58}$$

Then III is bounded by

$$III \lesssim \left(\sum_{i=1}^{a_{M+1}-1} l_i \right)^{-1} \sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \sum_{p=1}^{a_m-1} e^{-\frac{2\eta\gamma}{1-\alpha} (a_m^{1-\alpha} - p^{1-\alpha})} \left(\frac{a_m}{p} \right)^{2\alpha} + 1. \tag{C.59}$$

To show III is also bounded by $O(1)$, it is suffices to show that

$$\sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} \sum_{p=1}^{a_m-1} e^{-\frac{2\eta\gamma}{1-\alpha} (a_m^{1-\alpha} - p^{1-\alpha})} \left(\frac{a_m}{p} \right)^{2\alpha} \lesssim \sum_{i=1}^{a_{M+1}-1} l_i. \tag{C.60}$$

Using partial integration we have the following:

$$\int_1^{a_m-1} e^{\frac{2\eta\gamma}{1-\alpha} p^{1-\alpha}} p^{-2\alpha} dp = \int_{\frac{2\eta\gamma}{1-\alpha}}^{\frac{2\eta\gamma}{1-\alpha}(a_m-1)^{1-\alpha}} e^u u^{-\frac{\alpha}{1-\alpha}} du \lesssim e^{\frac{2\eta\gamma}{1-\alpha} a_{m-1}^{1-\alpha}} (a_m - 1)^{-\alpha}.$$

Then we have

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{i=a_m}^{a_{m+1}-1}\sum_{p=1}^{a_m-1} e^{-\frac{2\eta\gamma}{1-\alpha}(a_m^{1-\alpha}-p^{1-\alpha})}\left(\frac{a_m}{p}\right)^{2\alpha} \\
\lesssim &\sum_{m=1}^{M}\sum_{i=a_m}^{a_{m+1}-1} e^{-\frac{2\eta\gamma}{1-\alpha}a_m^{1-\alpha}} a_m^{2\alpha}\int_1^{a_m-1} e^{\frac{2\eta\gamma}{1-\alpha}p^{1-\alpha}}p^{-2\alpha}dp \\
\lesssim &\sum_{m=1}^{M} n_m a_m^{\alpha}.
\end{aligned} \tag{C.61}$$

Note that $a_m^\alpha \lesssim n_m$ since $\beta > 1/(1-\alpha)$, so we have the following

$$\sum_{m=1}^{M}\sum_{i=a_m}^{a_{m+1}-1}\sum_{p=1}^{a_m-1} e^{-\frac{2\eta\gamma}{1-\alpha}(a_m^{1-\alpha}-p^{1-\alpha})}\left(\frac{a_m}{p}\right)^{2\alpha} \lesssim \sum_{m=1}^{M} n_m a_m^{\alpha} \lesssim \sum_{m=1}^{M} n_m^2 \asymp \sum_{i=1}^{a_{M+1}-1} l_i. \tag{C.62}$$

□

C.3 Proof of Main Theorems

C.3.1 Proof of Theorem 4.3.5

Proof. In Lemma C.2.1, we demonstrate the convergence property of the estimator $\tilde{\Sigma}$, which is constructed based on linear process $\{U_n\}_{n\in\mathbb{N}}$. Let $s_n = \delta_n - U_n$ be the difference between the error sequence δ_n and the linear sequence U_n. It has the following recursion form:

$$\begin{aligned}
s_n &= \delta_{n-1} - \eta_n \nabla F(x_{n-1}) - (\mathbf{I} - \eta_n A)U_{n-1} \\
&= (\mathbf{I} - \eta_n A)(\delta_{n-1} - U_{n-1}) - \eta_n(\nabla F(x_{n-1}) - A\delta_{n-1}) \\
&= (\mathbf{I} - \eta_n A)s_{n-1} - \eta_n(\nabla F(x_{n-1}) - A\delta_{n-1}).
\end{aligned} \tag{C.63}$$

When n is big enough, x_{n-1} is close to the minimizer x^*. Based on Taylor's expansion around x^*, $\nabla F(x_{n-1}) \approx A\delta_{n-1}$ since $\nabla F(x^*)$ is zero. So

$$s_n \approx (\mathbf{I} - \eta_n A)s_{n-1}. \tag{C.64}$$

It takes a similar linear form as U_n and its value is small especially when δ_n is small. So the difference between U_n and δ_n, i.e., $x_n - x^*$, decays quickly as $n \to \infty$. We expect that the covariance matrix estimator $\tilde{\Sigma}$ and the recursive estimator $\hat{\Sigma}$ are asymptotically close.

To show Theorem 4.3.5, it is suffices to show that $\mathbb{E}\|\tilde{\Sigma}_n - \hat{\Sigma}_n\|_2$ can be bounded with the same order as $\mathbb{E}\|\tilde{\Sigma}_n - \Sigma\|_2$. Note that $\delta_n = x_n - x^*$ and $\hat{\Sigma}_n$ can be rewritten as

$$\hat{\Sigma}_n = \left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i \delta_k - l_i\bar{\delta}_n\right)\left(\sum_{k=t_i}^i \delta_k - l_i\bar{\delta}_n\right)^T.$$

Plug in the difference $s_n = \delta_n - U_n$, we can expand $\mathbb{E}\|\tilde{\Sigma}_n - \hat{\Sigma}_n\|_2$ as

$$\begin{aligned}\mathbb{E}\|\tilde{\Sigma}_n - \hat{\Sigma}_n\|_2 \leq 2\mathbb{E}&\left\|\left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i U_k - l_i\bar{U}_n\right)\left(\sum_{k=t_i}^i s_k - l_i\bar{s}_n\right)^T\right\|_2 \\ +\mathbb{E}&\left\|\left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i s_k - l_i\bar{s}_n\right)\left(\sum_{k=t_i}^i s_k - l_i\bar{s}_n\right)^T\right\|_2.\end{aligned} \tag{C.65}$$

We further claim that

$$\mathbb{E}\left\|\left(\sum_{i=1}^n l_i\right)^{-1} \sum_{i=1}^n \left(\sum_{k=t_i}^i s_k - l_i\bar{s}_n\right)\left(\sum_{k=t_i}^i s_k - l_i\bar{s}_n\right)^T\right\|_2 \lesssim M^{-1}. \tag{C.66}$$

Apply Cauchy's inequality twice, the first part in LHS of (C.65) can be bounded as following:

$$\begin{aligned}
&\mathbb{E}\left\|\left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\left(\sum_{k=t_i}^{i} U_k - l_i\bar{U}_n\right)\left(\sum_{k=t_i}^{i} s_k - l_i\bar{s}_n\right)^T\right\|_2 \\
&\leq \sqrt{\mathbb{E}\|\tilde{\Sigma}_n\|_2}\sqrt{\mathbb{E}\left\|\left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\left(\sum_{k=t_i}^{i} s_k - l_i\bar{s}_n\right)\left(\sum_{k=t_i}^{i} s_k - l_i\bar{s}_n\right)^T\right\|_2} \qquad (C.67)\\
&\lesssim M^{-1/2},
\end{aligned}$$

since $\sqrt{\mathbb{E}\|\tilde{\Sigma}_n\|_2}$ is bounded by some constant. Then $\mathbb{E}\|\tilde{\Sigma}_n - \hat{\Sigma}_n\|_2 \lesssim M^{-1/2}$ and we have Theorem 4.3.5. All we need to prove now is the claim in (C.66). By triangle inequality and the fact $\|C\|_2 \leq \operatorname{tr}(C)$ for any positive semi-definite matrix C,

$$\begin{aligned}
&\mathbb{E}\left\|\left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\left(\sum_{k=t_i}^{i} s_k - l_i\bar{s}_n\right)\left(\sum_{k=t_i}^{i} s_k - l_i\bar{s}_n\right)^T\right\|_2 \\
&\leq \left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\mathbb{E}\mathrm{tr}\left(\left(\sum_{k=t_i}^{i} s_k - l_i\bar{s}_n\right)\left(\sum_{k=t_i}^{i} s_k - l_i\bar{s}_n\right)^T\right) \\
&= \left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{k=t_i}^{i} s_k - l_i\bar{s}_n\right\|_2^2 \qquad (C.68)\\
&\lesssim \left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{k=t_i}^{i} s_k\right\|_2^2 + \left(\sum_{i=1}^{n} l_i\right)^{-1}\sum_{i=1}^{n} l_i^2\mathbb{E}\left\|\bar{s}_n\right\|_2^2 .
\end{aligned}$$

Note that s_n takes the form

$$s_n = (\mathbf{I} - \eta_n A)s_{n-1} - \eta_n(\nabla F(x_{n-1}) - A\delta_{n-1}), \delta_0 = 0.$$

First, we shall prove that

$$\left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} l_i^2 \mathbb{E} \left\|\bar{s}_n\right\|_2^2 = O(M^{-1}).$$

Based on the definition of Y_p^k we have

$$s_k = \sum_{p=1}^{k} Y_p^k \eta_p [A\delta_{p-1} - \nabla F(x_{p-1})],$$

and

$$\begin{aligned}\bar{s}_n &= n^{-1} \sum_{k=1}^{n} \sum_{p=1}^{k} Y_p^k \eta_p [A\delta_{p-1} - \nabla F(x_{p-1})] \\ &= n^{-1} \sum_{p=1}^{n} \left(\mathbf{I} + S_p^n\right) \eta_p [A\delta_{p-1} - \nabla F(x_{p-1})].\end{aligned}$$

By Cauchy's inequality

$$\begin{aligned}\mathbb{E} \left\|\bar{s}_n\right\|_2^2 &= n^{-2} \mathbb{E} \left\| \sum_{p=1}^{n} \left(\mathbf{I} + S_p^n\right) \eta_p [A\delta_{p-1} - \nabla F(x_{p-1})] \right\|_2^2 \\ &\leq n^{-2} \mathbb{E} \left(\sum_{p=1}^{n} \left\|\mathbf{I} + S_p^n\right\|_2 \eta_p \left\|A\delta_{p-1} - \nabla F(x_{p-1})\right\|_2 \right)^2 \qquad \text{(C.69)} \\ &\leq n^{-2} \left(\sum_{p=1}^{n} \left\|\mathbf{I} + S_p^n\right\|_2^2 \eta_p^2 \right) \left(\sum_{p=1}^{n} \mathbb{E} \left\|A\delta_{p-1} - \nabla F(x_{p-1})\right\|_2^2 \right).\end{aligned}$$

From Lemma C.1.2, $\|S_p^n\|_2 \lesssim (p+1)^{\alpha}$, and therefore $\sum_{p=1}^{n} \left\|\mathbf{I} + S_p^n\right\|_2^2 \eta_p^2 \lesssim O(n)$. By Taylor's expansion around x^*, $\left\|A\delta_p - \nabla F(x_p)\right\|_2 = O(\|\delta_p\|_2^2)$. Then using Lemma 4.3.4

$$\sum_{p=1}^{n} \mathbb{E} \left\|A\delta_{p-1} - \nabla F(x_{p-1})\right\|_2^2 \asymp \sum_{p=1}^{n} \mathbb{E}\|\delta_{p-1}\|_2^4 \lesssim \sum_{p=1}^{n} (p-1)^{-2\alpha}. \qquad \text{(C.70)}$$

Since $\alpha > 1/2$, $\sum_{p=1}^{n} (p-1)^{-2\alpha} = O(1)$. Then $\mathbb{E} \left\|\bar{s}_n\right\|_2^2 \lesssim n^{-1}$. Recall that $n_k = Ck^{\beta-1}$and

$n \asymp M^{1/\beta}$. We then have,

$$\left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} l_i^2 \mathbb{E} \left\| \bar{s}_n \right\|_2^2 \lesssim n^{-1} n_M \asymp M^{-1}. \tag{C.71}$$

Next we shall prove $\left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \mathbb{E} \left\| \sum_{k=t_i}^{i} s_k \right\|_2^2$ is bounded by $O(M^{-1})$. For $t_i \leq k \leq i$, where t_i is defined in Section 4.2, we have

$$\begin{aligned} s_k &= \prod_{p=t_i}^{k} \left(\mathbf{I} - \eta_p A\right) s_{t_i - 1} + \sum_{p=t_i}^{k} \prod_{i=p+1}^{k} \left(\mathbf{I} - \eta_i A\right) \eta_p \left(A\delta_{p-1} - \nabla F(x_{p-1})\right) \\ &= Y_{t_i-1}^{k} s_{t_i-1} + \sum_{p=t_i}^{k} Y_p^k \eta_p \left(A\delta_{p-1} - \nabla F(x_{p-1})\right), \end{aligned}$$

and

$$\sum_{k=t_i}^{i} s_k = S_{t_i-1}^{k} s_{t_i-1} + \sum_{p=t_i}^{i} \left(\mathbf{I} + S_p^i\right) \eta_p \left(A\delta_{p-1} - \nabla F(x_{p-1})\right).$$

Using triangle inequality and Cauchy's inequality ,

$$\begin{aligned} \mathbb{E} \left\| \sum_{k=t_i}^{i} s_k \right\|_2^2 &\lesssim \mathbb{E} \left(\left\| S_{t_i-1}^{i} s_{t_i-1} \right\|_2^2 + \left(\sum_{p=t_i}^{i} \left\| \mathbf{I} + S_p^i \right\|_2 \eta_p \left\| A\delta_{p-1} - \nabla F(x_{p-1}) \right\|_2 \right)^2 \right) \\ &\lesssim \left\| S_{t_i-1}^{i} \right\|_2^2 \mathbb{E} \left\| s_{t_i-1} \right\|_2^2 + \left(\sum_{p=t_i}^{i} \left\| \mathbf{I} + S_p^i \right\|_2^2 \eta_p^2 \right) \left(\sum_{p=t_i}^{i} \mathbb{E} \left\| A\delta_{p-1} - \nabla F(x_{p-1}) \right\|_2^2 \right). \end{aligned} \tag{C.72}$$

From Lemma C.1.2 $\|S_p^i\|_2 \lesssim (p+1)^\alpha$, therefore we have

$$\sum_{p=t_i}^{i} \left\| \mathbf{I} + S_p^n \right\|_2^2 \eta_p^2 \lesssim l_i.$$

According to Taylor's expansion around x^*, $\left\| A\delta_p - \nabla F(x_p) \right\|_2 = O(\|\delta_p\|_2^2)$. Then using

Lemma 4.3.4 we have,

$$\sum_{p=t_i}^{i} \mathbb{E}\left\|A\delta_{p-1} - \nabla F(x_{p-1})\right\|_2^2 \asymp \sum_{p=t_i}^{i} \mathbb{E}\|\delta_{p-1}\|_2^4 \lesssim l_i t_i^{-2\alpha}. \tag{C.73}$$

Note that $s_k = \delta_k - U_k$. From Lemma 4.3.4 and C.1.3, $\mathbb{E}\left\|\delta_k\right\|_2 \asymp \mathbb{E}\left\|U_k\right\|_2 \lesssim k^{-\alpha}$. So,

$$\mathbb{E}\left\|s_k\right\|_2^2 \leq 2\mathbb{E}\left\|\delta_k\right\|_2^2 + 2\mathbb{E}\left\|U_k\right\|_2^2 \lesssim k^{-2\alpha}.$$

Thus,

$$\mathbb{E}\left\|\sum_{k=t_i}^{i} s_k\right\|_2^2 \lesssim t_i^{2\alpha} t_i^{-2\alpha} + l_i^2 t_i^{-2\alpha} = 1 + l_i^2 t_i^{-2\alpha}.$$

Since $(\sum_{i=1}^{n} l_i)^{-1} \asymp \left(\sum_{m=1}^{M} n_m^2\right)^{-1}$ and $\alpha > 1/2$, we have

$$\begin{aligned} \left(\sum_{i=1}^{n} l_i\right)^{-1} \sum_{i=1}^{n} \mathbb{E}\left\|\sum_{k=t_i}^{i} s_k\right\|_2^2 &\lesssim \left(\sum_{m=1}^{M} n_m^2\right)^{-1} \left(\sum_{m=1}^{M} \sum_{i=a_m}^{a_{m+1}-1} (1 + l_i^2 a_m^{-2\alpha})\right) \\ &\lesssim n_M^{-1} + \left(\sum_{m=1}^{M} n_m^2\right)^{-1} \left(\sum_{m=1}^{M} n_m^3 a_m^{-2\alpha}\right) \\ &\lesssim M^{-1}. \end{aligned} \tag{C.74}$$

The claim is proved through (C.68), (C.71) and (C.74).

Finally, using the fact $M = O(n^{1/\beta})$, we can obtain the upper bound in terms of n. □

C.3.2 Proof of Theorem 4.3.8

The proof for the non-overlapping version is slightly simpler than but almost the same as that for the overlapping version. Instead of writing down the similar long proof, we will provide a high level clarification when changes are needed.

In the proof of Theorem 4.3.5, we break down the estimation error into several parts in the following form:

$$\left(\sum_{m=1}^{M}\sum_{i=a_m}^{a_{m+1}-1}|B_i|\right)^{-1}\sum_{m=1}^{M}\sum_{i=a_m}^{a_{m+1}-1}T_i, \tag{C.75}$$

where T_i is the term associated with batch B_i, the explicit formula may vary from parts to parts. In the proof for the non-overlapping version, we break down the estimation error into similar parts as above but in the form:

$$\left(\sum_{m=1}^{M}|B_{a_{m+1}-1}|\right)^{-1}\sum_{m=1}^{M}T_{a_{m+1}-1}. \tag{C.76}$$

So, in comparison with the proof for the overlapping version, fewer terms are needed to bound in the non-overlapping version proof. As we can see from previous proof, for large m, T_i's for $i \in [a_m, a_{m+1}-1]$ are usually bounded by the same order, in other words $\sum_{i=a_m}^{a_{m+1}-1} T_i$ are proportion to $T_{a_{m+1}-1}$. That means the upper bound for $\sum_{m=1}^{M} T_{a_{m+1}-1}$ can be easily generated from the upper bound for $\sum_{m=1}^{M}\sum_{i=a_m}^{a_{m+1}-1} T_i$. Since a_m are polynomially increasing, term in (C.76) and term in (C.75) are of the same order.

Next, we shall give an example to show how we can leverage pervious proofs. Define

$$\hat{S}_{n,NOL} = n^{-1}\left[\sum_{m=1}^{M-1}\left(\sum_{k=a_m}^{a_{m+1}-1}\epsilon_k\right)\left(\sum_{k=a_m}^{a_{m+1}-1}\epsilon_k\right)^T + \left(\sum_{k=a_M}^{n}\epsilon_k\right)\left(\sum_{k=a_M}^{n}\epsilon_k\right)^T\right]. \tag{C.77}$$

We shall follow the same proof of Lemma C.2.2 to show the corresponding result

$$\mathbb{E}\|\hat{S}_{n,NOL} - S\|_2 \leq M^{-\alpha\beta/2} + M^{-1/2}. \tag{C.78}$$

Proof. We define

$$\hat{S}^*_{n,NOL} = n^{-1}\left[\sum_{m=1}^{M-1}\left(\sum_{k=a_m}^{a_{m+1}-1}\epsilon^*_k\right)\left(\sum_{k=a_m}^{a_{m+1}-1}\epsilon^*_k\right)^T + \left(\sum_{k=a_M}^{n}\epsilon^*_k\right)\left(\sum_{k=a_M}^{n}\epsilon^*_k\right)^T\right].$$

Then we can bound $\mathbb{E}\|\hat{S}_{n,NOL} - S\|_2$ through triangle inequality

$$\mathbb{E}\|\hat{S}_{n,NOL} - S\|_2 \leq \mathbb{E}\|\hat{S}^*_{n,NOL} - S\|_2 + \mathbb{E}\|\hat{S}_{n,NOL} - \hat{S}^*_{n,NOL}\|_2. \tag{C.79}$$

Step 1: Bound $\mathbb{E}\|\hat{S}^*_{n,NOL} - S\|_2$.

Same as in the proof of Lemma C.2.2, we have

$$\mathbb{E}\|\hat{S}^*_{n,NOL} - S\|_2 \leq \sqrt{d\|\mathbb{E}(\hat{S}^*_{n,NOL} - S)^2\|_2}. \tag{C.80}$$

Note that by definition of S,

$$\mathbb{E}(\hat{S}^*_{n,NOL}) = n^{-1}\left[\sum_{m=1}^{M-1}\sum_{k=a_m}^{a_{m+1}-1}\mathbb{E}(\epsilon^*_k\epsilon^{*T}_k) + \sum_{k=a_M}^{n}\mathbb{E}(\epsilon^*_k\epsilon^{*T}_k)\right] = S.$$

Then

$$\|\mathbb{E}(\hat{S}^*_{n,NOL} - S)^2\|_2 = \|\mathbb{E}\hat{S}^{*2}_{n,NOL} - S^2\|_2.$$

Note that $\mathbb{E}(\epsilon_{p_1}\epsilon^T_{p_2}\epsilon_{p_3}\epsilon^T_{p_4})$ is nonzero if and only if for any r there exist $r' \neq r$ such that $p_r = p_{r'}$, $r, r' \in \{1, 2, 3, 4\}$. There are two cases we can consider. The first case is $p_1 = p_3 \neq p_2 = p_4$ or $p_1 = p_4 \neq p_2 = p_3$. This requires i and j in the same block. The second case is $p_1 = p_2$ and $p_3 = p_4$. So we can expand $\mathbb{E}\hat{S}^2_{n,NOL}$ and rewrite it into two parts,

$$\mathbb{E}\hat{S}^{*2}_{n,NOL} = n^{-2}I + n^{-2}II, \tag{C.81}$$

where

$$\begin{aligned}I =&\mathbb{E}\sum_{m=1}^{M-1}\sum_{a_m\leq p_1\neq p_2\leq a_{m+1}-1}\left(\epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_1}\epsilon^{*T}_{p_2} + \epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_2}\epsilon^{*T}_{p_1}\right)\\ &+\mathbb{E}\sum_{a_M\leq p_1\neq p_2\leq n}\left(\epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_1}\epsilon^{*T}_{p_2} + \epsilon^*_{p_1}\epsilon^{*T}_{p_2}\epsilon^*_{p_2}\epsilon^{*T}_{p_1}\right),\end{aligned}$$

$$II = \sum_{i \in SET_n} \sum_{j \in SET_n} \sum_{p=t_i}^{i} \sum_{q=t_j}^{j} \mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T}), (SET_n = \{a_2 - 1, a_3 - 1, ..., a_M - 1\} \cup \{n\}).$$

Let $\|\mathbb{E}(\epsilon_{p_1}^* \epsilon_{p_2}^{*T} \epsilon_{p_3}^* \epsilon_{p_4}^{*T})\|_2$ be bounded by constant C for any p_r,$r \in \{1, 2, 3, 4\}$. Then we can bound I as follows,

$$\|I\|_2 \leq \sum_{m=1}^{M} \left[\sum_{a_m \leq p_1 \neq p_2 \leq a_{m+1}-1} (C + C) \right] \lesssim \sum_{m=1}^{M} n_m^2. \tag{C.82}$$

Since $n \asymp M^{\beta}$, we have,

$$n^{-2} \|I\|_2 \lesssim \frac{\sum_{m=1}^{M} n_m^2}{n^2} \lesssim M^{-1}. \tag{C.83}$$

Next, notice that $\sum_{i \in SET_n} \sum_{j \in SET_n} \sum_{p=t_i}^{i} \sum_{q=t_j}^{j} 1 = n^2$. Then,

$$\begin{aligned} \left\| n^{-2} II - S^2 \right\|_2 =& n^{-2} \left\| \sum_{i \in SET_n} \sum_{j \in SET_n} \sum_{p=t_i}^{i} \sum_{q=t_j}^{j} (\mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T}) - S^2) \right\|_2 \\ \leq& n^{-2} \sum_{m=1}^{M} \sum_{k=1}^{M} \sum_{p=a_m}^{a_{m+1}-1} \sum_{q=a_k}^{a_{k+1}-1} \left\| \mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T}) - S^2 \right\|_2 \\ =& n^{-2} III + n^{-2} IV. \end{aligned} \tag{C.84}$$

We consider two cases here. One is when p and q are in the same block. Let

$$III = \sum_{m=1}^{M} \sum_{p=a_m}^{a_{m+1}-1} \sum_{q=a_m}^{a_{m+1}-1} \left\| \mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T}) - S^2 \right\|_2.$$

Here $\|\mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T})\|_2$ is still bounded by constant C. Then we have

$$\begin{aligned} n^{-2} III \leq& n^{-2} \sum_{m=1}^{M} \sum_{p=a_m}^{a_{m+1}-1} \sum_{q=a_m}^{a_{m+1}-1} \left(C + \left\| S^2 \right\|_2 \right) \\ \lesssim& n^{-2} \sum_{m=1}^{M} n_m^2 \lesssim M^{-1}. \end{aligned} \tag{C.85}$$

The other case is when p and q are in different blocks. Let

$$IV = \sum_{m \neq k} \sum_{q=a_k}^{a_{k+1}-1} \sum_{p=a_m}^{a_{m+1}-1} \left\| \mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T}) - S^2 \right\|_2 .$$

For $p > q$, we have

$$\left\| \mathbb{E}(\epsilon_p^* \epsilon_p^{*T} \epsilon_q^* \epsilon_q^{*T}) - S^2 \right\|_2 = \left\| \mathbb{E}(\epsilon_p^* \epsilon_p^{*T}) \mathbb{E}(\epsilon_q^* \epsilon_q^{*T}) - S^2 \right\|_2 = 0. \tag{C.86}$$

Therefore $IV = 0$. Combining above results, we have

$$\left\| n^{-2} II - S^2 \right\|_2 \lesssim n^{-2} III + n^{-2} IV \lesssim M^{-1}. \tag{C.87}$$

Thus,

$$\mathbb{E}\|\hat{S}^*_{n,NOL} - S\|_2 \leq \sqrt{d\mathbb{E}\|\hat{S}^{*2}_{n,NOL} - S^2\|_2} \lesssim \sqrt{n^{-2}\left\|I\right\|_2 + \left\|n^{-2}II - S^2\right\|_2} \lesssim M^{-1/2}.$$

Step 2: Bound $\mathbb{E}\|\hat{S}_{n,NOL} - \hat{S}^*_{n,NOL}\|_2$.

