Compute Faster and Learn Better:

Model-based Nonconvex

Optimization for Machine Learning

by

Tuo Zhao

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

Nov, 2016

Abstract

Nonconvex optimization naturally arises in many machine learning problems. Machine learning researchers exploit various nonconvex formulations to gain modeling flexibility, estimation robustness, adaptivity, and computational scalability. Although classical computational complexity theory has shown that solving nonconvex optimization is generally NP-hard in the worst case, practitioners have proposed numerous heuristic optimization algorithms, which achieve outstanding empirical performance in real-world applications.

To bridge this gap between practice and theory, we propose a new generation of model-based optimization algorithms and theory, which incorporate the statistical thinking into modern optimization. Particularly, when designing practical computational algorithms, we take the underlying statistical models into consideration. Our novel algorithms exploit hidden geometric structures behind many nonconvex optimization problems, and can obtain global optima with the desired statistics properties in polynomial time with high probability.

Acknowledgments

I would like to thank my advisors Dr. Han Liu at Princeton University and Dr. Raman Arora at Johns Hopkins University for their patience, enthusiasm in research and valuable advices in life. Their competence and humbleness inspired me over the past five years and will influence me forever. I am deeply grateful for support and friendship. This work would not have been possible without their help.

I would also like to express my special appreciation to my collaborators Dr. Tong Zhang at Rutgers University and Dr. Kathryn Roeder at Carnegie Mellon University. You have also been tremendous mentors for me. Your advices on both research as well as on my career have been priceless.

I am grateful to the members of my GBO committee – Dr. Brian Caffo, Dr. Michael Rosenblum, Dr. Mark Dredze – for their support throughout my doctoral program, and constructive comments and suggestions during my doctoral dissertation work.

My special thank to the wonderful faculty and staff members at Johns Hopkins

ACKNOWLEDGMENTS

University and Princeton University. A special gratitude I give to Dr. Vladimir Braverman, Dr. Randal Burns, Dr. Jianqing Fan, Dr. Gregory Hager, Dr. Samory Kpotufe, Dr. Xin Li, Dr. Mengdi Wang, Dr. David Yarovsky for their determination in sharing their expertise without restrictions. Thanks to Michael Bino, Connie Brown, Zack Burwell, Debbie DeFord, Laura Graham, Melissa Holmes, Kimberly Lupinacci, Tracy Marshall, Tonette McClamy, Shani P. McPherson, Tabitha Mischler, Dr. Joanne Selinski, Carol Smith, Javonnia Thomas, Cathy Thornton, and Tara Zigler for their exceptional efforts to help me in every way.

I would like to thank the friends that been with me at Johns Hopkins University, Princeton University, University of Minnesota, and Iowa State University. Each of you helped me learn and grow in the past five years: Dr. Xingyuan Fang, Jian Ge, Dr. Quanquan Gu, Dr. Fang Han, Dr. Mingyi Hong, Xingguo Li, Zaoxing Liu, Huanran Lu, Junwei Lu, Cong Ma, Dr. Yang Ning, Dr. Qiang Sun, Dr. Weichen Wang, Yiming Wang, Zhaoran Wang, Zhuoran Yang, Lin Yang, Dr. Mo Yu, and Tianqi Zhao.

Finally, I want to express my deepest gratitude to my parents, Jianhua Zhao and Guirong Zhang, and my wife, Yanbo Xu, who have always done their best to encourage and support me through the process of seeking my degree. Nothing I say will suffice to describe how much their efforts mean to me.

Dedication

This thesis is dedicated to my parents Jianhua Zhao and Guirong Zhang, my wife Yanbo Xu, and my son Boyi Zhao.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Nonconvex optimization naturally arises in many statistical machine learning problems. Statisticians and machine learning scientists exploit various nonconvex formulations to gain desired computational and statistical properties (e.g. estimation robustness, modeling flexibility, computational efficiency, and scalability, [5, 6, 7, 8, 9, 10, 11, 12, 13]). Typical real-world applications include, for instance: analyzing sequencing data from high throughput genomic experiments, image data from fMRI (functional Magnetic Resonance Imaging), proteomic data from tandem mass spectrometry analysis, climate data from geographically distributed data centers, and social media data from eBusiness [14, 15, 16].

Most work on these nonconvex problems treats the statistical properties and practical algorithms separately. On one hand, practitioners proposed numerous heuristic nonconvex optimization algorithms, many of which have been corrobo-

rated to achieve very good empirical performance in real-world applications. On the other hand, existing statistical theory only establishes the statistical properties for a small set of these nonconvex problems. Even worse, most of these statistical properties are established only on hypothetical global optimum, which have been shown to be intractable to obtain in the worst case by theoretical computer scientists. Thus, there exists a significant gap between theory and practice: **What has been proved is not the same as what is being widely used!**

To address this crucial computational and statistical challenge, we focus on developing a new generation of *statistical optimization algorithms* and *model-based computational theory*, which incorporates the statistical thinking into modern optimization. These new algorithms and theory naturally bridges researchers from different areas, including machine learning, statistics, optimization, and stochastic analysis. More specifically, we address the following two important nonconvex problems in statistical machine learning:

Problem (1) Nonconvex Sparsity-inducing Regularization and Constraint: The SCAD (Smooth Clipped Absolute Deviation) and MCP (Minimax Concavity Penalty) regularizers, and the ℓ_0 constraint have been widely used for variable selection in high dimensional regularized M-estimation problems. They can effectively reduce the estimation bias, and make the obtained estimator attain significantly better statistical performance in both parameter estimation and support recovery than

the convex ℓ_1-regularized estimator [9, 13].

Problem (2) Biconvex Loss: These loss functions are very popular in low rank matrix factorization problems such as matrix completion, noisy matrix decomposition, and matrix regression. Compared with related convex approaches, they avoid intensive singular value decompositions, and therefore gain significant improvement in computational efficiency and scalability [11, 12].

The above two nonoconvex optimization problems have been extensively studied by researchers from conventional optimization community. However, their theory does not take the *underlying statistical models* into consideration so they do not help statisticians to establish statistical guarantees. More precisely, the underlying statistical models contain very rich *distributional information*, which enables us to develop new algorithms and more refined theory to establish computational and statistical guarantees for nonconvex optimization problems. This unconventional research weaves the knowledge of statistics and optimization at a fundamental level.

To tackle the nonconvexity in **Problem (1)** (Nonconvex Regularized or Constrained M-estimation), we exploit the **restricted strong convexity** of their nonconvex objective functions. Particularly, when restricted to a sparse set involving only a few coordinates, these nonconvex objective functions mimic the behavior of a strongly convex function. Thus the key to tackle the nonconvexity is to de-

vise a mechanism, under which the solution path achieved by the optimization algorithm always falls within the **restricted convex regions**. The restricted strong convexity has been considered as one of the most important conditions in existing statistical literature on developing high dimensional statistical guarantees. This condition can also provide us new insights on developing computational guarantees for nonconvex optimization algorithms. Particularly, we exploits the restricted strong convexity and develops theoretical guarantees for *pathwise coordinate optimization* in Chapter 2[9, 10].

The pathwise coordinate optimization has gained significant success in practice, and has been widely recognized as one of the most important computational frameworks for solving high dimensional sparse learning problems with the MCP and SCAD regularizers. It differs from the classical coordinate optimization in three salient features: *warm start initialization*, *strong rule for coordinate preselection*, and *active set strategy*. These three features grant superior empirical performance, but also pose significant challenge to theoretical analysis. To close this long lasting problem, we proposed a novel analytical framework. This framework shows that these three features play pivotal roles in guaranteeing the outstanding statistical and computational performance of the pathwise coordinate optimization framework. In particular, we analyzed the existing pathwise coordinate optimization algorithms, and developed a precise characterization of the solution sparsity patterns. Our analysis lead to several **new active set updating rules** and **initializa-**

tion strategies to improve existing pathwise coordinate optimization algorithms. Through a simple but elegant proof, we showed that for the nonconvex optimization problems in **Problem (1)**, the proposed improved algorithms guarantee *linear convergence* to a *unique sparse local optimum* with the same *optimal statistical properties* as the global optimum. This is the **first** result establishing the strong computational and statistical guarantees of the pathwise coordinate optimization framework in high dimensions [9, 10].

In addition, we apply our model-based optimization technique to high dimensional sparse learning problems with the ℓ_0 constraint in Chapter 3. Specifically, we propose a novel stochastic variance reduced gradient hard thresholding algorithm. By exploiting the restricted strong convexity, we show that the proposed stochastic optimization algorithm also enjoys strong linear convergence guarantees and nearly optimal statistical accuracy. We further extend our proposed algorithm to an asynchronous variant for parallel nonconvex optimization with a provable linear speedup. This is the **first** result establishing the strong computational and statistical guarantees for nonconvex stochastic optimization with variance reduction in high dimensions[13].

To address the nonconvexity in **Problem (2)** (Low Rank Matrix Factorization), we propose a novel analytical framework for analyzing popular nonconvex optimization algorithms such as *alternating minimization* and *alternating gradient descent algorithms* in Chapter 4. Specifically, our proposed framework shows that

these algorithms are essentially solving a sequence of convex optimization problems but using **inexact gradient** information. Then by exploiting our proposed model-based computational theory, we show that given a proper initialization, these algorithms can guarantee the error of the inexact gradient diminishes with the iterations. This eventually allows me to establish *global linear convergence* to *global optima* for these algorithms. To the best of our knowledge, this is the **first** unified computational and statistical theory for a broad class of nonconvex low rank matrix factorization algorithms [11, 12].

Chapter 2

Pathwise Coordinate Optimization for Nonconvex Sparse Learning

This chapter introduces our proposed novel pathwise coordinate optimization algorithm for solving nonconvex sparse learning problems. By investigating the data generating process (underlying statistical models) of sparse learning problems, we show that the resulting nonconvex optimization problem shows strong convexity and smoothness over a sparse domain. Therefore, by exploiting such hidden convex structures, we establish new computational and statistical theory for our proposed optimization algorithm.

2.1 Background

Modern data acquisition routinely produces massive amount of high dimensional data, where the number of variables d greatly exceeds the sample size n, such as high throughput genomic data [14] and image data from functional Magnetic Resonance Imaging [15]. To handle high dimensionality, we often assume that only a small subset of variables are relevant in modeling [17]. Such a parsimonious assumption motivates various sparse learning approaches. Taking sparse linear regression as an example, we consider a linear model $y = X\theta^* + \epsilon$, where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n\times d}$ is the design matrix, $\theta^* = (\theta_1, ..., \theta_d)^\top \in \mathbb{R}^d$ is the unknown sparse regression coefficient vector, $\epsilon \sim N(0, \sigma^2 I)$ is the random noise, and $I \in \mathbb{R}^{n\times n}$ is the identity matrix. Let $\|\cdot\|_2$ denote the ℓ_2 norm, and $\mathcal{R}_\lambda(\theta)$ denote a sparsity-inducing regularizer with a regularization parameter $\lambda > 0$. We can obtain a sparse estimator of θ^* by solving the following regularized least square optimization problem

$$\min_{\theta\in\mathbb{R}^d} \mathcal{F}_\lambda(\theta), \text{ where } \mathcal{F}_\lambda(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2 + \mathcal{R}_\lambda(\theta). \tag{2.1.1}$$

Popular choices of $\mathcal{R}_\lambda(\theta)$ are usually coordinate decomposable, $\mathcal{R}_\lambda(\theta) = \sum_{j=1}^d r_\lambda(\theta_j)$, including the ℓ_1 (Lasso, [18]), SCAD (Smooth Clipped Absolute Deviation, [19]), and MCP (Minimax Concavity Penalty, [20]) regularizers. For example, the ℓ_1 regularizer takes $\mathcal{R}_\lambda(\theta) = \lambda\|\theta\|_1 = \lambda\sum_j |\theta_j|$ with $r_\lambda(|\theta_j|) = \lambda|\theta_j|$ for $j = 1, ..., d$.

The ℓ_1 regularizer is convex and computationally tractable, but often induces large estimation bias, and requires a restrictive irrepresentable condition to attain variable selection consistency [21, 22]. To address this issue, nonconvex regularizers such as SCAD and MCP have been proposed to obtain nearly unbiased estimators. Throughout the rest of the chapter, we only consider MCP as an example due to space limit, but the extension to SCAD is straightforward. Particularly, let $\mathcal{E}$ be an event, we define $\mathbb{1}_{\{\mathcal{E}\}}$ as an indicator function with $\mathbb{1}_{\{\mathcal{E}\}} = 1$ if $\mathcal{E}$ holds and $\mathbb{1}_{\{\mathcal{E}\}} = 0$ otherwise. Given $\gamma > 1$, MCP has

$$r_\lambda(|\theta_j|) = \lambda\left(|\theta_j| - \frac{\theta_j^2}{2\lambda\gamma}\right) \cdot \mathbb{1}_{\{|\theta_j| < \lambda\gamma\}} + \frac{\lambda^2\gamma}{2} \cdot \mathbb{1}_{\{|\theta_j| \geq \lambda\gamma\}}. \tag{2.1.2}$$

We call γ the concavity parameter of MCP, since it essentially characterizes the concavity of the MCP regularizer: A larger γ implies that the regularizer is less concave. We observe that the MCP regularizer can be written as

$$\mathcal{R}_\lambda(\theta) = \lambda\|\theta\|_1 + \mathcal{H}_\lambda(\theta), \tag{2.1.3}$$

where $\mathcal{H}_\lambda(\theta) = \sum_{j=1}^d h_\lambda(|\theta_j|)$ is a smooth, concave, and also coordinate decomposable function with

$$h_\lambda(|\theta_j|) = -\frac{\theta_j^2}{2\gamma} \cdot \mathbb{1}_{\{|\theta_j| < \lambda\gamma\}} + \frac{\lambda^2\gamma - 2\lambda|\theta_j|}{2} \cdot \mathbb{1}_{\{|\theta_j| \geq \lambda\gamma\}}. \tag{2.1.4}$$

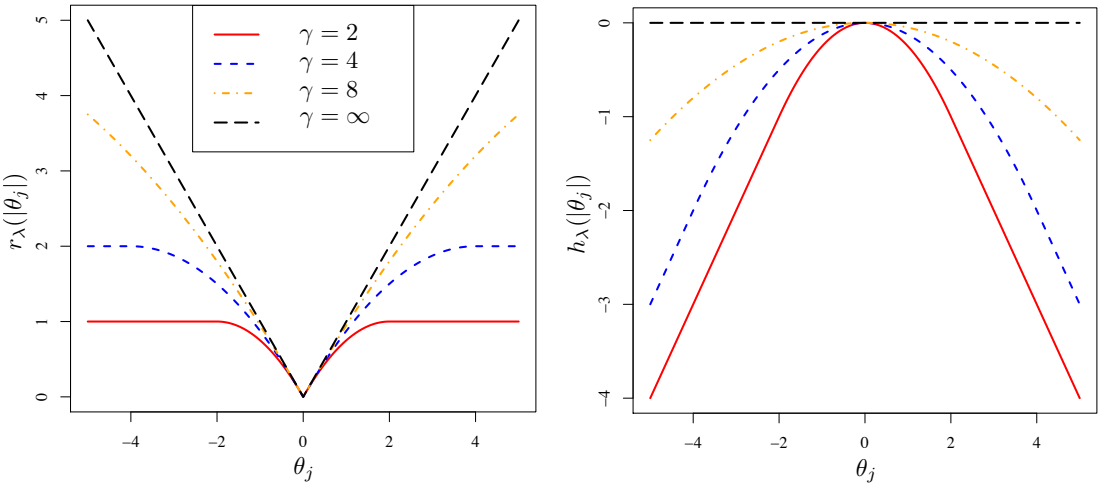


Figure 2.1: Several examples of the MCP regularizer with $\lambda = 1$ and $\gamma = 2, 4, 8$, and ∞ (Lasso). The MCP regularizer reduces the estimation bias and achieve better performance than the ℓ_1 regularizer in both parameter estimation and support recovery, but imposes great computational challenge.

We present several examples of the MCP regularizer in Figure 2.1.

[19, 20] show that the nonconvex regularizer effectively reduces the estimation bias, and achieve better performance than the ℓ_1 regularizer in both parameter estimation and support recovery. Particularly, given a suitable chosen $\gamma < \infty$, they show that there exits a local optimum to (2.1.1), which attains the oracle properties under much weaker conditions. However, they cannot not provide specific algorithms that guarantee such a local optimum in polynomial time due to the nonconvexity.

Typical algorithms for solving (2.1.1) developed in existing optimization literature include proximal gradient algorithms [23] and coordinate optimization algorithms [24]. The proximal gradient algorithms need to access all entries of the

design matrix X in each iteration for computing a full gradient and a sophisticated line search step. Thus, they are often not scalable and efficient in practice when d is large. To address this issue, many researchers resort to the coordinate optimization algorithms for better computational efficiency and scalability.

The classical coordinate optimization algorithm is straightforward and much simpler than the proximal gradient algorithms in each iteration: Given $\theta^{(t)}$ at the t-th iteration, we select a coordinate j, and then take an exact coordinate minimization step

$$\theta_j^{(t+1)} = \underset{\theta_j}{\operatorname{argmin}} \ \mathcal{F}_\lambda(\theta_j, \theta_{\setminus j}^{(t)}), \tag{2.1.5}$$

where $\theta_{\setminus j}$ is a subvector of θ with the j-th entry removed. For the ℓ_1, SCAD, and MCP regularizers, (2.1.5) admits a closed form solution. For notational simplicity, we denote $\theta_j^{(t+1)} = \mathcal{T}_{\lambda,j}(\theta^{(t)})$. Then (2.1.5) can be rewritten as

$$\theta_j^{(t+1)} = \mathcal{T}_{\lambda,j}(\theta^{(t)}) = \underset{\theta_j}{\operatorname{argmin}} \frac{1}{2n}\|z^{(t)} - X_{*j}\theta_j\|_2^2 + r_\lambda(\theta_j), \tag{2.1.6}$$

where X_{*j} denotes the j-th column of X and $z^{(t)} = y - X\theta^{(t)} + X_{*j}\theta_j^{(t)}$ is the partial residual. Without loss of generality, we assume that X satisfies the column normalization condition $\|X_{*j}\|_2 = \sqrt{n}$ for all $j = 1, \ldots, d$. Let $\widetilde{\theta}_j^{(t)} = \frac{1}{n} X_{*j}^\top z^{(t)}$. Then for

MCP, we obtain $\theta_j^{(t+1)}$ by

$$\theta_j^{(t+1)} = \widetilde{\theta}_j^{(t)} \cdot \mathbb{1}_{\{|\widetilde{\theta}_j^{(t)}| \geq \gamma\lambda\}} + \frac{\mathcal{S}_\lambda(\widetilde{\theta}_j^{(t)})}{1 - 1/\gamma} \cdot \mathbb{1}_{\{|\widetilde{\theta}_j^{(t)}| < \gamma\lambda\}}, \tag{2.1.7}$$

where $\mathcal{S}_\lambda(a) = \text{sign}(a) \cdot \max\{|a| - \lambda, 0\}$. As shown in Appendix A.1, (2.1.7) can be efficiently calculated by a simple partial residual update trick, which only requires the access to one single column of the design matrix X_{*j} (Recall the proximal gradient algorithms need to access the entire design matrix). Once we obtain $\theta_j^{(t+1)}$, we take $\theta_{\backslash j}^{(t+1)} = \theta_{\backslash j}^{(t)}$. Such a coordinate optimization algorithm, though simple, is not necessarily efficient in theory and practice. Existing optimization theory only shows its sublinear convergence to local optima in high dimensions if we select coordinates from 1 to d in a cyclic order throughout all iterations [25]. Moreover, no theoretical guarantee has been established on statistical properties of the obtained estimators for nonconvex regularizers in parameter estimation and support recovery. Thus, the coordinate optimization algorithms were almost neglected until recent rediscovery by [3, 4, 26].

Remark 2.1.1 (Connection between MCP and Lasso)**.** Let $\frac{c}{\infty} = 0$ for any constant c. As can be seen from (2.1.2), for $\gamma = \infty$, MCP is reduced to the ℓ_1 regularizer, i.e., $r_\lambda(|\theta_j|) = \lambda|\theta_j|$ with $h_\lambda(|\theta_j|) = 0$. Accordingly, (2.1.7) is reduced to $\theta_j^{(t+1)} = \mathcal{S}_\lambda(\widetilde{\theta}_j^{(t)})$, which is identical to the updating formula of the coordinate optimization algorithm proposed in [27] for Lasso. Thus, throughout the rest of the chapter, we

just simply consider the ℓ_1 regularizer as a special case of MCP, unless we clearly specify the difference between $\gamma < \infty$ and $\gamma = \infty$ for MCP.

As illustrated in Figure 2.2, [28, 4, 26] propose a pathwise coordinate optimization framework with three nested loops, which integrates the warm start initialization, active set updating strategy, and strong rule for coordinate preselection into the classical coordinate optimization.

Particularly, in the *outer loop*, the warm start initialization optimizes (2.1.1) with a sequence of decreasing regularization parameters in a multistage manner, and yields solutions from sparse to dense. Within each stage of the warm start initialization (an iteration of the outer loop), the algorithm uses the solution from the previous stage for initialization, and then adopts the active set updating strategy to exploit the solution sparsity to speed up computation. The active set updating strategy contains two consequent nested loops: In the *middle loop*, the algorithm first divides all coordinates into active ones (active set) and inactive ones (inactive set) based on some heuristic coordinate gradient thresholding rule (strong rule, [26]). Then within each iteration of the middle loop, an *inner loop* is called to conduct coordinate optimization. In general, the algorithm runs an inner loop on the current active coordinates until convergence, with all inactive coordinates remain zero. The algorithm then exploits some heuristic rule to identify a new active set, which further decreases the objective value and repeats the inner loops. The iteration within each stage terminates when the active set in the middle loop no longer

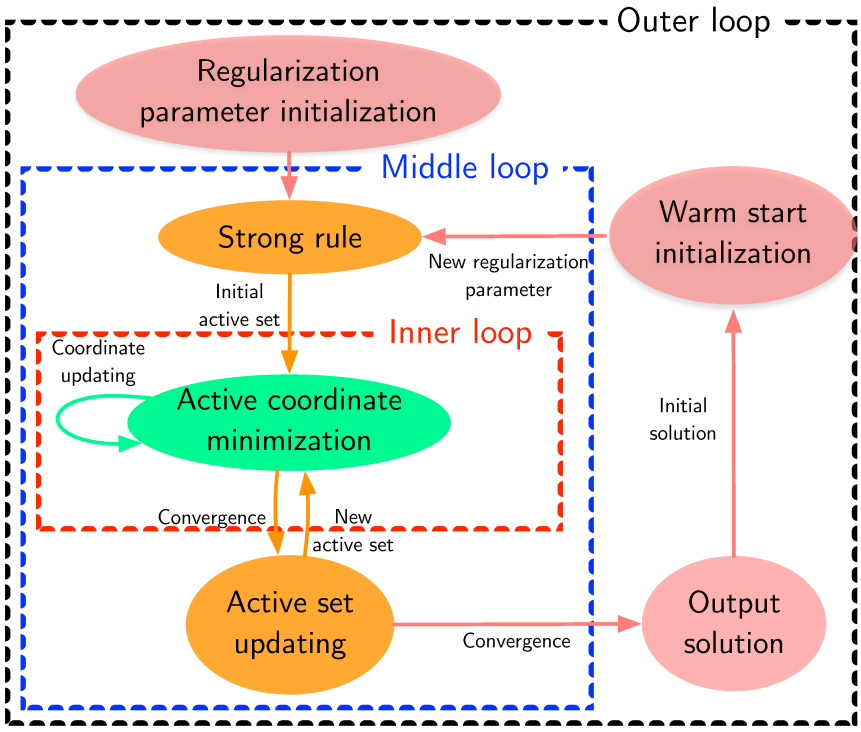


Figure 2.2: The pathwise coordinate optimization framework contains 3 nested loops: (I) Warm start initialization; (II) Active set updating and strong rule for coordinate preselection; (III) Active coordinate minimization. Many empirical results have corroborated its outstanding performance. Detailed descriptions of the three loops is presented in Section 3.2.

changes. In practice, the warm start initialization, active set updating strategies, and strong rule for coordinate preselection encourage the algorithm to iterate over a small active set involving only a small number of coordinates, and therefore significantly boost the computational efficiency and scalability. Software packages such as `GLMNET` and `SparseNet` have been developed and widely applied to many research areas.

Despite of the popularity of the pathwise coordinate optimization framework, we are still in lack of adequate theory to justify its superior computational per-

formance due to its complex algorithmic structure. The warm start initialization, active set updating strategy, and strong rule for coordinate preselection are only considered as engineering heuristics in existing literature. On the other hand, many experimental results have shown that the pathwise coordinate optimization framework is effective at finding local optima with good empirical performance, yet no theoretical guarantee has been established. Thus, a gap exists between theory and practice.

To bridge this gap, we propose a new algorithm, named PICASSO (Path-wIse CalibrAted Sparse Shooting algOrithm), which improves the existing pathwise coordinate optimization framework. Particularly, we propose a new greedy selection rule for active set updating and a new convex relaxation based warm start initialization strategy (for sparse learning problems using general loss functions beyond the least square loss). These modifications though simple, have a profound impact: The solution sparsity and restricted strong convexity can be ensured throughout all iterations, which allows us to establish statistical and computational guarantees of PICASSO in high dimensions [29, 30]. Eventually, we prove that PICASSO attains a linear convergence to a unique sparse local optimum with optimal statistical properties in parameter estimation and support recovery (See more details in Section 2.3). To the best of our knowledge, this is the first result on the computational and statistical guarantees for the pathwise coordinate optimization framework in high dimensions. Besides algorithm and theory, we also have the proposed

algorithm implemented using C with a R wrapper. The latest version is available on `https://cran.r-project.org/web/packages/picasso`.

Several proximal gradient algorithms are closely related to PICASSO. By exploiting similar sparsity structures of the optimization problem, [31, 10, 32] show that these proximal gradient algorithms also attain linear convergence to (approximate) local optima with guaranteed statistical properties. We will compare these algorithms with PICASSO in Section 2.6.

The rest of this chapter is organized as follows: In Section 3.2, we present the PICASSO algorithm; In Section 2.3 we present a new theory for analyzing the pathwise coordinate optimization framework, and establish the computational and statistical properties of PICASSO for sparse linear regression; In Section 2.4, we extend PICASSO to other sparse learning problems with general loss functions, and provide theoretical guarantees; In Section 2.5, we present thorough numerical experiments to support our theory; In Section 2.6, we discuss related work; In Section 2.7, we present the proofs of the theorems. Due to space limit, the proofs of all lemmas are deferred to Appendix A.

Notations: Given a vector $v = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$, we define vector norms: $\|v\|_1 = \sum_j |v_j|$, $\|v\|_2^2 = \sum_j v_j^2$, and $\|v\|_\infty = \max_j |v_j|$. We denote the number of nonzero entries in v as $\|v\|_0 = \sum_j \mathbb{1}_{\{v_j \neq 0\}}$. We define the soft-thresholding function and operator as $\mathcal{S}_\lambda(v_j) = \text{sign}(v_j) \cdot \max\{|v_j| - \lambda, 0\}$ and $\mathcal{S}_\lambda(v) = \Big(\mathcal{S}_\lambda(v_1), ..., \mathcal{S}_\lambda(v_d)\Big)^\top$. We denote $v_{\setminus j} = (v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_d)^\top \in \mathbb{R}^{d-1}$ as the subvector of v with the j-th entry removed.

Let $\mathcal{A} \subseteq \{1,...,d\}$ be an index set. We use $\overline{\mathcal{A}}$ to denote the complementary set to $\mathcal{A}$, i.e. $\overline{\mathcal{A}} = \{j \mid j \in \{1,...,d\},\ j \notin \mathcal{A}\}$. We use $v_{\mathcal{A}}$ to denote a subvector of v by extracting all entries of v with indices in $\mathcal{A}$. Given a matrix $A \in \mathbb{R}^{d\times d}$, we use $A_{*j} = (A_{1j},...,A_{dj})^\top$ to denote the j-th column of A, and $A_{k*} = (A_{k1},...,A_{kd})^\top$ to denote the k-th row of A. Let $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ be the largest and smallest eigenvalues of A. We define the matrix norms $\|A\|_{\mathrm{F}}^2 = \sum_j \|A_{*j}\|_2^2$ and $\|A\|_2$ as the largest singular value of A. We denote $A_{\backslash i\backslash j}$ as the submatrix of A with the i-th row and the j-th column removed. We denote $A_{i\backslash j}$ as the i-th row of A with its j-th entry removed. Let $\mathcal{A} \subseteq \{1,...,d\}$ be an index set. We use $A_{\mathcal{A}\mathcal{A}}$ to denote a submatrix of A by extracting all entries of A with both row and column indices in $\mathcal{A}$.

2.2 Pathwise Calibrated Sparse Shooting Algorithm

We introduce the PICASSO algorithm for sparse linear regression. PICASSO is a pathwise coordinate optimization algorithm and contains three nested loops (as illustrated in Figure 2). For simplicity, we first introduce its inner loop, then its middle loop, and at last its outer loop.

2.2.1 Inner Loop: Iterates over Coordinates within an Active Set

We start with the inner loop of PICASSO, which is the active coordinate minimization (ActCooMin) algorithm. The iteration index for the inner loop is (t), where $t = 0, 1, 2, \ldots$. Recall we are interested in the following nonconvex optimization problem

$$\min_{\theta \in \mathbb{R}^d} \mathcal{F}_\lambda(\theta), \text{ where } \mathcal{F}_\lambda(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2 + \mathcal{R}_\lambda(\theta). \tag{2.2.1}$$

As illustrated in Algorithm 1, the ActCooMin algorithm solves (2.2.1) by iteratively conducting exact coordinate minimization, but it is only allowed to iterate over a subset of all coordinates, which is called "the active set". Accordingly, the complementary set to the active set is called "the inactive set", because the values of these coordinates do not change throughout all iterations of the inner loop. Since the active set usually contains a very small number of coordinates, the active set coordinate minimization algorithm is very scalable and efficient.

For notational simplicity, we denote the active and inactive sets by $\mathcal{A}$ and $\overline{\mathcal{A}}$ respectively. Here we select $\mathcal{A}$ and $\overline{\mathcal{A}}$ based on the sparsity pattern of the initial

solution of the inner loop $\theta^{(0)}$,

$$\mathcal{A} = \{j \mid \theta_j^{(0)} \neq 0\} \quad \text{and} \quad \overline{\mathcal{A}} = \{j \mid \theta_j^{(0)} = 0\}.$$

The ActCooMin algorithm then minimizes (2.2.1) with all coordinates of $\overline{\mathcal{A}}$ staying at zero values,

$$\min_{\theta \in \mathbb{R}^d} \mathcal{F}_\lambda(\theta) \quad \text{subject to} \quad \theta_{\overline{\mathcal{A}}} = 0. \tag{2.2.2}$$

The ActCooMin algorithm iterates over all active coordinates in a cyclic order at each iteration. Without loss of generality, we assume

$$|\mathcal{A}| = s \quad \text{and} \quad \mathcal{A} = \{j_1, ..., j_s\} \subseteq \{1, ..., d\},$$

where $j_1 \leq j_2 \leq ... \leq j_s$. Given a solution $\theta^{(t)}$ at the t-th iteration, we construct a sequence of auxiliary solutions $\{w^{(t+1,k)}\}_{k=0}^s$ to obtain $\theta^{(t+1)}$. Particularly, for $k = 0$, we take $w^{(t+1,0)} = \theta^{(t)}$; For $k = 1, ..., s$, we take

$$w_{j_k}^{(t+1,k)} = \mathcal{T}_{\lambda, j_k}(w^{(t+1,k-1)}) \quad \text{and} \quad w_{\setminus j_k}^{(t+1,k)} = w_{\setminus j_k}^{(t+1,k-1)},$$

where $\mathcal{T}_{\lambda, j_k}(\cdot)$ is defined in (2.1.6). We then set $\theta^{(t+1)} = w^{(t+1,s)}$ for the next iteration. Given τ as a small convergence parameter (e.g. 10^{-5}), we terminate the ActCooMin

Algorithm 1: The active coordinate minimization algorithm (ActCooMin) is the inner loop of PICASSO. It iterates over only a small subset of all coordinates in a cyclic order. Thus, its computation is scalable and efficient. Without loss of generality, we assume $|\mathcal{A}| = s$ and $\mathcal{A} = \{j_1, ..., j_s\} \subseteq \{1, ..., d\}$, where $j_1 \leq j_2 \leq ... \leq j_s$.

Algorithm: $\widehat{\theta} \leftarrow \text{ActCooMin}(\lambda, \theta^{(0)}, \mathcal{A}, \tau)$
Initialize: $t \leftarrow 0$
Repeat
  $w^{(t+1,0)} \leftarrow \theta^{(t)}$
  For $k \leftarrow 1, ..., s$
    $w_{j_k}^{(t+1,k)} \leftarrow \mathcal{T}_{\lambda, j_k}(w^{(t+1,k-1)})$, $w_{\backslash j_k}^{(t+1,k)} \leftarrow w_{\backslash j_k}^{(t+1,k-1)}$
  $\theta^{(t+1)} \leftarrow w^{(t+1,s)}$
  $t \leftarrow t + 1$
Until $\|\theta^{(t+1)} - \theta^{(t)}\|_2 \leq \tau\lambda$
Return: $\widehat{\theta} \leftarrow \theta^{(t)}$

algorithm when

$$\|\theta^{(t+1)} - \theta^{(t)}\|_2 \leq \tau\lambda. \tag{2.2.3}$$

We then take the output solution as $\widehat{\theta} = \theta^{(t+1)}$.

The ActCooMin algorithm only converges to a local optimum of (2.2.2), which is not necessarily a local optimum of (2.2.1). Thus, PICASSO needs to combine this inner loop with some active set updating scheme, which allows the active set to change. This leads to the middle loop of PICASSO.

2.2.2 Middle Loop: Iteratively Updates Active Sets

We then introduce the middle loop of PICASSO, which is the iterative active set updating (IteActUpd) algorithm. The iteration index of the middle loop is $[m]$, where $m = 0, 1, 2, \ldots$. As illustrated in Algorithm 2, the IteActUpd algorithm simultaneously decreases the objective value and iteratively changes the active set to ensure convergence to a local optimum to (2.2.1). For notational simplicity, we denote the least square loss function and its gradient as $\mathcal{L}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$ and $\nabla\mathcal{L}(\theta) = \frac{1}{n}X^\top(X\theta - y)$.

(I) Active Set Initialization by Strong Rule: We first introduce how PICASSO initializes the active set for each middle loop. Suppose an initial solution $\theta^{[0]}$ is supplied to the middle loop of PICASSO. [3] suggest a straightforward "simple rule" to initialize the active set based on the sparsity pattern of $\theta^{[0]}$,

$$\mathcal{A}_0 = \{j \mid \theta_j^{[0]} \neq 0\} \quad \text{and} \quad \overline{\mathcal{A}}_0 = \{j \mid \theta_j^{[0]} = 0\}. \tag{2.2.4}$$

[26] further show that (2.2.4) is sometimes too conservative, and suggest a more aggressive active set initialization procedure using a "strong rule", which often leads to better computational performance in practice. Specifically, given an active set initialization parameter $\varphi \in (0, 1)$, the strong rule[1] for PICASSO initializes $\mathcal{A}_0$

[1] Our proposed strong rule for PICASSO is sightly different from the sequential strong rule proposed in [26]. See more details in Remark 2.2.1.

and $\overline{\mathcal{A}}_0$ as

$$\mathcal{A}_0 = \{j \mid \theta_j^{[0]} = 0,\ |\nabla_j \mathcal{L}(\theta^{[0]})| \geq (1-\varphi)\lambda\} \cup \{j \mid \theta_j^{[0]} \neq 0\}, \tag{2.2.5}$$

$$\overline{\mathcal{A}}_0 = \{j \mid \theta_j^{[0]} = 0,\ |\nabla_j \mathcal{L}(\theta^{[0]})| < (1-\varphi)\lambda\}, \tag{2.2.6}$$

where $\nabla_j \mathcal{L}(\theta^{[0]})$ denotes the j-th entry of $\nabla \mathcal{L}(\theta^{[0]})$. As can be seen from (2.2.5), the strong rule yields an active set, which is no smaller than the simple rule. Note that we need the initialization parameter φ to be a reasonably small value (e.g. 0.1). Otherwise, the strong rule may select too many active coordinates and compromise the solution sparsity.

(II) Active Set Updating Strategy: We then introduce how PICASSO updates the active set at each iteration of the middle loop. Suppose at the m-th iteration ($m \geq 1$), we are supplied with a solution $\theta^{[m]}$ with a pair of active and inactive sets defined as

$$\mathcal{A}_m = \{j \mid \theta_j^{[m]} \neq 0\} \quad \text{and} \quad \overline{\mathcal{A}}_m = \{j \mid \theta_j^{[m]} = 0\}.$$

Each iteration of the IteActUpd algorithm contains two stages. The first stage conducts the active coordinate minimization algorithm over the active set $\mathcal{A}_m$ until convergence, and returns a solution $\theta^{[m+0.5]}$. Note that the active coordinate minimization algorithm may yield zero values for some active coordinates. Accord-

ingly, we remove these coordinates from the active set, and obtain a new pair of active and inactive sets as

$$\mathcal{A}_{m+0.5} = \{j \mid \theta_j^{[m+0.5]} \neq 0\} \quad \text{and} \quad \overline{\mathcal{A}}_{m+0.5} = \{j \mid \theta_j^{[m+0.5]} = 0\}.$$

The second stage checks which inactive coordinates of $\overline{\mathcal{A}}_{m+0.5}$ should be added into the active set. Existing pathwise coordinate optimization algorithms usually add inactive coordinates into the active set based on a *cyclic selection rule* [3, 4]. Particularly, they conduct the exact coordinate minimization over all coordinates of $\overline{\mathcal{A}}_{m+0.5}$ in a cyclic order. Accordingly, an inactive coordinate is added into the active set if the corresponding exact coordinate minimization yields a nonzero value. Such a cyclic selection rule, however, has no control over the solution sparsity. It may add too many inactive coordinates into the active set, and compromise the solution sparsity.

To address this issue, we propose a new greedy selection rule for updating the active set. Particularly, let $\nabla_j \mathcal{L}(\theta^{[m+0.5]})$ denote the j-th entry of $\nabla \mathcal{L}(\theta^{[m+0.5]})$. We select a coordinate by

$$k_m = \text{argmax}_{k \in \overline{\mathcal{A}}_{m+0.5}} |\nabla_k \mathcal{L}(\theta^{[m+0.5]})|.$$

Algorithm 2: The iterative active set updating (IteActUpd) algorithm is the middle loop of PICASSO. It simultaneously decreases the objective value and iteratively changes the active set. To encourage the sparsity of the active set, the greedy selection rule moves only one inactive coordinate to the active set in each iteration.

Algorithm: $\widehat{\theta} \leftarrow \text{IteActUpd}(\lambda, \theta^{[0]}, \delta, \tau, \varphi)$
Initialize: $m \leftarrow 0$, $\mathcal{A}_0 \leftarrow \{j \mid \theta_j^{[0]} = 0,\ |\nabla_j \mathcal{L}(\theta^{[0]})| \geq (1-\varphi)\lambda\} \cup \{j \mid \theta_j^{[0]} \neq 0\}$
Repeat
  $\theta^{[m+0.5]} \leftarrow \text{ActCooMin}(\lambda, \theta^{[m]}, \mathcal{A}_m, \tau)$
  $\mathcal{A}_{m+0.5} \leftarrow \{j \mid \theta_j^{[m+0.5]} \neq 0\}$, $\overline{\mathcal{A}}_{m+0.5} \leftarrow \{j \mid \theta_j^{[m+0.5]} = 0\}$
  $k_m \leftarrow \text{argmax}_{k \in \overline{\mathcal{A}}_{m+0.5}} |\nabla_k \mathcal{L}(\theta^{[m+0.5]})|$
  $\theta_{k_m}^{[m+1]} \leftarrow \mathcal{T}_{\lambda, k_m}(\theta^{[m+0.5]})$, $\theta_{\backslash k_m}^{[m+1]} \leftarrow \theta_{\backslash k_m}^{[m+0.5]}$
  $\mathcal{A}_{m+1} \leftarrow \mathcal{A}_{m+0.5} \cup \{k_m\}$, $\overline{\mathcal{A}}_{m+1} \leftarrow \overline{\mathcal{A}}_{m+0.5} \setminus \{k_m\}$
  $m \leftarrow m + 1$
Until $|\nabla_{k_m} \mathcal{L}(\theta^{[m+0.5]})| \leq (1+\delta)\lambda$
Return: $\widehat{\theta} \leftarrow \theta^{[m]}$

We then terminate the IteActUpd algorithm if

$$|\nabla_{k_m} \mathcal{L}(\theta^{[m+0.5]})| \leq (1+\delta)\lambda, \tag{2.2.7}$$

where δ is a small convergence parameter (e.g. 10^{-5}). Otherwise, we take

$$\theta_{k_m}^{[m+1]} = \mathcal{T}_{\lambda, k_m}(\theta^{[m+0.5]}) \quad \text{and} \quad \theta_{\backslash k_m}^{[m+1]} = \theta_{\backslash k_m}^{[m+0.5]},$$

and set the new active and inactive sets as

$$\mathcal{A}_{m+1} = \mathcal{A}_{m+0.5} \cup \{k_m\} \quad \text{and} \quad \overline{\mathcal{A}}_{m+1} = \overline{\mathcal{A}}_{m+0.5} \setminus \{k_m\}.$$

The IteActUpd algorithm, though equipped with the proposed greedy selection rule and strong rule for coordinate preselection, ensures the solution sparsity throughout iterations only for a sufficiently large regularization parameter[2]. Otherwise, given an insufficiently large regularization parameter, the IteActUpd algorithm may still overselect active coordinates. To address this issue, we combine the IteActUpd algorithm with a sequence of decreasing regularization parameters, which leads to the outer loop of PICASSO.

2.2.3 Outer Loop: Iterates over Regularization Parameters

The outer loop of PICASSO is the warm start initialization (WarmStartInt). The iteration index of the outer loop is $\{K\}$, where $K = 1, ..., N$. As illustrated in Algorithm 3, the warm start initialization solves (2.1.1) indexed by a geometrically decreasing sequence of regularization parameters $\{\lambda_K = \lambda_0 \eta^K\}_{K=0}^N$ with a common ratio $\eta \in (0, 1)$, and outputs a sequence of $N + 1$ solutions $\{\widehat{\theta}^{\{K\}}\}_{K=0}^N$, which is also called the solution path.

For sparse linear regression[3], the warm start initialization chooses the leading regularization parameter λ_0 as $\lambda_0 = \|\nabla \mathcal{L}(0)\|_\infty = \|\frac{1}{n} X^\top y\|_\infty$. Recall $\mathcal{H}_\lambda(\theta)$ is defined

[2] As will be shown in Section 2.3, the choice of λ is determined by the initial solution of the middle loop.

[3] When dealing with general loss functions, we need a new convex relaxation based warm start initialization approach, which will be introduced in Section 2.4.2.

in (2.1.3). By verifying the KKT condition, we have

$$\min_{\xi \in \partial \|0\|_1} \|\nabla \mathcal{L}(0) + \nabla \mathcal{H}_{\lambda_0}(0) + \lambda_0 \xi\|_\infty = \min_{\xi \in \partial \|0\|_1} \|\nabla \mathcal{L}(0) + \lambda_0 \xi\|_\infty = 0,$$

where the first equality comes from $\nabla \mathcal{H}_{\lambda_0}(0) = 0$ for the MCP regularizer (See more details in Appendix A.2). This indicates that 0 is a local optimum of (2.1.1). Accordingly, we set $\widehat{\theta}^{\{0\}} = 0$. Then for $K = 1, 2, ..., N$, we solve (2.1.1) for λ_K using $\widehat{\theta}^{\{K-1\}}$ as initialization.

The warm start initialization starts with large regularization parameters to suppress the overselection of irrelevant coordinates $\{j \mid \theta_j^* = 0\}$ (in conjunction with the IteActUpd algorithm). Thus, the solution sparsity ensures the restricted convexity throughout all iterations, making the algorithm behaves as if minimizing a strongly convex function. Though large regularization parameters may also yield zero values for many relevant coordinates $\{j \mid \theta_j^* \neq 0\}$ and result in larger estimation errors, this can be compensated by the decreasing regularization sequence. Eventually, PICASSO gradually recovers the relevant coordinates, reduces the estimation error of each output solution, and attains a sparse output solution with optimal statistical properties in parameter estimation and support recovery.

Remark 2.2.1 (Connection to the sequential strong rule)**.** [26] propose a sequential

Algorithm 3: The warm start initialization is the **outer loop** of PICASSO. It solves (2.1.1) with respect to a decreasing sequence of regularization parameters $\{\lambda_K\}_{K=0}^N$. The leading regularization parameter λ_0 is chosen as $\lambda_0 = \|\nabla \mathcal{L}(0)\|_\infty$, which yields an all zero output solution $\widehat{\theta}^{\{0\}} = 0$. For $K = 1, ..., N$, we solve (2.1.1) for λ_K using $\widehat{\theta}^{\{K-1\}}$ as an initial solution. $\{\tau_K\}_{K=1}^N$ and $\{\delta_K\}_{K=1}^N$ are two sequences of small convergence parameters, where τ_K and δ_K correspond to the K-th outer loop iteration with the regularization parameter λ_K.

Algorithm: $\{\widehat{\theta}^{\{K\}}\}_{K=0}^N \leftarrow \text{WarmStartInt}(\{\lambda_K\}_{K=0}^N)$
Parameter: η, φ, $\{\tau_K\}_{K=1}^N$, $\{\delta_K\}_{K=1}^N$
Initialize: $\lambda_0 \leftarrow \|\nabla \mathcal{L}(0)\|_\infty$, $\widehat{\theta}^{\{0\}} \leftarrow 0$
For $K \leftarrow 1, 2, ..., N$
  $\lambda_K \leftarrow \eta \lambda_{K-1}$
  $\widehat{\theta}^{\{K\}} \leftarrow \text{IteActUpd}(\lambda_K, \widehat{\theta}^{\{K-1\}}, \delta_K, \tau_K, \varphi)$
Return: $\{\widehat{\theta}^{\{K\}}\}_{K=0}^N$

strong rule for coordinate preselection, which initializes the active set for λ_K as

$$\mathcal{A}_0 = \{j \mid \theta_j^{[0]} = 0,\ |\nabla_j \mathcal{L}(\theta^{[0]})| \geq 2\lambda_K - \lambda_{K-1}\} \cup \{j \mid \theta_j^{[0]} \neq 0\}, \tag{2.2.8}$$

$$\overline{\mathcal{A}}_0 = \{j \mid \theta_j^{[0]} = 0,\ |\nabla_j \mathcal{L}(\theta^{[0]})| < 2\lambda_K - \lambda_{K-1}\}. \tag{2.2.9}$$

Recall $\lambda_K = \eta \lambda_{K-1}$. Then we have $2\lambda_K - \lambda_{K-1} = (1 - (1-\eta)/\eta)\lambda_K$. This indicates that the sequential strong rule is a special case of our strong rule for PICASSO with $\varphi = (1-\eta)/\eta$.

2.3 Computational and Statistical Theory

We develop a new theory to analyze the pathwise coordinate optimization framework, and establish the computational and statistical properties of PICASSO for

sparse linear regression. Recall our linear model assumption is $y = X\theta^* + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I)$[4]. Moreover, in (2.1.3), we rewrite the nonconvex regularizer as $\mathcal{R}_\lambda(\theta) = \lambda\|\theta\|_1 + \mathcal{H}_\lambda(\theta)$, where $\mathcal{H}_\lambda(\theta) = \sum_{j=1}^d h_\lambda(|\theta_j|)$ is a smooth, concave, and coordinate decomposable function. For notational simplicity, we define $\widetilde{\mathcal{L}}_\lambda(\theta) = \mathcal{L}(\theta) + \mathcal{H}_\lambda(\theta)$. Accordingly, we rewrite $\mathcal{F}_\lambda(\theta)$ as

$$\mathcal{F}_\lambda(\theta) = \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta) = \widetilde{\mathcal{L}}_\lambda(\theta) + \lambda\|\theta\|_1.$$

2.3.1 Computational Theory

We first introduce three assumptions. The first assumption requires λ_N to be sufficiently large.

Assumption 2.3.1. We require that the regularization sequence satisfies

$$\lambda_N = 8\sigma\sqrt{\frac{\log d}{n}} \geq 4\|\nabla\mathcal{L}(\theta^*)\|_\infty = \frac{4}{n}\|X^\top\epsilon\|_\infty. \tag{2.3.1}$$

Moreover, we require $\eta \in [0.96, 1)$.

Assumption 2.3.1 ensures that all regularization parameters are sufficiently large to eliminate irrelevant coordinates for PICASSO.

Remark 2.3.2. Note that Assumption 2.3.1 is a deterministic bound for our chosen

[4]For simplicity, we only consider the Gaussian noise setting, but it is straight forward to extend our analysis to the subGaussian noise setting.

λ_N. As will be shown in Lemma 2.3.13, since $\|X^\top \epsilon\|_\infty$ is random, we need to verify that (2.3.1) holds with high probability when applying PICASSO to sparse linear regression.

Before we present the second assumption, we define the largest and smallest s sparse eigenvalues of the Hessian matrix $\nabla^2 \mathcal{L}(\theta) = \frac{1}{n} X^\top X$ as follows.

Definition 2.3.3. Given an integer $s \geq 1$, we define the largest and smallest s sparse eigenvalues as

$$\rho_+(s) = \sup_{\|v\|_0 \leq s} \frac{v^\top \nabla^2 \mathcal{L}(\theta) v}{\|v\|_2^2} \quad \text{and} \quad \rho_-(s) = \inf_{\|v\|_0 \leq s} \frac{v^\top \nabla^2 \mathcal{L}(\theta) v}{\|v\|_2^2}.$$

The next lemma connects the largest and smallest s sparse eigenvalues to the restricted strong convexity and smoothness.

Lemma 2.3.4. Suppose there exists an integer s such that $0 < \rho_-(s) \leq \rho_+(s) < \infty$. For any $\theta, \theta' \in \mathbb{R}^d$ satisfying $\|\theta - \theta'\|_0 \leq s$, $\mathcal{L}(\theta)$ is restricted strongly convex and smooth,

$$\frac{\rho_-(s)}{2} \|\theta' - \theta\|_2^2 \leq \mathcal{L}(\theta') - \mathcal{L}(\theta) - (\theta' - \theta)^\top \nabla \mathcal{L}(\theta) \leq \frac{\rho_+(s)}{2} \|\theta' - \theta\|_2^2. \tag{2.3.2}$$

Moreover, given $\alpha = 1/\gamma \leq \rho_-(s)$ and $\widetilde{\rho}_-(s) = \rho_-(s) - \alpha > 0$, where γ is the concavity parameter of MCP defined in (2.1.2), for any $\theta, \theta' \in \mathbb{R}^d$ satisfying $\|\theta - \theta'\|_0 \leq s$,

$\widetilde{\mathcal{L}}_\lambda(\theta)$ is restricted strongly convex and smooth,

$$\frac{\widetilde{\rho}_-(s)}{2}\|\theta' - \theta\|_2^2 \leq \widetilde{\mathcal{L}}_\lambda(\theta') - \widetilde{\mathcal{L}}_\lambda(\theta) - (\theta' - \theta)^\top \nabla \widetilde{\mathcal{L}}_\lambda(\theta) \leq \frac{\rho_+(s)}{2}\|\theta' - \theta\|_2^2.$$

Meanwhile, for any $\xi \in \partial\|\theta\|_1$, $\mathcal{F}_\lambda(\theta)$ is restricted strongly convex,

$$\frac{\widetilde{\rho}_-(s)}{2}\|\theta' - \theta\|_2^2 \leq \mathcal{F}_\lambda(\theta') - \mathcal{F}_\lambda(\theta) - (\theta' - \theta)^\top (\nabla \widetilde{\mathcal{L}}_\lambda(\theta) + \lambda\xi).$$

Lemma 2.3.4 indicates the importance of the solution sparsity: When θ is sufficiently sparse, the restricted strong convexity of $\mathcal{L}(\theta)$ dominates the concavity of $\mathcal{H}_\lambda(\theta)$. Thus, if an algorithm ensures the solution sparsity throughout all iterations, it will behave like minimizing a strongly convex optimization problem. Accordingly, a linear convergence can be established. Note that Lemma 2.3.4 is also applicable to Lasso, since Lasso satisfies $\alpha = 1/\gamma = 1/\infty = 0$. Now we introduce the second assumption.

Assumption 2.3.5. Given $\|\theta^*\|_0 \leq s^*$, there exists an integer $\widetilde{s}$ such that

$$\widetilde{s} \geq (484\kappa^2 + 100\kappa)s^*,\ \rho_+(s^* + 2\widetilde{s}) < \infty, \text{ and } \widetilde{\rho}_-(s^* + 2\widetilde{s}) > 0,$$

where κ is defined as $\kappa = \rho_+(s^* + 2\widetilde{s})/\widetilde{\rho}_-(s^* + 2\widetilde{s})$.

Assumption 2.3.5 guarantees that the optimization problem satisfies the restricted strong convexity as long as the number of active irrelevant coordinates

never exceeds $\widetilde{s}$ throughout all iterations.

Remark 2.3.6. Assumptions 2.3.1 and 2.3.5 are closely related to high dimensional statistical theories for sparse linear regression in existing literature. See more details in [29, 30, 20, 33].

Now we introduce the last assumption on the computational parameters.

Assumption 2.3.7. Recall the convergence parameters δ_K's and τ_K's are defined in Algorithm 3, and the active set initialization parameter φ is defined in (2.2.5). We require for all $K = 1, ..., N$,

$$\delta_K \leq \frac{1}{8}, \quad \tau_K \leq \frac{\delta_K}{\rho_+(s^* + 2\widetilde{s})} \sqrt{\frac{\widetilde{\rho}_-(1)}{\rho_+(1)(s^* + 2\widetilde{s})}}, \quad \text{and} \quad \varphi \leq \frac{1}{8}.$$

Assumption 2.3.7 guarantees that all middle and inner loops of PICASSO attain adequate precision such that their output solutions satisfy the desired computational and statistical properties.

Remark 2.3.8. All constants in our technical assumptions and proofs are for providing insights of PICASSO. We do not make efforts on optimizing any of these constants. Taking Assumption 2.3.1 as an example, we choose $\eta \in [0.96, 1)$ just for easing our analysis. However, η can also be chosen as any other constant, e.g. 0.95, as long as it is sufficiently close to 1. Such a change in η only affects the required sample complexity, iteration complexity, and statistical rates of convergence up to a small constant factor.

Now, we start with the convergence analysis for the inner loop of PICASSO. The following theorem presents the convergence rate in terms of the objective value. For notational simplicity, we omit the outer loop index K, and denote λ_K and τ_K by λ and τ respectively.

Theorem 2.3.9. [Inner Loop] Suppose Assumption 2.3.5 holds. If the initial active set satisfies $|\mathcal{A}| = s \leq s^* + 2\widetilde{s}$, then (2.2.2) is essentially strongly convex. For $t = 1, 2...$, we have

$$\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta}) \leq \left(\frac{s\rho_+^2(s)}{s\rho_+^2(s) + \widetilde{\rho}_-(s)\widetilde{\rho}_-(1)} \right)^t [\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\overline{\theta})],$$

where $\overline{\theta}$ is a unique global optimum to (2.2.2). Moreover, we need at most

$$\left(1 + \frac{s\rho_+^2(s)}{\widetilde{\rho}_-(s)\widetilde{\rho}_-(1)} \right) \cdot \log \left(\frac{2[\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\overline{\theta})]}{\widetilde{\rho}_-(1)\tau^2\lambda^2} \right)$$

iterations to terminate the ActCooMin algorithm, where τ is defined in (2.2.3).

Theorem 2.3.9 guarantees that given a sufficiently sparse active set, Algorithm 1 essentially minimizes a strongly convex optimization problem, though (2.1.1) is globally nonconvex. Thus, it attains a linear convergence to a unique global optimum.

Then, we proceed with the convergence analysis for the middle loop of PICASSO. The following theorem presents the convergence rate in terms of the ob-

jective value. For notational simplicity, we omit the outer loop index K, and denote λ_K and δ_K by λ and δ. Moreover, we define

$$\triangle_\lambda = \frac{4\lambda^2 s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})}, \quad \mathcal{S} = \{j \mid \theta_j^* \neq 0\}, \quad \text{and} \quad \overline{\mathcal{S}} = \{j \mid \theta_j^* = 0\}. \tag{2.3.3}$$

Theorem 2.3.10. [Middle Loop] Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. For any $\lambda \geq \lambda_N$, if the initial solution $\theta^{[0]}$ satisfies $\|\theta^{[0]}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$ and $\mathcal{F}_\lambda(\theta^{[0]}) \leq \mathcal{F}_\lambda(\theta^*) + \triangle_\lambda$, then regardless the active set initialized by either the strong rule or simple rule, we have $|\mathcal{A}_0 \cap \overline{\mathcal{S}}| \leq \widetilde{s}$. Meanwhile, for $m = 0, 1, 2, \ldots$, we also have $\|\theta^{[m]}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s} + 1$, $\|\theta^{[m+0.5]}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$, and

$$\mathcal{F}_\lambda(\theta^{[m]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda) \leq \left(1 - \frac{\widetilde{\rho}_-(s^* + 2\widetilde{s})}{(s^* + 2\widetilde{s})\rho_+(1)}\right)^m [\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda)],$$

where $\overline{\theta}^\lambda$ is a unique sparse local optimum of (2.1.1) satisfying

$$\mathcal{K}_\lambda(\overline{\theta}^\lambda) = \min_{\xi \in \partial\|\overline{\theta}^\lambda\|_1} \|\nabla\widetilde{\mathcal{L}}_\lambda(\overline{\theta}^\lambda) + \lambda\xi\|_\infty = 0 \quad \text{and} \quad \|\overline{\theta}^\lambda_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}. \tag{2.3.4}$$

Moreover, recall δ is defined in (2.2.7), we need at most

$$\frac{(s^* + 2\widetilde{s})\rho_+(1)}{\widetilde{\rho}_-(s^* + 2\widetilde{s})} \cdot \log\left(\frac{\delta\lambda}{3\rho_+(1)[\mathcal{F}_\lambda(\theta^{[0]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda)]}\right).$$

active set updating iterations to terminate the IteActUpd algorithm. Meanwhile,

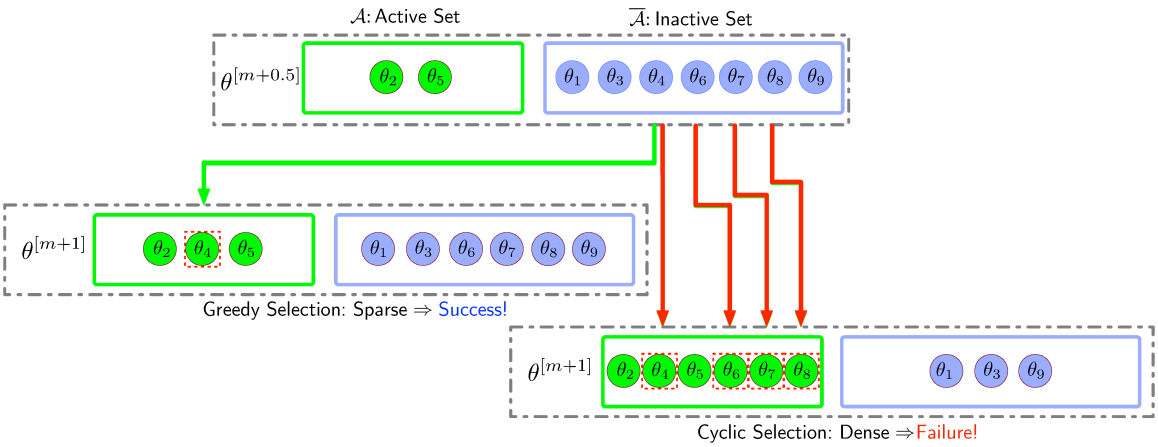


Figure 2.3: An illustration of the failure of the cyclic selection rule. The green and blue circles denote the active and inactive coordinates respectively. Suppose we have 9 coordinates and the maximum number of active coordinates we can tolerate is 4. The greedy selection rule is conservative, and only add one coordinate to the active set each time. Thus, it eventually increases the number of active coordinates from 2 to 3, and prevents the overselecting coordinates. In contrast, the cyclic selection rule used in [3, 4] leads to overselecting coordinates, which eventually increases the number of active coordinates to 6. Thus, it fails to preserve the restricted strong convexity.

we have the output solution $\widehat{\theta}^{\lambda}$ satisfying $\mathcal{K}_{\lambda}(\widehat{\theta}^{\lambda}) \leq \delta\lambda$.

Theorem 2.3.10 guarantees that when supplied a proper initial solution, the middle loop of PICASSO attains a linear convergence to a unique sparse local optimum. Moreover, Theorem 2.3.10 has three important implications:

(I) The greedy rule is conservative and only select one coordinate each time. This mechanism prevents the overselection of irrelevant coordinates and encourages the solution sparsity. In contrast, the cyclic selection rule in [4] may overselect irrelevant coordinates and compromise the restricted convexity. An illustration is provided in Figure 2.3.

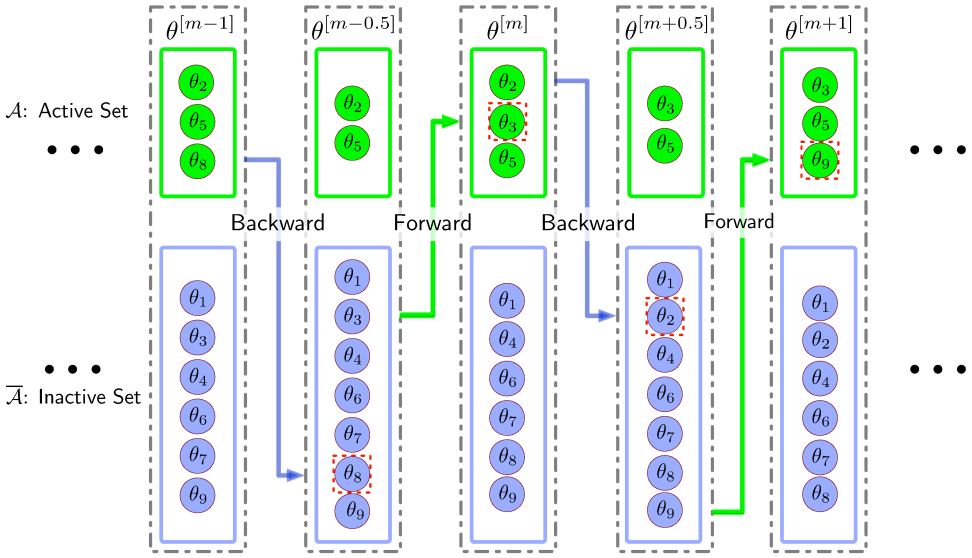


Figure 2.4: An illustration of the active set updating algorithm. The green and blue circles denote the active and inactive coordinates respectively. Suppose we have 9 coordinates, and the maximum number of active coordinates we can tolerate is 4. The active set updating iteration first removes some active coordinates from the active set, then add some inactive coordinates into the active set. Thus, the number of active coordinates is ensured to never exceed 4 throughout all iterations. To the best of our knowledge, such a "forward-backward" phenomenon has not been discovered and rigorously characterized in existing literature.

(II) Besides decreasing the objective value, the active coordinate minimization algorithm can remove some irrelevant coordinates from the active set. Thus, in conjunction with the greedy selection rule, the solution sparsity is ensured throughout all iterations. An illustration is provided in Figure 2.4. To the best of our knowledge, such a "forward-backward" phenomenon has not been discovered and rigorously characterized in existing literature.

(III) The strong rule for coordinate preselection in PICASSO put some coordinates with zero values to the active set, only when their corresponding coordinate gradients have sufficiently large magnitudes. Thus, it prevents the overselection of

irrelevant coordinates and ensure the solution sparsity.

Next, we proceed with the convergence analysis for the outer loop of PICASSO. As has been shown in Theorem 2.3.10, each middle loop of PICASSO requires a proper initialization. Since θ^* and $\mathcal{S}$ are unknown in practice, it is difficult to manually pick such an initial solution. The next theorem shows that the warm start initialization guides PICASSO to attain such a proper initialization for every middle loop without any prior knowledge.

Lemma 2.3.11. [Outer Loop] Recall $\triangle_{\lambda_K}$ and $\mathcal{K}_{\lambda_K}(\theta)$ are defined in (2.3.3) and (2.3.4) respectively. Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. If θ satisfies $\|\theta_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$ and $\mathcal{K}_{\lambda_{K-1}}(\theta) \leq \delta_{K-1}\lambda_{K-1}$, then we have

$$\|\widehat{\Delta}\|_1 \leq 11\|\widehat{\Delta}_{\mathcal{S}}\|_1 \leq 11\sqrt{s^*}\|\widehat{\Delta}\|_2,\ \mathcal{K}_{\lambda_K}(\theta) \leq \frac{\lambda_K}{4},\ \mathcal{F}_{\lambda_K}(\theta) \leq \mathcal{F}_{\lambda_K}(\theta^*) + \triangle_{\lambda_K}.$$

The warm start initialization starts with an all zero local optimum and a sufficiently large λ_0, which naturally satisfy all requirements

$$\|0_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s} \quad \text{and} \quad \mathcal{K}_{\lambda_0}(0) = 0.$$

Thus, $\theta^{[0]} = 0$ is a proper initial solution for λ_1. Then combining Theorems 2.3.10 and 2.3.11, we show by induction that the output solution of each middle loop is always a proper initial solution for the next middle loop. The warm start initial-

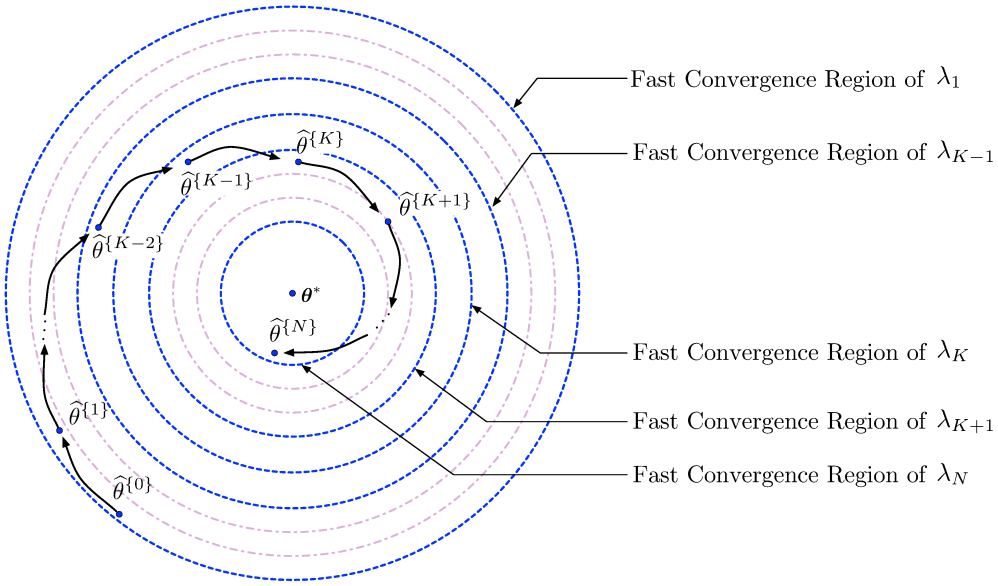


Figure 2.5: An illustration of the warm start initialization (the outer loop). From an intuitive geometric perspective, the warm start initialization yields a sequence of nested fast convergence regions. We start with large regularization parameters. This suppresses the overselection of irrelevant coordinates $\{j \mid \theta_j^* = 0\}$ and yields highly sparse solutions. With the decrease of the regularization parameter, PICASSO gradually recovers the relevant coordinates, and eventually obtains a sparse estimator $\widehat{\theta}^{\{N\}}$ with optimal statistical properties in both parameter estimation and support recovery.

ization is illustrated in Figure 2.5.

Combining Theorems 2.3.9 and 2.3.10 with Lemma 2.3.11, we establish the global convergence in terms of the objective value for PICASSO.

Theorem 2.3.12. [Main Theorem] Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. Recall $\alpha = 1/\gamma$ and γ is the concavity parameter defined in (2.1.2), δ_K's and τ_K's are the convergence parameters for the middle and inner loops within the K-th iteration of the outer loop, and κ and $\widetilde{s}$ are defined in Assumption 2.3.5. For $K = 1, \cdots, N$, we have:

(I) At the K-th iteration of the outer loop, the number of exact coordinate minimization iterations within each inner loop is at most

$$\left(s^* + 2\widetilde{s} + \frac{(s^* + 2\widetilde{s})^2 \rho_+^2 (s^* + 2\widetilde{s})}{\widetilde{\rho}_-(s^* + 2\widetilde{s})\widetilde{\rho}_-(1)}\right) \cdot \log\left(\frac{50 s^*}{\widetilde{\rho}_-(1)\tau_K^2 \widetilde{\rho}_-(s^* + \widetilde{s})}\right);$$

(II) At the K-th iteration of the outer loop, the number of active set updating iterations is at most

$$\frac{(s^* + 2\widetilde{s})\rho_+(1)}{\widetilde{\rho}_-(s^* + 2\widetilde{s})} \cdot \log\left(\frac{75 s^* \rho_+(1)}{\delta_K^2 \widetilde{\rho}_-(s^* + \widetilde{s})}\right);$$

(III) At the K-th iteration of the outer loop, we have

$$\mathcal{F}_{\lambda_N}(\widehat{\theta}^{[K]}) - \mathcal{F}_{\lambda_N}(\overline{\theta}^{\lambda_N}) \leq \left[\mathbb{1}_{\{K<N\}} + \mathbb{1}_{\{K=N\}} \cdot \delta_N\right] \frac{50 \lambda_K^2 s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})}.$$

Theorem 2.3.12 guarantees that PICASSO attains a linear convergence to a unique sparse local optimum, which is a significant improvement over sublinear convergence of the randomized coordinate minimization algorithms established in existing literature. To the best of our knowledge, this is the first result establishing the convergence properties of the pathwise coordinate optimization framework in high dimensions.

2.3.2 Statistical Theory

Finally, we analyze the statistical properties of the estimator obtained by PICASSO for sparse linear regression. We assume $\|\theta^*\|_0 \leq s^*$, and for any $v \neq 0$, the design matrix X satisfies

$$\psi_\ell \|v\|_2^2 - \gamma_\ell \cdot \frac{\log d}{n}\|v\|_1^2 \leq \frac{\|Xv\|_2^2}{n} \leq \psi_u \|v\|_2^2 + \gamma_u \cdot \frac{\log d}{n}\|v\|_1^2, \tag{2.3.5}$$

where γ_ℓ, γ_u, ψ_ℓ, and ψ_u are positive constants, and do not scale with (s^*, n, d). Existing literature has shown that (2.3.5) is satisfied by many common examples of sub-Gaussian random design with high probability [34, 33].

We then verify Assumptions 2.3.1 and 2.3.5 by the following lemma.

Lemma 2.3.13. Suppose $\epsilon \sim N(0, \sigma^2 I)$ and (2.3.5) holds. Given $\lambda_N = 8\sigma\sqrt{\log d/n}$, we have

$$\mathbb{P}\left(\lambda_N \geq 4\|\nabla \mathcal{L}(\theta^*)\|_\infty = \frac{4}{n}\|X^\top \epsilon\|_\infty\right) \geq 1 - 2d^{-2}.$$

Moreover, given $\|\frac{1}{n}X^\top X\|_1 = \mathcal{O}(d)$, $\|\theta^*\|_\infty = \mathcal{O}(d)$, $\gamma \geq 4/\psi_\ell$, and large enough n, there exists a generic constant C_1 such that we have $N = \mathcal{O}_P(\log d)$,

$$\widetilde{s} = C_1 s^* \geq [484\kappa^2 + 100\kappa] \cdot s^*, \ \widetilde{\rho}_-(s^* + 2\widetilde{s}) \geq \frac{\psi_\ell}{4}, \text{ and } \rho_+(s^* + 2\widetilde{s}) \leq \frac{5\psi_u}{4}.$$

Lemma 2.3.13 guarantees that the regularization sequence satisfies Assumption 2.3.1 with high probability, and Assumption 2.3.5 holds when the design ma-

trix satisfies (2.3.5). Thus, by Theorem 2.3.12, we know that with high probability, PICASSO attains a linear convergence to a unique sparse local optimum for sparse linear regression. Moreover, Lemma 2.3.13 also implies that the number of regularization parameters only needs to be the order of $\log d$. Thus, solving the optimization problem with a sequence of regularization parameters does not require much additional efforts.

We then characterize the statistical rate of convergence in parameter estimation for the estimator obtained by PICASSO.

Theorem 2.3.14 (Parameter Estimation)**.** Suppose $\epsilon \sim N(0, \sigma^2 I)$ and (2.3.5) holds. Given $\gamma \geq 4/\psi_\ell$ and $\lambda_N = 8\sigma\sqrt{\log d/n}$, for small enough δ_N and large enough n such that $n \geq C_2 s^* \log d$ for a generic constant C_2, we have

$$\|\widehat{\theta}^{\{N\}} - \theta^*\|_2 = \mathcal{O}_P\left(\sigma\sqrt{\frac{s_1^*}{n}} + \sigma\sqrt{\frac{s_2^* \log d}{n}}\right),$$

where $s_1^* = |\{j \mid |\theta_j^*| \geq \gamma\lambda_N\}|$ and $s_2^* = |\{j \mid 0 < |\theta_j^*| < \gamma\lambda_N\}|$.

By dividing all nonzero θ_j^*'s into strong signals and weak signals by their magnitudes, Theorem 2.3.14 shows that the MCP regularizer reduces the estimation error for strong signal with magnitudes larger than $\gamma\lambda_N$, and therefore attains a faster statistical rate of convergence than Lasso.

Remark 2.3.15 (Parameter Estimation for Lasso)**.** Theorem 2.3.14 is also applicable to Lasso with $\gamma = \infty$. As a result, all nonzero θ_j^*'s are considered as weak signals

$|\theta_j^*| < \infty$ for all $j = 1,..,d$, i.e., $s_1^* = 0$ and $s_2^* = s^*$. Theorem 2.3.14 only guarantees a slower statistical rate of convergence for Lasso,

$$\|\widehat{\theta}^{\{N\}} - \theta^*\|_2 = \mathcal{O}_P\left(\sigma\sqrt{\frac{s_2^* \log d}{n}}\right) = \mathcal{O}_P\left(\sigma\sqrt{\frac{s^* \log d}{n}}\right) \quad \text{for} \quad \gamma = \infty.$$

We then proceed to show that the statistical rate of convergence in Theorem 2.3.14 is minimax optimal in parameter estimation for a suitably chosen $\gamma < \infty$. Particularly, we consider a class of sparse vectors:

$$\Theta(s_1^*, s_2^*, d) = \Big\{\theta^* \;\Big|\; \theta^* \in \mathbb{R}^d,\ \textstyle\sum_{j=1}^d \mathbb{1}_{\{|\theta_j^*| \geq \theta_{\min}\}} \leq s_1^*, \quad \sum_{j=1}^d \mathbb{1}_{\{0<|\theta_j^*|<\theta_{\min}\}} \leq s_2^*\Big\}. \tag{2.3.6}$$

where $\theta_{\min} = \frac{8\gamma\sigma}{\sqrt{C_2(s_1^*+s_2^*)}}$ is the threshold between strong and week signals for some generic constant C_2 and $\gamma < \infty$. Given $s^* = s_1^* + s_2^*$ and $n \geq C_2 s^* \log d$, we have

$$\theta_{\min} = \frac{8\gamma\sigma}{\sqrt{C_2(s_1^* + s_2^*)}} \geq 8\gamma\sigma\sqrt{\frac{\log d}{n}} = \gamma\lambda_N,$$

which matches the threshold for dividing signals in Theorem 2.3.14. The next theorem establishes a lower bound for parameter estimation.

Theorem 2.3.16 (Lower Bound)**.** Let $\widehat{\theta}$ denote any estimator of θ^* based on $y \sim N(X\theta^*, \sigma^2 I)$, where $\theta^* \in \Theta(s_1^*, s_2^*, d)$. Then there exists a generic constant C_4 such

that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta(s_1^*, s_2^*, d)} \mathbb{E}\|\widehat{\theta} - \theta^*\|_2 \geq C_4 \left(\sigma \sqrt{\frac{s_1^*}{n}} + \sigma \sqrt{\frac{s_2^* \log d}{n}} \right).$$

Theorem 2.3.16 guarantees that the estimator obtained by PICASSO attains the minimax optimal rates of convergence over $\Theta(s_1^*, s_2^*, d)$. The convex ℓ_1 regularizer, however, only attains a suboptimal statistical rate of convergence due to the universal estimation bias regardless the signal strength. See more details in [29, 30].

To analyze the support recovery performance for the estimator obtained by PICASSO, we define the oracle least square estimator $\widehat{\theta}^{\mathrm{o}}$ as

$$\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}} = \operatorname*{argmin}_{\theta_{\mathcal{S}}} \frac{1}{2n} \|y - X_{*\mathcal{S}} \theta_{\mathcal{S}}\|_2^2 \quad \text{and} \quad \widehat{\theta}^{\mathrm{o}}_{\overline{\mathcal{S}}} = 0, \tag{2.3.7}$$

where $\mathcal{S}$ and $\overline{\mathcal{S}}$ are defined in (2.3.3). Recall $\overline{\theta}^{\lambda_N}$ is the unique sparse local minimizer to (2.1.1) with λ_N. The following theorem shows that $\overline{\theta}^{\lambda_N}$ is identical to the oracle least square estimator $\widehat{\theta}^{\mathrm{o}}$ with high probability.

Theorem 2.3.17 (Support Recovery)**.** Suppose (2.3.5) holds,

$$\epsilon \sim N(0, \sigma^2 I), \quad \text{and} \quad \min_{j \in \mathcal{S}} |\theta_j^*| \geq C_5 \gamma \sigma \sqrt{\frac{\log d}{n}} \tag{2.3.8}$$

for a generic constant C_5. Given $4/\psi_\ell \leq \gamma < \infty$ and $\lambda_N = 8\sigma\sqrt{\log d/n}$, for large enough n, there exits a generic constant C_3 such that $\mathbb{P}(\overline{\theta}^{\lambda_N} = \widehat{\theta}^{\mathrm{o}}) \geq 1 - 4d^{-2}$. Mean-

while, with probability at least $1 - 4d^{-2}$, we also have

$$\|\widehat{\theta}^{\{N\}} - \theta^*\|_2 \leq C_3\sigma\sqrt{\frac{s^*}{n}} \quad \text{and} \quad \text{supp}(\widehat{\theta}^{\{N\}}) = \text{supp}(\theta^*).$$

Theorem 2.3.17 guarantees that PICASSO converges to $\widehat{\theta}^{\text{o}}$ with high probability, which is often referred to the oracle property in existing literature [19]. Besides, we also guarantee that the estimator $\widehat{\theta}^{\{N\}}$ obtained by PICASSO is nearly unbiased and correctly identifies the true support with high probability. Although the ℓ_1 regularizer can be viewed as a special case of MCP, such an oracle property does not hold Lasso. This is because we require $\gamma < \infty$ such that the estimation bias can be eliminated for strong signals. Thus Lasso cannot guarantee the correct support recovery (unless the design matrix satisfies a restrictive irrepresentable condition–see more details in [21, 22]). We present a illustration of Theorems 2.3.14 and 2.3.17 in Figure 2.6.

Remark 2.3.18. There are several differences between [35] and our theory: (I) [35] is only applicable to global optima or some local optima. But they do not provide any algorithm, which can guarantee these optima. (II) Our theory is specifically developed for the estimator obtained by PICASSO, which is an output solution in a finite number of iterations. (III) [35] only analyze sparse linear regression using the least square loss function, but our theory is also applicable to general loss functions, as will be shown in the next section.

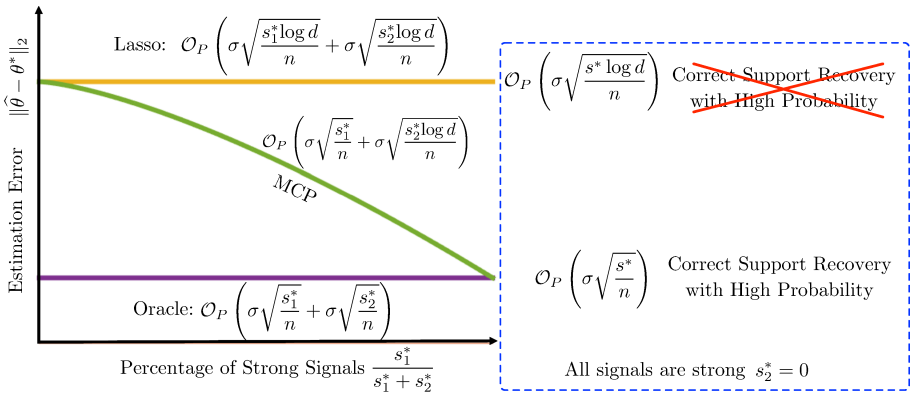


Figure 2.6: An illustration of the statistical rates of convergence in parameter estimation and support recovery for the Lasso, MCP, and oracle estimators. Recall s_1^* and s_2^* are defined in (2.3.6), and $s^* = s_1^* + s_2^*$. When all the signals are weak ($s_1^* = 0, s^* = s_2^*$), both the Lasso and MCP estimators attain the same estimation error bound $\mathcal{O}_P(\sigma\sqrt{s^*\log d/n})$. When some signals are strong, the MCP-regularized estimator attains a better estimation error bound $\mathcal{O}_P(\sigma\sqrt{s_1^*/n} + \sigma\sqrt{s_2^*\log d/n})$ than Lasso, because it reduces the estimation bias for the strong signals. Eventually, when all the signals are strong ($s_2^* = 0, s^* = s_2^*$), the MCP estimator attains the same estimation error bound as the oracle estimator $\mathcal{O}_P(\sigma\sqrt{s^*/n})$.

2.4 Extension to General Loss Functions

PICASSO can be further extended to other regularized M-estimation problems. Taking sparse logistic regression as an example, we denote the binary response vector by $y = (y_1, ..., y_n)^\top \in \mathbb{R}^n$, and the design matrix by $X \in \mathbb{R}^{n\times d}$. We consider a logistic model with $\mathbb{P}(y_i = 1) = \pi_i(\theta^*)$ and $\mathbb{P}(y_i = -1) = 1 - \pi_i(\theta^*)$, where

$$\pi_i(\theta) = \frac{1}{1 + e^{-X_{i*}^\top \theta}} \text{ for } i = 1, ..., n. \tag{2.4.1}$$

When θ^* is sparse, we consider the optimization problem

$$\min_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta)+\mathcal{R}_\lambda(\theta), \quad \text{where } \mathcal{L}(\theta)=\frac{1}{n}\sum_{i=1}^{n}\Big[\log\Big(1+e^{-y_i X_{i*}^\top\theta}\Big)\Big]. \tag{2.4.2}$$

For notational simplicity, we denote the logistic loss function in (2.4.2) as $\mathcal{L}(\theta)$, and define $\widetilde{\mathcal{L}}_\lambda(\theta)=\mathcal{L}(\theta)+\mathcal{H}_\lambda(\theta)$. Then similar to sparse linear regression, we write $\mathcal{F}_\lambda(\theta)$ as

$$\mathcal{F}_\lambda(\theta)=\mathcal{L}(\theta)+\mathcal{R}_\lambda(\theta)=\widetilde{\mathcal{L}}_\lambda(\theta)+\lambda\|\theta\|_1.$$

The logistic loss function is twice differentiable with

$$\nabla\mathcal{L}(\theta)=\frac{1}{n}\sum_{i=1}^{n}[1-\pi_i(\theta)]y_i X_{i*} \quad \text{and} \quad \nabla^2\mathcal{L}(\theta)=\frac{1}{n}X^\top PX,$$

where $P=\mathrm{diag}([1-\pi_1(\theta)]\pi_1(\theta),...,[1-\pi_n(\theta)]\pi_n(\theta))\in\mathbb{R}^{n\times n}$. Similar to sparse linear regression, we also assume that the design matrix X satisfies the column normalization condition $\|X_{*j}\|_2=\sqrt{n}$ for all $j=1,...,d$.

2.4.1 Proximal Coordinate Gradient Descent

For sparse logistic regression, directly taking the minimum with respect to a selected coordinate does not admit a closed form solution, and therefore may involve some sophisticated algorithm such as the root-finding method.

To address this issue, [25] suggest a more convenient approach, which takes a proximal coordinate gradient descent iteration. For example, we select a coordinate j at the t-th iteration and consider a quadratic approximation of $\mathcal{F}_\lambda(\theta_j; \theta_{\setminus j}^{(t)})$,

$$\mathcal{Q}_{\lambda,j,L}(\theta_j; \theta^{(t)}) = \mathcal{V}_{\lambda,j,L}(\theta_j; \theta^{(t)}) + \lambda|\theta_j| + \lambda\|\theta_{\setminus j}^{(t)}\|_1,$$

where $L > 0$ is a step size parameter, and $\mathcal{V}_{\lambda,j,L}(\theta_j; \theta^{(t)})$ is defined as

$$\mathcal{V}_{\lambda,j,L}(\theta_j; \theta^{(t)}) = \widetilde{\mathcal{L}}_\lambda(\theta^{(t)}) + (\theta_j - \theta_j^{(t)})\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{(t)}) + \frac{L}{2}(\theta_j - \theta_j^{(t)})^2.$$

Here we choose the step size parameter L such that $\mathcal{Q}_{\lambda,j,L}(\theta_j; \theta^{(t)}) \geq \mathcal{F}_\lambda(\theta_j, \theta_{\setminus j}^{(t)})$ for all $j = 1, ...d$. We then take

$$\theta_j^{(t+1)} = \underset{\theta_j}{\operatorname{argmin}}\, \mathcal{Q}_{\lambda,j,L}(\theta_j; \theta^{(t)}) = \underset{\theta_j}{\operatorname{argmin}}\, \mathcal{V}_{\lambda,j,L}(\theta_j; \theta^{(t)}) + \lambda|\theta_j|. \tag{2.4.3}$$

Different from the exact coordinate minimization, (2.4.3) always has a closed form solution obtained by soft thresholding. Particularly, we define $\widetilde{\theta}_j^{(t)} = \theta_j^{(t)} - \nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{(t)})/L$. Then we have

$$\theta_j^{(t+1)} = \underset{\theta_j}{\operatorname{argmin}}\, \frac{1}{2}(\theta_j - \widetilde{\theta}_j^{(t)})^2 + \frac{\lambda}{L}|\theta_j| = \mathcal{S}_{\lambda/L}(\widetilde{\theta}_j^{(t)}) \quad \text{and} \quad \theta_{\setminus j}^{(t+1)} = \theta_{\setminus j}^{(t)}.$$

For notational convenience, we write $\theta_j^{(t+1)} = \mathcal{T}_{\lambda,j,L}(\theta^{(t)})$. When applying PICASSO

to solve sparse logistic regression, we only need to replace $\mathcal{T}_{\lambda,j}(\cdot)$ with $\mathcal{T}_{\lambda,j,L}(\cdot)$ in Algorithms 1-3.

Remark 2.4.1. For sparse logistic regression, we have $\nabla^2_{jj}\mathcal{L}(\theta) = \frac{1}{n}X_{*j}^\top P X_{*j}$. Since P is a diagonal matrix, and $\pi_i(\theta) \in (0,1)$ for any $\theta \in \mathbb{R}^d$, we have $\|P\|_2 = \max_i P_{ii} \in (0, 1/4]$ for all $i = 1, ..., n$. Then we have $X_{*j}^\top P X_{*j} \leq \|P\|_2 \|X_{*j}\|_2^2 \leq n/4$, where the last inequality comes from the column normalization condition of X. Thus, we choose $L = \sup_\theta \max_j \nabla^2_{jj}\mathcal{L}(\theta) = 1/4$.

We then analyze the computational and statistical properties of the estimator obtained by PICASSO for sparse logistic regression.

2.4.2 Convex Relaxation based Warm Start Initialization

We assume that $\|\theta^*\|_0 \leq s^*$, and for any $v \neq 0$ and any θ such that $\|\theta - \theta^*\|_2 \leq R$, we have

$$\psi_\ell \|v\|_2^2 - \gamma_\ell \sqrt{\frac{\log d}{n}} \|v\|_1^2 \leq v^\top \nabla^2 \mathcal{L}(\theta) v \leq \psi_u \|v\|_2^2 + \gamma_u \sqrt{\frac{\log d}{n}} \|v\|_1^2, \tag{2.4.4}$$

where γ_ℓ, γ_u, ψ_ℓ, ψ_u, and R are positive constants, and do not scale with (s^*, n, d). Existing literature has shown that many common examples of sub-Gaussian random design satisfy (2.4.4) with high probability [34, 33, 32].

Similar to sparse linear regression, we need to verify Assumptions 2.3.1 and 2.3.5 for sparse logistic regression by the following lemma.

Lemma 2.4.2. Suppose (2.4.4) holds. Given $\lambda_N = 16\sqrt{\log d/n}$, we have

$$\mathbb{P}\bigg(\lambda_N \geq 4\|\nabla\mathcal{L}(\theta^*)\|_\infty = \frac{4}{n}\|X^\top w\|_\infty\bigg) \geq 1 - d^{-7},$$

where $w = ([1-\pi_1(\theta^*)]y_1, ..., [1-\pi_n(\theta^*)]y_n)^\top$ with $\pi_i(\theta)$'s defined in (2.4.1). Moreover, given $\gamma \geq 4/\psi_\ell$ and $\|\theta - \theta^*\|_2 \leq R$, there exists some generic constant C_1 such that for large enough n, we have

$$\widetilde{s} = C_1 s^* \geq [484\kappa^2 + 100\kappa]s^*,\ \widetilde{\rho}_-(s^* + 2\widetilde{s}) \geq \frac{\psi_\ell}{2},\ \rho_+(s^* + 2\widetilde{s}) \leq \frac{5\psi_u}{4}.$$

The proof of Lemma 2.4.2 directly follows Appendix A.9 and [32], and therefore is omitted. Lemma 2.4.2 guarantees that the regularization sequence satisfies Assumption 2.3.1 with high probability, and Assumption 2.3.5 holds when the design matrix satisfies (2.4.4).

Different from sparse linear regression, however, the restricted convexity and smoothness only hold over an ℓ_2 ball centered at θ^* for sparse logistic regression. Thus, directly choosing $\widehat{\theta}^{\{0\}} = 0$ may violate the restricted strong convexity. A simple counter example is $\|\theta^*\|_2 > R$, which results in $\|0 - \theta^*\|_2 > R$. To address this issue, we propose a new convex relaxation based warm start initialization to obtain an initial solution for λ_0. Particularly, we solve the following convex relaxation of

(2.1.1):

$$\min_{\theta\in\mathbb{R}^d} \widetilde{\mathcal{F}}_{\lambda_0}(\theta), \text{ where } \widetilde{\mathcal{F}}_{\lambda_0}(\theta) = \mathcal{L}(\theta) + \lambda_0\|\theta\|_1 \tag{2.4.5}$$

up to an adequate precision. For example, we choose θ^{relax} satisfying the approximate KKT condition of (2.4.5) as follows,

$$\min_{\xi\in\partial\|\theta^{\text{relax}}\|_1} \|\nabla\mathcal{L}(\theta^{\text{relax}}) + \lambda_0\xi\|_\infty \le \delta_0\lambda_0, \tag{2.4.6}$$

where $\delta_0 \in (0,1)$ is the initial precision parameter for λ_0. Since δ_0 in (2.4.6) can be chosen as a sufficiently large value (e.g. $\delta_0 = 1/8$), computing θ^{relax} becomes very efficient even for algorithms with only sublinear rates of convergence to global optima, e.g., classical coordinate minimization and proximal gradient algorithms. For notational convenience, we call the above initialization procedure the convex relaxation based warm initialization.

Lemma 2.4.3. Suppose Assumption 2.3.5 holds only for $\|\theta-\theta^*\|_2 \le R$. Given $\rho_-(s^*+\widetilde{s})R \ge 9\lambda_0\sqrt{s^*} \ge 18\lambda_N\sqrt{s^*}$ and $\delta_0 = 1/8$, we have

$$\|\theta^{\text{relax}}_{\widetilde{\mathcal{S}}}\|_0 \le \widetilde{s}, \quad \|\theta^{\text{relax}} - \theta^*\|_2 \le R, \quad \text{and} \quad \mathcal{F}_{\lambda_0}(\theta^{\text{relax}}) \le \mathcal{F}_{\lambda_0}(\theta^*) + \triangle_{\lambda_0}.$$

Lemma 2.4.3 guarantees that θ^{relax} is a proper initial solution for λ_0. Thus, all convergence analysis in Theorem 2.3.12 directly follows, and PICASSO attains a

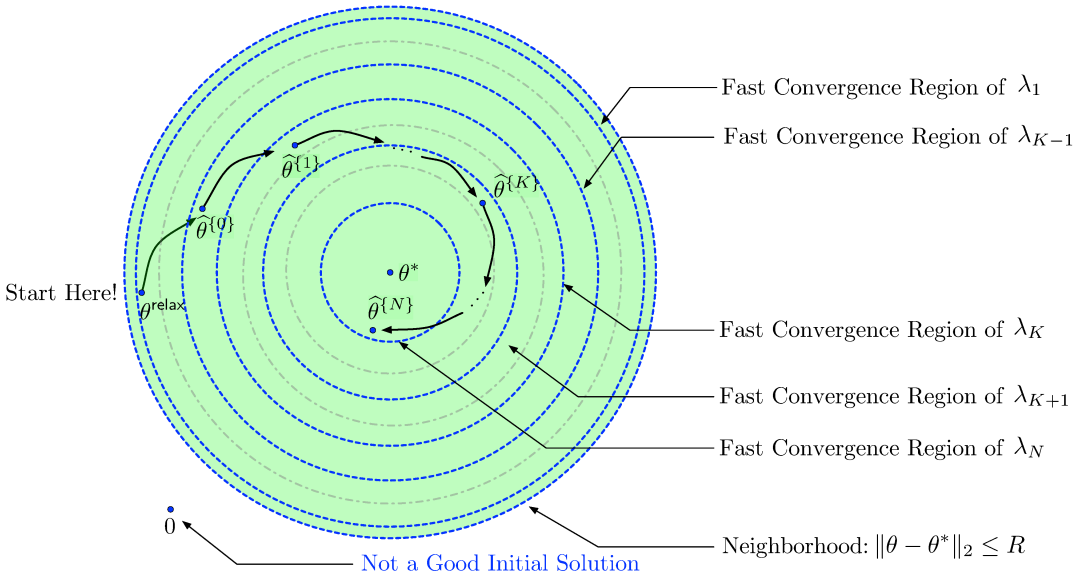


Figure 2.7: An illustration of the convex relaxation based warm start initialization. When the restricted convexity and smoothness only hold over a neighborhood around θ^* (Green Region). Directly choosing 0 as the initial solution may violate the restricted strong convexity. Thus, we adopt a convex relaxation approach to obtain an initial solution, which is ensured to be sparse and belong to the desired neighborhood.

linear convergence to a unique sparse local optimum with high probability. The statistical properties can also be established accordingly. An illustration of the convex relaxation based warm start initialization is provided in Figure 2.7.

2.5 Numerical Experiments

We evaluate the computational and statistical performance of PICASSO through numerical simulations. We compare PICASSO with five competitors: (1) SparseNet [4]; (2) Path-following Iterative Shrinkage Thresholding Algorithm (PISTA, [31]); (3) Accelerated PISTA (A-PISTA, [10]); (4) Multistage Convex Relaxation Method

(Mcvx, [36]); (5) Local Linear Approximation (LLA, [37]). Note that all subproblems of Mcvx and LLA are solved by proximal gradient algorithms with backtracking line search.

All experiments are conducted on a PC with Intel Core i5 3.3 GHz and 16GB memory. All programs are coded in double precision `C`, called from a `R` wrapper. We optimize the computation by exploiting the vector and matrix sparsity, which gains a significant speedup in vector and matrix manipulations (e.g. computing the gradient and evaluating the objective value). We apply PICASSO to sparse linear regression with the MCP regularizer.

Simulated Data: We generate each row of the design matrix X_{i*} independently from a d-dimensional Gaussian distribution with mean 0 and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, where $\Sigma_{kj} = 0.75$ and $\Sigma_{kk} = 1$ for all $j,k = 1,...,d$ and $k \neq j$. We then normalize each column of the design matrix X_{*j} such that $\|X_{*j}\|_2 = \sqrt{n}$. The response vector is generated from the linear model $y = X\theta^* + \epsilon$, where $\theta^* \in \mathbb{R}^d$ is the regression coefficient vector, and ϵ is generated from a n-dimensional Gaussian distribution with mean 0 and covariance matrix $\sigma^2 I$. We set $n = 300$, $d = 18000$, $s^* = 18$, and $\sigma^2 = 4$. θ^* has 18 nonzero entries, which are $\theta^*_{1000} = \theta^*_{7000} = \theta^*_{13000} = 3$, $\theta^*_{2000} = \theta^*_{8000} = \theta^*_{14000} = 2$, $\theta^*_{3000} = \theta^*_{9000} = \theta^*_{15000} = 1.5$, $\theta^*_{4000} = \theta^*_{10000} = \theta^*_{16000} = -3$, $\theta^*_{5000} = \theta^*_{11000} = \theta^*_{17000} = -2$, and $\theta_{6000} = \theta^*_{12000} = \theta_{18000} = -1.5$ for $k = 0,...,2$. We then set $\gamma = 1.25$, $N = 70$, $\lambda_N = 0.25\sigma\sqrt{\log d/n}$, $\varphi = 0.05$, $\delta_K = 10^{-3}$, and $\tau_K = 10^{-6}$ for all $1 \leq K \leq N$.

We present the numerical results averaged over 1000 simulations. Specifically, we create a validation set using the same design matrix as the training set for regularization parameter selection. We then tune the regularization parameter over the selected regularization sequence. We denote the response vector of the validation set as $\widetilde{y} \in \mathbb{R}^n$. Let $\widehat{\theta}^{\lambda}$ denote the obtained estimator using the regularization parameter λ. We then choose the optimal regularization parameter $\widehat{\lambda}$ by

$$\widehat{\lambda} = \text{argmin}_{\lambda \in \{\lambda_1, \ldots, \lambda_N\}} \|\widetilde{y} - X\widehat{\theta}^{\lambda}\|_2^2.$$

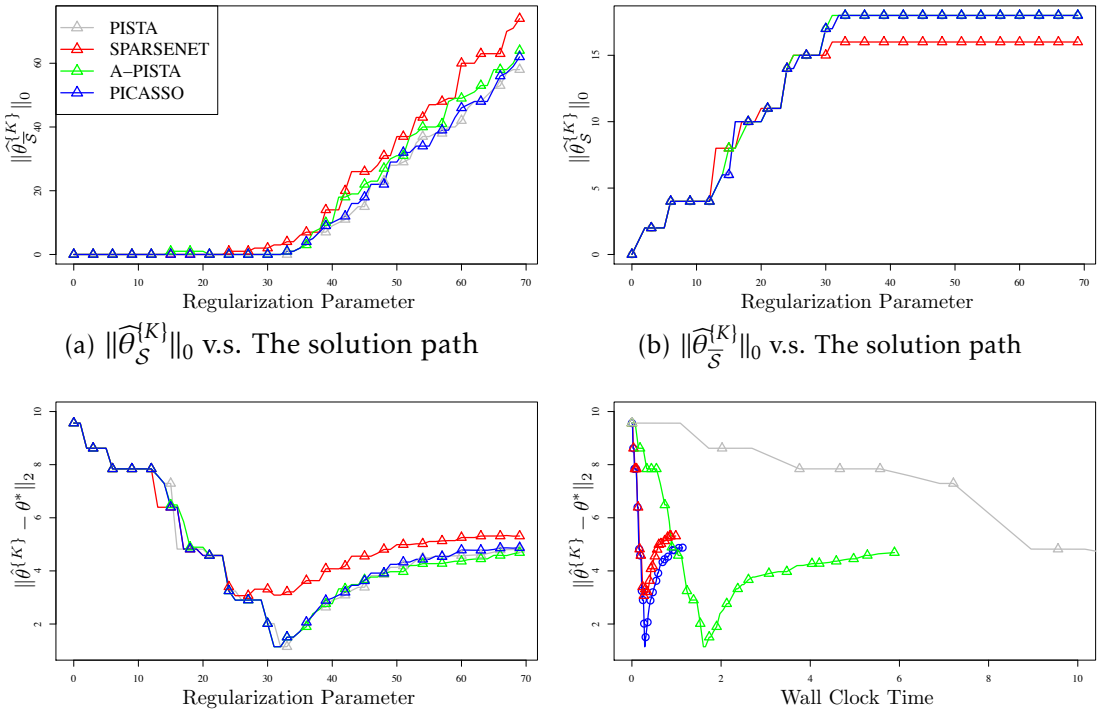
We repeat the simulation for 1000 times and summarize the averaged results in Table 1. In terms of timing performance, PICASSO slightly outperforms SparseNet, outperforms A-PISTA, and greatly outperforms PISTA, LLA, and Mcvx respectively. In terms of support recovery and parameter estimation, PICASSO slightly outperforms A-PISTA, PISTA, and Mcvx, and greatly outperforms SparseNet and LLA.

To further demonstrate the superiority of PICASSO, we present a typical failure example of SparseNet using the heuristic cyclic selection rule. This example is chosen from our 1000 simulations, and illustrated in Figure 2.8. We see that the heuristic cyclic selection rule in SparseNet always needs to iterate over many irrelevant variables before getting to the relevant variable when identifying a new active set. Since these irrelevant variables are highly correlated with

Table 2.1: Quantitative comparison on the simulated data set ($n = 300$, $d = 18000$, $s^* = 18$, $\sigma^2 = 4$). In terms of timing performance, PICASSO slightly outperforms SparseNet, outperforms A-PISTA, and greatly outperforms PISTA, LLA, and Mcvx respectively. In terms of support recovery and parameter estimation, PICASSO slightly outperforms A-PISTA, PISTA, and Mcvx, and greatly outperforms SparseNet and LLA.

Method	$\|\widehat{\theta} - \theta^*\|_2$	$\|\widehat{\theta}_{\mathcal{S}}\|_0$	$\|\widehat{\theta}_{\mathcal{S}^c}\|_0$	Correct	Timing
PICASSO	1.258(0.515)	17.79(0.54)	0.48(0.52)	616/1000	1.062(0.084)
SparseNet	1.602(0.791)	17.64(0.85)	2.07(1.41)	248/1000	1.109(0.088)
PISTA	1.267(0.528)	17.76(0.54)	0.55(0.51)	614/1000	52.358(5.920)
A-PISTA	1.276(0.530)	17.76(0.54)	0.57(0.57)	613/1000	6.358(0.865)
Mcvx	1.293(0.529)	17.76(0.52)	0.58(0.52)	615/1000	67.247(7.128)
LLA	1.517(0.949)	17.50(0.61)	1.28(0.85)	365/1000	31.247(3.870)

the relevant variables in our experiment, the heuristic cyclic selection rule tends to overselect the irrelevant variables and miss some relevant variables. In contrast, PICASSO, PISTA, and A-PISTA have mechanisms to prevent overselecting irrelevant variables when identifying active sets. This eventually makes them outperform SparseNet in both parameter estimation and support recovery. Moreover, we also see that PISTA is much slower than other algorithms, because PISTA needs to calculate a full gradient and conduct a sophisticated line search in every iteration, which are computationally expensive. Though A-PISTA adopts the coordinate minimization to further accelerate PISTA, it still suffers from the computationally expensive line search when identifying active sets. This eventually leads to less competitive timing performance than PICASSO.



(a) $\|\widehat{\theta}^{\{K\}}_{\mathcal{S}}\|_0$ v.s. The solution path

(b) $\|\widehat{\theta}^{\{K\}}_{\overline{\mathcal{S}}}\|_0$ v.s. The solution path

(c) $\|\widehat{\theta}^{\{K\}} - \theta^*\|_2$ v.s. The solution path

(d) $\|\widehat{\theta}^{\{K\}} - \theta^*\|_2$ v.s. The wall clock time

Figure 2.8: A typical failure example of SparseNet using the heuristic cyclic selection rule, which is chosen from our 1000 simulations. We see that cyclic selection rule tends to overselect the irrelevant coordinate and miss some relevant coordinates when updating the active set. Thus SparseNet eventually yields denser solutions with worse performance in parameter estimation and support recovery than PICASSO, PISTA, and A-PISTA.

Real Data: We adopt the gene expression data set in [38]. The original data set contains 31,042 gene expression values of 120 rats. Our goal is to identify genes with expression values related to that of gene TRIM32, which is known to be associated with human diseases of the retina (corresponding to Probe 1389163_at). Following the same preprocessing procedure as [39] and [40], we remove genes lacking sufficient variation or expression, and then choose 4,000 genes with the largest sample variances in expression values.

Table 2.2: Quantitative comparison on the real data example. PICASSO attains better prediction error and smaller average model sizes than those of other competing algorithms. Moreover, PICASSO attains much better timing performance than PISTA, Mcvx, and LLA.

Method	Average model size	Prediction Error	Timing
PICASSO	12.35(5.33)	0.2789(0.0705)	0.759(0.278)
SparseNet	14.71(5.86)	0.2922(0.0854)	0.901(0.606)
PISTA	12.99(5.56)	0.2797(0.0803)	31.511(2.041)
A-PISTA	12.85(5.56)	0.2796(0.0803)	5.729(2.741)
Mcvx	14.15(3.61)	0.2825(0.0822)	36.672(4.464)
LLA	14.30(3.66)	0.2844(0.0861)	24.250(3.105)

We set $\gamma = 1.05$, $N = 70$, $\lambda_N = 0.01\lambda_0$, $\delta_K = 10^{-3}$, and $\tau_K = 10^{-6}$ for all $1 \leq K \leq N$. We randomly split the 120 rats into a training set of 90 rats for fitting the model, a validation set of 15 rats for tuning parameter selection, and a testing set of 15 rats for evaluating the prediction performance. The optimal tuning parameter is selected based on minimizing the prediction error on the validation set. Table 2 summarizes the numerical results averaged over 100 random splits. We see that PICASSO attains better prediction error and smaller average model sizes than those of the other competing algorithms. Moreover, PICASSO attains much better timing performance than PISTA and Mcvx. Besides, PICASSO identifies a few genes, which are not identified by Lasso, SparseNet, and LLA. These identified genes may be worth further investigation in genomic studies.

2.6 Discussions and Future Work

Here we discuss several existing methods related to PICASSO, including the multistage convex relaxation method (Mcvx), local linear approximation method (LLA), path-following iterative shrinkage thresholding algorithm (PISTA), accelerated path-followi-ng iterative shrinkage thresholding algorithm (A-PISTA), and proximal gradient algorithm.

The multistage convex relaxation method is proposed in [36]. It solves a sequence of convex relaxation problems of (2.1.1). [36] show that the obtained estimator enjoys similar statistical guarantees to those of PICASSO for sparse linear regression. However, there is online sublinear guarantee on its convergence rate to a local optimum. Moreover, since each relaxed problem is still lack of strong convexity, the multistage convex relaxation method needs to be combined with some efficient computational algorithms such as PICASSO.

The local linear approximation method is proposed in [37, 40, 41]. It is essentially a special case of the multistage convex relaxation with only two iterations. Similar to the multistage convex relaxation method, it also needs an efficient computational algorithm to solve each relaxed problem. Moreover, in order to obtain the variable selection consistency, the local linear approximation method requires a stronger minimum signal strength. Taking sparse linear regression as an example, [40, 41] requires a minimum signal strength of order of $\sigma\sqrt{s^* \log d/n}$, while PICASSO only requires a minimum signal strength of order of $\sigma\sqrt{\log d/n}$.

The path-following iterative shrinkage thresholding algorithm (PISTA) is proposed in [31]. PISTA is essentially a proximal gradient algorithm combined with the warm start initialization. PISTA needs to calculate the entire (d-dimensional) gradient vector and requires a sophisticated backtracking line search procedure in every iteration. Thus, PICASSO is computationally much more efficient than PISTA in practice, although PISTA and PICASSO enjoy similar theoretical guarantees. Besides, the implementation of PISTA requires subtle control over the step size, and often yield slow empirical convergence. An accelerated PISTA algorithm (A-PISTA) is proposed in [10], which uses coordinate minimization algorithms to accelerated PISTA. It shows an improved computational performance over PISTA in our numerical simulations, but not as competitive as PICASSO.

Moreover, when extending PISTA to general loss functions, [31] propose a contained formulation. Particularly, they solve (2.1.1) with an additional constraint

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta) \quad \text{subject to } \|\theta\|_2 \leq R/2. \tag{2.6.1}$$

The additional constraint guarantees that the solution always stays in the restricted strongly convex region (a small neighborhood around θ^*), only under the assumption $\|\theta^*\|_2 \leq R/2$, where R is a constant and cannot scale with (n, s^*, d). This assumption is very restrictive, and also introduces an additional tuning parameter. In contrast, our proposed convex relaxation based warm start initialization avoids

this assumption, and allows $\|\theta^*\|_2$ to be arbitrarily large. Furthermore, we want to emphasize that PISTA exploits an explicit soft-thresholding procedure to directly control the solution sparsity in each iteration, while PICASSO adopts an algorithmic strategy to control the sparsity of the active set.

Other researchers focus on solving (2.1.1) with an additional constraint,

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta) \quad \text{subject to } \|\theta\|_1 \leq M, \tag{2.6.2}$$

where $M > 0$ is an extra tuning parameter. [32] show that the proximal gradient algorithm attains a linear convergence to a ball centered at θ^* to (2.6.2) with a radius approximately equal to the statistical error. However, the analysis of [32] does not justify the advantage of nonconvex regularization: They only provides a slower statistical rate of convergence than PICASSO in parameter estimation for their obtained estimator, and no support recovery guarantee is established. Besides, their analysis for general loss functions also requires the restrictive assumption: $\|\theta^*\|_2 \leq R/2$, where R is a constant and does not scale with (n, s^*, d). Nevertheless, PICASSO does not require this assumption.

For future work, we are interested in possible extensions: (I) Extension to more complicated regularizers such as grouping regularizers for variable clustering; (II) Extension to more complicated (possibly nonconvex) loss functions such as sparse phase retrieval and sparse coding problems; (III) Extension to asynchronous par-

allel optimization setting with shared memory or communication-efficient distributed optimization setting; (IV) Extension to second order algorithms such as the regularized iterative reweighed least square optimization algorithm for sparse generalized linear model estimation (proximal Newton). These extensions will lead to more efficient and scalable coordinate optimization algorithms for more sophisticated nonconvex optimization problems.

2.7 Proof of Main Results

We present the proof sketch of our computational and statistical theories. Some lemmas are deferred to Section A. To unify the convergence analysis of PICASSO using the exact coordinate minimization (2.1.6) and proximal coordinate gradient descent (2.4.3), we define two auxiliary parameters $\nu_+(1)$ and $\nu_-(1)$. Specifically, we choose $\nu_+(1) = \nu_-(1) = L$ for the proximal coordinate gradient descent, and $\nu_+(1) = \rho_+(1)$ and $\nu_-(1) = \widetilde{\rho}_-(1)$ for the exact coordinate minimization.

2.7.1 Proof of Theorem 2.3.9

Proof. Since $\|\theta^{(0)}\|_0 = s \leq s^* + 2\widetilde{s}$, by Assumption 2.3.5 and Lemma 2.3.4, we know that (2.2.2) is a strongly convex optimization problem. Thus, its minimizer $\overline{\theta}$ is unique. We then introduce the following lemmas.

Lemma 2.7.1. Suppose Assumption and 2.3.5 holds, and $|\mathcal{A}| = s \leq s^* + 2\widetilde{s}$. For $t = 0, 1, 2, \ldots$, we have $\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\theta^{(t+1)}) \geq \frac{\nu_-(1)}{2}\|\theta^{(t)} - \theta^{(t+1)}\|_2^2$.

Lemma 2.7.2. Suppose Assumption and 2.3.5 holds, and $|\mathcal{A}| = s \leq s^* + 2\widetilde{s}$. For $t = 0, 1, 2, \ldots$, we have $\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(\overline{\theta}) \leq \frac{s\rho_+^2(s)}{2\widetilde{\rho}_-(s)}\|\theta^{(t+1)} - \theta^{(t)}\|_2^2$.

Lemmas 2.7.1 and 2.7.2 characterize the successive descent and the gap towards the optimal objective value after each iteration respectively.

[Linear Convergence] Combining Lemmas 2.7.1 and 2.7.2, we obtain

$$\begin{aligned}\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(\overline{\theta}) \leq \frac{s\rho_+^2(s)}{\widetilde{\rho}_-(s)\nu_-(1)}[\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})] \\ - \frac{s\rho_+^2(s)}{\widetilde{\rho}_-(s)\nu_-(1)}[\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(\overline{\theta})].\end{aligned} \tag{2.7.1}$$

By simple manipulation, (2.7.1) implies

$$\begin{aligned}\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(\overline{\theta}) &\overset{\text{(i)}}{\leq} \left(\frac{s\rho_+^2(s)}{\widetilde{\rho}_-(s)\nu_-(1) + s\rho_+^2(s)}\right)[\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})] \\ &\overset{\text{(ii)}}{\leq} \left(\frac{s\rho_+^2(s)}{\widetilde{\rho}_-(s)\nu_-(1) + s\rho_+^2(s)}\right)^{t+1}[\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\overline{\theta})],\end{aligned} \tag{2.7.2}$$

where (ii) comes from recursively using (i).

[Number of Iterations] Combining (2.7.2) with Lemma 2.7.1, we obtain

$$\begin{aligned}\|\theta^{(t)} - \theta^{(t+1)}\|_2^2 &\overset{\text{(i)}}{\leq} \frac{2[\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})]}{\nu_-(1)} \\ &\leq \left(\frac{s\rho_+^2(s)}{\widetilde{\rho}_-(s)\nu_-(1) + s\rho_+^2(s)}\right)^t \frac{2[\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\overline{\theta})]}{\nu_-(1)},\end{aligned}$$

where (i) comes from $\mathcal{F}_\lambda(\theta^{(t)}) \geq \mathcal{F}_\lambda(\overline{\theta})$. Thus, we need at most

$$t = \log^{-1}\left(\frac{s\rho_+^2(s)}{\widetilde{\rho}_-(s)\nu_-(1) + s\rho_+^2(s)}\right)\log\left(\frac{\nu_-(1)\tau^2\lambda^2}{2[\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\overline{\theta})]}\right)$$

iterations such that

$$\|\theta^{(t+1)} - \theta^{(t)}\|_2^2 \leq \left(\frac{s\rho_+^2(s)}{\widetilde{\rho}_-(s)\nu_-(1) + s\rho_+^2(s)}\right)^t \frac{2[\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\overline{\theta})]}{\nu_-(1)} \leq \tau^2\lambda^2.$$

□

2.7.2 Proof of Theorem 2.3.10

Proof. Before the proof starts, we first introduce the following lemmas.

Lemma 2.7.3. Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. There exists a unique sparse local optimum $\overline{\theta}^\lambda$ satisfying $\|\overline{\theta}^\lambda_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$ and $\mathcal{K}_\lambda(\overline{\theta}^\lambda) = 0$.

Lemma 2.7.4. Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. If the initial solution $\theta^{(0)}$ in Algorithm 1 satisfies $\|\theta^{(0)}_{\overline{\mathcal{S}}}\|_0 \leq 2\widetilde{s}$ and $\mathcal{F}_\lambda(\theta^{(0)}) \leq \mathcal{F}_\lambda(\theta^*) + \triangle_\lambda$, the output

solution $\widehat{\theta}$ satisfies

$$\min_{\xi_{\mathcal{A}} \in \partial \|\widehat{\theta}_{\mathcal{A}}\|_1} \|\nabla_{\mathcal{A}} \widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) + \lambda \xi_{\mathcal{A}}\|_\infty \leq \delta\lambda \quad \text{and} \quad \|\widehat{\theta}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}. \tag{2.7.3}$$

Lemma 2.7.5. Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. If the initial solution $\theta^{[0]}$ satisfies $\|\theta^{[0]}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$ and $\mathcal{F}_\lambda(\theta^{[0]}) \leq \mathcal{F}_\lambda(\theta^*) + \triangle_\lambda$. Then regardless the simple rule or strong rule, we have $|\mathcal{A}_0 \cap \overline{\mathcal{S}}| \leq \widetilde{s}$.

The proof of Lemmas 2.7.3, 2.7.4, and 2.7.5 are provided in Appendices A.3.5, A.3.6, and A.3.8 respectively. Lemma 2.7.3 verifies the existence of the unique sparse local optimum. Lemma 2.7.4 implies that the inner loop of PICASSO removes irrelevant coordinates, and encourages the output solution sparsity. Lemma 2.7.5 implies that the initial active set is sufficiently sparse.

[Solution Sparsity] Since the objective always decreases, we have

$$\mathcal{F}_\lambda(\theta^{[m+1]}) \leq \mathcal{F}_\lambda(\theta^{[m+0.5]}) \leq \mathcal{F}_\lambda(\theta^{[0]}) \leq \mathcal{F}_\lambda(\theta^*) + \triangle_\lambda \tag{2.7.4}$$

for all $m = 0, 1, 2, \ldots$. Since $\theta^{[0]}$ satisfies $\|\theta^{[0]}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$, by Lemma 2.7.5, we have $|\mathcal{A}_0 \cap \overline{\mathcal{S}}| \leq \widetilde{s}$. Then by Lemma 2.7.4, we have $\|\theta^{[0.5]}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$. Moreover, the greedy selection rule moves only one inactive coordinate to the active set, and therefore guarantees $\|\theta^{[1]}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s} + 1$. By induction, we prove $\|\theta^{[m]}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s} + 1$ and $\|\theta^{[m+0.5]}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$ for all $m = 0, 1, 2, \ldots$.

[Linear Convergence] We first prove the linear convergence for the proximal coordinate gradient descent. We need to construct an auxiliary solution

$$\begin{aligned} w^{[m+1]} &= \underset{w\in\mathbb{R}^d}{\operatorname{argmin}}\, \mathcal{J}_{\lambda,L}(w;\theta^{[m+0.5]}) \\ &= \underset{w\in\mathbb{R}^d}{\operatorname{argmin}}\, \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) + (w-\theta^{[m+0.5]})^\top \nabla\widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) \\ &\qquad + \frac{L}{2}\|w-\theta^{[m+0.5]}\|_2^2 + \lambda\|w\|_1. \end{aligned}$$

We can verify $w_k^{[m+1]} = \operatorname{argmin}_{\theta_k} \mathcal{Q}_{\lambda,k,L}(\theta_k;\theta^{[m+0.5]})$ for $j = 1,...,d$. For notational simplicity, we define $w^{[m+1]} = \mathcal{T}_{\lambda,L}(\theta^{[m+0.5]})$. Before we proceed, we introduce the following lemmas.

Lemma 2.7.6. Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. For the proximal coordinate gradient descent and $m = 0, 1, 2...$, we have

$$\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\theta^{[m+1]}) \geq \frac{1}{s^* + 2\widetilde{s}}\Big[\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{J}_{\lambda,L}(w^{[m+1]};\theta^{[m+0.5]})\Big].$$

Lemma 2.7.7. Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. For the proximal coordinate gradient descent and $m = 0, 1, 2...$, we have

$$\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda) \leq \frac{L}{\widetilde{\rho}_-(s^* + 2\widetilde{s})}\Big[\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{J}_{\lambda,L}(w^{[m+1]};\theta^{[m+0.5]})\Big].$$

The proofs of Lemmas 2.7.6 and 2.7.7 are presented in Appendices A.3.9 and

A.3.12. Lemmas 2.7.6 and 2.7.7 characterize the successive descent in each iteration and the gap towards the optimal objective value after each iteration respectively. Combining Lemmas 2.7.6 and 2.7.7, we obtain

$$\begin{aligned} &\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda) \\ &\leq \frac{(s^* + 2\widetilde{s})L}{\widetilde{\rho}_-(s^* + 2\widetilde{s})}\Big([\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda)] - [\mathcal{F}_\lambda(\theta^{[m+1]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda]\Big). \end{aligned} \tag{2.7.5}$$

By simple manipulation, (2.7.5) implies

$$\begin{aligned} \mathcal{F}_\lambda(\theta^{[m+1]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda) &\leq \left(1 - \frac{\widetilde{\rho}_-(s^* + 2\widetilde{s})}{(s^2 + 2\widetilde{s})L}\right)[\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda)] \\ &\overset{\text{(i)}}{\leq} \left(1 - \frac{\widetilde{\rho}_-(s^* + 2\widetilde{s})}{(s^* + 2\widetilde{s})L}\right)[\mathcal{F}_\lambda(\theta^{[m]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda)] \\ &\overset{\text{(ii)}}{\leq} \left(1 - \frac{\widetilde{\rho}_-(s^* + 2\widetilde{s})}{(s^* + 2\widetilde{s})L}\right)^{m+1} [\mathcal{F}_\lambda(\theta^{[0]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda)], \end{aligned} \tag{2.7.6}$$

where (i) comes from (2.7.4), and (ii) comes from recursively applying (i).

For the exact coordinate minimization, at the m-th iteration, we only need to conduct a proximal coordinate gradient descent iteration with $L = \rho_+(1)$, and obtain an auxiliary solution $\widetilde{\theta}^{[m+1]}$. Since $\mathcal{F}_\lambda(\theta^{[m+1]}) \leq \mathcal{F}_\lambda(\widetilde{\theta}^{[m+1]})$, by (2.7.6), we further have

$$\mathcal{F}_\lambda(\theta^{[m+1]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda) \leq \left(1 - \frac{\widetilde{\rho}_-(s^* + 2\widetilde{s})}{(s^* + 2\widetilde{s})\rho_+(1)}\right)\Big[\mathcal{F}_\lambda(\theta^{[m]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda)\Big]. \tag{2.7.7}$$

[Number of Iterations] Before we proceed, we introduce the following lemma.

Lemma 2.7.8. Suppose Assumption 2.3.5 holds. For any θ, we conduct an exact coordinate minimization or proximal coordinate gradient descent iteration over a coordinate k, and obtain w. Then we have $\mathcal{F}_\lambda(\theta)-\mathcal{F}_\lambda(w) \geq \frac{\nu_-(1)}{2}(w_k-\theta_k)^2$. Moreover, if $\theta_k = 0$ and $|\nabla_k \mathcal{L}(\theta)| \geq (1+\delta)\lambda$, we have

$$|w_k| \geq \frac{\delta\lambda}{L} \quad \text{and} \quad \mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(w) \geq \frac{\delta^2\lambda^2}{2\nu_+(1)}.$$

Lemma 2.7.8 characterizes the sufficient descent when adding the selected inactive coordinate k into the active set. Assume that the selected coordinate k_m satisfies $|\nabla_{k_m}\mathcal{L}(\theta^{[m+0.5]})| \geq (1+\delta)\lambda$. Then by Lemma 2.7.8, we have

$$\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda) \geq \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\theta^{[m+1]}) \geq \frac{\delta^2\lambda^2}{2\nu_+(1)}. \tag{2.7.8}$$

Moreover, by (2.7.6) and (2.7.7), we need at most

$$m = \log^{-1}\left(1 - \frac{\widetilde{\rho}_-(s^*+2\widetilde{s})}{(s^*+2\widetilde{s})\nu_+(1)}\right)\log\left(\frac{\delta^2\lambda^2}{3\nu_+(1)[\mathcal{F}_\lambda(\theta^{[0]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda)]}\right)$$

iterations such that $\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda) \leq \frac{\delta^2\lambda^2}{3\nu_+(1)}$, which is contradicted by (2.7.8). Thus, we must have $\max_{k\in\overline{\mathcal{A}}_m}|\nabla_k\mathcal{L}(\theta^{[m+0.5]})| \leq (1+\delta)\lambda$, and the algorithm is terminated.

[Approximately Optimal Output Solution] By Lemma 2.7.4, we know that when

every inner loop terminates, the approximate KKT condition must hold over the active set. Since $\nabla_{\overline{\mathcal{A}}_m} \mathcal{H}_\lambda(\theta^{[m+0.5]}) = 0$, the stopping criterion $\max_{k \in \overline{\mathcal{A}}_m} |\nabla_k \mathcal{L}(\theta^{[m+0.5]})| \leq (1+\delta)\lambda$ implies that the approximate KKT condition holds over the inactive set,

$$\min_{\xi_{\overline{\mathcal{A}}_m} \in \partial \|\theta^{[m+0.5]}_{\overline{\mathcal{A}}_m}\|_1} \|\nabla_{\overline{\mathcal{A}}_m} \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) + \lambda \xi_{\overline{\mathcal{A}}_m}\|_\infty \leq \delta\lambda.$$

The above two approximate KKT conditions implies that $\theta^{[m+0.5]}$ satisfies the approximate KKT condition $\mathcal{K}_\lambda(\theta^{[m+0.5]}) \leq \delta\lambda$. □

2.7.3 Proof of Theorem 2.3.12

Proof. [Result (I)] Before we proceed, we introduce the following lemma.

Lemma 2.7.9. Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. For any $\lambda \geq \lambda_N$, if θ satisfies $\|\theta_{\overline{S}}\|_0 \leq \widetilde{s}$ and $\mathcal{K}_\lambda(\theta) \leq \delta\lambda$, where $\delta \leq 1/8$, then for any $\lambda' \in [\lambda_N, \lambda]$, we have

$$\mathcal{F}_{\lambda'}(\theta) - \mathcal{F}_{\lambda'}(\overline{\theta}^{\lambda'}) \leq \frac{40(\mathcal{K}_\lambda(\theta) + 3(\lambda - \lambda'))(\lambda + \lambda')s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})}.$$

The proof of Lemma 2.7.9 is provided in Appendix A.6. If we take $\lambda = \lambda' = \lambda_K$ and $\theta = \widehat{\theta}^{\{K-1\}}$, then Lemma 2.7.9 implies

$$\mathcal{F}_{\lambda_K}(\widehat{\theta}^{\{K-1\}}) - \mathcal{F}_{\lambda_K}(\overline{\theta}^{\lambda_K}) \leq \frac{25s^*\lambda_K^2}{\widetilde{\rho}_-(s^* + \widetilde{s})}. \tag{2.7.9}$$

Since the objective value always decreases within each middle loop, for any inner loop with λ_K, we have $\mathcal{F}_{\lambda_K}(\theta^{(0)}) - \mathcal{F}_{\lambda_K}(\overline{\theta}) \leq \mathcal{F}_{\lambda_K}(\widehat{\theta}^{\{K-1\}}) - \mathcal{F}_{\lambda_K}(\overline{\theta}^{\lambda_K})$. Thus, by Theorem 2.3.9 and (2.7.9), we know that the number of iterations within each inner loop is at most

$$\log^{-1}\left(\frac{\widetilde{\rho}_-(s)\nu_-(1)+s\rho_+^2(s)}{s\rho_+^2(s)}\right)\log\left(\frac{\nu_-(1)\tau_K^2\widetilde{\rho}_-(s^*+\widetilde{s})}{25s^*}\right).$$

[Results (II)] Combining Theorem 2.3.10 with (2.7.9), we know that the number of active set updating iterations within each middle loop is at most

$$\log^{-1}\left(1-\frac{\widetilde{\rho}_-(s^*+2\widetilde{s})}{(s^*+2\widetilde{s})\nu_+(1)}\right)\log\left(\frac{\delta_K^2\widetilde{\rho}_-(s^*+\widetilde{s})}{75\nu_+(1)s^*}\right).$$

[Results (III)] For $K<N$, we take $\lambda'=\lambda_N$, $\lambda=\lambda_K$, and $\theta=\widehat{\theta}^{\{K\}}$. Then by Lemma 2.7.9, we have

$$\mathcal{F}_{\lambda_N}(\widehat{\theta}^{\{K\}}) - \mathcal{F}_{\lambda_N}(\overline{\theta}^{\lambda_N}) \leq \frac{25(\lambda_K+\lambda_N)(\mathcal{K}_{\lambda_K}(\widehat{\theta}^{\{K\}})+3(\lambda_K-\lambda_N))s^*}{\widetilde{\rho}_-(s^*+\widetilde{s})},$$

which completes the proof due to $\lambda_K > \lambda_N$ for $K=0,\ldots,N-1$.

□

2.7.4 Proof of Theorem 2.3.16

Proof. For any θ^*, we consider a partition of $\mathbb{R}^d$ as

$$\mathcal{S}_1 = \left\{ j \;\middle|\; \theta_j^* \geq \frac{C_2\sigma}{\sqrt{s_1^* + s_2^*}} \right\}, \text{ and } \mathcal{S}_{2,3} = \left\{ j \;\middle|\; \theta_j^* < \frac{C_2\sigma}{\sqrt{s_1^* + s_2^*}} \right\}.$$

We consider the first scenario, where $\mathcal{S}_3 = \emptyset$. Then we establish the lower bound for estimating $\theta^*_{\mathcal{S}_1}$ only. Let $\widetilde{\theta}_{\mathcal{S}_1}$ denote any estimator of $\theta^*_{\mathcal{S}_1}$ based on $y \sim N(X_{*\mathcal{S}_1}\theta^*_{\mathcal{S}_1}, \sigma^2 I)$. This is essentially a low dimensional linear regression problem since $s_1^* < n$. By the minimax lower bound for standard linear regression model in [42], we have

$$\inf_{\widetilde{\theta}_{\mathcal{S}_1}} \sup_{\theta \in \Theta(s_1^*, s_2^*, d)} \mathbb{E}\|\widetilde{\theta}_{\mathcal{S}_1} - \theta^*_{\mathcal{S}_1}\|_2 \geq C_6\sigma\sqrt{\frac{s_1^*}{n}}$$

for a generic constant C_6. We then consider a second scenario, where $\mathcal{S}_1 = \emptyset$. Then we establish the lower bound for estimating $\theta^*_{\mathcal{S}_{2,3}}$ only. Let $\widetilde{\theta}_{\mathcal{S}_{2,3}}$ denote any estimator of $\theta^*_{\mathcal{S}_{2,3}}$ based on $y \sim N(X_{*\mathcal{S}_{2,3}}\theta^*_{\mathcal{S}_{2,3}}, \sigma^2 I)$. This is essentially a high dimensional sparse linear regression problem. By the lower bound for sparse linear regression model established in [43], we have

$$\inf_{\widetilde{\theta}_{\mathcal{S}_{2,3}}} \sup_{\theta \in \Theta(s_1^*, s_2^*, d)} \mathbb{E}\|\widetilde{\theta}_{\mathcal{S}_{2,3}} - \theta^*_{\mathcal{S}_{2,3}}\|_2 \geq 2C_7\sigma\sqrt{\frac{s_2^*\log(d - s_2^*)}{n}} \geq C_7\sigma\sqrt{\frac{s_2^*\log d}{n}},$$

where C_7 is a generic constant and the last inequality comes from the fact $s_2^* \ll d$. Combining two scenarios, we have

$$\begin{aligned}
&\inf_{\widehat{\theta}} \sup_{\theta\in\Theta(s_1^*,s_2^*,d)} \mathbb{E}\|\widehat{\theta}-\theta^*\|_2 \\
&\geq \max\left\{\inf_{\widetilde{\theta}_{\mathcal{S}_1}} \sup_{\theta\in\Theta(s_1^*,s_2^*,d)} \mathbb{E}\|\widetilde{\theta}_{\mathcal{S}_1}-\theta^*_{\mathcal{S}_1}\|_2,\ \inf_{\widetilde{\theta}_{\mathcal{S}_{2,3}}} \sup_{\theta\in\Theta(s_1^*,s_2^*,d)} \mathbb{E}\|\widetilde{\theta}_{\mathcal{S}_{2,3}}-\theta^*_{\mathcal{S}_{2,3}}\|_2\right\} \\
&\geq \frac{1}{2}\inf_{\widetilde{\theta}_{\mathcal{S}_1}} \sup_{\theta\in\Theta(s_1^*,s_2^*,d)} \mathbb{E}\|\widetilde{\theta}_{\mathcal{S}_1}-\theta^*_{\mathcal{S}_1}\|_2 + \frac{1}{2}\inf_{\widetilde{\theta}_{\mathcal{S}_{2,3}}} \sup_{\theta\in\Theta(s_1^*,s_2^*,d)} \mathbb{E}\|\widetilde{\theta}_{\mathcal{S}_{2,3}}-\theta^*_{\mathcal{S}_{2,3}}\|_2 \\
&\geq \frac{C_6}{2}\sigma\sqrt{\frac{s_1^*}{n}} + \frac{C_7}{2}\sigma\sqrt{\frac{s_2^*\log d}{n}} \geq C_4\left(\sigma\sqrt{\frac{s_1^*}{n}} + \sigma\sqrt{\frac{s_2^*\log d}{n}}\right),
\end{aligned}$$

where $C_4 = \min\{\frac{C_6}{2}, \frac{C_7}{2}\}$. □

2.7.5 Proof of Theorem 2.3.17

Proof. For notational simplicity, we denote λ_N by λ, $\widehat{\theta}^{\{N\}}$ by $\widehat{\theta}$, and $\overline{\theta}^{\lambda_N}$ by $\overline{\theta}^{\lambda}$. Before we proceed, we first introduce the following lemmas.