Let $v_k = \epsilon_k - \epsilon_k^*, k \geq 1$. We can expand $\mathbb{E}\|\hat{S}_{n,NOL} - \hat{S}^*_{n,NOL}\|_2$ as

$$\begin{aligned} &\mathbb{E}\|\hat{S}_{n,NOL} - \hat{S}^*_{n,NOL}\|_2 \\ =&\mathbb{E}\left\| n^{-1} \sum_{i \in SET_n} \left[\left(\sum_{k=t_i}^{i} \epsilon_k\right) \left(\sum_{k=t_i}^{i} \epsilon_k\right)^T - \left(\sum_{k=t_i}^{i} \epsilon_k^*\right) \left(\sum_{k=t_i}^{i} \epsilon_k^*\right)^T \right] \right\|_2 \\ \leq& 2\mathbb{E}\left\| n^{-1} \sum_{i \in SET_n} \left(\sum_{k=t_i}^{i} v_k\right) \left(\sum_{k=t_i}^{i} \epsilon_k^*\right)^T \right\|_2 + \mathbb{E}\left\| n^{-1} \sum_{i \in SET_n} \left(\sum_{k=t_i}^{i} v_k\right) \left(\sum_{k=t_i}^{i} v_k\right)^T \right\|_2 . \end{aligned} \tag{C.88}$$

Apply Cauchy's inequality

$$\mathbb{E}\left\|n^{-1}\sum_{i\in SET_n}\left(\sum_{k=t_i}^{i}v_k\right)\left(\sum_{k=t_i}^{i}\epsilon_k^*\right)^T\right\|_2$$
$$\leq\sqrt{\mathbb{E}\|\hat{S}_{n,NOL}^*\|_2}\sqrt{\mathbb{E}\left\|n^{-1}\sum_{i\in SET_n}\left(\sum_{k=t_i}^{i}v_k\right)\left(\sum_{k=t_i}^{i}v_k\right)^T\right\|_2}.$$

By triangle inequality and the fact $\|C\|_2\leq\mathrm{tr}(C)$ for any positive semi-definite matrix C,

$$\mathbb{E}\left\|n^{-1}\sum_{i\in SET_n}\left(\sum_{k=t_i}^{i}v_k\right)\left(\sum_{k=t_i}^{i}v_k\right)^T\right\|_2\leq n^{-1}\sum_{i\in SET_n}\mathbb{E}\mathrm{tr}\left(\left(\sum_{k=t_i}^{i}v_k\right)\left(\sum_{k=t_i}^{i}v_k\right)^T\right)$$
$$=n^{-1}\sum_{i\in SET_n}\mathbb{E}\left\|\sum_{k=t_i}^{i}v_k\right\|_2^2\leq n^{-1}\sum_{m=1}^{M}\sum_{k=a_m}^{a_{m+1}-1}\mathbb{E}\left\|v_k\right\|_2^2\lesssim n^{-1}\sum_{m=1}^{M}n_m(a_m-1)^{-\alpha}. \tag{C.89}$$

The last inequality comes from the fact $\mathbb{E}\|v_k\|_2^2\lesssim(k-1)^{-\alpha}$. Since $a_m\asymp m^\beta, n_m\asymp m^{\beta-1}$ and $n\asymp M_\beta$, we have

$$\mathbb{E}\left\|n^{-1}\sum_{i\in SET_n}\left(\sum_{k=t_i}^{i}v_k\right)\left(\sum_{k=t_i}^{i}v_k\right)^T\right\|_2\lesssim M^{-\alpha\beta}$$

Then

$$\mathbb{E}\|\hat{S}_{n,NOL}-\hat{S}_{n,NOL}^*\|_2\lesssim M^{-\alpha\beta/2}.$$

Finally, we reach the result

$$\mathbb{E}\|\hat{S}_{n,NOL}-S\|_2\lesssim\mathbb{E}\|\hat{S}_{n,NOL}^*-S\|_2+\mathbb{E}\|\hat{S}_{n,NOL}-\hat{S}_{n,NOL}^*\|_2\lesssim M^{-\alpha\beta/2}+M^{-1/2}. \tag{C.90}$$

□

C.4 Proof of Proposition 4.3.1

Without loss of generality, we assume $x^* = 0$. Then in the mean estimation model, the SGD ietrate x_i takes the form

$$x_i = (1 - \eta_i)x_{i-1} + \eta_i e_i, \tag{C.91}$$

where $\eta_i = i^{-\alpha}, 1/2 < \alpha < 1$. And e_i are i.i.d from $N(0, 1)$. Let $x_0 = 0$, then

$$x_i = \sum_{p=1}^{i} \prod_{k=p+1}^{i} (1 - k^{\alpha})p^{-\alpha}e_p. \tag{C.92}$$

Let $W_k = \sum_{i=a_k}^{a_{k+1}-1} x_i$, $1 \leq k \leq M - 1$, and $W_M = \sum_{i=a_M}^{n} x_i$ where M satisfies $a_M \leq n < a_{M+1}$. We can rewrite the covariance of $\sqrt{n}\bar{x}_n$ as

$$\mathrm{Var}(\sqrt{n}\bar{x}_n) = \frac{\mathbb{E}(W_1 + ... + W_M)^2}{n}. \tag{C.93}$$

We can rewrite the estimator as

$$\hat{\Sigma}_{n,NOL} = \frac{W_1^2 + W_2^2 + ... + W_M^2}{n}. \tag{C.94}$$

For simplicity, we ignore the $\bar{x}_n$ term in the estimator since $\bar{x}_n$ converge to 0 at rate of $O(n^{-1/2})$, which is much faster than the convergence rate of the variance estimator. Then

$$n\mathrm{Bias}(\hat{\Sigma}_{n,NOL}) = 2 \sum_{1 \leq f < g \leq M} \mathrm{Cov}(W_f, W_g), \tag{C.95}$$

and

$$n^2\mathrm{Var}(\hat{\Sigma}_{n,NOL}) = \sum_{f=1}^{M} \mathrm{Var}(W_f^2) + 2 \sum_{1 \leq f < g \leq M} \mathrm{Cov}(W_f^2, W_g^2). \tag{C.96}$$

Next, we shall approximate $\mathrm{Var}(W_f)$ and $\mathrm{Cov}(W_f, W_g)$, $f < g$.

$$\begin{aligned}
\mathrm{Var}(W_f) &= \sum_{i=a_f}^{a_{f+1}-1} \mathrm{Var}(x_i) + \sum_{j=a_f+1}^{a_{f+1}-1} \sum_{i=a_f}^{j-1} \mathrm{Cov}(x_i, x_j) \\
&\overset{*}{\asymp} \sum_{i=a_f}^{a_{f+1}-1} i^{-\alpha} + \sum_{j=a_f+1}^{a_{f+1}-1} j^{-\alpha}(1 - j^{-\alpha}) \sum_{k=0}^{j-a_f-1} (1 - j^{-\alpha})^k \\
&= \sum_{i=a_f}^{a_{f+1}-1} i^{-\alpha} + \sum_{j=a_f+1}^{a_{f+1}-1} (1 - j^{-\alpha}) \\
&= a_{f+1} - 1 - a_f.
\end{aligned} \tag{C.97}$$

The second line $*$ in (C.97) follows from some simple calculations with $\mathrm{Var}(x_i) \asymp i^{-\alpha}$ and $\mathrm{Cov}(x_i, x_j) \asymp j^{-\alpha}(1 - j^{-\alpha})^{j-i}$ for $i < j$. Then,

$$\begin{aligned}
\mathrm{Cov}(W_f, W_g) &= \sum_{i=a_f}^{a_{f+1}-1} \sum_{j=a_g}^{a_{g+1}-1} \mathrm{Cov}(x_i, x_j) \\
&\asymp \sum_{i=a_f}^{a_{f+1}-1} a_g^{-\alpha} \sum_{j=a_g}^{a_{g+1}-1} (1 - a_g^{-\alpha})^{j-i} \\
&= \sum_{i=a_f}^{a_{f+1}-1} a_g^{-\alpha}(1 - a_g^{-\alpha})^{a_g-i} \sum_{l=0}^{a_{g+1}-1-i} (1 - a_g^{-\alpha})^l \\
&= \sum_{i=a_f}^{a_{f+1}-1} (1 - a_g^{-\alpha})^{a_g-i} = (1 - a_g^{-\alpha})^{a_g-a_{f+1}+1}/a_{g+1}^{-a}.
\end{aligned} \tag{C.98}$$

Since W_f is normal, we have $\mathrm{Var}(W_f^2) = 2\mathrm{Var}(W_f)^2$ and $\mathrm{Cov}(W_f^2, W_g^2) = 2\mathrm{Cov}(W_f, W_g)^2$. Then,

$$\begin{aligned} n\mathrm{Bias}(\hat{\Sigma}_{n,NOL}) &\asymp \sum_{g=2}^{m} \frac{1 - a_{g+1}^{-\alpha}}{a_{g+1}^{-\alpha}} \sum_{f=1}^{g-1} (1 - a_{g+1}^{-\alpha})^{a_g - a_{f+1}} \\ &= \sum_{g=2}^{m} \frac{1 - a_{g+1}^{-\alpha}}{a_{g+1}^{-\alpha}} O(1) \asymp \sum_{g=2}^{m} \frac{1 - g^{-\alpha\beta}}{g^{-\alpha\beta}} \asymp m^{\alpha\beta+1} \asymp n^{\alpha+1/\beta}. \end{aligned} \tag{C.99}$$

Also the variance

$$\begin{aligned} n^2\mathrm{Var}(\hat{\Sigma}_{n,NOL}) &= \sum_{f=1}^{M} (a_{f+1} - a_f - 1)^2 + 2 \sum_{1 \le f < g \le M} (1 - a_g^{-\alpha})^{a_g - a_{f+1} + 1} / a_{g+1}^{-a} \\ &\asymp \sum_{f=1}^{n^{1/\beta}} f^{2\beta-2} \asymp n^{2-1/\beta}. \end{aligned} \tag{C.100}$$

Then we have the mean squared error

$$MSE(\hat{\Sigma}_{n,NOL}) = \mathrm{Bias}^2(\hat{\Sigma}_{n,NOL}) + \mathrm{Var}(\hat{\Sigma}_{n,NOL}) \asymp n^{-1/\beta} + n^{2\alpha+2/\beta-2}. \tag{C.101}$$

C.5 Simulation for stopping rule

In this section, we include a simple simulation study applying the fixed-width sequential stopping rule. We set the tolerance $\epsilon_i = 0.01$ for $i = 1, ..., d$. The rule is applied to our online approach SGD inference procedure for both linear and logistic regressions with same settings discussed in Section 4.4. We present termination iterations and coverage probabilities at termination in Table C.1.

Table C.1: Apply fixed-width sequential stopping rule with the tolerance 0.01 (discussed in Section 4.3.4). We present termination iterations and coverage probabilities at termination. Standard errors are reported in the brackets.

	Linear	
	$d = 5$	$d = 20$
Termination iteration	47,737 (13,594)	98,644 (52,424)
Coverage probabilities	0.881 (0.022)	0.906 (0.020)
	Logistic	
	$d = 5$	$d = 20$
Termination iteration	249,962 (63,507)	446,016 (84,910)
Coverage probabilities	0.865 (0.024)	0.843 (0.025)

APPENDIX D

APPENDIX FOR CHAPTER 5

D.1 Proof

For notaton simplicity in this section, for any vector $\nu = (\nu_1, \dots, \nu_m)^\top \in \mathbb{R}^m$, we use $|\nu| = \sqrt{\sum_{\ell=1}^m \nu_\ell^2}$ to denote its Euclidean norm. For any random vector $X \in \mathbb{R}^m$ and constant $q > 0$, we write $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}$ if $\mathbb{E}|X|^q < \infty$.

D.1.1 Proof of Theorem 5.3.4

Proof. Without loss of generality, we assume $x^* = 0$. Observe that

$$\begin{aligned} x_i &= x_{i-1} - \eta_i \nabla F(x_{i-1}) + \eta_i \{\nabla F(x_{i-1}) - \nabla f(x_{i-1}, \xi_i)\} \\ &= x_{i-1} - \eta_i \nabla F(x_{i-1}) + \eta_i \Delta(x_{i-1}, \xi_i), \end{aligned}$$

where $\{\Delta(x_{i-1}, \xi_i)\}_{i \in \mathbb{N}}$ is a sequence of martingale differences with respect to the filtration $\mathcal{F}_i = \sigma(\xi_1, \dots, \xi_i)$, $i \geq 1$. Then, by Assumption 5.3.1, we have

$$\begin{aligned} \|x_i\|_q^2 &\leq \|x_{i-1} - \eta_i \nabla F(x_{i-1})\|_q^2 + (q-1)\eta_i^2 \|\Delta(x_{i-1}, \xi_i)\|_q^2 \\ &\leq (1 - \eta_i c_1)\|x_{i-1}\|_q^2 + 2(q-1)\eta_i^2 (\|\Delta(0, \xi_i)\|_q^2 + \gamma^2 \|x_{i-1}\|_q^2) \\ &\leq (1 - \eta_i c_2)\|x_{i-1}\|_q^2 + 2(q-1)\eta_i^2 \|\Delta(0, \xi_i)\|_q^2, \end{aligned}$$

where κ and $\tilde{\kappa}$ are positive constants depending only on γ, q and η. Consequently, it follows that

$$\|x_i\|_q^2 \leq \prod_{k=1}^{i} (1 - \eta_k c_2)|x_0|^2 + 2(q-1)\|\Delta(0, \xi)\|_q^2 \sum_{k=1}^{i} \eta_k^2 \prod_{l=k+1}^{i} (1 - \eta_l c_2).$$

Let $y_0 = 0$ and

$$y_i = y_{i-1} - \eta_i \nabla F(y_{i-1}) + \eta_i \Delta(0, \xi_i), \quad i = 1, 2, \ldots.$$

Then, by Assumption 5.3.2,

$$\begin{aligned}\|x_i - y_i\|_q^2 &\leq (1 - \eta_i c_3)\|x_{i-1} - y_{i-1}\|_q^2 + (q-1)\eta_i^2 \|\Delta(x_{i-1}, \xi_i) - \Delta(0, \xi_i)\|_q^2 \\ &\leq (1 - \eta_i c_3)\|x_{i-1} - y_{i-1}\|_q^2 + (q-1)\eta_i^2 \gamma^2 \|x_{i-1}\|_q^2.\end{aligned}$$

Similar to Lemma 14 in Gadat and Panloup [2023], it is straightforward to derive that

$$\|\bar{x}_n - \bar{y}_n\|_q \lesssim \max\left(\frac{|x_0 - x^*|}{n}, \frac{1}{n^\beta}\right), \tag{D.1}$$

where $\bar{y}_n = n^{-1}\sum_{i=1}^n y_i$. Let $w_0 = 0$ and

$$\begin{aligned}w_i &= w_{i-1} - \eta_i \nabla^2 F(0) w_{i-1} + \eta_i \Delta(0, \xi_i) \\ &=: A_i w_{i-1} + \eta_i \Delta(0, \xi_i), \quad i = 1, 2, \ldots,\end{aligned}$$

where $A_i = I_d - \eta_i \nabla^2 F(0)$. Let $\lambda_0 = \lambda_{\min}\{\nabla^2 F(0)\} > 0$ denote the minimal eigenvalue of the Hessian matrix $\nabla^2 F(0)$. By Assumption 5.3.3, we have

$$\begin{aligned}\|y_i - w_i\|_{q/2} &= \|A_i(y_{i-1} - w_{i-1}) + \eta_i\{\nabla^2 F(0) y_{i-1} - \nabla F(y_{i-1})\}\|_{q/2} \\ &\leq (1 - \eta_i \lambda_0)\|y_{i-1} - w_{i-1}\|_{q/2} + \eta_i \|\nabla^2 F(0) y_{i-1} - \nabla F(y_{i-1})\|_{q/2} \\ &\leq (1 - \eta_i \lambda_0)\|y_{i-1} - w_{i-1}\|_{q/2} + \mathcal{L}\eta_i \|y_{i-1}\|_q^2.\end{aligned}$$

Let $\bar{w}_n = n^{-1}\sum_{i=1}^n w_i$. Then, similar to (D.1), it follows that

$$\|\bar{y}_n - \bar{w}_n\|_{q/2} \lesssim \max\left(\frac{|x_0 - x^*|}{n}, \frac{1}{n^\beta}\right).$$

Now it remains to derive the strong Gaussian approximation for $\bar{w}_n$. Notice that

$$\begin{aligned}\sum_{i=1}^n w_i &= \sum_{i=1}^n \sum_{k=1}^i \prod_{l=k+1}^i A_l \eta_k \Delta(0, \xi_k) \\ &= \sum_{k=1}^n \sum_{i=k}^n \prod_{l=k+1}^i A_l \eta_k \Delta(0, \xi_k) \\ &=: \sum_{k=1}^n D_k,\end{aligned}$$

where $D_1, \ldots, D_n$ are independent and for each $k \in \{1, \ldots, n\}$, we have

$$\begin{aligned}\|D_k\|_q &\le \eta_k \sum_{i=k}^n \prod_{l=k+1}^i (1 - \eta_l \kappa^\circ)\|\Delta(0, \xi_k)\|_q \\ &\le \eta_k \sum_{i=k}^n \exp\left(-c\kappa^\circ \sum_{l=k+1}^i l^{-\beta}\right)\|\Delta(0, \xi)\|_q.\end{aligned}$$

Elementary calculations imply that $\sum_{k=1}^n \|D_k\|_q^2 \lesssim n$ for any $\beta \in (1/2, 1)$. Consequently, by Theorem 2.1 in Mies and Steland [2023], on a sufficiently rich probability space, there exist independent random vectors $W_n \overset{\mathcal{D}}{=} \sqrt{n}\bar{w}_n$ and $Z_n \sim N(0, \Gamma_n)$ such that

$$\|W_n - Z_n\|_2 \lesssim n^{1/q-1/2}\sqrt{\log n}.$$

Putting all these pieces together, we obtain (5.12). □

D.1.2 Proof of Theorem 5.3.8

Proof. Let $n = N/K$. Under Assumption 5.3.7, we have $Z_{n,k}, k = 1, ..., K$, which are *i.i.d* Gaussian $\mathcal{N}(0, \Sigma_n)$ such that

$$(\mathbb{E}|\sqrt{n}(\hat{x}_n^{(k)} - x^*) - Z_{n,k})|^2)^{1/2} = O(\delta(n)).$$

For notation simplicity we use Z_k to denote $Z_{n,k}$ in the rest of the proof. Define

$$S = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \upsilon^\top Z_k, \quad \text{and} \quad R = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (\upsilon^\top (Z_k - \bar{Z}_K))^2}.$$

It can be shown that

$$\frac{S}{R} \sim t_{K-1}.$$

Further define

$$\hat{S}_n = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \upsilon^\top \sqrt{n}(\hat{x}_n^{(k)} - x^*), \quad \text{and} \quad \hat{R}_n = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (\upsilon^\top \sqrt{n}(\hat{x}_n^{(k)} - \bar{x}_{K,n}))^2}.$$

Then $\hat{t}_\upsilon$ can be rewritten as $\hat{t}_\upsilon = \hat{S}_n / \hat{R}_n$. Now it is suffice to show

$$\sup_t \left| \mathbb{P}\left(\frac{\hat{S}_n}{\hat{R}_n} \geq t \right) - \mathbb{P}\left(\frac{S}{R} \geq t \right) \right| \lesssim c_1(n)^{1/4}.$$

Step 1: Bound $\mathbb{E}(\hat{S}_n - S)^2$ and $\mathbb{E}(\hat{R}_n - R)^2$. We first show that $\hat{S}_n - S$ and $\hat{R}_n - R$ have the same convergence rate in Assumption 5.3.7.