Lemma 2.7.10. Suppose $\epsilon \sim N(0,\sigma^2 I)$ and $\|X_{*j}\|_2 = \sqrt{n}$ for $j = 1,...,d$. Then we have

$$\mathbb{P}\left(\frac{1}{n}\|X^\top\epsilon\|_\infty \geq 2\sigma\sqrt{\frac{\log d}{n}}\right) \leq 2d^{-2}.$$

Lemma 2.7.11. Suppose Assumptions 2.3.1 and 2.3.5, and the following event

$$\mathcal{E}_1 = \left\{ \frac{1}{n} \|X^\top \epsilon\|_\infty \geq 2\sigma \sqrt{\frac{\log d}{n}} \right\}$$

hold. We have

$$\frac{1}{n} X_{*\mathcal{S}}(y - X\widehat{\theta}^{\mathrm{o}}) + \nabla_{\mathcal{S}} \mathcal{H}_\lambda(\widehat{\theta}^{\mathrm{o}}) + \lambda \nabla \|\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}}\|_1 = 0.$$

Lemma 2.7.12. Suppose Assumptions 2.3.1, and 2.3.5, and the following event

$$\mathcal{E}_2 = \left\{ \frac{1}{n} \|U^\top \epsilon\|_\infty \geq 2\sigma \sqrt{\frac{\log d}{n}} \right\}$$

hold, where $U = X^\top (I - X_{*\mathcal{S}}(X_{*\mathcal{S}}^\top X_{*\mathcal{S}})^{-1} X_{*\mathcal{S}}^\top)$. There exists some $\widehat{\xi}^{\mathrm{o}}_{\overline{\mathcal{S}}} \in \partial \|\widehat{\theta}^{\mathrm{o}}_{\overline{\mathcal{S}}}\|_1$ such that

$$\frac{1}{n} X_{*\overline{\mathcal{S}}}^\top (y - X\widehat{\theta}^{\mathrm{o}}) + \nabla_{\overline{\mathcal{S}}} \mathcal{H}_\lambda(\widehat{\theta}^{\mathrm{o}}) + \lambda \widehat{\xi}^{\mathrm{o}}_{\overline{\mathcal{S}}} = 0.$$

The proof of Lemma 2.7.10 is provided in [33], therefore is omitted. The proofs of Lemmas 2.7.11 and 2.7.12 are presented in Appendices A.11 and A.12. Lemmas 2.7.11 and 2.7.12 imply that $\widehat{\theta}^{\mathrm{o}}$ satisfies the KKT condition of (2.1.1) over $\mathcal{S}$ and $\overline{\mathcal{S}}$ respectively. Note that the above results only depend on Conditions $\mathcal{E}_1$ and

$\mathcal{E}_2$. Meanwhile, we also have

$$\begin{aligned}\|U_{*j}\|_2 &= \|X_{*j}^\top (I - X_{*\mathcal{S}}(X_{*\mathcal{S}}^\top X_{*\mathcal{S}})^{-1} X_{*\mathcal{S}}^\top)\|_2 \\ &\leq \|I - X_{*\mathcal{S}}(X_{*\mathcal{S}}^\top X_{*\mathcal{S}})^{-1} X_{*\mathcal{S}}^\top\|_2 \|X_{*j}\|_2 \leq \|X_{*j}\|_2 = \sqrt{n}, \end{aligned} \tag{2.7.10}$$

where the last inequality comes from $\|I - X_{*\mathcal{S}}(X_{*\mathcal{S}}^\top X_{*\mathcal{S}})^{-1} X_{*\mathcal{S}}^\top\|_2 \leq 1$. Thus, (2.7.10) implies that Lemma 2.7.10 is also applicable to $\mathcal{E}_2$. Moreover, since both $\widehat{\theta}^{\{N\}}$ and $\widehat{\theta}^{\mathrm{o}}$ are sparse local optima, by Lemma A.3.1, we further have $\mathbb{P}(\widehat{\theta}^{\mathrm{o}} = \overline{\theta}^{\lambda}) \geq 1 - 4d^{-2}$.

Moreover, since $\widehat{\theta}$ converges to $\overline{\theta}^{\lambda}$, given a sufficiently small δ_N, we have

$$\|\nabla \widetilde{\mathcal{L}}_\lambda(\overline{\theta}^{\lambda}) - \nabla \widetilde{\mathcal{L}}_\lambda(\widehat{\theta})\|_\infty \leq \|\widetilde{\mathcal{L}}_\lambda(\overline{\theta}^{\lambda}) - \widetilde{\mathcal{L}}_\lambda(\widehat{\theta})\|_2 \leq \rho_+(s^*)\|\overline{\theta}^{\lambda} - \widehat{\theta}\|_2 \leq \omega \ll \frac{\lambda}{4}.$$

Since we have proved $\|\nabla_{\overline{\mathcal{S}}} \widetilde{\mathcal{L}}_\lambda(\overline{\theta}^{\lambda})\|_\infty \leq \lambda/4$ in Lemma 2.7.12, we have

$$\|\widetilde{\mathcal{L}}_\lambda(\widehat{\theta})\|_\infty \leq \|\nabla_{\overline{\mathcal{S}}} \widetilde{\mathcal{L}}_\lambda(\overline{\theta}^{\lambda})\|_\infty + \|\nabla \widetilde{\mathcal{L}}_\lambda(\overline{\theta}^{\lambda}) - \nabla \widetilde{\mathcal{L}}_\lambda(\widehat{\theta})\|_\infty \leq \frac{\lambda}{4} + \omega.$$

Since $\widehat{\theta}$ also satisfies the approximate KKT condition and $\delta \leq 1/8$, then we must have $\widehat{\theta}_{\overline{\mathcal{S}}} = 0$. Moreover, since we have also proved that there exists some constant C_8 such that $\min_{j \in \mathcal{S}} |\overline{\theta}_j^{\lambda}| \geq C_8 \sigma \sqrt{\log d / n}$ in Lemma 2.7.11, then for $\omega / \rho_-(s^*) \ll C_8 \sigma \sqrt{\log d / n}$, we have

$$\min_{j \in \mathcal{S}} |\widehat{\theta}_j| = \min_{j \in \mathcal{S}} |\overline{\theta}_j^{\lambda}| - \omega \geq C_8 \sigma \sqrt{\frac{\log d}{n}} > 0.$$

Combining with the fact $\widehat{\theta}_{\overline{S}} = 0$, we have $\text{supp}(\widehat{\theta}) = \text{supp}(\overline{\theta}^{\lambda}) = \text{supp}(\theta^*)$. Meanwhile, since all signals are strong enough, then by Theorem 2.3.14, we also have $\|\widehat{\theta} - \theta^*\|_2 \leq C_3\sigma\sqrt{\frac{s^*}{n}}$. □

Chapter 3

Stochastic Variance Reduced Optimization for Nonconvex Sparse Learning

This chapter proposes a stochastic variance reduced optimization algorithm for solving sparse learning problems with cardinality constraints. Sufficient conditions are provided, under which the proposed algorithm enjoys strong linear convergence guarantees and nearly optimal estimation accuracy in high dimensions. We further extend the proposed algorithm to an asynchronous parallel variant with a nearly linear speedup. Numerical experiments demonstrate the efficiency of our algorithm in terms of both parameter estimation and computational performance.

3.1 Background

High dimensionality in learning tasks is challenging from both the statistical and computational perspectives. Based on the principle of parsimony, we usually assume that only a small number of variables are relevant for modeling the response variable. In the past decade, a large family of ℓ_1-regularized or ℓ_1-constrained sparse estimators have been proposed, including Lasso [18], Logistic Lasso [44], Group Lasso [45], Graphical Lasso [46, 47], and more. The ℓ_1-norm serves as a convex surrogate for controlling the cardinality of the parameters, and a large family of algorithms, such as proximal gradient algorithms [48], have been developed for finding the ℓ_1-norm based estimators in polynomial time. The ℓ_1-regularization or constraint, however, often incurs large estimation bias, and attains worse empirical performance than the ℓ_0-regularization and constraint [19, 20]. This motivates us to study a family of cardinality constrained M-estimators. Formally, we consider the following nonconvex optimization problem:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{F}(\theta) \quad \text{subject to } \|\theta\|_0 \leq k, \tag{3.1.1}$$

where $\mathcal{F}(\theta)$ is a smooth and non-strongly convex loss function, and $\|\theta\|_0$ denotes the number of nonzero entries in θ [49, 1].

To solve (4.3.1), a (full) gradient hard thresholding (FG-HT) algorithm has been proposed in the statistics and machine learning communities over the past few

years [49, 1, 50, 51]. FG-HT iteratively performs a gradient update followed by a hard thresholding operation. Let $\mathcal{H}_k(\theta)$ denote a hard thresholding operator that keeps the largest k entries in magnitude and sets the other entries equal to zero. Then, at the t-th iteration, FG-HT performs the update:

$$\theta^{(t)} = \mathcal{H}_k\Big(\theta^{(t-1)} - \eta \nabla \mathcal{F}(\theta^{(t-1)})\Big),$$

where $\eta > 0$ is the step size. Existing literature has shown that under suitable conditions, FG-HT attains linear convergence to an approximately global optimum with optimal estimation accuracy with high probability [49, 1].

Despite these good properties, FG-HT is not suitable for solving large-scale problems. The computational bottleneck is that FG-HT evaluates the (full) gradient at each iteration; its computational complexity depends linearly on the number of samples. Therefore, FG-HT becomes computationally expensive for high-dimensional problems with large sample sizes.

To address the scalability issue, a scenario that is typical in machine learning wherein the objective function decomposes over samples is considered in [2], i.e. the objective function $\mathcal{F}(\theta)$ takes an additive form over many smooth component functions:

$$\mathcal{F}(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta) \quad \text{and} \quad \nabla \mathcal{F}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\theta),$$

and each $f_i(\theta)$ is associated with a few samples of the entire data set (i.e., the mini-batch setting). In such settings, we can exploit the additive nature of $\mathcal{F}(\theta)$ and consider a stochastic gradient hard thresholding (SG-HT) algorithm based on unbiased estimates of the gradient rather than computing the full gradient. In particular, SG-HT uses a stochastic gradient $\nabla f_{i_t}(\theta^{(t)})$ as an estimate of the full gradient $\nabla\mathcal{F}(\theta^{(t)})$, where i_t is randomly sampled from $\{1,\ldots,n\}$ with equal probabilities at each iteration. Though SG-HT greatly reduces the computational cost at each iteration, it can only obtain an estimator with suboptimal estimation accuracy, owing to the variance of the stochastic gradient introduced by random sampling. Moreover, the convergence analysis of SG-HT in [2] requires $\mathcal{F}(\theta)$ to satisfy the Restricted Isometry Property (RIP) with parameter 1/7, i.e., the restricted condition number of the Hessian matrix $\nabla^2\mathcal{F}(\theta)$ cannot exceed 4/3 (see more details in Section 3.3). Taking sparse linear regression as an example, such an RIP condition requires the design matrix to be nearly orthogonal, which is not satisfied even by some simple random correlated Gaussian designs [34].

To address the suboptimal estimation accuracy and the restrictive requirement on $\mathcal{F}(\theta)$ in the stochastic setting, we propose a stochastic variance reduced gradient hard thresholding (SVRG-HT) algorithm. More specifically, we exploit a semi-stochastic optimization scheme to reduce the variance introduced by the random sampling [52, 53]. SVRG-HT contains two nested loops: at each iteration of the outer loop, SVRG-HT calculates the full gradient. In the subsequent inner loop,

the stochastic gradient update is adjusted by the full gradient followed by hard thresholding at each iteration. This simple modification enables the algorithm to attain linear convergence to an approximately global optimum with optimal estimation accuracy, and meanwhile the amortized computational complexity remains similar to that of conventional stochastic optimization. Moreover, our theoretical analysis is applicable to an arbitrarily large restricted condition number of the Hessian matrix $\nabla^2 \mathcal{F}(\theta)$. To further boost the computational performance, we extend SVRG-HT to an asynchronous parallel variant via a lock-free approach for parallelization [54, 55, 56]. We establish theoretically that a near linear speedup is achieved for asynchronous SVRG-HT.

Several existing algorithms are closely related to our proposed algorithm, including the proximal stochastic variance reduced gradient algorithm [57], stochastic averaging gradient algorithm [58], and stochastic dual coordinate ascent algorithm [59]. However, these algorithms guarantee global linear convergence only for strongly convex optimization problems. Several statistical methods in existing literature are also closely related to cardinality constrained M-estimators, including nonconvex constrained M-estimators [60] and nonconvex regularized M-estimators [32]. These methods usually require somewhat complicated computational formulation and often involve many tuning parameters. We discuss these methods in more details in Section 3.6.

The rest of the paper is organized as follows: in Section 2, we derive the SVRG-

HT algorithm; in Section 3, we present the computational and statistical theory; in Section 4, we introduce the parallel variant of SVRG-HT; in Section 5, we present the numerical experiments; in Section 6, we discuss related algorithms and optimization problems; and in Section 7, we present the technical proof of all theorems. Due to space limit, we defer some technical details to Appendix.

3.2 Algorithm

Before we present the proposed algorithm, we introduce some notations. Given an integer $n \geq 1$, we define $[n] = \{1, \ldots, n\}$. Given a vector $v = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$, we define vector norms: $\|v\|_1 = \sum_j |v_j|$, $\|v\|_2^2 = \sum_j v_j^2$, and $\|v\|_\infty = \max_j |v_j|$. Given an index set $\mathcal{I} \subseteq [d]$, we define $\mathcal{I}^C$ as the complement set of $\mathcal{I}$, and $v_\mathcal{I} \in \mathbb{R}^d$, where $[v_\mathcal{I}]_j = v_j$ if $j \in \mathcal{I}$ and $[v_\mathcal{I}]_j = 0$ if $j \notin \mathcal{I}$. We use $\mathrm{supp}(v)$ to denote the index set of nonzero entries of v. Given two vectors $v, w \in \mathbb{R}^d$, we use $\langle v, w \rangle = \sum_{i=1}^d v_i w_i$ to denote the inner product. Given a matrix $A \in \mathbb{R}^{n \times d}$, we use $A^\top$ to denote the transpose, A_{i*} and A_{*j} to denote the i-th row and j-th column respectively, $\sigma_i(A)$ to denote the i-th largest singular value, $\mathrm{rank}(A)$ to denote the rank, $\|A\|_* = \sum_{i=1}^{\mathrm{rank}(A)} \sigma_i(A)$ to denote the nuclear norm, and $\mathrm{vec}(A)$ to denote a vector obtained by concatenating the columns of A. Given an index set $\mathcal{I} \subseteq [d]$, we denote the submatrix of A with all row indices in $\mathcal{I}$ by $A_{\mathcal{I}*}$, and denote the submatrix of A with all column indices in $\mathcal{I}$ by $A_{*\mathcal{I}}$. Given two matrices $A, B \in \mathbb{R}^{n \times d}$, we use

Algorithm 4: Stochastic Variance Reduced Gradient Hard Thresholding Algorithm (SVRG-HT). $\mathcal{H}_k(\cdot)$ is the hard thresholding operator, which keeps the largest k (in magnitude) entries and sets the other entries equal to zero.

Parameter: update frequency m, step size parameter η, sparsity k
Initialize: $\widetilde{\theta}^{(0)}$
For $r = 1, 2, \ldots$
  $\widetilde{\theta} = \widetilde{\theta}^{(r-1)}$
  $\widetilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\widetilde{\theta})$
  $\theta^{(0)} = \widetilde{\theta}$
  For $t = 0, 1, \ldots, m-1$
    (S1) Randomly sample i_t from $[n]$
    (S2) $\overline{\theta}^{(t+1)} = \theta^{(t)} - \eta\left(\nabla f_{i_t}(\theta^{(t)}) - \nabla f_{i_t}(\widetilde{\theta}) + \widetilde{\mu}\right)$
    (S3) $\theta^{(t+1)} = \mathcal{H}_k(\overline{\theta}^{(t+1)})$
  $\widetilde{\theta}^{(r)} = \theta^{(m)}$
Return: $\widetilde{\theta}^{(r)}$

$\langle A, B\rangle = \mathrm{Trace}(A^\top B) = \sum_{i=1}^{n}\sum_{j=1}^{d} A_{ij}B_{ij}$. Moreover, we use the common notations of $\Omega(\cdot)$ and $\mathcal{O}(\cdot)$ to characterize the asymptotics of two real sequences. For logarithmic functions, we denote $\log(\cdot)$ as the natural logarithm when we do not specify the base.

The proposed stochastic variance reduced gradient hard thresholding (SVRG-HT) algorithm is presented in Algorithm 4. Different from the stochastic gradient hard thresholding (SG-HT) algorithm proposed in [2], our SVRG-HT algorithm adopts the semi-stochastic optimization scheme proposed in [52], which can guarantee that the variance introduced by stochastic sampling over component functions decreases with the optimization error.

Next, we sketch a concrete example for illustrating the details of SVRG-HT.

Specifically, we consider a sparse linear model

$$y = A\theta^* + z, \tag{3.2.1}$$

where $A \in \mathbb{R}^{nb\times d}$ is the design matrix, $y \in \mathbb{R}^{nb}$ is the response vector, $\theta^* \in \mathbb{R}^d$ is the unknown sparse regression coefficient vector with $\|\theta^*\|_0 = k^*$, and $z \in \mathbb{R}^{nb}$ is a random noise vector sampled from $N(0, \sigma^2 I)$. We are interested in estimating θ^* by sovling the following nonconvex optimization problem:

$$\min_{\theta\in\mathbb{R}^d} \mathcal{F}(\theta) = \frac{1}{2nb}\|y - A\theta\|_2^2 \quad \text{subject to } \|\theta\|_0 \le k. \tag{3.2.2}$$

To solve (3.2.2) in the stochastic mini-batch optimization regime, we divide A into n submatrices such that each submatrix contains b rows of A, i.e., we have n mini-batches and b is the mini-batch size. For notational simplicity, we define the i-th submatrix as $A_{\mathcal{S}_i *}$, where $\mathcal{S}_i$ is the set of the corresponding row indices with $|\mathcal{S}_i| = b$ for all $i = 1, ..., n$. Accordingly, we have

$$f_i(\theta) = \frac{1}{2b}\|y_{\mathcal{S}_i} - A_{\mathcal{S}_i *}\theta\|_2^2 \quad \text{and} \quad \mathcal{F}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2b}\|y_{\mathcal{S}_i} - A_{\mathcal{S}_i *}\|_2^2.$$

Let us consider the computational cost of SVRG-HT per iteration. Note that the full gradient $\widetilde{\mu} = \nabla\mathcal{F}(\theta)$ remains unchanged through the inner loop, and we only calculate the full gradient once every m inner iterations. We can verify that the

Algorithm 5: Stochastic Average Gradient Hard Thresholding Algorithm (SAGA-HT). The SAGA-HT algorithm has similar computational and statistical performance to SVRG-HT in both theory and practice.

Parameter: step size parameter η, sparsity k
Initialize: $\widetilde{\theta}^{(0)}$
For $r = 1, 2, \ldots$
  Randomly sample i_r from $[n]$
  $\theta_{i_r}^{(r)} = \widetilde{\theta}^{(r-1)}$, and store $\nabla f_{i_t}(\theta_{i_t}^{(r)})$ in the table of stochastic gradients. All other entries
    in the table remain unchanged
  $\overline{\theta}^{(r)} = \widetilde{\theta}^{(r-1)} - \eta\left(\nabla f_{i_t}(\theta^{(t)}) - \nabla f_{i_t}(\theta_i^{(r-1)}) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\theta_i^{(r)})\right)$
  $\widetilde{\theta}^{(r)} = \mathcal{H}_k(\overline{\theta}^{(t+1)})$
Return: $\widetilde{\theta}^{(r)}$

average per iteration computational cost is $\mathcal{O}((n+m)bd/m)$. When m is of the same order of n for some constant $c > 0$, it is further reduced to $\mathcal{O}(bd)$, which matches that of SG-HT up to a constant factor.

A closely related algorithm to SVRG is stochastic average gradient algorithm (SAGA); we refer the reader to [61] for further details. In Algorithm 5, we present an extension of SAGA to SAGA hard thresholding (SAGA-HT) algorithm for nonconvex sparse learning. As for SVRG-HT, the average per iteration computational cost for SAGA-HT is $\mathcal{O}(bd)$. However, unlike SAGA-HT, which needs to maintain n stochastic gradients in the memory resulting in a space complexity of $\mathcal{O}(nd)$, SVRG-HT only maintains a batch gradient in memory relaxing the space requirements to $\mathcal{O}(d)$. This is an enormous advantage for SVRG-HT over SAGA-HT for large n.

3.3 Theory

We are interested in analyzing the convergence of our proposed algorithm to the unknown sparse parameter θ^* of the underlying statistical model. For example, for sparse linear regression in (3.2.1), θ^* is the unknown regression coefficient vector. This is different from the conventional optimization theory, which analyzes the convergence properties of the algorithm to an optimum of the optimization problem.

Our proposed theoretical analysis is applicable to both SVRG-HT and SAGA-HT. As mentioned in Section 3.2, SVRG-HT has an advantage over SAGA-HT in space complexity. Therefore, we focus only on the analysis for SVRG-HT in this section, and an extension to SAGA-HT is straightforward.

Throughout the analysis, we make two important assumptions on the objective function, which are defined as follows.

Definition 3.3.1 (Restricted Strong Convexity Condition)**.** A differentiable function $\mathcal{F}$ is restricted ρ_s^--strongly convex at sparsity level s if there exists a generic constant $\rho_s^- > 0$ such that for any $\theta, \theta' \in \mathbb{R}^d$ with $\|\theta - \theta'\|_0 \leq s$, we have

$$\mathcal{F}(\theta) - \mathcal{F}(\theta') - \langle \nabla \mathcal{F}(\theta'), \theta - \theta' \rangle \geq \frac{\rho_s^-}{2} \|\theta - \theta'\|_2^2. \tag{3.3.1}$$

Definition 3.3.2 (Restricted Strong Smoothness Condition)**.** For any $i \in [n]$, a differentiable function f_i is restricted ρ_s^+-strongly smooth at sparsity level s if there

exists a generic constant $\rho_s^+ > 0$ such that for any $\theta, \theta' \in \mathbb{R}^d$ with $\|\theta - \theta'\|_0 \leq s$, we have

$$f_i(\theta) - f_i(\theta') - \langle \nabla f_i(\theta'), \theta - \theta' \rangle \leq \frac{\rho_s^+}{2} \|\theta - \theta'\|_2^2. \tag{3.3.2}$$

We assume that the objective function $\mathcal{F}(\theta)$ satisfies the restricted strong convexity (RSC) condition, and all component functions $\{f_i(\theta)\}_{i=1}^n$ satisfy the restricted strong smoothness (RSS) condition. Moreover, we define the restricted condition number $\kappa_s = \rho_s^+ / \rho_s^-$. The restricted strong convexity and smoothness have been widely studied in high dimensional statistical theory [34, 32, 62]. They guarantee that the objective function behaves like a strongly convex and smooth function over a sparse domain, which is extremely important for establishing the computational theory.

The restricted isometry property (RIP) is closely related to the RSC and RSS conditions [63, 64]. However, RIP is more restrictive, since it requires $\rho_s^+ < 2$, which can be easily violated by simple random correlated sub-Gaussian designs. Moreover, RIP is only applicable to linear regression, while the RSC and RSS conditions are applicable to more general problems such as sparse generalized linear models estimation.

3.3.1 Computational Theory

We present two key technical lemmas which will be instrumental in developing a computational theory for SVRG-HT. Recall that $\theta^* \in \mathbb{R}^d$ is the unknown sparse vector of interest with $\|\theta^*\|_0 \leq k^*$, and $\mathcal{H}_k(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is a hard thresholding operator that keeps the largest k entries (in magnitude) setting other entries to zero.

Lemma 3.3.3. For $k > k^*$ and for any vector $\theta \in \mathbb{R}^d$, we have

$$\|\mathcal{H}_k(\theta) - \theta^*\|_2^2 \leq \left(1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}\right)\|\theta - \theta^*\|_2^2. \tag{3.3.3}$$

Lemma 3.3.3 shows that the hard thresholding operator is nearly non-expansive for k sufficiently larger than k^* such that $\frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$ is small. The proof of Lemma 3.3.3 is presented in Appendix B.1.

Remark 3.3.4. It is important to note that while Lemma 3.3.3 may seem related to Lemma 1 in [1], there is an important difference. Lemma 1 in [1] characterizes the effect of the hard thresholding operator by bounding the distance $\|\mathcal{H}_k(\theta) - \theta\|_2$ between a vector and its thresholded version. Lemma 3.3.3, on the other hand, bounds the increase in distance of a vector from a fixed target vector (of sparsity k^*) due to thresholding. The latter, we argue, makes more intuitive sense from an optimization perspective.

For notational simplicity, we denote the full gradient and the stochastic vari-

ance reduced gradient by

$$\widetilde{\mu} = \nabla \mathcal{F}(\widetilde{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\widetilde{\theta}) \quad \text{and} \quad g^{(t)}(\theta^{(t)}) = \nabla f_{i_t}(\theta^{(t)}) - \nabla f_{i_t}(\widetilde{\theta}) + \widetilde{\mu}. \tag{3.3.4}$$

The next lemma shows that $g^{(t)}(\theta^{(t-1)})$ is an unbiased estimator of $\nabla \mathcal{F}(\theta^{(t-1)})$ with a well controlled second moment over a sparse support.

Lemma 3.3.5. Suppose that $\mathcal{F}(\theta)$ satisfies the RSC condition and that functions $\{f_i(\theta)\}_{i=1}^n$ satisfy the RSS condition with $s = 2k + k^*$. Let $\mathcal{I}^* = \text{supp}(\theta^*)$ denote the support of θ^*. Let $\theta^{(t)}$ be a sparse vector with $\|\theta^{(t)}\|_0 \leq k$ and support $\mathcal{I}^{(t)} = \text{supp}(\theta^{(t)})$. Then conditioning on $\theta^{(t)}$, for any $\mathcal{I} \supseteq (\mathcal{I}^* \cup \mathcal{I}^{(t)})$, we have $\mathbb{E}[g^{(t)}(\theta^{(t)})] = \nabla \mathcal{F}(\theta^{(t)})$ and

$$\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2 \leq 12\rho_s^+ \left[\mathcal{F}(\theta^{(t)}) - \mathcal{F}(\theta^*) + \mathcal{F}(\widetilde{\theta}) - \mathcal{F}(\theta^*)\right] + 3\|\nabla_{\mathcal{I}} \mathcal{F}(\theta^*)\|_2^2. \tag{3.3.5}$$

The proof of Lemma 3.3.5 is presented in Appendix B.2.

Remark 3.3.6. For smooth convex problems, we have $\nabla \mathcal{F}(\theta^*) = 0$ if θ^* is a global minimizer. However, given that the problem of interest here, Problem 4.3.1, is nonconvex, the second term on the R.H.S of (3.3.5) is nonzero. This results in a setting different from existing work using variance reduction [52, 53].

We now present our first main result characterizing the quality of solution given by Algorithm 4 both in terms of the error in the objective value as well as

error in terms of the parameter estimation.

Theorem 3.3.7. Let θ^* denote the unknown sparse parameter vector of the underlying statistical model, with $\|\theta^*\|_0 \leq k^*$. Assume that the objective function $\mathcal{F}(\theta)$ satisfies the RSC condition and functions $\{f_i(\theta)\}_{i=1}^n$ satisfy the RSS condition with $s = 2k + k^*$, where $k \geq C_1 \kappa_s^2 k^*$ and C_1 is a generic constant. Define

$$\widetilde{\mathcal{I}} = \operatorname{supp}\left(\mathcal{H}_{2k}(\nabla \mathcal{F}(\theta^*))\right) \cup \operatorname{supp}(\theta^*).$$

There exist generic constants C_2, C_3, and C_4 such that if we set $\eta\rho_s^+ \in [C_2, C_3]$ and $m \geq C_4 \kappa_s$, then we have

$$\frac{\left(1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}\right)^m \cdot \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}}{\eta\rho_s^-(1-6\eta\rho_s^+)\left(\left(1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}\right)^m - 1\right)} + \frac{6\eta\rho_s^+}{1-6\eta\rho_s^+} \leq \frac{3}{4}.$$

Furthermore, the parameter $\widetilde{\theta}^{(r)}$ at the r-th iteration of SVRG-HT satisfies

$$\mathbb{E}\left[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\right] \leq \left(\frac{3}{4}\right)^r \cdot \left[\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)\right] + \frac{6\eta}{(1-6\eta\rho_s^+)} \|\nabla_{\widetilde{\mathcal{I}}} \mathcal{F}(\theta^*)\|_2^2 \quad \text{and} \quad (3.3.6)$$

$$\mathbb{E}\|\widetilde{\theta}^{(r)} - \theta^*\|_2 \leq \sqrt{\frac{2\left(\frac{3}{4}\right)^r \left[\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)\right]}{\rho_s^-}} + \frac{2\sqrt{s}\|\nabla \mathcal{F}(\theta^*)\|_\infty}{\rho_s^-} + \|\nabla_{\widetilde{\mathcal{I}}} \mathcal{F}(\theta^*)\|_2 \sqrt{\frac{12\eta}{(1-6\eta\rho_s^+)\rho_s^-}}. \quad (3.3.7)$$

Moreover, given a constant $\delta \in (0,1)$ and a pre-specified accuracy $\varepsilon > 0$, we need at

most

$$r = \left\lceil 4 \log\left(\frac{\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)}{\varepsilon\delta} \right) \right\rceil$$

outer iterations such that with probability at least $1 - \delta$, we have

$$\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*) \leq \varepsilon + \frac{6\eta}{(1 - 6\eta\rho_s^+)} \|\nabla_{\widetilde{\mathcal{I}}} \mathcal{F}(\theta^*)\|_2^2 \quad \text{and} \tag{3.3.8}$$

$$\|\widetilde{\theta}^{(r)} - \theta^*\|_2 \leq \sqrt{\frac{2\varepsilon}{\rho_s^-}} + \frac{2\sqrt{s}\|\nabla \mathcal{F}(\theta^*)\|_\infty}{\rho_s^-} + \|\nabla_{\widetilde{\mathcal{I}}} \mathcal{F}(\theta^*)\|_2 \sqrt{\frac{12\eta}{(1 - 6\eta\rho_s^+)\rho_s^-}}. \tag{3.3.9}$$

The proof of Theorem 3.3.7 is presented in Section 3.7.1.

Remark 3.3.8. Theorem 3.3.7 has two important implications: **(I)** Our analysis for SVRG-HT allows an arbitrary large κ_s as long as $\mathcal{F}(\theta)$ and $\{f_i(\theta)\}_{i=1}^n$ satisfy the RSC and RSS conditions respectively with $s = \Omega(\kappa_s^2 k^*)$. In contrast, the theoretical analysis for SG-HT in [2] requires κ_s not to exceed 4/3, which is very restrictive; **(II)** To get $\widetilde{\theta}^{(r)}$ to satisfy (3.3.8) and (3.3.9), we need $\mathcal{O}(\log(1/\varepsilon))$ outer iterations. Since within each outer iteration, we need to calculate a full gradient and m stochastic variance reduced gradients, the overall computational complexity of SVRG-HT is

$$\mathcal{O}\left([n + \kappa_s] \cdot \log\left(\frac{1}{\varepsilon} \right) \right).$$

In contrast, the overall computational complexity of the full gradient hard thresholding algorithm (FG-HT) is $\mathcal{O}(\kappa_s n \log(1/\varepsilon))$. Thus SVRG-HT yields a significant

improvement over FG-HT when κ_s is large.

3.3.2 Statistical Theory

SVRG-HT is applicable to a large family of sparse learning problems. Here, we present theoretical results for three popular examples of constrained M-estimation problems: sparse linear regression, sparse generalized linear model estimation, and low-rank matrix estimation (where the cardinality constraint is replaced by a rank constraint).

3.3.2.1 Sparse Linear Regression

Consider the sparse linear model

$$y = A\theta^* + z,$$

as introduced in Section 3.2. We want to estimate θ^* by solving the optimization problem in (3.2.2). We assume that for any $v \in \mathbb{R}^d$ with $\|v\|_0 \leq s$, the design matrix A satisfies

$$\frac{\|Av\|_2^2}{nb\|v\|_2^2} \geq \psi_1 - \varphi_1 \frac{\log d}{nb} \frac{\|v\|_1^2}{\|v\|_2^2} \quad \text{and} \quad \frac{\|A_{\mathcal{S}_i *} v\|_2^2}{b\|v\|_2^2} \leq \psi_2 + \varphi_2 \frac{\log d}{b} \frac{\|v\|_1^2}{\|v\|_2^2}, \forall i \in [n], \quad (3.3.10)$$

where ψ_1, ψ_2, φ_1, and φ_2 are constants that do not scale with (n, b, k^*, d). Existing literature has shown that (3.3.10) is satisfied by many common examples of sub-Gaussian random design [34, 62]. The next lemma shows that (3.3.10) implies the RSC and RSS conditions.

Lemma 3.3.9. Suppose that the design matrix A satisfies (3.3.10). Then, given large enough n and b, there exist a constant C_5 and an integer k such that $\mathcal{F}(\theta)$ and $\{f_i(\theta)\}_{i=1}^n$ satisfy the RSC and RSS conditions respectively with $s = 2k + k^*$, where

$$k = C_5 k^* \geq C_1 \kappa_s^2 k^*, \quad \rho_s^- \geq \psi_1/2, \quad \text{and} \quad \rho_s^+ \leq 2\psi_2.$$

A proof of Lemma 3.3.9 can be found in Appendix B.3. Combining Lemma 3.3.9 and Theorem 3.3.7, we get the following computational and statistical guarantees for the estimator obtained by SVRG-HT.

Corollary 3.3.10. Suppose that the design matrix A satisfies (3.3.10) with $\frac{\max_j \|A_{*j}\|_2}{\sqrt{nb}} \leq 1$, and k, η and m are as specified in Theorem 3.3.7. Then, for any confidence parameter $\delta \in (0,1)$, a sufficiently small accuracy parameter $\varepsilon > 0$, and large enough n and b, we need at most $r = \left\lceil 4 \log\left(\frac{\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)}{\varepsilon\delta}\right)\right\rceil$ outer iterations in SVRG-HT to guarantee that with high probability, we have

$$\|\widetilde{\theta}^{(r)} - \theta^*\|_2 = \mathcal{O}\left(\sigma\sqrt{\frac{k^* \log d}{nb}}\right). \tag{3.3.11}$$

See Section 3.7.2 for a proof of Corollary 3.3.10.

Remark 3.3.11. Corollary 3.3.10 guarantees that the proposed SVRG-HT estimator attains the optimal statistical rate of convergence in parameter estimation [43] when $\varepsilon = \mathcal{O}\left(\sigma\sqrt{\frac{k^*\log d}{nb}}\right)$. In contrast, previous work, for instance see Corollary 5 in [2], shows that the estimator obtained by the SGHT algorithm attains the statistical rate of convergence

$$\mathcal{O}\left(\sigma\sqrt{\frac{k^*\log d}{b}}\right)$$

with high probability, and hence is suboptimal when n scales with (b, k^*, d).

3.3.2.2 Sparse Generalized Linear Models

We next consider sparse generalized linear models (GLM) defined by the following conditional distribution

$$\mathbb{P}(y_i|A_{i*}, \theta^*, \sigma) = \exp\left\{\frac{y_i A_{i*}\theta^* - h(A_{i*}\theta^*)}{a(\sigma)}\right\},$$

where $a(\sigma)$ is a fixed and known scale parameter, $\theta^* \in \mathbb{R}^d$ is the unknown sparse regression coefficient with $\|\theta^*\|_0 = k^*$, and $h(\cdot)$ is the cumulant function[65] satisfy-

ing

$$h'(A_{i*}\theta^*) = \mathbb{E}[y_i | A_{i*}, \theta^*, \sigma].$$

We further assume that there exists some constant c_u such that $h''(x) \leq c_u$ for all $x \in \mathbb{R}$. Such a boundedness assumption is necessary to establish the RSC and RSS conditions for GLM [32]. Note that this assumption holds for various popular settings, including linear regression, logistic regression, and multinomial regression.

Analogous to sparse linear regression, we divide A into n mini-batches, where each mini-batch is denoted by $A_{\mathcal{S}_i *}$ and $\mathcal{S}_i$ denotes the corresponding row indices of A, with $|\mathcal{S}_i| = b$, for all $i = 1, ..., n$. Then, our objective is essentially the negative log-likelihood, i.e.,

$$\min_{\theta \in \mathbb{R}^d} \mathcal{F}(\theta) = \frac{1}{a(\sigma) \cdot n} \sum_{i=1}^{n} f_i(\theta) \text{ subject to } \|\theta\|_0 \leq k, \ \|\theta\|_2 \leq \tau, \tag{3.3.12}$$

for some $\tau > 0$, where $f_i(\theta) = \frac{1}{b} \sum_{\ell \in \mathcal{S}_i} (h(A_{\ell *}\theta) - y_\ell A_{\ell *}\theta)$, for all $i = 1, \ldots, n$. The additional constraint $\|\theta\|_2 \leq \tau$ in (3.3.12) may not be necessary in practice, but it is essential for our theoretical analysis; we further expand on this later in this section.

For concreteness, we consider sparse logistic regression as a special case of the setup above. We want to estimate θ^* from nb independent responses $y_\ell \sim$ Bernoulli$(\pi_\ell(\theta^*))$, $\ell \in [nb]$, where $\pi_\ell(\theta^*) = \left(\frac{\exp(A_{\ell *}^\top \theta^*)}{1 + \exp(A_{\ell *}^\top \theta^*)} \right)$. The resulting optimiza-

tion problem is as follows:

$$\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{b}\sum_{\ell\in\mathcal{S}_i}(\log[1+\exp(A_{\ell *}\theta)]-y_\ell A_{\ell *}\theta) \text{ subject to } \|\theta\|_0\le k,\ \|\theta\|_2\le\tau.$$

Assume that for any $v\in\mathbb{R}^d$ with $\|v\|_0\le s$ and $\|v\|_2\le 2\tau$, the design matrix A satisfies $\frac{\max_j\|A_{*j}\|_2}{\sqrt{nb}}\le 1$, and the objective $\mathcal{F}(\theta)$ and $\{f_i(\theta)\}_{i=1}^n$ satisfy

$$\begin{aligned} v^\top\nabla^2\mathcal{F}(\theta)v &\ge \psi_1\|v\|_2^2-\varphi_1\frac{\log d}{nb}\|v\|_1^2 \quad \text{and} \\ v^\top\nabla^2 f_i(\theta)v &\le \psi_2\|v\|_2^2+\varphi_2\frac{\log d}{b}\|v\|_1^2, \end{aligned} \tag{3.3.13}$$

where ψ_1, ψ_2, φ_1 and φ_2 are constants that do not scale with (n,b,k^*,d) – (3.3.13) is satisfied by many common examples of sub-Gaussian random design [32]. We show that (3.3.13) implies the RSC and RSS conditions over an ℓ_2 ball centered at θ^* with radius 2τ.

Lemma 3.3.12. Suppose that $\mathcal{F}(\theta)$ and $\{f_i(\theta)\}_{i=1}^n$ satisfy (3.3.13). Then, given large enough n and b, for any θ with $\|\theta-\theta^*\|_2\le 2\tau$, there exist a constant C_6 and an integer k such that $\mathcal{F}(\theta)$ and $\{f_i(\theta)\}_{i=1}^n$ satisfy the RSC and RSS conditions respectively with $s=2k+k^*$, where

$$k=C_6k^*\ge C_1\kappa_s^2k^*,\quad \rho_s^-\ge\psi_1/2,\quad \text{and}\quad \rho_s^+\le 2\psi_2.$$

The proof of Lemma 3.3.12 is analogous to the proof of Lemma 3.3.9, thus is

omitted. Lemma 3.3.12 guarantees that the RSC and RSS conditions hold over a neighborhood of θ^*. For sparse GLM, we further assume $\|\theta^*\|_2 \leq \tau$. This implies that for any $\theta \in \mathbb{R}^d$ with $\|\theta\|_2 \leq \tau$, we have $\|\theta - \theta^*\|_2 \leq \|\theta\|_2 + \|\theta^*\|_2 \leq 2\tau$.

Remark 3.3.13 (SVRG-HT with Projection)**.** Due to the additional ℓ_2-constraint, we need a projection step in SVRG-HT. In particular, we replace Step (S3) in Algorithm 4 with the following update:

$$\theta^{(t+1)} = \Pi_\tau(\mathcal{H}_k(\overline{\theta}^{(t+1)})),$$

where $\Pi_\tau(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is an ℓ_2-norm projection operator defined as $\Pi_\tau(v) = \max\{\|v\|_2, \tau\} \cdot v/\|v\|_2$ for any $v \in \mathbb{R}^d$. Since $\Pi_\tau(\cdot)$ is strictly contractive, i.e., $\|\Pi_\tau(\theta) - \theta^*\|_2 \leq \|\theta - \theta^*\|_2$, Theorem 3.3.7 still holds[1] for SVRG-HT with this additional projection step.

Our next result gives the statistical rate of convergence of the obtained estimator for sparse GLM estimation.

Corollary 3.3.14. Suppose that A_{i*}'s have i.i.d. sub-Gaussian rows, and k, η and m are as specified in Theorem 3.3.7. In addition, suppose $\|\theta^*\|_2 \leq \tau$. Then, given a constant $\delta \in (0, 1)$, a sufficiently small accuracy parameter $\varepsilon > 0$, and large enough n and b, we need at most $r = \left\lceil 4\log\left(\frac{\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)}{\varepsilon\delta}\right)\right\rceil$ outer iterations of SVRG-HT so

[1]The gap of the objective value is also contractive after projection due to the convexity of $\mathcal{F}$.

as to guarantee that, with high probability, we have

$$\|\widetilde{\theta}^{(r)} - \theta^*\|_2 = \mathcal{O}\left(\sqrt{\frac{k^* \log d}{nb}}\right). \tag{3.3.14}$$

We note that the statistical rate of convergence above matches the state-of-the-art result in parameter estimation for GLM; see [32] for more details. A proof of Corollary 3.3.14 is given in Section 3.7.3.

3.3.2.3 Low-rank Matrix Recovery

Next, we consider a low-rank matrix linear model

$$y = \mathcal{A}(\Theta^*) + z,$$

where $y \in \mathbb{R}^{nb}$ is the response vector, $\Theta^* \in \mathbb{R}^{d \times p}$ is the unknown low-rank matrix with $\text{rank}(\Theta^*) = k^*$, $\mathcal{A}(\cdot) : \mathbb{R}^{d \times p} \to \mathbb{R}^{nb}$ is a linear operator defined as $\mathcal{A}(\Theta) = [\langle A_1, \Theta \rangle, \ldots, \langle A_{nb}, \Theta \rangle]^\top$ for any matrix $\Theta \in \mathbb{R}^{d \times p}$, $A_i \in \mathbb{R}^{d \times p}$ is a measurement matrix for all $i = 1, \ldots, nb$, and $z \in \mathbb{R}^{nb}$ is a random noise vector sampled from $N(0, \sigma^2 I)$.

As before, we divide the observations into n blocks, indexed by $y_{\mathcal{S}_i}$, where $\mathcal{S}_i$ denotes the corresponding indices of y, with $|\mathcal{S}_i| = b$, for all $i = 1, \ldots, n$. Then, the

resulting optimization problem is

$$\min_{\Theta \in \mathbb{R}^{d \times p}} \mathcal{F}(\Theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\Theta) \quad \text{subject to } \operatorname{rank}(\Theta) \leq k, \tag{3.3.15}$$

where $f_i(\Theta) = \frac{1}{2b} \|y_{\mathcal{S}_i} - \mathcal{A}_{\mathcal{S}_i}(\Theta)\|_2^2$ and $\mathcal{A}_{\mathcal{S}_i}(\Theta)$ denotes a sub-vector of $\mathcal{A}(\Theta)$ indexed by $\mathcal{S}_i$, for all $i = 1, \ldots, n$.

For low-rank matrix problems, we consider the following matrix RSC and RSS conditions that are simple generalization of the RSC and RSS conditions for sparse vectors in Definitions 3.3.1 and 3.3.2. These matrix RSC and RSS conditions were studied recently in high-dimensional statistical analyses for low-rank matrix recovery [66, 33, 67].

Definition 3.3.15 (Matrix Restricted Strong Convexity Condition)**.** A differentiable function $\mathcal{F} : \mathbb{R}^{d \times p} \to \mathbb{R}$ is restricted ρ_s^--strongly convex at rank level s if there exists a generic constant $\rho_s^- > 0$ such that for any $\Theta, \Theta' \in \mathbb{R}^{d \times p}$ with $\operatorname{rank}(\Theta - \Theta') \leq s$, we have

$$\mathcal{F}(\Theta) - \mathcal{F}(\Theta') - \langle \nabla \mathcal{F}(\Theta'), \Theta - \Theta' \rangle \geq \frac{\rho_s^-}{2} \|\Theta - \Theta'\|_{\mathrm{F}}^2. \tag{3.3.16}$$

Definition 3.3.16 (Matrix Restricted Strong Smoothness Condition)**.** For any $i \in [n]$, a differentiable function $f_i : \mathbb{R}^{d \times p} \to \mathbb{R}$ is restricted ρ_s^+-strongly smooth at rank level s if there exists a generic constant $\rho_s^+ > 0$ such that for any $\Theta, \Theta' \in \mathbb{R}^{d \times p}$ with

$\text{rank}(\Theta - \Theta') \leq s$, we have

$$f_i(\Theta) - f_i(\Theta') - \langle \nabla f_i(\Theta'), \Theta - \Theta' \rangle \leq \frac{\rho_s^+}{2} \|\Theta - \Theta'\|_{\text{F}}^2. \tag{3.3.17}$$

As with the RSC and RSS conditions, the matrix RSC and RSS conditions can be verified for $\mathcal{F}(\Theta)$ and $\{f_i(\Theta)\}_{i=1}^n$ by studying sub-Gaussian random design [66]. Specifically, if $\{A_i\}_{i=1}^{nb}$ in the linear operator $\mathcal{A}(\cdot)$ are drawn i.i.d. from the $\Sigma_{\mathcal{A}}$-Gaussian ensemble, i.e., $\text{vec}(A_i) \sim N(0, \Sigma_{\mathcal{A}})$ with $\Sigma_{\mathcal{A}} \in \mathbb{R}^{dp \times dp}$, then, with high probability, we have

$$\begin{aligned}
\frac{\mathcal{A}(\Theta)}{\sqrt{nb}} &\geq \psi_1 \|\sqrt{\Sigma_{\mathcal{A}}}\text{vec}(\Theta)\|_2 - \varphi_1 \rho(\Sigma_{\mathcal{A}}) \left(\sqrt{\frac{d}{nb}} + \sqrt{\frac{p}{nb}} \right) \|\Theta\|_* \quad \text{and} \\
\frac{\mathcal{A}_{\mathcal{S}_i}(\Theta)}{\sqrt{b}} &\leq \psi_2 \|\sqrt{\Sigma_{\mathcal{A}}}\text{vec}(\Theta)\|_2 - \varphi_2 \rho(\Sigma_{\mathcal{A}}) \left(\sqrt{\frac{d}{b}} + \sqrt{\frac{p}{b}} \right) \|\Theta\|_* \quad \text{for all} \quad i = 1, \ldots, n,
\end{aligned}$$

where $\rho^2(\Sigma_{\mathcal{A}}) = \sup_{\|u\|_2=1, \|v\|_2=1} \text{var}(u^\top X v)$, and the random matrix X is sampled from the $\Sigma_{\mathcal{A}}$-Gaussian ensemble. This further implies that $\mathcal{F}(\Theta)$ and $\{f_i(\Theta)\}_{i=1}^n$ satisfy the matrix RSC and RSS conditions respectively for large enough k, following the result in Lemma 3.3.12.

Remark 3.3.17 (SVRG-HT for Singular Value Thresholding)**.** For low-rank matrix recovery, we need to replace the hard thresholding operator $\mathcal{H}_k(\cdot)$ in Step (S3) of Algorithm 4 by the singular value thresholding operator $\mathcal{R}_k(\cdot)$. In particular, we

replace Step (S3) with the following update:

$$\Theta^{(t+1)} = \mathcal{R}_k(\overline{\Theta}^{(t+1)}) = \sum_{i=1}^{r} \overline{\sigma}_i \overline{U}_i \overline{V}_i^\top,$$

where $\overline{\sigma}_i$, $\overline{U}_i$, and $\overline{V}_i$ are the i-th largest singular value, and the corresponding left and right singular vectors of $\overline{\Theta}^{(t+1)}$ respectively.

For sparse vectors, Lemma 3.3.3 guarantees that the hard thresholding operation is nearly non-expansive when k is sufficiently larger than k^*. We provide a similar result for the singular value thresholding operation on matrices.

Lemma 3.3.18. Recall that $\Theta^* \in \mathbb{R}^{d\times p}$ is the unknown low-rank matrix of interest with $\text{rank}(\Theta^*) \leq k^*$, and $\mathcal{R}_k(\cdot): \mathbb{R}^{d\times p} \to \mathbb{R}^{d\times p}$ is the singular value thresholding operator, which keeps the largest k singular values and sets the other singular values equal to zero. Given $k > k^*$, for any matrix $\Theta \in \mathbb{R}^{d\times p}$, we have

$$\|\mathcal{R}_k(\Theta) - \Theta^*\|_{\text{F}}^2 \leq \left(1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}\right) \cdot \|\Theta - \Theta^*\|_{\text{F}}^2. \tag{3.3.18}$$

See Appendix B.4 for a proof of Lemma 3.3.18. Given Lemma 3.3.18, the computational theory follows directly from Theorem 3.3.7. This further allows us to characterize the statistical properties of the obtained estimator for low-rank matrix recovery as follows.

Corollary 3.3.19. Suppose that in the linear operator $\mathcal{A}(\cdot)$, $\text{vec}(A_i)$ is drawn i.i.d.

from $N(0, \Sigma_{\mathcal{A}})$, and k, η and m are as specified in Theorem 3.3.7. Then, given a constant $\delta \in (0,1)$, a sufficiently small accuracy parameter $\varepsilon > 0$, and large enough n and b, we need at most $r = \left\lceil 4\log\left(\frac{\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)}{\varepsilon\delta}\right)\right\rceil$ outer iterations of SVRG-HT to guarantee that, with high probability, we have

$$\|\widetilde{\Theta}^{(r)} - \Theta^*\|_{\mathrm{F}} = \mathcal{O}\left(\sigma\sqrt{\frac{k^*(d+p)}{nb}}\right).$$

The statistical rate of the convergence in Corollary 3.3.19 matches with the state-of-the-art result in parameter estimation for low-rank matrix recovery [66]. The analysis follows directly from Corollary 3.3.10 and [66].

3.4 Asynchronous SVRG-HT

We extend SVRG-HT to an asynchronous parallel variant, named asynchronous SVRG-HT (ASVRG-HT). Here, we assume a parallel computing procedure with a multicore architecture, where each processor makes a stochastic gradient update on a global parameter stored in a shared memory in an asynchronous and lock-free mode. This setup is similar to that used in many asynchronous algorithms [54, 55, 56, 68].

Compared with SVRG-HT, the algorithmic difference is as follows: at the t-th iteration of inner loop, we randomly sample an index $i_t \in [n]$ of the component function with equal probability and an index set $e_t \subset [d]$ over all subsets of $[d]$ with

Algorithm 6: Asynchronous Stochastic Variance Reduced Gradient Hard Thresholding Algorithm. We assume a parallel computing procedure with a multicore architecture, where each processor makes a stochastic gradient update of a global parameter stored in a shared memory via an asynchronous and lock-free mode.

Parameter: update frequency m, step size parameter η, sparsity k
Initialize: $\widetilde{\theta}^{(0)}$
For $r = 1, 2, \ldots$
  $\widetilde{\theta} = \widetilde{\theta}^{(r-1)}$
  $\widetilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\widetilde{\theta})$
  $\theta^{(0)} = \widetilde{\theta}$
  For $t = 0, 1, \ldots, m-1$
    (S1) Randomly sample i_t from $[n]$ and $e_t \subset [d]$ with $|e_t| \leq k$
    (S2) $\overline{\theta}^{(t+1)} = \theta^{(t)} - \eta \cdot [g^{(t)}(\theta^{(t)})]_{e_t}$, where $g^{(t)}(\theta^{(t)}) = \nabla f_{i_t}(\theta^{(t)}) - \nabla f_{i_t}(\widetilde{\theta}) + \widetilde{\mu}$
    (S3) $\theta^{(t+1)} = \mathcal{H}_k(\overline{\theta}^{(t+1)})$
  $\widetilde{\theta}^{(r)} = \theta^{(m)}$
Return: $\widetilde{\theta}^{(r)}$

equal probability, where e_t has a fixed cardinality upper bounded by k for any t. Then we only update $\theta^{(t)}$ over the index set e_t of the variance reduced gradient $g^{(t)}(\theta^{(t)})$. The full algorithm is presented in Algorithm 6.

We first introduce two important parameters following the notions in [54]. The first parameter ς captures the degree of parallelism in the asynchronous algorithm. Let t' be the actual evaluation of θ performed at the t-th iteration, then ς is the smallest positive integer such that $t - t' \leq \varsigma$ for any t. This is an upper bound of delay that the actual evaluation of parameter is performed at the current iteration. The more parallel computations are adopted, the larger value of ς can be. The value of ς is approximately linear on the number of cores in parallel computing architecture [54, 56].

The second parameter Δ captures the sparsity of data. Suppose $f_i(\theta)$ only depends on θ_{e_i}, where $e_i \subset [d]$ and $|e_i| = k_i$ for some positive integer k_i. Then $\Delta \in [0,1]$ is the smallest constant such that $\mathbb{E}\|\theta_e\|_2^2 \leq \Delta\|\theta\|_2^2$, where $e \subseteq [d]$ is a subset of $[d]$ with $|e| = k_i$ sampled with equal probabilities. The sparser the parameter is, on which the function depends, the smaller Δ is. We are interested in the setting $\Delta \ll 1$.

We now present our main result characterizing the error of the objective value and estimation error for Algorithm 6.

Theorem 3.4.1. Let θ^* be the unknown sparse vector of our interest with $\|\theta^*\|_0 \leq k^*$. Suppose $\mathcal{F}(\theta)$ satisfies the RSC condition and $\{f_i(\theta)\}_{i=1}^n$ satisfy RSS condition with $s = 2k + k^*$, where $k \geq C_1\kappa_s^2 k^*$ and C_1 is a generic constant. We define

$$\widetilde{\mathcal{I}} = \mathrm{supp}\left(\mathcal{H}_{2k}(\nabla\mathcal{F}(\theta^*))\right) \cup \mathrm{supp}(\theta^*).$$

There exist generic constants C_2, C_3, C_4, and C_5 such that if we set $\eta\rho_s^+ \in [C_2, C_3]$, $m \geq C_4\kappa_s$ and $\Delta\varsigma^2 \leq C_5$, then

$$\frac{\left(1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}\right)^m \cdot \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}}{\eta\rho_s^-(1 - 12\eta\rho_s^+\Gamma)\left(\left(1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}\right)^m - 1\right)} + \frac{12\eta\rho_s^+\Gamma}{1 - 12\eta\rho_s^+\Gamma} \leq \frac{5}{6}.$$

where $\Gamma = \frac{1+\rho_s^+\Delta\varsigma^2\eta}{1-2\rho_s^{+2}\Delta\varsigma^2\eta^2}$. Further, the parameter $\widetilde{\theta}^{(r)}$ at the r-th iteration of ASVRG-

HT satisfies

$$\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\Big] \leq \left(\frac{5}{6}\right)^r \Big[\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)\Big] + \frac{18\eta\Gamma}{1-12\eta\rho_s^+\Gamma}\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2^2 \quad \text{and}$$

$$\mathbb{E}\|\widetilde{\theta}^{(r)} - \theta^*\|_2 \leq \sqrt{\frac{2\left(\frac{5}{6}\right)^r \Big[\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)\Big]}{\rho_s^-}} + \frac{2\sqrt{s}\|\nabla\mathcal{F}(\theta^*)\|_\infty}{\rho_s^-} + \|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2\sqrt{\frac{36\eta\Gamma}{(1-12\eta\rho_s^+\Gamma)\rho_s^-}}.$$

Moreover, given a constant $\delta \in (0,1)$ and a pre-specified accuracy $\varepsilon > 0$, we need at most

$$r = \left\lceil 4\log\left(\frac{\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)}{\varepsilon\delta}\right)\right\rceil$$

outer iterations such that with probability at least $1-\delta$, we have simultaneously

$$\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*) \leq \varepsilon + \frac{18\eta\Gamma}{1-12\eta\rho_s^+\Gamma}\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2^2 \quad \text{and}$$

$$\|\widetilde{\theta}^{(r)} - \theta^*\|_2 \leq \sqrt{\frac{2\varepsilon}{\rho_s^-}} + \frac{2\sqrt{s}\|\nabla\mathcal{F}(\theta^*)\|_\infty}{\rho_s^-} + \|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2\sqrt{\frac{36\eta\Gamma}{(1-12\eta\rho_s^+\Gamma)\rho_s^-}}.$$

The proof of Theorem 3.4.1 is presented in Section 3.7.4. Theorem 3.4.1 indicates that ASVRG-HT has a similar iteration complexity to SVRG-HT. Therefore, when $\Delta\varsigma^2 = \mathcal{O}(1)$, ASVRG-HT can be ς times faster than SVRG-HT due to the parallelism. For example, if $\Delta = \mathcal{O}(k/d)$, then we achieve a speedup of $\varsigma = \Omega(\sqrt{d/k})$ times, which is analogous to ASVRG in [55]. Since Theorem 3.4.1 provides similar computational guarantees for ASVRG-HT, we can further establish similar statis-

tical guarantees for ASVRG-HT by following Section 3.3.2.

Remark 3.4.2. To ease the analysis, we assume a sampling scheme of with-replacement for parallelism, where only one component of the gradient is used to update the parameter to avoid using locks in practice. However, in practice, a scheme of without-replacement can be applied to significantly improve the efficiency [54].

3.5 Experiments

We compare the empirical performance of SVRG-HT with two other competitors: FG-HT proposed in [1] and SG-HT proposed in [2] on both synthetic data and real data. We also compare the performance of parameter estimation between the ℓ_0-constrained problem (4.3.1) and an ℓ_1-regularized problem solved by the proximal stochastic variance reduced gradient (Prox-SVRG) algorithm [57].

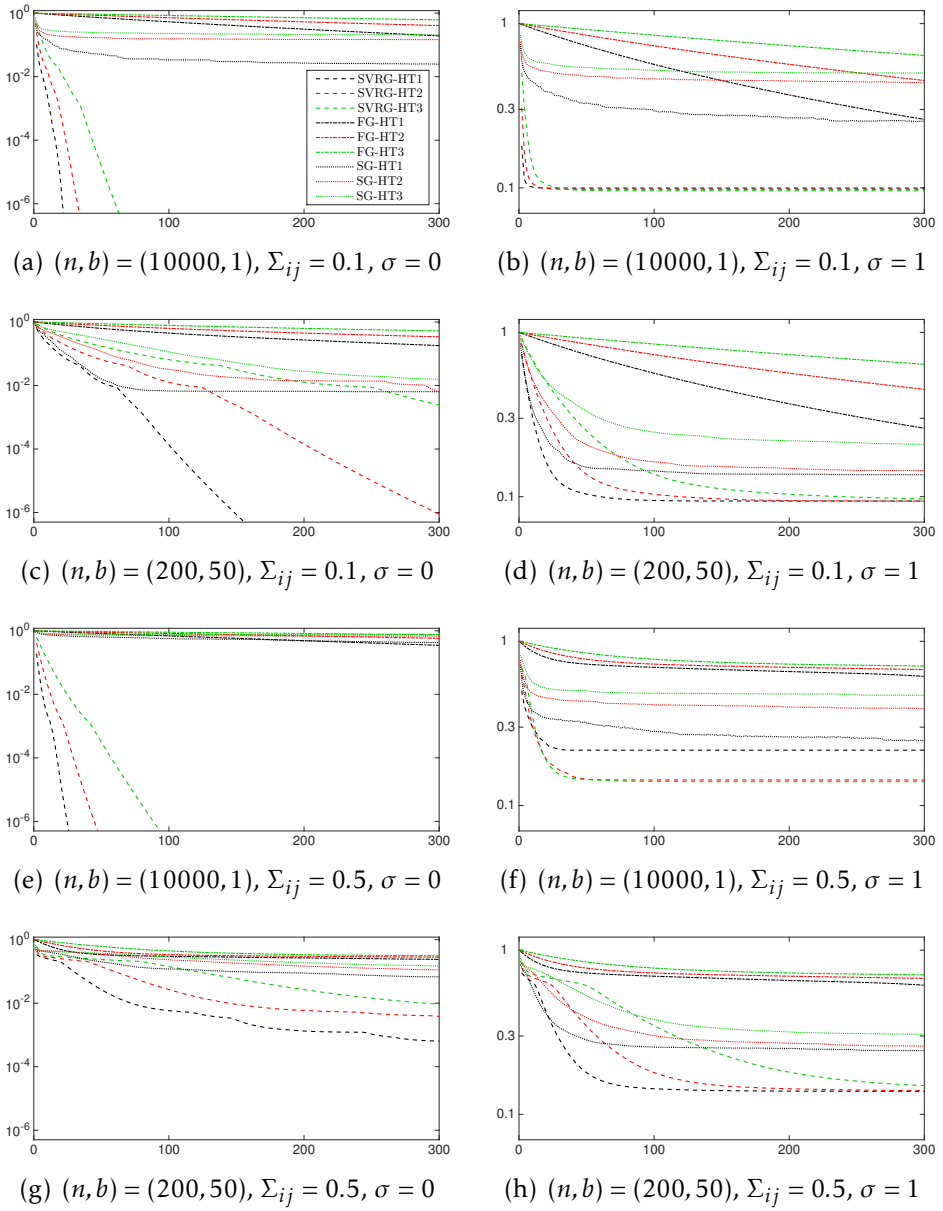
3.5.1 Synthetic Data

We consider a sparse linear regression problem. We generate each row of the design matrix A_{i*}, $i \in [nb]$, independently from a d-dimensional Gaussian distribution with mean 0 and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. The response vector is generated from the linear model $y = A\theta^* + z \in \mathbb{R}^{nb}$, where $\theta^* \in \mathbb{R}^d$ is the k^*-sparse regression coefficient vector, and z is generated from an n-dimensional Gaussian distribution with mean 0 and covariance matrix $\sigma^2 I$. We set $nb = 10000$, $d = 25000$, $k^* = 200$

and $k = 500$. For Σ, we set $\Sigma_{ii} = 1$ and $\Sigma_{ij} = c$ for some constant $c \in (0,1)$ for all $i \neq j$. The nonzero entries in θ^* are sampled independently from a uniform distribution over the interval $(-2,+2)$. We divide 10000 samples into n mini batches, and each mini batch contains $b = 10000/n$ samples.

Figure 3.1 illustrates the computational performance of FG-HT, SG-HT, and SVRG-HT for eight different settings of (n,b) and Σ_{ij}, each with step sizes $\eta = 1/256, 1/512$, and $1/1024$. The first four settings are noiseless, i.e., $\sigma = 0$ with (1) $(n,b) = (10000,1)$, $\Sigma_{ij} = 0.1$; (2) $(n,b) = (10000,1)$, $\Sigma_{ij} = 0.5$; (3) $(n,b) = (200,50)$, $\Sigma_{ij} = 0.1$; (4) $(n,b) = (200,50)$, $\Sigma_{ij} = 0.5$. For simplicity, we choose the update frequency of the inner loop as $m = n$ throughout our experiments[2]. The last four settings are noisy with $\sigma = 1$ and identical choices of (n,b), Σ_{ij} and m as in (1)-(4). For all algorithms, we plot the objective values averaged over 50 different runs. The horizontal axis corresponds to the number of passes over the entire dataset; computing a full gradient is counted as 1 pass, while computing a stochastic gradient is counted as $1/n$-th of a pass. The vertical axis corresponds to the ratio of current objective value over the objective value using $\widetilde{\theta}^{(0)} = 0$. We further provide the optimal relative estimation error $\|\widetilde{\theta}^{(10^6)} - \theta^*\|_2 / \|\theta^*\|_2$ after 10^6 effective passes of the entire dataset for all settings of the three algorithms in Table 3.1. The estimation error is obtained by averaging over 50 different runs, each of which is chosen from a sequence of step sizes $\eta \in \{1/2^5, 1/2^6, \ldots, 1/2^{14}\}$.

[2]Larger m results in increasing number of effective passes of the entire dataset required to achieve the same decrease of objective values, which is also observed in Prox-SVRG [57]



(a) $(n, b) = (10000, 1)$, $\Sigma_{ij} = 0.1$, $\sigma = 0$ (b) $(n, b) = (10000, 1)$, $\Sigma_{ij} = 0.1$, $\sigma = 1$

(c) $(n, b) = (200, 50)$, $\Sigma_{ij} = 0.1$, $\sigma = 0$ (d) $(n, b) = (200, 50)$, $\Sigma_{ij} = 0.1$, $\sigma = 1$

(e) $(n, b) = (10000, 1)$, $\Sigma_{ij} = 0.5$, $\sigma = 0$ (f) $(n, b) = (10000, 1)$, $\Sigma_{ij} = 0.5$, $\sigma = 1$

(g) $(n, b) = (200, 50)$, $\Sigma_{ij} = 0.5$, $\sigma = 0$ (h) $(n, b) = (200, 50)$, $\Sigma_{ij} = 0.5$, $\sigma = 1$

Figure 3.1: Comparison among the three algorithms in all settings on the simulated data. The horizontal axis corresponds to the number of passes over the entire dataset. The vertical axis corresponds to the ratio of current objective value over the objective value using $\widetilde{\theta}^{(0)} = 0$. For each algorithm, option 1, 2 and 3 correspond to the step sizes $\eta = 1/256, 1/512$, and $1/1024$ respectively. It is evident from the plots that SVRG-HT outperforms the other competitors in terms of the convergence rate over all settings.

Table 3.1: Comparison of optimal relative estimation errors among the three algorithms in all settings on the simulated data. We denote $(n,b)_1 = (10000,1)$ and $(n,b)_2 = (200,50)$. SVRG-HT achieves comparable result with FG-HT, both of which outperforms SG-HT over all settings.

Method	$\sigma = 0$				$\sigma = 1$			
	$\Sigma_{ij} = 0.1$		$\Sigma_{ij} = 0.5$		$\Sigma_{ij} = 0.1$		$\Sigma_{ij} = 0.5$	
	$(n,b)_1$	$(n,b)_2$	$(n,b)_1$	$(n,b)_2$	$(n,b)_1$	$(n,b)_2$	$(n,b)_1$	$(n,b)_2$
FG-HT	$< 10^{-20}$		$< 10^{-20}$		0.00851		0.02940	
SG-HT	$< 10^{-20}$	$< 10^{-20}$	$< 10^{-20}$	0.13885	0.02490	0.06412	0.21676	0.18764
SVRG-HT	$< 10^{-20}$	$< 10^{-20}$	$< 10^{-20}$	$< 10^{-20}$	0.00968	0.00970	0.02614	0.02823

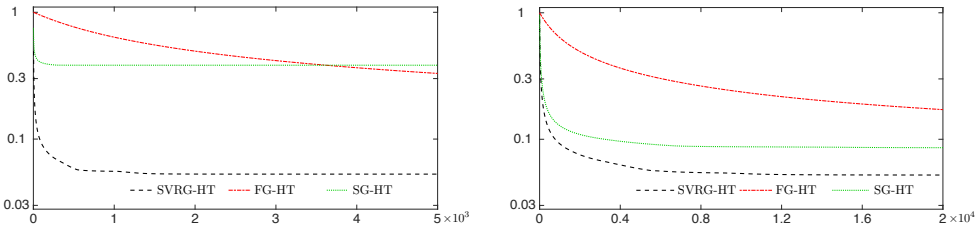
We see from Figure 3.1 that SVRG-HT outperforms the other competitors in terms of the convergence rate in all settings. While FG-HT also enjoys linear converge guarantees, its computational cost at each iteration is n times larger than that of SVRG-HT. Consequently, its performance is much worse than that of SVRG-HT. Besides, we also see that SG-HT converges slower than SVRG-HT in all settings. This is because the largest eigenvalue of any 500 by 500 submatrix of the covariance matrix is large (larger than 50 or 250) such that the underlying design matrix violates the Restricted Isometry Property (RIP) required by SG-HT. On the other hand, Table 3.1 indicates that the optimal estimation error of SVRG-HT is comparable to FG-HT, both of which outperform SG-HT, especially in noisy settings. It is important to note that with the optimal step size, the estimation of FG-HT usually becomes stable after $> 10^5$ passes, while the estimation of SVRG-HT usually becomes stable within a few dozen to a few hundred passes, which validates the significant improvement of SVRG-HT over FG-HT in terms of the computational cost.

3.5.2 Real Data

We adopt a subset of RCV1 dataset with 9625 documents and 29992 distinct words, including the classes of "C15", "ECAT", "GCAT", and "MCAT" [69]. We apply logistic regression to perform a binary classification for all classes, each of which uses 5000 documents for training, i.e., $nb = 5000$ and $d = 29992$, with the same proportion of documents from each class, and the rest for testing. We illustrate the computational performance of FG-HT, SG-HT, and SVRG-HT in two different settings: Setting (1) has $(n, b) = (5000, 1)$; Setting (2) has $(n, b) = (100, 50)$. We choose $k = 200$ and $m = n$ for both settings. For all three algorithms, we plot their objective values and provide the optimal classification errors averaged over 10 different runs using random data separations. Figure 3.2 demonstrates the computational performance for "C15" on the training dataset, and the other classes have similar performance. The horizontal axis corresponds to the number of passes over the entire training dataset. The vertical axis corresponds to the ratio of current objective value over the initial objective value using $\widetilde{\theta}^{(0)} = 0$. Similar to the synthetic data, SVRG-HT outperforms the other competitors in terms of the convergence rate in both settings.

We further provide the optimal misclassification rates of all classes for the three algorithms in Table 3.2, where the optimal step size η for each algorithm is chosen from a sequence of values $\{1/2^5, 1/2^6, \ldots, 1/2^{14}\}$. Similar to the synthetic data again, the optimal misclassification rate of SVRG-HT is comparable to FG-HT, both

of which outperform SG-HT. The estimation of FG-HT generally requires $> 10^6$ passes to become stable, while the estimation of SVRG-HT generally requires a few hundred to a few thousand passes to be stable, which validates the significant improvement of SVRG-HT over FG-HT on this real dataset in terms of the computational cost.



(a) $(n, b) = (5000, 1)$ (b) $(n, b) = (100, 50)$

Figure 3.2: Comparison among the three algorithms in two different settings on the training dataset of RCV1 for the class "C15". The horizontal axis corresponds to the number of passes over the entire training dataset. The vertical axis corresponds to the ratio of current objective value over the initial objective. It is evident from the plots that SVRG-HT outperforms the other competitors in both settings.

Table 3.2: Comparison of optimal classification errors on the test dataset of RCV1 among the three algorithms for both settings and all four classes. We denote $(n, b)_1 = (5000, 1)$ and $(n, b)_2 = (100, 50)$. SVRG-HT achieves comparable result with FG-HT, both of which outperform SG-HT over all settings.

	C15		ECAT		GCAT		MCAT	
	$(n, b)_1$	$(n, b)_2$	$(n, b)_1$	$(n, b)_2$	$(n, b)_1$	$(n, b)_2$	$(n, b)_1$	$(n, b)_2$
FG-HT	0.02844		0.05581		0.03028		0.05703	
SG-HT	0.03259	0.03361	0.06851	0.07179	0.06263	0.09142	0.07638	0.08228
SVRG-HT	0.02826	0.02867	0.05628	0.05631	0.03354	0.03444	0.05877	0.05927

3.5.3 ℓ_0-Norm/SVRG-HT vs. ℓ_1-Norm/Prox-SVRG

We further discuss the empirical performance of sparsity induced problems using the ℓ_0-norm and the ℓ_1-norm respectively. Specifically, we consider the sparse linear regression problem (3.2.2) for the ℓ_0-constrained problem and the following ℓ_1-regularized problem,

$$\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{b}\|y_{\mathcal{S}_i} - A_{\mathcal{S}_i *}\theta\|_2^2 + \lambda\|\theta\|_1, \tag{3.5.1}$$

where $\lambda > 0$ is a regularization parameter. The ℓ_0-constrained problem (3.2.2) is solved by SVRG-HT, and the ℓ_1-regularized problem (3.5.1) is solved by Prox-SVRG [57]. We follow the same settings as in Section 3.5.1 for data generation and the choice of parameters for SVRG-HT. For the ℓ_1-regularized problem (3.5.1), we choose an optimal regularization parameter λ from a sequence of values $\{1/2^2, 1/2^4, 1/2^6, \ldots, 1/2^{20}\}$, which returns the optimal relative estimation error $\|\widetilde{\theta}^{(10^6)} - \theta^*\|_2/\|\theta^*\|_2$.

Table 3.3: Comparison of optimal relative estimation errors between (3.2.2) and (3.5.1) in all settings on the synthetic data. We denote $(n,b)_1 = (10000, 1)$ and $(n,b)_2 = (200, 50)$.

Method	$\sigma = 0$				$\sigma = 1$			
	$\Sigma_{ij} = 0.1$		$\Sigma_{ij} = 0.5$		$\Sigma_{ij} = 0.1$		$\Sigma_{ij} = 0.5$	
	$(n,b)_1$	$(n,b)_2$	$(n,b)_1$	$(n,b)_2$	$(n,b)_1$	$(n,b)_2$	$(n,b)_1$	$(n,b)_2$
ℓ_0-norm	$< 10^{-20}$	$< 10^{-20}$	$< 10^{-20}$	$< 10^{-20}$	0.00968	0.00970	0.02614	0.02823
ℓ_1-norm	$\approx 10^{-6}$	$\approx 10^{-7}$	$\approx 10^{-6}$	$\approx 10^{-7}$	0.01715	0.01306	0.08475	0.08177

Table 3.3 provides the optimal estimation errors in all settings, each of which is averaged over 50 different runs. We observe that the ℓ_0-norm problem uniformly

outperforms the ℓ_1-norm problem in terms of statistical accuracy. Besides, it is important to note that we only need to tune the step size η for the ℓ_0-norm problem (3.2.2), which is insensitive in different settings, and the sparsity parameter k is fixed throughout. On the other hand, for the ℓ_1-norm problem (3.5.1), we need to tune both the step size η and the regularization parameter λ to obtain the optimal estimation, which require much more tuning efforts. Moreover, we observe that SVRG-HT converges faster than Prox-SVRG, where SVRG-HT typically requires a few dozen to a few hundred passes of data to converge. This is because SVRG-HT always guarantees the solution sparsity, and the restricted strong convexity enables the fast convergence. In contrast, Prox-SVRG requires a few thousand passes of data to converge, because Prox-SVRG often yields dense solutions, especially at the first few iterations.

3.6 Discussion

We provide a summary of comparison between our proposed algorithm SVRG-HT with FG-HT [1] and SG-HT [2] in Table 3.4. We want to remark that though the computational complexity of SG-HT may seem lower than SVRG-HT, the RSC and RSS conditions of SG-HT are very restrictive, and it generally converges much slower than SVRG-HT in practice.

SVRG-HT is closely related to some recent work on stochastic optimization al-

Table 3.4: Comparison with FG-HT [1] and SG-HT [2]. Our contributions are manifold: (1) less restrictive assumptions on the RSC and RSS conditions than SG-HT; (2) improving the iteration complexity and computational complexity over FG-HT; and (3) improving the statistical performance over SG-HT. We only provide the statistical error of sparse linear regression for illustration.

Method	Restrictions on κ_s	Ite. Complexity	Comp. Complexity	Statistical Error
FG-HT	No: κ_s bounded	$\mathcal{O}(\kappa_s \log(1/\varepsilon))$	$\mathcal{O}(n\kappa_s \cdot \log(1/\varepsilon))$	$\mathcal{O}\left(\sigma\sqrt{k^* \log d/(nb)}\right)$
SG-HT	Yes: $\kappa_s \leq \frac{4}{3}$	$\mathcal{O}(\log(1/\varepsilon))$	$\mathcal{O}(\log(1/\varepsilon))$	$\mathcal{O}\left(\sigma\sqrt{k^* \log d/b}\right)$
SVRG-HT	No: κ_s bounded	$\mathcal{O}(\log(1/\varepsilon))$	$\mathcal{O}([n+\kappa_s] \cdot \log(1/\varepsilon))$	$\mathcal{O}\left(\sigma\sqrt{k^* \log d/(nb)}\right)$

gorithms, including Prox-SVRG [57], stochastic averaging gradient (SAG) [58] and stochastic dual coordinate ascent (SDCA, [59]). However, the focus in these previous works has been on establishing global linear convergence for optimization problems involving strongly convex objective with a convex constraint, whereas SVRG-HT guarantees linear convergence for optimization problems involving a non-strongly convex objective with nonconvex cardinality constraint.

Other related work includes nonconvex regularized M-estimators proposed in [32]. In particular, the following nonconvex optimization problem is considered in [32]:

$$\min_{\theta} \ \mathcal{F}(\theta) + \mathcal{P}_{\lambda,\gamma}(\theta) \quad \text{subject to } \|\theta\|_1 \leq R, \tag{3.6.1}$$

where $\mathcal{P}_{\lambda,\gamma}(\theta)$ is a nonconvex regularization function with tuning parameters λ and γ; Popular choices for $\mathcal{P}_{\lambda,\gamma}(\theta)$ are the SCAD and MCP regularization functions studied in [19, 20]. It is shown in [32] that under restricted strong convexity and

restricted strong smoothness conditions, similar to those studied here, the proximal gradient descent attains linear convergence to approximate global optima with optimal estimation accuracy. Accordingly, one could adopt the Prox-SVRG to solve (3.6.1) in a stochastic fashion, and trim the analyses in [57] and [32] to establish similar convergence guarantees. We remark, however, that Problem (3.6.1) involves three tuning parameters, λ, γ, and R which, in practice, requires a large amount of tuning effort to attain good empirical performance. In contrast, Problem (4.3.1) involves a single tuning parameter, k, which makes tuning more efficient.

3.7 Proofs of Main Results

We present the proofs of our main theoretical results in this section.

3.7.1 Proof of Theorem 3.3.7

Part 1: We first demonstrate (3.3.6) and (3.3.7). let $v = \theta^{(t)} - \eta g_{\mathcal{I}}^{(t)}(\theta^{(t)})$ and $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^{(t)} \cup \mathcal{I}^{(t+1)}$, where $\mathcal{I}^* = \text{supp}(\theta^*)$, $\mathcal{I}^{(t)} = \text{supp}(\theta^{(t)})$ and $\mathcal{I}^{(t+1)} = \text{supp}(\theta^{(t+1)})$.

Conditioning on $\theta^{(t)}$, we have the following expectation:

$$
\begin{aligned}
\mathbb{E}\|v-\theta^*\|_2^2 &= \mathbb{E}\|\theta^{(t)}-\eta g_{\mathcal{I}}^{(t)}(\theta^{(t)})-\theta^*\|_2^2 \\
&= \mathbb{E}\|\theta^{(t)}-\theta^*\|_2^2+\eta^2\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2-2\eta\langle\theta^{(t)}-\theta^*,\mathbb{E}g_{\mathcal{I}}^{(t)}(\theta^{(t)})\rangle \\
&= \mathbb{E}\|\theta^{(t)}-\theta^*\|_2^2+\eta^2\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2-2\eta\langle\theta^{(t)}-\theta^*,\nabla_{\mathcal{I}}\mathcal{F}(\theta^{(t)})\rangle \\
&\le \mathbb{E}\|\theta^{(t)}-\theta^*\|_2^2+\eta^2\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2-2\eta\Big[\mathcal{F}(\theta^{(t)})-\mathcal{F}(\theta^*)\Big] \\
&\le \mathbb{E}\|\theta^{(t)}-\theta^*\|_2^2-2\eta\Big[\mathcal{F}(\theta^{(t)})-\mathcal{F}(\theta^*)\Big] \\
&\qquad +12\eta^2\rho_s^+\Big[\mathcal{F}(\theta^{(t)})-\mathcal{F}(\theta^*)+\mathcal{F}(\widetilde{\theta})-\mathcal{F}(\theta^*)\Big]+3\eta^2\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2 \\
&= \mathbb{E}\|\theta^{(t)}-\theta^*\|_2^2-2\eta(1-6\eta\rho_s^+)\Big[\mathcal{F}(\theta^{(t)})-\mathcal{F}(\theta^*)\Big] \\
&\qquad +12\eta^2\rho_s^+\Big[\mathcal{F}(\widetilde{\theta})-\mathcal{F}(\theta^*)\Big]+3\eta^2\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2, \qquad (3.7.1)
\end{aligned}
$$

where the first inequality follows from the convexity of $\mathcal{F}(\theta)$ and the second inequality follows from Lemma 3.3.5.

Since $\theta^{(t+1)}=\overline{\theta}_k^{(t+1)}=v_k$, i.e. $\theta^{(t+1)}$ is the best k-sparse approximation of v, then we have the following from Lemma 3.3.3

$$
\|\theta^{(t+1)}-\theta^*\|_2^2\le\left(1+\frac{2\sqrt{k^*}}{\sqrt{k-k^*}}\right)\cdot\|v-\theta^*\|_2^2. \qquad (3.7.2)
$$

Let $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$. Combining (3.7.1) and (3.7.2), we have

$$\mathbb{E}\|\theta^{(t+1)} - \theta^*\|_2^2 \leq \alpha\mathbb{E}\|\theta^{(t)} - \theta^*\|_2^2 - 2\alpha\eta(1 - 6\eta\rho_s^+)\Big[\mathcal{F}(\theta^{(t)}) - \mathcal{F}(\theta^*)\Big] + 12\alpha\eta^2\rho_s^+\Big[\mathcal{F}(\widetilde{\theta}) - \mathcal{F}(\theta^*)\Big] + 3\alpha\eta^2\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2. \quad (3.7.3)$$

Notice that $\widetilde{\theta} = \theta^{(0)} = \widetilde{\theta}^{(r-1)}$. By summing (3.7.3) over $t = 0, 1, \ldots, m-1$ and taking expectation with respect to all t's, we have

$$\begin{aligned}
&\mathbb{E}\|\theta^{(m)} - \theta^*\|_2^2 + \frac{2\eta(1 - 6\eta\rho_s^+)(\alpha^m - 1)}{\alpha - 1}\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\Big] \\
&\leq \alpha^m\mathbb{E}\|\widetilde{\theta}^{(r-1)} - \theta^*\|_2^2 + \frac{12\eta^2\rho_s^+(\alpha^m - 1)}{\alpha - 1}\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)\Big] + \frac{3\eta^2(\alpha^m - 1)}{\alpha - 1}\mathbb{E}\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2 \\
&\leq \frac{2\alpha^m}{\rho_s^-}\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)\Big] + \frac{12\eta^2\rho_s^+(\alpha^m - 1)}{\alpha - 1}\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)\Big] \\
&\qquad + \frac{3\eta^2(\alpha^m - 1)}{\alpha - 1}\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2^2, \qquad (3.7.4)
\end{aligned}$$

where the last inequality follows from the RSC condition (3.3.1) and the definition of $\widetilde{\mathcal{I}}$. It further follows from (3.7.4)

$$\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\Big] \leq \left(\frac{\alpha^m(\alpha - 1)}{\eta\rho_s^-(1 - 6\eta\rho_s^+)(\alpha^m - 1)} + \frac{6\eta\rho_s^+}{1 - 6\eta\rho_s^+}\right)\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)\Big] + \frac{3\eta}{2(1 - 6\eta\rho_s^+)}\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2^2. \quad (3.7.5)$$

Let $\beta = \frac{\alpha^m(\alpha-1)}{\eta\rho_s^-(1-6\eta\rho_s^+)(\alpha^m-1)} + \frac{6\eta\rho_s^+}{1-6\eta\rho_s^+}$ and apply (3.7.5) recursively, then we have the desired bound (3.3.6) when $\beta \leq \frac{3}{4}$.

We then demonstrate (3.3.7). The RSC condition implies

$$\mathcal{F}(\theta^*) \leq \mathcal{F}(\widetilde{\theta}^{(r)}) + \langle \nabla \mathcal{F}(\theta^*), \theta^* - \widetilde{\theta}^{(r)} \rangle - \frac{\rho_s^-}{2} \|\widetilde{\theta}^{(r)} - \theta^*\|_2^2. \tag{3.7.6}$$

Let $\zeta = \left(\frac{5}{6}\right)^r \left[\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)\right] + \frac{6\eta}{(1-6\eta\rho_s^+)} \|\nabla_{\widetilde{\mathcal{I}}} \mathcal{F}(\theta^*)\|_2^2$. Combining (3.3.6) and (3.7.6), we have

$$\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r)}) - \zeta\Big] \leq \mathcal{F}(\theta^*) \leq \mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r)}) + \langle \nabla \mathcal{F}(\theta^*), \theta^* - \widetilde{\theta}^{(r)} \rangle - \frac{\rho_s^-}{2} \|\widetilde{\theta}^{(r)} - \theta^*\|_2^2\Big]. \tag{3.7.7}$$

Using the duality of norms, we have

$$\begin{aligned} \mathbb{E}\langle \nabla \mathcal{F}(\theta^*), \theta^* - \widetilde{\theta}^{(r)} \rangle &\leq \|\nabla \mathcal{F}(\theta^*)\|_\infty \mathbb{E}\|\widetilde{\theta}^{(r)} - \theta^*\|_1 \\ &\leq \sqrt{s} \|\nabla \mathcal{F}(\theta^*)\|_\infty \mathbb{E}\|\widetilde{\theta}^{(r)} - \theta^*\|_2. \end{aligned} \tag{3.7.8}$$

Combining (3.7.7), (3.7.8), and $(\mathbb{E}[x])^2 \leq \mathbb{E}[x^2]$, we have

$$\frac{\rho_s^-}{2} (\mathbb{E}\|\widetilde{\theta}^{(r)} - \theta^*\|_2)^2 \leq \sqrt{s} \|\nabla \mathcal{F}(\theta^*)\|_\infty \mathbb{E}\|\widetilde{\theta}^{(r)} - \theta^*\|_2 + \zeta. \tag{3.7.9}$$

Let $a = \mathbb{E}\|\widetilde{\theta}^{(r)} - \theta^*\|_2$, then (3.7.9) is equivalent to solving the following quadratic function of a:

$$\frac{\rho_s^-}{2} a^2 - \sqrt{s} \|\nabla \mathcal{F}(\theta^*)\|_\infty a - \zeta \leq 0,$$

which yields the bound (3.3.7).

Now we show that with k, η and m specified in the theorem, we guarantee $\beta \leq \frac{3}{4}$. More specifically, let $\eta \leq \frac{C_3}{\rho_s^+} \leq \frac{1}{18\rho_s^+}$, then we have

$$\frac{6\eta\rho_s^+}{1-6\eta\rho_s^+} \leq \frac{6C_3}{1-6C_3} \leq \frac{1}{2}.$$

If $k \geq C_1\kappa_s^2 k^*$ and $\eta \geq \frac{C_2}{\rho_s^+}$ with $C_2 \leq C_3$, then we have $\alpha \leq 1 + \frac{2}{\sqrt{C_1-1}\cdot\kappa_s}$ and

$$\begin{aligned}\frac{\alpha^m(\alpha-1)}{\eta\rho_s^-(1-6\eta\rho_s^+)(\alpha^m-1)} &\leq \frac{\frac{2}{\sqrt{C_1-1}\cdot\kappa_s}}{\frac{2C_2}{3\kappa_s}\left(1-(1+\frac{2}{\sqrt{C_1-1}\cdot\kappa_s})^{-m}\right)} \\ &= \frac{3}{C_2\sqrt{C_1-1}\left(1-(1+\frac{2}{\sqrt{C_1-1}\cdot\kappa_s})^{-m}\right)}. \end{aligned} \tag{3.7.10}$$

Then (3.7.10) is guaranteed to be strictly smaller than $\frac{1}{2}$ if we have

$$m \geq \log_{1+\frac{2}{\sqrt{C_1-1}\cdot\kappa_s}} \frac{C_2\sqrt{C_1-1}}{C_2\sqrt{C_1-1}-6}. \tag{3.7.11}$$

Using the the fact that $\ln(1+x) > x/2$ for $x \in (0,1)$, it follows that

$$\log_{1+\frac{2}{\sqrt{C_1-1}\cdot\kappa_s}} \frac{C_2\sqrt{C_1-1}}{C_2\sqrt{C_1-1}-6} = \frac{\log\frac{C_2\sqrt{C_1-1}}{C_2\sqrt{C_1-1}-6}}{\log 1+\frac{2}{\sqrt{C_1-1}\cdot\kappa_s}} \leq \log\frac{C_2\sqrt{C_1-1}}{C_2\sqrt{C_1-1}-6}\cdot\sqrt{C_1-1}\cdot\kappa_s.$$

Then (3.7.11) holds if m satisfies

$$m \geq \log \frac{C_2\sqrt{C_1-1}}{C_2\sqrt{C_1-1}-6} \cdot \sqrt{C_1-1} \cdot \kappa_s$$

If we choose $C_1 = 161^2$, $C_2 = \frac{1}{20}$, $C_3 = \frac{1}{18}$ and $C_4 = 222$, then we have $\beta \leq \frac{3}{4}$.

Part 2: Next, we demonstrate (3.3.8) and (3.3.9). It follows from (3.3.6)

$$\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\Big] - \frac{6\eta}{(1-6\eta\rho_s^+)}\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2^2 \leq \left(\frac{3}{4}\right)^r \Big[\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)\Big]. \tag{3.7.12}$$

Let $\xi_1, \xi_2, \xi_3, \ldots$ be a non-negative sequence of random variables, which is defined as

$$\xi_r \triangleq \max\left\{\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*) - \frac{6\eta}{(1-6\eta\rho_s^+)}\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2^2,\ 0\right\}.$$

For a fixed $\varepsilon > 0$, it follows from Markov's Inequality and (3.7.12)

$$\mathbb{P}(\xi_r \geq \varepsilon) \leq \frac{\mathbb{E}\xi_r}{\varepsilon} \leq \frac{\left(\frac{3}{4}\right)^r \Big[\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)\Big]}{\varepsilon}. \tag{3.7.13}$$

Given $\delta \in (0,1)$, let the R.H.S. of (3.7.13) be no greater than δ, which requires

$$r \geq \log_{\left(\frac{3}{4}\right)^{-1}} \frac{\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)}{\varepsilon\delta}.$$

Therefore, we have that if $r = \left\lceil 4\log\left(\frac{\mathcal{F}(\widetilde{\theta}^{(0)})-\mathcal{F}(\theta^*)}{\varepsilon\delta}\right)\right\rceil$, then (3.3.8) holds with probability at least $1-\delta$. Finally, (3.3.9) holds by combining (3.3.7) and (3.3.8).

3.7.2 Proof of Corollary 3.3.10

For sparse linear model, we have $\nabla\mathcal{F}(\theta^*) = A^\top z/(nb)$. Since z has i.i.d. $N(0,\sigma^2)$ entries, then $A_{*j}^\top z/(nb) \sim N(0,\sigma^2\|A_{*j}\|_2^2/(nb)^2)$ for any $j \in [d]$. Using Mill's Inequality for tail bounds of the normal distribution (Theorem 4.7 in [70]), we have

$$\mathbb{P}\left(\left|\frac{A_{*j}^\top z}{nb}\right| > 2\sigma\sqrt{\frac{\log d}{nb}}\right) = \mathbb{P}\left(\left|\frac{A_{*j}^\top z}{\sigma\|A_{*j}\|_2}\right| > 2\frac{\sqrt{nb\log d}}{\|A_{*j}\|_2}\right) \leq \|A_{*j}\|_2\sqrt{\frac{1}{2\pi nb\log d}}\exp\left(-4\frac{nb\log d}{\|A_{*j}\|_2^2}\right).$$

Using union bound and the assumption $\frac{\max_j\|A_{*j}\|_2}{\sqrt{nb}} \leq 1$, this implies

$$\mathbb{P}\left(\left\|\frac{A_{*j}^\top z}{nb}\right\|_\infty > 2\sigma\sqrt{\frac{\log d}{nb}}\right) \leq \frac{d^{-4}}{\sqrt{2\pi\log d}}.$$

Then with probability at least $1 - \frac{1}{\sqrt{2\pi\log d}}\cdot d^{-4}$, we have

$$\|\nabla\mathcal{F}(\theta^*)\|_\infty \leq \left\|\frac{A^\top z}{nb}\right\|_\infty \leq 2\sigma\sqrt{\frac{\log d}{nb}}. \tag{3.7.14}$$

Conditioning on (3.7.14), we have

$$\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2^2 \leq s\|\nabla\mathcal{F}(\theta^*)\|_\infty^2 \leq \frac{4\sigma^2 s\log d}{nb}. \tag{3.7.15}$$

We have from Lemma 3.3.9 that $s = 2k+k^* = (2C_5+1)k^*$ for some constant C_5 when n and b are large enough. Given $\varepsilon > 0$ and $\delta \in (0,1)$, if

$$r \geq 4\log\left(\frac{\mathcal{F}(\widetilde{\theta}^{(0)}) - \mathcal{F}(\theta^*)}{\varepsilon\delta}\right),$$

and $\varepsilon = \mathcal{O}\left(\sigma\sqrt{\frac{k^*\log d}{nb}}\right)$, then with probability at least $1-\delta-\frac{1}{\sqrt{2\pi\log d}}\cdot d^{-4}$, we have from (3.3.9), (3.7.14), and (3.7.15)

$$\|\widetilde{\theta}^{(r)} - \theta^*\|_2 \leq c_3\sigma\sqrt{\frac{k^*\log d}{nb}}, \tag{3.7.16}$$

where c_3 is a constant. This completes the proof.

3.7.3 Proof of Corollary 3.3.14

The only difference between the proof of Corollary 3.3.14 and the proof of Corollary 3.3.10 is the upper bounds of $\|\nabla\mathcal{F}(\theta^*)\|_\infty$ and $\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2^2$. When $\{A_i*\}_{i=1}^{nb}$ are independent sub-Gaussian vectors, it follows from [32] that $\mathcal{F}(\theta)$ and $\{f_i(\theta)\}_{i=1}^n$ satisfy (3.3.13). Besides, there exist constants c_4, c_5, and c_6, such that with probability at least $1-c_4d^{-c_5}$, we have

$$\|\nabla\mathcal{F}(\theta^*)\|_\infty \leq c_6\sqrt{\frac{\log d}{nb}}. \tag{3.7.17}$$

Conditioning on (3.7.17), we have

$$\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2^2 \leq s\|\nabla\mathcal{F}(\theta^*)\|_\infty^2 \leq \frac{c_6^2 s\log d}{nb}. \tag{3.7.18}$$

The rest of the proof follows immediately from the proof of Corollary 3.3.10.

3.7.4 Proof of Theorem 3.4.1

Recall from (3.3.4) that $g^{(t)}(\theta^{(t)}) = \nabla f_{i_t}(\theta^{(t)}) - \nabla f_{i_t}(\widetilde{\theta}) + \nabla\mathcal{F}(\widetilde{\theta})$. We also denote $u = \theta^{(t)} - \eta h_{\mathcal{I}}^{(t)}(\theta^{(t)})$, where $h^{(t)}(\theta^{(t)}) = \nabla f_{i_t}(\theta^{(t')}) - \nabla f_{i_t}(\widetilde{\theta}) + \nabla\mathcal{F}(\widetilde{\theta})$ and t' is the actual evaluation used at the t-th iteration. Then we have

$$\begin{aligned}\mathbb{E}\|u-\theta^*\|_2^2 &= \mathbb{E}\|\theta^{(t)} - \eta h_{\mathcal{I}}^{(t)}(\theta^{(t)}) - \theta^*\|_2^2 \\ &= \mathbb{E}\Big[\|\theta^{(t)}-\theta^*\|_2^2 + \eta^2\|h_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2 - 2\eta\langle\theta^{(t)}-\theta^*, h_{\mathcal{I}}^{(t)}(\theta^{(t)})\rangle\Big]\end{aligned} \tag{3.7.19}$$

We first bound $\mathbb{E}\|h_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2$ in terms of $\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2$ as

$$\begin{aligned}\mathbb{E}\|h_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2 &\leq 2\mathbb{E}\Big[\|h_{\mathcal{I}}^{(t)}(\theta^{(t)}) - g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2 + \|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2\Big] \\ &= 2\mathbb{E}\Big[\|\nabla_{\mathcal{I}} f_{i_t}(\theta^{(t)}) - \nabla_{\mathcal{I}} f_{i_t}(\theta^{(t')})\|_2^2 + \|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2\Big] \\ &\leq 2(\rho_s^+)^2\varsigma\sum_{j=t'}^{t-1}\mathbb{E}\|\theta_{e_t}^{(j+1)} - \theta_{e_t}^{(j)}\|^2 + 2\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2 \\ &\leq 2(\rho_s^+)^2\Delta\varsigma\eta^2\sum_{j=t'}^{t-1}\mathbb{E}\|h_{\mathcal{I}}^{(j)}(\theta^{(j)})\|^2 + 2\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2,\end{aligned}$$

where the first inequality is from $\|a\|_2^2 \leq 2\|a-b\|_2^2 + 2\|b\|_2^2$ for any vector a and b, the second inequality is from the definition of ς, triangle inequality, and $\|f_i(\theta) - f_i(\theta')\|_2 \leq \rho_s^+ \|\theta - \theta'\|_2$ implied by the RSS condition [71], and the last inequality is from the definition of Δ. Take the summation of the inequality above from $t = 0$ to $m-1$, we have

$$\begin{aligned}\sum_{t=0}^{m-1} \mathbb{E}\|h_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2 &\leq \sum_{t=0}^{m-1}\left[2(\rho_s^+)^2\Delta\varsigma\eta^2\sum_{j=t'}^{t-1}\mathbb{E}\|h_{\mathcal{I}}^{(j)}(\theta^{(j)})\|^2 + 2\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2\right] \\ &\leq 2(\rho_s^+)^2\Delta\varsigma^2\eta^2\sum_{t=0}^{m-1}\mathbb{E}\|h_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2 + \sum_{t=0}^{m-1}\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2,\end{aligned}$$

where the second inequality is from the definition of ς. The inequality above implies

$$\sum_{t=0}^{m-1}\mathbb{E}\|h_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2 \leq \frac{2}{1-2\rho_s^{+2}\Delta\varsigma^2\eta^2}\sum_{t=0}^{m-1}\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2. \tag{3.7.20}$$

Next, we bound $\mathbb{E}\langle\theta^{(t)} - \theta^*, h_{\mathcal{I}}^{(t)}(\theta^{(t)})\rangle$. This can be written as

$$\begin{aligned}\mathbb{E}\langle\theta^* - \theta^{(t)}, h_{\mathcal{I}}^{(t)}(\theta^{(t)})\rangle &= \mathbb{E}\langle\theta^* - \theta^{(t)}, \nabla_{\mathcal{I}} f_{i_t}(\theta^{(t')})\rangle \\ &= \mathbb{E}\langle\theta^* - \theta^{(t')}, \nabla_{\mathcal{I}} f_{i_t}(\theta^{(t')})\rangle + \sum_{j=t'}^{t-1}\mathbb{E}\langle\theta^{(j)} - \theta^{(j+1)}, \nabla_{\mathcal{I}} f_{i_t}(\theta^{(j)})\rangle \\ &\quad + \sum_{j=t'}^{t-1}\mathbb{E}\langle\theta^{(j)} - \theta^{(j+1)}, \nabla_{\mathcal{I}} f_{i_t}(\theta^{(t')}) - \nabla_{\mathcal{I}} f_{i_t}(\theta^{(j)})\rangle. \end{aligned}\tag{3.7.21}$$

From the convexity of f_{i_t}, we have

$$\mathbb{E}\langle \theta^* - \theta^{(t')}, \nabla_{\mathcal{I}} f_{i_t}(\theta^{(t')})\rangle \leq \mathbb{E}\Big[f_{i_t}(\theta^*) - f_{i_t}(\theta^{(t')})\Big]. \tag{3.7.22}$$

Besides, the RSS condition implies

$$\begin{aligned}\sum_{j=t'}^{t-1} \mathbb{E}\langle \theta^{(j)} - \theta^{(j+1)}, \nabla_{\mathcal{I}} f_{i_j}(\theta^{(j)})\rangle &\leq \sum_{j=t'}^{t-1} \mathbb{E}\left[f_{i_t}(\theta^{(j)}) - f_{i_t}(\theta^{(j+1)}) + \frac{\rho_s^+}{2}\|\theta^{(j)} - \theta^{(j+1)}\|_2^2\right] \\ &\leq \mathbb{E}\Big[f_{i_t}(\theta^{(t')}) - f_{i_t}(\theta^{(t)})A\Big] + \frac{\rho_s^+ \Delta}{2}\sum_{j=t'}^{t-1}\mathbb{E}\|\theta^{(j)} - \theta^{(j+1)}\|_2^2. \end{aligned} \tag{3.7.23}$$

Moreover, we have

$$\begin{aligned}&\sum_{j=t'}^{t-1} \mathbb{E}\langle \theta^{(j)} - \theta^{(j+1)}, \nabla_{\mathcal{I}} f_{i_t}(\theta^{(t')}) - \nabla_{\mathcal{I}} f_{i_t}(\theta^{(j)})\rangle \\ &\leq \mathbb{E}\left[\sum_{j=t'}^{t-1} \|\theta_{e_t}^{(j)} - \theta_{e_t}^{(j+1)}\|_2 \cdot \|\nabla_{\mathcal{I}} f_{i_t}(\theta^{(t')}) - \nabla_{\mathcal{I}} f_{i_t}(\theta^{(j)})\|_2\right] \\ &\leq \mathbb{E}\left[\sum_{j=t'}^{t-1} \|\theta_{e_t}^{(j)} - \theta_{e_t}^{(j+1)}\|_2 \cdot \sum_{l=t'}^{j-1} \|\nabla_{\mathcal{I}} f_{i_t}(\theta^{(l)}) - \nabla_{\mathcal{I}} f_{i_t}(\theta^{(l+1)})\|_2\right] \\ &\leq \mathbb{E}\left[\sum_{j=t'}^{t-1}\sum_{l=t'}^{j-1} \frac{\rho_s^+}{2}\left(\|\theta_{e_t}^{(j)} - \theta_{e_t}^{(j+1)}\|_2 + \|\theta_{e_t}^{(l)} - \theta_{e_t}^{(l+1)}\|_2\right)\right] \\ &\leq \frac{\rho_s^+ \Delta(\varsigma - 1)}{2}\sum_{j=t'}^{t-1}\mathbb{E}\|\theta^{(j)} - \theta^{(j+1)}\|_2^2, \end{aligned} \tag{3.7.24}$$

where the first inequality is from Cauchy-Schwarz inequality, the second inequality is from the triangle inequality, the third inequality is from the RSS condition

and the inequality of arithmetic and geometric means, and the last inequality is from a counting argument.

Combining (3.7.21) – (3.7.24), we have

$$\mathbb{E}\langle\theta^{(t)}-\theta^*, h_{\mathcal{I}}^{(t)}(\theta^{(t)})\rangle \geq \mathbb{E}\left[\mathcal{F}(\theta^{(t)}-\mathcal{F}(\theta^*)-\rho_s^+\Delta\varsigma\eta^2\sum_{j=t'}^{t-1}\mathbb{E}\|\theta^{(j)}-\theta^{(j+1)}\|_2^2\right]. \quad (3.7.25)$$

Combing (3.7.19), (3.7.20), and (3.7.25), we have

$$\begin{aligned}\mathbb{E}\|u-\theta^*\|_2^2 \leq \mathbb{E}\Big[&\|\theta^{(t)}-\theta^*\|_2^2+\eta^2\|h_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2 \\ &-2\eta\left(\mathcal{F}(\theta^{(t)}-\mathcal{F}(\theta^*)\right)+\rho_s^+\Delta\varsigma\eta^2\sum_{j=t'}^{t-1}\|h_{\mathcal{I}}^{(j)}(\theta^{(j)})\|_2^2\Big].\end{aligned} \quad (3.7.26)$$

The rest of the proof follows analogously from the proof of Theorem 3.3.7. Specifically, by summing (3.7.26) over $t = 0, 1, \ldots, m-1$, taking expectation with respect to all t's, and combining Lemma 3.3.3, Lemma 3.3.5, and (3.7.25), we have

$$\begin{aligned}&\mathbb{E}\|\theta^{(m)}-\theta^*\|_2^2+\frac{2\eta\,(1-12\rho_s^+\eta\Gamma)\,(\alpha^m-1)}{\alpha-1}\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r)})-\mathcal{F}(\theta^*)\Big] \\ &\quad\leq\left(\frac{2\alpha^m}{\rho_s^-}+\frac{24\rho_s^+\eta^2\Gamma(\alpha^m-1)}{\alpha-1}\right)\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r-1)})-\mathcal{F}(\theta^*)\Big] \\ &\quad\quad+\frac{6\eta^2\Gamma(\alpha^m-1)}{\alpha-1}\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2^2,\end{aligned} \quad (3.7.27)$$

where $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$ and $\Gamma = \frac{1+\rho_s^+ \Delta\varsigma^2 \eta}{1-2\rho_s^{+2}\Delta\varsigma^2\eta^2}$. It further follows from (3.7.27)

$$\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\Big] \leq \left(\frac{\alpha^m(\alpha-1)}{\eta\rho_s^-(1-12\eta\rho_s^+\Gamma)(\alpha^m-1)} + \frac{12\eta\rho_s^+\Gamma}{1-12\eta\rho_s^+\Gamma}\right)\mathbb{E}\Big[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)\Big] + \frac{3\eta\Gamma}{1-12\eta\rho_s^+\Gamma}\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\theta^*)\|_2^2. \quad (3.7.28)$$

Finally, $\frac{\alpha^m(\alpha-1)}{\eta\rho_s^-(1-12\eta\rho_s^+\Gamma)(\alpha^m-1)} + \frac{12\eta\rho_s^+\Gamma}{1-12\eta\rho_s^+\Gamma} \leq \frac{5}{6}$ holds with the same choices of constants C_1 to C_4 as in Theorem 3.3.7 and $C_5 = \frac{1}{2}$.

Chapter 4

Alternating Optimization for Matrix Factorization

This chapter introduces our novel computational theory on alternating optimization for matrix factorization. By investigating the data generating process (underlying statistical models) of matrix factorization problems, we show that the resulting nonconvex optimization problem shows strong bi-convexity and smoothness over. Therefore, by exploiting such hidden convex structures, we establish new computational and statistical theory for a broad family of alternating optimization algorithms.

4.1 Background

Let $M^* \in \mathbb{R}^{m\times n}$ be a rank k matrix with k much smaller than m and n. Our goal is to estimate M^* based on partial observations of its entries. For example, matrix completion is based on a subsample of M^*'s entries, while matrix sensing is based on linear measurements $\langle A_i, M^*\rangle$, where $i \in \{1,\ldots,d\}$ with d much smaller than mn and A_i is the sensing matrix. In the past decade, significant progress has been made on the recovery of low rank matrix [72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 66, 87, 88, 67, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104]. Among these works, most are based upon convex relaxation with nuclear norm constraint or regularization. Nevertheless, solving these convex optimization problems can be computationally prohibitive in high dimensional regimes with large m and n [105]. A computationally more efficient alternative is nonconvex optimization. In particular, we reparameterize the $m \times n$ matrix variable M in the optimization problem as $UV^\top$ with $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$, and optimize over U and V. Such a reparametrization automatically enforces the low rank structure and leads to low computational cost per iteration. Due to this reason, the nonconvex approach is widely used in large scale applications such as recommendation systems or collaborative filtering [106, 107].

Despite the superior empirical performance of the nonconvex approach, the understanding of its theoretical guarantees is rather limited in comparison with the convex relaxation approach. The classical nonconvex optimization theory can

only show its sublinear convergence to local optima. But many empirical results have corroborated its exceptional computational performance and convergence to global optima. Only until recently has there been theoretical analysis of the block coordinate descent-type nonconvex optimization algorithm, which is known as alternating minimization [94, 96, 97, 98]. In particular, the existing results show that, provided a proper initialization, the alternating minimization algorithm attains a linear rate of convergence to a global optimum $U^* \in \mathbb{R}^{m\times k}$ and $V^* \in \mathbb{R}^{n\times k}$, which satisfy $M^* = U^*V^{*\top}$. Meanwhile, [77, 78] establish the convergence of the gradient-type methods, and [99] further establish the convergence of a broad class of nonconvex optimization algorithms including both gradient-type and block coordinate descent-type methods. However, [77, 78, 99] only establish the asymptotic convergence for an infinite number of iterations, rather than the explicit rate of convergence. Besides these works, [76, 79, 95] consider projected gradient-type methods, which optimize over the matrix variable $M \in \mathbb{R}^{m\times n}$ rather than $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$. These methods involve calculating the top k singular vectors of an $m \times n$ matrix at each iteration. For k much smaller than m and n, they incur much higher computational cost per iteration than the aforementioned methods that optimize over U and V. All these works, except [99], focus on specific algorithms, while [99] do not establish the explicit optimization rate of convergence.

In this chapter, we propose a new theory for analyzing a broad class of nonconvex optimization algorithms for low rank matrix estimation. The core of our theory

is the notion of inexact first order oracle. Based on the inexact first order oracle, we establish sufficiently conditions under which the iteration sequences converge geometrically to the global optima. For both matrix sensing and completion, a direct consequence of our threoy is that, a broad family of nonconvex optimization algorithms, including gradient descent, block coordinate gradient descent, and block coordinate minimization, attain linear rates of convergence to the true low rank matrices U^* and V^*. In particular, our proposed theory covers alternating minimization as a special case and recovers the results of [94, 96, 97, 98] under suitable conditions. Meanwhile, our approach covers gradient-type methods, which are also widely used in practice [108, 109, 107, 110, 90, 111]. To the best of our knowledge, our analysis is the first one that establishes exact recovery guarantees and geometric rates of convergence for a broad family of nonconvex matrix sensing and completion algorithms.

To achieve maximum generality, our unified analysis significantly differs from previous works. In detail, [94, 96, 97, 98] view alternating minimization as an approximate power method. However, their point of view relies on the closed form solution of each iteration of alternating minimization, which makes it difficult to generalize to other algorithms, e.g., gradient-type methods. Meanwhile, [99] take a geometric point of view. In detail, they show that the global optimum of the optimization problem is the unique stationary point within its neighborhood and thus a broad class of algorithms succeed. However, such geometric analysis of

the objective function does not characterize the convergence rate of specific algorithms towards the stationary point. Unlike existing results, we analyze nonconvex optimization algorithms as approximate convex counterparts. For example, our analysis views alternating minimization on a nonconvex objective function as an approximate block coordinate minimization on some convex objective function. We use the key quantity, the inexact first order oracle, to characterize such a perturbation effect, which results from the local nonconvexity at intermediate solutions. This eventually allows us to establish explicit rate of convergence in an analogous way as existing convex optimization analysis.

Our proposed inexact first order oracle is closely related to a series previous work on inexact or approximate gradient descent algorithms: [112, 113, 114, 115, 116, 117, 118]. Different from these existing results focusing on convex minimization, we show that the inexact first order oracle can also sharply captures the evolution of generic optimization algorithms even with the presence of nonconvexity. More recently, [119, 120, 121] respectively analyze the Wirtinger Flow algorithm for phase retrieval, the expectation maximization (EM) Algorithm for latent variable models, and the gradient descent algorithm for sparse coding based on a similar idea to ours. Though their analysis exploits similar nonconvex structures, they work on completely different problems, and the delivered technical results are also fundamentally different.

A conference version of this chapter was presented in the Annual Conference

on Neural Information Processing Systems 2015 [11]. During our conference version was under review, similar work was released on arXiv.org by [122, 123, 124, 125]. These works focus on symmetric positive semidefinite low rank matrix factorization problems. In contrast, our proposed methodologies and theory do not require the symmetry and positive semidefiniteness, and therefore can be applied to rectangular low rank matrix factorization problems.

The rest of this section is organized as follows. In Section 4.2, we review the matrix sensing problems, and then introduce a general class of nonconvex optimization algorithms. In Section 4.3, we present the convergence analysis of the algorithms. In Section 4.4, we lay out the proof. In Section 4.5, we extend the proposed methodology and theory to the matrix completion problems. In Section 4.6, we provide numerical experiments. All supplementary proof is provided in Appendix C.

Notation: For $v = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, we define the vector ℓ_q norm as $\|v\|_q^q = \sum_j v_j^q$. We define e_i as an indicator vector, where the i-th entry is one, and all other entries are zero. For a matrix $A \in \mathbb{R}^{m\times n}$, we use $A_{*j} = (A_{1j}, ..., A_{mj})^\top$ to denote the j-th column of A, and $A_{i*} = (A_{i1}, ..., A_{in})^\top$ to denote the i-th row of A. Let $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ be the largest and smallest nonzero singular values of A. We define the following matrix norms: $\|A\|_{\mathrm{F}}^2 = \sum_j \|A_{*j}\|_2^2$, $\|A\|_2 = \sigma_{\max}(A)$. Moreover, we define $\|A\|_*$ to be the sum of all singular values of A. We define as the Moore-Penrose pseudoinverse of A as $A^\dagger$. Given another matrix $B \in \mathbb{R}^{m\times n}$, we define the inner

product as $\langle A, B\rangle = \sum_{i,j} A_{ij}B_{ij}$. For a bivariate function $f(u,v)$, we define $\nabla_u f(u,v)$ to be the gradient with respect to u. Moreover, we use the common notations of $\Omega(\cdot)$, $O(\cdot)$, and $o(\cdot)$ to characterize the asymptotics of two real sequences.

4.2 Matrix Sensing

We start with the matrix sensing problem. Let $M^* \in \mathbb{R}^{m\times n}$ be the unknown low rank matrix of interest. We have d sensing matrices $A_i \in \mathbb{R}^{m\times n}$ with $i \in \{1,\dots,d\}$. Our goal is to estimate M^* based on $b_i = \langle A_i, M^*\rangle$ in the high dimensional regime with d much smaller than mn. Under such a regime, a common assumption is $\text{rank}(M^*) = k \ll \min\{d,m,n\}$. Existing approaches generally recover M^* by solving the following convex optimization problem

$$\min_{M\in\mathbb{R}^{m\times n}} \|M\|_* \quad \text{subject to } b = \mathcal{A}(M), \tag{4.2.1}$$

where $b = [b_1,...,b_d]^\top \in \mathbb{R}^d$, and $\mathcal{A}(M):\ \mathbb{R}^{m\times n} \to \mathbb{R}^d$ is an operator defined as

$$\mathcal{A}(M) = [\langle A_1, M\rangle,...,\langle A_i, M\rangle]^\top \in \mathbb{R}^d. \tag{4.2.2}$$

Existing convex optimization algorithms for solving (4.2.1) are computationally inefficient, since they incur high per-iteration computational cost and only attain sublinear rates of convergence to the global optimum [94, 105]. Therefore in large

scale settings, we usually consider the following nonconvex optimization problem instead

$$\min_{U\in\mathbb{R}^{m\times k},V\in\mathbb{R}^{n\times k}} \mathcal{F}(U,V), \text{ where } \mathcal{F}(U,V)=\frac{1}{2}\|b-\mathcal{A}(UV^\top)\|_2^2. \tag{4.2.3}$$

The reparametrization of $M = UV^\top$, though making the problem in (4.2.3) nonconvex, significantly improves the computational efficiency. Existing literature [106, 107, 108, 109, 107, 110, 90, 111] has established convincing evidence that (4.2.3) can be effectively solved by a broad variety of gradient-based nonconvex optimization algorithms, including gradient descent, alternating exact minimization (i.e., alternating least squares or block coordinate minimization), as well as alternating gradient descent (i.e., block coordinate gradient descent), as illustrated in Algorithm 7.

It is worth noting that the QR decomposition and rank k singular value decomposition in Algorithm 7 can be accomplished efficiently. In particular, the QR decomposition can be accomplished in $O(k^2\max\{m,n\})$ operations, while the rank k singular value decomposition can be accomplished in $O(kmn)$ operations. In fact, the QR decomposition is not necessary for particular update schemes, e.g., [94] prove that the alternating exact minimization update schemes with or without the QR decomposition are equivalent.

Algorithm 7: A family of nonconvex optimization algorithms for matrix sensing. Here $(\overline{U}, D, \overline{V}) \leftarrow \mathrm{KSVD}(M)$ is the rank k singular value decomposition of M. D is a diagonal matrix containing the top k singular values of M in decreasing order, and $\overline{U}$ and $\overline{V}$ contain the corresponding top k left and right singular vectors of M. $(\overline{V}, R_{\overline{V}}) \leftarrow \mathrm{QR}(V)$ is the QR decomposition, where $\overline{V}$ is the corresponding orthonormal matrix and $R_{\overline{V}}$ is the corresponding upper triangular matrix.

Input: $\{b_i\}_{i=1}^d$, $\{A_i\}_{i=1}^d$

Parameter: Step size η, Total number of iterations T

$(\overline{U}^{(0)}, D^{(0)}, \overline{V}^{(0)}) \leftarrow \mathrm{KSVD}(\sum_{i=1}^d b_i A_i)$, $V^{(0)} \leftarrow \overline{V}^{(0)} D^{(0)}$, $U^{(0)} \leftarrow \overline{U}^{(0)} D^{(0)}$

For $t \leftarrow 0, 1, ..., T-1$

Updating V:

- Alternating Exact Minimization : $V^{(t+0.5)} \leftarrow \mathrm{argmin}_V \mathcal{F}(\overline{U}^{(t)}, V)$
  $(\overline{V}^{(t+1)}, R_{\overline{V}}^{(t+0.5)}) \leftarrow \mathrm{QR}(V^{(t+0.5)})$
- Alternating Gradient Descent : $V^{(t+0.5)} \leftarrow V^{(t)} - \eta \nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)})$
  $(\overline{V}^{(t+1)}, R_{\overline{V}}^{(t+0.5)}) \leftarrow \mathrm{QR}(V^{(t+0.5)})$, $U^{(t)} \leftarrow \overline{U}^{(t)} R_{\overline{V}}^{(t+0.5)\top}$
- Gradient Descent : $V^{(t+0.5)} \leftarrow V^{(t)} - \eta \nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)})$
  $(\overline{V}^{(t+1)}, R_{\overline{V}}^{(t+0.5)}) \leftarrow \mathrm{QR}(V^{(t+0.5)})$, $U^{(t+1)} \leftarrow \overline{U}^{(t)} R_{\overline{V}}^{(t+0.5)\top}$

Updating U:

- Alternating Exact Minimization : $U^{(t+0.5)} \leftarrow \mathrm{argmin}_U \mathcal{F}(U, \overline{V}^{(t+1)})$
  $(\overline{U}^{(t+1)}, R_{\overline{U}}^{(t+0.5)}) \leftarrow \mathrm{QR}(U^{(t+0.5)})$
- Alternating Gradient Descent : $U^{(t+0.5)} \leftarrow U^{(t)} - \eta \nabla_U \mathcal{F}(U^{(t)}, \overline{V}^{(t+1)})$
  $(\overline{U}^{(t+1)}, R_{\overline{U}}^{(t+0.5)}) \leftarrow \mathrm{QR}(U^{(t+0.5)})$, $V^{(t+1)} \leftarrow \overline{V}^{t+1} R_{\overline{U}}^{(t+0.5)\top}$
- Gradient Descent : $U^{(t+0.5)} \leftarrow U^{(t)} - \eta \nabla_U \mathcal{F}(U^{(t)}, \overline{V}^{(t)})$
  $(\overline{U}^{(t+1)}, R_{\overline{U}}^{(t+0.5)}) \leftarrow \mathrm{QR}(U^{(t+0.5)})$, $V^{(t+1)} \leftarrow \overline{V}^{t} R_{\overline{U}}^{(t+0.5)\top}$

Return: $M^{(T)} \leftarrow U^{(T-0.5)} \overline{V}^{(T)\top}$ (for gradient descent we use $\overline{U}^{(T)} V^{(T)\top}$)

4.3 Convergence Analysis

We analyze the convergence of the algorithms illustrated in Section 4.2. Before we present the main results, we first introduce a unified analytical framework based on a key quantity named the approximate first order oracle. Such a unified framework equips our theory with the maximum generality. Without loss of

generality, we assume $m \le n$ throughout the rest of this chapter.

4.3.1 Main Idea

We first provide an intuitive explanation for the success of nonconvex optimization algorithms, which forms the basis of our later analysis of the main results in §4. Recall that (4.2.3) can be written as a special instance of the following optimization problem,

$$\min_{U \in \mathbb{R}^{m\times k}, V \in \mathbb{R}^{n\times k}} f(U, V). \tag{4.3.1}$$

A key observation is that, given fixed U, $f(U, \cdot)$ is strongly convex and smooth in V under suitable conditions, and the same also holds for U given fixed V correspondingly. For the convenience of discussion, we summarize this observation in the following technical condition, which will be later verified for matrix sensing and completion under suitable conditions.

Assumption 4.3.1 (Strong Biconvexity and Bismoothness)**.** There exist universal constants $\mu_+ > 0$ and $\mu_- > 0$ such that

$$\frac{\mu_-}{2}\|U' - U\|_{\mathrm{F}}^2 \le f(U', V) - f(U, V) - \langle U' - U, \nabla_U f(U, V)\rangle \le \frac{\mu_+}{2}\|U' - U\|_{\mathrm{F}}^2 \text{ for all } U, U',$$
$$\frac{\mu_-}{2}\|V' - V\|_{\mathrm{F}}^2 \le f(U, V') - f(U, V) - \langle V' - V, \nabla_V f(U, V)\rangle \le \frac{\mu_+}{2}\|V' - V\|_{\mathrm{F}}^2 \text{ for all } V, V'.$$

4.3.1.1 Ideal First Order Oracle

To ease presentation, we assume that U^* and V^* are the unique global minimizers to the generic optimization problem in (4.3.1). Assuming that U^* is given, we can obtain V^* by

$$V^* = \underset{V \in \mathbb{R}^{n \times k}}{\operatorname{argmin}} f(U^*, V). \tag{4.3.2}$$

Assumption 4.3.1 implies the objective function in (4.3.2) is strongly convex and smooth. Hence, we can choose any gradient-based algorithm to obtain V^*. For example, we can directly solve for V^* in

$$\nabla_V f(U^*, V) = 0, \tag{4.3.3}$$

or iteratively solve for V^* using gradient descent, i.e.,

$$V^{(t)} = V^{(t-1)} - \eta \nabla_V f(U^*, V^{(t-1)}), \tag{4.3.4}$$

where η is a step size. Taking gradient descent as an example, we can invoke classical convex optimization results [126] to prove that

$$\|V^{(t)} - V^*\|_{\mathrm{F}} \leq \kappa \|V^{(t-1)} - V^*\|_{\mathrm{F}} \text{ for all } t = 0, 1, 2, \ldots,$$

where $\kappa \in (0,1)$ and only depends on μ_+ and μ_- in Assumption 4.3.1. For notational simplicity, we call $\nabla_V f(U^*, V^{(t-1)})$ the ideal first order oracle, since we do not know U^* in practice.

4.3.1.2 Inexact First Order Oracle

Though the ideal first order oracle is not accessible in practice, it provides us insights to analyze nonconvex optimization algorithms. Taking gradient descent as an example, at the t-th iteration, we take a gradient descent step over V based on $\nabla_V f(U, V^{(t-1)})$. Now we can treat $\nabla_V f(U, V^{(t-1)})$ as an approximation of $\nabla_V f(U^*, V^{(t-1)})$, where the approximation error comes from approximating U^* by U. Then the relationship between $\nabla_V f(U^*, V^{(t-1)})$ and $\nabla_V f(U, V^{(t-1)})$ is similar to that between gradient and approximate gradient in existing literature on convex optimization. For simplicity, we call $\nabla_V f(U, V^{(t-1)})$ the inexact first order oracle.

To characterize the difference between $\nabla_V f(U^*, V^{(t-1)})$ and $\nabla_V f(U, V^{(t-1)})$, we define the approximation error of the inexact first order oracle as

$$\mathcal{E}(V, V', U) = \|\nabla_V f(U^*, V') - \nabla_V f(U, V')\|_{\mathrm{F}}, \tag{4.3.5}$$

where V' is the current decision variable for evaluating the gradient. In the above example, it holds for $V' = V^{(t-1)}$. Later we will illustrate that $\mathcal{E}(V, V', U)$ is critical to our analysis. In the above example of alternating gradient descent, we will prove

later that for $V^{(t)} = V^{(t-1)} - \eta \nabla_V f(U, V^{(t-1)})$, we have

$$\|V^{(t)} - V^*\|_{\mathrm{F}} \leq \kappa \|V^{(t-1)} - V^*\|_{\mathrm{F}} + \frac{2}{\mu_+} \mathcal{E}(V^{(t)}, V^{(t-1)}, U). \tag{4.3.6}$$

In other words, $\mathcal{E}(V^{(t)}, V^{(t-1)}, U)$ captures the perturbation effect by employing the inexact first order oracle $\nabla_V f(U, V^{(t-1)})$ instead of the ideal first order oracle $\nabla_V f(U^*, V^{(t-1)})$. For $V^{(t+1)} = \operatorname{argmin}_V f(U, V)$, we will prove that

$$\|V^{(t)} - V^*\|_{\mathrm{F}} \leq \frac{1}{\mu_-} \mathcal{E}(V^{(t)}, V^{(t)}, U). \tag{4.3.7}$$

According to the update schemes shown in Algorithms 7 and 8, for alternating exact minimization, we set $U = U^{(t)}$ in (4.3.7), while for gradient descent or alternating gradient descent, we set $U = U^{(t-1)}$ or $U = U^{(t)}$ in (4.3.6) respectively. Due to symmetry, similar results also hold for $\|U^{(t)} - U^*\|_{\mathrm{F}}$.

To establish the geometric rate of convergence towards the global minima U^* and V^*, it remains to establish upper bounds for the approximate error of the inexact first oder oracle. Taking gradient decent as an example, we will prove that given an appropriate initial solution, we have

$$\frac{2}{\mu_+} \mathcal{E}(V^{(t)}, V^{(t-1)}, U^{(t-1)}) \leq \alpha \|U^{(t-1)} - U^*\|_{\mathrm{F}} \tag{4.3.8}$$

for some $\alpha \in (0, 1 - \kappa)$. Combining with (4.3.6) (where we take $U = U^{(t-1)}$), (4.3.8)

further implies

$$\|V^{(t)} - V^*\|_{\mathrm{F}} \leq \kappa \|V^{(t-1)} - V^*\|_{\mathrm{F}} + \alpha \|U^{(t-1)} - U^*\|_{\mathrm{F}}. \tag{4.3.9}$$

Correspondingly, similar results hold for $\|U^{(t)} - U^*\|_{\mathrm{F}}$, i.e.,

$$\|U^{(t)} - U^*\|_{\mathrm{F}} \leq \kappa \|U^{(t-1)} - U^*\|_{\mathrm{F}} + \alpha \|V^{(t-1)} - V^*\|_{\mathrm{F}}. \tag{4.3.10}$$

Combining (4.3.9) and (4.3.10) we then establish the contraction

$$\max\{\|V^{(t)} - V^*\|_{\mathrm{F}}, \|U^{(t)} - U^*\|_{\mathrm{F}}\} \leq (\alpha + \kappa) \cdot \max\{\|V^{(t-1)} - V^*\|_{\mathrm{F}}, \|U^{(t-1)} - U^*\|_{\mathrm{F}}\},$$

which further implies the geometric convergence, since $\alpha \in (0, 1-\kappa)$. Respectively, we can establish similar results for alternating exact minimization and alternating gradient descent. Based upon such a unified analysis, we now present the main results.

4.3.2 Main Results

Before presenting the main results, we first introduce an assumption known as the restricted isometry property (RIP). Recall that k is the rank of the target low rank matrix M^*.

Assumption 4.3.2 (Restricted Isometry Property)**.** The linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{m \times n} \to$

$\mathbb{R}^d$ defined in (4.2.2) satisfies $2k$-RIP with parameter $\delta_{2k} \in (0,1)$, i.e., for all $\Delta \in \mathbb{R}^{m\times n}$ such that $\text{rank}(\Delta) \leq 2k$, it holds that

$$(1-\delta_{2k})\|\Delta\|_{\text{F}}^2 \leq \|\mathcal{A}(\Delta)\|_2^2 \leq (1+\delta_{2k})\|\Delta\|_{\text{F}}^2.$$

Several random matrix ensembles satisfy $2k$-RIP for a sufficiently large d with high probability. For example, suppose that each entry of A_i is independently drawn from a sub-Gaussian distribution, $\mathcal{A}(\cdot)$ satisfies $2k$-RIP with parameter δ_{2k} with high probability for $d = \Omega(\delta_{2k}^{-2} kn \log n)$.

The following theorem establishes the geometric rate of convergence of the nonconvex optimization algorithms summarized in Algorithm 7.

Theorem 4.3.3. Assume there exists a sufficiently small constant C_1 such that $\mathcal{A}(\cdot)$ satisfies $2k$-RIP with $\delta_{2k} \leq C_1/k$, and the largest and smallest nonzero singular values of M^* are constants, which do not scale with (d,m,n,k). For any pre-specified precision ϵ, there exist an η and universal constants C_2 and C_3 such that for all $T \geq C_2 \log(C_3/\epsilon)$, we have $\|M^{(T)} - M^*\|_{\text{F}} \leq \epsilon$.

The proof of Theorems 4.3.3 is provided in Section 4.4.2, Section 4.4.3, and Section 4.4.4. Theorem 4.3.3 implies that all three nonconvex optimization algorithms converge geometrically to the global optimum. Moreover, assuming that each entry of A_i is independently drawn from a sub-Gaussian distribution with mean zero and variance proxy one, our result further suggests that, to achieve ex-

act low rank matrix recovery, our algorithm requires the number of measurements d to satisfy

$$d = \Omega(k^3 n \log n), \tag{4.3.11}$$

since we assume that $\delta_{2k} \leq C_1/k$. This sample complexity result matches the state-of-the-art result for nonconvex optimization methods, which is established by [94]. In comparison with their result, which only covers the alternating exact minimization algorithm, our results holds for a broader variety of nonconvex optimization algorithms.

Note that the sample complexity in (4.3.11) depends on a polynomial of $\frac{\sigma_{\max}(M^*)}{\sigma_{\min}(M^*)}$, which is treated as a constant in our chapter. If we allow $\frac{\sigma_{\max}(M^*)}{\sigma_{\min}(M^*)}$ to increase, we can plug the nonconvex optimization algorithms into the multi-stage framework proposed by [94]. Following similar lines to the proof of Theorem 4.3.3, we can derive a new sample complexity, which is independent of $\frac{\sigma_{\max}(M^*)}{\sigma_{\min}(M^*)}$. See more details in [94].

4.4 Proof of Main Results

We sketch the proof of Theorems 4.3.3. The proof of all related lemmas are provided in Appendix C. For notational simplicity, let $\sigma_1 = \sigma_{\max}(M^*)$ and $\sigma_k = \sigma_{\min}(M^*)$. Recall the nonconvex optimization algorithms are symmetric about the

updates of U and V. Hence, the following lemmas for the update of V also hold for updating U. We omit some statements for conciseness. Theorem 4.5.2 can be proved in a similar manner, and its proof is provided in Appendix C.5.

Before presenting the proof, we first introduce the following lemma, which verifies Assumption 4.3.1.

Lemma 4.4.1. Suppose that $\mathcal{A}(\cdot)$ satisfies $2k$-RIP with parameter δ_{2k}. Given an arbitrary orthonormal matrix $\overline{U} \in \mathbb{R}^{m\times k}$, for any $V,\ V' \in \mathbb{R}^{n\times k}$, we have

$$\frac{1+\delta_{2k}}{2}\|V'-V\|_{\mathrm{F}}^2 \geq \mathcal{F}(\overline{U},V') - \mathcal{F}(\overline{U},V) - \langle \nabla_V \mathcal{F}(\overline{U},V), V'-V\rangle \geq \frac{1-\delta_{2k}}{2}\|V'-V\|_{\mathrm{F}}^2.$$

The proof of Lemma 4.4.1 is provided in Appendix C.1.1. Lemma 4.4.1 implies that $\mathcal{F}(\overline{U},\cdot)$ is strongly convex and smooth in V given a fixed orthonormal matrix $\overline{U}$, as specified in Assumption 4.3.1. Equipped with Lemma 4.4.1, we now lay out the proof for each update scheme in Algorithm 7.

4.4.1 Rotation Issue

Given a factorization of $M^* = \overline{U}^* V^{*\top}$, we can equivalently represent it as $M^* = \overline{U}^*_{\mathrm{new}} V^{*\top}_{\mathrm{new}}$, where

$$\overline{U}^*_{\mathrm{new}} = \overline{U}^* O_{\mathrm{new}} \quad \text{and} \quad V^{*\top}_{\mathrm{new}} = V^{*\top} O_{\mathrm{new}}$$

for an arbitrary unitary matrix $O_{\text{new}} \in \mathbb{R}^{k\times k}$. This implies that directly calculating $\|\overline{U} - \overline{U}^*\|_{\text{F}}$ is not desirable and the algorithm may converge to an arbitrary factorization of M^*.