Using Cauchy–Schwarz inequality and Assumption 5.3.7, we have

$$\mathbb{E}(\hat{S}_n - S)^2 \leq \sum_{k=1}^{K} \mathbb{E}|\sqrt{n}(\hat{x}_n^{(k)} - x^*) - Z_k)|^2 = \delta^2,$$

where $\delta = O(\delta(n))$, which converges to 0. Similarly, applying triangle inequality

$$(\hat{R}_n - R)^2 \leq \frac{1}{K-1}\sum_{k=1}^{K}\left[(v^\top(\sqrt{n}(\hat{x}_n^{(k)} - x^*) - Z_k) + \frac{1}{K}(S_n - S))^2\right]$$
$$\leq \frac{2}{K-1}\sum_{k=1}^{K}\left[|\sqrt{n}(\hat{x}_n^{(k)} - x^*) - Z_k)|^2 + \frac{1}{K}(S_n - S)^2\right].$$

So for R_n we still have

$$\mathbb{E}(\hat{R}_n - R)^2 \lesssim \delta^2.$$

Step 2: Bound $\mathbb{P}(|\hat{S}_n/\hat{R}_n - S/R| \geq \epsilon)$. Our next step is to bound the tail of the difference between $\hat{S}_n/\hat{R}_n$ and S/R. $\mathbb{P}(|\hat{S}_n/\hat{R}_n - S/R| \geq \epsilon)$ can be decompose as

$$\mathbb{P}\left(\left|\frac{\hat{S}_n}{\hat{R}_n} - \frac{S}{R}\right| \geq \epsilon\right) \leq \mathbb{P}\left(\left|\frac{\hat{S}_n}{\hat{R}_n} - \frac{\hat{S}_n}{R}\right| \geq \epsilon\right) + \mathbb{P}\left(\left|\frac{\hat{S}_n}{R} - \frac{S}{R}\right| \geq \epsilon\right)$$

To deal with the first term in the above inequality, we first look at $\left|\frac{1}{\hat{R}_n} - \frac{1}{R}\right|$. For any $a > 0, y > 0, z > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{\hat{R}_n} - \frac{1}{R}\right| \geq a\right) \leq& \mathbb{P}\left(\left|\frac{1}{\hat{R}_n} - \frac{1}{R}\right| \geq a, |R| \geq y, \left|\hat{R}_n - R\right| \geq z\right)\\ &+ \mathbb{P}(|R| < y) + \mathbb{P}(|\hat{R}_n - R| > z)\\ \leq& \mathbb{P}(|\hat{R}_n - R| \geq ay(y-z)) + \mathbb{P}(|R| < y) + \mathbb{P}(|\hat{R}_n - R| > z)\\ \lesssim& \frac{\delta^2}{(ay(y-z))^2} + y + \frac{\delta^2}{z^2}\end{aligned}$$

The last line is derived from Markov's inequality and the probability density function (pdf) of the chi-square distribution. By choosing $y = (\delta/a)^{2/5}, z = y/2$, when $a \leq \delta^{-2/3}$ we have

$$\frac{\delta^2}{(ay(y-z))^2} + y + \frac{\delta^2}{z^2} = \frac{4\delta^2}{a^2y^4} + y + \frac{4\delta^2}{y^2} \lesssim \left(\frac{\delta}{a}\right)^{2/5}.$$

Then we have

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\hat{S}_n}{\hat{R}_n}-\frac{\hat{S}_n}{R}\right|\geq\epsilon\right) \leq & \mathbb{P}\left(|\hat{S}_n|\left|\frac{1}{\hat{R}_n}-\frac{1}{R}\right|\geq\epsilon, \left|\frac{1}{\hat{R}_n}-\frac{1}{R}\right|\geq a\right) \\ & + \mathbb{P}\left(|\hat{S}_n|\left|\frac{1}{\hat{R}_n}-\frac{1}{R}\right|\geq\epsilon, \left|\frac{1}{\hat{R}_n}-\frac{1}{R}\right|\leq a\right) \\ \leq & \mathbb{P}\left(\left|\frac{1}{\hat{R}_n}-\frac{1}{R}\right|\geq a\right)+\mathbb{P}\left(|\hat{S}_n|\geq\frac{\epsilon}{a}\right) \\ \lesssim & \left(\frac{\delta}{a}\right)^{2/5}+\left(\frac{a}{\epsilon}\right)^2\end{aligned}$$

Similarly, for any $b > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\hat{S}_n}{R}-\frac{S}{R}\right|\geq\epsilon\right) \leq & \mathbb{P}\left(\frac{|\hat{S}_n-S|}{R}\geq\epsilon, R\geq b\right)+\mathbb{P}\left(\frac{|\hat{S}_n-S|}{R}\geq\epsilon, R\leq b\right) \\ \leq & P(|\hat{S}_n-S|\leq b\epsilon)+P(R\leq b) \\ \lesssim & \frac{\delta^2}{b^2\epsilon^2}+b.\end{aligned}$$

The last step is derived from Markov's inequality and the probability density function (pdf) of the chi-square distribution. Then combine everything we have

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\hat{S}_n}{\hat{R}_n}-\frac{S}{R}\right|\geq\epsilon\right) & \leq \mathbb{P}\left(\left|\frac{\hat{S}_n}{\hat{R}_n}-\frac{\hat{S}_n}{R}\right|\geq\epsilon\right)+\mathbb{P}\left(\left|\frac{\hat{S}_n}{R}-\frac{S}{R}\right|\geq\epsilon\right) \\ & \lesssim \left(\frac{\delta}{a}\right)^{2/5}+\left(\frac{a}{\epsilon}\right)^2+\frac{\delta^2}{b^2\epsilon^2}+b.\end{aligned}$$

Step 3: Bound $|\mathbb{P}((\hat{S}_n/\hat{R}_n)\geq t)-\mathbb{P}((S/R)\geq t)|$. Let $f_{t_{K-1}}$ denote the pdf of t_{K-1}, then for any t and $0 < \epsilon < t$,

$$\mathbb{P}(t_{K-1}\geq t-\epsilon)-\mathbb{P}(t_{K-1}\geq t)\leq \epsilon f_{t_{K-1}}(t-\epsilon)\lesssim\epsilon.$$

Then we can bound $|\mathbb{P}((\hat{S}_n/\hat{R}_n) \geq t) - \mathbb{P}((S/R) \geq t)|$ as following

$$\begin{aligned}\left|\mathbb{P}\left(\frac{\hat{S}_n}{\hat{R}_n} \geq t\right) - \mathbb{P}\left(\frac{S}{R} \geq t\right)\right| &\leq \mathbb{P}(t_{K-1} \geq t - \epsilon) - \mathbb{P}(t_{K-1} \geq t) + \mathbb{P}\left(\left|\frac{\hat{S}_n}{\hat{R}_n} - \frac{S}{R}\right| \geq \epsilon\right) \\ &\lesssim \epsilon + \left(\frac{\delta}{a}\right)^{2/5} + \left(\frac{a}{\epsilon}\right)^2 + \frac{\delta^2}{b^2\epsilon^2} + b\end{aligned}$$

Choose $a = (\delta\epsilon^5)^{1/6}$, $b = (\delta/\epsilon)^{2/3}$, $\epsilon \asymp \delta^{1/4}$, we obtain

$$\epsilon + \left(\frac{\delta}{a}\right)^{2/5} + \left(\frac{a}{\epsilon}\right)^2 + \frac{\delta^2}{b^2\epsilon^2} + b \lesssim \left(\frac{\delta}{\epsilon}\right)^{1/3} + \left(\frac{\delta}{\epsilon}\right)^{2/3} + \epsilon \lesssim \delta^{1/4}.$$

We therefore have

$$\sup_t \left|\mathbb{P}\left(\frac{\hat{S}_n}{\hat{R}_n} \geq t\right) - \mathbb{P}\left(\frac{S}{R} \geq t\right)\right| \lesssim \delta(n)^{1/4}.$$

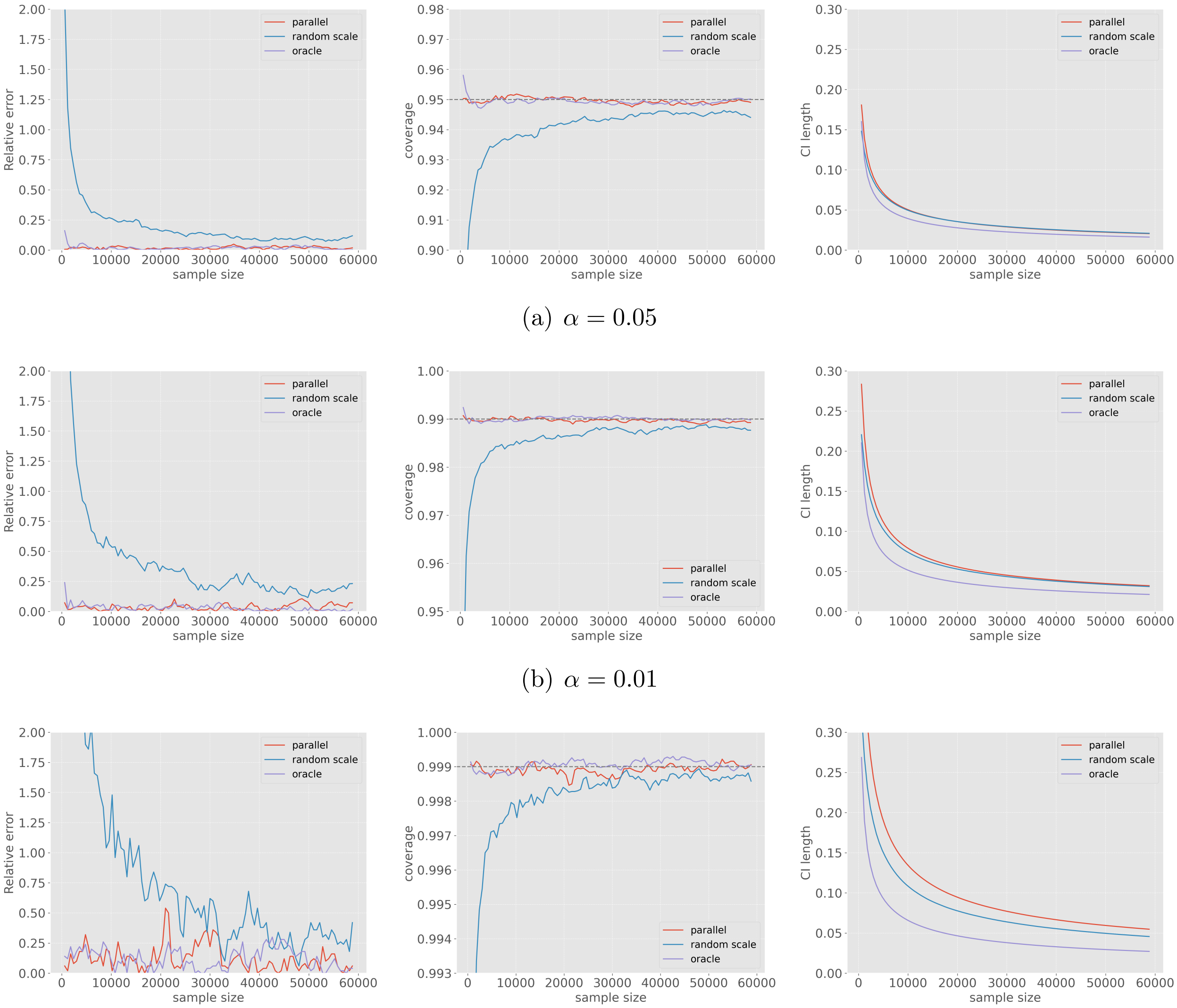
□

D.2 Additional numerical results

In Table D.1, we provide critical values used in the random scaling method. In Figure D.1, D.2 and D.3, we provide results for linear regression and logistic regression when $d = 5$ with same settings as described in Section 5.4.

Probability	97.5%	99.5%	99.95%
Critical Value	6.474	10.0544	14.76972

Table D.1: Asymptotic one-sided critical values for asymptotic pivotal statistic in the random scaling method (5.4) via Monte Carlo simulation with 1,000,000 samples.

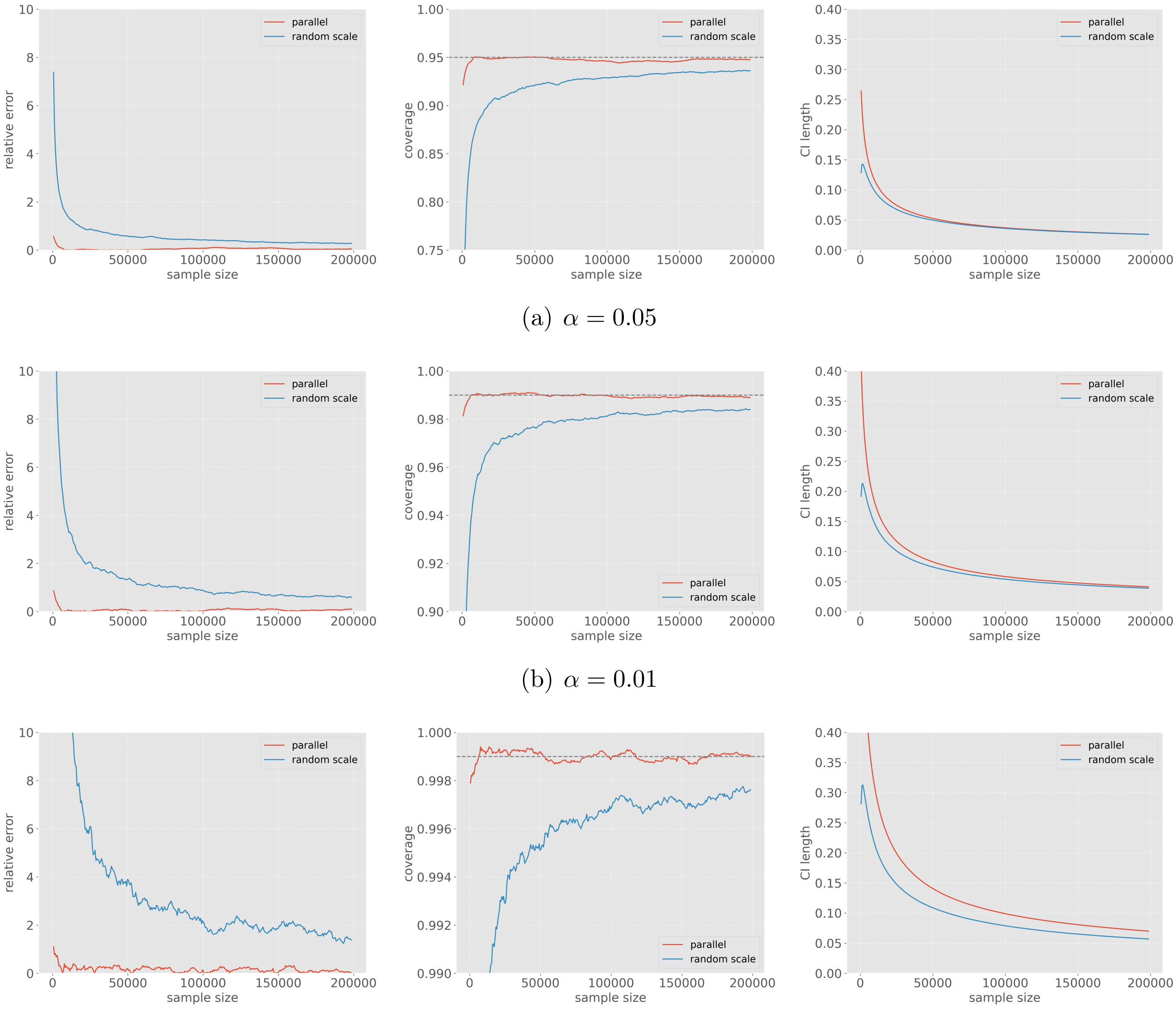


(a) $\alpha = 0.05$

(b) $\alpha = 0.01$

(c) $\alpha = 0.001$

Figure D.1: Linear Regression $d = 5$: Left: relative error of coverage; Middle: empirical coverage; Right: length of confidence intervals.

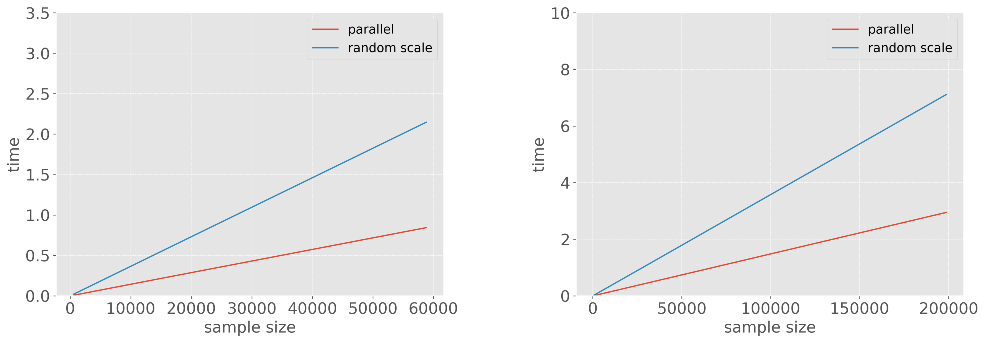


(a) $\alpha = 0.05$

(b) $\alpha = 0.01$

(c) $\alpha = 0.001$

Figure D.2: Logistic Regression $d = 5$: Left: relative error of coverage; Middle: empirical coverage; Right: length of confidence intervals.



(a) Linear, $d = 5$

(b) Logistic, $d = 5$

Figure D.3: Computation time: d = 5

APPENDIX E

APPENDIX FOR CHAPTER 6

E.1 Proofs of main results

In this section, we provide proofs for our main results: Theorems 6.4.3, 6.4.5, 6.4.7, 6.5.3 for establishing Type I error control, and Lemmas 6.3.3, 6.5.2 for computing the conditional density. For simplicity, we will write $\|\cdot\|$ to denote the usual Euclidean norm on vectors, and the operator norm on matrices.

E.1.1 Proof of Theorems 6.4.3, 6.4.5: error control for constrained aCSS

Proof. As mentioned in Section 6.4, Theorem 6.4.3 is a special case of Theorem 6.4.5, achieved by taking $k(\theta_0) = d$ and taking $v_i = \mathbf{e}_i$ for $i = 1, \ldots, d$. Therefore, it is sufficient to prove Theorem 6.4.5. Moreover, it is sufficient to bound the distance to exchangeability, since as argued in Barber and Janson [2022] we have

$$\mathbb{P}\left(\text{pval}_T(X, \tilde{X}^{(1)}, ..., \tilde{X}^{(M)}) \le \alpha\right) \le \alpha + d_{\text{exch}}(X, \tilde{X}^{(1)}, ..., \tilde{X}^{(M)}).$$

From this point on, then, we only need to establish the bound on $d_{\text{exch}}(X, \tilde{X}^{(1)}, ..., \tilde{X}^{(M)})$.

Step 1: reduce to total variation distance

We first show that we can obtain the upper bound of the distance to exchangeability through the total variation distance between $P_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g})$ and its plug-in version. This part of the proof follows the same arguments as the analogous part of the proof of [Barber and Janson, 2022, Theorem 1] for unconstrained aCSS. Let

$$\Omega_{\text{SSOSP}} = \left\{(x, w) \in X \times \mathbb{R}^d : \hat{\theta}(x, w) \text{ is a SSOSP of } (6.3)\right\},$$

and $P^*_{\theta_0}$ be the distribution of $(X, W) \sim P_{\theta_0} \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ conditional on the event $(X, W) \in \Omega_{\text{SSOSP}}$. Consider the joint distribution (a)

$$\text{Distrib. (a)} \begin{cases} (X, W) \sim P^*_{\theta_0}, \\ \hat{\theta} = \hat{\theta}(X, W), \hat{g} = \nabla(\hat{\theta}; X, W) = \nabla(\hat{\theta}; X) + \sigma W \\ \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)} \mid X, \hat{g}, \hat{\theta} \sim \tilde{P}_M(\cdot; X, \hat{\theta}, \hat{g}), \end{cases}$$

which is equivalent to the aCSS procedure conditional on the event $(X, W) \in \Omega_{\text{SSOSP}}$. On the other hand, if $(X, W) \notin \Omega_{\text{SSOSP}}$, then $\tilde{X}^{(1)} = \cdots = \tilde{X}^{(M)} = X$ according to definition and therefore $(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)})$ is exchangeable. Thus, the exchangeability is violated only on the event $(X, W) \in \Omega_{\text{SSOSP}}$. Combined with convex property of distance-to-exchangeability, we have

$$d_{\text{exch}}(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \leq d_{\text{exch}}(\text{Distribution of } X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)} \text{ under Distrib. (a)}),$$

Let $Q^*_{\theta_0}$ be the joint distribution of $(\hat{\theta}(X, W), \hat{g}(X, W))$ under $(X, W) \sim P^*_{\theta_0}$. Define distribution (b)

$$\text{Distrib. (b)} \begin{cases} (\hat{\theta}, \hat{g}) \sim Q^*_{\theta_0}, \\ X \mid \hat{\theta}, \hat{g} \sim p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}), \\ \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)} \mid X, \hat{\theta}, \hat{g} \sim \tilde{P}_M(\cdot; X, \hat{\theta}, \hat{g}), \end{cases}$$

where $p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g})$ is defined in Lemma 6.3.3. By definition of $p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g})$, it is clear that Distrib. (b) is equivalent to Distrib. (a), and then

$$d_{\text{exch}}(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \leq d_{\text{exch}}(\text{Distribution of } X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)} \text{ under Distrib. (b)}),$$

Further let $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ be the plug-in version of $p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g})$ and define

$$\text{Distrib. (c)} \begin{cases} (\hat{\theta}, \hat{g}) \sim Q^*_{\theta_0}, \\ X \mid \hat{\theta}, \hat{g} \sim p_{\hat{\theta}}(\cdot|\hat{\theta}, \hat{g}), \\ \tilde{X}^{(1)}, \dots, \tilde{X}^{(M)} \mid X, \hat{\theta}, \hat{g} \sim \tilde{P}_M(\cdot; X, \hat{\theta}, \hat{g}). \end{cases}$$

From the definition of $\tilde{P}_M$, $(X, \tilde{X}^{(1)}, \dots, \tilde{X}^{(M)})$ is exchangeable under Distrib. (c). Then,

$$d_{\text{exch}}(\text{Distribution of } X, \tilde{X}^{(1)}, \dots, \tilde{X}^{(M)} \text{ under Distrib. (b)}) \leq d_{\text{TV}}(\text{Distrib. (b)}, \text{Distrib. (c)}).$$

Since the only difference between Distrib. (b) and Distrib. (c) lies in the conditional distribution $X|\hat{\theta}, \hat{g}$,

$$d_{\text{TV}}(\text{Distrib. (b)}, \text{Distrib. (c)}) = \mathbb{E}_{Q^*_{\theta_0}}\left[d_{\text{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}))\right].$$

Therefore we can bound the distance to exchangeability as

$$d_{\text{exch}}(X, \tilde{X}^{(1)}, \dots, \tilde{X}^{(M)}) \leq \mathbb{E}_{Q^*_{\theta_0}}\left[d_{\text{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}))\right], \tag{E.1}$$

i.e., the distance to exchangeability of $X, \tilde{X}^{(1)}, \dots, \tilde{X}^{(M)}$ from the constrained aCSS procedure is bounded by the expected total variation distance between the true conditional distribution and the plug-in conditional distribution.

Step 2: bound the total variation distance

Our next step is to bound this expected total variation distance. Here our arguments will need to address a more challenging setting than the corresponding part of the proof of [Barber and Janson, 2022, Theorem 1], as we need to handle constrained rather than unconstrained optimization, as well as the issue of the sparse structure reflected by $k(\theta_0)$.