To address this issue, existing analysis usually chooses subspace distances to evaluate the difference between subspaces spanned by columns of $\overline{U}^*$ and $\overline{U}$, because these subspaces are invariant to rotations [94]. For example, let $\overline{U}_\perp \in \mathbb{R}^{m\times(m-k)}$ denote the orthonormal complement to $\overline{U}$, we can choose the subspace distance as $\|\overline{U}_\perp^\top \overline{U}^*\|_{\text{F}}$. For any $O_{\text{new}} \in \mathbb{R}^{k\times k}$ such that $O_{\text{new}}^\top O_{\text{new}} = I_k$, we have

$$\|\overline{U}_\perp^\top \overline{U}_{\text{new}}^*\|_{\text{F}} = \|\overline{U}_\perp^\top \overline{U}^* O_{\text{new}}\|_{\text{F}} = \|\overline{U}_\perp^\top \overline{U}^*\|_{\text{F}}.$$

In this chapter, we consider a different subspace distance defined as

$$\min_{O^\top O = I_k} \|\overline{U} - \overline{U}^* O\|_{\text{F}}. \tag{4.4.1}$$

We can verify that (4.4.1) is also invariant to rotation. The next lemma shows that (4.4.1) is equivalent to $\|\overline{U}_\perp^\top \overline{U}^*\|_{\text{F}}$.

Lemma 4.4.2. Given two orthonormal matrices $\overline{U} \in \mathbb{R}^{m\times k}$ and $\overline{U}^* \in \mathbb{R}^{m\times k}$, we have

$$\|\overline{U}_\perp^\top \overline{U}^*\|_{\text{F}} \le \min_{O^\top O = I} \|\overline{U} - \overline{U}^* O\|_{\text{F}} \le \sqrt{2}\|\overline{U}_\perp^\top \overline{U}^*\|_{\text{F}}.$$

The proof of Lemma 4.4.2 is provided in [127], therefore omitted. Equipped

with Lemma 4.4.2, our convergence analysis guarantees that there always exists a factorization of M^* satisfying the desired computational properties for each iteration (See Lemma 4.4.5, Corollaries 4.4.7 and 4.4.8). Similarly, the above argument can also be generalized to gradient descent and alternating gradient descent algorithms.

4.4.2 Proof of Theorem 4.3.3 (Alternating Exact Minimization)

Proof. Throughout the proof for alternating exact minimization, we define a constant $\xi \in (1, \infty)$ to simplify the notation. Moreover, we assume that at the t-th iteration, there exists a matrix factorization of M^*

$$M^* = \overline{U}^{*(t)} V^{*(t)\top},$$

where $\overline{U}^{*(t)} \in \mathbb{R}^{m \times k}$ is an orthonormal matrix. We define the approximation error of the inexact first order oracle as

$$\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}) = \|\nabla_V \mathcal{F}(\overline{U}^{*(t)}, V^{(t+0.5)}) - \nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t+0.5)})\|_{\mathrm{F}}.$$

The following lemma establishes an upper bound for the approximation error of the approximation first order oracle under suitable conditions.

Lemma 4.4.3. Suppose that δ_{2k} and $\overline{U}^{(t)}$ satisfy

$$\delta_{2k} \leq \frac{(1-\delta_{2k})^2\sigma_k}{12\xi k(1+\delta_{2k})\sigma_1} \quad \text{and} \quad \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}. \tag{4.4.2}$$

Then we have

$$\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}) \leq \frac{(1-\delta_{2k})\sigma_k}{2\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}.$$

The proof of Lemma 4.4.3 is provided in Appendix C.1.2. Lemma 4.4.3 shows that the approximation error of the inexact first order oracle for updating V diminishes with the estimation error of $\overline{U}^{(t)}$, when $\overline{U}^{(t)}$ is sufficiently close to $\overline{U}^{*(t)}$. The following lemma quantifies the progress of an exact minimization step using the inexact first order oracle.

Lemma 4.4.4. We have

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{1}{1-\delta_{2k}}\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}).$$

The proof of Lemma 4.4.4 is provided in Appendix C.1.3. Lemma 4.4.4 illustrates that the estimation error of $V^{(t+0.5)}$ diminishes with the approximation error of the inexact first order oracle. The following lemma characterizes the effect of the renormalization step using QR decomposition, i.e., the relationship between $V^{(t+0.5)}$ and $\overline{V}^{(t+1)}$ in terms of the estimation error.

Lemma 4.4.5. Suppose that $V^{(t+0.5)}$ satisfies

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{4}. \tag{4.4.3}$$

Then there exists a factorization of $M^* = U^{*(t+1)}\overline{V}^{*(t+1)}$ such that $\overline{V}^{*(t+0.5)} \in \mathbb{R}^{n\times k}$ is an orthonormal matrix, and satisfies

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{2}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}.$$

The proof of Lemma 4.4.5 is provided in Appendix C.1.4. The next lemma quantifies the accuracy of the initialization $\overline{U}^{(0)}$.

Lemma 4.4.6. Suppose that δ_{2k} satisfies

$$\delta_{2k} \leq \frac{(1-\delta_{2k})^2\sigma_k^4}{192\xi^2 k(1+\delta_{2k})^2\sigma_1^4}. \tag{4.4.4}$$

Then there exists a factorization of $M^* = \overline{U}^{*(0)}V^{*(0)\top}$ such that $\overline{U}^{*(0)} \in \mathbb{R}^{m\times k}$ is an orthonormal matrix, and satisfies

$$\|\overline{U}^{(0)} - \overline{U}^{*}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}.$$

The proof of Lemma 4.4.6 is provided in Appendix C.1.5. Lemma 4.4.6 implies that the initial solution $\overline{U}^{(0)}$ attains a sufficiently small estimation error.

Combining Lemmas 4.4.3, 4.4.4, and 4.4.5, we obtain the following corollary for a complete iteration of updating V.

Corollary 4.4.7. Suppose that δ_{2k} and $\overline{U}^{(t)}$ satisfy

$$\delta_{2k} \leq \frac{(1-\delta_{2k})^2\sigma_k^4}{192\xi^2 k(1+\delta_{2k})^2\sigma_1^4} \quad \text{and} \quad \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}. \tag{4.4.5}$$

We then have

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}.$$

Moreover, we also have

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{1}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \quad \text{and} \quad \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{2\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}.$$

The proof of Corollary 4.4.7 is provided in Appendix C.1.6. Since the alternating exact minimization algorithm updates U and V in a symmetric manner, we can establish similar results for a complete iteration of updating U in the next corollary.

Corollary 4.4.8. Suppose that δ_{2k} and $\overline{V}^{(t+1)}$ satisfy

$$\delta_{2k} \leq \frac{(1-\delta_{2k})^2\sigma_k^4}{192\xi^2 k(1+\delta_{2k})^2\sigma_1^4} \quad \text{and} \quad \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}. \tag{4.4.6}$$

Then there exists a factorization of $M^* = \overline{U}^{*(t+1)} V^{*(t+1)\top}$ such $\overline{U}^{*(t+1)}$ is an orthonormal matrix, and satisfies

$$\|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}.$$

Moreover, we also have

$$\|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{1}{\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \text{ and } \|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{2\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}.$$

The proof of Corollary 4.4.8 directly follows Appendix C.1.6, and is therefore omitted.

We then proceed with the proof of Theorem 4.3.3 for alternating exact minimization. Lemma 4.4.6 ensures that (4.4.5) of Corollary 4.4.7 holds for $\overline{U}^{(0)}$. Then Corollary 4.4.7 ensures that (4.4.6) of Corollary 4.4.8 holds for $\overline{V}^{(1)}$. By induction, Corollaries 4.4.7 and 4.4.8 can be applied recursively for all T iterations. Thus we obtain

$$\begin{aligned}\|\overline{V}^{(T)} - \overline{V}^{*(T)}\|_{\mathrm{F}} &\leq \frac{1}{\xi}\|\overline{U}^{(T-1)} - \overline{U}^{*(T-1)}\|_{\mathrm{F}} \leq \frac{1}{\xi^2}\|\overline{V}^{(T-1)} - \overline{V}^{*(T-1)}\|_{\mathrm{F}} \\ &\leq \cdots \leq \frac{1}{\xi^{2T-1}}\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi^{2T}(1+\delta_{2k})\sigma_1}, \end{aligned} \tag{4.4.7}$$

where the last inequality comes from Lemma 4.4.6. Therefore, for a pre-specified

accuracy ϵ, we need at most

$$T = \left\lceil \frac{1}{2} \log\left(\frac{(1-\delta_{2k})\sigma_k}{2\epsilon(1+\delta_{2k})\sigma_1} \right) \log^{-1} \xi \right\rceil \tag{4.4.8}$$

iterations such that

$$\|\overline{V}^{(T)} - \overline{V}^{*(T)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi^{2T}(1+\delta_{2k})\sigma_1} \leq \frac{\epsilon}{2}. \tag{4.4.9}$$

Moreover, Corollary 4.4.8 implies

$$\|U^{(T-0.5)} - U^{*(T)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{2\xi}\|\overline{V}^{(T)} - \overline{V}^{*(T)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k^2}{8\xi^{2T+1}(1+\delta_{2k})\sigma_1},$$

where the last inequality comes from (4.4.7). Therefore, we need at most

$$T = \left\lceil \frac{1}{2} \log\left(\frac{(1-\delta_{2k})\sigma_k^2}{4\xi\epsilon(1+\delta_{2k})} \right) \log^{-1} \xi \right\rceil \tag{4.4.10}$$

iterations such that

$$\|U^{(T-0.5)} - U^{*}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k^2}{8\xi^{2T+1}(1+\delta_{2k})\sigma_1} \leq \frac{\epsilon}{2\sigma_1}. \tag{4.4.11}$$

Then combining (4.4.9) and (4.4.11), we obtain

$$\begin{aligned}\|M^{(T)} - M^*\| &= \|U^{(T-0.5)}\overline{V}^{(T)\top} - U^{*(T)}\overline{V}^{*(T)\top}\|_{\mathrm{F}} \\ &= \|U^{(T-0.5)}\overline{V}^{(T)\top} - U^{*(T)}\overline{V}^{(T)\top} + U^{*(T)}\overline{V}^{(T)\top} - U^{*(T)}\overline{V}^{*(T)\top}\|_{\mathrm{F}} \\ &\leq \|\overline{V}^{(T)}\|_2\|U^{(T-0.5)} - U^{*(T)}\|_{\mathrm{F}} + \|U^{*(T)}\|_2\|\overline{V}^{(T)} - \overline{V}^{*(T)}\|_{\mathrm{F}} \leq \epsilon, \quad (4.4.12)\end{aligned}$$

where the last inequality comes from $\|\overline{V}^{(T)}\|_2 = 1$ (since $\overline{V}^{(T)}$ is orthonormal) and $\|U^*\|_2 = \|M^*\|_2 = \sigma_1$ (since $U^{*(T)}\overline{V}^{*(T)\top} = M^*$ and $\overline{V}^{*(T)}$ is orthonormal). Thus combining (4.4.8) and (4.4.10) with (4.4.12), we complete the proof. □

4.4.3 Proof of Theorem 4.3.3 (Alternating Gradient Descent)

Proof. Throughout the proof for alternating gradient descent, we define a sufficiently large constant ξ. Moreover, we assume that at the t-th iteration, there exists a matrix factorization of M^*

$$M^* = \overline{U}^{*(t)}V^{*(t)\top},$$

where $\overline{U}^{*(t)} \in \mathbb{R}^{m \times k}$ is an orthonormal matrix. We define the approximation error of the inexact first order oracle as

$$\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) = \|\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)}) - \nabla_V \mathcal{F}(\overline{U}^{*(t)}, V^{(t)})\|_{\mathrm{F}}.$$

The first lemma is parallel to Lemma 4.4.3 for alternating exact minimization.

Lemma 4.4.9. Suppose that δ_{2k}, $\overline{U}^{(t)}$, and $V^{(t)}$ satisfy

$$\delta_{2k} \leq \frac{(1-\delta_{2k})\sigma_k}{24\xi k\sigma_1}, \ \ \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{4\xi\sigma_1^2}, \ \text{ and } \ \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_1\sqrt{k}}{2}. \quad (4.4.13)$$

Then we have

$$\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \leq \frac{(1+\delta_{2k})\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}.$$

The proof of Lemma 4.4.9 is provided in Appendix C.2.1. Lemma 4.4.9 illustrates that the approximation error of the inexact first order oracle diminishes with the estimation error of $\overline{U}^{(t)}$, when $\overline{U}^{(t)}$ and $V^{(t)}$ are sufficiently close to $\overline{U}^{*(t)}$ and $V^{*(t)}$.

Lemma 4.4.10. Suppose that the step size parameter η satisfies

$$\eta = \frac{1}{1+\delta_{2k}}. \quad (4.4.14)$$

Then we have

$$\|V^{(t+0.5)} - V^*\|_{\mathrm{F}} \leq \sqrt{\delta_{2k}}\|V^{(t)} - V^*\|_{\mathrm{F}} + \frac{2}{1+\delta_{2k}}\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}).$$

The proof of Lemma 4.4.10 is provided in Appendix C.2.2. Lemma 4.4.10 characterizes the progress of a gradient descent step with a pre-specified fixed step size. A more practical option is adaptively selecting η using the backtracking line search procedure, and similar results can be guaranteed. See [126] for details. The following lemma characterizes the effect of the renormalization step using QR decomposition.

Lemma 4.4.11. Suppose that $V^{(t+0.5)}$ satisfies

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{4}. \tag{4.4.15}$$

Then there exists a factorization of $M^* = U^{*(t+1)}\overline{V}^{*(t+1)}$ such that $\overline{V}^{*(t+1)} \in \mathbb{R}^{n\times k}$ is an orthonormal matrix, and

$$\begin{aligned}
\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} &\leq \frac{2}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}, \\
\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} &\leq \frac{3\sigma_1}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}},
\end{aligned}$$

The proof of Lemma 4.4.11 is provided in Appendix C.2.3. The next lemma quantifies the accuracy of the initial solutions.

Lemma 4.4.12. Suppose that δ_{2k} satisfies

$$\delta_{2k} \leq \frac{\sigma_k^6}{192\xi^2 k\sigma_1^6}. \tag{4.4.16}$$

Then we have

$$\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{4\xi\sigma_1^2} \quad \text{and} \quad \|V^{(0)} - V^{*(0)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{2\xi\sigma_1} \leq \frac{\sigma_1\sqrt{k}}{2}.$$

The proof of Lemma 4.4.12 is provided in Appendix C.2.4. Lemma 4.4.12 indicates that the initial solutions $\overline{U}^{(0)}$ and $V^{(0)}$ attain sufficiently small estimation errors.

Combining Lemmas 4.4.9, 4.4.10, 4.4.5, , we obtain the following corollary for a complete iteration of updating V.

Corollary 4.4.13. Suppose that δ_{2k}, $\overline{U}^{(t)}$, and $V^{(t)}$ satisfy

$$\delta_{2k} \leq \frac{\sigma_k^6}{192\xi^2 k\sigma_1^6}, \quad \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{4\xi\sigma_1^2}, \quad \text{and} \quad \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{2\xi\sigma_1}. \tag{4.4.17}$$

We then have

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{4\xi\sigma_1^2} \quad \text{and} \quad \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{2\xi\sigma_1}.$$

Moreover, we have

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \sqrt{\delta_{2k}}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{4.4.18}$$

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{2\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{4}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{4.4.19}$$

$$\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \frac{3\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \left(\frac{6}{\xi} + 1\right)\sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}. \tag{4.4.20}$$

The proof of Corollary 4.4.13 is provided in Appendix C.2.5. Since the alternating gradient descent algorithm updates U and V in a symmetric manner, we can establish similar results for a complete iteration of updating U in the next corollary.

Corollary 4.4.14. Suppose that δ_{2k}, $\overline{V}^{(t+1)}$, and $U^{(t)}$ satisfy

$$\delta_{2k} \leq \frac{\sigma_k^6}{192\xi^2 k\sigma_1^6}, \ \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{4\xi\sigma_1^2}, \ \text{and} \ \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{2\xi\sigma_1}. \tag{4.4.21}$$

We then have

$$\|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{4\xi\sigma_1^2} \quad \text{and} \quad \|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{2\xi\sigma_1}.$$

Moreover, we have

$$\|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \sqrt{\delta_{2k}}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \frac{2\sigma_k}{\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{4.4.22}$$

$$\|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{2\sqrt{\delta_{2k}}}{\sigma_k}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \frac{4}{\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{4.4.23}$$

$$\|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}} \leq \frac{3\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \left(\frac{6}{\xi} + 1\right)\sigma_1\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}. \tag{4.4.24}$$

The proof of Corollary 4.4.14 directly follows Appendix C.2.5, and is therefore omitted.

Now we proceed with the proof of Theorem 4.3.3 for alternating gradient descent. Recall that Lemma 4.4.12 ensures that (4.4.17) of Corollary 4.4.13 holds for $\overline{U}^{(0)}$ and $V^{(0)}$. Then Corollary 4.4.13 ensures that (4.4.21) of Corollary 4.4.14 holds for $U^{(0)}$ and $\overline{V}^{(1)}$. By induction, Corollaries 4.4.7 and 4.4.8 can be applied recursively for all T iterations. For notational simplicity, we write (4.4.18)-(4.4.24)

as

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \alpha_1 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \gamma_1 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{4.4.25}$$

$$\sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_2 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \gamma_2 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{4.4.26}$$

$$\|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_3 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \gamma_3 \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{4.4.27}$$

$$\sigma_1 \|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_4 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \gamma_4 \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{4.4.28}$$

$$\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_5 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \gamma_5 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{4.4.29}$$

$$\|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_6 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \gamma_6 \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}. \tag{4.4.30}$$

Note that we have $\gamma_5, \gamma_6 \in (1, 2)$, but α_1,...,α_6, γ_1,..., and γ_4 can be sufficiently small as long as ξ is sufficiently large. We then have

$$\begin{aligned}
\|U^{(t+1)} - U^{*(t+2)}\|_{\mathrm{F}} &\overset{\text{(i)}}{\leq} \alpha_5 \|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}} + \gamma_5 \sigma_1 \|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \\
&\overset{\text{(ii)}}{\leq} \alpha_5 \alpha_6 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \alpha_5 \gamma_6 \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} + \gamma_5 \sigma_1 \|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \\
&\overset{\text{(iii)}}{\leq} (\alpha_5 \alpha_6 + \gamma_5 \alpha_4) \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + (\gamma_5 \gamma_4 \sigma_1 + \alpha_5 \gamma_6) \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \\
&\overset{\text{(iv)}}{\leq} (\alpha_5 \alpha_6 + \gamma_5 \alpha_4) \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + (\gamma_5 \gamma_4 \sigma_1 + \alpha_5 \gamma_6) \alpha_2 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} \\
&\qquad + (\gamma_5 \gamma_4 \sigma_1 + \alpha_5 \gamma_6) \gamma_2 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}},
\end{aligned} \tag{4.4.31}$$

where (i) comes from (4.4.29), (ii) comes from (4.4.30), (iii) comes from (4.4.28),

and (iv) comes from (4.4.26). Similarly, we can obtain

$$\|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}} \le \alpha_6 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \gamma_6 \alpha_2 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \gamma_6 \gamma_2 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \quad (4.4.32)$$

$$\sigma_1 \|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \le \alpha_4 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \gamma_4 \alpha_2 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \gamma_4 \gamma_2 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \quad (4.4.33)$$

$$\|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}} \le \alpha_3 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \gamma_3 \alpha_2 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \gamma_3 \gamma_2 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}. \quad (4.4.34)$$

For simplicity, we define

$$\phi_{V^{(t+1)}} = \|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}},\ \phi_{V^{(t+0.5)}} = \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}},\ \phi_{\overline{V}^{(t+1)}} = \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}},$$

$$\phi_{U^{(t+1)}} = \|U^{(t+1)} - U^{*(t+2)}\|_{\mathrm{F}},\ \phi_{U^{(t+0.5)}} = \|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}},\ \phi_{\overline{U}^{(t+1)}} = \sigma_1 \|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}}.$$

Then combining (4.4.25), (4.4.26) with (4.4.31)–(4.4.34), we obtain

$$\max\left\{\phi_{V^{(t+1)}}, \phi_{V^{(t+0.5)}}, \phi_{\overline{V}^{(t+1)}}, \phi_{U^{(t+1)}}, \phi_{U^{(t+0.5)}}, \phi_{\overline{U}^{(t+1)}}\right\} \le \beta \max\left\{\phi_{V^{(t)}}, \phi_{U^{(t)}}, \phi_{\overline{U}^{(t)}}\right\}, \quad (4.4.35)$$

where β is a contraction coefficient defined as

$$\beta = \max\{\alpha_5\alpha_6 + \gamma_5\alpha_4, \alpha_6, \alpha_4, \alpha_3\} + \max\{\alpha_1, \alpha_2, (\gamma_5\gamma_4\sigma_1 + \alpha_5\gamma_6), \gamma_6\alpha_2, \gamma_4\alpha_2, \gamma_3\alpha_2\}$$
$$+ \max\{\gamma_1, \gamma_2, (\gamma_5\gamma_4\sigma_1 + \alpha_5\gamma_6)\gamma_2, \gamma_6\gamma_2, \gamma_4\gamma_2, \gamma_3\gamma_2\}.$$

Then we can choose ξ as a sufficiently large constant such that $\beta < 1$. By recursively applying (4.4.35) for $t = 0, ..., T$, we obtain

$$\max\left\{\phi_{V^{(T)}}, \phi_{V^{(T-0.5)}}, \phi_{\overline{V}^{(T)}}, \phi_{U^{(T)}}, \phi_{U^{(T-0.5)}}, \phi_{\overline{U}^{(T)}}\right\} \le \beta \max\left\{\phi_{V^{(T-1)}}, \phi_{U^{(T-1)}}, \phi_{\overline{U}^{(T-1)}}\right\}$$
$$\le \beta^2 \max\left\{\phi_{V^{(T-2)}}, \phi_{U^{(T-2)}}, \phi_{\overline{U}^{(T-2)}}\right\} \le \ldots \le \beta^T \max\left\{\phi_{V^{(0)}}, \phi_{U^{(0)}}, \phi_{\overline{U}^{(0)}}\right\}.$$

By Corollary 4.4.13, we obtain

$$\begin{aligned}\|U^{(0)} - U^{*(1)}\|_{\mathrm{F}} &\le \frac{3\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(0)} - V^{*(0)}\|_{\mathrm{F}} + \left(\frac{6}{\xi} + 1\right)\sigma_1\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \\ &\overset{\text{(i)}}{\le} \frac{3\sigma_1}{\sigma_k} \cdot \frac{\sigma_k^3}{12\xi\sigma_1^3} \cdot \frac{\sigma_k^2}{2\xi\sigma_1} + \left(\frac{6}{\xi} + 1\right)\frac{\sigma_k^2}{4\xi\sigma_1} \\ &\overset{\text{(ii)}}{=} \frac{\sigma_k^4}{8\xi^2\sigma_1^3} + \frac{3\sigma_k^2}{2\xi^2\sigma_1} + \frac{\sigma_k^2}{4\xi\sigma_1} \overset{\text{(iii)}}{\le} \frac{\sigma_k^2}{2\xi\sigma_1}, \end{aligned} \tag{4.4.36}$$

where (i) and (ii) come from Lemma 4.4.12, and (iii) comes from the definition of ξ and $\sigma_1 \ge \sigma_k$. Combining (4.4.36) with Lemma 4.4.12, we have

$$\left\{\phi_{V^{(0)}}, \phi_{U^{(0)}}, \phi_{\overline{U}^{(0)}}\right\} \le \max\left\{\frac{\sigma_k^2}{2\xi\sigma_1}, \frac{\sigma_k^2}{4\xi\sigma_1^2}\right\}.$$

Then we need at most

$$T = \left\lceil \log\left(\max\left\{ \frac{\sigma_k^2}{\xi\sigma_1}, \frac{\sigma_k^2}{2\xi\sigma_1^2}, \frac{\sigma_k^2}{\xi}, \frac{\sigma_k^2}{2\xi\sigma_1} \right\} \cdot \frac{1}{\epsilon} \right) \log^{-1}(\beta^{-1}) \right\rceil$$

iterations such that

$$\|\overline{V}^{(T)} - \overline{V}^*\|_{\mathrm{F}} \leq \beta^T \max\left\{ \frac{\sigma_k^2}{2\xi\sigma_1}, \frac{\sigma_k^2}{4\xi\sigma_1^2} \right\} \leq \frac{\epsilon}{2} \text{ and } \|U^{(T)} - U^*\|_{\mathrm{F}} \leq \beta^T \max\left\{ \frac{\sigma_k^2}{2\xi\sigma_1}, \frac{\sigma_k^2}{4\xi\sigma_1^2} \right\} \leq \frac{\epsilon}{2\sigma_1}.$$

We then follow similar lines to (4.4.12) in Section 4.4.2, and show $\|M^{(T)} - M^*\|_{\mathrm{F}} \leq \epsilon$, which completes the proof. □

4.4.4 Proof of Theorem 4.3.3 (Gradient Descent)

Proof. The convergence analysis of the gradient descent algorithm is similar to that of the alternating gradient descent. The only difference is that for updating U, the gradient descent algorithm employs $V = \overline{V}^{(t)}$ instead of $V = \overline{V}^{(t+1)}$ to calculate the gradient at $U = U^{(t)}$. Then everything else directly follows Section 4.4.3, and is therefore omitted. □

4.5 Extensions to Matrix Completion

We then extend our methodology and theory to matrix completion problems. Let $M^* \in \mathbb{R}^{m \times n}$ be the unknown low rank matrix of interest. We observe a subset

of the entries of M^*, namely, $\mathcal{W} \subseteq \{1,\dots,m\} \times \{1,\dots,n\}$. We assume that $\mathcal{W}$ is drawn uniformly at random, i.e., $M^*_{i,j}$ is observed independently with probability $\bar{\rho} \in (0,1]$. To exactly recover M^*, a common assumption is the incoherence of M^*, which will be specified later. A popular approach for recovering M^* is to solve the following convex optimization problem

$$\min_{M \in \mathbb{R}^{m\times n}} \|M\|_* \quad \text{subject to } \mathcal{P}_{\mathcal{W}}(M^*) = \mathcal{P}_{\mathcal{W}}(M), \tag{4.5.1}$$

where $\mathcal{P}_{\mathcal{W}}(M) : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is an operator defined as

$$[\mathcal{P}_{\mathcal{W}}(M)]_{ij} = \begin{cases} M_{ij} & \text{if } (i,j) \in \mathcal{W}, \\ 0 & \text{otherwise.} \end{cases}$$

Similar to matrix sensing, existing algorithms for solving (4.5.1) are computationally inefficient. Hence, in practice we usually consider the following nonconvex optimization problem

$$\min_{U\in\mathbb{R}^{m\times k}, V\in\mathbb{R}^{n\times k}} \mathcal{F}_{\mathcal{W}}(U,V), \text{ where } \mathcal{F}_{\mathcal{W}}(U,V) = \frac{1}{2}\|\mathcal{P}_{\mathcal{W}}(M^*) - \mathcal{P}_{\mathcal{W}}(UV^\top)\|_{\mathrm{F}}^2. \tag{4.5.2}$$

Similar to matrix sensing, (4.5.2) can also be efficiently solved by gradient-based algorithms illustrated in Algorithm 8. For the convenience of later convergence analysis, we partition the observation set $\mathcal{W}$ into $2T+1$ subsets $\mathcal{W}_0$,...,$\mathcal{W}_{2T}$ by Algorithm 10. However, in practice we do not need the partition scheme, i.e., we

simply set $\mathcal{W}_0 = \cdots = \mathcal{W}_{2T} = \mathcal{W}$.

Algorithm 8: A family of nonconvex optimization algorithms for matrix completion. The incoherence factorization algorithm IF$(\cdot)$ is illustrated in Algorithm 9, and the partition algorithm Partition$(\cdot)$, which is proposed by [98], is provided in Algorithm 10 of Appendix C.3 for the sake of completeness. The initialization procedures $\mathrm{INT}_{\overline{U}}(\cdot)$ and $\mathrm{INT}_{\overline{U}}(\cdot)$ are provided in Algorithm 11 and Algorithm 12 of Appendix C.4 for the sake of completeness. Here $\mathcal{F}_{\mathcal{W}}(\cdot)$ is defined in (4.5.2).

Input: $\mathcal{P}_{\mathcal{W}}(M^*)$ **Parameter**: Step size η, Total number of iterations T
$(\{\mathcal{W}_t\}_{t=0}^{2T}, \widetilde{\rho}) \leftarrow \mathrm{Partition}(\mathcal{W})$, $\mathcal{P}_{\mathcal{W}_0}(\widetilde{M}) \leftarrow \mathcal{P}_{\mathcal{W}_0}(M^*)$, and $\widetilde{M}_{ij} \leftarrow 0$ for all $(i,j) \notin \mathcal{W}_0$ $(\overline{U}^{(0)}, V^{(0)}) \leftarrow \mathrm{INT}_{\overline{U}}(\widetilde{M})$, $(\overline{V}^{(0)}, U^{(0)}) \leftarrow \mathrm{INT}_{\overline{V}}(\widetilde{M})$ **For** $t \leftarrow 0, 1, ..., T-1$

Updating V:

- Alternating Exact Minimization : $V^{(t+0.5)} \leftarrow \mathrm{argmin}_V \mathcal{F}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V)$
  $(\overline{V}^{(t+1)}, R_{\overline{V}}^{(t+0.5)}) \leftarrow \mathrm{IF}(V^{(t+0.5)})$
- Alternating Gradient Descent : $V^{(t+0.5)} \leftarrow V^{(t)} - \eta \nabla_V \mathcal{F}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V^{(t)})$
  $(\overline{V}^{(t+1)}, R_{\overline{V}}^{(t+0.5)}) \leftarrow \mathrm{IF}(V^{(t+0.5)})$, $U^{(t)} \leftarrow \overline{U}^{(t)} R_{\overline{V}}^{(t+0.5)\top}$
- Gradient Descent : $V^{(t+0.5)} \leftarrow V^{(t)} - \eta \nabla_V \mathcal{F}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V^{(t)})$
  $(\overline{V}^{(t+1)}, R_{\overline{V}}^{(t+0.5)}) \leftarrow \mathrm{IF}(V^{(t+0.5)})$, $U^{(t+1)} \leftarrow \overline{U}^{(t)} R_{\overline{V}}^{(t+0.5)\top}$

Updating U:

- Alternating Exact Minimization : $U^{(t+0.5)} \leftarrow \mathrm{argmin}_U \mathcal{F}_{\mathcal{W}_{2t+2}}(U, \overline{V}^{(t+1)})$
  $(\overline{U}^{(t+1)}, R_{\overline{U}}^{(t+0.5)}) \leftarrow \mathrm{IF}(U^{(t+0.5)})$
- Alternating Gradient Descent : $U^{(t+0.5)} \leftarrow U^{(t)} - \eta \nabla_U \mathcal{F}_{\mathcal{W}_{2t+2}}(U^{(t)}, \overline{V}^{(t+1)})$
  $(\overline{U}^{(t+1)}, R_{\overline{U}}^{(t+0.5)}) \leftarrow \mathrm{IF}(U^{(t+0.5)})$, $V^{(t+1)} \leftarrow \overline{V}^{(t+1)} R_{\overline{U}}^{(t+0.5)\top}$
- Gradient Descent : $U^{(t+0.5)} \leftarrow U^{(t)} - \eta \nabla_U \mathcal{F}_{\mathcal{W}_{2t+2}}(U^{(t)}, \overline{V}^{(t)})$
  $(\overline{U}^{(t+1)}, R_{\overline{U}}^{(t+0.5)}) \leftarrow \mathrm{IF}(U^{(t+0.5)})$, $V^{(t+1)} \leftarrow \overline{V}^{(t)} R_{\overline{U}}^{(t+0.5)\top}$

Return: $M^{(T)} \leftarrow U^{(T-0.5)} \overline{V}^{(T)\top}$ (for gradient descent we use $\overline{U}^{(T)} V^{(T)\top}$)

Before we present the convergence analysis, we first introduce an assumption known as the incoherence property.

Assumption 4.5.1 (Incoherence Property)**.** The target rank k matrix M^* is incoherent with parameter μ, i.e., given the rank k singular value decomposition of $M^* =$

Algorithm 9: The incoherence factorization algorithm for matrix completion. It guarantees that the solutions satisfy the incoherence condition throughout all iterations.

Input: W^{in}
$r \leftarrow$ Number of rows of W^{in}
$k \leftarrow$ Number of columns of W^{in}
Parameter: Incoherence parameter μ
$(\overline{W}^{\text{in}}, R^{\text{in}}_{\overline{W}}) \leftarrow \text{QR}(W^{\text{in}})$
$\widetilde{W} \leftarrow \underset{W}{\text{argmin}} \|W - \overline{W}^{\text{in}}\|_{\text{F}}^2$ subject to $\max_j \|W_{j*}\|_2 \le \mu\sqrt{k/r}$
$W^{\text{out}} \leftarrow \widetilde{W} R^{\text{in}}_W$ $(\overline{W}^{\text{out}}, R^{\text{tmp}}_{\overline{W}}) \leftarrow \text{QR}(W^{\text{out}})$
$R^{\text{out}}_{\overline{W}} = \overline{W}^{\text{out}\top} W^{\text{in}}$
Return: $\overline{W}^{\text{out}}, R^{\text{out}}_{\overline{W}}$

$\overline{U}^* \Sigma^* \overline{V}^{*\top}$, we have

$$\max_i \|\overline{U}^*_{i*}\|_2 \le \mu\sqrt{\frac{k}{m}} \quad \text{and} \quad \max_j \|\overline{V}^*_{j*}\|_2 \le \mu\sqrt{\frac{k}{n}}.$$

Roughly speaking, the incoherence assumption guarantees that each entry of M^* contains similar amount of information, which makes it feasible to complete M^* when its entries are missing uniformly at random. The following theorem establishes the iteration complexity and the estimation error under the Frobenius norm.

Theorem 4.5.2. Suppose that there exists a universal constant C_4 such that $\bar{\rho}$ satisfies

$$\bar{\rho} \ge \frac{C_4 \mu^2 k^3 \log n \log(1/\epsilon)}{m}, \tag{4.5.3}$$

where ϵ is the pre-specified precision. Then there exist an η and universal constants C_5 and C_6 such that for any $T \geq C_5 \log(C_6/\epsilon)$, we have $\|M^{(T)} - M\|_{\mathrm{F}} \leq \epsilon$ with high probability.

The proof of Theorem 4.5.2 is provided in Appendices C.5.1, C.5.2, and C.5.3. Theorem 4.5.2 implies that all three nonconvex optimization algorithms converge to the global optimum at a geometric rate. Furthermore, our results indicate that the completion of the true low rank matrix M^* up to ϵ-accuracy requires the entry observation probability $\bar{\rho}$ to satisfy

$$\bar{\rho} = \Omega(\mu^2 k^3 \log n \log(1/\epsilon)/m). \tag{4.5.4}$$

This result matches the result established by [96], which is the state-of-the-art result for alternating minimization. Moreover, our analysis covers three nonconvex optimization algorithms.

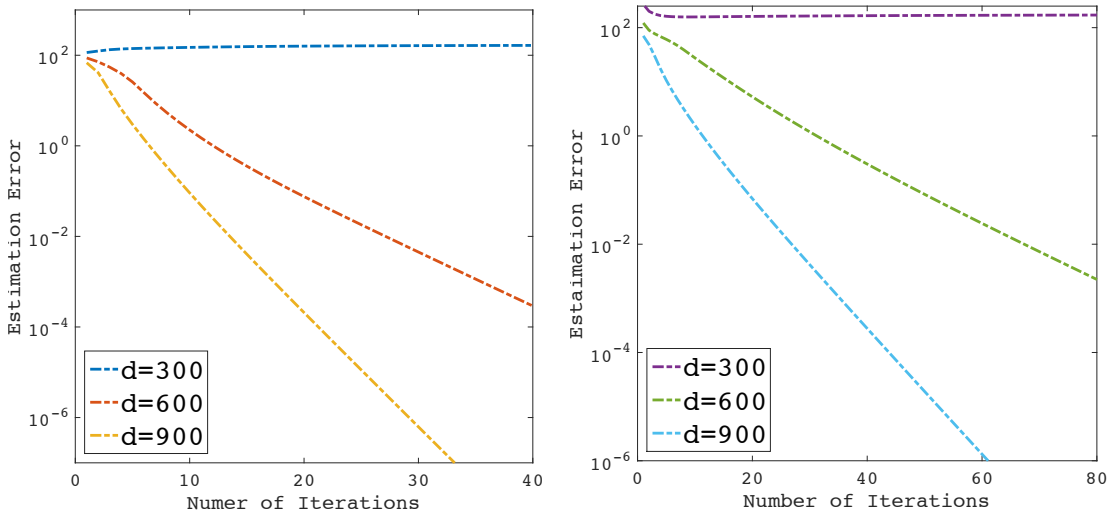
In fact, the sample complexity in (4.5.4) depends on a polynomial of $\frac{\sigma_{\max}(M^*)}{\sigma_{\min}(M^*)}$, which is a constant since in this chapter we assume that $\sigma_{\max}(M^*)$ and $\sigma_{\min}(M^*)$ are constants. If we allow $\frac{\sigma_{\max}(M^*)}{\sigma_{\min}(M^*)}$ to increase, we can replace the QR decomposition in Algorithm 9 with the smooth QR decomposition proposed by [98] and achieve a dependency of $\log\left(\frac{\sigma_{\max}(M^*)}{\sigma_{\min}(M^*)}\right)$ on the condition number with a more involved proof. See more details in [98]. However, in this chapter, our primary focus is on the dependency on k, n and m, rather than optimizing over the dependency on condition

number.

4.6 Numerical Experiments

We present numerical experiments to support our theoretical analysis. We first consider a matrix sensing problem with $m = 30$, $n = 40$, and $k = 5$. We vary d from 300 to 900. Each entry of A_i's are independent sampled from $N(0,1)$. We then generate $M = UV^\top$, where $\widetilde{U} \in \mathbb{R}^{m\times k}$ and $\widetilde{V} \in \mathbb{R}^{n\times k}$ are two matrices with all their entries independently sampled from $N(0,1/k)$. We then generate d measurements by $b_i = \langle A_i, M\rangle$ for $i = 1,...,d$. Figure 4.1 illustrates the empirical performance of the alternating exact minimization and alternating gradient descent algorithms for a single realization. The step size for the alternating gradient descent algorithm is determined by the backtracking line search procedure. We see that both algorithms attain linear rate of convergence for $d = 600$ and $d = 900$. Both algorithms fail for $d = 300$, because $d = 300$ is below the minimum requirement of sample complexity for the exact matrix recovery.

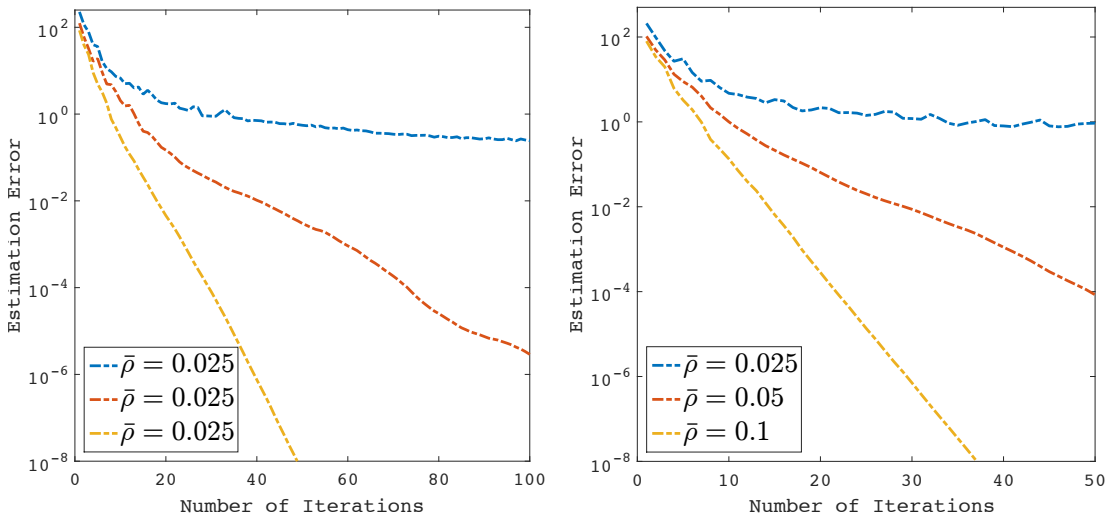
We then consider a matrix completion problem with $m = 1000$, $n = 50$, and $k = 5$. We vary $\bar{\rho}$ from 0.025 to 0.1. We then generate $M = UV^\top$, where $\widetilde{U} \in \mathbb{R}^{m\times k}$ and $\widetilde{V} \in \mathbb{R}^{n\times k}$ are two matrices with all their entries independently sampled from $N(0,1/k)$. The observation set is generated uniformly at random with probability $\bar{\rho}$. Figure 4.2 illustrates the empirical performance of the alternating exact mini-



(a) Alternating Exact Minimization Algorithm (b) Alternating Gradient Descent Algorithm

Figure 4.1: Two illustrative examples for matrix sensing. The vertical axis corresponds to estimation error $\|M^{(t)} - M\|_{\mathrm{F}}$. The horizontal axis corresponds to numbers of iterations. Both the alternating exact minimization and alternating gradient descent algorithms attain linear rate of convergence for $d = 600$ and $d = 900$. But both algorithms fail for $d = 300$, because the sample size is not large enough to guarantee proper initial solutions.

mization and alternating gradient descent algorithms for a single realization. The step size for the alternating gradient descent algorithm is determined by the backtracking line search procedure. We see that both algorithms attain linear rate of convergence for $\bar{\rho} = 0.05$ and $\bar{\rho} = 0.1$. Both algorithms fail for $\bar{\rho} = 0.025$, because the entry observation probability is below the minimum requirement of sample complexity for the exact matrix recovery.



(a) Alternating Exact Minimization Algorithm (b) Alternating Gradient Descent Algorithm

Figure 4.2: Two illustrative examples for matrix completion. The vertical axis corresponds to estimation error $\|M^{(t)} - M\|_{\mathrm{F}}$. The horizontal axis corresponds to numbers of iterations. Both the alternating exact minimization and alternating gradient descent algorithms attain linear rate of convergence for $\bar{\rho} = 0.05$ and $\bar{\rho} = 0.1$. But both algorithms fail for $\bar{\rho} = 0.025$, because the entry observation probability is not large enough to guarantee proper initial solutions.

Chapter 5

Conclusions

We propose a new class of model-based nonconvex optimization algorithms for solving various machine learning problems, including high dimensional sparse learning and matrix factorization. By analyzing data generating process of the underlying statistical distribution, we exploit the hidden convexity behind the nonconvex optimization problem to tackle computational challenges. Specifically, we show two types of hidden convexity: Restricted Strong Convexity in Chapters 2 and 3 as well as Strong Bi-convexity in Chapter 4. Different from the worse-case analysis in existing optimization and computational theory, our theoretical analysis shows that with high probability, our proposed algorithms attain linear convergence to global or approximately global optima, which enjoys strong statistical guarantees.

Moreover, by investigating the gap between convex and nonconvex approaches,

we show that the convex approaches may lead to sub-optimal statistical performance. Specifically, in Chapter 2, we show that Lasso only attains suboptimal statistical rates of convergence in parameter estimation for sparse linear regression. In contrast, the nonconvex approaches attains the optimal performance.

In summary, our proposed model-based nonconvex optimization framework is a systematic approach for designing and analyzing algorithms to learn from large-scale data.

Appendix A

Supporting Proof for Chapter 2

A.1 Computational Complexity Comparison

We first show that the computational complexity of each proximal gradient iteration is $\mathcal{O}(nd)$. At the t-th iteration, we calculate

$$\theta^{(t+1)} = \mathcal{S}_{\lambda/L}\left(\theta^{(t+1)} - \frac{1}{Ln} X^\top \left(y^{(t)} - X\theta^{(t)}\right)\right),$$

where L is the step size parameter. Thus, the computational complexity is $\mathcal{O}(ns + nd + d + d) = \mathcal{O}(nd)$, where $s = \|\theta^{(t)}\|_0 \leq d$.

We then show that the overall computational complexity of each coordinate minimization iteration is only $\mathcal{O}(n)$. Suppose we maintain $\widetilde{y}^{(t)} = X_{*\setminus j}\theta^{(t)}_{\setminus j}$ for the

t-th iteration. Then we calculate $\theta_j^{(t+1)}$ by

$$\theta_j^{(t+1)} = \widetilde{\theta}_j^{(t)} \cdot \mathbb{1}_{\{|\widetilde{\theta}_j^{(t)}| \geq \gamma\lambda\}} + \frac{\mathcal{S}_\lambda(\widetilde{\theta}_j^{(t)})}{1 - 1/\gamma} \cdot \mathbb{1}_{\{|\widetilde{\theta}_j^{(t)}| < \gamma\lambda\}}, \tag{A.1.1}$$

where $\widetilde{\theta}_j^{(t)} = \frac{1}{n} X_{*j}^\top (y - \widetilde{y}^{(t)})$. Thus, the computational complexity of (A.1.1) is $\mathcal{O}(n)$. Once we have $\widetilde{\theta}_j^{(t)}$, we obtain $\widetilde{y}^{(t+1)}$ for the $(t+1)$ iteration by

$$\widetilde{y}^{(t+1)} = \widetilde{y}^{(t)} + X_{*j}(\theta_j^{(t+1)} - \theta_j^{(t)}),$$

and the computational complexity is also $\mathcal{O}(n)$. Thus the overall computational complexity is $\mathcal{O}(n)$. For proximal coordinate gradient algorithms, the coordinate gradient can be computed using a similar strategy, and therefore its overall computational complexity is also $\mathcal{O}(n)$ for each iteration.

A.2 The MCP regularizer

Throughout our analysis, we frequently use the following properties of the MCP regularizer.

Lemma A.2.1. For the MCP regularizer, $h(\cdot)$ and $h'(\cdot)$ satisfy:

(R.1) For any $a > b \geq 0$, we have

$$-\alpha(a - b) \leq h'_\lambda(a) - h'_\lambda(b) \leq 0,$$

where $\alpha = 1/\gamma \geq 0$;

(R.2) For some $\gamma > 0$ and $\forall\ a \geq 0$, we have $h'_\lambda(a) \in [-\lambda, 0]$ if $a \leq \lambda\gamma$, and $h'_\lambda(a) = -\lambda$ otherwise;

(R.3) $h_\lambda(\cdot)$ and $h'_\lambda(\cdot)$ pass through the origin, i.e., $h_\lambda(0) = 0$ and $h'_\lambda(0) = 0$;

(R.4) For $\forall\ a \geq 0$, we have $|h'_{\lambda_1}(a) - h'_{\lambda_2}(a)| \leq |\lambda_1 - \lambda_2|$.

The proof of Lemma A.2.1 is straightforward, and therefore omitted. Note that all above properties also hold for Lasso, i.e., $\gamma = \infty$ and $h_\lambda(\cdot) = 0$.

A.3 Lemmas for Computational Theory

A.3.1 Proof of Lemma 2.3.4

Proof. Since $\mathcal{L}(\theta)$ is twice differentiable and $\|\theta - \theta'\|_0 \leq s$, by the mean value theorem, we have

$$\mathcal{L}(\theta') - \mathcal{L}(\theta) - (\theta' - \theta)^\top \nabla \mathcal{L}(\theta) = \frac{1}{2}(\theta' - \theta)^\top \nabla^2 \mathcal{L}(\widetilde{\theta})(\theta' - \theta), \tag{A.3.1}$$

where $\widetilde{\theta} = (1 - \beta)\theta' + \beta\theta$ for some $\beta \in (0, 1)$. By Definition 2.3.3, we have

$$\frac{\rho_-(s)}{2}\|\theta' - \theta\|_2^2 \leq \frac{1}{2}(\theta' - \theta)^\top \nabla^2 \mathcal{L}(\widetilde{\theta})(\theta' - \theta) \leq \frac{\rho_+(s)}{2}\|\theta' - \theta\|_2^2. \tag{A.3.2}$$

Combining (A.3.1) with (A.3.2), we have

$$\frac{\rho_-(s)}{2}\|\theta' - \theta\|_2^2 \leq \mathcal{L}(\theta') - \mathcal{L}(\theta) - (\theta' - \theta)^\top \nabla \mathcal{L}(\theta) \leq \frac{\rho_+(s)}{2}\|\theta' - \theta\|_2^2. \tag{A.3.3}$$

By (R.1) in Assumption A.2.1, we have

$$-\frac{\alpha}{2}\|\theta' - \theta\|_2^2 \leq \mathcal{H}_\lambda(\theta') - \mathcal{H}_\lambda(\theta) - (\theta' - \theta)^\top \nabla \mathcal{H}_\lambda(\theta) \leq 0. \tag{A.3.4}$$

Combining (A.3.3) with (A.3.4), we have

$$\begin{aligned}\frac{\rho_-(s) - \alpha}{2}\|\theta' - \theta\|_2^2 \leq \widetilde{\mathcal{L}}_\lambda(\theta') - \widetilde{\mathcal{L}}_\lambda(\theta) - (\theta' - \theta)^\top \nabla \widetilde{\mathcal{L}}_\lambda(\theta) \\ \leq \frac{\rho_+(s)}{2}\|\theta' - \theta\|_2^2.\end{aligned} \tag{A.3.5}$$

By the convexity of $\|\theta\|_1$, we have

$$\|\theta'\|_1 \geq \|\theta\|_1 + (\theta' - \theta)^\top \xi \tag{A.3.6}$$

for any $\xi \in \partial\|\theta\|_1$. Combining (A.3.6) with (A.3.5), we obtain

$$\mathcal{F}_\lambda(\theta') \geq \mathcal{F}_\lambda(\theta) + (\theta' - \theta)^\top (\nabla \widetilde{\mathcal{L}}_\lambda(\theta) + \lambda\xi) + \frac{\rho_-(s)}{2}\|\theta' - \theta\|_2^2.$$

□

A.3.2 Proof of Lemma 2.7.1

Proof. By Lemma 2.7.8, we have

$$\begin{aligned}\mathcal{F}_\lambda(w^{(t+1,k-1)}) - \mathcal{F}_\lambda(w^{(t+1,k)}) &\geq \frac{\nu_-(1)}{2}(w_k^{(t+1,k-1)} - w_k^{(t+1,k)})^2 \\ &= \frac{\nu_-(1)}{2}(\theta_k^{(t+1)} - \theta_k^{(t)})^2,\end{aligned}$$

which further implies

$$\begin{aligned}\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\theta^{(t+1)}) &= \sum_{k=1}^{s}[\mathcal{F}_\lambda(w^{(t+1,k-1)}) - \mathcal{F}_\lambda(w^{(t+1,k)})] \\ &\geq \frac{\nu_-(1)}{2}\|\theta^{(t)} - \theta^{(t+1)}\|_2^2.\end{aligned}$$

□

A.3.3 Proof of Lemma 2.7.2

Proof. We first analyze the gap for the proximal coordinate gradient descent. Let $\theta \in \mathbb{R}^d$ be a vector satisfying $\theta_{\overline{\mathcal{A}}} = 0$. By the restricted convexity of $\mathcal{F}_\lambda(\theta)$, we have

$$\begin{aligned}\mathcal{F}_\lambda(\theta) \geq \mathcal{F}_\lambda(\theta^{(t+1)}) + (\nabla_{\mathcal{A}}\widetilde{\mathcal{L}}_\lambda(\theta^{(t+1)}) + \lambda\xi_{\mathcal{A}}^{(t+1)})^\top(\theta - \theta^{(t+1)}) \\ + \frac{\widetilde{\rho}_-(s)}{2}\|\theta - \theta^{(t+1)}\|_2^2,\end{aligned} \tag{A.3.7}$$

where $\xi_{\mathcal{A}}^{(t+1)}$ satisfies the optimality condition of the proximal coordinate gradient descent,

$$\nabla \mathcal{V}_{\lambda,k,L}(\theta_k^{(t+1)}; w^{(t+1,k-1)}) + \lambda \xi_k^{(t+1)} = 0 \text{ for any } k \in \mathcal{A}. \tag{A.3.8}$$

By setting $\theta_{\overline{\mathcal{A}}} = 0$ and minimizing both sides of (A.3.7) over $\theta_{\mathcal{A}}$, we obtain

$$\begin{aligned} \mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(\overline{\theta}) &\leq \frac{1}{2\widetilde{\rho}_-(s)} \|\nabla_{\mathcal{A}} \widetilde{\mathcal{L}}_\lambda(\theta^{(t+1)}) + \lambda \xi_{\mathcal{A}}^{(t+1)}\|_2^2 \\ &\overset{\text{(i)}}{=} \frac{1}{2\widetilde{\rho}_-(s)} \sum_{k=1}^{s} \|\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{(t+1)}) - \nabla \mathcal{V}_{\lambda,k,L}(\theta_k^{(t+1)}; w^{(t+1,k-1)})\|_2^2 \\ &\overset{\text{(ii)}}{\leq} \frac{\rho_+^2(s)}{2\widetilde{\rho}_-(s)} \sum_{k=1}^{s} \|\theta^{(t+1)} - w^{(t+1,k-1)}\|_2^2 \leq \frac{s\rho_+^2(s)}{2\widetilde{\rho}_-(s)} \|\theta^{(t+1)} - \theta^{(t)}\|_2^2, \end{aligned} \tag{A.3.9}$$

where (i) comes from (A.3.8), and (ii) comes from $\nabla \mathcal{V}_{\lambda,k,L}(\theta_k^{(t+1)}; w^{(t+1,k-1)}) = \nabla \widetilde{\mathcal{L}}_\lambda(w^{(t+1,k-1)})$ and the restricted smoothness of $\widetilde{\mathcal{L}}_\lambda(\theta)$.

For the exact coordinate minimization, we have $\nabla \mathcal{V}_{\lambda,k,L}(\theta_k^{(t+1)}; w^{(t+1,k-1)}) = \nabla \mathcal{Y}_{\lambda,k}(\theta_k^{(t+1)}; w^{(t+1,k-}$ Thus, (A.3.9) also holds.

□

A.3.4 Proof of Lemma 2.7.8

Proof. For the proximal coordinate gradient descent, we have

$$\mathcal{F}_\lambda(\theta) = \mathcal{V}_{\lambda,k,L}(\theta_k;\theta) + \lambda|\theta_k| + \lambda\|\theta_{\setminus k}\|_1, \tag{A.3.10}$$

$$\mathcal{F}_\lambda(w) \le \mathcal{V}_{\lambda,k,L}(w_k;\theta) + \lambda|\theta'_k| + \lambda\|\theta_{\setminus k}\|_1. \tag{A.3.11}$$

Since $\mathcal{V}_{\lambda,k,L}(\theta_k;\theta)$ is strongly convex in θ_k, we have

$$\begin{aligned}\mathcal{V}_{\lambda,k,L}(\theta_k;\theta) - \mathcal{V}_{\lambda,k,L}(w_k;\theta) \\ \ge (\theta_k - w_k)\nabla\mathcal{V}_{\lambda,k,L}(w_k;\theta) + \frac{L}{2}(w_k-\theta_k)^2.\end{aligned} \tag{A.3.12}$$

By the convexity of the absolute value function, we have

$$|\theta_k| - |w_k| \ge (\theta_k - w_k)\xi_k, \tag{A.3.13}$$

where $\xi_k \in \partial|w_k|$ satisfies the optimality condition of the proximal coordinate gradient descent,

$$\nabla\mathcal{V}_{\lambda,k,L}(w_k;\theta) + \lambda\xi_k = 0. \tag{A.3.14}$$

Subtracting (A.3.10) by (A.3.11), we have

$$\begin{aligned}\mathcal{F}_\lambda(\theta)-\mathcal{F}_\lambda(w) &\geq \mathcal{V}_{\lambda,k,L}(\theta_k;\theta)-\mathcal{V}_{\lambda,k,L}(w_k;\theta)+\lambda|\theta_k|-\lambda|w_k| \\ &\overset{\text{(i)}}{\geq}(\theta_k-w_k)(\nabla\mathcal{V}_{\lambda,k,L}(w_k;\theta)+\lambda\xi_k)+\frac{L}{2}(w_k-\theta_k)^2 \overset{\text{(ii)}}{\geq}\frac{L}{2}(w_k-\theta_k)^2.\end{aligned}$$

where (i) comes from (A.3.12) and (A.3.13), and (ii) comes from (A.3.14).

For the exact coordinate minimization, we only need to slightly trim the above analysis. Specifically, we replace $\mathcal{V}_{\lambda,k,L}(w_k;\theta)$ with

$$\mathcal{Y}_{\lambda,k}(w_k;\theta)=\widetilde{\mathcal{L}}_\lambda(w_k,\theta_{\backslash k}).$$

Since $\widetilde{\mathcal{L}}_\lambda(\theta)$ is restrictedly convex, we have

$$\mathcal{Y}_{\lambda,k}(\theta_k;\theta)-\mathcal{Y}_{\lambda,k}(w_k;\theta)\geq(\theta_k-w_k)\nabla\mathcal{Y}_{\lambda,k}(\theta_k';\theta)+\frac{\widetilde{\rho}_-(1)}{2}(w_k-\theta_k)^2.$$

Eventually, we obtain

$$\mathcal{F}_\lambda(w)-\mathcal{F}_\lambda(\theta)\geq\frac{\widetilde{\rho}_-(1)}{2}(w_k-\theta_k)^2.$$

We then proceed to analyze the descent for the proximal coordinate gradient

descent when $\theta_k = 0$ and $|\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta)| \geq (1+\delta)\lambda$. Then we have

$$|w_k| = |\mathcal{S}_{\lambda/L}(-\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta)/L)| \geq \frac{\delta\lambda}{L},$$

where the last inequality comes from the definition of the soft thresholding function. Thus, we obtain

$$\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(w) \geq \frac{L}{2} w_k^2 \geq \frac{\delta^2\lambda^2}{2L}.$$

For the exact coordinate minimization, we construct an auxiliary solution w' by a proximal coordinate gradient descent iteration using $L = \rho_+(1)$. Since w is obtained by the exact minimization, we have

$$\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(w) \geq \mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(w') \geq \frac{\delta^2\lambda^2}{2\rho_+(1)}.$$

□

A.3.5 Proof of Lemma 2.7.3

Proof. Before we proceed, we first introduce the following lemma.

Lemma A.3.1. Suppose Assumption (2.3.5) holds. If $\overline{\theta}^{\lambda}$ satisfies

$$\|\overline{\theta}^{\lambda}_{\overline{S}}\|_0 \leq \widetilde{s} \quad \text{and} \quad \mathcal{K}_{\lambda}(\overline{\theta}^{\lambda}) = 0,$$

then $\overline{\theta}^{\lambda}$ is a unique sparse local optimum to (2.1.1).

The proof of Lemma is provided in Appendix A.3.13. We then proceed with the proof. We consider a sequence of auxiliary solutions obtained by the proximal gradient algorithm. The details for generating such a sequence are provided in [31]. By Theorem 5.1 in [31], we know that such a sequence of solutions converges to a sparse local optimum $\overline{\theta}^{\lambda}$. By Lemma A.3.1, we know that the sparse local optimum is unique. □

A.3.6 Proof of Lemma 2.7.4

Proof. Before we proceed, we first introduce the following lemma.

Lemma A.3.2. Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. For any $\lambda \geq \lambda_N$, if θ satisfies

$$\|\theta_{\overline{S}}\|_0 \leq s \quad \text{and} \quad \mathcal{F}_{\lambda}(\theta) \leq \mathcal{F}_{\lambda}(\theta^*) + \frac{4\lambda^2 s^*}{\widetilde{\rho}_{-}(s^* + s)}, \tag{A.3.15}$$

where $s \leq 2\widetilde{s}$, then we have

$$\|\theta - \theta^*\|_2 \leq \frac{9\lambda\sqrt{s^*}}{\widetilde{\rho}_-(s^*+s)} \quad \text{and} \quad \|\theta - \theta^*\|_1 \leq \frac{25\lambda s^*}{\widetilde{\rho}_-(s^*+s)}.$$

The proof of Lemma A.3.2 is provided in Appendix A.3.7. Lemma A.3.2 characterizes the estimation errors of any sufficiently sparse solution with a sufficiently small objective value.

When the inner loop terminates, we have the output solution as $\widehat{\theta} = \theta^{(t+1)}$. Since both the exact coordinate minimization and proximal coordinate gradient descent iterations always decrease the objective value, we have

$$\mathcal{F}_\lambda(\theta^{(t+1)}) \leq \mathcal{F}_\lambda(\theta^*) + \frac{4\lambda^2 s^*}{\widetilde{\rho}_-(s^*+2\widetilde{s})}. \tag{A.3.16}$$

By (A.3.9) in Appendix A.3.3, we have shown

$$\|\nabla_{\mathcal{A}}\widetilde{\mathcal{L}}_\lambda(\theta^{(t+1)}) + \lambda\xi_{\mathcal{A}}^{(t+1)}\|_2^2 \leq (s^*+2\widetilde{s})\rho_+^2(s^*+2\widetilde{s})\|\theta^{(t+1)} - \theta^{(t)}\|_2^2. \tag{A.3.17}$$

Since Assumption 2.3.7 holds and $\widetilde{\rho}_-(1) \leq \nu_+(1)$, we have

$$\|\theta^{(t+1)} - \theta^{(t)}\|_2^2 \leq \tau^2\lambda^2 \leq \frac{\delta^2\lambda^2}{(s^*+2\widetilde{s})\rho_+^2(s^*+2\widetilde{s})}. \tag{A.3.18}$$

Combining (A.3.17) with (A.3.18), we have $\theta^{(t+1)}$ satisfying the approximate KKT

condition over the active set,

$$\min_{\xi_{\mathcal{A}} \in \partial \|\theta_{\mathcal{A}}^{(t+1)}\|_1} \|\nabla_{\mathcal{A}} \widetilde{\mathcal{L}}_\lambda(\theta^{(t+1)}) + \lambda \xi_{\mathcal{A}}\|_\infty \leq \|\nabla_{\mathcal{A}} \widetilde{\mathcal{L}}_\lambda(\theta^{(t+1)}) + \lambda \xi_{\mathcal{A}}^{(t+1)}\|_2 \leq \delta\lambda.$$

We now proceed to characterize the sparsity of $\widehat{\theta} = \theta^{(t+1)}$ by exploiting the above approximate KKT condition. By Assumption 2.3.1, we have $\lambda \geq 4\|\nabla \widetilde{\mathcal{L}}_\lambda(\theta^*)\|_\infty$, which implies

$$\left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^*)| \geq \lambda/4,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| = 0. \tag{A.3.19}$$

We then consider an arbitrary set $\mathcal{S}'$ such that

$$\mathcal{S}' = \left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) - \nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^*)| \geq \lambda/2,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}.$$

Let $s' = |\mathcal{S}'|$. There exists a $v \in \mathbb{R}^d$ such that

$$\|v\|_\infty = 1,\ \ \|v\|_0 \leq s', \quad \text{and} \quad s'\lambda/2 \leq v^\top(\nabla \widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\theta^*)). \tag{A.3.20}$$

By Cauchy-Schwarz inequality, (A.3.20) implies

$$\begin{aligned} \frac{s'\lambda}{2} &\leq \|v\|_2 \|\nabla \widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\theta^*)\|_2 \leq \sqrt{s'}\|\nabla \widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\theta^*)\|_2 \\ &\overset{\text{(i)}}{\leq} \rho_+(s^* + 2\widetilde{s})\sqrt{s'}\|\widehat{\theta} - \theta^*\|_2 \overset{\text{(ii)}}{\leq} \rho_+(s^* + 2\widetilde{s})\sqrt{s'}\frac{9\lambda\sqrt{s^*}}{\widetilde{\rho}_-(s^* + 2\widetilde{s})}, \end{aligned} \tag{A.3.21}$$

where (i) comes from the restricted smoothness of $\widetilde{\mathcal{L}}_\lambda(\theta)$, and (ii) comes from (A.3.16) and Lemma A.3.2. (A.3.21) further implies

$$\sqrt{s'} \leq \frac{18\rho_+(s^* + 2\widetilde{s})\sqrt{s^*}}{\widetilde{\rho}_-(s^* + 2\widetilde{s})}. \tag{A.3.22}$$

Since $\mathcal{S}'$ is arbitrary defined, by simple manipulation, (A.3.22) implies

$$\left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) - \nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^*)| \geq \lambda/2,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| \leq 364\kappa^2 s^*. \tag{A.3.23}$$

Combining (A.3.19) with (A.3.23), we have

$$\begin{aligned} &\left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\widehat{\theta})| \geq 3\lambda/4,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| \\ &\leq \left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^*)| \geq \lambda/4,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| \\ &+ \left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) - \nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^*)| \geq \lambda/2,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| \leq 364\kappa^2 s^* < \widetilde{s}, \end{aligned} \tag{A.3.24}$$

where the last inequality comes from Assumption 2.3.5. Since we require $\delta \leq 1/8$ in Assumption 2.3.7, (A.3.24) implies that for any $u \in \mathbb{R}^d$ satisfying $\|u\|_\infty \leq 1$, we have

$$\left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) + \delta\lambda u_j| \geq 7\lambda/8,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| \leq \widetilde{s}.$$

Then for any $j \in \overline{\mathcal{S}} \cap \mathcal{A}$ satisfying $|\nabla_j \widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) + \delta\lambda u_j| \leq 7\lambda/8$, there exists a ξ_j such that

$$|\xi_j| \leq 1 \quad \text{and} \quad \nabla_j \widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) + \delta\lambda u_j + \lambda\xi_j = 0,$$

which further implies $\widehat{\theta}_j = 0$. Thus, we must have $\|\widehat{\theta}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$. □

A.3.7 Proof of Lemma A.3.2

Proof. For notational simplicity, we define $\Delta = \theta - \theta^*$. We first rewrite (A.3.15) as

$$\lambda\|\theta^*\|_1 - \lambda\|\theta\|_1 + \frac{4\lambda^2 s^*}{\widetilde{\rho}_-(s^*+s)} \geq \widetilde{\mathcal{L}}_\lambda(\theta) - \widetilde{\mathcal{L}}_\lambda(\theta^*). \tag{A.3.25}$$

By the restricted convexity of $\widetilde{\mathcal{L}}_\lambda(\theta)$, we have

$$\begin{aligned} &\widetilde{\mathcal{L}}_\lambda(\theta) - \widetilde{\mathcal{L}}_\lambda(\theta^*) - \frac{\widetilde{\rho}_-(s^*+s)}{2}\|\Delta\|_2^2 \qquad (A.3.26)\\ &\overset{\text{(i)}}{\geq} \Delta_{\mathcal{S}}^\top[\nabla_{\mathcal{S}}\mathcal{L}(\theta^*) + \nabla_{\mathcal{S}}\mathcal{H}_\lambda(\theta^*)] + \Delta_{\overline{\mathcal{S}}}^\top \nabla_{\overline{\mathcal{S}}}\mathcal{L}(\theta^*)\\ &\overset{\text{(ii)}}{\geq} -\|\Delta_{\mathcal{S}}\|_1\|\nabla\mathcal{L}(\theta^*)\|_\infty - \|\Delta_{\overline{\mathcal{S}}}\|_1\|\nabla\mathcal{L}(\theta^*)\|_\infty - \|\Delta_{\mathcal{S}}\|_1\|\nabla_{\mathcal{S}}\mathcal{H}_\lambda(\theta^*)\|_\infty, \end{aligned}$$

where (i) comes from $\nabla_{\overline{\mathcal{S}}}\mathcal{H}_\lambda(\theta^*) = 0$ by (R.3) of Lemma A.2.1, and (ii) comes from Hölder's inequality. Assumption 2.3.1 and (R.2) of Lemma A.2.1 imply

$$\|\nabla\mathcal{L}(\theta^*)\|_\infty \leq \frac{\lambda}{4} \quad \text{and} \quad \|\nabla_{\mathcal{S}}\mathcal{H}_\lambda(\theta^*)\|_\infty \leq \lambda. \tag{A.3.27}$$

Combining (A.3.26) with (A.3.27), we obtain

$$\widetilde{\mathcal{L}}_\lambda(\theta) - \widetilde{\mathcal{L}}_\lambda(\theta^*) \geq -\frac{5\lambda}{4}\|\Delta_{\mathcal{S}}\|_1 - \frac{\lambda}{4}\|\Delta_{\overline{\mathcal{S}}}\|_1 + \frac{\widetilde{\rho}_-(s^*+s)}{2}\|\Delta\|_2^2. \tag{A.3.28}$$

Plugging (A.3.28) and

$$\|\theta^*\|_1 - \|\theta\|_1 = \|\theta^*_{\mathcal{S}}\|_1 - (\|\theta_{\mathcal{S}}\|_1 + \|\Delta_{\overline{\mathcal{S}}}\|_1) \leq \|\Delta_{\mathcal{S}}\|_1 - \|\Delta_{\overline{\mathcal{S}}}\|_1$$

into (A.3.25), we obtain

$$\frac{9\lambda}{4}\|\Delta_{\mathcal{S}}\|_1 + \frac{4\lambda^2 s^*}{\widetilde{\rho}_-(s^*+s)} \geq \frac{3\lambda}{4}\|\Delta_{\overline{\mathcal{S}}}\|_1 + \frac{\widetilde{\rho}_-(s^*+s)}{2}\|\Delta\|_2^2. \tag{A.3.29}$$

We consider the first case: $\widetilde{\rho}_-(s^*+s)\|\Delta\|_1 > 16\lambda s^*$. Then we have

$$\frac{5\lambda}{2}\|\Delta_{\mathcal{S}}\|_1 \geq \frac{\lambda}{2}\|\Delta_{\overline{\mathcal{S}}}\|_1 + \frac{\widetilde{\rho}_-(s^*+s)}{2}\|\Delta\|_2^2. \tag{A.3.30}$$

By simple manipulation, (A.3.30) implies

$$\frac{\widetilde{\rho}_-(s^*+s)}{2}\|\Delta\|_2^2 \leq \frac{5\lambda}{2}\|\Delta_{\mathcal{S}}\|_1 \leq \frac{5\lambda}{2}\sqrt{s^*}\|\Delta_{\mathcal{S}}\|_2 \leq \frac{5\lambda}{2}\sqrt{s^*}\|\Delta\|_2, \tag{A.3.31}$$

where the second inequality comes from the fact that $\Delta_{\mathcal{S}}$ only contains s^* entries. By simple manipulation, (A.3.31) further implies

$$\|\Delta\|_2 \leq \frac{5\lambda\sqrt{s^*}}{\widetilde{\rho}_-(s^*+s)}. \tag{A.3.32}$$

Meanwhile, (A.3.30) also implies

$$\|\Delta_{\overline{\mathcal{S}}}\|_1 \leq 5\|\Delta_{\mathcal{S}}\|_1. \tag{A.3.33}$$

Combining (A.3.32) with (A.3.33), we obtain

$$\|\Delta\|_1 \leq 5\|\Delta_{\mathcal{S}}\|_1 \leq 5\sqrt{s^*}\|\Delta_{\mathcal{S}}\|_2 \leq 5\sqrt{s^*}\|\Delta\|_2 \leq \frac{25\lambda s^*}{\widetilde{\rho}_-(s^*+s)}. \tag{A.3.34}$$

We consider the second case: $\widetilde{\rho}_-(s^*+s)\|\Delta\|_1 \leq 16\lambda s^*$. Then (A.3.29) implies

$$\|\Delta\|_2 \leq \frac{9\lambda\sqrt{s^*}}{\widetilde{\rho}_-(s^*+s)}.$$

Combining two cases, we obtain

$$\|\Delta\|_2 \leq \frac{9\lambda\sqrt{s^*}}{\widetilde{\rho}_-(s^*+s)} \quad \text{and} \quad \|\Delta\|_1 \leq \frac{25\lambda s^*}{\widetilde{\rho}_-(s^*+s)}.$$

□

A.3.8 Proof of Lemma 2.7.5

Proof. By Assumption 2.3.1, we have $\lambda \geq 4\|\nabla \widetilde{\mathcal{L}}_\lambda(\theta^*)\|_\infty$, which implies

$$\left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^*)| \geq \lambda/4,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| = 0. \tag{A.3.35}$$

We then consider an arbitrary set $\mathcal{S}'$ such that

$$\mathcal{S}' = \left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{[0]}) - \nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^*)| \geq \lambda/2,\ j \in \overline{\mathcal{S}}\right\}.$$

Let $s' = |\mathcal{S}'|$. Then there exists a $v \in \mathbb{R}^d$ such that

$$\|v\|_\infty = 1,\ \ \|v\|_0 \leq s', \quad \text{and} \quad s'\lambda/2 \leq v^\top(\nabla \widetilde{\mathcal{L}}_\lambda(\theta^{[0]}) - \nabla \widetilde{\mathcal{L}}_\lambda(\theta^*)). \tag{A.3.36}$$

By Cauchy-Schwarz inequality, (A.3.36) implies

$$\begin{aligned}\frac{s'\lambda}{2} &\leq \|v\|_2 \|\nabla \widetilde{\mathcal{L}}_\lambda(\theta^{[0]}) - \nabla \widetilde{\mathcal{L}}_\lambda(\theta^*)\|_2 \leq \sqrt{s'}\|\nabla \widetilde{\mathcal{L}}_\lambda(\theta^{[0]}) - \nabla \widetilde{\mathcal{L}}_\lambda(\theta^*)\|_2 \\ &\overset{\text{(i)}}{\leq} \rho_+(s^* + 2\widetilde{s})\sqrt{s'}\|\theta^{[0]} - \theta^*\|_2 \overset{\text{(ii)}}{\leq} \rho_+(s^* + 2\widetilde{s})\sqrt{s'}\frac{9\lambda\sqrt{s^*}}{\widetilde{\rho}_-(s^* + 2\widetilde{s})},\end{aligned} \tag{A.3.37}$$

where (i) comes from the restricted smoothness of $\widetilde{\mathcal{L}}_\lambda(\theta)$, and (ii) comes from Lemma A.3.2. By simple manipulation, (A.3.37) is rewritten as

$$\sqrt{s'} \leq \frac{18\rho_+(s^* + 2\widetilde{s})\sqrt{s^*}}{\widetilde{\rho}_-(s^* + 2\widetilde{s})}. \tag{A.3.38}$$

Since $\mathcal{S}'$ is arbitrary defined, by simple manipulation, (A.3.22) implies

$$\left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{[0]}) - \nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^*)| \geq \lambda/2,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| \leq 364\kappa^2 s^*. \tag{A.3.39}$$

Combining (A.3.35) with (A.3.39), we have

$$\begin{aligned} &\left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{[0]})| \geq 3\lambda/4,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| \\ &\leq \left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^*)| \geq \lambda/4,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| \\ &+ \left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{[0]}) - \nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^*)| \geq \lambda/2,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| \leq 364\kappa^2 s^* < \widetilde{s}, \end{aligned} \tag{A.3.40}$$

where the last inequality comes from Assumption 2.3.5. Since Assumption 2.3.7 requires $\varphi \leq 1/8$, we have $(1-\varphi)\lambda > 3\lambda/4$. Thus, (A.3.40) implies that the strong rule selects at most $\widetilde{s}$ irrelevant coordinates. □

A.3.9 Proof of Lemma 2.7.6

Proof. Before we proceed, we first introduce the following lemmas.

Lemma A.3.3. Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. For any $\lambda \geq \lambda_N$, if θ satisfies

$$\|\theta_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s} \quad \text{and} \quad \mathcal{F}_\lambda(\theta) \leq \mathcal{F}_\lambda(\theta^*) + \frac{4\lambda^2 s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})}, \tag{A.3.41}$$

then we have $\|[\mathcal{T}_{\lambda,L}(\theta)]_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$.

The proof of Lemma A.3.3 is provided in Appendix A.3.10. Since $\theta^{[m+0.5]}$ satisfies (A.3.41) for all $m = 0, 1, 2, \ldots$, by Lemma A.3.3, we have $\|w_{\overline{\mathcal{S}}}^{[m+0.5]}\|_0 \leq \widetilde{s}$ for all $m = 0, 1, 2, \ldots$.

Lemma A.3.4. Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. For every active set updating iteration, if we select a coordinate as

$$k_m = \text{argmax}_{k \in \overline{\mathcal{A}}_m} |\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})|,$$

then we have

$$k_m = \operatorname*{argmin}_k \mathcal{Q}_{\lambda,k,L}(\mathcal{T}_{\lambda,k,L}(\theta^{[m+0.5]}); \theta^{[m+0.5]}).$$

The proof of Lemma A.3.4 is provided in Appendix A.3.11. Lemma A.3.4 guarantees that our selected coordinate k_m leads to a sufficient descent in the objective value. Thus, we have

$$\begin{aligned} &\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\theta^{[m+1]}) \qquad (A.3.42)\\ &\geq \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{Q}_{\lambda,k_m,L}(\theta_{k_m}^{[m+1]}; \theta^{[m+0.5]})\\ &\geq \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \frac{1}{|\mathcal{B}_m|} \sum_{k \in \mathcal{B}_m} \mathcal{Q}_{\lambda,k,L}(w_k^{[m+0.5]}; \theta^{[m+0.5]}), \end{aligned}$$

where $\mathcal{B}_m = \left\{k \mid w_k^{[m+1]} \neq 0 \quad \text{or} \quad \theta_k^{[m+0.5]} \neq 0\right\}$ and $|\mathcal{B}_m| \leq s^* + 2\widetilde{s}$. By rearranging (A.3.42), we obtain

$$\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\theta^{[m+1]}) \geq \frac{1}{s^* + 2\widetilde{s}}\Big[\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{J}_{\lambda,L}(w^{[m+1]}; \theta^{[m+0.5]})\Big].$$

□

A.3.10 Proof of Lemma A.3.3

Proof. We define an auxiliary solution

$$\widetilde{\theta} = \theta - \frac{1}{L}\nabla\widetilde{\mathcal{L}}_\lambda(\theta) = \theta - \frac{1}{L}\nabla\widetilde{\mathcal{L}}_\lambda(\theta^*) + \frac{1}{L}(\nabla\widetilde{\mathcal{L}}_\lambda(\theta) - \nabla\widetilde{\mathcal{L}}_\lambda(\theta^*)).$$

For notational simplicity, we denote $\Delta = \theta - \theta^*$. We first consider

$$\begin{aligned}\left|\left\{j \in \overline{\mathcal{S}} \, |\theta_j| \geq L^{-1}\lambda/4\right\}\right| &\leq \left|\left\{j \in \overline{\mathcal{S}} \;\middle|\; |\Delta_j| \geq L^{-1}\lambda/4\right\}\right| \\ &\leq \frac{4L}{\lambda}\|\Delta_{\overline{\mathcal{S}}}\|_1 \leq \frac{4L}{\lambda}\|\Delta\|_1 \leq \frac{100Ls^*}{\widetilde{\rho}_-(s^* + \widetilde{s})},\end{aligned} \tag{A.3.43}$$

where the last inequality comes from Lemma A.3.2. By Assumption 2.3.1, we have $\|\nabla\widetilde{\mathcal{L}}_\lambda(\theta^*)\|_{\infty,2} \leq \lambda/4$, which implies

$$\left|\left\{j \in \overline{\mathcal{S}} \;\middle|\; |\nabla_j\widetilde{\mathcal{L}}_\lambda(\theta^*)| \geq \lambda/4\right\}\right| = 0. \tag{A.3.44}$$

Recall in Appendix A.3.6, we have shown that

$$\left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta)| \geq \frac{\lambda}{2},\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| \leq 364\kappa^2 s^*. \tag{A.3.45}$$

Combining (A.3.43) and (A.3.44) with (A.3.45), we have

$$\begin{aligned} \left|\left\{j \in \overline{\mathcal{S}} \;\middle|\; |\widetilde{\theta}_j| \geq L^{-1}\lambda\right\}\right| &\leq \left|\left\{j \in \overline{\mathcal{S}} \;\middle|\; |\theta_j| \geq L^{-1}\lambda/4\right\}\right| \\ + \left|\left\{j \in \overline{\mathcal{S}} \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^*)| \geq \lambda/4\right\}\right| &+ \left|\left\{j \;\middle|\; |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta)| \geq \lambda/2,\ j \in \overline{\mathcal{S}} \cap \mathcal{A}\right\}\right| \\ &\leq \left(364\kappa^2 + \frac{100Ls^*}{\widetilde{\rho}_-(s^* + \widetilde{s})}\right) s^* \leq \widetilde{s}, \end{aligned} \tag{A.3.46}$$

where the last inequality comes from $L \leq \rho_+(s^* + 2\widetilde{s})$ and Assumption 2.3.5. By definition of the soft thresholding operator, we have $[\mathcal{T}_{\lambda,L}(\theta)]_j = \mathcal{S}_{\lambda/L}(\widetilde{\theta}_j)$. Thus, (A.3.46) further implies $\|[\mathcal{T}_{\lambda,L}(\theta)]_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$. □

A.3.11 Proof of Lemma A.3.4

Proof. Suppose there exists a coordinate k such that

$$\theta_k^{[m+0.5]} = 0 \quad \text{and} \quad |\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})| \geq (1+\delta)\lambda. \tag{A.3.47}$$

We conduct a proximal coordinate gradient descent iteration over the coordinate k, and obtain an auxiliary solution $w_k^{[m+1]}$. Since $w_k^{[m+1]}$ is obtained by the proximal

coordinate gradient descent over the coordinate k, we have

$$w_k^{[m+1]} = \underset{w_k}{\operatorname{argmin}} \mathcal{Q}_{\lambda,k,L}(w_k; \theta^{[m+0.5]}). \tag{A.3.48}$$

We then derive an upper bound for $\mathcal{Q}_{\lambda,k,L}(w_k^{[m+1]}; \theta^{[m+0.5]})$. We consider

$$\begin{aligned}\mathcal{Q}_{\lambda,k,L}(w_k^{[m+1]}; \theta^{[m+0.5]}) &= \lambda|w_k^{[m+1]}| + \lambda\|\theta_{\setminus k}^{[m+0.5]}\|_1 + \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) \\ &+ (w_k^{[m+1]} - \theta_k^{[m+0.5]})\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) + \frac{L}{2}(w_k^{[m+1]} - \theta_k^{[m+0.5]})^2.\end{aligned} \tag{A.3.49}$$

By the convexity of the absolute value function, we have

$$|\theta_k^{[m+0.5]}| \geq |w_k^{[m+1]}| + (\theta_k^{[m+0.5]} - w_k^{[m+1]})\xi_k, \tag{A.3.50}$$

where $\xi_k \in \partial|w_k^{[m+1]}|$ satisfies the optimality condition of (A.3.48), i.e.,

$$w_k^{[m+1]} - \theta_k^{[m+0.5]} + \frac{1}{L}\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) + \frac{\lambda}{L}\xi_k = 0 \tag{A.3.51}$$

for some $\xi_k \in \partial|w_k^{[m+1]}|$. Combining (A.3.50) with (A.3.49), we have

$$\begin{aligned}&\mathcal{Q}_{\lambda,k,L}(w_k^{[m+1]}; \theta^{[m+0.5]}) - \mathcal{F}_\lambda(\theta^{[m+0.5]}) \\ &\leq (w_k^{[m+1]} - \theta_k^{[m+0.5]})(\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) + \lambda\xi_k) + \frac{L}{2}(w_k^{[m+1]} - \theta_k^{[m+0.5]})^2 \\ &\overset{\text{(i)}}{=} -\frac{L}{2}(w_k^{[m+1]} - \theta_k^{[m+0.5]})^2 \overset{\text{(ii)}}{\leq} -\frac{\delta^2\lambda^2}{2L},\end{aligned} \tag{A.3.52}$$

where (i) comes from (A.3.51) and (ii) comes from Lemma 2.7.8 and (A.3.47).

Assume that there exists another coordinate j with $\theta_j^{[m+0.5]} = 0$ such that

$$|\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})| > |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})|. \tag{A.3.53}$$

Similarly, we conduct a proximal coordinate gradient descent iteration over the coordinate j, and obtain an auxiliary solution $w_j^{[m+1]}$. By definition of the soft thresholding function, we rewrite $w_k^{[m+1]}$ and $w_j^{[m+1]}$ as

$$w_k^{[m+1]} = -\frac{z_k}{L}\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) \quad \text{and} \quad w_j^{[m+1]} = -\frac{z_j}{L}\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}),$$

where z_k and z_j are defined as

$$z_k = 1 - \frac{\lambda}{|\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})|} \quad \text{and} \quad z_j = 1 - \frac{\lambda}{|\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})|}.$$

By (A.3.53), we know $z_k \geq z_j$. Moreover, we define

$$z = \frac{|\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})|}{|\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})|} \cdot z_j \quad \text{and} \quad \widetilde{w}_k^{[m+1]} = -\frac{z}{L}\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}). \tag{A.3.54}$$

Note that we have $|\widetilde{w}_k^{[m+1]}| = |w_j^{[m+1]}|$. We then consider

$$
\begin{aligned}
&\mathcal{Q}_{\lambda,k,L}(\widetilde{w}_k^{[m+1]};\theta^{[m+0.5]}) - \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) \\
&= -\frac{z}{L}|\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})|^2 + \frac{L}{2}|\widetilde{w}_k^{[m+1]}|^2 + \lambda|\widetilde{w}_k^{[m+1]}| + \lambda\|\theta_{\backslash k}^{[m+0.5]}\|_1 \\
&\overset{\text{(i)}}{=} -\frac{z_j}{L}|\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})| \cdot |\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})| + \frac{L}{2}|\widetilde{w}_k^{[m+1]}|^2 + \lambda|\widetilde{w}_k^{[m+1]}| + \lambda\|\theta_{\backslash k}^{[m+0.5]}\|_1 \\
&\overset{\text{(ii)}}{<} -\frac{z_j}{L}|\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})|^2 + \frac{L}{2}|w_j^{[m+1]}|^2 + \lambda|w_j^{[m+1]}| + \lambda\|\theta_{\backslash k}^{[m+0.5]}\|_1 \\
&= \mathcal{Q}_{\lambda,k,L}(w_j^{[m+1]};\theta^{[m+0.5]}) - \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}),
\end{aligned}
$$

where (i) comes from (A.3.54) and (ii) comes from (A.3.47). We then have

$$
\begin{aligned}
\mathcal{Q}_{\lambda,k,L}(w_k^{[m+1]};\theta^{[m+0.5]}) &\le \mathcal{Q}_{\lambda,k,L}(\widetilde{w}_k^{[m+1]};\theta^{[m+0.5]}) \\
&\le \mathcal{Q}_{\lambda,j,L}(w_j^{[m+1]};\theta^{[m+0.5]}),
\end{aligned} \tag{A.3.55}
$$

where the last inequality comes from (A.3.48). Thus, (A.3.55) guarantees

$$
\mathcal{Q}_{\lambda,k_m,L}(w_{k_m}^{[m+0.5]};\theta^{[m+0.5]}) = \min_{j\in\overline{\mathcal{A}}_m} \mathcal{Q}_{\lambda,j,L}(w_j^{[m+1]};\theta^{[m+0.5]}), \tag{A.3.56}
$$

where $k_m = \operatorname{argmax}_{k\in\overline{\mathcal{A}}_m} |\nabla\widetilde{\mathcal{L}}_k(\theta)^{[m+0.5]}|$.

For any $j \in \mathcal{A}_m$, we construct two auxiliary solutions $w^{[m+1]}$ and $v^{[m+1]}$,

$$
w_j^{[m+1]} = \operatorname*{argmin}_{v_j} \mathcal{Q}_{\lambda,j,L}(v_j;\theta^{[m+0.5]}) \ \text{ and } \ v_j^{[m+1]} = \operatorname*{argmin}_{v_j} \mathcal{F}_\lambda(v_j,\theta_{\backslash j}^{[m+0.5]}).
$$

Recall $\theta^{[m+0.5]}$ is the output solution of the previous inner loop, i.e, $\theta^{[m+0.5]} = \theta^{(t+1)}$. By the restricted convexity of $\mathcal{F}_\lambda(\theta)$, we have

$$\begin{aligned}
&\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(v_j^{[m+1]}, \theta_{\setminus j}^{(t+1)}) \\
&\qquad \leq \frac{(\nabla_j \widetilde{\mathcal{L}}_\lambda(\theta^{(t+1)}) + \lambda \xi_j)^2}{2\widetilde{\rho}_-(1)} \leq \frac{\|\nabla_{\mathcal{A}} \widetilde{\mathcal{L}}_\lambda(\theta^{(t+1)}) + \lambda \xi_{\mathcal{A}}\|_2^2}{2\widetilde{\rho}_-(1)},
\end{aligned}$$

for some $\xi_{\mathcal{A}} \in \partial \|\theta_{\mathcal{A}}^{(t+1)}\|_1$. Since the inner loop terminates when $\|\theta^{(t+1)} - \theta^{(t)}\|_2^2 \leq \tau^2 \lambda^2$, we have

$$\begin{aligned}
&\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(v_j^{[m+1]}, \theta_{\setminus j}^{(t+1)}) \\
&\qquad \leq \frac{(s^* + 2\widetilde{s})\rho_+^2(s^* + 2\widetilde{s})\|\theta^{(t+1)} - \theta^{(t)}\|_2^2}{2\widetilde{\rho}_-(1)} \leq \frac{\delta^2 \lambda^2}{2L},
\end{aligned} \tag{A.3.57}$$

where the last equality comes from Assumption 2.3.7. Thus, (A.3.57) implies

$$\begin{aligned}
&\mathcal{Q}_{\lambda,j,L}(w_j^{[m+1]}; \theta^{[m+0.5]}) - \mathcal{F}_\lambda(\theta^{[m+0.5]}) \\
&\qquad \geq \mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(v_j^{[m+1]}, \theta_{\setminus j}^{(t+1)}) \geq -\frac{\delta^2 \lambda^2}{2L}.
\end{aligned} \tag{A.3.58}$$

Since j is arbitrarily selected from $\mathcal{A}_m$, by (A.3.52) and (A.3.58), we have

$$\mathcal{Q}_{\lambda,k_m,L}(w_{k_m}^{[m+0.5]}; \theta^{[m+0.5]}) \leq \min_{j \in \mathcal{A}_m} \mathcal{Q}_{\lambda,j,L}(w_j^{[m+1]}; \theta^{[m+0.5]}). \tag{A.3.59}$$

Combining (A.3.56) with (A.3.59), we have

$$\mathcal{Q}_{\lambda,k_m,L}(w_{k_m}^{[m+0.5]};\theta^{[m+0.5]}) = \min_j \mathcal{Q}_{\lambda,j,L}(w_j^{[m+1]};\theta^{[m+0.5]}).$$

□

A.3.12 Proof of Lemma 2.7.7

Proof. Define $\mathcal{D}_m = \{w \mid w \in \mathbb{R}^d,\ w_{\overline{\mathcal{B}}_m} = 0\}$, we have

$$\begin{aligned}\mathcal{J}_{\lambda,L}(w^{[m+1]};\theta^{[m+0.5]}) &= \min_{w\in\mathcal{D}_m} \mathcal{J}_{\lambda,L}(w;\theta^{[m+0.5]}) = \min_{w\in\mathcal{D}_m} \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) \\ &\quad + (w-\theta^{[m+0.5]})^\top \nabla\widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) + \lambda\|w\|_1 + \frac{L}{2}\|w-\theta^{[m+0.5]}\|_2^2 \\ &\leq \min_{w\in\mathcal{D}_m} \mathcal{F}_\lambda(w) + \frac{(L-\rho_-(s^*+2\widetilde{s}))}{2}\|w-\theta^{[m+0.5]}\|_2^2,\end{aligned}$$

where the last inequality coms from the restricted convexity of $\widetilde{\mathcal{L}}_\lambda(\theta)$, i.e.,

$$\widetilde{\mathcal{L}}_\lambda(w) \leq \widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) + (w-\theta^{[m+0.5]})^\top \nabla\widetilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) + \frac{\rho_-(s^*+2\widetilde{s})}{2}\|w-\theta^{[m+0.5]}\|_2^2.$$

Let $w = z\overline{\theta}^{\lambda} + (1-z)\theta^{[m+0.5]})$ for $z \in [0,1]$. Then we have

$$\begin{aligned}
&\mathcal{J}_{\lambda,L}(w^{[m+1]}; \theta^{[m+0.5]}) \qquad (A.3.60)\\
&\leq \min_{z\in[0,1]} \mathcal{F}_{\lambda}(z\overline{\theta}^{\lambda} + (1-z)\theta^{[m+0.5]}) + \frac{z^2(L - \rho_{-}(s^* + 2\widetilde{s}))}{2}\|\overline{\theta}^{\lambda} - \theta^{[m+0.5]}\|_2^2\\
&\leq \mathcal{F}_{\lambda}(\theta^{[m+0.5]}) + \min_{z\in[0,1]} z[\mathcal{F}_{\lambda}(\overline{\theta}^{\lambda}) - \mathcal{F}_{\lambda}(\theta^{[m+0.5]})]\\
&\qquad\qquad + \frac{(z^2 L - z\rho_{-}(s^* + 2\widetilde{s}))}{2}\|\overline{\theta}^{\lambda} - \theta^{[m+0.5]}\|_2^2,
\end{aligned}$$

where the last inequality comes from the restricted convexity of $\mathcal{F}_{\lambda}(\theta)$, i.e.,

$$\begin{aligned}
\mathcal{F}_{\lambda}(z\overline{\theta}^{\lambda} + (1-z)\theta^{[m+0.5]}) + \frac{z(1-z)\rho_{-}(s^* + 2\widetilde{s})}{2}\|\overline{\theta}^{\lambda} - \theta^{[m+0.5]}\|_2^2\\
\leq z\mathcal{F}_{\lambda}(\overline{\theta}^{\lambda}) + (1-z)\mathcal{F}_{\lambda}(\theta^{[m+0.5]}).
\end{aligned}$$

By the restricted convexity of $\mathcal{F}_{\lambda}(\theta)$, we have

$$\|\overline{\theta}^{\lambda} - \theta^{[m+0.5]}\|_2^2 \leq \frac{2[\mathcal{F}_{\lambda}(\theta^{[m+0.5]}) - \mathcal{F}_{\lambda}(\overline{\theta}^{\lambda})]}{\rho_{-}(s^* + 2\widetilde{s})}. \qquad (A.3.61)$$

Combining (A.3.61) with (A.3.60), we obtain

$$\begin{aligned}
&\mathcal{J}_{\lambda,L}(w^{[m+1]}; \theta^{(t)}) - \mathcal{F}_{\lambda}(\theta^{[m+0.5]}) \qquad (A.3.62)\\
&\qquad \leq \min_{z\in[0,1]} \left(\frac{z^2 L}{\rho_{-}(s^* + 2\widetilde{s})} - 2z\right)[\mathcal{F}_{\lambda}(\theta^{[m+0.5]}) - \mathcal{F}_{\lambda}(\overline{\theta}^{\lambda})].
\end{aligned}$$

By setting $z = \widetilde{\rho}_-(s^* + 2\widetilde{s})/L$, we minimize the R.H.S of (A.3.62) and obtain

$$\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{J}_{\lambda,L}(w^{[m+1]}; \theta^{(t)}) \geq \frac{\widetilde{\rho}_-(s^* + 2\widetilde{s})}{L}[\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\overline{\theta}^\lambda)].$$

□

A.3.13 Proof of Lemma A.3.1

Proof. We prove the uniqueness of $\overline{\theta}^\lambda$ by contradiction. Assume that there exist two different local optima $\overline{\theta}^\lambda$ and $\widetilde{\theta}^\lambda$. Let $\bar{\xi} \in \partial\|\overline{\theta}^\lambda\|_1$ and $\widetilde{\xi} \in \partial\|\widetilde{\theta}^\lambda\|_1$ be two sub-gradient vectors satisfying

$$\nabla\widetilde{\mathcal{L}}_\lambda(\overline{\theta}^\lambda) + \lambda\bar{\xi} = 0 \quad \text{and} \quad \nabla\widetilde{\mathcal{L}}_\lambda(\widetilde{\theta}^\lambda) + \lambda\widetilde{\xi} = 0. \tag{A.3.63}$$

By the restricted strong convexity of $\mathcal{F}_\lambda(\theta)$, we obtain

$$\mathcal{F}_\lambda(\overline{\theta}^\lambda) \geq \mathcal{F}_\lambda(\widetilde{\theta}^\lambda) + (\overline{\theta}^\lambda - \widetilde{\theta}^\lambda)^\top(\nabla\widetilde{\mathcal{L}}_\lambda(\widetilde{\theta}^\lambda) + \lambda\widetilde{\xi}) + \frac{\widetilde{\rho}_-(s^* + 2\widetilde{s})}{2}\|\overline{\theta}^\lambda - \widetilde{\theta}^\lambda\|_2^2,$$
$$\mathcal{F}_\lambda(\widetilde{\theta}^\lambda) \geq \mathcal{F}_\lambda(\overline{\theta}^\lambda) + (\widetilde{\theta}^\lambda - \overline{\theta}^\lambda)^\top(\nabla\bar{\mathcal{L}}_\lambda(\widetilde{\theta}^\lambda) + \lambda\bar{\xi}) + \frac{\widetilde{\rho}_-(s^* + 2\widetilde{s})}{2}\|\overline{\theta}^\lambda - \widetilde{\theta}^\lambda\|_2^2,$$

since $\|\overline{\theta}^\lambda_{\widetilde{S}}\|_0 \leq \widetilde{s}$ and $\|\widetilde{\theta}^\lambda_{\widetilde{S}}\|_0 \leq \widetilde{s}$. Combining the above two inequalities with (A.3.63), we have $\|\overline{\theta}^\lambda - \widetilde{\theta}^\lambda\|_2^2 = 0$ implying $\overline{\theta}^\lambda = \widetilde{\theta}^\lambda$. That is contradicted by our assumption. Thus, the local optimum $\overline{\theta}^\lambda$ is unique.

□

A.3.14 Proof of Lemma 2.3.11

Proof. For notational simplicity, we define $\Delta = \theta - \theta^*$. Let $\widetilde{\xi} \in \partial\|\theta\|_1$ be a subgradient vector satisfying

$$\mathcal{K}_{\lambda_{K-1}}(\theta) = \|\nabla\widetilde{\mathcal{L}}_{\lambda_{K-1}}(\theta) + \lambda_{K-1}\widetilde{\xi}\|_\infty.$$

We then consider the following decomposition

$$\begin{aligned}
&\mathcal{K}_{\lambda_K}(\theta) \le \|\nabla\widetilde{\mathcal{L}}_{\lambda_K}(\theta) + \lambda_K\widetilde{\xi}\|_\infty \qquad \text{(A.3.64)}\\
&\le \|\nabla\widetilde{\mathcal{L}}_{\lambda_{K-1}}(\theta) + \lambda_{K-1}\widetilde{\xi}\|_\infty + \|\lambda_K\widetilde{\xi} - \lambda_{K-1}\widetilde{\xi}\|_\infty\\
&+ \|\nabla\mathcal{H}_{\lambda_K}(\theta) - \nabla\mathcal{H}_{\lambda_{K-1}}(\theta)\|_\infty \overset{\text{(i)}}{\le} \delta_{K-1}\lambda_{K-1} + 3(1-\eta)\lambda_{K-1} \overset{\text{(ii)}}{\le} \frac{\lambda_K}{4},
\end{aligned}$$

where (i) comes from (R.4) in Lemma A.2.1, and (ii) comes from $\delta_{K-1} \le 1/8$ and $1 - \eta \le 1/24$ in Assumption 2.3.1.

We then proceed to characterize the statistical error of θ in terms of λ_K. For notational simplicity, we omit the index K and denote λ_K by λ. Since (A.3.64) implies that θ satisfies the approximate KKT condition for λ, then by the restricted convexity of $\widetilde{\mathcal{L}}_\lambda(\theta)$, we have

$$\begin{aligned}
\mathcal{F}_\lambda(\theta^*) - \frac{\widetilde{\rho}_-(s^*+\widetilde{s})}{2}\|\Delta\|_2^2 &\ge \mathcal{F}_\lambda(\theta) - \Delta^\top(\nabla\widetilde{\mathcal{L}}_\lambda(\theta) + \lambda\widetilde{\xi}) \qquad \text{(A.3.65)}\\
&\overset{\text{(i)}}{\ge} \mathcal{F}_\lambda(\theta) - \|\nabla\widetilde{\mathcal{L}}_\lambda(\theta) + \lambda\widetilde{\xi}\|_\infty \cdot \|\Delta\|_1 \overset{\text{(ii)}}{\ge} \mathcal{F}_\lambda(\theta) - \frac{\lambda}{4}\|\Delta\|_1,
\end{aligned}$$

where (i) comes from Hölder's inequality and (ii) comes from (A.3.64). We then rewrite (A.3.65) as

$$\lambda\|\theta^*\|_1 - \lambda\|\theta\|_1 + \frac{\lambda}{4}\|\Delta\|_1 \geq \widetilde{\mathcal{L}}_\lambda(\theta) - \widetilde{\mathcal{L}}_\lambda(\theta^*) + \frac{\widetilde{\rho}_-(s^* + \widetilde{s})}{2}\|\Delta\|_2^2. \tag{A.3.66}$$

By the restricted convexity of $\widetilde{\mathcal{L}}_\lambda(\theta)$ again, we have

$$\begin{aligned}
&\widetilde{\mathcal{L}}_\lambda(\theta) - \widetilde{\mathcal{L}}_\lambda(\theta^*) - \frac{\widetilde{\rho}_-(s^* + \widetilde{s})}{2}\|\Delta\|_2^2 \geq \Delta^\top \nabla \widetilde{\mathcal{L}}_\lambda(\theta^*) \qquad (A.3.67)\\
&\overset{\text{(i)}}{=} \Delta_{\mathcal{S}}^\top \nabla_{\mathcal{S}} \mathcal{L}(\theta^*) + \Delta_{\overline{\mathcal{S}}}^\top \nabla_{\overline{\mathcal{S}}} \mathcal{L}(\theta^*) + \Delta_{\mathcal{S}}^\top \nabla_{\mathcal{S}} \mathcal{H}_\lambda(\theta^*)\\
&\overset{\text{(ii)}}{\geq} -\|\Delta_{\mathcal{S}}\|_1 \|\nabla \mathcal{L}(\theta^*)\|_\infty - \|\Delta_{\overline{\mathcal{S}}}\|_1 \|\nabla \mathcal{L}(\theta^*)\|_\infty - \|\Delta_{\mathcal{S}}\|_1 \|\nabla_{\mathcal{S}} \mathcal{H}_\lambda(\theta^*)\|_\infty,
\end{aligned}$$

where (i) comes from $\nabla_{\overline{\mathcal{S}}} \mathcal{H}_\lambda(\theta^*) = 0$ by (R.3) of Lemma A.2.1, and (ii) comes from Hölder's inequality. Assumption 2.3.1 and (R.2) of Lemma A.2.1 imply

$$\|\nabla \mathcal{L}(\theta^*)\|_\infty \leq \lambda/4 \quad \text{and} \quad \|\nabla_{\mathcal{S}} \mathcal{H}_\lambda(\theta^*)\|_\infty \leq \lambda. \tag{A.3.68}$$

Combining (A.3.67) with (A.3.68), we obtain

$$\widetilde{\mathcal{L}}_\lambda(\theta) - \widetilde{\mathcal{L}}_\lambda(\theta^*) \geq -\frac{3}{2}\lambda\|\Delta_{\mathcal{S}}\|_1 - \frac{\lambda}{2}\|\Delta_{\overline{\mathcal{S}}}\|_1 + \widetilde{\rho}_-(s^* + \widetilde{s})\|\Delta\|_2^2. \tag{A.3.69}$$

Plugging (A.3.69) and

$$\|\theta^*\|_1 - \|\theta\|_1 = \|\theta^*_{\mathcal{S}}\|_1 - (\|\theta_{\mathcal{S}}\|_1 + \|\Delta_{\overline{\mathcal{S}}}\|_1) \leq \|\Delta_{\mathcal{S}}\|_1 - \|\Delta_{\overline{\mathcal{S}}}\|_1$$

into (A.3.66), we obtain

$$\frac{11\lambda}{4}\|\Delta_{\mathcal{S}}\|_1 \geq \frac{\lambda}{4}\|\Delta_{\overline{\mathcal{S}}}\|_1 + \widetilde{\rho}_-(s^* + \widetilde{s})\|\Delta\|_2^2. \tag{A.3.70}$$

By simple manipulation, (A.3.70) implies

$$\widetilde{\rho}_-(s^* + \widetilde{s})\|\Delta\|_2^2 \leq \frac{11\lambda}{4}\|\Delta_{\mathcal{S}}\|_1 \leq \frac{11\lambda}{4}\sqrt{s^*}\|\Delta_{\mathcal{S}}\|_2 \leq \frac{11\lambda}{4}\sqrt{s^*}\|\Delta\|_2, \tag{A.3.71}$$

where the second inequality comes from the fact that $\Delta_{\mathcal{S}}$ only contains s^* rows. By simple manipulation again, (A.3.71) implies

$$\|\Delta\|_2 \leq \frac{11\lambda\sqrt{s^*}}{4\widetilde{\rho}_-(s^* + \widetilde{s})}. \tag{A.3.72}$$

Meanwhile, (A.3.70) also implies

$$\|\Delta_{\overline{\mathcal{S}}}\|_1 \leq 11\|\Delta_{\mathcal{S}}\|_1. \tag{A.3.73}$$

Combining (A.3.72) with (A.3.73), we obtain

$$\|\Delta\|_1 \leq 11\|\Delta_{\mathcal{S}}\|_1 \leq 11\sqrt{s^*}\|\Delta_{\mathcal{S}}\|_2 \leq 11\sqrt{s^*}\|\Delta\|_2 \leq \frac{31\lambda s^*}{\widetilde{\rho}_-(s^*+\widetilde{s})}. \tag{A.3.74}$$

Plugging (A.3.74) and (A.3.72) into (A.3.65), we have

$$\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(\theta^*) \leq \delta\lambda\|\Delta\|_1 \leq \frac{4\lambda^2 s^*}{\widetilde{\rho}_-(s^*+\widetilde{s})}.$$

□

A.4 Lemmas for General Loss Functions

A.5 Proof of Lemma 2.4.3

Proof. For notational simplicity, we denote θ^{relax} by θ and write $\widetilde{\mathcal{F}}_\lambda(\theta) = \mathcal{L}(\theta) + \lambda\|\theta\|_1$. Let $\widetilde{\xi} \in \partial\|\theta\|_1$ be a subgradient vector satisfying

$$\|\nabla\mathcal{L}(\theta) + \lambda\widetilde{\xi}\|_\infty = \min_{\xi\in\partial\|\theta\|_1} \|\nabla\mathcal{L}(\theta) + \lambda\xi\|_\infty.$$

For notational simplicity, we define $\Delta = \theta^* - \theta$. Since $\widetilde{\mathcal{F}}_\lambda(\theta)$ is a convex function, we have

$$\begin{aligned}\widetilde{\mathcal{F}}_\lambda(\theta^*) &\geq \widetilde{\mathcal{F}}_\lambda(\theta) - \Delta^\top(\nabla\mathcal{L}(\theta) + \lambda\widetilde{\xi}) \\ &\geq \widetilde{\mathcal{F}}_\lambda(\theta) - \|\Delta\|_1\|\nabla\mathcal{L}(\theta) + \lambda\widetilde{\xi}\|_\infty \geq \widetilde{\mathcal{F}}_\lambda(\theta) - \frac{\lambda}{8}\|\Delta\|_1,\end{aligned} \tag{A.5.1}$$

where the second inequality comes from Hölder's inequality, and the last inequality comes from (2.4.6).

To establish the statistical properties of θ, we need to verify that θ satisfies $\|\theta - \theta^*\|_2 \leq R$ such that the restricted strong convexity holds for θ. We prove it by contradiction. We first assume $\|\theta - \theta^*\|_2 \geq R$. Then there exists some $z \in (0,1)$ such that

$$\widetilde{\theta} = (1-z)\theta + z\theta^* \quad \text{and} \quad \|\widetilde{\theta} - \theta^*\|_2 = R. \tag{A.5.2}$$

Then by the convexity of $\widetilde{\mathcal{F}}_\lambda(\theta)$ again, (A.5.1) and (A.5.2) imply

$$\begin{aligned}\widetilde{\mathcal{F}}_\lambda(\widetilde{\theta}) &\leq (1-z)\widetilde{\mathcal{F}}_\lambda(\theta) + z\widetilde{\mathcal{F}}_\lambda(\theta^*) \\ &\leq (1-z)\widetilde{\mathcal{F}}_\lambda(\theta^*) + \frac{(1-z)\lambda}{8}\|\Delta\|_1 + z\widetilde{\mathcal{F}}_\lambda(\theta^*) \leq \widetilde{\mathcal{F}}_\lambda(\theta^*) + \frac{\lambda}{8}\|\widetilde{\Delta}\|_1,\end{aligned} \tag{A.5.3}$$

where the last inequality comes from the fact

$$\|\widetilde{\Delta}\|_1 = \|\widetilde{\theta} - \theta^*\|_1 = \|(1-z)\theta + z\theta^* - \theta^*\|_1 = (1-z)\|\Delta\|_1.$$

By simple manipulation, we can rewrite (A.5.3) as

$$\mathcal{L}(\widetilde{\theta}) - \mathcal{L}(\theta^*) \leq \lambda\|\theta^*\|_1 - \lambda\|\widetilde{\theta}\|_1 + \frac{\lambda}{8}\|\widetilde{\Delta}\|_1. \tag{A.5.4}$$

By the convexity of $\mathcal{L}(\theta)$, we have

$$\begin{aligned} \mathcal{L}(\widetilde{\theta}) - \mathcal{L}(\theta^*) &\geq \widetilde{\Delta}^\top \nabla\mathcal{L}(\theta^*) \\ &\geq -\|\widetilde{\Delta}\|_1\|\nabla\mathcal{L}(\theta^*)\|_\infty \geq -\frac{\lambda}{8}\|\Delta_{\mathcal{S}}\|_1 - \frac{\lambda}{8}\|\Delta_{\overline{\mathcal{S}}}\|_1, \end{aligned} \tag{A.5.5}$$

where the last inequality comes from our assumption $\lambda \geq 8\|\nabla\mathcal{L}(\theta^*)\|_\infty$. By the decomposability of the ℓ_1 norm, we have

$$\begin{aligned} &\|\theta^*\|_1 - \|\theta\|_1 + \frac{1}{8}\|\widetilde{\Delta}\|_1 \\ &= \|\theta^*_{\mathcal{S}}\|_1 - (\|\theta_{\mathcal{S}}\|_1 + \|\Delta_{\overline{\mathcal{S}}}\|_1) + \frac{1}{8}\|\widetilde{\Delta}_{\mathcal{S}}\|_1 + \frac{1}{8}\|\widetilde{\Delta}_{\overline{\mathcal{S}}}\|_1 \end{aligned} \tag{A.5.6}$$

$$\leq \frac{9}{8}\|\Delta_{\mathcal{S}}\|_1 - (1-\delta)\|\widetilde{\Delta}_{\overline{\mathcal{S}}}\|_1 \leq \frac{9}{8}\|\widetilde{\Delta}_{\mathcal{S}}\|_1 - \frac{7}{8}\|\widetilde{\Delta}_{\overline{\mathcal{S}}}\|_1.$$

Combining (A.5.4) with (A.5.5) and (A.5.6), we obtain

$$\|\widetilde{\Delta}_{\overline{\mathcal{S}}}\|_1 \leq \frac{5}{3}\|\widetilde{\Delta}_{\mathcal{S}}\|_1. \tag{A.5.7}$$

To establish the statistical properties of $\widetilde{\theta}$, we define the following sets:

$$\begin{aligned}
\mathcal{S}_0 &= \left\{ j \;\middle|\; j \in \overline{\mathcal{S}},\ \textstyle\sum_{k\in\overline{\mathcal{S}}} \mathbb{1}_{\{|\widetilde{\theta}_k|\geq|\widetilde{\theta}_j|\}} \leq \widetilde{s} \right\},\\
\mathcal{S}_1 &= \left\{ j \;\middle|\; j \in \overline{\mathcal{S}} \setminus \mathcal{S}_0,\ \textstyle\sum_{k\in\overline{\mathcal{S}}\setminus\mathcal{S}_0} \mathbb{1}_{\{|\widetilde{\theta}_k|\geq|\widetilde{\theta}_j|\}} \leq \widetilde{s} \right\},\\
\mathcal{S}_2 &= \left\{ j \;\middle|\; j \in \overline{\mathcal{S}} \setminus (\mathcal{S}_0 \cup \mathcal{S}_1),\ \textstyle\sum_{k\in\overline{\mathcal{S}}\setminus(\mathcal{S}_0\cup\mathcal{S}_1)} \mathbb{1}_{\{|\widetilde{\theta}_k|\geq|\widetilde{\theta}_j|\}} \leq \widetilde{s} \right\},\\
\mathcal{S}_3 &= \left\{ j \;\middle|\; j \in \overline{\mathcal{S}} \setminus (\mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2),\ \textstyle\sum_{k\in\overline{\mathcal{S}}\setminus(\mathcal{S}_0\cup\mathcal{S}_1\cup\mathcal{S}_2)} \mathbb{1}_{\{|\widetilde{\theta}_k|\geq|\widetilde{\theta}_j|\}} \leq \widetilde{s} \right\}, \ldots.
\end{aligned}$$

Before we proceed with the proof, we introduce the following lemma.

Lemma A.5.1 (Lemma 6.9 in [128])**.** Let $b_1 \geq b_2 \geq \ldots \geq 0$. For $s \in \{1, 2, \ldots\}$, we have

$$\sqrt{\textstyle\sum_{j\geq i+1} b_j^2} \leq \textstyle\sum_{k=1}^{\infty} \sqrt{\sum_{j=ks+1}^{(k+1)s} b_j^2} \leq \sqrt{s}\sum_{k=1}^{\infty} b_j.$$

The proof of Lemma A.5.1 is provided in [128], and therefore is omitted. By Lemma A.5.1 and (A.5.7), we have

$$\sum_{j\geq 1}\|\widetilde{\Delta}_{\mathcal{S}_j}\|_1 \leq \frac{1}{\sqrt{s}}\|\widetilde{\Delta}_{\overline{\mathcal{S}}}\|_1 \leq \frac{5}{3}\sqrt{\frac{s^*}{s}}\|\widetilde{\Delta}_{\mathcal{S}}\|_2 \leq \frac{5}{3}\sqrt{\frac{s^*}{s}}\|\widetilde{\Delta}_{\mathcal{A}}\|_2,$$

where $\mathcal{A} = \mathcal{S} \cup \mathcal{S}_0$. By definition of the largest sparse eigenvalue and Assumption

2.3.5, given $\ddot{\theta} = z\widetilde{\theta} + (1-z)\theta^*$ for any $z \in [0,1]$ and $j \geq 1$, we have

$$\left|\widetilde{\Delta}_{\mathcal{S}_j}^\top \nabla^2_{\mathcal{S}_j\mathcal{A}} \mathcal{L}(\ddot{\theta})\widetilde{\Delta}_{\mathcal{A}}\right| \leq \rho_+(s^* + \widetilde{s})\|\widetilde{\Delta}_{\mathcal{S}_j}\|_2 \|\widetilde{\Delta}_{\mathcal{A}}\|_2,$$

which further implies

$$\begin{aligned} |\widetilde{\Delta}_{\overline{\mathcal{A}}}^\top \nabla^2_{\overline{\mathcal{A}}\mathcal{A}} \mathcal{L}(\ddot{\theta})\widetilde{\Delta}_{\mathcal{A}}| &\leq \textstyle\sum_{j\geq 1} |\widetilde{\Delta}_{\mathcal{S}_j}^\top \nabla^2_{\mathcal{S}_j\mathcal{A}} \mathcal{L}(\ddot{\theta})\widetilde{\Delta}_{\mathcal{A}}| \\ &= \frac{5\rho_+(s^* + 2\widetilde{s})}{3}\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2 \sqrt{\frac{s^*}{\widetilde{s}}}. \end{aligned} \tag{A.5.8}$$

By definition of the smallest sparse eigenvalue and Assumption 2.3.5, we have

$$\frac{\widetilde{\Delta}_{\mathcal{A}}^\top \nabla^2_{\mathcal{A}\mathcal{A}} \mathcal{L}(\ddot{\theta})\widetilde{\Delta}_{\mathcal{A}}}{\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2} \geq \rho_-(s^* + \widetilde{s}). \tag{A.5.9}$$

Combining (A.5.8) with (A.5.9), we have

$$|\widetilde{\Delta}_{\overline{\mathcal{A}}}^\top \nabla^2_{\overline{\mathcal{A}}\mathcal{A}} \mathcal{L}(\ddot{\theta})\widetilde{\Delta}_{\mathcal{A}}| \leq \frac{5\rho_+(s^* + \widetilde{s})}{3\rho_-(s^* + \widetilde{s})} \sqrt{\frac{s^*}{\widetilde{s}}} \widetilde{\Delta}_{\mathcal{A}}^\top \nabla^2_{\mathcal{A}\mathcal{A}} \mathcal{L}(\ddot{\theta})\widetilde{\Delta}_{\mathcal{A}},$$

which further implies

$$\frac{|\widetilde{\Delta}_{\overline{\mathcal{A}}}^\top \nabla^2_{\overline{\mathcal{A}}\mathcal{A}} \mathcal{L}(\ddot{\theta})\widetilde{\Delta}_{\mathcal{A}}|}{|\widetilde{\Delta}_{\mathcal{A}}^\top \nabla^2_{\mathcal{A}\mathcal{A}} \mathcal{L}(\ddot{\theta})\widetilde{\Delta}_{\mathcal{A}}|} \leq \frac{5\rho_+(s^* + \widetilde{s})}{3\rho_-(s^* + \widetilde{s})} \sqrt{\frac{s^*}{\widetilde{s}}}.$$

Eventually, we have

$$\begin{aligned}\frac{\Delta^\top \nabla^2 \mathcal{L}(\ddot{\theta})\Delta}{\|\Delta_{\mathcal{A}}\|_2^2} &\geq \left(1 - \frac{|\widetilde{\Delta}_{\overline{\mathcal{A}}}^\top \nabla^2_{\overline{\mathcal{A}}\mathcal{A}} \mathcal{L}(\ddot{\theta})\widetilde{\Delta}_{\mathcal{A}}|}{|\widetilde{\Delta}_{\mathcal{A}}^\top \nabla^2_{\mathcal{A}\mathcal{A}} \mathcal{L}(\ddot{\theta})\widetilde{\Delta}_{\mathcal{A}}|}\right)\rho_-(s^* + \widetilde{s}) \\ &\geq \left(1 - \frac{9\rho_+(s^* + \widetilde{s})}{7\rho_-(s^* + \widetilde{s})}\sqrt{\frac{s^*}{\widetilde{s}}}\right)\rho_-(s^* + \widetilde{s}) \geq \frac{7\rho_-(s^* + \widetilde{s})}{8}, \end{aligned} \tag{A.5.10}$$

where the last inequality comes from Assumption 2.3.5. Then by the mean value theorem, we choose some z such that

$$\mathcal{L}(\widetilde{\theta}) - \mathcal{L}(\theta^*) - \widetilde{\Delta}^\top \nabla \mathcal{L}(\theta^*) = \frac{1}{2}\widetilde{\Delta}^\top \nabla^2 \mathcal{L}(\ddot{\theta})\widetilde{\Delta} \geq \frac{7\rho_-(s^* + \widetilde{s})}{16}\|\widetilde{\Delta}_{\mathcal{S}}\|_2^2,$$

which implies

$$\begin{aligned}\mathcal{L}(\widetilde{\theta}) - \mathcal{L}(\theta^*) &\geq \widetilde{\Delta}^\top \nabla \mathcal{L}(\theta^*) + \frac{7\rho_-(s^* + \widetilde{s})}{16}\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2 \\ &\geq \frac{7\rho_-(s^* + \widetilde{s})}{16}\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2 - \frac{\lambda}{8}\|\widetilde{\Delta}_{\mathcal{S}}\|_1 - \frac{\lambda}{8}\|\widetilde{\Delta}_{\overline{\mathcal{S}}}\|_1.\end{aligned}$$

Then by (A.5.4) and (A.5.6), we have

$$\begin{aligned}\rho_-(s^* + \widetilde{s})\|\widetilde{\Delta}_{\mathcal{S}}\|_2^2 \leq \rho_-(s^* + \widetilde{s})\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2 &\leq \frac{20}{7}\lambda\|\Delta_{\mathcal{S}}\|_1 \\ &\leq \frac{20}{7}\sqrt{s^*}\lambda\|\Delta_{\mathcal{S}}\|_2 \leq \frac{20}{7}\sqrt{s^*}\lambda\|\Delta_{\mathcal{A}}\|_2,\end{aligned}$$

which further implies

$$\|\Delta_{\mathcal{S}}\|_2 \le \|\Delta_{\mathcal{A}}\|_2 \le \frac{20\sqrt{s^*}\lambda}{7\rho_-(s^*+\widetilde{s})} \quad \text{and} \quad \|\Delta_{\mathcal{S}}\|_1 \le \frac{20s^*\lambda}{7\rho_-(s^*+\widetilde{s})}. \tag{A.5.11}$$

By Lemma A.5.1, (A.5.11) implies

$$\|\widetilde{\Delta}_{\overline{\mathcal{A}}}\|_2 \le \frac{\|\widetilde{\Delta}_{\overline{\mathcal{S}}}\|_1}{\sqrt{s^*}} \le \frac{5\|\widetilde{\Delta}_{\mathcal{S}}\|_1}{3\sqrt{s^*}} = \frac{24\sqrt{s^*}\lambda}{5\rho_-(s^*+\widetilde{s})}.$$

Combining the above results, we have

$$\|\widetilde{\Delta}\|_2 = \sqrt{\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2 + \|\widetilde{\Delta}_{\overline{\mathcal{A}}}\|_2^2} \le \frac{17\sqrt{s^*}\lambda}{3\rho_-(s^*+\widetilde{s})} < R.$$

where the last inequality comes from the intial condition of θ. This conflicts with our assumption $\|\widetilde{\Delta}\|_2 = R$. Therefore we must have $\|\theta - \theta^*\|_2 \le R$. Consequently, we repeat the above proof for θ, and obtain

$$\|\Delta\|_2 \le \frac{17\sqrt{s^*}\lambda}{3\rho_-(s^*+\widetilde{s})} \quad \text{and} \quad \|\Delta\|_1 = \|\Delta_{\mathcal{S}}\|_1 + \|\Delta_{\overline{\mathcal{S}}}\|_1 \le \frac{23\sqrt{s^*}\lambda}{3\rho_-(s^*+\widetilde{s})}.$$

We now characterize the sparsity of θ. By Assumption 2.3.1 and the intial condition of θ, we have $\lambda = 2\lambda_N \ge 8\|\nabla\mathcal{L}(\theta^*)\|_\infty$, which further implies

$$\left|\left\{j \;\middle|\; |\nabla_j\mathcal{L}(\theta^*)| \ge \lambda/8,\ j \in \overline{\mathcal{S}}\right\}\right| = 0. \tag{A.5.12}$$

We then consider an arbitrary set $\mathcal{S}'$ such that

$$\mathcal{S}' = \left\{ j \;\middle|\; |\nabla_j \mathcal{L}(\theta) - \nabla_j \mathcal{L}(\theta^*)| \geq 5\lambda/8,\ j \in \overline{\mathcal{S}} \right\}.$$

Let $s' = |\mathcal{S}'|$. Then there exists v such that

$$\|v\|_\infty = 1,\ \ \|v\|_0 \leq s', \quad \text{and} \quad 5s'\lambda/8 \leq v^\top (\nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(\theta^*)).$$

Since $\mathcal{L}(\theta)$ is twice differentiable, then by the mean value theorem, there exists some $z_1 \in [0,1]$ such that

$$\ddot{\theta} = z_1 \theta + (1 - z_1)\theta^* \quad \text{and} \quad \nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(\theta^*) = \nabla^2 \mathcal{L}(\ddot{\theta})\Delta.$$

Then we have

$$\frac{5s'\lambda}{8} \leq v^\top \nabla^2 \mathcal{L}(\ddot{\theta})\Delta \leq \sqrt{v^\top \nabla^2 \mathcal{L}(\ddot{\theta}) v} \sqrt{\Delta^\top \nabla^2 \mathcal{L}(\ddot{\theta})\Delta}.$$

Since we have $\|v\|_0 \le s'$, then we obtain

$$\begin{aligned}
\frac{3s'\lambda}{4} &\le \sqrt{\rho_+(s')}\sqrt{s'}\sqrt{\Delta^\top(\nabla\mathcal{L}(\theta)-\nabla\mathcal{L}(\theta^*))} \\
&\le \sqrt{\rho_+(s')}\sqrt{s'}\sqrt{\|\Delta\|_1\cdot\|\nabla\mathcal{L}(\theta)-\nabla\mathcal{L}(\theta^*)\|_\infty} \\
&\le \sqrt{\rho_+(s')}\sqrt{s'}\sqrt{\|\Delta\|_1(\|\nabla\mathcal{L}(\theta)\|_\infty+\|\nabla\mathcal{L}(\theta^*)\|_\infty)} \\
&\le \sqrt{\rho_+(s')}\sqrt{s'}\sqrt{\|\Delta\|_1(\|\nabla\mathcal{L}(\theta)-\lambda\xi\|_\infty+\lambda\|\widetilde{\xi}\|_\infty+\|\nabla\mathcal{L}(\theta^*)\|_\infty)} \\
&\le \sqrt{\rho_+(s')}\sqrt{s'}\sqrt{\frac{115s^*\lambda^2}{12\rho_-(s^*+\widetilde{s})}}.
\end{aligned}$$

By simple manipulation, we have $\frac{5\sqrt{s'}}{8} \le \sqrt{\rho_+(s')}\sqrt{\frac{115s^*}{12\rho_-(s^*+\widetilde{s})}}$, which implies $s' \le \frac{184\rho_+(s')}{15\rho_-(s^*+\widetilde{s})}\cdot s^*$.

Since $\mathcal{S}'$ is arbitrary defined, by simple manipulation, we have

$$\left|\left\{j \;\middle|\; |\nabla_j\mathcal{L}(\theta)-\nabla_j\mathcal{L}(\theta^*)| \ge 5\lambda/8,\ j\in\overline{\mathcal{S}}\right\}\right| \le 13\kappa s^* < \widetilde{s}. \tag{A.5.13}$$

Thus, (A.5.12) and (A.5.13) imply

$$\left|\left\{j \;\middle|\; |\nabla_j\mathcal{L}(\widehat{\theta})+\frac{\lambda}{8}u_j| \ge 7\lambda/8,\ j\in\overline{\mathcal{S}}\cap\mathcal{A}\right\}\right| \le \widetilde{s}$$

for any $u\in\mathbb{R}^d$ satisfying $\|u\|_\infty \le 1$. Then there exists a $\xi_j\in\mathbb{R}$ satisfying

$$|\xi_j| \le 1 \quad \text{and} \quad \nabla_j\widetilde{\mathcal{L}}_\lambda(\widehat{\theta})+\lambda u_j/8+\lambda\xi_j = 0,$$

for any $j \in \overline{\mathcal{S}} \cap \mathcal{A}$ satisfying $|\nabla_j \mathcal{L}(\widehat{\theta}) + \lambda u_j/8| \leq 7\lambda/8$. This further implies $\theta_j = 0$. Thus, we have $\|\theta_{\overline{\mathcal{A}}}\|_0 \leq \widetilde{s}$.

Since θ is sufficiently sparse, we know that the restricted convexity holds for θ and θ^*. Then we refine our analysis for θ. By the restricted convexity of $\widetilde{\mathcal{F}}_\lambda(\theta)$, we have

$$\begin{aligned} &\widetilde{\mathcal{F}}_\lambda(\theta^*) - \frac{\rho_-(s^* + \widetilde{s})}{2}\|\Delta\|_2^2 \\ &\qquad \geq \widetilde{\mathcal{F}}_\lambda(\theta) - \Delta^\top(\nabla\mathcal{L}(\theta) + \lambda\widetilde{\xi}) \geq \widetilde{\mathcal{F}}_\lambda(\theta) - \frac{\lambda}{8}\|\Delta\|_1. \end{aligned} \tag{A.5.14}$$

By simple manipulation, we rewrite (A.5.14) as

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) \leq \lambda\|\theta^*\|_1 - \lambda\|\theta\|_1 + \frac{\lambda}{8}\|\Delta\|_1.$$

By the restricted convexity of $\mathcal{L}(\theta)$, we have

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) - \rho_-(s^* + \widetilde{s})\|\Delta\|_2^2 \geq -\frac{\lambda}{8}\|\Delta_{\mathcal{S}}\|_1 - \frac{\lambda}{8}\|\Delta_{\overline{\mathcal{S}}}\|_1, \tag{A.5.15}$$

where the last inequality comes from our assumption $\lambda \geq 8\|\nabla\mathcal{L}(\theta^*)\|_\infty$. By the

decomposability of the ℓ_1 norm, we have

$$\begin{aligned} &\|\theta^*\|_1 - \|\theta\|_1 + \frac{1}{8}\|\Delta\|_1 \\ &\quad = \|\theta^*_{\mathcal{S}}\|_1 - (\|\theta_{\mathcal{S}}\|_1 + \|\Delta_{\overline{\mathcal{S}}}\|_1) + \frac{1}{8}\|\Delta_{\mathcal{S}}\|_1 + \frac{1}{8}\|\Delta_{\overline{\mathcal{S}}}\|_1 \\ &\quad \leq \frac{9}{8}\|\Delta_{\mathcal{S}}\|_1 - (1-\delta)\|\Delta_{\overline{\mathcal{S}}}\|_1 \leq \frac{9}{8}\|\Delta_{\mathcal{S}}\|_1 - \frac{7}{8}\|\Delta_{\overline{\mathcal{S}}}\|_1, \end{aligned} \tag{A.5.16}$$

where the last inequality comes from $\delta < 1/8$ in Assumption 2.3.1. Combining (A.5.7) and (A.5.4) with (A.5.15) and (A.5.16), we obtain

$$\rho_-(s^* + \widetilde{s})\|\Delta\|_2^2 \leq \frac{5\lambda}{4}\|\Delta_{\mathcal{S}}\|_1 \leq \frac{5\lambda\sqrt{s^*}}{4}\|\Delta_{\mathcal{S}}\|_2 \leq \frac{5\lambda\sqrt{s^*}}{4}\|\Delta_{\mathcal{S}}\|_2,$$

which implies that

$$\|\Delta\|_2 \leq \frac{5\lambda\sqrt{s^*}}{4\rho_-(s^* + \widetilde{s})} \quad \text{and} \quad \|\Delta_{\mathcal{S}}\|_1 \leq \sqrt{s^*}\|\Delta_{\mathcal{S}}\|_2 \leq \frac{5\lambda s^*}{4\rho_-(s^* + \widetilde{s})}.$$

By (A.5.7), we further have

$$\|\Delta\|_1 \leq \frac{8}{3}\|\Delta_{\mathcal{S}}\|_1 \leq \frac{10\lambda s^*}{3\rho_-(s^* + \widetilde{s})}. \tag{A.5.17}$$

Plugging (A.5.17) into (A.5.14), we have

$$\widetilde{\mathcal{F}}_\lambda(\theta^*) \geq \widetilde{\mathcal{F}}_\lambda(\theta) + \frac{8\lambda^2 s^*}{7\rho_-(s^* + \widetilde{s})}.$$

By the concavity of $\mathcal{H}_\lambda(\theta)$ and Hölder's inequality, we have

$$\begin{aligned}\mathcal{H}_\lambda(\theta^{\text{relax}}) &\leq \mathcal{H}_\lambda(\theta^*) + (\theta^{\text{relax}} - \theta^*)^\top \nabla \mathcal{H}_\lambda(\theta^*) \\ &\leq \mathcal{H}_\lambda(\theta^*) + \|\theta^{\text{relax}} - \theta^*\|_1 \|\nabla \mathcal{H}_\lambda(\theta^*)\|_\infty.\end{aligned}$$

Since we have $\|\mathcal{H}_\lambda(\theta)\|_\infty \leq \lambda$, by Lemma 2.4.3, we have

$$\mathcal{H}_\lambda(\theta^{\text{relax}}) \leq \mathcal{H}_\lambda(\theta^*) + \lambda\|\theta^{\text{relax}} - \theta^*\|_1 \leq \mathcal{H}_\lambda(\theta^*) + \triangle_{\lambda_0}.$$

Since $\mathcal{F}_{\lambda_0}(\theta) = \widetilde{\mathcal{F}}_{\lambda_0}(\theta) + \mathcal{H}_{\lambda_0}(\theta)$, by Lemma 2.4.3 again, we have $\mathcal{F}_{\lambda_0}(\theta^{\text{relax}}) \leq \mathcal{F}_\lambda(\theta^*) + \triangle_{\lambda_0}$. Thus, θ^{relax} is a proper initial solution for solving (2.1.1) with λ_0 by PICASSO.