To begin, we calculate

$$
\begin{aligned}
d_{\mathrm{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})) &= \mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[\left(1 - \frac{p_{\hat{\theta}}(X \mid \hat{\theta}, \hat{g})}{p_{\theta_0}(X \mid \hat{\theta}, \hat{g})}\right)_+\right] \\
&= \mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[\left(1 - \frac{\frac{f(X;\hat{\theta})}{f(X;\theta_0)}}{\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\frac{f(X';\hat{\theta})}{f(X';\theta_0)}}\right)_+\right],
\end{aligned}
\tag{E.2}
$$

where $(x)_+ = \max\{x, 0\}$. Here the first step holds by properties of the total variation distance, while the second step holds by the density calculation in (6.7). To bound this quantity, we first want to show that $\frac{f(X;\hat{\theta})}{f(X;\theta_0)}$ is almost a constant over $p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g})$. For any x, θ, we take a Taylor series for the function $\theta \to \log f(x;\theta)$:

$$
\log f(x;\theta_0) - \log f(x;\theta) = (\theta_0 - \theta)^\top \nabla_\theta \log f(x, \theta) + \int_{t=0}^1 t(\theta - \theta_0)^\top \nabla_\theta^2 \log f(x; \theta_t)(\theta - \theta_0)\ \mathsf{d}t,
$$

where we write $\theta_t = (1-t)\theta_0 + t\theta$. Therefore, we have

$$
\begin{aligned}
&\frac{f(x;\theta)}{f(x;\theta_0)} = \exp\left\{\log f(x;\theta) - \log f(x;\theta_0)\right\} \\
&= \exp\left\{-(\theta_0 - \theta)^\top \nabla_\theta \log f(x;\theta) - \int_{t=0}^1 t(\theta - \theta_0)^\top \nabla_\theta^2 \log f(x;\theta_t)(\theta - \theta_0)\ \mathsf{d}t\right\} \\
&= \exp\left\{(\theta_0 - \theta)^\top (\nabla_\theta(x;\theta) - g) + \int_{t=0}^1 t(\theta - \theta_0)^\top \left(H(\theta_t; x) - H(\theta_t)\right)(\theta - \theta_0)\ \mathsf{d}t\right. \\
&\qquad \left. + (\theta_0 - \theta)^\top (g - \nabla_\theta R(\theta)) + \int_{t=0}^1 t(\theta - \theta_0)^\top H(\theta_t)(\theta - \theta_0)\ \mathsf{d}t\right\},
\end{aligned}
$$

where the last step holds for any fixed value $g \in \mathbb{R}^d$ (which will be chosen later), using the fact that $-\nabla_\theta \log f(x;\theta) = \nabla_\theta(x;\theta) - \nabla_\theta R(\theta)$ by definition of .

Next let $\Theta_0 = \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta \cap \{\theta : \|\theta - \theta_0\|_{v,0} \le k(\theta_0)\}$. If $\theta \in \Theta_0$, then by definition of $\|\theta - \theta_0\|_{v,0}$, there exists a subset $S(\theta, \theta_0) \subseteq [p]$ with $|S(\theta, \theta_0)| \le k(\theta_0)$, such that $(\theta - \theta_0) \in \operatorname{span}(\{v_i\}_{i \in S(\theta,\theta_0)})$. Recall that for any set $S \subseteq [p]$, $\mathcal{P}_{v_S}$ denotes the projection to span($\{v_i\}_{i \in S}$). Then we have

$$\begin{aligned}
&\left|(\theta_0 - \theta)^\top (\nabla_\theta(x;\theta) - g)\right| = \left|(\theta_0 - \theta)^\top \mathcal{P}_{v_{S(\theta,\theta_0)}} (\nabla_\theta(x;\theta) - g)\right| \\
&\qquad \leq \|\theta_0 - \theta\| \max_{S:|S| \leq k(\theta_0)} \|\mathcal{P}_{v_S}(\nabla_\theta(\theta;x) - g)\| \leq r(\theta_0) \max_{S:|S| \leq k(\theta_0)} \|\mathcal{P}_{v_S}(\nabla_\theta(\theta;x) - g)\|.
\end{aligned}$$

We also calculate, for $\theta \in \Theta_0$,

$$\begin{aligned}
&\int_{t=0}^1 t(\theta - \theta_0)^\top \left(H(\theta_t;x) - H(\theta_t)\right)(\theta - \theta_0)\, \mathsf{d}t \\
&\leq \int_{t=0}^1 t\|\theta - \theta_0\|^2 \cdot \lambda_{\max}(H(\theta_t;x) - H(\theta_t))\, \mathsf{d}t \\
&\leq \frac{1}{2} \sup_{\theta' \in \Theta_0} \left(\lambda_{\max}(H(\theta';x) - H(\theta'))\right)_+ \cdot \|\theta - \theta_0\|^2 \\
&\leq \frac{r(\theta_0)^2}{2} \sup_{\theta' \in \Theta_0} \left(\lambda_{\max}(H(\theta';x) - H(\theta'))\right)_+,
\end{aligned}$$

and similarly,

$$\int_{t=0}^1 t(\theta - \theta_0)^\top \left(H(\theta_t;x) - H(\theta_t)\right)(\theta - \theta_0)\, \mathsf{d}t \geq -\frac{r(\theta_0)^2}{2} \sup_{\theta' \in \Theta_0} \left(\lambda_{\max}(H(\theta') - H(\theta';x))\right)_+.$$

Combining all these calculations, for any $\theta \in \Theta_0$ we have

$$\begin{aligned}
\frac{f(x;\theta)}{f(x;\theta_0)} \leq \exp\Bigg\{ & r(\theta_0) \max_{S:|S| \leq k(\theta_0)} \|\mathcal{P}_{v_S}(\nabla_\theta(\theta;x) - g)\| \\
&+ \frac{r(\theta_0)^2}{2} \sup_{\theta' \in \Theta_0} \left(\lambda_{\max}(H(\theta';x) - H(\theta'))\right)_+ \\
&+ (\theta_0 - \theta)^\top (g - \nabla_\theta R(\theta)) + \int_{t=0}^1 t(\theta - \theta_0)^\top H(\theta_t)(\theta - \theta_0)\, \mathsf{d}t \Bigg\},
\end{aligned}$$

and similarly,

$$\frac{f(x;\theta)}{f(x;\theta_0)} \geq \exp\bigg\{ - r(\theta_0) \max_{S:|S|\leq k(\theta_0)} \|\mathcal{P}_{v_S}(\nabla_\theta(\theta;x) - g)\| \\ - \frac{r(\theta_0)^2}{2} \sup_{\theta'\in\Theta_0} \left(\lambda_{\max}(H(\theta') - H(\theta';x))\right)_+ \\ + (\theta_0 - \theta)^\top (g - \nabla_\theta R(\theta)) + \int_{t=0}^1 t(\theta-\theta_0)^\top H(\theta_t)(\theta-\theta_0) \;\mathsf{d}t\bigg\},$$

Now let

$$\Delta_1(\theta,g;x) = r(\theta_0) \max_{S:|S|\leq k(\theta_0)} \|\mathcal{P}_{v_S}(\nabla_\theta(\theta;x) - g)\| + \frac{r(\theta_0)^2}{2} \sup_{\theta'\in\Theta_0} \left(\lambda_{\max}\left(H(\theta';x) - H(\theta')\right)\right)_+,$$

and

$$\Delta_1'(\theta,g;x) = r(\theta_0) \max_{S:|S|\leq k(\theta_0)} \|\mathcal{P}_{v_S}(\nabla_\theta(\theta;x) - g)\| + \frac{r(\theta_0)^2}{2} \sup_{\theta'\in\Theta_0} \left(\lambda_{\max}\left(H(\theta') - H(\theta';x)\right)\right)_+.$$

Then in our work above, we have shown that

$$e^{-\Delta_1'(\theta,g;x)} \leq \frac{f(x;\theta)}{f(x;\theta_0)} \cdot e^{-(\theta_0-\theta)^\top (g-\nabla_\theta R(\theta)) - \int_{t=0}^1 t(\theta-\theta_0)^\top H(\theta_t)(\theta-\theta_0)\;\mathsf{d}t} \leq e^{\Delta_1(\theta,g;x)}$$

holds for all x, all g, and all $\theta\in\Theta_0$. This means that, for all $x,x'\in\mathcal{X}$, all g, and all $\theta\in\Theta_0$,

$$\frac{\frac{f(x';\theta)}{f(x';\theta_0)}}{\frac{f(x;\theta)}{f(x;\theta_0)}} \leq \frac{e^{\Delta_1(\theta,g;x')}}{e^{-\Delta_1'(\theta,g;x)}}.$$

In particular, on the event that $\hat\theta\in\Theta_0$, plugging in $g=\hat g$, we have

$$\frac{\frac{f(x';\hat\theta)}{f(x';\theta_0)}}{\frac{f(x;\hat\theta)}{f(x;\theta_0)}} \leq \frac{e^{\Delta_1(\hat\theta,\hat g;x')}}{e^{-\Delta_1'(\hat\theta,\hat g;x)}},$$

again for all $x, x' \in \mathcal{X}$. Taking an expected value with respect to $X' \sim p_{\theta_0}(\cdot; \hat{\theta}, \hat{g})$, then,

$$\frac{\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[\frac{f(X';\hat{\theta})}{f(X';\theta_0)}\right]}{\frac{f(x;\hat{\theta})}{f(x;\theta_0)}} = \mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[\frac{\frac{f(X';\hat{\theta})}{f(X';\theta_0)}}{\frac{f(x;\hat{\theta})}{f(x;\theta_0)}}\right]$$
$$\leq \mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[\frac{e^{\Delta_1(\hat{\theta},\hat{g};X')}}{e^{-\Delta_1'(\hat{\theta},\hat{g};x)}}\right] = \frac{\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[e^{\Delta_1(\hat{\theta},\hat{g};X')}\right]}{e^{-\Delta_1'(\hat{\theta},\hat{g};x)}}.$$

Therefore, on the event that $\hat{\theta} \in \Theta_0$, we have shown that

$$\left(1 - \frac{\frac{f(x;\hat{\theta})}{f(x;\theta_0)}}{\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[\frac{f(X';\hat{\theta})}{f(X';\theta_0)}\right]}\right)_+ \leq 1 - \frac{e^{-\Delta_1'(\hat{\theta},\hat{g};x)}}{\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[e^{\Delta_1(\hat{\theta},\hat{g};X')}\right]}.$$

(Note that the right-hand side is always nonnegative, since the functions Δ_1, Δ_1' both return only nonnegative values.) In particular, on the event that $\hat{\theta} \in \Theta_0$,

$$d_{\text{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})) = \mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[\left(1 - \frac{\frac{f(x;\hat{\theta})}{f(x;\theta_0)}}{\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[\frac{f(X';\hat{\theta})}{f(X';\theta_0)}\right]}\right)_+\right]$$
$$\leq \mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[1 - \frac{e^{-\Delta_1'(\hat{\theta},\hat{g};X)}}{\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[e^{\Delta_1(\hat{\theta},\hat{g};X')}\right]}\right].$$

Combining both cases (i.e., $\hat{\theta} \in \Theta_0$ and $\hat{\theta} \not\in \Theta_0$), we see that

$$d_{\text{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})) \leq \mathbb{1}_{\hat{\theta}\not\in\Theta_0} + \mathbb{1}_{\hat{\theta}\in\Theta_0}\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[1 - \frac{e^{-\Delta_1'(\hat{\theta},\hat{g};X)}}{\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[e^{\Delta_1(\hat{\theta},\hat{g};X')}\right]}\right].$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{Q^*_{\theta_0}}\left[d_{\mathrm{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta},\hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta},\hat{g}))\right] \\
&\leq \mathbb{P}_{Q^*_{\theta_0}}\{\hat{\theta}\notin\Theta_0\} + \mathbb{E}_{Q^*_{\theta_0}}\left[\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[1-\frac{e^{-\Delta'_1(\hat{\theta},\hat{g};x)}}{\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[e^{\Delta_1(\hat{\theta},\hat{g};X')}\right]}\right]\right] \\
&\leq \mathbb{P}_{Q^*_{\theta_0}}\{\hat{\theta}\notin\Theta_0\} + \mathbb{E}_{Q^*_{\theta_0}}\left[\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[\Delta'_1(\hat{\theta},\hat{g};X)\right]\right] + 1 - \frac{1}{\mathbb{E}_{Q^*_{\theta_0}}\left[\mathbb{E}_{p_{\theta_0}(\cdot|\hat{\theta},\hat{g})}\left[e^{\Delta_1(\hat{\theta},\hat{g};X)}\right]\right]},
\end{aligned}$$

where the last step follows the same calculation as in the analogous part of the proof of [Barber and Janson, 2022, Theorem 1]. Next, by definition, $(\hat{\theta},\hat{g}) \sim Q^*_{\theta_0}$ and $X \mid \hat{\theta},\hat{g} \sim p_{\theta_0}(\cdot \mid \hat{\theta},\hat{g})$ is equivalent to the joint distribution of $(X, \hat{\theta}(X,W), \hat{g}(X,W))$ when $(X,W) \sim P^*_{\theta_0}$. Therefore

$$\begin{aligned}
&\mathbb{E}_{Q^*_{\theta_0}}\left[d_{\mathrm{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta},\hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta},\hat{g}))\right] \\
&\leq \mathbb{P}_{P^*_{\theta_0}}\{\hat{\theta}\notin\Theta_0\} + \mathbb{E}_{P^*_{\theta_0}}\left[\Delta'_1(\hat{\theta}(X,W),\hat{g}(X,W);X)\right] + 1 - \frac{1}{\mathbb{E}_{P^*_{\theta_0}}\left[e^{\Delta_1(\hat{\theta}(X,W),\hat{g}(X,W);X)}\right]}.
\end{aligned}$$

Now define

$$\Delta_2(x,w) = r(\theta_0)\sigma \max_{S:|S|\leq k(\theta_0)} \|\mathcal{P}_{v_S} w\| + \frac{r(\theta_0)^2}{2} \sup_{\theta'\in\Theta_0} \left(\lambda_{\max}\left(H(\theta';x) - H(\theta')\right)\right)_+,$$

and

$$\Delta'_2(x,w) = r(\theta_0)\sigma \max_{S:|S|\leq k(\theta_0)} \|\mathcal{P}_{v_S} w\| + \frac{r(\theta_0)^2}{2} \sup_{\theta'\in\Theta_0} \left(\lambda_{\max}\left(H(\theta') - H(\theta';x)\right)\right)_+.$$

Observe that $\hat{g}(X,W) = \nabla(\hat{\theta};X,W) = \nabla(\hat{\theta};X) + \sigma W$ by definition, and so we must have

$$\Delta_1(\hat{\theta}(X,W),\hat{g}(X,W);X) = \Delta_2(X,W),\ \ \Delta'_1(\hat{\theta}(X,W),\hat{g}(X,W);X) = \Delta'_2(X,W).$$

Consequently,

$$\mathbb{E}_{Q^*_{\theta_0}}\left[d_{\text{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}))\right]$$
$$\leq \mathbb{P}_{P^*_{\theta_0}}\{\hat{\theta} \notin \Theta_0\} + \mathbb{E}_{P^*_{\theta_0}}\left[\Delta_2'(X, W)\right] + \left(1 - \frac{1}{\mathbb{E}_{P^*_{\theta_0}}\left[e^{\Delta_2(X,W)}\right]}\right).$$

Next let $\mathcal{E}_{\text{SSOSP}}$ be the event that $(X, W) \in \Omega_{\text{SSOSP}}$. Recall that $P^*_{\theta_0}$ is the distribution of $(X, W) \sim P_{\theta_0} \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ conditional on $\mathcal{E}_{\text{SSOSP}}$. Then, following the exact same steps as the analogous part of the proof of [Barber and Janson, 2022, Theorem 1], it holds that

$$\mathbb{E}_{Q^*_{\theta_0}}\left[d_{\text{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}))\right]$$
$$\leq \frac{\mathbb{P}\{\{\hat{\theta} \notin \Theta_0\} \cap \mathcal{E}_{\text{SSOSP}}\} + \mathbb{E}\left[\Delta_2'(X, W)\right] + \log \mathbb{E}\left[e^{\Delta_2(X,W)}\right]}{1 - \mathbb{P}\{\mathcal{E}^{\complement}_{\text{SSOSP}}\}}$$
$$\leq \frac{\delta(\theta_0) + \tilde{\delta}(\theta_0) - \mathbb{P}(\mathcal{E}^{\complement}_{\text{SSOSP}}) + \mathbb{E}\left[\Delta_2'(X, W)\right] + \log \mathbb{E}\left[e^{\Delta_2(X,W)}\right]}{1 - \mathbb{P}\{\mathcal{E}^{\complement}_{\text{SSOSP}}\}},$$

where now probability and expectation are taken with respect to $(X, W) \sim P_{\theta_0} \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$, and where the last step holds by Assumption 6.4.1, together with the assumption in the theorem.

Next, for a standard normal vector $Z \sim \mathcal{N}(0, \mathbf{I}_d)$ and 1-Lipschitz function f, we have $\log \mathbb{E}e^{\lambda f(Z)} \leq \frac{\lambda^2}{2} + \lambda \mathbb{E}[f(Z)]$ for all λ [Boucheron et al., 2013]. We can verify that $f(z) = \max_{S \subseteq [p]:|S| \leq k(\theta_0)} \|\mathcal{P}_{v_S} z\|$ is a 1-Lipschitz function, and by definition of h_v, we have $\mathbb{E}[f(Z)^2] = h_v(k(\theta_0))$. Then, since $\sqrt{d}W$ is a standard normal random vector, we have

$$\log \mathbb{E}\left[e^{2r(\theta_0)\sigma \max_{S:|S| \leq k(\theta_0)} \|\mathcal{P}_{v_S} W\|}\right] = \log \mathbb{E}\left[e^{\frac{2r(\theta_0)\sigma}{\sqrt{d}} f(\sqrt{d}W)}\right]$$
$$\leq \frac{2r(\theta_0)^2\sigma^2}{d} + 2r(\theta_0)\sigma\sqrt{\frac{h_v(k(\theta_0))}{d}}.$$

Next, we can assume that $2\sigma r(\theta_0) \leq d\sqrt{\frac{h_v(k(\theta_0))}{d}}$. (To see why, observe that $h_v(k(\theta_0)) \geq h_v(1) \geq 1$. If this inequality fails, then $3\sigma r(\theta_0)\sqrt{\frac{h_v(k(\theta_0)}{d}} \geq \frac{3\sigma r(\theta_0)}{\sqrt{d}} \geq 1$, and so the bound in the theorem holds trivially since total variation distance can never exceed 1.) Then we have

$$\log \mathbb{E}\left[e^{2r(\theta_0)\sigma \max_{S:|S|\leq k(\theta_0)} \|\mathcal{P}_{v_S} W\|}\right] \leq 3r(\theta_0)\sigma\sqrt{\frac{h_v(k(\theta_0))}{d}}.$$

Next, combining Cauchy–Schwarz and Assumption 6.4.2 we have

$$\begin{aligned}
&\log \mathbb{E}\left[e^{\Delta_2(X,W)}\right] \\
&\leq \frac{1}{2}\log \mathbb{E}\left[e^{2r(\theta_0)\sigma \max_{S:|S|\leq k(\theta_0)} \|\mathcal{P}_{v_S} W\|}\right] + \frac{1}{2}\log \mathbb{E}\left[e^{r(\theta_0)^2 \sup_{\theta' \in \Theta_0} \left(\lambda_{\max}\left(H(\theta';x) - H(\theta')\right)\right)_+}\right] \\
&\leq 1.5r(\theta_0)\sigma\sqrt{\frac{h_v(k(\theta_0))}{d}} + \frac{\epsilon(\theta_0)}{2}.
\end{aligned}$$

Similarly, by Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\Delta_2'(X,W)\right] \\
&= \mathbb{E}\left[r(\theta_0)\sigma \max_{S:|S|\leq k(\theta_0)} \|\mathcal{P}_{v_S} W\|\right] + \frac{1}{2}\mathbb{E}\left[r(\theta_0)^2 \sup_{\theta' \in \Theta_0} \left(\lambda_{\max}\left(H(\theta') - H(\theta';x)\right)\right)_+\right] \\
&\leq \frac{1}{2}\log \mathbb{E}\left[e^{2r(\theta_0)\sigma \max_{S:|S|\leq k(\theta_0)} \|\mathcal{P}_{v_S} W\|}\right] + \frac{1}{2}\mathbb{E}\left[r(\theta_0)^2 \sup_{\theta' \in \Theta_0} \left(\lambda_{\max}\left(H(\theta') - H(\theta';x)\right)\right)_+\right] \\
&\leq 1.5r(\theta_0)\sigma\sqrt{\frac{h_v(k(\theta_0))}{d}} + \frac{\epsilon(\theta_0)}{2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{Q^*_{\theta_0}}\left[d_{\mathrm{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}))\right] \\
&\qquad\qquad \leq \frac{\delta(\theta_0) + \tilde{\delta}(\theta_0) - \mathbb{P}(\mathcal{E}^{\complement}_{\mathrm{SSOSP}}) + 3\sigma r(\theta_0)\sqrt{\frac{h_v(k(\theta_0))}{d}} + \epsilon(\theta_0)}{1 - \mathbb{P}\{\mathcal{E}^{\complement}_{\mathrm{SSOSP}}\}} \\
&\qquad\qquad\qquad\qquad \leq 3\sigma r(\theta_0)\sqrt{\frac{h_v(k(\theta_0))}{d}} + \epsilon(\theta_0) + \delta(\theta_0) + \tilde{\delta}(\theta_0),
\end{aligned}$$

where to verify the last step, we can apply the fact that $\frac{a-b}{1-b} \leq a$ for any $a \in [0,1]$ and $b \in [0,1)$ (note that we can assume that $3\sigma r(\theta_0)\sqrt{\frac{h_v(k(\theta_0))}{d}} + \epsilon(\theta_0) + \delta(\theta_0) + \tilde{\delta}(\theta_0) \leq 1$, as otherwise the bound holds trivially since total variation distance can never exceed 1). This completes the proof. □

E.1.2 Proof of Theorem 6.4.7: constrained aCSS for the Gaussian linear model

Following the same reasoning as in the proof of Theorem 6.4.5, we only need to bound

$$\mathbb{E}_{Q^*_{\theta_0}}\left[d_{\mathrm{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}))\right],$$

where, as in that proof, $Q^*_{\theta_0}$ is the joint distribution of $(\hat{\theta}(X,W), \hat{g}(X,W))$ under $(X,W) \sim P^*_{\theta_0}$, where $P^*_{\theta_0}$ is the distribution of $(X,W) \sim P_{\theta_0} \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ conditional on the event $(X,W) \in \Omega_{\mathrm{SSOSP}}$. For the Gaussian case, by our assumption (6.14) on $R(\theta)$, the event $(X,W) \in \Omega_{\mathrm{SSOSP}}$ holds almost surely, and so $Q^*_{\theta_0}$ is in fact the joint distribution of $(\hat{\theta}(X,W), \hat{g}(X,W))$ under $(X,W) \sim \mathcal{N}(Z\theta_0, \nu^2\mathbf{I}_n) \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$.