□

A.6 Proof of Theorem 2.7.9

Proof. Let $\widetilde{\xi} \in \partial\|\theta\|_1$ be a subgradient vector satisfying $\mathcal{K}_\lambda(\theta) = \|\nabla\widetilde{\mathcal{L}}_\lambda(\theta) + \lambda\widetilde{\xi}\|_\infty$. By the restricted convexity of $\widetilde{\mathcal{L}}_{\lambda'}(\theta)$, we have

$$\begin{aligned}
\mathcal{F}_{\lambda'}(\theta) - \mathcal{F}_{\lambda'}(\overline{\theta}^{\lambda'}) &\leq (\theta - \overline{\theta}^{\lambda'})^\top(\nabla\mathcal{L}(\theta) + \nabla\mathcal{H}_{\lambda'}(\theta) + \lambda'\widetilde{\xi}) \qquad \text{(A.6.1)}\\
&= (\theta - \overline{\theta}^{\lambda'})^\top(\nabla\mathcal{L}(\theta) + \nabla\mathcal{H}_{\lambda}(\theta)\\
&\qquad + \lambda\widetilde{\xi} - \lambda\widetilde{\xi} + \lambda'\widetilde{\xi} - \nabla\mathcal{H}_{\lambda}(\theta) + \nabla\mathcal{H}_{\lambda'}(\theta))\\
&\overset{\text{(i)}}{\leq} \|\theta - \overline{\theta}^{\lambda'}\|_1(\|\nabla\mathcal{L}(\theta) + \nabla\mathcal{H}_{\lambda}(\theta) + \lambda\widetilde{\xi}\|_\infty\\
&\qquad + (\lambda - \lambda') + \|\nabla\mathcal{H}_{\lambda}(\theta) - \nabla\mathcal{H}_{\lambda'}(\theta)\|_\infty)\\
&\overset{\text{(ii)}}{\leq} (\mathcal{K}_\lambda(\theta) + 3(\lambda - \lambda'))\|\theta - \overline{\theta}^{\lambda'}\|_1,
\end{aligned}$$

where (i) comes from Hölder's inequality and $\|\widetilde{\xi}\|_\infty \leq 1$, and (ii) comes from (R.3) of Lemma A.2.1. Meanwhile, since we have

$$\|\overline{\theta}^{\lambda'}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}, \;\; \mathcal{K}_{\lambda'}(\overline{\theta}^{\lambda'}) = 0 \leq \lambda'/4, \;\; \|\theta_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}, \text{ and } \;\; \mathcal{K}_\lambda(\theta) \leq \lambda/4,$$

following similar lines to the proof of Theorem 2.3.11, we have

$$\|\overline{\theta}^{\lambda'} - \theta^*\|_1 \leq \frac{25\lambda' s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})} \quad \text{and} \quad \|\theta - \theta^*\|_1 \leq \frac{25\lambda s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})},$$

which further implies

$$\|\theta - \overline{\theta}^{\lambda'}\|_1 \leq \|\theta^* - \theta\|_1 + \|\theta^* - \overline{\theta}^{\lambda'}\|_1 \leq \frac{50(\lambda + \lambda')s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})}. \tag{A.6.2}$$

Plugging (A.6.2) into (A.6.1), we obtain

$$\mathcal{F}_{\lambda'}(\theta) - \mathcal{F}_{\lambda'}(\overline{\theta}^{\lambda'}) \leq \frac{50(\mathcal{K}_\lambda(\theta) + 3(\lambda - \lambda'))(\lambda + \lambda')s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})}.$$

□

A.7 Lemmas for Statistical Theory

A.8 Proof of Theorem 2.3.14

Before we proceed with the main proof, we first introduce the following lemmas.

Lemma A.8.1. Suppose Assumptions 2.3.1, 2.3.5, and 2.3.7 hold. Then we have

$$\|\widehat{\theta}^{\{N\}} - \theta^*\|_2 = \mathcal{O}\Bigg(\underbrace{\frac{\|\nabla_{\mathcal{S}_1}\mathcal{L}(\theta^*)\|_2}{\widetilde{\rho}_-(s^* + 2\widetilde{s})}}_{V_1} + \underbrace{\frac{\lambda\sqrt{|\mathcal{S}_2|}}{\widetilde{\rho}_-(s^* + \widetilde{s})}}_{V_2} + \underbrace{\frac{\delta_N \lambda\sqrt{s^*}}{\widetilde{\rho}_-(s^* + \widetilde{s})}}_{V_3} \Bigg),$$

where $\mathcal{S}_1 = \{j \mid |\theta_j^*| \geq \gamma\lambda_N\}$ and $\mathcal{S}_2^* = \{j \mid 0 < |\theta_j^*| < \gamma\lambda_N\}$.

The proof of Lemma A.8.1 is provided in Appendix A.10. Lemma A.8.1 divides the estimation error of $\widehat{\theta}^{\{N\}}$ into three parts: V_1 is the error for strong signals; V_2 is the error for weak signals; V_3 is the optimization error.

Lemma A.8.2. Suppose Assumption 2.3.5 holds, X satisfies the column normalization condition, and the observation noise $\epsilon \sim N(0, \sigma^2 I)$ is Gaussian. We then have

$$\mathbb{P}\left(\frac{1}{n}\|X_{*\mathcal{S}_1}^{\top}\epsilon\|_2 \geq 3\sigma\sqrt{\frac{\rho_+(|\mathcal{S}_1|)\cdot|\mathcal{S}_1|}{n}}\right) \leq 2\exp\left(-2|\mathcal{S}_1|\right).$$

Lemma A.8.2 is a direct result of Hanson-Wright inequality [129], and therefore its proof is omitted. Lemma A.8.2 characterizes the large deviation properties of $\|\nabla_{\mathcal{S}_1}\mathcal{L}(\theta^*)\|_2$ in Lemma A.8.1 for sparse linear regression.

We the proceed with the main proof. For notational simplicity, we omit the index N and denote $\widehat{\theta}^{\{N\}}$, λ_N, and δ_N by $\widehat{\theta}$, λ, and δ respectively. If we choose a sufficiently small δ such that $\delta \leq \frac{1}{40\sqrt{s^*}}$, then we apply Lemmas A.8.1 and A.8.2, and obtain

$$\|\widehat{\Delta}\|_2 \leq \frac{3\sqrt{|\mathcal{S}_1|}\sigma}{\widetilde{\rho}_-(s^*+2\widetilde{s})}\sqrt{\frac{\rho_+(|\mathcal{S}_1|)\cdot|\mathcal{S}_1|}{n}} + \frac{3\lambda\sqrt{|\mathcal{S}_2|}}{\widetilde{\rho}_-(s^*+2\widetilde{s})} + \frac{0.3\lambda}{\widetilde{\rho}_-(s^*+2\widetilde{s})}.$$

Since all above results rely on Assumptions 2.3.1 and 2.3.5, by Lemma 2.3.13, we

have

$$\|\widehat{\Delta}\|_2 \leq \frac{15\sqrt{|\mathcal{S}_1|}\sigma}{\psi_\ell}\sqrt{\frac{\rho_+(|\mathcal{S}_1|)|\mathcal{S}_1|}{n}} + \frac{(96\sqrt{|\mathcal{S}_2|}+10)\sigma}{\psi_\ell}\sqrt{\frac{\log d}{n}}$$

with probability at least $1 - 2\exp(-2\log d) - 2\exp(-2\cdot|\mathcal{S}_1|)$.

A.9 Proof of Lemma 2.3.13

Proof. By Lemma 2.7.10, we have

$$\|\nabla\mathcal{L}(\theta^*)\|_\infty = \|\frac{1}{n}X^\top(y - X\theta^*)\|_\infty = \frac{1}{n}\|X^\top\epsilon\|_\infty. \tag{A.9.1}$$

Since we take $\lambda = 8\sigma\sqrt{\log d/n}$, combining (A.9.1) with Lemma 2.7.10, we obtain

$$\mathbb{P}(\lambda \geq 4\|\nabla\mathcal{L}(\theta^*)\|_\infty) \leq 1 - \frac{2}{d^2}.$$

Moreover, for any $v \in \mathbb{R}^d$ and $\|v\|_0 \leq s$, $\|v\|_1 \leq \sqrt{s}\|v\|_2$. Then (2.3.5) implies

$$\frac{\|Xv\|_2^2}{n} \geq \psi_\ell\|v\|_2^2 - \gamma_\ell\frac{s\log d}{n}\|v\|_2^2. \tag{A.9.2}$$

By simple manipulation, (A.9.2) implies

$$\frac{\|Xv\|_2^2}{n} \geq \frac{3\psi_\ell}{4}\|v\|_2^2 \tag{A.9.3}$$

for n large enough such that $\gamma_\ell \frac{s\log d}{n} \le \frac{\psi_\ell}{4}$. Similarly, (2.3.5) implies

$$\frac{\|Xv\|_2^2}{n} \le \frac{5\psi_u}{4}\|v\|_2^2 \tag{A.9.4}$$

for n large enough such that $\gamma_u \frac{s\log d}{n} \le \frac{\psi_u}{4}$. Since v is an arbitrary sparse vector, for $\alpha \le \psi_\ell/4$, (A.9.3) and (A.9.4) guarantee

$$\widetilde{\rho}_-(s) = \rho_-(s) - \alpha \ge \psi_\ell/2 \quad \text{and} \quad \rho_+(s) = \rho_-(s) \le 5\psi_u/4. \tag{A.9.5}$$

Let $s = s^* + 2\widetilde{s}$. (A.9.5) implies

$$484\kappa^2 + 100\kappa \le 484 \cdot \frac{25\psi_u^2}{4\psi_\ell^2} + 100 \cdot \frac{5\psi_u}{2\psi_\ell}.$$

Then we can choose C_1 as $C_1 = 3025 \cdot \frac{\psi_u^2}{\psi_\ell^2} + 250 \cdot \frac{\psi_u}{\psi_\ell}$ such that $\widetilde{s} = C_1 s^* \ge (484\kappa^2 + 100\kappa)s^*$. Meanwhile, we need a large enough n satisfying

$$\frac{\log d}{n} \le \frac{\psi_\ell}{4\gamma_\ell(s^* + 2C_1 s^*)} \quad \text{and} \quad \frac{\log d}{n} \le \frac{\psi_u}{4\gamma_u s^* + 2C_1 s^*}.$$

Moreover, we have

$$\begin{aligned}\lambda_0 = \|\frac{1}{n}Xy\|_\infty &\le \|\frac{1}{n}X^\top X\theta^*\|_\infty + \|\frac{1}{n}X^\top \epsilon\|_\infty \\ &\le \|\frac{1}{n}X^\top X\|_1 \|\theta^*\|_\infty + \mathcal{O}_P\left(\sigma\sqrt{\frac{\log d}{n}}\right).\end{aligned}$$

Given $\|\frac{1}{n}X^\top X\|_1 = \mathcal{O}(d)$ and $\|\theta^*\|_\infty = \mathcal{O}(d)$, for large enough n, we have

$$\lambda_0 = \mathcal{O}_P(d^2) \text{ and } N = \frac{\log \lambda_0/\lambda_N}{\log \eta} = \mathcal{O}_P\left(\log\left(\frac{d^2}{\sigma}\sqrt{\frac{n}{\log d}}\right)\right) = \mathcal{O}_P(\log d).$$

□

A.10 Proof of Lemma A.8.1

Proof. For notational simplicity, we omit the index N and denote $\widehat{\theta}^{[N]}$, λ_N, and δ_N by $\widehat{\theta}$, λ, and δ respectively. We define $\widehat{\Delta} = \widehat{\theta} - \theta^*$. Let $\widehat{\xi} \in \partial\|\widehat{\theta}\|_1$ be a subgradient vector satisfying $\mathcal{K}_\lambda(\widehat{\theta}) = \|\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) + \lambda\widehat{\xi}\|_\infty \le \delta\lambda$. Then by the restricted convexity of $\mathcal{F}_\lambda(\theta)$, we have

$$\mathcal{F}_\lambda(\widehat{\theta}) \ge \mathcal{F}_\lambda(\theta^*) + \widehat{\Delta}^\top(\nabla\widetilde{\mathcal{L}}_\lambda(\theta^*) + \lambda\widetilde{\xi}) + \frac{\widetilde{\rho}_-(s^* + \widetilde{s})}{2}\|\widehat{\Delta}\|_2^2, \tag{A.10.1}$$

$$\mathcal{F}_\lambda(\theta^*) \ge \mathcal{F}_\lambda(\widehat{\theta}) - \widehat{\Delta}^\top(\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) + \lambda\widehat{\xi}) + \frac{\widetilde{\rho}_-(s^* + \widetilde{s})}{2}\|\widehat{\Delta}\|_2^2, \tag{A.10.2}$$

where $\widetilde{\xi} \in \partial\|\theta^*\|_1$. Combining (A.10.1) with (A.10.2), we have

$$\begin{aligned}\widetilde{\rho}_-(s^* + 2\widetilde{s})\|\widehat{\Delta}\|_2^2 &\le \|\widehat{\Delta}\|_1\|\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\theta}) + \lambda\widehat{\xi}\|_\infty - \widehat{\Delta}^\top(\nabla\mathcal{L}(\theta^*) + \nabla\mathcal{H}_\lambda(\theta^*) + \lambda\widetilde{\xi}) \\ &\le \underbrace{|\widehat{\Delta}^\top(\nabla\mathcal{L}(\theta^*) + \nabla\mathcal{H}_\lambda(\theta^*) + \lambda\widetilde{\xi})|}_{V_0} + \underbrace{\delta\lambda\|\widehat{\Delta}\|_1}_{V4}.\end{aligned} \tag{A.10.3}$$

[Bounding V_0] We consider the following decomposition

$$|\widehat{\Delta}^\top(\nabla\mathcal{L}(\theta^*)+\nabla\mathcal{H}_\lambda(\theta^*)+\lambda\widetilde{\xi})| \\ \leq \sum_{\mathcal{A}\in\{\mathcal{S}_1,\mathcal{S}_2,\overline{\mathcal{S}}\}}|\widehat{\Delta}_{\mathcal{A}}^\top(\nabla_{\mathcal{A}}\mathcal{L}(\theta^*)+\nabla_{\mathcal{A}}\mathcal{H}_\lambda(\theta^*)+\lambda\widetilde{\xi}_{\mathcal{A}})|,$$

where $\mathcal{S}_1=\{j\mid|\theta_j^*|\geq\gamma\lambda\}$ and $\mathcal{S}_2=\{j\mid 0<|\theta_j^*|<\gamma\lambda\}$. For $\overline{\mathcal{S}}$, we have $\|\nabla_{\overline{\mathcal{S}}}\mathcal{L}(\theta^*)\|_\infty\leq\lambda/4$ and $\nabla_{\overline{\mathcal{S}}}\mathcal{H}_\lambda(\theta^*)=0$. Thus, there exists some $\widetilde{\xi}_{\overline{\mathcal{S}}}\in\partial\|\theta^*_{\overline{\mathcal{S}}}\|_1$ such that $\nabla_{\overline{\mathcal{S}}}\mathcal{L}(\theta^*)+\nabla_{\overline{\mathcal{S}}}\mathcal{H}_\lambda(\theta^*)+\lambda\widetilde{\xi}_{\overline{\mathcal{S}}}=0$, which implies

$$|\widehat{\Delta}^\top(\nabla_{\overline{\mathcal{S}}}\mathcal{L}(\theta^*)+\nabla_{\overline{\mathcal{S}}}\mathcal{H}_\lambda(\theta^*)+\lambda\widetilde{\xi}_{\overline{\mathcal{S}}})|=0. \tag{A.10.4}$$

For all $j\in\mathcal{S}_1$, we have $|\theta_j^*|>\gamma\lambda$ and $|\theta_j|$ is smooth at $\theta_j=\theta_j^*$. Thus, by (R.2) of Lemma A.2.1, we have $\nabla_{\mathcal{S}_1}\mathcal{H}_\lambda(\theta^*)+\lambda\widetilde{\xi}_{\mathcal{S}_1}=0$, which implies

$$|\widehat{\Delta}_{\mathcal{S}_1}^\top(\nabla_{\mathcal{S}_1}\mathcal{L}(\theta^*)+\nabla_{\mathcal{S}_1}\mathcal{H}_\lambda(\theta^*)+\lambda\widetilde{\xi}_{\mathcal{S}_1})|=|\widehat{\Delta}_{\mathcal{S}_1}^\top\nabla_{\mathcal{S}_1}\mathcal{L}(\theta^*)| \tag{A.10.5}$$
$$\leq\|\widehat{\Delta}_{\mathcal{S}_1}\|_2\|\nabla_{\mathcal{S}_1}\mathcal{L}(\theta^*)\|_2\leq\|\widehat{\Delta}\|_2\|\nabla_{\mathcal{S}_1}\mathcal{L}(\theta^*)\|_2.$$

We then consider $\mathcal{S}_2$. Then we have

$$|\widehat{\Delta}_{\mathcal{S}_2}^\top(\nabla_{\mathcal{S}_2}\mathcal{L}(\theta^*)+\nabla_{\mathcal{S}_2}\mathcal{H}_\lambda(\theta^*)+\lambda\widetilde{\xi}_{\mathcal{S}_2})| \tag{A.10.6}$$
$$\leq\|\widehat{\Delta}_{\mathcal{S}_2}\|_1(\|\nabla_{\mathcal{S}_2}\mathcal{L}(\theta^*)\|_\infty+\|\nabla_{\mathcal{S}_2}\mathcal{H}_\lambda(\theta^*)\|_\infty+\|\lambda\widetilde{\xi}_{\mathcal{S}_2}\|_\infty)\leq 3\lambda\sqrt{|\mathcal{S}_2|}\|\widehat{\Delta}\|_2.$$

Combining (A.10.4) and (A.10.5) with (A.10.6), we have

$$V_0 \leq \|\nabla_{\mathcal{S}_1}\mathcal{L}(\theta^*)\|_2\|\widehat{\Delta}\|_2 + 3\lambda\sqrt{|\mathcal{S}_2|}\|\widehat{\Delta}\|_2. \tag{A.10.7}$$

[Bounding V_4] We then proceed to bound V_4. Since θ satisfies the approximate KKT condition, by Theorem 2.3.11, we have $\|\widehat{\Delta}\|_1 \leq 11\sqrt{s^*}\|\widehat{\Delta}\|_2$. Thus, by (A.10.7) into (A.10.3), we have

$$\widetilde{\rho}_-(s^* + \widetilde{s})\|\widehat{\Delta}\|_2^2 \leq \|\nabla_{\mathcal{S}_1}\mathcal{L}(\theta^*)\|_2\|\widehat{\Delta}\|_2 + 3\lambda\sqrt{|\mathcal{S}_2|}\|\widehat{\Delta}\|_2 + 11\delta\lambda\sqrt{s^*}\|\widehat{\Delta}\|_2.$$

Solving the above inequality, we complete the proof. □

A.11 Proof of Lemma 2.7.11

Proof. We then proceed to establish the error bound of the oracle estimator under the ℓ_∞ norm. Since Lemma 2.3.13 guarantees that $\rho_-(s) > 0$, (2.3.7) is a strongly convex problem over $\theta_{\mathcal{S}}$ with a unique optimum

$$\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}} = (X_{*\mathcal{S}}^\top X_{*\mathcal{S}})^{-1} X_{*\mathcal{S}}^\top y. \tag{A.11.1}$$

Then conditioning on the event $\mathcal{E}_1 = \left\{\|X^\top \epsilon\|_\infty / n \le 2\sigma\sqrt{\log d/n}\right\}$, we rewrite (A.11.1) as

$$\begin{aligned} \|\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}} - \theta^*_{\mathcal{S}}\|_\infty &= \|(X^\top_{*\mathcal{S}} X_{*\mathcal{S}})^{-1} X^\top_{*\mathcal{S}}(y - \mathbb{E}y)\|_\infty \\ &= \|(X^\top_{*\mathcal{S}} X_{*\mathcal{S}})^{-1} X^\top_{*\mathcal{S}}\epsilon\|_\infty \le \frac{1}{\rho_-(s^*)n}\|X^\top_{*\mathcal{S}}\epsilon\|_\infty \le \frac{2\sigma}{\rho_-(s^*)}\sqrt{\frac{\log d}{n}}. \end{aligned} \tag{A.11.2}$$

Since θ^* satisfies (2.3.8), (A.11.2) implies

$$\begin{aligned} \min_{j\in\mathcal{S}}|\widehat{\theta}^{\mathrm{o}}_j| &= \min_{j\in\mathcal{S}}|\widehat{\theta}^{\mathrm{o}}_j - \theta^*_j + \theta^*_j| \ge \min_{j\in\mathcal{S}}|\theta^*_j| - \|\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}} - \theta^*_{\mathcal{S}}\|_\infty \\ &\ge \left(C_5\gamma - \frac{2}{\rho_-(s^*)}\right)\sigma\sqrt{\frac{\log d}{n}} \ge \left(C_5\gamma - \frac{4}{\psi_\ell}\right)\sigma\sqrt{\frac{\log d}{n}}, \end{aligned} \tag{A.11.3}$$

where the last inequality comes from Lemma 2.3.13. Taking $C_5 = 8 + \frac{4}{\gamma\psi_\ell}$, (A.11.3) implies

$$\min_{j\in\mathcal{S}}|\widehat{\theta}^{\mathrm{o}}_j| \ge \left(C_5\gamma - \frac{4}{\psi_\ell}\right)\sigma\sqrt{\frac{\log d}{n}} \ge 8\gamma\sigma\sqrt{\frac{\log d}{n}} = \gamma\lambda,$$

where the last equality comes from $\gamma \ge 4/\psi_\ell$. Then by (R.2) of Lemma A.2.1, we have

$$\nabla_{\mathcal{S}}\mathcal{H}_\lambda(\widehat{\theta}^{\mathrm{o}}) + \lambda\nabla\|\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}}\|_1 = 0. \tag{A.11.4}$$

Combining (A.11.4) with the optimality condition of (2.3.7), we have

$$\frac{1}{n}X_{*\mathcal{S}}(y - X\widehat{\theta^{\mathrm{o}}}) + \nabla_{\mathcal{S}}\mathcal{H}_{\lambda}(\widehat{\theta^{\mathrm{o}}}) + \lambda\nabla\|\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}}\|_1 = 0. \tag{A.11.5}$$

□

A.12 Proof of Lemma 2.7.12

Proof. We consider the decomposition

$$\begin{aligned}
\|X^{\top}_{*\overline{\mathcal{S}}}(y - X\widehat{\theta^{\mathrm{o}}})\|_{\infty} &= \|X^{\top}_{*\overline{\mathcal{S}}}(y - X_{*\mathcal{S}}\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}})\|_{\infty} \qquad \text{(A.12.1)}\\
&= \|X^{\top}_{*\overline{\mathcal{S}}}[X_{*\mathcal{S}}\theta^{*}_{\mathcal{S}} + \epsilon + X_{*\mathcal{S}}(X^{\top}_{*\mathcal{S}}X_{*\mathcal{S}})^{-1}X^{\top}_{*\mathcal{S}}(X_{*\mathcal{S}}\theta^{*}_{\mathcal{S}} + \epsilon)]\|_{\infty}\\
&= \|X^{\top}_{*\overline{\mathcal{S}}}(I - X_{*\mathcal{S}}(X^{\top}_{*\mathcal{S}}X_{*\mathcal{S}})^{-1}X^{\top}_{*\mathcal{S}})\epsilon\|_{\infty} \leq \|U^{\top}_{*\overline{\mathcal{S}}}\epsilon\|_{\infty},
\end{aligned}$$

where $U = X^{\top}(I - X_{*\mathcal{S}}(X^{\top}_{*\mathcal{S}}X_{*\mathcal{S}})^{-1}X^{\top}_{*\mathcal{S}})$. Conditioning on the event $\mathcal{E}_2 = \left\{\|U^{\top}\epsilon\|_{\infty}/n \leq 2\sigma\sqrt{\log d/n}\right\}$, (A.12.1) implies

$$\frac{1}{n}\|X^{\top}_{*\overline{\mathcal{S}}}(y - X\widehat{\theta^{\mathrm{o}}})\|_{\infty} \leq \frac{\lambda}{4}. \tag{A.12.2}$$

By (R.3) of Lemma A.2.1, we have $\nabla \mathcal{H}_\lambda(\widehat{\theta}^{\mathrm{o}}_{\overline{\mathcal{S}}}) = 0$. Since $|\theta_j|$ is non-differentiable at $\theta_j = 0$, then (A.12.2) implies that there exists some $\widehat{\xi}^{\mathrm{o}}_{\overline{\mathcal{S}}} \in \partial \|\widehat{\theta}^{\mathrm{o}}_{\overline{\mathcal{S}}}\|_1$ such that

$$\frac{1}{n} X^{\top}_{*\overline{\mathcal{S}}}(y - X\widehat{\theta}^{\mathrm{o}}) + \nabla_{\overline{\mathcal{S}}} \mathcal{H}_\lambda(\widehat{\theta}^{\mathrm{o}}) + \lambda \widehat{\xi}^{\mathrm{o}}_{\overline{\mathcal{S}}} = 0. \tag{A.12.3}$$

□

Appendix B

Supporting Proof for Chapter 3

B.1 Proof of Lemma 3.3.3

For notational convenience, denote $\theta' = \mathcal{H}_k(\theta)$. Let $\text{supp}(\theta^*) = \mathcal{I}^*$, $\text{supp}(\theta) = \mathcal{I}$, $\text{supp}(\theta') = \mathcal{I}'$, and $\theta'' = \theta - \theta'$ with $\text{supp}(\theta'') = \mathcal{I}''$. Clearly we have $\mathcal{I}' \cup \mathcal{I}'' = \mathcal{I}$, $\mathcal{I}' \cap \mathcal{I}'' = \emptyset$, and $\|\theta\|_2^2 = \|\theta'\|_2^2 + \|\theta''\|_2^2$. Then we have

$$\begin{aligned} \|\theta' - \theta^*\|_2^2 - \|\theta - \theta^*\|_2^2 &= \|\theta'\|_2^2 - 2\langle\theta', \theta^*\rangle - \|\theta\|_2^2 + 2\langle\theta, \theta^*\rangle \\ &= 2\langle\theta'', \theta^*\rangle - \|\theta''\|_2^2. \end{aligned} \tag{B.1.1}$$

If $2\langle\theta'', \theta^*\rangle - \|\theta''\|_2^2 \leq 0$, then (3.3.3) holds naturally. From this point on, we will discuss the situation when $2\langle\theta'', \theta^*\rangle - \|\theta''\|_2^2 > 0$.

Let $\mathcal{I}^* \cap \mathcal{I}' = \mathcal{I}^{*1}$ and $\mathcal{I}^* \cap \mathcal{I}'' = \mathcal{I}^{*2}$, and denote $(\theta^*)_{\mathcal{I}^{*1}} = \theta^{*1}$, $(\theta^*)_{\mathcal{I}^{*2}} = \theta^{*2}$,

$(\theta')_{\mathcal{I}^{*1}} = \theta^{1*}$, and $(\theta'')_{\mathcal{I}^{*2}} = \theta^{2*}$. Then we have

$$\begin{aligned} 2\langle \theta'', \theta^* \rangle - \|\theta''\|_2^2 &= 2\langle \theta^{2*}, \theta^{*2} \rangle - \|\theta''\|_2^2 \\ &\leq 2\langle \theta^{2*}, \theta^{*2} \rangle - \|\theta^{2*}\|_2^2 \leq 2\|\theta^{2*}\|_2 \|\theta^{*2}\|_2 - \|\theta^{2*}\|_2^2. \end{aligned} \tag{B.1.2}$$

Let $|\text{supp}(\theta^{2*})| = |\mathcal{I}^{*2}| = k^{**}$ and $\theta_{2,\max} = \|\theta^{2*}\|_\infty$, then consequently we have $\|\theta^{2*}\|_2 = m \cdot \theta_{2,\max}$ for some $m \in [1, \sqrt{k^{**}}]$. Notice that we are interested in $1 \leq k^{**} \leq k^*$, since (3.3.3) holds naturally if $k^{**} = 0$. In terms of $\|\theta^{*2}\|_2$, the R.H.S. of (B.1.2) is maximized in the following three cases.

Case 1: $m = 1$, if $\|\theta^{*2}\|_2 \leq \theta_{2,\max}$;

Case 2: $m = \frac{\|\theta^{*2}\|_2}{\theta_{2,\max}}$, if $\theta_{2,\max} < \|\theta^{*2}\|_2 < \sqrt{k^{**}}\theta_{2,\max}$, ;

Case 3: $m = \sqrt{k^{**}}$, if $\|\theta^{*2}\|_2 \geq \sqrt{k^{**}}\theta_{2,\max}$.

Case 1: If $\|\theta^{*2}\|_2 \leq \theta_{2,\max}$, then the R.H.S. of (B.1.2) is maximized when $m = 1$, i.e. θ^{2*} has only one nonzero element $\theta_{2,\max}$. From (B.1.2), we have

$$2\langle \theta'', \theta^* \rangle - \|\theta''\|_2^2 \leq 2\theta_{2,\max}\|\theta^{*2}\|_2 - \theta_{2,\max}^2 \leq 2\theta_{2,\max}^2 - \theta_{2,\max}^2 = \theta_{2,\max}^2. \tag{B.1.3}$$

Denote $\theta_{1,\min}$ as the smallest element of θ^{1*} (in magnitude), which indicates that $|\theta_{1,\min}| \geq |\theta_{2,\max}|$ as θ' contains the largest k entries and θ'' contains the smallest

$d-k$ entries of θ. For $\|\theta-\theta^*\|_2^2$, we have

$$\begin{aligned}\|\theta-\theta^*\|_2^2 &= \|\theta'-\theta^{*1}\|_2^2+\|\theta''-\theta^{*2}\|_2^2 \\ &= \|\theta_{(\mathcal{I}^{*1})^C}\|_2^2+\|\theta_{\mathcal{I}^{*1}}-\theta^{*1}\|_2^2+\|\theta^{*2}\|_2^2-(2\langle\theta'',\theta^*\rangle-\|\theta''\|_2^2) \qquad \text{(B.1.4)} \\ &\geq (k-k^*+k^{**})\theta_{1,\min}^2-\theta_{2,\max}^2, \qquad \text{(B.1.5)}\end{aligned}$$

where the last inequality follows from the fact that $\theta_{(\mathcal{I}^{*1})^C}$ has $k-k^*+k^{**}$ entries larger than $\theta_{1,\min}$ (in magnitude). Combining (B.1.1), (B.1.3), and (B.1.5), we have

$$\begin{aligned}\frac{\|\theta'-\theta^*\|_2^2-\|\theta-\theta^*\|_2^2}{\|\theta-\theta^*\|_2^2} &\leq \frac{\theta_{2,\max}^2}{(k-k^*+k^{**})\theta_{1,\min}^2-\theta_{2,\max}^2} \\ &\leq \frac{\theta_{2,\max}^2}{(k-k^*+k^{**})\theta_{2,\max}^2-\theta_{2,\max}^2} \leq \frac{1}{k-k^*}. \qquad \text{(B.1.6)}\end{aligned}$$

Case 2: If $\theta_{2,\max}<\|\theta^{*2}\|_2<\sqrt{k^{**}}\theta_{2,\max}$, then the R.H.S. of (B.1.2) is maximized when $m=\frac{\|\theta^{*2}\|_2}{\theta_{2,\max}}$. From (B.1.2), we have

$$2\langle\theta'',\theta^*\rangle-\|\theta''\|_2^2 \leq 2\sqrt{k^{**}}\theta_{2,\max}\cdot m\theta_{2,\max}-\theta_{2,\max}^2 \leq k^{**}\theta_{2,\max}^2. \qquad \text{(B.1.7)}$$

From (B.1.4), we have

$$\|\theta-\theta^*\|_2^2 \geq (k-k^*+k^{**})\theta_{1,\min}^2+m^2\theta_{2,\max}^2-\theta_{2,\max}^2 \geq (k-k^*+k^{**})\theta_{1,\min}^2. \qquad \text{(B.1.8)}$$

Combining (B.1.1), (B.1.7), and (B.1.8), we have

$$\frac{\|\theta' - \theta^*\|_2^2 - \|\theta - \theta^*\|_2^2}{\|\theta - \theta^*\|_2^2} \leq \frac{k^{**}\theta_{2,\max}^2}{(k - k^* + k^{**})\theta_{1,\min}^2} \leq \frac{k^{**}}{k - k^* + k^{**}}. \tag{B.1.9}$$

Case 3: If $\|\theta^{*2}\|_2 \geq \sqrt{k^{**}}\theta_{2,\max}$, then the R.H.S. of (B.1.2) is maximized when $m = \sqrt{k^{**}}$. Let $\|\theta^{*2}\|_2 = \gamma\theta_{2,\max}$ for some $\gamma \geq \sqrt{k^{**}}$. From (B.1.2), we have

$$2\langle\theta'', \theta^*\rangle - \|\theta''\|_2^2 \leq 2\gamma\sqrt{k^{**}}\theta_{2,\max}^2 - k^{**}\theta_{2,\max}^2. \tag{B.1.10}$$

From (B.1.4), we have

$$\|\theta - \theta^*\|_2^2 \geq (k - k^* + k^{**})\theta_{1,\min}^2 + \gamma^2\theta_{2,\max}^2 - \gamma\sqrt{k^{**}}\theta_{2,\max}^2 + k^{**}\theta_{2,\max}^2. \tag{B.1.11}$$

Combining (B.1.1), (B.1.10), and (B.1.11), we have

$$\begin{aligned}\frac{\|\theta' - \theta^*\|_2^2 - \|\theta - \theta^*\|_2^2}{\|\theta - \theta^*\|_2^2} &\leq \frac{2\gamma\sqrt{k^{**}}\theta_{2,\max}^2 - k^{**}\theta_{2,\max}^2}{(k - k^* + k^{**})\theta_{1,\min}^2 + \gamma^2\theta_{2,\max}^2 - \gamma\sqrt{k^{**}}\theta_{2,\max}^2 + k^{**}\theta_{2,\max}^2} \\ &\leq \frac{2\gamma\sqrt{k^{**}} - k^{**}}{k - k^* + 2k^{**} + \gamma^2 - 2\gamma\sqrt{k^{**}}}. \end{aligned} \tag{B.1.12}$$

Inspecting the R.H.S. of (B.1.12) carefully, we can see that it is either a bell shape function or a monotone decreasing function when $\gamma \geq \sqrt{k^{**}}$. Setting the first derivative of the R.H.S. in terms of γ to zero, we have $\gamma = \frac{1}{2}\sqrt{k^{**}} + \sqrt{k - k^* + \frac{5}{4}k^{**}}$ (the other root is smaller than $\sqrt{k^{**}}$). Denoting $\gamma_* = \max\{\sqrt{k^{**}}, \frac{1}{2}\sqrt{k^{**}} + \sqrt{k - k^* + \frac{5}{4}k^{**}}\}$ and

plugging it into the R.H.S. of (B.1.12), we have

$$\frac{\|\theta' - \theta^*\|_2^2 - \|\theta - \theta^*\|_2^2}{\|\theta - \theta^*\|_2^2} \leq \max\left\{\frac{k^{**}}{k - k^* + k^{**}}, \frac{2\sqrt{k^{**}}}{2\sqrt{k - k^* + \frac{5}{4}k^{**}} - \sqrt{k^{**}}}\right\}. \tag{B.1.13}$$

Combining (B.1.6), (B.1.9), and (B.1.13), and taking $k > k^*$ and $k^* \geq k^{**} \geq 1$ into consideration, we have

$$\begin{aligned}\max\left\{\frac{1}{k - k^*}, \frac{k^{**}}{k - k^* + k^{**}}, \frac{2\sqrt{k^{**}}}{2\sqrt{k - k^* + \frac{5}{4}k^{**}} - \sqrt{k^{**}}}\right\} &\leq \frac{2\sqrt{k^{**}}}{2\sqrt{k - k^* + \frac{5}{4}k^{**}} - \sqrt{k^{**}}} \\ &\leq \frac{2\sqrt{k^*}}{2\sqrt{k - k^*} - \sqrt{k^*}} \leq \frac{2\sqrt{k^*}}{\sqrt{k - k^*}},\end{aligned}$$

which finishes the proof.

B.2 Proof of Lemma 3.3.5

It is straightforward that the stochastic variance reduced gradient (3.3.4) satisfies

$$\mathbb{E}g^{(t)}(\theta^{(t)}) = \mathbb{E}\nabla f_{i_t}(\theta^{(t)}) - \mathbb{E}\nabla f_{i_t}(\widetilde{\theta}) + \widetilde{\mu} = \nabla\mathcal{F}(\theta^{(t)}).$$

Thus $g^{(t)}(\theta^{(t)})$ is a unbiased estimator of $\nabla\mathcal{F}(\theta^{(t)})$ and the first claim is verified.

Next, we bound $\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2$. For any $i \in [n]$ and θ with $\mathrm{supp}(\theta) \subseteq \mathcal{I}$, consider

$$\phi_i(\theta) = f_i(\theta) - f_i(\theta^*) - \langle \nabla f_i(\theta^*), \theta - \theta^* \rangle.$$

Since $\nabla \phi_i(\theta^*) = \nabla f_i(\theta^*) - \nabla f_i(\theta^*) = 0$, we have $\phi_i(\theta^*) = \min_\theta \phi_i(\theta)$, which implies

$$\begin{aligned} 0 = \phi_i(\theta^*) &\le \min_\eta \phi_i(\theta - \eta \nabla_{\mathcal{I}} \phi_i(\theta)) \le \min_\eta \phi_i(\theta) - \eta \|\nabla_{\mathcal{I}} \phi_i(\theta)\|_2^2 + \frac{\rho_s^+ \eta^2}{2} \|\nabla_{\mathcal{I}} \phi_i(\theta)\|_2^2 \\ &= \phi_i(\theta) - \frac{1}{2\rho_s^+} \|\nabla_{\mathcal{I}} \phi_i(\theta)\|_2^2, \end{aligned} \tag{B.2.1}$$

where the second inequality follows from the RSS condition and the last equality follows from the fact that $\eta = 1/\rho_s^+$ minimizes the function. From (B.2.1), we have

$$\|\nabla_{\mathcal{I}} f_i(\theta) - \nabla_{\mathcal{I}} f_i(\theta^*)\|_2^2 \le 2\rho_s^+ \left[f_i(\theta) - f_i(\theta^*) - \langle \nabla_{\mathcal{I}} f_i(\theta^*), \theta - \theta^* \rangle \right]. \tag{B.2.2}$$

Since the sampling of i from $[n]$ is uniform, we have from (B.2.2)

$$\begin{aligned} \mathbb{E}\|\nabla_{\mathcal{I}} f_i(\theta) - \nabla_{\mathcal{I}} f_i(\theta^*)\|_2^2 &= \frac{1}{n} \sum_{i=1}^n \|\nabla_{\mathcal{I}} f_i(\theta) - \nabla_{\mathcal{I}} f_i(\theta^*)\|_2^2 \\ &\le 2\rho_s^+ \left[\mathcal{F}(\theta) - \mathcal{F}(\theta^*) - \langle \nabla_{\mathcal{I}} \mathcal{F}(\theta^*), \theta - \theta^* \rangle \right] \\ &\le 2\rho_s^+ \left[\mathcal{F}(\theta) - \mathcal{F}(\theta^*) + |\langle \nabla_{\mathcal{I}} \mathcal{F}(\theta^*), \theta - \theta^* \rangle| \right] \\ &\le 4\rho_s^+ \left[\mathcal{F}(\theta) - \mathcal{F}(\theta^*) \right], \end{aligned} \tag{B.2.3}$$

where the last inequality is from the convexity of $\mathcal{F}(\theta)$.

By the definition of $g_{\mathcal{I}}^{(t)}$ in (3.3.4), we can verify the second claim as

$$\begin{aligned}
\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2 &\leq 3\mathbb{E}\|\Big[\nabla_{\mathcal{I}} f_{i_t}(\widetilde{\theta}) - \nabla_{\mathcal{I}} f_{i_t}(\theta^*)\Big] - \nabla_{\mathcal{I}}\mathcal{F}(\widetilde{\theta}) + \nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2 \\
&\quad + 3\mathbb{E}\|\nabla_{\mathcal{I}} f_{i_t}(\theta^{(t)}) - \nabla_{\mathcal{I}} f_{i_t}(\theta^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2 \\
&\leq 3\mathbb{E}\|\nabla_{\mathcal{I}} f_{i_t}(\theta^{(t)}) - \nabla_{\mathcal{I}} f_{i_t}(\theta^*)\|_2^2 + 3\mathbb{E}\|\nabla_{\mathcal{I}} f_{i_t}(\widetilde{\theta}) - \nabla_{\mathcal{I}} f_{i_t}(\theta^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2 \\
&\leq 12\rho_s^+\Big[\mathcal{F}(\theta^{(t)}) - \mathcal{F}(\theta^*) + \mathcal{F}(\widetilde{\theta}) - \mathcal{F}(\theta^*)\Big] + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2, \qquad \text{(B.2.4)}
\end{aligned}$$

where the first inequality follows from $\|a+b+c\|_2^2 \leq 3\|a\|_2^2 + 3\|b\|_2^2 + 3\|c\|_2^2$, the second inequality follows from $\mathbb{E}\|x - \mathbb{E}x\|_2^2 \leq \mathbb{E}\|x\|_2^2$ with $\mathbb{E}\Big[\nabla_{\mathcal{I}} f_{i_t}(\widetilde{\theta}) - \nabla_{\mathcal{I}} f_{i_t}(\theta^*)\Big] = \nabla_{\mathcal{I}}\mathcal{F}(\widetilde{\theta}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^*)$, and the last inequality follows from (B.2.3).

B.3 Proof of Lemma 3.3.9

For any $\theta, \theta' \in \mathbb{R}^d$ in sparse linear model, we have $\nabla^2\mathcal{F}(\theta) = A^\top A$ and there exists some θ'' such that

$$\mathcal{F}(\theta) - \mathcal{F}(\theta') - \langle \nabla\mathcal{F}(\theta'), \theta - \theta' \rangle = \frac{1}{2}(\theta - \theta')^\top \nabla^2\mathcal{F}(\theta'')(\theta - \theta') = \frac{1}{2}\|A(\theta - \theta')\|_2^2,$$

where $\|\theta - \theta'\|_0 \leq 2k \leq s$. Let $v = \theta - \theta'$, then $\|v\|_0 \leq s$ and $\|v\|_1^2 \leq s\|v\|_2^2$. From (3.3.10), we have

$$\frac{\|Av\|_2^2}{nb} \geq \psi_1\|v\|_2^2 - \varphi_1\frac{s\log d}{nb}\|v\|_2^2 \ \text{ and } \ \frac{\|A_{\mathcal{S}_i *}v\|_2^2}{b} \leq \psi_2\|v\|_2^2 + \varphi_2\frac{s\log d}{b}\|v\|_2^2, \forall i \in [n].$$

The inequality above further imply

$$\rho_s^- = \inf_{\|v\|_0 \le s} \frac{\|Av\|_2^2}{nb\|v\|_2^2} \ge \psi_1 - \varphi_1 \frac{s \log d}{nb} \quad \text{and}$$
$$\rho_s^+ = \sup_{\|v\|_0 \le s, i \in [n]} \frac{\|A_{\mathcal{S}_i *} v\|_2^2}{b\|v\|_2^2} \le \psi_2 + \varphi_2 \frac{s \log d}{b}. \tag{B.3.1}$$

If $b \ge \frac{\varphi_2 s \log d}{\psi_2}$ and $n \ge \frac{2\varphi_1 \psi_2}{\psi_1 \varphi_2}$, then we have $nb \ge \frac{2\varphi_1 s \log d}{\psi_1}$. Combining these with (B.3.1), we have

$$\rho_s^- \ge \frac{1}{2}\psi_1, \text{ and } \rho_s^+ \le 2\psi_2.$$

This implies $\kappa_s = \frac{\rho_s^+}{\rho_s^-} \le \frac{4\psi_2}{\psi_1}$. Then there exists some $C_5 \ge \frac{16 C_1 \psi_2^2}{\psi_1^2}$ such that

$$k = C_5 k^* \ge C_1 \kappa_s^2 k^*.$$

B.4 Proof of Lemma 3.3.18

Let $\Theta = U\Sigma V^\top$ and $\Theta^* = U^*\Sigma^* V^{*\top}$ be the singular value decomposition of Θ and Θ^* respectively. Since Σ and Σ^* are diagonal, if $k > k^*$, we have from Lemma 3.3.3

$$\|\mathcal{R}_k(\Sigma) - \Sigma^*\|_F^2 \le \left(1 + \frac{2\sqrt{k^*}}{\sqrt{k - k^*}}\right) \|\Sigma - \Sigma^*\|_F^2. \tag{B.4.1}$$

Then we have

$$\|\mathcal{R}_k(\Theta)-\Theta^*\|_F^2-\|\Theta-\Theta^*\|_F^2=\|\mathcal{R}_k(\Theta)\|_F^2-\|\Theta\|_F^2+2\langle\Theta-\mathcal{R}_k(\Theta),\Theta^*\rangle$$

$$=\|\mathcal{R}_k(\Sigma)\|_F^2-\|\Sigma\|_F^2+2\langle\Theta-\mathcal{R}_k(\Theta),\Theta^*\rangle\leq\|\mathcal{R}_k(\Sigma)\|_F^2-\|\Sigma\|_F^2+2\sum_{i=1}^{k^*}\sigma_i(\Theta-\mathcal{R}_k(\Theta))\cdot\sigma_i(\Theta^*)$$

$$=\|\mathcal{R}_k(\Sigma)\|_F^2-\|\Sigma\|_F^2+2\sum_{i=1}^{k^*}(\sigma_i(\Theta)-\sigma_i(\mathcal{R}_k(\Theta)))\cdot\sigma_i(\Theta^*)=\|\mathcal{R}_k(\Sigma)-\Sigma^*\|_F^2-\|\Sigma-\Sigma^*\|_F^2$$

$$\leq\frac{2\sqrt{k^*}}{\sqrt{k-k^*}}\cdot\|\Sigma-\Sigma^*\|_F^2\leq\frac{2\sqrt{k^*}}{\sqrt{k-k^*}}\cdot\|\Theta-\Theta^*\|_F^2,$$

where the first and last inequalities are from $\langle A,B\rangle\leq\sum_{i=1}^{\min\{\mathrm{rank}(A),\mathrm{rank}(B)\}}\sigma_i(A)\cdot\sigma_i(B)$ for matrices $A,B\in\mathbb{R}^{d\times p}$ and the second inequality is from (B.4.1). This finishes the proof.

Appendix C

Supporting Proof for Chapter 4

C.1 Lemmas for Theorem 4.3.3 (Alternating Exact Minimization)

C.1.1 Proof of Lemma 4.4.1

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^{*}$ and V^{*} respectively. Then we define two $nk \times nk$ matrices

$$S^{(t)} = \begin{bmatrix} S_{11}^{(t)} & \cdots & S_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ S_{k1}^{(t)} & \cdots & S_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad S_{pq}^{(t)} = \sum_{i=1}^{d} A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{(t)\top} A_i^{\top},$$

$$G^{(t)} = \begin{bmatrix} G_{11}^{(t)} & \cdots & G_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ G_{k1}^{(t)} & \cdots & G_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad G_{pq}^{(t)} = \sum_{i=1}^{d} A_i \overline{U}_{*p}^{*} \overline{U}_{*q}^{*\top} A_i^{\top}$$

for $1 \leq p, q \leq k$. Note that $S^{(t)}$ and $G^{(t)}$ are essentially the partial Hessian matrices $\nabla_V^2 \mathcal{F}(\overline{U}^{(t)}, V)$ and $\nabla_V^2 \mathcal{F}(\overline{U}^{*}, V)$ for a vectorized V, i.e., $\mathrm{vec}(V) \in \mathbb{R}^{nk}$. Before we proceed with the main proof, we first introduce the following lemma.

Lemma C.1.1. Suppose that $\mathcal{A}(\cdot)$ satisfies $2k$-RIP with parameter δ_{2k}. We then have

$$1 + \delta_{2k} \geq \sigma_{\max}(S^{(t)}) \geq \sigma_{\min}(S^{(t)}) \geq 1 - \delta_{2k}.$$

The proof of Lemma C.1.1 is provided in Appendix C.1.7. Note that Lemma C.1.1 is also applicable $G^{(t)}$, since $G^{(t)}$ shares the same structure with $S^{(t)}$.

We then proceed with the proof of Lemma 4.4.1. Given a fixed $\overline{U}$, $\mathcal{F}(\overline{U}, V)$ is a quadratic function of V. Therefore we have

$$\begin{aligned} \mathcal{F}(\overline{U}, V') = \mathcal{F}(\overline{U}, V) \\ + \langle \nabla_V \mathcal{F}(\overline{U}, V), V' - V \rangle + \langle \mathrm{vec}(V') - \mathrm{vec}(V), \nabla_V^2 F(\overline{U}, V)(\mathrm{vec}(V') - \mathrm{vec}(V)) \rangle, \end{aligned}$$

which further implies implies

$$\mathcal{F}(\overline{U}, V') - \mathcal{F}(\overline{U}, V) - \langle \nabla F_V(\overline{U}, V), V' - V \rangle \leq \sigma_{\max}(\nabla_V^2 F(\overline{U}, V)) \|V' - V\|_{\mathrm{F}}^2$$

$$\mathcal{F}(\overline{U}, V') - \mathcal{F}(\overline{U}, V) - \langle \nabla F_V(\overline{U}, V), V' - V \rangle \geq \sigma_{\min}(\nabla_V^2 F(\overline{U}, V)) \|V' - V\|_{\mathrm{F}}^2.$$

Then we can verify that $\nabla_V^2 F(U, V)$ also shares the same structure with $S^{(t)}$. Thus applying Lemma C.1.1 to the above two inequalities, we complete the proof. □

C.1.2 Proof of Lemma 4.4.3

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^*$ and V^* respectively. We define two $nk \times nk$ matrices

$$J^{(t)} = \begin{bmatrix} J_{11}^{(t)} & \cdots & J_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ J_{k1}^{(t)} & \cdots & J_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad J_{pq}^{(t)} = \sum_{i=1}^{d} A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{*\top} A_i^{\top},$$

$$K^{(t)} = \begin{bmatrix} K_{11}^{(t)} & \cdots & K_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ K_{k1}^{(t)} & \cdots & K_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad K_{pq}^{(t)} = \overline{U}_{*p}^{(t)\top} \overline{U}_{*q}^{*} I_n$$

for $1 \leq p, q \leq k$. Before we proceed with the main proof, we first introduce the following lemmas.

Lemma C.1.2. Suppose that $\mathcal{A}(\cdot)$ satisfies $2k$-RIP with parameter δ_{2k}. We then have

$$\|S^{(t)}K^{(t)} - J^{(t)}\|_2 \leq 3\delta_{2k}\sqrt{k}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}.$$

The proof of Lemma C.1.2 is provided in Appendix C.1.8. Note that Lemma C.1.2 is also applicable to $G^{(t)}K^{(t)} - J^{(t)}$, since $G^{(t)}$ and $S^{(t)}$ share the same structure.

Lemma C.1.3. Given $F \in \mathbb{R}^{k\times k}$, we define a $nk \times nk$ matrix

$$\mathbb{F} = \begin{bmatrix} F_{11}I_n & \cdots & F_{1k}I_n \\ \vdots & \ddots & \vdots \\ F_{k1}I_n & \cdots & F_{kk}I_n \end{bmatrix}.$$

For any $V \in \mathbb{R}^{n\times k}$, let $v = \mathrm{vec}(V) \in \mathbb{R}^{nk}$, then we have $\|\mathbb{F}v\|_2 = \|FV^\top\|_{\mathrm{F}}$.

Proof. By linear algebra, we have

$$[FV]_{ij} = F_{i*}^\top V_{j*} = \sum_{\ell=1}^{k} F_{i\ell}V_{j\ell} = \sum_{\ell=1}^{k} F_{i\ell}I_{*\ell}^\top V_{*\ell},$$

which completes the proof. □

We then proceed with the proof of Lemma 4.4.3. Since $b_i = \mathrm{tr}(V^{*\top}A_iU^*)$, then

we rewrite $\mathcal{F}(\overline{U}, V)$ as

$$\mathcal{F}(\overline{U}, V) = \frac{1}{2}\sum_{i=1}^{d}\Big(\operatorname{tr}(V^\top A_i \overline{U}) - b_i\Big)^2 = \frac{1}{2}\sum_{i=1}^{d}\Bigg(\sum_{j=1}^{k} V_{j*}^\top A_i \overline{U}_{*j} - \sum_{j=1}^{k} V_{j*}^{*\top} A_i \overline{U}_{*j}^*\Bigg)^2.$$

For notational simplicity, we define $v = \operatorname{vec}(V)$. Since $V^{(t+0.5)}$ minimizes $\mathcal{F}(\overline{U}^{(t)}, V)$, we have

$$\operatorname{vec}\Big(\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t+0.5)})\Big) = S^{(t)} v^{(t+0.5)} - J^{(t)} v^* = 0.$$

Solving the above system of equations, we obtain

$$v^{(t+0.5)} = (S^{(t)})^{-1} J^{(t)} v^*. \tag{C.1.1}$$

Meanwhile, we have

$$\begin{aligned}\operatorname{vec}(\nabla_V \mathcal{F}(\overline{U}^*, V^{(t+0.5)})) &= G^{(t)} v^{(t+0.5)} - G^{(t)} v^* \\ &= G^{(t)} (S^{(t)})^{-1} J^{(t)} v^* - G^{(t)} v^* = G^{(t)}\Big((S^{(t)})^{-1} J^{(t)} - I_{nk}\Big) v^*, \end{aligned} \tag{C.1.2}$$

where the second equality come from (C.1.1). By triangle inequality, (C.1.2) fur-

ther implies

$$\begin{aligned}\|((S^{(t)})^{-1}J^{(t)} - I_{nk})v^*\|_2 &\le \|(K^{(t)} - I_{nk})v^*\|_2 + \|(S^{(t)})^{-1}(J^{(t)} - S^{(t)}K^{(t)})v^*\|_2 \\ &\le \|(\overline{U}^{(t)\top}\overline{U}^* - I_k)V^{*\top}\|_{\mathrm{F}} + \|(S^{(t)})^{-1}\|_2\|(J^{(t)} - S^{(t)}K^{(t)})v^*\|_2 \\ &\le \|\overline{U}^{(t)\top}\overline{U}^* - I_k\|_{\mathrm{F}}\|V^*\|_2 + \|(S^{(t)})^{-1}\|_2\|(J^{(t)} - S^{(t)}K^{(t)})v^*\|_2,\end{aligned} \tag{C.1.3}$$

where the second inequality comes from Lemma C.1.3. Plugging (C.1.3) into (C.1.2), we have

$$\begin{aligned}\|\mathrm{vec}(\nabla_V \mathcal{F}(\overline{U}^*, V^{(t+0.5)}))\|_2 &\le \|G^{(t)}\|_2\|((S^{(t)})^{-1}J^{(t)} - I_{nk})v^*\|_2 \\ &\overset{\text{(i)}}{\le}(1+\delta_{2k})(\sigma_1\|\overline{U}^{(t)\top}\overline{U}^* - I_k\|_2 + \|(S^{(t)})^{-1}\|_2\|S^{(t)}K^{(t)} - J^{(t)}\|_2\sigma_1\sqrt{k}) \\ &\overset{\text{(ii)}}{\le}(1+\delta_{2k})\sigma_1\left(\|(\overline{U}^{(t)} - \overline{U}^*)^\top(\overline{U}^{(t)} - \overline{U}^*)\|_{\mathrm{F}} + \frac{3\delta_{2k}k}{1-\delta_{2k}}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}\right) \\ &\overset{\text{(iii)}}{\le}(1+\delta_{2k})\sigma_1\left(\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}^2 + \frac{3\delta_{2k}k}{1-\delta_{2k}}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}\right) \overset{\text{(iv)}}{\le} \frac{(1-\delta_{2k})\sigma_k}{2\xi}\|\overline{U}^* - \overline{U}^{(t)}\|_{\mathrm{F}},\end{aligned}$$

where (i) comes from Lemma C.1.1 and $\|V^*\|_2 = \|M^*\| = \sigma_1$ and $\|V^*\|_{\mathrm{F}} = \|v^*\|_2 \le \sigma_1\sqrt{k}$, (ii) comes from Lemmas C.1.1 and C.1.2, (iii) from Cauchy-Schwartz inequality, and (iv) comes from (4.4.2). Since we have $\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t+0.5)}) = \mathbf{0}$, we further btain

$$\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}) \le \frac{(1-\delta_{2k})\sigma_k}{2\xi}\|\overline{U}^* - \overline{U}^{(t)}\|_{\mathrm{F}},$$

which completes the proof. □

C.1.3 Proof of Lemma 4.4.4

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^*$ and V^* respectively. By the strong convexity of $\mathcal{F}(\overline{U}^*, \cdot)$, we have

$$\mathcal{F}(\overline{U}^*, V^*) - \frac{1-\delta_{2k}}{2}\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2 \geq \mathcal{F}(\overline{U}^*, V^{(t+0.5)}) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t+0.5)}), V^* - V^{(t+0.5)} \rangle. \quad \text{(C.1.4)}$$

By the strong convexity of $\mathcal{F}(\overline{U}^*, \cdot)$ again, we have

$$\begin{aligned}\mathcal{F}(\overline{U}^*, V^{(t+0.5)}) &\geq \mathcal{F}(\overline{U}^*, V^*) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^*), V^* - V^{(t+0.5)} \rangle + \frac{1-\delta_{2k}}{2}\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2 \\ &\geq \mathcal{F}(\overline{U}^*, V^*) + \frac{1-\delta_{2k}}{2}\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2, \end{aligned} \quad \text{(C.1.5)}$$

where the last inequality comes from the optimality condition of $V^* = \operatorname{argmin}_V \mathcal{F}(\overline{U}^*, V)$, i.e.

$$\langle \nabla_V \mathcal{F}(\overline{U}^*, V^*), V^{(t+0.5)} - V^* \rangle \geq 0.$$

Meanwhile, since $V^{(t+0.5)}$ minimizes $\mathcal{F}(\overline{U}^{(t)}, \cdot)$, we have the optimality condition

$$\langle \nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t+0.5)}), V^* - V^{(t+0.5)} \rangle \geq 0,$$

which further implies

$$\begin{aligned} &\langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t+0.5)}), V^* - V^{(t+0.5)} \rangle \\ &\qquad \geq \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t+0.5)}) - \nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t+0.5)}), V^* - V^{(t+0.5)} \rangle. \end{aligned} \tag{C.1.6}$$

Combining (C.1.4) and (C.1.5) with (C.1.6), we obtain

$$\|V^{(t+0.5)} - V^*\|_2 \leq \frac{1}{1-\delta_{2k}} \mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}),$$

which completes the proof. □

C.1.4 Proof of Lemma 4.4.5

Proof. Before we proceed with the proof, we first introduce the following lemma.

Lemma C.1.4. Suppose that $A^* \in \mathbb{R}^{n\times k}$ is a rank k matrix. Let $E \in \mathbb{R}^{n\times k}$ satisfy $\|E\|_2 \|A^{*\dagger}\|_2 < 1$. Then given a QR decomposition $(A^* + E) = QR$, there exists a fac-

torization of $A^* = Q^*O^*$ such that $Q^* \in \mathbb{R}^{n\times k}$ is an orthonormal matrix, and satisfies

$$\|Q - Q^*\|_{\mathrm{F}} \leq \frac{\sqrt{2}\|A^{*\dagger}\|_2\|E\|_{\mathrm{F}}}{1 - \|E\|_2\|A^{*\dagger}\|_2}.$$

The proof of Lemma C.1.4 is provided in [127], therefore omitted.

We then proceed with the proof of Lemma 4.4.5. We consider $A^* = V^{*(t)}$ and $E = V^{(t+0.5)} - V^{*(t)}$ in Lemma C.1.4 respectively. We can verify that

$$\|V^{(t+0.5)} - V^{*(t)}\|_2\|V^{*(t)\dagger}\|_2 \leq \frac{\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}}{\sigma_k} \leq \frac{1}{4}.$$

Then there exists a $V^{*(t)} = \overline{V}^{*(t+1)}O^*$ such that $\overline{V}^{*(t+1)}$ is an orthonormal matrix, and satisfies

$$\|\overline{V}^{*(t+0.5)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq 2\|V^{*(t)\dagger}\|_2\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{2}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}.$$

□

C.1.5 Proof of Lemma 4.4.6

Proof. Before we proceed with the main proof, we first introduce the following lemma.

Lemma C.1.5. Let $b = \mathcal{A}(M^*)$, M is a rank-k matrix, and $\mathcal{A}$ is a linear measurement

operator that satisfies $2k$-RIP with constant $\delta_{2k} < 1/3$. Let $X^{(t+1)}$ be the $(t+1)$-th step iterate of SVP, then we have

$$\|\mathcal{A}(X^{(t+1)}) - b\|_2^2 \leq \|\mathcal{A}(M^*) - b\|_2^2 + 2\delta_{2k}\|\mathcal{A}(X^{(t)}) - b\|_2^2$$

The proof of Lemma C.1.5 is provided in [79], therefore omitted. We then explain the implication of Lemma C.1.5. [79] show that $X^{(t+1)}$ is obtained by taking a projected gradient iteration over $X^{(t)}$ using step size $\frac{1}{1+\delta_{2k}}$. Then taking $X^{(t)} = 0$, we have

$$X^{(t+1)} = \frac{\overline{U}^{(0)}\overline{\Sigma}^{(0)}\overline{V}^{(0)\top}}{1+\delta_{2k}}.$$

Then Lemma C.1.5 implies

$$\left\|\mathcal{A}\left(\frac{\overline{U}^{(0)}\overline{\Sigma}^{(0)}\overline{V}^{(0)\top}}{1+\delta_{2k}} - M^*\right)\right\|_2^2 \leq 4\delta_{2k}\|\mathcal{A}(M^*)\|_2^2. \tag{C.1.7}$$

Since $\mathcal{A}(\cdot)$ satisfies $2k$-RIP, then (C.1.7) further implies

$$\left\|\frac{\overline{U}^{(0)}\overline{\Sigma}^{(0)}\overline{V}^{(0)\top}}{1+\delta_{2k}} - M^*\right\|_{\mathrm{F}}^2 \leq 4\delta_{2k}(1+3\delta_{2k})\|M^*\|_{\mathrm{F}}^2. \tag{C.1.8}$$

We then project each column of M^* into the subspace spanned by $\{\overline{U}_{*i}^{(0)}\}_{i=1}^k$, and

obtain

$$\|\overline{U}^{(0)}\overline{U}^{(0)\top}M^* - M^*\|_{\mathrm{F}}^2 \le 6\delta_{2k}\|M^*\|_{\mathrm{F}}^2.$$

Let $\overline{U}_{\perp}^{(0)}$ denote the orthonormal complement of $\overline{U}^{(0)}$, i.e.,

$$\overline{U}_{\perp}^{(0)\top}\overline{U}_{\perp}^{(0)} = I_{n-k} \quad \text{and} \quad \overline{U}_{\perp}^{(0)\top}\overline{U}^{(0)} = 0.$$

Then given a compact singular value decomposition of $M^* = \widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top}$, we have

$$\frac{6\delta_{2k}k\sigma_1^2}{\sigma_k^2} \ge \|(\overline{U}^{(0)}\overline{U}^{(0)\top} - I_n)\widetilde{U}^*\|_{\mathrm{F}}^2 = \|\overline{U}_{\perp}^{(0)\top}\widetilde{U}^*\|_{\mathrm{F}}^2.$$

Thus Lemma 4.4.2 guarantees that for $O^* = \mathrm{argmin}_{O^\top O = I_k}\|\overline{U}^{(0)} - \widetilde{U}^*O\|_{\mathrm{F}}$, we have

$$\|\overline{U}^{(0)} - \widetilde{U}^*O^*\|_{\mathrm{F}} \le \sqrt{2}\|\overline{U}_{\perp}^{(0)\top}\widetilde{U}^*\|_{\mathrm{F}} \le 2\sqrt{3\delta_{2k}k}\cdot\frac{\sigma_1}{\sigma_k}.$$

We define $\overline{U}^{*(0)} = \widetilde{U}^*O^*$. Then combining the above inequality with (4.4.4), we have

$$\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \le \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}.$$

Meanwhile, we define $V^{*(0)} = \widetilde{V}^*\widetilde{D}^*O^*$. Then we have $\overline{U}^{*(0)}V^{*(0)\top} = \widetilde{U}^*OO^{*\top}\widetilde{D}^*\widetilde{V}^* = M^*$. □

C.1.6 Proof of Corollary 4.4.7

Proof. Since (4.4.5) ensures that (4.4.2) of Lemma 4.4.3 holds, then we have

$$\begin{aligned}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} &\leq \frac{1}{1-\delta_{2k}}\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}) \overset{\text{(i)}}{\leq} \frac{1}{1-\delta_{2k}} \cdot \frac{(1-\delta_{2k})\sigma_k}{2\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\overset{\text{(ii)}}{\leq} \frac{1}{1-\delta_{2k}} \cdot \frac{(1-\delta_{2k})\sigma_k}{2\xi} \cdot \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1} \\ &\leq \left(\frac{(1-\delta_{2k})\sigma_k}{8\xi^2(1+\delta_{2k})\sigma_1}\right)\sigma_k \overset{\text{(iii)}}{\leq} \frac{\sigma_k}{4}, \end{aligned} \tag{C.1.9}$$

where (i) comes from Lemma 4.4.4, (ii) comes from (4.4.5), and (iii) comes from the definition of ξ and $\sigma_k \leq \sigma_1$. Since (C.1.9) ensures that (4.4.3) of Lemma 4.4.5 holds for $V^{(t+0.5)}$, then we obtain

$$\begin{aligned}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} &\leq \frac{2}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \overset{\text{(i)}}{\leq} \frac{1}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\overset{\text{(ii)}}{\leq} \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}, \end{aligned} \tag{C.1.10}$$

where (i) comes from (C.1.9), and (ii) comes from the definition of ξ and (4.4.5).

□

C.1.7 Proof of Lemma C.1.1

Proof. We consider an arbitrary $W \in \mathbb{R}^{n\times k}$ such that $\|W\|_{\mathrm{F}} = 1$. Let $w = \mathrm{vec}(W)$. Then we have

$$\begin{aligned} w^\top B w &= \sum_{p,q=1}^{k} W_{*p}^\top S_{pq}^{(t)} W_{*p} = \sum_{p,q=1}^{k} W_{*p}^\top \Bigg(\sum_{i=1}^{d} A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{(t)\top} A_i^\top \Bigg) W_{*q} \\ &= \sum_{i=1}^{d} \Bigg(\sum_{p=1}^{k} W_{*p}^\top A_i \overline{U}_{*p}^{(t)} \Bigg) \Bigg(\sum_{q=1}^{k} W_{*q}^\top A_i \overline{U}_{*q}^{(t)} \Bigg) = \sum_{i=1}^{n} \mathrm{tr}(W^\top A_i \overline{U}^{(t)})^2 = \|\mathcal{A}(\overline{U}^{(t)} W^\top)\|_2^2. \end{aligned}$$

Since $\mathcal{A}(\cdot)$ satisfies $2k$-RIP, then we have

$$\begin{aligned} \|\mathcal{A}(\overline{U}^{(t)} W^\top)\|_2^2 &\geq (1-\delta_{2k}) \|\overline{U}^{(t)} W^\top\|_{\mathrm{F}} = (1-\delta_{2k}) \|W\|_{\mathrm{F}} = 1-\delta_{2k}, \\ \|\mathcal{A}(\overline{U}^{(t)} W^\top)\|_2^2 &\leq (1+\delta_{2k}) \|\overline{U}^{(t)} W^\top\|_{\mathrm{F}} = (1+\delta_{2k}) \|W\|_{\mathrm{F}} = 1+\delta_{2k}. \end{aligned}$$

Since W is arbitrary, then we have

$$\sigma_{\min}(S^{(t)}) = \min_{\|w\|_2=1} w^\top S^{(t)} w \geq 1-\delta_{2k} \quad \text{and} \quad \sigma_{\max}(S^{(t)}) = \max_{\|w\|_2=1} w^\top S^{(t)} w \leq 1+\delta_{2k}.$$

□

C.1.8 Proof of Lemma C.1.2

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^*$ and V^* respectively. Before we proceed with the main proof, we

first introduce the following lemma.