Next, applying Lemma 6.3.3, we calculate

$$p_{\theta_0}(x \mid \hat{\theta}, \hat{g}) \propto \exp\left\{-\frac{1}{2\nu^2}\|x - Z\theta_0\|^2 - \frac{1}{2\sigma^2/d}\left\|\hat{g} - \left(\frac{1}{\nu^2}Z^\top(Z\hat{\theta} - x) + \nabla_\theta R(\hat{\theta})\right)\right\|^2\right\}$$

and

$$p_{\hat{\theta}}(x \mid \hat{\theta}, \hat{g}) \propto \exp\left\{-\frac{1}{2\nu^2}\|x - Z\hat{\theta}\|^2 - \frac{1}{2\sigma^2/d}\left\|\hat{g} - \left(\frac{1}{\nu^2}Z^\top(Z\hat{\theta} - x) + \nabla_\theta R(\hat{\theta})\right)\right\|^2\right\},$$

which simplifies to the normal distributions

$$\mathcal{N}\left(Z\hat{\theta} + \left(\mathbf{I}_n + \frac{d}{\sigma^2\nu^2}ZZ^\top\right)^{-1}\left[\frac{d}{\sigma^2}Z(\nabla_\theta R(\hat{\theta}) - \hat{g}) + Z(\theta_0 - \hat{\theta})\right], \nu^2\left(\mathbf{I}_n + \frac{d}{\sigma^2\nu^2}ZZ^\top\right)^{-1}\right)$$

and

$$\mathcal{N}\left(Z\hat{\theta} + \left(\mathbf{I}_n + \frac{d}{\sigma^2\nu^2}ZZ^\top\right)^{-1}\left[\frac{d}{\sigma^2}Z(\nabla_\theta R(\hat{\theta}) - \hat{g})\right], \nu^2\left(\mathbf{I}_n + \frac{d}{\sigma^2\nu^2}ZZ^\top\right)^{-1}\right),$$

respectively. For any $\mu, \mu' \in \mathbb{R}^n$ and any positive definite $\Sigma \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} d_{\mathrm{TV}}\big(\mathcal{N}(\mu,\Sigma), \mathcal{N}(\mu',\Sigma)\big) &\le \sqrt{\frac{1}{2}d_{\mathrm{KL}}\big(\mathcal{N}(\mu,\Sigma)\big\|\mathcal{N}(\mu',\Sigma)\big)} \\ &= \sqrt{\frac{1}{2}\cdot\frac{1}{2}(\mu-\mu')^\top\Sigma^{-1}(\mu-\mu')} = \frac{1}{2}\|\Sigma^{-1/2}(\mu-\mu')\|, \end{aligned}$$

where d_{KL} is the Kullback–Leibler divergence, and the first step holds by Pinsker's inequality. Applying this calculation to the distributions $p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g})$ and $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ computed above, we have

$$\begin{aligned} d_{\mathrm{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta},\hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta},\hat{g})) &\le \frac{1}{2\nu}\left\|\left(\mathbf{I}_n + \frac{d}{\sigma^2\nu^2}ZZ^\top\right)^{1/2}\cdot\left(\mathbf{I}_n + \frac{d}{\sigma^2\nu^2}ZZ^\top\right)^{-1}Z(\hat{\theta}-\theta_0)\right\| \\ &\le \frac{1}{2\nu}\left\|\left(\mathbf{I}_n + \frac{d}{\sigma^2\nu^2}ZZ^\top\right)^{-1/2}Z\right\|\cdot\|\hat{\theta}-\theta_0\| \\ &= \frac{\sigma}{2\sqrt{d}}\left\|\left(\frac{\sigma^2\nu^2}{d}\mathbf{I}_n + ZZ^\top\right)^{-1/2}Z\right\|\cdot\|\hat{\theta}-\theta_0\| \le \frac{\sigma}{2\sqrt{d}}\cdot\|\hat{\theta}-\theta_0\|. \end{aligned}$$

On the event that $\|\hat{\theta}-\theta_0\| \le r(\theta_0)$ we therefore have $d_{\mathrm{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta},\hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta},\hat{g})) \le \frac{\sigma}{2\sqrt{d}}r(\theta_0)$. Since total variation distance is always bounded by 1, and we therefore have

$$\begin{aligned} &\mathbb{E}_{Q^*_{\theta_0}}\left[d_{\mathrm{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta},\hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta},\hat{g}))\right] \\ &\qquad \le \frac{\sigma}{2\sqrt{d}}r(\theta_0)\cdot\mathbb{P}_{Q^*_{\theta_0}}\{\|\hat{\theta}-\theta_0\| \le r(\theta_0)\} + \mathbb{P}_{Q^*_{\theta_0}}\{\|\hat{\theta}-\theta_0\| > r(\theta_0)\} \\ &\qquad \le \frac{\sigma}{2\sqrt{d}}r(\theta_0) + \delta(\theta_0), \end{aligned}$$

since $\|\hat{\theta}-\theta_0\| \le r(\theta_0)$ holds with probability at least $1-\delta(\theta_0)$ by assumption.

E.1.3 Proof of Lemma 6.3.3: conditional density

We begin by introducing some notation for remaining proofs. For $A \in \mathbb{R}^{r\times d}, b \in \mathbb{R}^r$, define a subset of Θ with active set $\mathcal{I} \subseteq [r]$ as follows:

$$\Theta_{A,b,\mathcal{I}} = \{\theta \in \Theta : A_i^\top \theta = b_i, \forall i \in \mathcal{I}; A_i^\top \theta < b_i, \forall i \in [r]\backslash\mathcal{I}\},$$

where A_i is the ith row of A. We will write $\Theta_{\mathcal{I}} = \Theta_{A,b,\mathcal{I}}$ when A, b are fixed. As before, we define $\mathcal{I}(\theta) = \{i \in [r] : A_i^\top \theta = b_i\}$, the active set for a given $\theta \in \Theta$, so that we have $\theta \in \Theta_{A,b,\mathcal{I}(\theta)}$ by definition.

Before proving Lemma 6.3.3, we need a preliminary result, which we will prove below.

Lemma E.1.1. *For index set $\mathcal{I} \in [r]$, define*

$$\Omega_{\text{SSOSP},\mathcal{I}} = \left\{(x,w) \in \mathcal{X} \times \mathbb{R}^d : \hat{\theta}(x,w) \text{ is a SSOSP of (6.3), and } \mathcal{I}(\hat{\theta}(x,w)) = \mathcal{I}\right\},$$

and

$$\Psi_{\text{SSOSP},\mathcal{I}} = \Big\{(x,\theta,g) \in \mathcal{X} \times \Theta_{\mathcal{I}} \times \mathbb{R}^d : \exists w \in \mathbb{R}^d \text{ such that}$$
$$\theta = \hat{\theta}(x,w) \text{ is a SSOSP of (6.3), and } g = \hat{g}(x,w)\Big\}.$$

Define a map $\psi_{\mathcal{I}}$ from $\Omega_{\text{SSOSP},\mathcal{I}}$ as

$$\psi_{\mathcal{I}} : (x,w) \to \left(x, \hat{\theta}(x,w), \hat{g}(x,w)\right).$$

Then $\psi_{\mathcal{I}}$ is a bijection between $\Omega_{\text{SSOSP},\mathcal{I}}$ and $\Psi_{\text{SSOSP},\mathcal{I}}$ with inverse

$$\psi_{\mathcal{I}}^{-1} : (x,\theta,g) \to \left(x, \frac{g - \nabla_\theta(\theta;x)}{\sigma}\right).$$

To give intuition for this result, the bijection between $\Omega_{\text{SSOSP},\mathcal{I}}$ and $\Psi_{\text{SSOSP},\mathcal{I}}$ helps us see why we need to condition on both $\hat{\theta}$ and $\hat{g}$, rather than on $\hat{\theta}$ alone as for the (unconditional) aCSS of Barber and Janson [2022]. Intuitively, the estimator $\hat{\theta}$ itself cannot reflect enough information for data (x, w) when constraints appear in the optimization step, because $\hat{\theta}$ may have lower effective dimension (e.g., if one constraint is active, then the value of $\hat{\theta}$ has $d-1$ degrees of freedom; this means that $(x, \hat{\theta})$ cannot contain sufficient information to recover (x, w), since w is d-dimensional). In the unconstrained case, $\hat{g} \equiv 0$ due to the first-order optimality conditions, so conditioning on $(\hat{\theta}, \hat{g})$ is equivalent to simply conditioning on $\hat{\theta}$, in that case.

With this result in place, we are now ready to prove Lemma 6.3.3, which calculates the conditional density.

Proof of Lemma 6.3.3. Consider the joint distribution $(X, W) \sim P_{\theta_0} \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$. By assumption in the lemma, the event $(X, W) \in \Omega_{\text{SSOSP},\mathcal{I}}$ has positive probability. Then the joint density of (X, W), conditioning on the event that $\hat{\theta}(X, W)$ is a SSOSP of (6.3) with active set $\mathcal{I}$, i.e., $(X, W) \in \Omega_{\text{SSOSP},\mathcal{I}}$, is proportional to the function

$$h_{\theta_0}(x, w) = f(x; \theta_0) \exp\left\{-\frac{d}{2}\|w\|^2\right\} \mathbb{1}_{(x,w)\in\Omega_{\text{SSOSP},\mathcal{I}}}.$$

By Lemma E.1.1, $\psi_{\mathcal{I}}$ is a bijection between $\Omega_{\text{SSOSP},\mathcal{I}}$ and $\Psi_{\text{SSOSP},\mathcal{I}}$. For any measurable set $I_{\mathcal{I}} \subseteq \Psi_{\text{SSOSP},\mathcal{I}}$, define

$$\psi_{\mathcal{I}}^{-1}(I_{\mathcal{I}}) = \{(x, w) \in \Omega_{\text{SSOSP},\mathcal{I}} : \psi_{\mathcal{I}}(x, w) \in I_{\mathcal{I}})\}.$$

Then, we calculate

$$\begin{aligned}
&\mathbb{P}\left\{(X,\hat{\theta}(X,W),\hat{g}(X,W)) \in I_{\mathcal{I}} \mid (X,\hat{\theta}(X,W),\hat{g}(X,W)) \in \Psi_{\text{SSOSP},\mathcal{I}}\right\}\\
&= \mathbb{P}\left\{(X,W) \in \psi_{\mathcal{I}}^{-1}(I_{\mathcal{I}}) \mid (X,W) \in \Omega_{\text{SSOSP},\mathcal{I}}\right\} \text{ by Lemma E.1.1}\\
&= \frac{\int_{\psi_{\mathcal{I}}^{-1}(I_{\mathcal{I}})} h_{\theta_0}(x,w)\, \mathsf{d}\nu_{\mathcal{X}}(x)\, \mathsf{d}w}{\int_{\mathcal{X}\times\mathbb{R}^d} h_{\theta_0}(x',w')\, \mathsf{d}\nu_{\mathcal{X}}(x')\, \mathsf{d}w'}\\
&= \frac{\int_{\psi_{\mathcal{I}}^{-1}(I_{\mathcal{I}})} f(x;\theta_0)\exp\left\{-\frac{d}{2}\|w\|^2\right\} \mathbb{1}_{(x,w)\in\Omega_{\text{SSOSP},\mathcal{I}}}\, \mathsf{d}\nu_{\mathcal{X}}(x)\, \mathsf{d}w}{\int_{\mathcal{X}\times\mathbb{R}^d} h_{\theta_0}(x',w')\, \mathsf{d}\nu_{\mathcal{X}}(x')\, \mathsf{d}w'} \text{ by definition of } h_{\theta_0}(x,w)\\
&= \frac{\int_{\psi_{\mathcal{I}}^{-1}(I_{\mathcal{I}})} f(x;\theta_0)e^{-\frac{d}{2\sigma^2}\|\hat{g}(x,w)-\nabla_\theta(\hat{\theta}(x,w);x)\|^2} \mathbb{1}_{(x,w)\in\Omega_{\text{SSOSP},\mathcal{I}}}\, \mathsf{d}\nu_{\mathcal{X}}(x)\, \mathsf{d}w}{\int_{\mathcal{X}\times\mathbb{R}^d} h_{\theta_0}(x',w')\, \mathsf{d}\nu_{\mathcal{X}}(x')\, \mathsf{d}w'}\\
&= \frac{\int_{\mathcal{X}} f(x;\theta_0)\int_{\mathbb{R}^d} e^{-\frac{d}{2\sigma^2}\|\hat{g}(x,w)-\nabla_\theta(\hat{\theta}(x,w);x)\|^2} \mathbb{1}_{(x,w)\in\psi_{\mathcal{I}}^{-1}(I_{\mathcal{I}})}\, \mathsf{d}w\, \mathsf{d}\nu_{\mathcal{X}}(x)}{\int_{\mathcal{X}\times\mathbb{R}^d} h_{\theta_0}(x',w')\, \mathsf{d}\nu_{\mathcal{X}}(x')\, \mathsf{d}w'},
\end{aligned}$$

where the last step holds since $\psi_{\mathcal{I}}^{-1}(I_{\mathcal{I}}) \subseteq \Omega_{\text{SSOSP},\mathcal{I}}$.

Next, we need to reparameterize θ and g, since given the active set $\mathcal{I}$, these variables must lie in lower-dimensional subspaces of Θ and of $\mathbb{R}^d$, respectively. Let $k = \text{rank}(\text{span}(A_{\mathcal{I}})^\perp)$, let $U_{\mathcal{I}} \in \mathbb{R}^{d\times k}$ be an orthonormal basis for $\text{span}(A_{\mathcal{I}})^\perp$ as before, and let $V_{\mathcal{I}} \in \mathbb{R}^{d\times(d-k)}$ be an orthonormal basis for $\text{span}(A_{\mathcal{I}})$, so that $(U_{\mathcal{I}}\ V_{\mathcal{I}}) \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. Define $\Theta' = \{U_{\mathcal{I}}^\top\theta : \theta \in \Theta_{\mathcal{I}}\} \subseteq \mathbb{R}^k$. Then $\theta' = U_{\mathcal{I}}^\top\theta$ and $g' = V_{\mathcal{I}}^\top g$ are a reparametrization of (θ, g), which now take values in Θ' and $\mathbb{R}^{d-k}$, respectively. To see why, let $\theta_* \in \mathbb{R}^{d-k}$ be the unique value such that $\theta_* = V_{\mathcal{I}}^\top\theta$ for all $\theta \in \Theta_{\mathcal{I}}$, i.e., θ_* is determined by the active constraints (specifically, if $A_{\mathcal{I}} = MDV_{\mathcal{I}}^\top$ is a singular value decomposition, then $\theta_* = D^{-1}M^\top b_{\mathcal{I}}$). Then $\theta = U_{\mathcal{I}}\theta' + V_{\mathcal{I}}\theta_*$, and $g = V_{\mathcal{I}}g'$, whenever (θ, g) corresponds to a SSOSP with active set $\mathcal{I}$ (i.e., for any $\theta \in \Theta_{\mathcal{I}}$ and $g \in \text{span}(A_{\mathcal{I}})$).

Next, for $\theta \in \Theta_{\mathcal{I}}$ and $g \in \text{span}(A_{\mathcal{I}})$, if $(x,\theta,g) \in \Psi_{\text{SSOSP},\mathcal{I}}$ then by the SSOSP conditions we must have some w such that $\theta = \hat{\theta}(x,w)$ is a SSOSP of (6.3), and $g = \hat{g}(x,w) =$

$\nabla_\theta(\theta; x, w) = \nabla_\theta(\theta; x) + \sigma w$. Combining with the work above, we can write

$$w = \phi_x(\theta', g') \text{ where } \phi_x(\theta', g') = \frac{V_\mathcal{I} g' - \nabla_\theta(U_\mathcal{I}\theta' + V_\mathcal{I}\theta_*; x)}{\sigma},$$

and so

$$\theta = \hat{\theta}(x, w) = \hat{\theta}\left(x, \phi_x(\theta', g')\right), \; g = \hat{g}(x, w) = \hat{g}\left(x, \phi_x(\theta', g')\right).$$

Therefore,

$$\theta' = U_\mathcal{I}^\top \hat{\theta}\left(x, \phi_x(\theta', g')\right), \; g' = V_\mathcal{I}^\top \hat{g}\left(x, \phi_x(\theta', g')\right).$$

We can also calculate

$$\nabla_{\theta'}\phi_x(\theta', g') = -\sigma^{-1} U_\mathcal{I}^\top \nabla_\theta^2(U_\mathcal{I}\theta' + V_\mathcal{I}\theta_*; x)$$

and

$$\nabla_{g'}\phi_x(\theta', g') = \sigma^{-1} V_\mathcal{I}^\top.$$

Therefore,

$$\begin{aligned}
\det\left(\nabla\phi_x(\theta',g')\right) &= \det\left(\begin{pmatrix} \nabla_{\theta'}\phi_x(\theta',g') \\ \nabla_{g'}\phi_x(\theta',g') \end{pmatrix}\right) \\
&= \det\left(\begin{pmatrix} \nabla_{\theta'}\phi_x(\theta',g') \\ \nabla_{g'}\phi_x(\theta',g') \end{pmatrix} \cdot (U_{\mathcal{I}}\ V_{\mathcal{I}})\right) \\
&= \det\left(\begin{pmatrix} \nabla_{\theta'}\phi_x(\theta',g')U_{\mathcal{I}} & \nabla_{\theta'}\phi_x(\theta',g')V_{\mathcal{I}} \\ \nabla_{g'}\phi_x(\theta',g')U_{\mathcal{I}} & \nabla_{g'}\phi_x(\theta',g')V_{\mathcal{I}} \end{pmatrix}\right) \\
&= \det\left(\begin{pmatrix} -\frac{U_{\mathcal{I}}^\top \nabla_\theta^2(U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*;x)U_{\mathcal{I}}}{\sigma} & -\frac{U_{\mathcal{I}}^\top \nabla_\theta^2(U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*;x)V_{\mathcal{I}}}{\sigma} \\ \sigma^{-1}V_{\mathcal{I}}^\top U_{\mathcal{I}} & \sigma^{-1}V_{\mathcal{I}}^\top V_{\mathcal{I}} \end{pmatrix}\right) \\
&= \det\left(\begin{pmatrix} -\frac{U_{\mathcal{I}}^\top \nabla_\theta^2(U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*;x)U_{\mathcal{I}}}{\sigma} & -\frac{U_{\mathcal{I}}^\top \nabla_\theta^2(U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*;x)V_{\mathcal{I}}}{\sigma} \\ 0 & \sigma^{-1}\mathbf{I}_{d-k} \end{pmatrix}\right) \\
&= (-1)^k\sigma^{-d}\cdot\det\left(U_{\mathcal{I}}^\top \nabla_\theta^2(U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*;x)U_{\mathcal{I}}\right).
\end{aligned}$$

From this point on, following similar arguments as [Barber and Janson, 2022, Section B.4] to verify the validity of applying the change-of-variables formula for integration, we calculate

$$\begin{aligned}
&\int_{\mathbb{R}^d} e^{-\frac{d}{2\sigma^2}\|\hat{g}(x,w)-\nabla_\theta(\hat{\theta}(x,w);x)\|^2}\mathbb{1}_{(x,w)\in\psi_{\mathcal{I}}^{-1}(I_{\mathcal{I}})}\,\mathsf{d}w \\
&= \sigma^{-d}\int_{\Theta'\times\mathbb{R}^{d-k}} e^{-\frac{d}{2\sigma^2}\|V_{\mathcal{I}}g'-\nabla_\theta(U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*;x)\|^2}\cdot\det{}_{\mathcal{I},\theta',x}\cdot\mathbb{1}_{(x,\phi_x(\theta',g'))\in\psi_{\mathcal{I}}^{-1}(I_{\mathcal{I}})}\,\mathsf{d}g'\,\mathsf{d}\theta',
\end{aligned}$$

where we write $\det_{\mathcal{I},\theta',x} = \det\left(U_{\mathcal{I}}^\top \nabla_\theta^2(U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*;x)U_{\mathcal{I}}\right)$ (note that this determinant must be positive, by the SSOSP conditions). We can also verify from our definitions that

$\mathbb{1}_{(x,\phi_x(\theta',g'))\in\psi_{\mathcal{I}}^{-1}(I_{\mathcal{I}})} = \mathbb{1}_{(x,U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*,V_{\mathcal{I}}g')\in I_{\mathcal{I}}}$. With this calculation in place we then have

$$\begin{aligned}\mathbb{P}\Big\{(X,\hat{\theta}(X,W),\hat{g}(X,W))\in I_{\mathcal{I}} \mid (X,\hat{\theta}(X,W),\hat{g}(X,W))\in\Psi_{\mathrm{SSOSP},\mathcal{I}}\Big\}&\\ = \left(\int_{\mathcal{X}\times\mathbb{R}^d} h_{\theta_0}(x',w')\,\mathsf{d}\nu_{\mathcal{X}}(x')\,\mathsf{d}w'\right)^{-1}&\\ \cdot\,\sigma^{-d}\int_{\mathcal{X}} f(x;\theta_0)\int_{\Theta'\times\mathbb{R}^{d-k}} e^{-\frac{d}{2\sigma^2}\|V_{\mathcal{I}}g'-\nabla_\theta(U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*;x)\|^2}&\\ \cdot\,\det{}_{\mathcal{I},\theta',x}\cdot\mathbb{1}_{(x,U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*,V_{\mathcal{I}}g')\in I_{\mathcal{I}}}\,\mathsf{d}g'\,\mathsf{d}\theta'\,\mathsf{d}\nu_{\mathcal{X}}(x),&\end{aligned}$$

In particular, this verifies that

$$\frac{\sigma^{-d}f(x;\theta_0)\cdot e^{-\frac{d}{2\sigma^2}\|V_{\mathcal{I}}g'-\nabla_\theta(U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*;x)\|^2}\cdot\det{}_{\mathcal{I},\theta',x}\cdot\mathbb{1}_{(x,U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*,V_{\mathcal{I}}g')\in\Psi_{\mathrm{SSOSP},\mathcal{I}}}}{\int_{\mathcal{X}\times\mathbb{R}^d} h_{\theta_0}(x',w')\,\mathsf{d}\nu_{\mathcal{X}}(x')\,\mathsf{d}w'}$$

is the joint density of $(X,U_{\mathcal{I}}^\top\hat{\theta},V_{\mathcal{I}}^\top\hat{g}) = (X,U_{\mathcal{I}}^\top\hat{\theta}(X,W),V_{\mathcal{I}}^\top\hat{g}(X,W))$, conditional on the event $(X,\hat{\theta}(X,W),\hat{g}(X,W))\in\Psi_{\mathrm{SSOSP},\mathcal{I}}$. Therefore, the conditional density of

$$X \mid (U_{\mathcal{I}}^\top\hat{\theta},V_{\mathcal{I}}^\top\hat{g})$$

(again conditioning on this same event) can be written as

$$\propto f(x;\theta_0)\cdot e^{-\frac{d}{2\sigma^2}\|V_{\mathcal{I}}g'-\nabla_\theta(U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*;x)\|^2}\cdot\det{}_{\mathcal{I},\theta',x}\cdot\mathbb{1}_{(x,U_{\mathcal{I}}\theta'+V_{\mathcal{I}}\theta_*,V_{\mathcal{I}}g')\in\Psi_{\mathrm{SSOSP},\mathcal{I}}}.$$

Moreover, $U_{\mathcal{I}}^\top\hat{\theta}$ and $V_{\mathcal{I}}^\top\hat{g}$ uniquely determine $\hat{\theta}$ and $\hat{g}$ on the event that $\mathcal{I}$ is the active set, as described earlier, so we can equivalently condition on $(\hat{\theta},\hat{g})$ and can rewrite this density as

$$p_{\theta_0}(\cdot\mid\hat{\theta},\hat{g})\propto f(x;\theta_0)\cdot e^{-\frac{d}{2\sigma^2}\|\hat{g}-\nabla_\theta(\hat{\theta};x)\|^2}\cdot\det\left(U_{\mathcal{I}}^\top\nabla_\theta^2(\hat{\theta};x)U_{\mathcal{I}}\right)\cdot\mathbb{1}_{(x,\hat{\theta},\hat{g})\in\Psi_{\mathrm{SSOSP},\mathcal{I}}}. \quad \text{(E.3)}$$

Finally, by definition, $(x, \hat{\theta}, \hat{g}) \in \Psi_{\text{SSOSP},\mathcal{I}}$ if and only if $\hat{\theta} \in \Theta_{\mathcal{I}}$ and $x \in \mathcal{X}_{\hat{\theta},\hat{g}}$, so

$$\mathbb{1}_{(x,\hat{\theta},\hat{g}) \in \Psi_{\text{SSOSP},\mathcal{I}}} = \mathbb{1}_{x \in \mathcal{X}_{\hat{\theta},\hat{g}}}$$

for $\hat{\theta} \in \Theta_{\mathcal{I}}$. □

E.1.4 Proof of Theorem 6.5.3: error control for aCSS with an ℓ_1 penalty

At a high level, the strategies underlying the proofs of Theorems 6.4.3, 6.4.5, and 6.4.7 are fundamentally the same. In the constrained case, first Lemma 6.3.3 is applied to calculate the conditional density of X given $(\hat{\theta}, \hat{g})$ as the expression $p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g})$ given in the lemma. This then justifies the sampling distribution used for the copies $\tilde{X}^{(m)}$, i.e., $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$, and the distance to exchangeability is then bounded by bounding $d_{\text{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}))$.

In examining the ℓ_1-penalized case, the arguments are exactly identical. First, by applying Lemma 6.5.2 in place of Lemma 6.3.3, the reasoning of Section E.1.1 verifies that it suffices to bound $\mathbb{E}_{Q^*_{\theta_0}}\left[d_{\text{TV}}(p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g}), p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}))\right]$, where $Q^*_{\theta_0}$ is now defined as the distribution of $(\hat{\theta}(X, W), \hat{g}(X, W))$ conditioning on the event that $(X, W) \in \Omega^{\text{pen}}_{\text{SSOSP},S}$ where

$$\Omega^{\text{pen}}_{\text{SSOSP}} = \left\{(x, w) \in \mathcal{X} \times \mathbb{R}^d : \hat{\theta}(x, w) \text{ is a SSOSP of (6.16)}\right\},$$

i.e., we are conditioning on the event of finding a SSOSP for the ℓ_1-penalized (rather than constrained) optimization problem. The calculation of the bound on this expected total variation distance is then identical to the constrained case.

E.1.5 Proof of Lemma 6.5.2: conditional density for aCSS with an ℓ_1 penalty

Now we revisit the proof of Lemma 6.3.3 and revise it for the ℓ_1-penalized case. Define a subset of Θ with support S as

$$\Theta_S = \{\theta \in \Theta : S(\theta) = S\}.$$

Further define

$$\Omega^{\text{pen}}_{\text{SSOSP},S} = \left\{(x, w) \in \mathcal{X} \times \mathbb{R}^d : \hat{\theta}(x, w) \text{ is a SSOSP of (6.16), and } S(\hat{\theta}(x, w)) = S\right\}.$$

By a result analogous to Lemma E.1.1, we have a bijection between $\Omega^{\text{pen}}_{\text{SSOSP},S}$ and $\Psi^{\text{pen}}_{\text{SSOSP},S}$, where

$$\Psi^{\text{pen}}_{\text{SSOSP},S} = \Big\{(x, \theta, g) \in \mathcal{X} \times \Theta_S \times \mathbb{R}^d : \exists w \in \mathbb{R}^d \text{ such that}$$
$$\theta = \hat{\theta}(x, w) \text{ is a SSOSP of (6.16), and } g = \hat{g}(x, w)\Big\},$$

which is defined by the map $\psi_S : (x, w) \to \Big(x, \hat{\theta}(x, w), \hat{g}(x, w)\Big)$, with inverse $\psi_S^{-1} : (x, \theta, g) \to \Big(x, \frac{g - \nabla_\theta(\theta; x)}{\sigma}\Big)$.

Consider the joint distribution $(X, W) \sim P_{\theta_0} \times \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$. By assumption, the event $(X, W) \in \Omega^{\text{pen}}_{\text{SSOSP},S}$ has positive probability. Then the joint density of (X, W), conditioning on the event that $\hat{\theta}(X, W)$ is a SSOSP of (6.16) with support S, i.e., $(X, W) \in \Omega^{\text{pen}}_{\text{SSOSP},S}$, is proportional to the function

$$h_{\theta_0}(x, w) = f(x; \theta_0) \exp\left\{-\frac{d}{2}\|w\|^2\right\} \mathbb{1}_{(x,w) \in \Omega^{\text{pen}}_{\text{SSOSP},S}}.$$

For any measurable set $I_S \subseteq \Psi^{\text{pen}}_{\text{SSOSP},S}$, define

$$\psi_S^{-1}(I_S) = \{(x,w) \in \Omega^{\text{pen}}_{\text{SSOSP},S} : \psi_S(x,w) \in I_S)\}.$$

Then, following the same calculation for

$$\mathbb{P}\left\{(X,\hat{\theta}(X,W),\hat{g}(X,W)) \in I_{\mathcal{I}} \mid (X,\hat{\theta}(X,W),\hat{g}(X,W)) \in \Psi_{\text{SSOSP},\mathcal{I}}\right\}$$

as in the proof of Lemma 6.3.3 (with $\Omega_{\text{SSOSP},\mathcal{I}}$ replaced by $\Omega^{\text{pen}}_{\text{SSOSP},S}$), we have

$$\mathbb{P}\left\{(X,\hat{\theta}(X,W),\hat{g}(X,W)) \in I_S \mid (X,\hat{\theta}(X,W),\hat{g}(X,W)) \in \Psi^{\text{pen}}_{\text{SSOSP},S}\right\} \\ = \frac{\int_{\mathcal{X}} f(x;\theta_0) \int_{\mathbb{R}^d} e^{-\frac{d}{2\sigma^2}\|\hat{g}(x,w)-\nabla_\theta(\hat{\theta}(x,w);x)\|^2} \mathbb{1}_{(x,w)\in\psi_S^{-1}(I_S)} \,\mathsf{d}w\,\mathsf{d}\nu_{\mathcal{X}}(x)}{\int_{\mathcal{X}\times\mathbb{R}^d} h_{\theta_0}(x',w')\,\mathsf{d}\nu_{\mathcal{X}}(x')\,\mathsf{d}w'}.$$

Next we need to reparametrize $(\hat{\theta},\hat{g})$, since, as in the constrained case, these parameters, which each have dimension d, actually contain only d degrees of freedom in total (i.e., since there is a bijection between (x,w) and $(x,\hat{\theta},\hat{g})$, and $w \in \mathbb{R}^d$). In fact, in the ℓ_1-penalized setting, this is simple: once we condition on the event that $S(\hat{\theta}) = S$, this implies that $\hat{\theta}_{S^\complement} = \mathbf{0}_{d-|S|}$, and that $\hat{g}_S = \lambda\text{sign}(\hat{\theta}_S)$. In other words, $(\hat{\theta}_S,\hat{g}_{S^\complement})$ captures the full information contained in $(\hat{\theta},\hat{g})$—which agrees with our calculation of degrees of freedom since $|S| + |S^\complement| = d$. For convenience, we now define $\mathbf{I}_S$ as the d-by-$|S|$ matrix obtained by taking the d-by-d identity and extracting columns corresponding to S, and $\mathbf{I}_{S^\complement}$ similarly for $S^\complement$. Then, for $(x,\theta,g) \in \Psi_{\text{SSOSP},S}$, we have calculated

$$\theta = \mathbf{I}_S\theta_S,\ \ g = \mathbf{I}_S \cdot \lambda\text{sign}(\theta_S) + \mathbf{I}_{S^\complement} \cdot g_{S^\complement}.$$

Next, if $(x,\theta,g) \in \Psi_{\text{SSOSP},S}$ then by the SSOSP conditions we must have some w such that $\theta = \hat{\theta}(x,w)$ is a SSOSP of (6.16), and $g = \hat{g}(x,w) = \nabla_\theta(\theta;x,w) = \nabla_\theta(\theta;x) + \sigma w$.

Combining with the work above, we can write

$$w = \phi_x(\theta_S, g_{S^\complement}) \text{ where } \phi_x(\theta_S, g_{S^\complement}) = \frac{\mathbf{I}_S \cdot \lambda \text{sign}(\theta_S) + \mathbf{I}_{S^\complement} \cdot g_{S^\complement} - \nabla_\theta(\mathbf{I}_S\theta_S; x)}{\sigma},$$

and so

$$\theta = \hat{\theta}(x, w) = \hat{\theta}\left(x, \phi_x(\theta_S, g_{S^\complement})\right),\ g = \hat{g}(x, w) = \hat{g}\left(x, \phi_x(\theta_S, g_{S^\complement})\right).$$

Therefore,

$$\theta_S = \mathbf{I}_S^\top \hat{\theta}\left(x, \phi_x(\theta_S, g_{S^\complement})\right),\ g_{S^\complement} = \mathbf{I}_{S^\complement}^\top \hat{g}\left(x, \phi_x(\theta_S, g_{S^\complement})\right).$$

We can also calculate

$$\nabla_{\theta_S}\phi_x(\theta_S, g_{S^\complement}) = -\sigma^{-1}\mathbf{I}_S^\top \nabla_\theta^2(\mathbf{I}_S\theta_S; x)$$

and

$$\nabla_{g_{S^\complement}}\phi_x(\theta_S, g_{S^\complement}) = \sigma^{-1}\mathbf{I}_{S^\complement}^\top.$$

Therefore,

$$\begin{aligned}
\det\Big(\nabla\phi_x(\theta_S, g_{S^\complement})\Big) &= \det\left(\begin{pmatrix} \nabla_{\theta_S}\phi_x(\theta_S, g_{S^\complement}) \\ \nabla_{g_{S^\complement}}\phi_x(\theta_S, g_{S^\complement}) \end{pmatrix}\right) \\
&= \det\left(\begin{pmatrix} \nabla_{\theta_S}\phi_x(\theta_S, g_{S^\complement}) \\ \nabla_{g_{S^\complement}}\phi_x(\theta_S, g_{S^\complement}) \end{pmatrix} \cdot (\mathbf{I}_S \ \mathbf{I}_{S^\complement})\right) \\
&= \det\left(\begin{pmatrix} \nabla_{\theta_S}\phi_x(\theta_S, g_{S^\complement})\mathbf{I}_S & \nabla_{\theta_S}\phi_x(\theta_S, g_{S^\complement})\mathbf{I}_{S^\complement} \\ \nabla_{g_{S^\complement}}\phi_x(\theta_S, g_{S^\complement})\mathbf{I}_S & \nabla_{g_{S^\complement}}\phi_x(\theta_S, g_{S^\complement})\mathbf{I}_{S^\complement} \end{pmatrix}\right) \\
&= \det\left(\begin{pmatrix} -\sigma^{-1}\mathbf{I}_S^\top\nabla^2_\theta(\mathbf{I}_S\theta_S; x)\mathbf{I}_S & -\sigma^{-1}\mathbf{I}_S^\top\nabla^2_\theta(\mathbf{I}_S\theta_S; x)\mathbf{I}_{S^\complement} \\ \sigma^{-1}\mathbf{I}_{S^\complement}^\top\mathbf{I}_S & \sigma^{-1}\mathbf{I}_{S^\complement}^\top\mathbf{I}_{S^\complement} \end{pmatrix}\right) \\
&= \det\left(\begin{pmatrix} -\sigma^{-1}\mathbf{I}_S^\top\nabla^2_\theta(\mathbf{I}_S\theta_S; x)\mathbf{I}_S & -\sigma^{-1}\mathbf{I}_S^\top\nabla^2_\theta(\mathbf{I}_S\theta_S; x)\mathbf{I}_{S^\complement} \\ 0 & \sigma^{-1}\mathbf{I}_{d-|S|} \end{pmatrix}\right) \\
&= (-1)^{|S|}\sigma^{-d} \cdot \det\Big(\mathbf{I}_S^\top\nabla^2_\theta(\mathbf{I}_S\theta_S; x)\mathbf{I}_S\Big) \\
&= (-1)^{|S|}\sigma^{-d} \cdot \det\Big(\nabla^2_\theta(\mathbf{I}_S\theta_S; x)_S\Big).
\end{aligned}$$

From this point on, following similar arguments as [Barber and Janson, 2022, Section B.4] to verify the validity of applying the change-of-variables formula for integration, we calculate

$$\begin{aligned}
&\int_{\mathbb{R}^d} e^{-\frac{d}{2\sigma^2}\|\hat{g}(x,w)-\nabla_\theta(\hat{\theta}(x,w);x)\|^2}\mathbb{1}_{(x,w)\in\psi_S^{-1}(I_S)}\ \mathsf{d}w \\
&\qquad = \sigma^{-d}\int_{\theta_S\times\mathbb{R}^{d-k}} e^{-\frac{d}{2\sigma^2}\|\mathbf{I}_S\cdot\lambda\mathrm{sign}(\theta_S)+\mathbf{I}_{S^\complement}g_{S^\complement}-\nabla_\theta(\mathbf{I}_S\theta_S;x)\|^2} \\
&\qquad\qquad \cdot \det\Big(\nabla^2_\theta(\mathbf{I}_S\theta_S; x)_S\Big)\cdot\mathbb{1}_{(x,\phi_x(\theta_S,g_{S^\complement}))\in\psi_S^{-1}(I_S)}\ \mathsf{d}g_{S^\complement}\ \mathsf{d}\theta_S,
\end{aligned}$$

where we note that $\det\left(\nabla^2_\theta(\mathbf{I}_S\theta_S; x)_S\right)$ must be positive, by the SSOSP conditions. We can also verify from our definitions that $\mathbb{1}_{(x,\phi_x(\theta_S,g_{S^\complement}))\in\psi_S^{-1}(I_S)} = \mathbb{1}_{(x,\mathbf{I}_S\theta_S,\mathbf{I}_S\cdot\lambda\mathrm{sign}(\theta_S)+\mathbf{I}_{S^\complement}g_{S^\complement})\in I_S}$.

With this calculation in place we then have

$$\begin{aligned}\mathbb{P}\Big\{(X,\hat{\theta}(X,W),\hat{g}(X,W))\in I_S \mid (X,\hat{\theta}(X,W),\hat{g}(X,W))\in\Psi_{\mathrm{SSOSP},S}\Big\}\\ =\left(\sigma^d\int_{\mathcal{X}\times\mathbb{R}^d} h_{\theta_0}(x',w')\,\mathsf{d}\nu_{\mathcal{X}}(x')\,\mathsf{d}w'\right)^{-1}\\ \cdot\int_{\mathcal{X}} f(x;\theta_0)\int_{\theta_S\times\mathbb{R}^{d-|S|}} e^{-\frac{d}{2\sigma^2}\|\mathbf{I}_S\cdot\lambda\mathrm{sign}(\theta_S)+\mathbf{I}_{S^{\complement}}g_{S^{\complement}}-\nabla_\theta(\mathbf{I}_S\theta_S;x)\|^2}\\ \cdot\det\Big(\nabla^2_\theta(\mathbf{I}_S\theta_S;x)_S\Big)\cdot\mathbb{1}_{(x,\mathbf{I}_S\theta_S,\mathbf{I}_S\cdot\lambda\mathrm{sign}(\theta_S)+\mathbf{I}_{S^{\complement}}g_{S^{\complement}})\in I_S}\;\mathsf{d}g_{S^{\complement}}\;\mathsf{d}\theta_S\;\mathsf{d}\nu_{\mathcal{X}}(x).\end{aligned}$$

In particular, this verifies that

$$\frac{f(x;\theta_0)\cdot e^{-\frac{d}{2\sigma^2}\|\mathbf{I}_S\cdot\lambda\mathrm{sign}(\theta_S)+\mathbf{I}_{S^{\complement}}g_{S^{\complement}}-\nabla_\theta(\mathbf{I}_S\theta_S;x)\|^2}\cdot\det\left(\nabla^2_\theta(\mathbf{I}_S\theta_S;x)_S\right)\cdot\mathbb{1}_{(x,\mathbf{I}_S\theta_S,\mathbf{I}_S\cdot\lambda\mathrm{sign}(\theta_S)+\mathbf{I}_{S^{\complement}}g_{S^{\complement}})\in\Psi_{\mathrm{SSOSP},S}}}{\sigma^d\int_{\mathcal{X}\times\mathbb{R}^d}h_{\theta_0}(x',w')\,\mathsf{d}\nu_{\mathcal{X}}(x')\,\mathsf{d}w'}$$

is the joint density of $(X,\hat{\theta}_S,\hat{g}_{S^{\complement}})=(X,\hat{\theta}(X,W)_S,\hat{g}(X,W)_{S^{\complement}})$, conditional on the event $(X,\hat{\theta}(X,W),\hat{g}(X,W))\in\Psi_{\mathrm{SSOSP},S}$. Therefore, the conditional density of $X\mid(\hat{\theta}_S,\hat{g}_{S^{\complement}})$ (again conditioning on this same event) can be written as

$$\begin{aligned}\propto f(x;\theta_0)\cdot e^{-\frac{d}{2\sigma^2}\|\mathbf{I}_S\cdot\lambda\mathrm{sign}(\theta_S)+\mathbf{I}_{S^{\complement}}g_{S^{\complement}}-\nabla_\theta(\mathbf{I}_S\theta_S;x)\|^2}\cdot\det\Big(\nabla^2_\theta(\mathbf{I}_S\theta_S;x)_S\Big)\\ \cdot\mathbb{1}_{(x,\mathbf{I}_S\theta_S,\mathbf{I}_S\cdot\lambda\mathrm{sign}(\theta_S)+\mathbf{I}_{S^{\complement}}g_{S^{\complement}})\in\Psi_{\mathrm{SSOSP},S}}.\end{aligned}$$

Moreover, $\hat{\theta}_S$ and $\hat{g}_{S^{\complement}}$ uniquely determine $\hat{\theta}$ and $\hat{g}$ on the event that S is the support, as described earlier, so we can equivalently condition on $(\hat{\theta},\hat{g})$ and can rewrite this density as

$$p_{\theta_0}(\cdot\mid\hat{\theta},\hat{g})\propto f(x;\theta_0)\cdot e^{-\frac{d}{2\sigma^2}\|\hat{g}-\nabla_\theta(\hat{\theta};x)\|^2}\cdot\det\Big(\nabla^2_\theta(\hat{\theta};x)_S\Big)\cdot\mathbb{1}_{(x,\hat{\theta},\hat{g})\in\Psi_{\mathrm{SSOSP},S}}.\tag{E.4}$$

Finally, by definition, $(x,\hat{\theta},\hat{g})\in\Psi_{\mathrm{SSOSP},S}$ if and only if $\hat{\theta}\in\Theta_S$ and $x\in\mathcal{X}_{\hat{\theta},\hat{g}}$, so $\mathbb{1}_{(x,\hat{\theta},\hat{g})\in\Psi_{\mathrm{SSOSP},S}}=\mathbb{1}_{x\in\mathcal{X}_{\hat{\theta},\hat{g}}}$ for $\hat{\theta}\in\Theta_S$.

E.2 Additional proofs

E.2.1 Verifying that the plug-in version of $p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g})$ defines a density

To ensure that our procedure is well-defined in both constrained and ℓ_1-penalized cases, we need to verify that the plug-in version of the conditional density

$$p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g}) \propto p^{\mathrm{un}}_{\hat{\theta},\hat{g}}(x)$$

defines a valid density with respect to $\nu_{\mathcal{X}}$, where $p^{\mathrm{un}}_{\theta,g}(x)$ represents the unnormalized density, namely,

$$p^{\mathrm{un}}_{\theta,g}(x) = f(x;\theta) \cdot e^{-\frac{d}{2\sigma^2}\|g-\nabla_\theta(\theta;x)\|^2} \cdot \det\left(U^\top_{\mathcal{I}(\theta)} \nabla^2_\theta(\theta;x) U_{\mathcal{I}(\theta)}\right) \cdot \mathbb{1}_{(x,\theta,g)\in\Psi_{\mathrm{SSOSP},\mathcal{I}(\theta)}}$$

in the constrained case as in (6.8); and

$$p^{\mathrm{un}}_{\theta,g}(x) = f(x;\theta) \cdot e^{-\frac{d}{2\sigma^2}\|g-\nabla_\theta(\theta;x)\|^2} \cdot \det\left(\nabla^2_\theta(\theta;x)_{S(\theta)}\right) \cdot \mathbb{1}_{(x,\theta,g)\in\Psi^{\mathrm{pen}}_{\mathrm{SSOSP},S(\theta)}}$$

in the ℓ_1-penalized case as in (6.18). To verify this we only need to check that this unnormalized density integrates to a finite and positive value (the analogous result for aCSS appears in [Barber and Janson, 2022, Section B.3]).

Lemma E.2.1. *If Assumption 6.3.2 and 6.4.2 hold, then for $\theta \in \Theta$ and $g \in \mathbb{R}^d$, the unnormalized density $p^{\mathrm{un}}_{\theta,g}(x)$ is nonnegative and integrable with respect to $\nu_{\mathcal{X}}$. Furthermore, if the event $\hat{\theta} = \hat{\theta}(X, W)$ is a SSOSP has positive probability, then conditional on this event, $\int_{\mathcal{X}} p^{\mathrm{un}}_{\hat{\theta},\hat{g}}(x)\mathsf{d}\nu_{\mathcal{X}}(x) > 0$ holds almost surely.*

Proof. **Constrained case:** We first check nonnegativity. For any $\theta \in \Theta$ and any x, we have $f(x;\theta) > 0$ by Assumption 6.3.2. Furthermore, if $x \in \mathcal{X}_{\theta,g}$ then $\det\left(U^\top_{\mathcal{I}(\theta)} \nabla^2_\theta(\theta;x) U_{\mathcal{I}(\theta)}\right) > 0$ by definition of $\mathcal{X}_{\theta,g}$ and the SSOSP conditions. This verifies the nonnegativity for $p^{\mathrm{un}}_{\theta,g}(x)$

for any (θ, g, x). Next we check integrability.

$$\begin{aligned}
&\int_{\mathcal{X}} p^{\mathrm{un}}_{\theta,g}(x)\mathsf{d}\nu_{\mathcal{X}}(x) \le \int_{\mathcal{X}} f(x;\theta)\cdot\det\left(U_{\mathcal{I}(\theta)}^{\top}\nabla_{\theta}^{2}(\theta;x)U_{\mathcal{I}(\theta)}\right)\cdot\mathbb{1}_{U_{\mathcal{I}(\theta)}^{\top}\nabla_{\theta}^{2}(\theta;x)U_{\mathcal{I}(\theta)}\succ 0}\mathsf{d}\nu_{\mathcal{X}}(x)\\
&\le \int_{\mathcal{X}} f(x;\theta)\cdot\left(\lambda_{\max}\left(U_{\mathcal{I}(\theta)}^{\top}\nabla_{\theta}^{2}(\theta;x)U_{\mathcal{I}(\theta)}\right)\right)^{d}\cdot\mathbb{1}_{U_{\mathcal{I}(\theta)}^{\top}\nabla_{\theta}^{2}(\theta;x)U_{\mathcal{I}(\theta)}\succ 0}\mathsf{d}\nu_{\mathcal{X}}(x)\\
&\le \int_{\mathcal{X}} f(x;\theta)\cdot\left(\lambda_{\max}\left(\nabla_{\theta}^{2}(\theta;x)\right)\right)_{+}^{d}\mathsf{d}\nu_{\mathcal{X}}(x)\\
&\le \frac{d!}{r(\theta)^{2d}}\int_{\mathcal{X}} f(x;\theta)\\
&\quad\cdot\exp\left\{r(\theta)^{2}(\lambda_{\max}\left(H(\theta,x)-H(\theta)\right))_{+}+r(\theta)^{2}(\lambda_{\max}\left(H(\theta)-\nabla_{\theta}^{2}\mathcal{R}(\theta)\right))_{+}\right\}\mathsf{d}\nu_{\mathcal{X}}(x)\\
&= \frac{d!}{r(\theta)^{2d}}\exp\left\{r(\theta)^{2}(\lambda_{\max}\left(H(\theta)-\nabla_{\theta}^{2}\mathcal{R}(\theta)\right))_{+}\right\}\\
&\qquad\cdot\mathbb{E}_{P_{\theta}}\left[\exp\left\{r(\theta)^{2}(\lambda_{\max}\left(H(\theta,x)-H(\theta)\right))_{+}\right\}\right]\\
&\le \frac{d!}{r(\theta)^{2d}}e^{\epsilon(\theta)}\exp\left\{r(\theta)^{2}(\lambda_{\max}\left(H(\theta)-\nabla_{\theta}^{2}\mathcal{R}(\theta)\right))_{+}\right\},
\end{aligned}$$

where the third-to-last step holds since $t^d \le d!e^d$ for any $t \ge 0$, and the last step holds by applying Assumption 6.4.2. This verifies that $\int_{\mathcal{X}} p^{\mathrm{un}}_{\theta,g}(x)\mathsf{d}\nu_{\mathcal{X}}(x)$ is finite. Finally, we check $\int_{\mathcal{X}} p^{\mathrm{un}}_{\hat{\theta},\hat{g}}(x)\mathsf{d}\nu_{\mathcal{X}}(x) > 0$ holds almost surely. For any x, we have $\frac{f(x,\theta_0)}{f(x,\hat{\theta})} > 0$ by Assumption 6.3.2. Combined with the fact that $p^{\mathrm{un}}_{\hat{\theta},\hat{g}}(x)$ is nonnegative as proved above, it is therefore equivalent to verify that $\int_{\mathcal{X}} \frac{f(x,\theta_0)}{f(x,\hat{\theta})}p^{\mathrm{un}}_{\hat{\theta},\hat{g}}(x)\mathsf{d}\nu_{\mathcal{X}}(x) > 0$. This last claim must hold since $p_{\theta_0}(x \mid \hat{\theta}, \hat{g}) \propto \frac{f(x,\theta_0)}{f(x,\hat{\theta})}p^{\mathrm{un}}_{\hat{\theta},\hat{g}}(x)$ is the conditional density of $X \mid \hat{\theta}, \hat{g}$.

ℓ_1-penalized case: The proof for this case mirrors that for the constrained case. For any $\theta \in \Theta$ and x, we have $f(x;\theta) > 0$ by Assumption 6.3.2. Furthermore, if $(x, \theta, g) \in \Psi^{\mathrm{pen}}_{\mathrm{SSOSP},S(\theta)}$ then $\det\left(\nabla^2_\theta(\theta;x)_{S(\theta)}\right) > 0$ by definition of $\Psi^{\mathrm{pen}}_{\mathrm{SSOSP},S(\theta)}$ and the SSOSP conditions. This verifies the nonnegativity of $p^{\text{un}}_{\theta,g}(x)$ for any (θ, g, x). To check integrability, we have

$$\int_{\mathcal{X}} p^{\text{un}}_{\theta,g}(x) \mathsf{d}\nu_{\mathcal{X}}(x) \le \int_{\mathcal{X}} f(x;\theta) \cdot \det\left(\nabla^2_\theta(\theta;x)_{S(\theta)}\right) \cdot \mathbb{1}_{\nabla^2_\theta(\theta;x)_{S(\theta)} \succ 0} \mathsf{d}\nu_{\mathcal{X}}(x)$$
$$\le \int_{\mathcal{X}} f(x;\theta) \cdot \left(\lambda_{\max}\left(\nabla^2_\theta(\theta;x)_{S(\theta)}\right)\right)^d \cdot \mathbb{1}_{\nabla^2_\theta(\theta;x)_{S(\theta)} \succ 0} \mathsf{d}\nu_{\mathcal{X}}(x)$$
$$\le \int_{\mathcal{X}} f(x;\theta) \cdot \left(\lambda_{\max}\left(\nabla^2_\theta(\theta;x)\right)\right)^d_+ \mathsf{d}\nu_{\mathcal{X}}(x)$$
$$\le \frac{d!}{r(\theta)^{2d}} \int_{\mathcal{X}} f(x;\theta) \exp\left\{r(\theta)^2(\lambda_{\max}(H(\theta,x) - H(\theta)))_+ + r(\theta)^2(\lambda_{\max}\left(H(\theta) - \nabla^2_\theta \mathcal{R}(\theta)\right))_+\right\}$$
$$\le \frac{d!}{r(\theta)^{2d}} e^{\epsilon(\theta)} \exp\left\{r(\theta)^2(\lambda_{\max}\left(H(\theta) - \nabla^2_\theta \mathcal{R}(\theta)\right))_+\right\}.$$

Finally, $\int_{\mathcal{X}} p^{\text{un}}_{\hat\theta,\hat g}(x) \mathsf{d}\nu_{\mathcal{X}}(x) > 0$ holds almost surely for the same reason as in the constrained case. □

E.2.2 Proof of Lemma E.1.1

Proof. First we check that $\psi_{\mathcal{I}}$ is injective on $\Omega_{\text{SSOSP},\mathcal{I}}$. For any $(x_1, w_2), (x_2, w_2) \in \Omega_{\text{SSOSP},\mathcal{I}}$, if $\psi_{\mathcal{I}}(x_1, w_1) = \psi_{\mathcal{I}}(x_2, w_2) = (x, \theta, g)$, then by definition of $\psi_{\mathcal{I}}$, we have $x_1 = x_2 = x$ trivially. By definition of $\psi_{\mathcal{I}}$ and $\hat g$,

$$\nabla_\theta(\theta; x) + \sigma w_1 = \hat g(x_1, w_1) = g = \hat g(x_2, w_2) = \nabla_\theta(\theta; x) + \sigma w_2,$$

therefore $w_1 = w_2 = \frac{g - \nabla_\theta(\theta;x)}{\sigma}$. This establishes that $\Psi_{\mathcal{I}}$ is injective and that the inverse function (on the image of $\psi_{\mathcal{I}}$) is given as claimed above.

Then we verify that $\Psi_{\text{SSOSP},\mathcal{I}}$ is the image of $\psi_{\mathcal{I}}$. Suppose $(x, \theta, g) \in \psi_{\mathcal{I}}(\Omega_{\text{SSOSP},\mathcal{I}})$, i.e, for some w such that $(x, w) \in \Omega_{\text{SSOSP},\mathcal{I}}$, we have $\theta = \hat\theta(x, w)$, which is a SSOSP with active set $\mathcal{I}$, and $g = \nabla_\theta(\hat\theta(x, w); x, w) = \hat g(x, w)$. Then for this w, $\theta = \hat\theta(x, w) \in \Theta_{\mathcal{I}}$, and $g = \hat g(x, w)$. Therefore, $(x, \theta, g) \in \Psi_{\text{SSOSP},\mathcal{I}}$, and so we have shown that $\psi_{\mathcal{I}}(\Omega_{\text{SSOSP},\mathcal{I}}) \subseteq \Psi_{\text{SSOSP},\mathcal{I}}$.

Conversely suppose that $(x, \theta, g) \in \Psi_{\text{SSOSP},\mathcal{I}}$. By definition of $\Psi_{\text{SSOSP},\mathcal{I}}$, there exists w such that $\theta = \hat{\theta}(x, w)$ is a SSOSP of (6.3) with active set $\mathcal{I}$, and $g = \hat{g}(x, w)$. Therefore, for this w we have $(x, w) \in \Omega_{\text{SSOSP},\mathcal{I}}$. Then $(x, \theta, g) = (x, \hat{\theta}(x, w), \hat{g}(x, w)) = \psi_{\mathcal{I}}(x, w) \in \psi_{\mathcal{I}}(\Omega_{\text{SSOSP},\mathcal{I}})$. This verifies that $\Psi_{\text{SSOSP},\mathcal{I}} \subseteq \psi_{\mathcal{I}}(\Omega_{\text{SSOSP},\mathcal{I}})$, and thus completes the proof.

□

E.2.3 Proof of Lemma 6.4.4

Proof. Fix any $\lambda \in (0, 1/2)$. We calculate

$$\begin{aligned}
e^{\lambda h_v(k)} &= \exp\left\{\lambda \mathbb{E}_{Z\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\max_{S\subseteq[p],|S|\le k} \|\mathcal{P}_{v_S}(Z)\|^2\right]\right\} \\
&\le \mathbb{E}_{Z\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\exp\left\{\lambda \max_{S\subseteq[p],|S|\le k} \|\mathcal{P}_{v_S}(Z)\|^2\right\}\right] \quad \text{by Jensen's inequality} \\
&= \mathbb{E}_{Z\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\max_{S\subseteq[p],|S|\le k} \exp\left\{\lambda\|\mathcal{P}_{v_S}(Z)\|^2\right\}\right] \\
&\le \mathbb{E}_{Z\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\sum_{S\subseteq[p],|S|=k} \exp\left\{\lambda\|\mathcal{P}_{v_S}(Z)\|^2\right\}\right] \\
&= \sum_{S\subseteq[p],|S|=k} \mathbb{E}_{Z\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\exp\left\{\lambda\|\mathcal{P}_{v_S}(Z)\|^2\right\}\right].
\end{aligned}$$

Since $\|\mathcal{P}_{v_S}(Z)\|^2 \sim \chi^2_{\dim(\text{span}(\{v_i\}_{i\in S}))}$, we have

$$\begin{aligned}
e^{\lambda h_v(k)} &\le \sum_{S\subseteq[p],|S|=k} (1-2\lambda)^{-\frac{1}{2}\dim(\text{span}(\{v_i\}_{i\in S}))} \\
&\le \sum_{S\subseteq[p],|S|=k} (1-2\lambda)^{-k/2} = \binom{p}{k}(1-2\lambda)^{-k/2} \le \left(\frac{ep}{k}\right)^k (1-2\lambda)^{-k/2}.
\end{aligned}$$

Therefore,

$$h_v(k) \le \inf_{\lambda\in(0,1/2)} \left\{\lambda^{-1}\log\left[\left(\frac{ep}{k}\right)^k (1-2\lambda)^{-k/2}\right]\right\} = \frac{k}{2}\inf_{\lambda\in(0,1/2)}\left\{\frac{2\log(ep/k) - \log(1-2\lambda)}{\lambda}\right\}.$$

Taking $\lambda = 1/4$,

$$h_v(k) \leq 2k\left(2\log(ep/k) - \log(1/2)\right) \leq 4k\log(4p/k).$$

Finally, we have $\max_{S\subseteq[p],|S|\leq k} \|\mathcal{P}_{v_S}(Z)\|^2 \leq \|Z\|^2$, and therefore,

$$h_v(k) = \mathbb{E}_{Z\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\max_{S\subseteq[p],|S|\leq k} \|\mathcal{P}_{v_S}(Z)\|^2\right] \leq \mathbb{E}_{Z\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\|Z\|^2\right] = d,$$

since $\|Z\|^2 \sim \chi^2_d$.

□

E.3 Checking assumptions for examples

In this section, we verify that Assumptions 6.3.2, 6.4.1, and 6.4.2 hold for the three examples considered in Section 6.6: the Gaussian mixture model (Example 1), isotonic Gaussian linear regression (Example 2), and sparse high-dimensional Gaussian linear regression (Example 3).

E.3.1 Verifying assumptions for Examples 2 (isotonic regression) and 3 (sparse regression)

We first verify the assumptions for the two examples in the Gaussian linear model setting, since these are more straightforwards. First, Assumption 6.3.2 holds trivially by construction—we have $\Theta = \mathbb{R}^d$, and twice-differentiability of $(\theta; x)$ holds both with and without the ridge penalty.

Next we check Assumption 6.4.1. In both examples, the optimization problem that defines $\hat{\theta}(X, W)$ is strongly convex, meaning that we can define $\hat{\theta}(X, W)$ as the unique minimizer, and the SSOSP conditions then hold surely. Next we need to verify a high probability bound on $\|\hat{\theta}(X, W) - \theta_0\|$. First, for isotonic regression, we see that $\hat{\theta}(X, W)$ can equivalently be written as

$$\hat{\theta}(X,W) = \arg\min_{\theta\in\mathbb{R}^d} \left\{ \frac{1}{2}\|\theta - (X - \sigma W)\|_2^2 : \theta_1 \leq \cdots \leq \theta_n \right\},$$

i.e., the isotonic projection of $X - \sigma W$. Since $X - \sigma W \sim \mathcal{N}(\theta_0, (\nu^2 + \sigma^2/n)\mathbf{I}_n)$, applying the result of [Yang and Barber, 2019, Theorem 5 and Appendix A.1] we have a high-probability bound on the error,

$$\|\hat{\theta}(X,W) - \theta_0\| \leq O\left(n^{1/6}(\log n)^{1/3}(1+\sigma^2)^{2/3}\right) \text{ with probability} \geq 1 - 1/n.$$

If we choose $\sigma = O(1)$, we can therefore take $r(\theta_0) = O\left(n^{1/6}(\log n)^{1/3}\right)$ and $\delta(\theta_0) = 1/n$.

Next, for sparse regression, the calculation is a bit more complex. Our argument closely follows the framework developed in [Negahban et al., 2012, Theorem 1]. Let $\Delta = \hat{\theta}(X,W) - \theta_0$. Then by optimality of $\hat{\theta}(X,W)$ we have

$$\begin{aligned}\frac{1}{2\nu^2}\|X - Z(\theta_0 + \Delta)\|_2^2 + \sigma(\theta_0+\Delta)^\top W + \frac{\lambda_{\text{ridge}}}{2}\|\theta_0+\Delta\|_2^2 + \lambda\|\theta_0+\Delta\|_1 \\ \leq \frac{1}{2\nu^2}\|X - Z\theta_0\|_2^2 + \sigma\theta_0^\top W + \frac{\lambda_{\text{ridge}}}{2}\|\theta_0\|_2^2 + \lambda\|\theta_0\|_1.\end{aligned}$$

Rearranging terms, and writing $v = X - Z\theta_0 \sim \mathcal{N}(0, \nu^2\mathbf{I}_n)$,

$$\begin{aligned}\frac{1}{2}\Delta^\top\left(\frac{Z^\top Z}{\nu^2} + \lambda_{\text{ridge}}\mathbf{I}_d\right)\Delta - \Delta^\top\left(\frac{Z^\top v}{\nu^2} - \sigma W - \lambda_{\text{ridge}}\theta_0\right) &\leq \lambda\left(\|\theta_0\|_1 - \|\theta_0+\Delta\|_1\right) \\ &\leq \lambda\|\Delta_{S(\theta_0)}\|_1 - \lambda\|\Delta_{S(\theta_0)^\complement}\|_1.\end{aligned}$$

Then, if the penalty parameter satisfies $\lambda \geq 2\left\|\frac{Z^\top v}{\nu^2} - \sigma W - \lambda_{\text{ridge}}\theta_0\right\|_\infty$, it holds that

$$\frac{1}{2}\Delta^\top\left(\frac{Z^\top Z}{\nu^2} + \lambda_{\text{ridge}}\mathbf{I}_d\right)\Delta \leq 1.5\lambda\|\Delta_{S(\theta_0)}\|_1 - 0.5\lambda\|\Delta_{S(\theta_0)^\complement}\|_1.$$

Standard assumptions on Z (namely, a restricted eigenvalue type property [Negahban et al., 2012]) will then ensure

$$\|\Delta\| \leq O\left(\sqrt{\frac{|S(\theta_0)| \log d}{n}}\right)$$

with probability $\geq 1 - 1/n$, when we take $\nu = O(1)$, $\|\theta_0\|_\infty = O(1)$, $\lambda_{\text{ridge}} \lesssim \sqrt{n \log d}$, and $\sigma \lesssim \sqrt{nd}$. Therefore, we can take $r(\theta_0) = O\left(\sqrt{\frac{|S(\theta_0)| \log d}{n}}\right)$ and $\delta(\theta_0) = 1/n$.

Finally, we check Assumption 6.4.2. For isotonic regression, we have $H(\theta; x) = \nu^{-2}\mathbf{I}_d$, and for sparse regression, $H(\theta; x) = \nu^{-2}Z^\top Z + \lambda_{\text{ridge}}\mathbf{I}_d$. In both cases, $H(\theta; x)$ does not depend on x, and therefore, Assumption 6.4.2 holds trivially with $\epsilon(\theta_0) = 0$.

E.3.2 Verifying assumptions for Example 1 (Gaussian mixture model)

In this section, we verify that the assumptions of Theorem 6.4.3 hold for the Gaussian mixture model setting, specifically in the case of $J = 2$ components as implemented in our simulation. Assumption 6.3.2 holds trivially by construction. For Assumption 6.4.1, the accuracy of $\hat{\theta}(X, W)$ can be established with $r(\theta_0) \asymp \sqrt{\frac{\log n}{n}}$ and $\delta(\theta_0) \asymp n^{-1}$ via known results in the literature. For instance, [Hardt and Price, 2015, Corollary 1.4] show this accuracy level obtained via the EM algorithm, and we can then use the EM solution as an initialization for gradient descent within a $O(r(\theta_0))$-radius neighborhood, to find an FOSP; since the expected Hessian is positive definite, with high probability this FOSP is also a SSOSP. We omit the details.

Finally, we check Assumption 6.4.2, which will require some substantial calculations. To verify Assumption 6.4.2, we will check the following stronger condition

$$\mathbb{E}_{\theta_0}\left[\exp\left\{\sup_{\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta} r(\theta_0)^2 \cdot \|H(\theta; X) - H(\theta)\|\right\}\right] \leq c' e^{\epsilon(\theta_0)},$$

for any $r(\theta_0) = o(n^{-1/4})$ and $\epsilon(\theta_0) \gtrsim r(\theta_0)^2 n^{1/2} + r(\theta_0)^3 n$. We first calculate, for parameter

$\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$,

$$(\theta; x) = -\sum_{i=1}^{n} \log\left(\pi_1 \phi(x_i; \mu_1, \sigma_1^2) + (1-\pi_1)\phi(x_i; \mu_2, \sigma_2^2)\right),$$

where $\phi(t; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(t-\mu)^2/2\sigma^2}$ is the density of the normal distribution. After some calculations, we can verify that the Hessian takes the form

$$\begin{aligned} H(\theta; x) = \sum_{i=1}^{n} \Bigg[\sum_{m=0}^{2} x_i^m \cdot \Big(& a_{1,m}(\theta) f_1(x_i; \theta) + a_{2,m}(\theta) f_2(x_i; \theta) \\ & + b_{1,m}(\theta) f_1(x_i; \theta)^2 + b_{2,m}(\theta) f_2(x_i; \theta)^2 \Big) + \sum_{m=0}^{4} x_i^m \cdot c_m(\theta) f_1(x_i; \theta) f_2(x_i; \theta) \Bigg], \end{aligned}$$

where we define

$$f_1(t; \theta) = \frac{\pi_1 \phi(t; \mu_1, \sigma_1^2)}{\pi_1 \phi(t; \mu_1, \sigma_1^2) + (1-\pi_1)\phi(t; \mu_2, \sigma_2^2)}$$

and

$$f_2(t; \theta) = \frac{(1-\pi_1) \phi(t; \mu_2, \sigma_2^2)}{\pi_1 \phi(t; \mu_1, \sigma_1^2) + (1-\pi_1)\phi(t; \mu_2, \sigma_2^2)},$$

and where $a_{1,m}, a_{2,m}, b_{1,m}, b_{2,m}, c_m : \Theta \to \mathbb{R}^{5\times 5}$ are continuously differentiable functions (whose details we omit for brevity). We can rewrite this as

$$H(\theta; x) = \sum_{i=1}^{n} g_0(x_i; \theta) + x_i g_1(x_i; \theta) + x_i^2 g_2(x_i; \theta)$$

where

$$\begin{aligned} g_0(t; \theta) = a_{1,0}(\theta) f_1(t; \theta) + a_{2,0}(\theta) f_2(t; \theta) + b_{1,0}(\theta) f_1(t; \theta)^2 + b_{2,0}(\theta) f_2(t; \theta)^2 \\ + \sum_{m=0}^{4} c_m(\theta) t^m f_1(t; \theta) f_2(t; \theta) \end{aligned}$$

and

$$g_m(t;\theta) = a_{1,m}(\theta)f_1(t;\theta) + a_{2,m}(\theta)f_2(t;\theta) + b_{1,m}(\theta)f_1(t;\theta)^2 + b_{2,m}(\theta)f_2(t;\theta)^2$$

for $m = 1, 2$. Some additional calculations prove that we can find finite $C_m(\theta_0), C'_m(\theta_0)$ such that, as long as $r(\theta_0)$ is bounded by some appropriately chosen constant,

$$\sup_{t\in\mathbb{R}} \sup_{\theta\in\mathbb{B}(\theta_0,r(\theta_0))\cap\Theta} \|g_m(t;\theta)\| \leq C_m(\theta_0)$$

and

$$\sup_{t\in\mathbb{R}} \sup_{\theta\in\mathbb{B}(\theta_0,r(\theta_0))\cap\Theta} \|\nabla_\theta g_m(t;\theta)\| \leq C'_m(\theta_0).$$

(To give some intuition for this—for example, for the zeroth-order term, i.e., finding $C_m(\theta_0)$, it is trivial to see that $\sup_{t\in\mathbb{R}} f_\ell(t;\theta) \leq 1$ for each $\ell = 1, 2$; what is more subtle is the observation that $\sup_{t\in\mathbb{R}} t^m f_1(t;\theta)f_2(t;\theta)$ is also finite, as long as $\mu_1 \neq \mu_2$—and this condition is ensured as long as we enforce $(\mu_1)_0 \neq (\mu_2)_0$, i.e., the means are unequal in the true parameter θ_0, and $r(\theta_0)$ is taken to be sufficiently small.)

We then calculate

$$\|H(\theta;x) - H(\theta)\| \leq \|H(\theta;x) - H(\theta_0;x)\| + \|H(\theta_0;x) - H(\theta_0)\| + \|H(\theta) - H(\theta_0)\|.$$

For the first term, for all $\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta$,

$$\begin{aligned}
&\|H(\theta; x) - H(\theta_0; x)\| \\
&= \left\| \sum_{i=1}^n (g_0(x_i; \theta) - g_0(x_i; \theta_0)) + x_i (g_1(x_i; \theta) - g_1(x_i; \theta_0)) + x_i^2 (g_2(x_i; \theta) - g_2(x_i; \theta_0)) \right\| \\
&\leq \sum_{i=1}^n \|g_0(x_i; \theta) - g_0(x_i; \theta_0)\| + |x_i| \, \|g_1(x_i; \theta) - g_1(x_i; \theta_0)\| + x_i^2 \, \|g_2(x_i; \theta) - g_2(x_i; \theta_0)\| \\
&\leq \sum_{i=1}^n C_0'(\theta_0) r(\theta_0) + |x_i| C_1'(\theta_0) r(\theta_0) + x_i^2 C_2'(\theta_0) r(\theta_0) \\
&\leq r(\theta_0) \left[n \left(C_0'(\theta_0) + 0.5 C_1'(\theta_0) \right) + \sum_{i=1}^n x_i^2 \left(C_2'(\theta_0) + 0.5 C_1'(\theta_0) \right) \right].
\end{aligned}$$

Similarly, for the third term,

$$\|H(\theta) - H(\theta_0)\| \leq r(\theta_0) \left[n \left(C_0'(\theta_0) + 0.5 C_1'(\theta_0) \right) + \sum_{i=1}^n \mathbb{E}_{\theta_0}[X_i^2] \left(C_2'(\theta_0) + 0.5 C_1'(\theta_0) \right) \right].$$

By Cauchy–Schwarz, then,

$$\begin{aligned}
\log \mathbb{E}_{\theta_0} &\left[\exp \left\{ \sup_{\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta} r(\theta_0)^2 \cdot \|H(\theta; X) - H(\theta)\| \right\} \right] \\
&\leq \frac{1}{2} \log \mathbb{E}_{\theta_0} \left[\exp \left\{ 2 r(\theta_0)^2 \cdot \|H(\theta_0; X) - H(\theta_0)\| \right\} \right] \\
+ \frac{1}{2} \log \mathbb{E}_{\theta_0} &\left[\exp \left\{ 2 \sup_{\theta \in \mathbb{B}(\theta_0, r(\theta_0)) \cap \Theta} r(\theta_0)^2 \cdot (\|H(\theta; X) - H(\theta_0; X)\| + \|H(\theta) - H(\theta_0)\|) \right\} \right] \\
&\leq \frac{1}{2} \log \mathbb{E}_{\theta_0} \left[\exp \left\{ 2 r(\theta_0)^2 \cdot \|H(\theta_0; X) - H(\theta_0)\| \right\} \right] + c(\theta_0) \cdot n r(\theta)^3,
\end{aligned}$$

for an appropriate function $c(\theta_0)$, since the X_i^2's are subexponential under P_{θ_0}.

Next we bound the remaining term. Since the Hessian is a 5×5 matrix, for any $c > 0$

we have

$$
\begin{aligned}
&\mathbb{E}_{\theta_0}\left[\exp\left\{c\cdot\|H(\theta_0;X)-H(\theta_0)\|\right\}\right]\\
&\quad\leq\mathbb{E}_{\theta_0}\left[\exp\left\{5c\cdot\|H(\theta_0;X)-H(\theta_0)\|_\infty\right\}\right]\\
&\quad=\mathbb{E}_{\theta_0}\left[\exp\left\{5c\cdot\max_{j=1,\dots,5}\max_{k=1,\dots,5}\max\left\{H(\theta_0;X)_{jk}-H(\theta_0)_{jk},H(\theta_0)_{jk}-H(\theta_0;X)_{jk}\right\}\right\}\right]\\
&\quad\leq\sum_{j=1}^5\sum_{k=1}^5\mathbb{E}_{\theta_0}\left[\exp\left\{5c\left|H(\theta_0;X)_{jk}-H(\theta_0)_{jk}\right|\right\}\right]\\
&\quad\leq\sum_{j=1}^5\sum_{k=1}^5\mathbb{E}_{\theta_0}\left[\exp\left\{5c(H(\theta_0;X)_{jk}-H(\theta_0)_{jk})\right\}\right]\\
&\qquad+\sum_{j=1}^5\sum_{k=1}^5\mathbb{E}_{\theta_0}\left[\exp\left\{5c(H(\theta_0)_{jk}-H(\theta_0;X)_{jk})\right\}\right].
\end{aligned}
$$

Now we handle each term individually. We have

$$
\begin{aligned}
&\mathbb{E}_{\theta_0}\left[\exp\left\{5c(H(\theta_0;X)_{jk}-H(\theta_0)_{jk})\right\}\right]\\
&\quad=\mathbb{E}_{\theta_0}\left[\exp\left\{5c\sum_{i=1}^n\sum_{m=0}^2\left[X_i^m g_m(X_i;\theta_0)_{jk}-\mathbb{E}_{\theta_0}[X_i^m g_m(X_i;\theta_0)_{jk}]\right]\right\}\right]\\
&\quad\leq\prod_{m=0}^2\mathbb{E}_{\theta_0}\left[\exp\left\{15c\sum_{i=1}^n\left[X_i^m g_m(X_i;\theta_0)_{jk}-\mathbb{E}_{\theta_0}[X_i^m g_m(X_i;\theta_0)_{jk}]\right]\right\}\right]^{1/3}.
\end{aligned}
$$

Since X_i^m is subexponential for each $m=0,1,2$ while $g_m(X_i;\theta_0)_{jk}$ is bounded, and the product of a bounded random variable and a subexponential random variable is subexponential, we have

$$
\mathbb{E}_{\theta_0}\left[\exp\left\{15c\sum_{i=1}^n\left[X_i^m g_m(X_i;\theta_0)_{jk}-\mathbb{E}_{\theta_0}[X_i^m g_m(X_i;\theta_0)_{jk}]\right]\right\}\right]\leq e^{c^2nc'_{m,jk}(\theta_0)}
$$

assuming $c\leq c''_{m,jk}(\theta_0)$, for some positive-valued functions $c'_{m,jk},c''_{m,jk}$. The same type of calculation holds for the terms of the form $\mathbb{E}_{\theta_0}\left[\exp\left\{5c(H(\theta_0)_{jk}-H(\theta_0;X)_{jk})\right\}\right]$, for some positive-valued functions $\tilde{c}'_{m,jk}, \tilde{c}''_{m,jk}$. Combining everything,

$$\mathbb{E}_{\theta_0}\left[\exp\left\{c \cdot \|H(\theta_0; X) - H(\theta_0)\|\right\}\right] \leq \sum_{j=1}^{5}\sum_{k=1}^{5}\prod_{m=0}^{2} e^{\frac{1}{3}c^2 n c'_{m,jk}(\theta_0)} + \sum_{j=1}^{5}\sum_{k=1}^{5}\prod_{m=0}^{2} e^{\frac{1}{3}c^2 n \tilde{c}'_{m,jk}(\theta_0)},$$

for $0 < c < c''(\theta_0) = \min_{m,j,k}\min\{c''_{m,jk}(\theta_0), \tilde{c}''_{m,jk}(\theta_0)\}$. Letting

$$c'(\theta_0) = \max_{m,j,k}\max\{c''_{m,jk}(\theta_0), \tilde{c}''_{m,jk}(\theta_0)\},$$

then,

$$\mathbb{E}_{\theta_0}\left[\exp\left\{c \cdot \|H(\theta_0; X) - H(\theta_0)\|\right\}\right] \leq 50e^{c^2 n c'(\theta_0)}.$$

Choosing $c > r(\theta_0)^2$, then, by Jensen's inequality,

$$\begin{aligned}\mathbb{E}_{\theta_0}\left[\exp\left\{r(\theta_0)^2 \cdot \|H(\theta_0; X) - H(\theta_0)\|\right\}\right] &\leq \mathbb{E}_{\theta_0}\left[\exp\left\{c \cdot \|H(\theta_0; X) - H(\theta_0)\|\right\}\right]^{r(\theta_0)^2/c} \\ &\leq (50e^{c^2 n c'(\theta_0)})^{r(\theta_0)^2/c} = \exp\left\{\frac{r(\theta_0)^2}{c}\log 50 + r(\theta_0)^2 c n c'(\theta_0)\right\}.\end{aligned}$$

Choosing $c = \sqrt{\frac{\log 50}{nc'(\theta_0)}}$, then, which (for sufficiently large n) satisfies $c > r(\theta_0)^2$ and $c < c''(\theta_0)$,

$$\mathbb{E}_{\theta_0}\left[\exp\left\{r(\theta_0)^2 \cdot \|H(\theta_0; X) - H(\theta_0)\|\right\}\right] \leq \exp\left\{r(\theta_0)^2 \cdot 2\sqrt{nc'(\theta_0)\log 50}\right\}.$$

Combining everything, the assumption holds with $r(\theta_0) = o(n^{-1/4})$ and $\epsilon(\theta_0) \gtrsim r(\theta_0)^2 n^{1/2} + r(\theta_0)^3 n$.

E.4 Sampling details

For Example 1, we use MCMC to generate the copies $\tilde{X}^{(m)}$; see details in Section E.4.1. For Example 2 and 3, the conditional distribution is tractable, and we sample directly from the conditional distribution; see details in Section E.4.2.

E.4.1 Implementation details for Example 1 (Gaussian mixture model

For the Gaussian mixture model, the copies $\tilde{X}^{(m)}$ are sampled via MCMC. Here we give the details for this process.

When sampling directly from $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ is infeasible, Barber and Janson [2022] discusses two schemes for constructing copies with MCMC sampling: the Hub-and-spoke sampler and the Permuted serial sampler. In our simulation for Example 1, aCSS (with and without constraints) is run with the hub-and-spoke sampler. Given X and $\hat{\theta}, \hat{g}$, we sample the copies as follows:

- Initialize at X, and run the Markov chain (specified below) for L steps to define the “hub" $\tilde{X}^*$.

- Independently for $m = 1, \ldots, M$, initialize at $\tilde{X}^*$ and run the Markov chain (specified below) for L steps to define the “spoke” $\tilde{X}^m$.

Similar to Barber and Janson [2022], we can use use the Metropolis–Hastings (MH) to construct an efficient sampling scheme. Given $\hat{\theta}$, the reversible MCMC is given by the following:

- Starting at state x', generate a proposal x according to a properly chosen proposal distribution $q_{\hat{\theta}}(x \mid x')$.
- With probability $A_{\hat{\theta}}(x \mid x') = \min\left\{1, \frac{q_{\hat{\theta}}(x'|x)}{q_{\hat{\theta}}(x|x')} \frac{p_{\hat{\theta}}(x|\hat{\theta},\hat{g})}{p_{\hat{\theta}}(x'|\hat{\theta},\hat{g})}\right\}$, set the next state to equal x. Otherwise, the next state is set to equal x'.

Next, we will describe the proposal distribution and MH acceptance probability; we also refer to [Barber and Janson, 2022, Appendix D.2] for more details.

Proposal distribution $q_{\hat{\theta}}(x \mid x')$

In Example 1, the model P_θ is a product distribution with density

$$f_\theta(x) = \prod_{i=1}^{n} f_\theta^i(x_i).$$

We then use the same proposal distribution as [Barber and Janson, 2022, Examples 1,2,4]. For $s \in [n]$, define $q_{\hat{\theta}}(x|x')$ as follows:

- Draw a subset $\mathcal{S} \subseteq \{1, \ldots, n\}$ of size s, uniformly at random.
- For each $i = 1, \ldots, n$,
  - Set $x_i = x_i'$, if $i \notin S$,
  - Draw $x_i \sim f_{\hat{\theta}}^{(i)}$, if $i \in S$.

Here s controls the tradeoff between two goals: (1) the acceptance probability $A_{\hat{\theta}}(x|x')$ should not be too close to zero; (2) the proposed state should not be too similar to the previous state. Note that we can tune this MCMC hyperparameter after looking at $\hat{\theta}$ without violating any of our theoretical assumptions. We can then choose s based on the following simulation:

- Let $\theta_0^{\mathrm{sim}} = \hat{\theta}$.
- Draw $X^{\mathrm{sim}} \sim P_{\theta_0^{\mathrm{sim}}}$, $W \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$; calculate $\hat{\theta}^{\mathrm{sim}} = \hat{\theta}\left(X^{\mathrm{sim}}, W\right)$, and $\hat{g}^{\mathrm{sim}} = \nabla(\hat{\theta}^{\mathrm{sim}}; X^{\mathrm{sim}}, W)$.
- For each candidate of s , run one step of Metropolis-Hasting initialized at X^{sim} to generate X^{new}.
- Repeat for 100 draws of X^{sim}, discarding any draws for which $\hat{\theta}^{\mathrm{sim}}$ is not a SSOSP, to get an average acceptance probability $\bar{A}_s$. Among all values of s where $\bar{A}_s \geq 0.05$, choose s that maximizes $s\bar{A}_s$.

Note that this choice of s only depends on $\hat{\theta}$, and completing our θ-dependent definition of the proposal distribution $q_{\hat{\theta}}(x \mid x')$. Then we choose $L = \min\{2000, \frac{2n}{s\hat{A}_s}\}$ to ensure that most entries will be resampled within L steps.

MH acceptance probability

Given $\hat{\theta}, \hat{g}$, and a properly chosen proposal distribution $q_{\hat{\theta}}(x \mid x')$, the MH acceptance probability $A_{\hat{\theta}}(x \mid x')$ can be written as

$$A_{\hat{\theta}}(x \mid x') = \min\left\{1, \frac{q_{\hat{\theta}}(x' \mid x)}{q_{\hat{\theta}}(x \mid x')} \frac{p_{\hat{\theta}}(x \mid \hat{\theta}, \hat{g})}{p_{\hat{\theta}}(x' \mid \hat{\theta}, \hat{g})}\right\},$$

where

$$p_{\hat{\theta}}(x|\hat{\theta}, \hat{g}) \propto f(x; \hat{\theta}) \exp\left\{-\frac{\|\hat{g} - \nabla(\hat{\theta}; x)\|^2}{2\sigma^2/d}\right\} \det\left(U_{\mathcal{I}(\hat{\theta})}^\top \nabla_\theta^2(\hat{\theta}; x) U_{\mathcal{I}(\hat{\theta})}\right) \mathbb{1}_{x \in \mathcal{X}_{\hat{\theta},\hat{g}}}$$

The ratio in the MH acceptance probability without the indicator variables are straightforward to calculate. The ratio with indicator variables $\mathbb{1}_{x \in \mathcal{X}_{\hat{\theta},\hat{g}}} / \mathbb{1}_{x' \in \mathcal{X}_{\hat{\theta},\hat{g}}}$ requires more careful consideration. First, we will always have $\mathbb{1}_{x' \in \mathcal{X}_{\hat{\theta},\hat{g}}} = 1$ since x' is sampled from (6.7) with $x' \in \mathcal{X}_{\hat{\theta},\hat{g}}$. To check $\mathbb{1}_{x \in \mathcal{X}_{\hat{\theta},\hat{g}}}$, we have

$$\begin{aligned}\mathbb{1}_{x \in \mathcal{X}_{\hat{\theta},\hat{g}}} &= \mathbb{1}\left\{\exists w \in \mathbb{R}^d \text{ s.t. } \hat{\theta} = \hat{\theta}(x, w) \text{ is a SSOSP of (6.3), and } \hat{g} = \nabla(\hat{\theta}; x, w)\right\} \\ &= \mathbb{1}\left\{\hat{\theta}\left(x, \frac{\hat{g} - \nabla_\theta(\hat{\theta}; x)}{\sigma}\right) = \hat{\theta}, \text{and } U_{\mathcal{I}(\hat{\theta})}^\top \nabla_\theta^2(\hat{\theta}; x) U_{\mathcal{I}(\hat{\theta})} \succ 0\right\}.\end{aligned}$$

This means given proposed x, we only need to verify (1) $U_{\mathcal{I}(\hat{\theta})}^\top \nabla_\theta^2(\hat{\theta}; x) U_{\mathcal{I}(\hat{\theta})} \succ 0$ and (2) the algorithm $\hat{\theta}\left(x, \frac{\hat{g} - \nabla_\theta(\hat{\theta}; x)}{\sigma}\right)$ returns value $\hat{\theta}$.

E.4.2 Implementation details for Examples 2 (isotonic regression) and 3 (sparse regression)

In this section, we derive the sampling distribution for the copies $\tilde{X}^{(m)}$ for the two Gaussian linear model examples.

Recall that the objective function $(\theta; x, w)$ is defined as

$$(\theta; x, w) = \frac{1}{2\nu^2}\|x - Z\theta\|^2 + \mathcal{R}(\theta) + \sigma w^\top \theta,$$

and

$$\begin{cases} \hat{\theta} = \hat{\theta}(X, W), \\ \hat{g} = \frac{1}{\nu^2} Z^\top (Z\hat{\theta} - X) + \nabla_\theta \mathcal{R}(\hat{\theta}) + \sigma W, \end{cases}$$

where $\hat{\theta}(X, W)$ is the minimizer of $(\theta; X, W)$ subject to arbitrary linear constraints or ℓ_1 penalty. Note that the original aCSS is a special case of the constrained aCSS with no constraints and $\hat{g} = 0$. When $(\theta; x, w)$ is strictly convex (like if we add ridge penalty), a unique SSOSP exists (and is computationally efficient to find), and we can then define $\hat{\theta}(x, w)$ to be equal to this unique SSOSP. Based on the conditional density derived in (6.7), we can efficiently compute the conditional distribution $p_{\theta_0}(\cdot \mid \hat{\theta}, \hat{g})$ as

$$\mathcal{N}\left(Z\hat{\theta} + \left(\mathbf{I}_n + \frac{d}{\sigma^2\nu^2} ZZ^\top\right)^{-1} Z(\theta_0 - \hat{\theta} + \frac{d}{\sigma^2}(\nabla_\theta \mathcal{R}(\hat{\theta}) - \hat{g})), \nu^2 \left(\mathbf{I}_n + \frac{d}{\sigma^2\nu^2} ZZ^\top\right)^{-1}\right).$$

The plug-in conditional distribution $\tilde{X}$, i.e., $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$, is

$$\tilde{X} \sim \mathcal{N}\left(Z\hat{\theta} + \left(\mathbf{I}_n + \frac{d}{\sigma^2\nu^2} ZZ^\top\right)^{-1} Z\frac{d}{\sigma^2}(\nabla_\theta \mathcal{R}(\hat{\theta}) - \hat{g}), \nu^2 \left(\mathbf{I}_n + \frac{d}{\sigma^2\nu^2} ZZ^\top\right)^{-1}\right).$$

In Example 2, we choose $\mathcal{R}(\theta) = 0$, $Z = \mathbf{I}_n$ and $\nu^2 = 1$. Details of sampling using the aCSS method, with and without constraints, are as follows:

- For Barber and Janson [2022]'s aCSS method, $\hat{\theta}$ is computed via perturbed and unconstrained maximum likelihood estimation,

$$\hat{\theta} = \hat{\theta}_{\text{OLS}} = \text{argmin}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|X - \theta\|^2 + \sigma W^\top \theta \right\} = X - \sigma W,$$

  and then the copies $\tilde{X}^{(m)}$ are sampled directly from $p_{\hat{\theta}}(\cdot \mid \hat{\theta})$ via the distribution

$$\tilde{X}^{(m)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\hat{\theta}, \left(1 + \frac{n}{\sigma^2}\right)^{-1} \mathbf{I}_n\right).$$

- For our proposed constrained aCSS method, $\hat{\theta}$ is computed with the isotonic constraint,

$$\hat{\theta} = \hat{\theta}_{\text{iso}} = \text{argmin}_{\substack{\theta \in \mathbb{R}^n \\ \theta_1 \leq \cdots \leq \theta_n}} \left\{ \frac{1}{2} \|X - \theta\|^2 + \sigma W^\top \theta \right\},$$

  the gradient is given by

$$\hat{g} = \hat{\theta} - X + \sigma W,$$

  and then the copies $\tilde{X}^{(m)}$ are sampled directly from $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ via the distribution

$$\tilde{X}^{(m)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\hat{\theta} - \frac{n/\sigma^2}{1 + n/\sigma^2} \hat{g}, \left(1 + \frac{n}{\sigma^2}\right)^{-1} \mathbf{I}_n\right).$$

In Example 3, we choose $\mathcal{R}(\theta) = \frac{\lambda_{\text{ridge}}}{2} \|\theta\|^2$ as a ridge penalization with $\lambda_{\text{ridge}} = 0.01$, $\nu^2 = 1$. Details of sampling using the aCSS method, with and without an ℓ_1 penalty, are as follows:

- For Barber and Janson [2022]'s aCSS method, we will use a ridge regularizer. The method is then defined by setting

$$\hat{\theta} = \hat{\theta}_{\text{ridge}} = \text{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2}\|X - Z\theta\|^2 + \frac{\lambda_{\text{ridge}}}{2}\|\theta\|^2 + \sigma W^\top \theta \right\}$$
$$= \left(\lambda_{\text{ridge}}\mathbf{I}_d + Z^T Z\right)^{-1} (Z^T X - \sigma W),$$

and then sampling the copies $\tilde{X}^{(m)}$ directly from $p_{\hat{\theta}}(\cdot \mid \hat{\theta})$ via the distribution

$$\tilde{X}^{(m)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(Z\hat{\theta} + \frac{\lambda_{\text{ridge}} d}{\sigma^2} \left(\mathbf{I}_n + \frac{d}{\sigma^2} ZZ^\top\right)^{-1} Z\hat{\theta}, \left(\mathbf{I}_n + \frac{d}{\sigma^2} ZZ^\top\right)^{-1} \right).$$

- For our proposed penalized aCSS method, in order to be more comparable to aCSS, we also add the regularizer $R(\theta)$. This means that our estimator is given by the elastic net, incorporating both ℓ_1 and ℓ_2 penalization:

$$\hat{\theta} = \hat{\theta}_{\text{elastic-net}} = \text{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2}\|X - Z\theta\|^2 + \frac{\lambda_{\text{ridge}}}{2}\|\theta\|^2 + \lambda\|\theta\|_1 + \sigma W^\top \theta \right\},$$

with $\lambda = 2$, and the gradient is then computed as

$$\hat{g} = Z^T(Z\hat{\theta} - X) + \sigma W + \lambda_{\text{ridge}}\hat{\theta}.$$

We then sample the copies $\tilde{X}^{(m)}$ directly from $p_{\hat{\theta}}(\cdot \mid \hat{\theta}, \hat{g})$ via the distribution

$$\tilde{X}^{(m)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(Z\hat{\theta} + \frac{d}{\sigma^2} \left(\mathbf{I}_n + \frac{d}{\sigma^2} ZZ^\top\right)^{-1} Z(\lambda_{\text{ridge}}\hat{\theta} - \hat{g}), \left(\mathbf{I}_n + \frac{d}{\sigma^2} ZZ^\top\right)^{-1} \right).$$