Lemma C.1.6. Suppose $\mathcal{A}(\cdot)$ satisfies $2k$-RIP. For any $U,\ U' \in \mathbb{R}^{m\times k}$ and $V,\ V' \in \mathbb{R}^{n\times k}$, we have

$$|\langle \mathcal{A}(UV^\top), \mathcal{A}(U'V'^\top)\rangle - \langle U^\top U', V^\top V'\rangle| \le 3\delta_{2k}\|UV^\top\|_{\mathrm{F}} \cdot \|U'V'^\top\|_{\mathrm{F}}.$$

The proof of Lemma C.1.6 is provided in [94], and hence omitted.

We now proceed with the proof of Lemma C.1.2. We consider arbitrary $W, Z \in \mathbb{R}^{n\times k}$ such that $\|W\|_{\mathrm{F}} = \|Z\|_{\mathrm{F}} = 1$. Let $w = \mathrm{vec}(W)$ and $z = \mathrm{vec}(Z)$. Then we have

$$w^\top (S^{(t)}K^{(t)} - J^{(t)})z = \sum_{p,q=1}^{k} W_{*p}^\top [S^{(t)}K^{(t)} - J^{(t)}]_{pq} Z_{*q}.$$

We consider a decomposition

$$\begin{aligned}
[S^{(t)}K^{(t)} - J^{(t)}]_{pq} &= \sum_{\ell=1}^{k} S_{p\ell}^{(t)} K_{\ell q}^{(t)} - J_{pq}^{(t)} = \sum_{\ell=1}^{k} S_{p\ell}^{(t)} \overline{U}_{*\ell}^{(t)\top} \overline{U}_{*q}^{*} I_n - J_{pq}^{(t)} \\
&= \sum_{\ell=1}^{k} \overline{U}_{*q}^{*\top} \overline{U}_{*\ell}^{(t)} \sum_{i=1}^{d} A_i \overline{U}_{*p}^{(t)} \overline{U}_{*\ell}^{(t)} A_i^\top - J_{pq}^{(t)} \\
&= \sum_{\ell=1}^{k} A_i \overline{U}_{*q}^{*\top} \overline{U}_{*\ell}^{(t)} \sum_{i=1}^{d} \overline{U}_{*p}^{(t)} \overline{U}_{*\ell}^{(t)} A_i^\top - \sum_{i=1}^{d} A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{*} A_i^\top \\
&= \sum_{i=1}^{d} A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{*} (\overline{U}^{(t)} \overline{U}^{(t)\top} - I_n) A_i^\top.
\end{aligned}$$

which further implies

$$\begin{aligned} w^\top (S^{(t)}K^{(t)} - J^{(t)})z &= \sum_{p,q} W_{*p}^\top \Bigg(\sum_{i=1}^{d} A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{*} (\overline{U}^{(t)} \overline{U}^{(t)\top} - I_m) A_i^\top \Bigg) Z_{*q} \\ &= \sum_{i=1}^{d} \sum_{p,q} W_{*p}^\top A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{*} (\overline{U}^{(t)} \overline{U}^{(t)\top} - I_m) A_i^\top Z_{*q} \\ &= \sum_{i=1}^{d} \operatorname{tr}(W^\top A_i \overline{U}^{(t)}) \operatorname{tr}\Big(Z^\top A_i (\overline{U}^{(t)} \overline{U}^{(t)\top} - I_m) \overline{U}^{*} \Big). \end{aligned} \tag{C.1.11}$$

Since $\mathcal{A}(\cdot)$ satisfies $2k$-RIP, then by Lemma C.1.6, we obtain

$$\begin{aligned} w^\top (S^{(t)}K^{(t)} - J^{(t)})z &\le \operatorname{tr}\Big(\overline{U}^{*} (\overline{U}^{(t)} \overline{U}^{(t)\top} - I_m) \overline{U}^{(t)} W^\top Z \Big) \\ &\quad + 3\delta_{2k} \|\overline{U}^{(t)} W^\top\|_{\mathrm{F}} \|(\overline{U}^{(t)} \overline{U}^{(t)\top} - I_m) \overline{U}^{*} Z^\top\|_{\mathrm{F}} \\ &\overset{\text{(i)}}{\le} 3\delta_{2k} \|W\|_{\mathrm{F}} \sqrt{\|\overline{U}^{*\top} (\overline{U}^{(t)} \overline{U}^{(t)\top} - I_m) \overline{U}^{*}\|_{\mathrm{F}} \|Z^\top Z\|_{\mathrm{F}}}, \end{aligned} \tag{C.1.12}$$

where the last inequality comes from $(\overline{U}^{(t)} \overline{U}^{(t)\top} - I_m) \overline{U}^{(t)} = 0$. Let $\overline{U}_{\perp}^{(t)} \in \mathbb{R}^{m-k}$ denote the orthogonal complement to $\overline{U}^{(t)}$ such that $\overline{U}^{(t)\top} \overline{U}_{\perp}^{(t)} = 0$ and $\overline{U}_{\perp}^{(t)\top} \overline{U}_{\perp}^{(t)} = I_{m-k}$. Then we have

$$I_m - \overline{U}^{(t)} \overline{U}^{(t)\top} = \overline{U}_{\perp}^{(t)} \overline{U}_{\perp}^{(t)\top},$$

which implies

$$\sqrt{\|\overline{U}^{*\top}(\overline{U}^{(t)}\overline{U}^{(t)\top} - I_m)\overline{U}^{*}\|_{\mathrm{F}}} = \sqrt{\|\overline{U}^{*\top}\overline{U}_{\perp}^{(t)}\overline{U}_{\perp}^{(t)\top}\overline{U}^{*}\|_{\mathrm{F}}} \leq \|\overline{U}_{\perp}^{(t)\top}\overline{U}^{*}\|_{\mathrm{F}}$$
$$= \|\overline{U}_{\perp}^{(t)\top}\overline{U}^{(t)} - \overline{U}_{\perp}^{(t)\top}\overline{U}^{*}\|_{\mathrm{F}} \leq \|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}. \qquad \text{(C.1.13)}$$

Combining (C.1.12) with (C.1.13), we obtain

$$w^{\top}(S^{(t)}K^{(t)} - J^{(t)})z \leq 3\delta_{2k}\sqrt{k}\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}. \qquad \text{(C.1.14)}$$

Since W and Z are arbitrary, then (C.1.14) implies

$$\sigma_{\max}(S^{(t)}K^{(t)} - J^{(t)}) = \max_{\|w\|_2=1, \|z\|_2=1} w^{\top}(S^{(t)}K^{(t)} - J^{(t)})w \leq 3\delta_{2k}\sqrt{k}\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}},$$

which completes the proof.

□

C.2 Lemmas for Theorem 4.3.3 (Alternating Gradient Descent)

C.2.1 Proof of Lemma 4.4.9

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^{*}$ and V^{*} respectively. We have

$$\operatorname{vec}(\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)})) = S^{(t)} v^{(t)} - J^{(t)} v^{*} \quad \text{and} \quad \operatorname{vec}(\nabla_V \mathcal{F}(\overline{U}^{*}, V^{(t)})) = G^{(t)} v^{(t)} - G^{(t)} v^{*}.$$

Therefore, we further obtain

$$\begin{aligned}
&\|\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)}) - \nabla_V \mathcal{F}(\overline{U}^{*}, V^{(t)})\|_{\mathrm{F}} \\
&\quad = \|(S^{(t)} - J^{(t)})(v^{(t)} - v^{*}) + (S^{(t)} - J^{(t)}) v^{*} + (J^{(t)} - G^{(t)})(v^{(t)} - v^{*})\|_2 \\
&\quad \le \|(S^{(t)} - J^{(t)})(v^{(t)} - v^{*})\|_2 + \|(S^{(t)} - J^{(t)}) v^{*}\|_2 + \|(J^{(t)} - G^{(t)})(v^{(t)} - v^{*})\|_2 \\
&\quad \le \|S^{(t)}\|_2 \cdot \|((S^{(t)})^{-1} J^{(t)} - I_{nk})(v^{(t)} - v^{*})\|_2 + \|S^{(t)}\|_2 \cdot \|((S^{(t)})^{-1} J^{(t)} - I_{nk}) v^{*}\|_2 \\
&\qquad\qquad + \|G\|_2 \cdot \|((G^{(t)})^{-1} J^{(t)} - I_{nk})(v^{(t)} - v^{*})\|_2. \quad \text{(C.2.1)}
\end{aligned}$$

Recall that Lemma C.1.2 is also applicable to $G^{(t)} K^{(t)} - J^{(t)}$. Since we have

$$\|V^{(t)} - V^{*}\|_2 \le \|V^{(t)} - V^{*}\|_{\mathrm{F}} = \|v^{(t)} - v^{*}\|_2 \le \sigma_1,$$

following similar lines to Appendix C.1.2, we can show

$$\|((S^{(t)})^{-1}J^{(t)} - I_{mn})v^*\|_2 \le \sigma_1\left(\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}^2 + \frac{3\delta_{2k}k}{1-\delta_{2k}}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}\right),$$
$$\|((G^{(t)})^{-1}J^{(t)} - I_{mn})(v^{(t)} - v^*)\|_2 \le \sigma_1\left(\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}^2 + \frac{3\delta_{2k}k}{1-\delta_{2k}}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}\right),$$
$$\|((S^{(t)})^{-1}J^{(t)} - I_{mn})(v^{(t)} - v^*)\|_2 \le \sigma_1\left(\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}^2 + \frac{3\delta_{2k}k}{1-\delta_{2k}}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}\right).$$

Combining the above three inequalities with (C.2.1), we have

$$\begin{aligned}&\|\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)}) - \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)})\|_{\mathrm{F}} \\ &\qquad \le 2(1+\delta_{2k})\sigma_1\left(\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}^2 + \frac{3\delta_{2k}k}{1-\delta_{2k}}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}\right). \qquad \text{(C.2.2)}\end{aligned}$$

Since $\overline{U}^{(t)}$, δ_{2k}, and ξ satisfy (4.4.13), then (C.2.2) further implies

$$\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) = \|\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)}) - \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)})\|_{\mathrm{F}} \le \frac{(1+\delta_{2k})\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}},$$

which completes the proof. □

C.2.2 Proof of Lemma 4.4.10

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^*$ and V^* respectively. By the strong convexity of $\mathcal{F}(\overline{U}^*, \cdot)$, we have

$$\begin{aligned}\mathcal{F}(\overline{U}^*, V^*) - \frac{1-\delta_{2k}}{2}\|V^{(t)} - V^*\|_{\mathrm{F}}^2 &\geq \mathcal{F}(\overline{U}^*, V^{(t)}) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)}), V^* - V^{(t)} \rangle \\ &= \mathcal{F}(\overline{U}^*, V^{(t)}) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)}), V^{(t+0.5)} - V^{(t)} \rangle \\ &\quad + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)}), V^* - V^{(t+0.5)} \rangle. \end{aligned} \tag{C.2.3}$$

Meanwhile, we define

$$\mathcal{Q}(V; \overline{U}^*, V^{(t)}) = \mathcal{F}(\overline{U}^*, V^{(t)}) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)}), V - V^{(t)} \rangle + \frac{1}{2\eta}\|V - V^{(t)}\|_{\mathrm{F}}^2.$$

Since η satisfies (4.4.14) and $\mathcal{F}(\overline{U}^*, V)$ is strongly smooth in V for a fixed orthonormal $\overline{U}^*$, we have

$$\mathcal{Q}(V; \overline{U}^*, V^{(t)}) \geq \mathcal{F}(\overline{U}^*, V^{(t)}).$$

Combining the above two inequalities, we obtain

$$\begin{aligned}\mathcal{F}(\overline{U}^*, V^{(t)}) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)}), V^{(t+0.5)} - V^{(t)} \rangle &= \mathcal{Q}(V^{(t+0.5)}; \overline{U}^*, V^{(t)}) - \frac{1}{2\eta}\|V^{(t+0.5)} - V^{(t)}\|_{\mathrm{F}}^2 \\ &\geq \mathcal{F}(\overline{U}^*, V^{(t+0.5)}) - \frac{1}{2\eta}\|V^{(t+0.5)} - V^{(t)}\|_{\mathrm{F}}^2. \end{aligned} \tag{C.2.4}$$

Moreover, by the strong convexity of $\mathcal{F}(\overline{U}^*, \cdot)$ again, we have

$$\begin{aligned}\mathcal{F}(\overline{U}^*, V^{(t+0.5)}) &\geq \mathcal{F}(\overline{U}^*, V^*) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^*), V^{(t+0.5)} - V^* \rangle + \frac{1-\delta_{2k}}{2} \|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2 \\ &\geq \mathcal{F}(\overline{U}^*, V^*) + \frac{1-\delta_{2k}}{2} \|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2, \end{aligned} \tag{C.2.5}$$

where the second equalities comes from the optimality condition of $V^* = \operatorname{argmin}_V \mathcal{F}(\overline{U}^*, V)$, i.e.

$$\langle \nabla_V \mathcal{F}(\overline{U}^*, V^*), V^{(t+0.5)} - V^* \rangle \geq 0.$$

Combining (C.2.3) and (C.2.4) with (C.2.5), we obtain

$$\begin{aligned}&\mathcal{F}(\overline{U}^*, V^{(t)}) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)}), V^{(t+0.5)} - V^{(t)} \rangle \\ &\qquad \geq \mathcal{F}(\overline{U}^*, V^*) + \frac{1-\delta_{2k}}{2} \|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2 - \frac{1}{2\eta} \|V^{(t+0.5)} - V^{(t)}\|_{\mathrm{F}}^2. \end{aligned} \tag{C.2.6}$$

On the other hand, since $V^{(t+0.5)}$ minimizes $\mathcal{Q}(V; \overline{U}^*, V^{(t)})$, we have

$$\begin{aligned}0 &\leq \langle \nabla \mathcal{Q}(V^{(t+0.5)}; \overline{U}^*, V^{(t)}), V^* - V^{(t+0.5)} \rangle \\ &\leq \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)}), V^* - V^{(t+0.5)} \rangle + (1+\delta_{2k}) \langle V^{(t+0.5)} - V^{(t)}, V^* - V^{(t+0.5)} \rangle. \end{aligned} \tag{C.2.7}$$

Meanwhile, we have

$$\begin{aligned}
&\langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)}), V^* - V^{(t+0.5)} \rangle \\
&\quad = \langle \nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)}), V^* - V^{(t+0.5)} \rangle - \mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \|V^* - V^{(t+0.5)}\|_2 \\
&\quad \geq (1+\delta_{2k}) \langle V^{(t)} - V^{(t+0.5)}, V^* - V^{(t+0.5)} \rangle - \mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \|V^* - V^{(t+0.5)}\|_2 \\
&\quad = (1+\delta_{2k}) \langle V^{(t)} - V^{(t+0.5)}, V^* - V^{(t)} \rangle + \frac{1}{2\eta} \|V^{(t)} - V^{(t+0.5)}\|_{\mathrm{F}}^2 \\
&\qquad\qquad - \mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \|V^* - V^{(t+0.5)}\|_2. \qquad \text{(C.2.8)}
\end{aligned}$$

Combining (C.2.7) with (C.2.8), we obtain

$$\begin{aligned}
2\langle V^{(t)} - V^{(t+0.5)}, V^* - V^{(t)} \rangle &\leq -\eta(1-\delta_{2k}) \|V^{(t)} - V^*\|_2^2 - \eta(1-\delta_{2k}) \|V^{(t+0.5)} - V^*\|_2^2 \\
&\quad - \|V^{(t+0.5)} - V^{(t)}\|_2^2 + \mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \|V^* - V^{(t+0.5)}\|_2. \qquad \text{(C.2.9)}
\end{aligned}$$

Therefore, combining (C.2.6) with (C.2.9), we obtain

$$\begin{aligned}
\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2 &\leq \|V^{(t+0.5)} - V^{(t)} + V^{(t)} - V^*\|_{\mathrm{F}}^2 \\
&= \|V^{(t+0.5)} - V^{(t)}\|_{\mathrm{F}}^2 + \|V^{(t)} - V^*\|_{\mathrm{F}}^2 + 2\langle V^{(t+0.5)} - V^{(t)}, V^{(t)} - V^* \rangle \\
&\leq 2\eta \|V^{(t)} - V^*\|_{\mathrm{F}}^2 - \eta(1-\delta_{2k}) \|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2 \\
&\qquad - \mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \|V^* - V^{(t+0.5)}\|_2.
\end{aligned}$$

Rearranging the above inequality, we obtain

$$\|V^{(t+0.5)} - V^*\|_{\mathrm{F}} \leq \sqrt{\delta_{2k}}\|V^{(t)} - V^*\|_{\mathrm{F}} + \frac{2}{1+\delta_{2k}}\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}),$$

which completes the proof. □

C.2.3 Proof of Lemma 4.4.11

Proof. Before we proceed with the main proof, we first introduce the following lemma.

Lemma C.2.1. For any matrix $U, \widetilde{U} \in \mathbb{R}^{m\times k}$ and $V, \widetilde{V} \in \mathbb{R}^{n\times k}$, we have

$$\|UV^\top - \widetilde{U}\widetilde{V}^\top\|_{\mathrm{F}} \leq \|U\|_2\|V - \widetilde{V}\| + \|\widetilde{V}\|_2\|U - \widetilde{U}\|_{\mathrm{F}}.$$

Proof. By linear algebra, we have

$$\begin{aligned}\|UV^\top - \widetilde{U}\widetilde{V}^\top\|_{\mathrm{F}} &= \|UV^\top - U\widetilde{V}^\top + U\widetilde{V}^\top - \widetilde{U}\widetilde{V}^\top\|_{\mathrm{F}} \\ &\leq \|UV^\top - U\widetilde{V}^\top\|_{\mathrm{F}} + \|U\widetilde{V}^\top - \widetilde{U}\widetilde{V}^\top\|_{\mathrm{F}} \\ &\leq \|U\|_2\|V - \widetilde{V}\|_{\mathrm{F}} + \|\widetilde{V}\|_2\|U - \widetilde{U}\|_{\mathrm{F}}.\end{aligned}$$

□

We then proceed with the proof of Lemma 4.4.11. By Lemma C.2.1, we have

$$\begin{aligned}\|R_{\overline{V}}^{(t+0.5)} - \overline{V}^{*(t+1)\top} V^{*(t)}\|_{\mathrm{F}} &= \|\overline{V}^{(t+0.5)\top} V^{(t+0.5)} - \overline{V}^{*(t+1)\top} V^{*(t)}\|_{\mathrm{F}} \\ &\le \|\overline{V}^{(t+0.5)}\|_2 \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \|V^{*(t)}\|_2 \|\overline{V}^{(t+0.5)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \\ &\le \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2\sigma_1}{\sigma_k} \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}, \end{aligned} \tag{C.2.10}$$

where the last inequality comes from Lemma 4.4.5. Moreover, we define $U^{*(t+1)} = \overline{U}^{*(t)} (\overline{V}^{*(t+1)\top} V^{*(t)})^\top$. Then we can verify

$$U^{*(t+1)} \overline{V}^{*(t+1)} = \overline{U}^{*(t)} V^{*(t)\top} \overline{V}^{*(t+1)} \overline{V}^{*(t+1)\top} = M^* \overline{V}^{*(t+1)} \overline{V}^{*(t+1)\top} = M^*,$$

where the last equality holds, since $\overline{V}^{*(t+1)} \overline{V}^{*(t+1)\top}$ is exactly the projection matrix for the row space of M^*. Thus by Lemma C.2.1, we have

$$\begin{aligned}\|U^{(t+1)} - U^{*(t+1)}\|_{\mathrm{F}} &= \|\overline{U}^{(t)} R_{\overline{V}}^{(t+0.5)\top} - \overline{U}^{*(t)} (\overline{V}^{*(t+1)\top} V^{*(t)})^\top\|_{\mathrm{F}} \\ &\le \|\overline{U}^{(t)}\|_2 \|R_{\overline{V}}^{(t+0.5)} - \overline{V}^{*(t+1)\top} V^{*(t)}\|_{\mathrm{F}} + \|\overline{V}^{*(t+1)\top} V^{*(t)}\|_2 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\le \left(1 + \frac{2\sigma_1}{\sigma_k}\right) \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \end{aligned}$$

where the last inequality comes from (C.2.10), $\|\overline{V}^{*(t+1)}\|_2 = 1$, $\|\overline{U}^{(t)}\|_2 = 1$, and $\|V^{*(t)}\|_2 = \sigma_1$. □

C.2.4 Proof of Lemma 4.4.12

Proof. Following similar lines to Appendix C.1.5, we have

$$\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{4\xi\sigma_1^2}. \tag{C.2.11}$$

In Appendix C.1.5, we have already shown

$$\left\|\frac{\overline{U}^{(0)}\overline{\Sigma}^{(0)}\overline{V}^{(0)\top}}{1+\delta_{2k}} - M^*\right\|_{\mathrm{F}} \leq 2\sqrt{\delta_{2k}(1+3\delta_{2k})}\|\overline{\Sigma}^*\|_{\mathrm{F}}. \tag{C.2.12}$$

Then by Lemma C.2.1 we have

$$\begin{aligned}
\left\|\frac{\overline{U}^{(0)}\overline{\Sigma}^{(0)}}{1+\delta_{2k}} - V^{*(0)}\right\|_{\mathrm{F}} &= \left\|\frac{\overline{U}^{(0)\top}\overline{U}^{(0)}\overline{\Sigma}^{(0)}\overline{V}^{(0)\top}}{1+\delta_{2k}} - \overline{U}^{*(0)\top}M^*\right\|_{\mathrm{F}} \\
&\leq \|\overline{U}^{(0)}\|_2\left\|\frac{\overline{U}^{(0)}\overline{\Sigma}^{(0)}\overline{V}^{(0)\top}}{1+\delta_{2k}} - M^*\right\|_{\mathrm{F}} + \|M^*\|_2\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \\
&\leq 2\sqrt{\delta_{2k}k(1+3\delta_{2k})}\sigma_1 + \frac{\sigma_k^2}{4\xi\sigma_1^1},
\end{aligned} \tag{C.2.13}$$

where the last inequality comes from (C.2.11), (C.2.12), $\|M^*\|_2 = \sigma_1$, and $\|\overline{U}^{(0)}\|_2 = 1$. By triangle inequality, we further have

$$\begin{aligned}\|\overline{U}^{(0)}\overline{\Sigma}^{(0)} - V^{*(0)}\|_{\mathrm{F}} &\leq (1+\delta_{2k})\left\|\frac{\overline{U}^{(0)}\overline{\Sigma}^{(0)}}{1+\delta_{2k}} - V^{*(0)}\right\|_{\mathrm{F}} + \delta_{2k}\|V^{*(0)}\|_{\mathrm{F}} \\ &\overset{\text{(i)}}{\leq}(1+\delta_{2k})\left(2\sqrt{\delta_{2k}k(1+3\delta_{2k})}\sigma_1 + \frac{\sigma_k^2}{4\xi\sigma_1}\right) + \delta_{2k}\sigma_1\sqrt{k} \\ &\overset{\text{(ii)}}{\leq}\left(\frac{\sigma_k^3}{9\sigma_1^3\xi} + \frac{\sigma_k^2}{3\sigma_1^3\xi^2} + \frac{\sigma_k^3}{192\xi^3\sigma_1^2}\right)\sigma_1 \overset{\text{(iii)}}{\leq} \frac{\sigma_k^2}{2\xi\sigma_1},\end{aligned}$$

where (i) comes from (C.2.13) and $\|V^{*(0)}\|_{\mathrm{F}} = \|M^*\|_{\mathrm{F}} \leq \sigma_1\sqrt{k}$, (ii) comes from (4.4.16), and (iii) comes from the definition of ξ and $\sigma_1 \geq \sigma_k$. □

C.2.5 Proof of Corollary 4.4.13

Proof. Since (4.4.17) ensures that (4.4.13) of Lemma 4.4.9 holds, we have

$$\begin{aligned}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} &\leq \sqrt{\delta_{2k}}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2}{1+\delta_{2k}}\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \\ &\overset{\text{(i)}}{\leq}\sqrt{\delta_{2k}}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2}{1+\delta_{2k}}\cdot\frac{(1+\delta_{2k})\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\overset{\text{(ii)}}{\leq}\frac{\sigma_k^2}{12\xi\sigma_1^2}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\overset{\text{(iii)}}{\leq}\frac{\sigma_k^2}{12\xi\sigma_1^2}\cdot\frac{\sigma_k^2}{2\xi\sigma_1} + \frac{2\sigma_k}{\xi}\cdot\frac{\sigma_k^2}{4\xi\sigma_1^2} \overset{\text{(iv)}}{\leq}\frac{13\sigma_k^3}{24\xi^2\sigma_1^2} \overset{\text{(v)}}{\leq}\frac{\sigma_k}{4},\end{aligned} \tag{C.2.14}$$

where (i) comes from Lemma 4.4.10, (ii) and (iii) come from (4.4.17), and (iv) and (v) come from the definition of ξ and $\sigma_k \leq \sigma_1$. Since (C.2.14) ensures that (4.4.3) of

Lemma 4.4.5, then we obtain

$$\begin{aligned}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{2}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} &\overset{\text{(i)}}{\leq} \frac{2\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{4}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\overset{\text{(ii)}}{\leq} \left(\frac{\sigma_k}{3\xi\sigma_1} + \frac{4}{\xi}\right)\cdot\frac{\sigma_k^2}{4\xi\sigma_1^2} \overset{\text{(iii)}}{\leq} \frac{\sigma_k^2}{4\xi\sigma_1^2}, \end{aligned} \tag{C.2.15}$$

where (i) and (ii) come from (C.2.14), and (iii) comes from the definition of ξ and $\sigma_1 > \sigma_k$. Moreover, since (C.2.14) ensures that (4.4.15) of Lemma 4.4.11 holds, then we have

$$\begin{aligned}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} &\leq \frac{3\sigma_1}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\overset{\text{(i)}}{\leq} \frac{3\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \left(\frac{6}{\xi} + 1\right)\sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\overset{\text{(ii)}}{\leq} \frac{3\sigma_1}{\sigma_k}\cdot\frac{\sigma_k^3}{12\xi\sigma_1^3}\cdot\frac{\sigma_k^2}{2\xi\sigma_1} + \left(\frac{6}{\xi} + 1\right)\cdot\frac{\sigma_k^2}{4\xi\sigma_1} \\ &= \left(\frac{\sigma_k^2}{4\xi^2\sigma_1^2} + \frac{3}{\xi} + \frac{1}{2}\right)\frac{\sigma_k^2}{2\xi\sigma_1} \overset{\text{(iii)}}{\leq} \frac{\sigma_k^2}{2\xi\sigma_1}, \end{aligned}$$

where (i) comes from (C.2.14), (ii) comes from (4.4.17), and (iii) comes from the definition of ξ and $\sigma_1 \geq \sigma_k$. □

C.3 Partition Algorithm for Matrix Completion

Algorithm 10: The observation set partition algorithm for matrix completion. It guarantees the independence among all $2T+1$ output observation sets.

Input: $\mathcal{W}$, $\bar{\rho}$

$\widetilde{\rho} = 1 - (1-\bar{\rho})^{\frac{1}{2T+1}}$

For $t \leftarrow 0, ..., 2T$

$$\widetilde{\rho}_t = \frac{(mn)!\bar{\rho}^{t+1}(1-\bar{\rho})^{mn-t-1}}{\bar{\rho}(mn-t-1)!(t+1)!}$$

$\mathcal{W}_0 = \emptyset$, ..., $\mathcal{W}_{2T} = \emptyset$

For *every* $(i,j) \in \mathcal{W}$

Sample t from $\{0,...,2T\}$ with probability $\{\widetilde{\rho}_0, ..., \widetilde{\rho}_{2T}\}$

Sample (w/o replacement) a set $\mathcal{B}$ such that $|\mathcal{B}| = t$ from $\{0,...,2T\}$ with equal probability

Add (i,j) to $\mathcal{W}_\ell$ for all $\ell \in \mathcal{B}$

Return: $\{\mathcal{W}_t\}_{t=0}^{2T}, \widetilde{\rho}$

C.4 Initialization Procedures for Matrix Completion

C.5 Proof of Theorem 4.5.2

We present the technical proof for matrix completion. Before we proceed with the main proof, we first introduce the following lemma.

Algorithm 11: The initialization procedure $\mathrm{INT}_{\overline{U}}(\cdot)$ for matrix completion. It guarantees that the initial solutions satisfy the incoherence condition throughout all iterations.

Input: $\widetilde{M}$
Parameter: Incoherence parameter μ
$(\widetilde{U}, \widetilde{D}, \widetilde{V}) \leftarrow \mathrm{KSVD}(\widetilde{M})$
$\widetilde{U}^{\mathrm{tmp}} \leftarrow \underset{U}{\operatorname{argmin}} \|U - \widetilde{U}\|_{\mathrm{F}}^2$ subject to $\max_i \|U_{i*}\|_2 \leq \mu\sqrt{k/m}$
$(\overline{U}^{\mathrm{out}}, R_{\overline{U}}^{\mathrm{out}}) \leftarrow \mathrm{QR}(\widetilde{U}^{\mathrm{tmp}})$
$\widetilde{V}^{\mathrm{tmp}} \leftarrow \underset{V}{\operatorname{argmin}} \|V - \widetilde{V}^{\mathrm{tmp}}\|_{\mathrm{F}}^2$ subject to $\max_j \|V_{j*}\|_2 \leq \mu\sqrt{k/n}$
$(\overline{V}^{\mathrm{out}}, R_{\overline{V}}^{\mathrm{out}}) \leftarrow \mathrm{QR}(\widetilde{V}^{\mathrm{tmp}})$
$V^{\mathrm{out}} = \overline{V}^{\mathrm{out}}(\overline{U}^{\mathrm{out}\top}\widetilde{M}\overline{V}^{\mathrm{out}})^{\top}$
Return: $\overline{U}^{\mathrm{out}}$, V^{out}

Algorithm 12: The initialization procedure $\mathrm{INT}_{\overline{V}}(\cdot)$ for matrix completion. It guarantees that the initial solutions satisfy the incoherence condition throughout all iterations.

Input: $\widetilde{M}$
Parameter: Incoherence parameter μ
$(\widetilde{U}, \widetilde{D}, \widetilde{V}) \leftarrow \mathrm{KSVD}(\widetilde{M})$
$\widetilde{V}^{\mathrm{tmp}} \leftarrow \underset{V}{\operatorname{argmin}} \|V - \widetilde{V}\|_{\mathrm{F}}^2$ subject to $\max_j \|V_{j*}\|_2 \leq \mu\sqrt{k/n}$
$(\overline{V}^{\mathrm{out}}, R_{\overline{V}}^{\mathrm{out}}) \leftarrow \mathrm{QR}(\widetilde{V}^{\mathrm{tmp}})$
$\widetilde{U}^{\mathrm{tmp}} \leftarrow \underset{U}{\operatorname{argmin}} \|U - \widetilde{U}^{\mathrm{tmp}}\|_{\mathrm{F}}^2$ subject to $\max_i \|U_{i*}\|_2 \leq \mu\sqrt{k/m}$
$(\overline{U}^{\mathrm{out}}, R_{\overline{U}}^{\mathrm{out}}) \leftarrow \mathrm{QR}(\widetilde{U}^{\mathrm{tmp}})$
$U^{\mathrm{out}} = \overline{U}^{\mathrm{out}}(\overline{U}^{\mathrm{out}\top}\widetilde{M}\overline{V}^{\mathrm{out}})$
Return: $\overline{V}^{\mathrm{out}}$, U^{out}

Lemma C.5.1. [[98]] Suppose that the entry observation probability $\bar{\rho}$ of $\mathcal{W}$ satisfies (4.5.3). Then the output sets $\{\mathcal{W}_t\}_{t=0}^{2T}$, of Algorithm 10 are equivalent to $2T+1$ observation sets, which are independently generated with the entry observation

probability

$$\widetilde{\rho} \geq \frac{C_7 \mu^2 k^3 \log n}{m} \tag{C.5.1}$$

for some constant C_7.

See [98] for the proof of Lemma C.5.1. Lemma C.5.1 ensures the independence among all observation sets generated by Algorithm 10. To make the convergence analysis for matrix completion comparable to that for matrix sensing, we rescale both the objective function $\mathcal{F}_{\mathcal{W}}$ and step size η by the entry observation probability $\widetilde{\rho}$ of each individual set, which is also obtained by Algorithm 10. In particular, we define

$$\widetilde{\mathcal{F}}_{\mathcal{W}}(U, V) = \frac{1}{2\widetilde{\rho}} \|\mathcal{P}_{\mathcal{W}}(UV^\top) - \mathcal{P}_{\mathcal{W}}(M^*)\|_{\mathrm{F}}^2 \quad \text{and} \quad \widetilde{\eta} = \widetilde{\rho}\eta. \tag{C.5.2}$$

For notational simplicity, we assume that at the t-th iteration, there exists a matrix factorization of M^* as

$$M^* = \overline{U}^{*(t)} V^{*(t)\top},$$

where $\overline{U}^{*(t)} \in \mathbb{R}^{m \times k}$ is an orthonormal matrix. Then we define several $nk \times nk$ ma-

trices

$$S^{(t)} = \begin{bmatrix} S_{11}^{(t)} & \cdots & S_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ S_{k1}^{(t)} & \cdots & S_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad S_{pq}^{(t)} = \begin{bmatrix} \frac{1}{\widetilde{\rho}} \sum_{i:(i,1)\in\mathcal{W}_{2t+1}} \overline{U}_{ip}^{(t)} \overline{U}_{iq}^{(t)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\widetilde{\rho}} \sum_{i:(i,n)\in\mathcal{W}_{2t+1}} \overline{U}_{ip}^{(t)} \overline{U}_{iq}^{(t)} \end{bmatrix},$$

$$G^{(t)} = \begin{bmatrix} G_{11}^{(t)} & \cdots & G_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ G_{k1}^{(t)} & \cdots & G_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad G_{pq}^{(t)} = \begin{bmatrix} \frac{1}{\widetilde{\rho}} \sum_{i:(i,1)\in\mathcal{W}_{2t+1}} \overline{U}_{ip}^{*(t)} \overline{U}_{iq}^{*(t)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\widetilde{\rho}} \sum_{i:(i,n)\in\mathcal{W}_{2t+1}} \overline{U}_{ip}^{*(t)} \overline{U}_{iq}^{*(t)} \end{bmatrix},$$

$$J^{(t)} = \begin{bmatrix} J_{11}^{(t)} & \cdots & J_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ J_{k1}^{(t)} & \cdots & J_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad J_{pq}^{(t)} = \begin{bmatrix} \frac{1}{\widetilde{\rho}} \sum_{i:(i,1)\in\mathcal{W}_{2t+1}} \overline{U}_{ip}^{(t)} \overline{U}_{iq}^{*(t)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\widetilde{\rho}} \sum_{i:(i,n)\in\mathcal{W}_{2t+1}} \overline{U}_{ip}^{(t)} \overline{U}_{iq}^{*(t)} \end{bmatrix},$$

$$K^{(t)} = \begin{bmatrix} K_{11}^{(t)} & \cdots & K_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ K_{k1}^{(t)} & \cdots & K_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad K_{pq}^{(t)} = \overline{U}_{*p}^{(t)\top} \overline{U}_{*q}^{*(t)} I_n,$$

where $1 \le p, q \le k$. Note that $S^{(t)}$ and $G^{(t)}$ are the partial Hessian matrices $\nabla_V^2 \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V)$

and $\nabla_V^2 \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)}, V)$ with respect to a vectorized V, i.e., $\mathrm{vec}(V)$.

C.5.1 Proof of Theorem 4.5.2 (Alternating Exact Minimization)

Proof. Throughout the proof for alternating exact minimization, we define a constant $\xi \in (2, \infty)$ to simplify the notation. We define the approximation error of the inexact first order oracle as

$$\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}) = \|\nabla_V \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)}, V^{(t+0.5)}) - \nabla_V \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V^{(t+0.5)})\|_{\mathrm{F}}.$$

To simplify our later analysis, we first introduce the following event.

$$\mathcal{E}_U^{(t)} = \left\{ \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \le \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1} \quad \text{and} \quad \max_i \|\overline{U}_{i*}^{(t)}\|_2 \le 2\mu\sqrt{\frac{k}{m}} \right\}.$$

We then present two important consequences of $\mathcal{E}_U^{(t)}$.

Lemma C.5.2. Suppose that $\mathcal{E}_U^{(t)}$ holds, and $\widetilde{\rho}$ satisfies (C.5.1). Then we have

$$\mathbb{P}\Big(1 + \delta_{2k} \ge \sigma_{\max}(S^{(t)}) \ge \sigma_{\min}(S^{(t)}) \ge 1 - \delta_{2k}\Big) \ge 1 - n^{-3},$$

where δ_{2k} is some constant satisfying

$$\delta_{2k} \leq \frac{\sigma_k^6}{192\xi^2 k\sigma_1^6}. \tag{C.5.3}$$

The proof of Lemma C.5.2 is provided in Appendix C.6.1. Lemma C.5.2 is also applicable to $G^{(t)}$, since $G^{(t)}$ shares the same structure with $S^{(t)}$, and $\overline{U}^{*(t)}$ is incoherent with parameter μ.

Lemma C.5.3. Suppose that $\mathcal{E}_U^{(t)}$ holds, and $\widetilde{\rho}$ satisfies (C.5.1). Then for an incoherent V with parameter $3\sigma_1\mu$, we have

$$\mathbb{P}\Big(\|(S^{(t)}K^{(t)} - J^{(t)}) \cdot \mathrm{vec}(V)\|_2 \leq 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - U^{*(t)}\|_{\mathrm{F}}\Big) \geq 1 - n^{-3},$$

where δ_{2k} is defined in (C.5.3).

The proof of Lemma C.5.3 is provided in Appendix C.6.2. Note that Lemma C.5.3 is also applicable to $\|(G^{(t)}K^{(t)} - J^{(t)}) \cdot \mathrm{vec}(V)\|_2$, since $G^{(t)}$ shares the same structure with $S^{(t)}$, and $\overline{U}^{*(t)}$ is incoherent with parameter μ.

We then introduce another two events:

$$\mathcal{E}_{U,1}^{(t)} = \{1 + \delta_{2k} \geq \sigma_{\max}(S^{(t)}) \geq \sigma_{\min}(S^{(t)}) \geq 1 - \delta_{2k}\},$$
$$\mathcal{E}_{U,2}^{(t)} = \{1 + \delta_{2k} \geq \sigma_{\max}(G^{(t)}) \geq \sigma_{\min}(G^{(t)}) \geq 1 - \delta_{2k}\},$$

where δ_{2k} is defined in $\mathcal{E}_U^{(t)}$. By Lemmas C.5.2, we can verify that $\mathcal{E}_U^{(t)}$ implies $\mathcal{E}_{U,1}^{(t)}$

and $\mathcal{E}_{U,2}^{(t)}$ with probability at least $1-2n^{-3}$. The next lemma shows that $\mathcal{E}_{U,1}^{(t)}$ and $\mathcal{E}_{U,2}^{(t)}$ imply the strong convexity and smoothness of $\widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(U,V)$ in V at $U=\overline{U}^{(t)}$ and $\overline{U}^{*(t)}$.

Lemma C.5.4. Suppose that $\mathcal{E}_{U,1}^{(t)}$ and $\mathcal{E}_{U,2}^{(t)}$ hold. Then for any $V,\ V'\in\mathbb{R}^{n\times k}$, we have

$$\begin{aligned}\frac{1+\delta_{2k}}{2}\|V'-V\|_{\mathrm{F}}^2 \geq \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)},V') - \mathcal{F}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)},V)\\ -\langle\nabla_V\widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)},V),V'-V\rangle \geq \frac{1-\delta_{2k}}{2}\|V'-V\|_{\mathrm{F}}^2,\\ \frac{1+\delta_{2k}}{2}\|V'-V\|_{\mathrm{F}}^2 \geq \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)},V') - \mathcal{F}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)},V)\\ -\langle\nabla_V\widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)},V),V'-V\rangle \geq \frac{1-\delta_{2k}}{2}\|V'-V\|_{\mathrm{F}}^2.\end{aligned}$$

Since $S^{(t)}$ and $G^{(t)}$ are essentially the partial Hessian matrices $\nabla_V^2\widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)},V)$ and $\nabla_V^2\widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)},V)$, the proof of C.5.4 directly follows Appendix C.1.1, and is therefore omitted.

We then introduce another two events:

$$\begin{aligned}\mathcal{E}_{U,3}^{(t)} &= \{\|(S^{(t)}K^{(t)}-J^{(t)})\cdot\mathrm{vec}(V^{*(t)})\|_2 \leq 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)}-U^{*(t)}\|_{\mathrm{F}}\},\\ \mathcal{E}_{U,4}^{(t)} &= \{\|(G^{(t)}K^{(t)}-J^{(t)})\cdot\mathrm{vec}(V^{*(t)})\|_2 \leq 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)}-U^{*(t)}\|_{\mathrm{F}}\},\end{aligned}$$

where δ_{2k} is defined in $\mathcal{E}_U^{(t)}$. We can verify that $\mathcal{E}_U^{(t)}$ implies $\mathcal{E}_{U,3}^{(t)}$ and $\mathcal{E}_{U,4}^{(t)}$ with probability at least $1-2n^{-3}$ by showing the incoherence of $V^{*(t)}$. More specifically,

let $V^{*(t)} = \overline{V}^{*(t)} R_V^{*(t)}$ denote the QR decomposition of $V^{*(t)}$. We have

$$\|V_{j*}^{*(t)}\|_2 = \|R_V^{*(t)\top} V^{*(t)\top} e_j\|_2 \leq \|R_V^{*(t)}\|_2 \|V^{*(t)\top} e_j\|_2 \leq \sigma_1 \|V_{j*}^{*(t)}\|_2 \leq \sigma_1 \mu \sqrt{\frac{k}{n}}. \tag{C.5.4}$$

Then Lemma C.5.3 are applicable to $\mathcal{E}_{U,3}^{(t)}$ and $\mathcal{E}_{U,4}^{(t)}$.

We then introduce the following key lemmas, which will be used in the main proof.

Lemma C.5.5. Suppose that $\mathcal{E}_U^{(t)}$, $\mathcal{E}_{U,1}^{(t)}$,..., and $\mathcal{E}_{U,4}^{(t)}$ hold. We then have

$$\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}) \leq \frac{(1-\delta_{2k})\sigma_k}{2\xi} \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}.$$

Lemma C.5.5 shows that the approximation error of the inexact first order oracle for updating V diminishes with the estimation error of $U^{(t)}$, when $U^{(t)}$ is sufficiently close to $U^{*(t)}$. It is analogous to Lemma 4.4.3 in the analysis of matrix sensing, and its proof directly follows C.1.2, and is therefore omitted.

Lemma C.5.6. Suppose that $\mathcal{E}_U^{(t)}$, $\mathcal{E}_{U,1}^{(t)}$,..., and $\mathcal{E}_{U,4}^{(t)}$ hold. We then have

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{1}{1-\delta_{2k}} \mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}).$$

Lemma C.5.6 illustrates that the estimation error of $V^{(t+0.5)}$ diminishes with the approximation error of the inexact first order oracle. It is analogous to Lemma 4.4.4 in the analysis of matrix sensing, and its proof directly follows Appendix

C.1.3, and is therefore omitted.

Lemma C.5.7. Suppose that $V^{(t+0.5)}$ satisfies

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{8}. \tag{C.5.5}$$

Then there exists a factorization of $M^* = U^{*(t+1)}\overline{V}^{*(t+1)\top}$ such that $\overline{V}^{*(t+1)} \in \mathbb{R}^{n\times k}$ is an orthonormal matrix, and satisfies

$$\max_j \|\overline{V}^{(t+1)}_{j*}\|_2 \leq 2\mu\sqrt{\frac{2k}{n}} \quad \text{and} \quad \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{4}{\sigma_k}\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}.$$

The proof of Lemma C.5.7 is provided in Appendix C.6.3. Lemma C.5.7 ensures that the incoherence factorization enforces $\overline{V}^{(t+1)}$ to be incoherent with parameter 2μ.

Lemma C.5.8. Suppose that $\widetilde{\rho}$ satisfies (C.5.1). Then $\mathcal{E}_U^{(0)}$ holds with high probability.

The proof of Lemma C.5.8 is provided in Appendix C.6.4. Lemma C.5.8 shows that the initial solution $\overline{U}^{(0)}$ is incoherent with parameter 2μ, while achieving a sufficiently small estimation error with high probability. It is analogous to Lemma 4.4.6 for matrix sensing.

Combining Lemmas C.5.5, C.5.6, and C.5.7, we obtain the next corollary for a complete iteration of updating V.

Corollary C.5.9. Suppose that $\mathcal{E}_U^{(t)}$ holds. Then

$$\mathcal{E}_V^{(t)} = \left\{ \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \le \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1} \quad \text{and} \quad \max_j \|\overline{V}_{j*}^{(t+1)}\|_2 \le 2\mu\sqrt{\frac{k}{m}} \right\}$$

holds with probability at least $1 - 4n^{-3}$. Moreover, we have

$$\|\overline{V}^{(t+1)} - \overline{V}^*\|_{\mathrm{F}} \le \frac{2}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \text{ and } \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \le \frac{\sigma_k}{2\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}$$

with probability at least $1 - 4n^{-3}$.

The proof of Corollary C.5.9 is provided in Appendix C.6.5. Since the alternating exact minimization algorithm updates U and V in a symmetric manner, we can establish similar results for a complete iteration of updating U in the next corollary.

Corollary C.5.10. Suppose $\mathcal{E}_V^{(t)}$ holds. Then $\mathcal{E}_U^{(t+1)}$ holds with probability at least $1 - 4n^{-3}$. Moreover, we have

$$\|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \le \frac{2}{\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \text{ and } \|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}} \le \frac{\sigma_k}{2\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}$$

with probability at least $1 - 4n^{-3}$.

The proof of Lemma C.5.10 directly follows Appendix C.6.5, and is therefore omitted.

We proceed with the proof of Theorem 4.5.2 conditioning on $\mathcal{E}_U^{(0)}$. Similar to Section 4.3.3, we can recursively apply Corollaries C.5.9 and C.5.10, and show that $\{\mathcal{E}_U^{(t)}\}_{t=1}^T$ and $\{\mathcal{E}_V^{(t)}\}_{t=0}^T$ simultaneously hold with probability at least $1-8Tn^{-3}$. Then conditioning on all $\{\mathcal{E}_U^{(t)}\}_{t=0}^T$ and $\{\mathcal{E}_V^{(t)}\}_{t=0}^T$, we have

$$\begin{aligned}\|\overline{V}^{(T)}-\overline{V}^{*(T)}\|_{\mathrm{F}} &\le \frac{2}{\xi}\|\overline{U}^{(T-1)}-\overline{U}^{*(T-1)}\|_{\mathrm{F}} \le \left(\frac{2}{\xi}\right)^2\|\overline{V}^{(T-1)}-\overline{V}^{*(T-1)}\|_{\mathrm{F}} \\ &\le \left(\frac{2}{\xi}\right)^{2T-1}\|\overline{U}^{(0)}-\overline{U}^{*(0)}\|_{\mathrm{F}} \le \left(\frac{2}{\xi}\right)^{2T}\frac{(1-\delta_{2k})\sigma_k}{8(1+\delta_{2k})\sigma_1}, \end{aligned} \tag{C.5.6}$$

where the last inequality comes from the definition of $\mathcal{E}_U^{(0)}$. Thus we only need

$$T=\left\lceil \frac{1}{2}\log^{-1}\left(\frac{\xi}{2}\right)\log\left(\frac{(1-\delta_{2k})\sigma_k}{4(1+\delta_{2k})\sigma_1}\cdot\frac{1}{\epsilon}\right)\right\rceil$$

iterations such that

$$\|\overline{V}^{(T)}-\overline{V}^{*(T)}\|_{\mathrm{F}} \le \left(\frac{2}{\xi}\right)^{2T}\frac{(1-\delta_{2k})\sigma_k}{8(1+\delta_{2k})\sigma_1} \le \frac{\epsilon}{2}. \tag{C.5.7}$$

Meanwhile, by (C.5.7) and Corollary C.5.10, we have

$$\|U^{(T-0.5)}-U^{*(T)}\|_{\mathrm{F}} \le \frac{\sigma_k}{2\xi}\|\overline{V}^{(T)}-\overline{V}^{*(T)}\|_{\mathrm{F}} \le \left(\frac{2}{\xi}\right)^{2T}\frac{(1-\delta_{2k})\sigma_k^2}{16\xi(1+\delta_{2k})\sigma_1},$$

where the last inequality comes from (C.5.6). Thus we only need

$$T = \left\lceil \frac{1}{2}\log^{-1}\left(\frac{\xi}{2}\right)\log\left(\frac{(1-\delta_{2k})\sigma_k^2}{8\xi(1+\delta_{2k})}\cdot\frac{1}{\epsilon}\right)\right\rceil$$

iterations such that

$$\|U^{(T-0.5)} - U^{*(T)}\|_{\mathrm{F}} \leq \left(\frac{2}{\xi}\right)^{2T}\frac{(1-\delta_{2k})\sigma_k^2}{16\xi(1+\delta_{2k})\sigma_1} \leq \frac{\epsilon}{2\sigma_1}. \tag{C.5.8}$$

We then combine (C.5.7) and (C.5.8) by following similar lines to Section 4.4.2, and show

$$\|M^{(T)} - M^*\|_{\mathrm{F}} \leq \epsilon. \tag{C.5.9}$$

The above analysis only depends on $\mathcal{E}_U^{(0)}$. Because Lemma C.5.8 guarantees that $\mathcal{E}_U^{(0)}$ holds with high probability, given $T \ll n^3$, (C.5.9) also holds with high probability. □

C.5.2 Proof of Theorem 4.5.2 (Alternating Gradient Descent)

Proof. Throughout the proof for alternating gradient descent, we define a sufficiently large constant ξ. Moreover, we assume that at the t-th iteration, there

exists a matrix factorization of M^* as

$$M^* = \overline{U}^{*(t)} V^{*(t)\top},$$

where $\overline{U}^{*(t)} \in \mathbb{R}^{m\times k}$ is an orthonormal matrix. We define the approximation error of the inexact first order oracle as

$$\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \leq \|\widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V^{(t)}) - \nabla_V \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)}, V^{(t)})\|_{\mathrm{F}}$$

To simplify our later analysis, we introduce the following event.

$$\mathcal{E}_U^{(t)} = \Bigg\{ \max_i \|\overline{U}_{i*}^{(t)}\|_2 \leq 2\mu\sqrt{\frac{k}{n}}, \quad \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{4\xi\sigma_1^2} \\ \max_i \|V_{j*}^{(t)}\|_2 \leq 2\sigma_1\mu\sqrt{\frac{k}{n}}, \quad \text{and} \quad \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{2\xi\sigma_1} \Bigg\}.$$

As has been shown in Appendix C.5.1, $\mathcal{E}_U^{(t)}$ implies the following four events with probability at least $1 - 4n^{-3}$,

$$\mathcal{E}_{U,1}^{(t)} = \Big\{1 + \delta_{2k} \geq \sigma_{\max}(S^{(t)}) \geq \sigma_{\min}(S^{(t)}) \geq 1 - \delta_{2k}\Big\},$$
$$\mathcal{E}_{U,2}^{(t)} = \Big\{1 + \delta_{2k} \geq \sigma_{\max}(G^{(t)}) \geq \sigma_{\min}(G^{(t)}) \geq 1 - \delta_{2k}\Big\},$$
$$\mathcal{E}_{U,3}^{(t)} = \Big\{\|(S^{(t)}K^{(t)} - J^{(t)}) \cdot \mathrm{vec}(V^{*(t)})\|_2 \leq 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - U^{*(t)}\|_{\mathrm{F}}\Big\},$$
$$\mathcal{E}_{U,4}^{(t)} = \Big\{\|(G^{(t)}K^{(t)} - J^{(t)}) \cdot \mathrm{vec}(V^{*(t)})\|_2 \leq 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - U^{*(t)}\|_{\mathrm{F}}\Big\},$$

where δ_{2k} is defined in (C.5.3). In Appendix C.5.1, we also show that $\mathcal{E}_{U,1}^{(t)}$ and $\mathcal{E}_{U,2}^{(t)}$ imply the strong convexity and smoothness of $\widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(U,V)$ at $U=\overline{U}^{(t)}$ and $\overline{U}^{*(t)}$.

Moreover, we introduce the following two events,

$$\mathcal{E}_{U,5}^{(t)} = \Big\{\|(S^{(t)}K^{(t)} - J^{(t)})\cdot \mathrm{vec}(V^{(t)} - V^{*(t)})\|_2 \le 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - U^{*(t)}\|_{\mathrm{F}}\Big\},$$
$$\mathcal{E}_{U,6}^{(t)} = \Big\{\|(G^{(t)}K^{(t)} - J^{(t)})\cdot \mathrm{vec}(V^{(t)} - V^{*(t)})\|_2 \le 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - U^{*(t)}\|_{\mathrm{F}}\Big\},$$

where δ_{2k} is defined in (C.5.3). We can verify that $\mathcal{E}_U^{(t)}$ implies $\mathcal{E}_{U,5}^{(t)}$ and $\mathcal{E}_{U,6}^{(t)}$ with probability at least $1-2n^{-3}$ by showing the incoherence of $V^{(t)}-V^{*(t)}$. More specifically, we have

$$\max_j \|V_{j*}^{(t)} - V_{j*}^{*(t)}\|_2 \le \max_i \|V_{j*}^{(t)}\|_2 + \max_j \|V_{j*}^{*(t)}\|_2 \le 3\sigma_1\mu\sqrt{\frac{k}{n}},$$

where the last inequality follows the definition of $\mathcal{E}_U^{(t)}$ and the incoherence of $V^{*(t)}$ as shown in (C.5.4). Then Lemma C.5.3 are applicable to $\mathcal{E}_{U,5}^{(t)}$ and $\mathcal{E}_{U,6}^{(t)}$

We then introduce the following key lemmas, which will be used in the main proof.

Lemma C.5.11. Suppose that $\mathcal{E}_U^{(t)}$, $\mathcal{E}_{U,1}^{(t)}$, ..., and $\mathcal{E}_{U,6}^{(t)}$ hold. Then we have

$$\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \le \frac{(1+\delta_{2k})\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}.$$

Lemma C.5.11 shows that the approximation error of the inexact first order

oracle for updating V diminishes with the estimation error of $U^{(t)}$, when $U^{(t)}$ is sufficiently close to $U^{*(t)}$. It is analogous to Lemma 4.4.9 in the analysis of matrix sensing, and its proof directly follows C.2.1, and is therefore omitted.

Lemma C.5.12. Suppose that $\mathcal{E}_U^{(t)}$, $\mathcal{E}_{U,1}^{(t)}$, ..., and $\mathcal{E}_{U,6}^{(t)}$ hold. Meanwhile, the rescaled step size parameter $\widetilde{\eta}$ satisfies

$$\widetilde{\eta} = \frac{1}{1+\delta_{2k}}.$$

Then we have

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \sqrt{\delta_{2k}}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2}{1+\delta_{2k}}\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}).$$

Lemma C.5.12 illustrates that the estimation error of $V^{(t+0.5)}$ diminishes with the approximation error of the inexact first order oracle. It is analogous to Lemma 4.4.10 in the analysis of matrix sensing. Its proof directly follows Appendix C.2.2, and is therefore omitted.

Lemma C.5.13. Suppose that $V^{(t+0.5)}$ satisfies

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{8}.$$

We then have

$$\max_j \|\overline{V}_{j*}^{(t+1)}\|_2 \le 2\mu\sqrt{\frac{2k}{n}} \quad \text{and} \quad \max_i \|U_{i*}^{(t+1)}\|_2 \le 2\sigma_1\mu\sqrt{\frac{2k}{m}}$$

Moreover, there exists a factorization of $M^* = U^{*(t+1)}\overline{V}^{*(t+1)\top}$ such that $\overline{V}^{*(t+1)}$ is an orthonormal matrix, and

$$\begin{aligned}
\|\overline{V}^{(t+1)} - \overline{V}^{*(t)}\|_{\mathrm{F}} &\le \frac{4}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}, \\
\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} &\le \frac{5\sigma_1}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}.
\end{aligned}$$

The proof of Lemma C.5.7 is provided in Appendix C.7.1. Lemma C.5.13 guarantees that the incoherence factorization enforces $\overline{V}^{(t+1)}$ and $U^{(t)}$ to be incoherent with parameters 2μ and $2\sigma_1\mu$ respectively. The next lemma characterizes the estimation error of the initial solutions.

Lemma C.5.14. Suppose that $\widetilde{\rho}$ satisfies (C.5.1). Then $\mathcal{E}_U^{(0)}$ holds with high probability.

The proof of Lemma C.5.14 is provided in Appendix C.7.2. Lemma C.5.14 ensures that the initial solutions $U^{(0)}$, and $V^{(0)}$ are incoherent with parameters 2μ and $2\sigma_1\mu$ respectively, while achieving sufficiently small estimation errors with high probability. It is analogous to Lemma 4.4.12 in the analysis of matrix sensing.

Combining Lemmas C.5.11, C.5.12, and C.5.7, we obtain the following corol-

lary for a complete iteration of updating V.

Corollary C.5.15. Suppose that $\mathcal{E}_U^{(t)}$ holds. Then

$$\mathcal{E}_V^{(t)} = \Bigg\{ \max_j \|\overline{V}_{j*}^{(t)}\|_2 \le 2\mu\sqrt{\frac{k}{m}}, \quad \|\overline{V}^{(t)} - \overline{V}^{*(t)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{4\xi\sigma_1^2}, \\ \max_i \|U_{i*}^{(t)}\|_2 \le 2\sigma_1\mu\sqrt{\frac{k}{m}}, \text{ and } \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{2\xi\sigma_1} \Bigg\}$$

holds with probability at least $1 - 6n^{-3}$. Moreover, we have

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \le \sqrt{\delta_{2k}}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{C.5.10}$$

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t)}\|_{\mathrm{F}} \le \frac{4\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{8}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{C.5.11}$$

$$\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \le \frac{5\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \left(\frac{10}{\xi} + 1\right)\sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{C.5.12}$$

with probability at least $1 - 6n^{-3}$.

The proof of Corollary C.5.15 is provided in Appendix C.7.3. Since the algorithm updates U and V in a symmetric manner, we can establish similar results for a complete iteration of updating U in the next corollary.

Corollary C.5.16. Suppose $\mathcal{E}_V^{(t)}$ holds. Then $\mathcal{E}_U^{(t+1)}$ holds with probability at least

$1 - 6n^{-3}$. Moreover, we have

$$\|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \sqrt{\delta_{2k}}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \frac{2\sigma_k}{\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{C.5.13}$$

$$\|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{4\sqrt{\delta_{2k}}}{\sigma_k}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \frac{8}{\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{C.5.14}$$

$$\|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}} \leq \frac{5\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \left(\frac{10\sigma_1}{\xi} + 1\right)\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{C.5.15}$$

with probability at least $1 - 6n^{-3}$.

The proof of Corollary C.5.16 directly follows Appendix C.7.3, and is therefore omitted.

We then proceed with the proof of Theorem 4.5.2 conditioning on $\mathcal{E}_U^{(0)}$. Similar to Section 4.3.3, we can recursively apply Corollaries C.5.15 and C.5.16, and show that $\{\mathcal{E}_U^{(t)}\}_{t=1}^T$ and $\{\mathcal{E}_V^{(t)}\}_{t=0}^T$ simultaneously hold with probability at least $1 - 12Tn^{-3}$. For simplicity, we define

$$\phi_{V^{(t+1)}} = \|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}}, \;\; \phi_{V^{(t+0.5)}} = \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}, \;\; \phi_{\overline{V}^{(t+1)}} = \sigma_1\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}},$$

$$\phi_{U^{(t+1)}} = \|U^{(t+1)} - U^{*(t+2)}\|_{\mathrm{F}}, \;\; \phi_{U^{(t+0.5)}} = \|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}}, \;\; \phi_{\overline{U}^{(t+1)}} = \sigma_1\|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}}.$$

We then follow similar lines to Section 4.4.3 and Appendix C.5.1, and show that $\|M^{(T)} - M\|_{\mathrm{F}} \leq \epsilon$ with high probability. □

C.5.3 Proof of Theorem 4.5.2 (Gradient Descent)

Proof. The convergence analysis of the gradient descent algorithm is similar to alternating gradient descent. The only difference is, for updating U, gradient descent uses $V = \overline{V}^{(t)}$ instead of $V = \overline{V}^{(t+1)}$ to calculate the gradient at $U = U^{(t)}$. Then everything else directly follows Appendix C.5.2, and is therefore omitted.

□

C.6 Lemmas for Theorem 4.5.2 (Alternating Exact Minimization)

C.6.1 Proof of Lemma C.5.2

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^*$ and V^* respectively. Before we proceed with the main proof, we first introduce the following lemma.

Lemma C.6.1. Suppose that $\widetilde{\rho}$ satisfies (C.5.1). For any $z \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$ such that $\sum_i z_i = 0$, and a $t \in \{0, ..., 2T\}$, there exists a universal constant C such that

$$\sum_{(i,j)\in\mathcal{W}_t} z_i w_j \leq C m^{1/4} n^{1/4} \widetilde{\rho}^{1/2} \|z\|_2 \|w\|_2$$

with probability at least $1 - n^{-3}$.

The proof of Lemma C.6.1 is provided in [77], and therefore omitted.

We then proceed with the proof of Lemma C.5.2. For $j = 1, ..., k$, we define $S^{(j,t)}$, $J^{(j,t)}$, and $K^{(j,t)}$ as

$$S^{(j,t)} = \frac{1}{\widetilde{\rho}} \sum_{i:(i,j)\in\mathcal{W}_{2t+1}} \overline{U}_{i*}^{(t)} \overline{U}_{i*}^{(t)\top}, \;\; J^{(j,t)} = \frac{1}{\widetilde{\rho}} \sum_{i:(i,j)\in\mathcal{W}_{2t+1}} \overline{U}_{i*}^{(t)} \overline{U}_{i*}^{*\top}, \;\; \text{and} \;\; K^{(j,t)} = \overline{U}^{(t)\top} \overline{U}^{*}.$$

We then consider an arbitrary $W \in \mathbb{R}^{n\times k}$ such that $\|W\|_{\mathrm{F}} = 1$ and $w = \mathrm{vec}(W)$. Then we have

$$\max_j \sigma_{\max}(S^{(j,t)}) \geq w^\top S^{(t)} w = \sum_{j=1} W_{j*}^\top S^{(j,t)} W_{j*} \geq \min_j \sigma_{\min}(S^{(j,t)}). \tag{C.6.1}$$

Since $\mathcal{W}_{2t+1}$ is drawn uniformly at random, we can use mn independent Bernoulli random variables δ_{ij}'s to describe $\mathcal{W}_{2t+1}$, i.e., $\delta_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\widetilde{\rho})$ with $\delta_{ij} = 1$ denoting $(i,j) \in \mathcal{W}_{2t+1}$ and 0 denoting $(i,j) \notin \mathcal{W}_{2t+1}$. We then consider an arbitrary $z \in \mathbb{R}^k$ with $\|z\|_2 = 1$, and define

$$Y = z^\top S^{(j,t)} z = \frac{1}{\widetilde{\rho}} \sum_{i:(i,j)\in\mathcal{W}_{2t+1}} \left(z^\top \overline{U}_{i*}^{(t)}\right)^2 = \frac{1}{\widetilde{\rho}} \sum_i \delta_{ij} \left(z^\top \overline{U}_{i*}^{(t)}\right)^2.$$

Then we have

$$\mathbb{E}Y = z^\top \overline{U}^{(t)} \overline{U}^{(t)\top} z = 1 \quad \text{and} \quad \mathbb{E}Y^2 = \frac{1}{\widetilde{\rho}} \sum_i \delta_{ij} \left(z^\top \overline{U}_{i*}^{(t)}\right)^4 \leq \frac{4\mu^2 k}{m\widetilde{\rho}} \sum_i \left(z^\top \overline{U}_{i*}^{(t)}\right)^2 = \frac{4\mu^2 k}{m\widetilde{\rho}},$$

where the last inequality holds, since $\overline{U}^{(t)}$ is incoherent with parameter 2μ. Similarly, we can show

$$\max_i \left(z^\top \overline{U}_{i*}^{(t)}\right)^2 \leq \frac{4\mu^2 k}{m\widetilde{\rho}}.$$

Thus by Bernstein's inequality, we obtain

$$\mathbb{P}(|Y - \mathbb{E}Y| \geq \delta_{2k}) \leq \exp\left(-\frac{3\delta_{2k}^2}{6 + 2\delta_{2k}} \frac{m\widetilde{\rho}}{4\mu^2 k}\right).$$

Since $\widetilde{\rho}$ and δ_{2k} satisfy (4.5.3) and (C.5.3), then for a sufficiently large C_7, we have

$$\mathbb{P}(|Y - \mathbb{E}Y| \geq \delta_{2k}) \leq \frac{1}{n^3},$$

which implies that for any z and j, we have

$$\mathbb{P}\left(1 + \delta_{2k} \geq z^\top S^{(j,t)} z \geq 1 - \delta_{2k}\right) \geq 1 - n^{-3}. \tag{C.6.2}$$

Combining (C.6.2) with (C.6.1), we complete the proof. □

C.6.2 Proof of Lemma C.5.3

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^*$ and V^* respectively. Let $H^{(j,t)} = S^{(j,t)}K^{(j,t)} - J^{(j,t)}$. We have

$$H^{(j,t)} = \frac{1}{\widetilde{\rho}} \sum_{i:(i,j)\in\mathcal{W}_{2t+1}} \overline{U}_{i*}^{(t)}\overline{U}_{i*}^{(t)\top}\overline{U}^{(t)\top}\overline{U}^* - \overline{U}_{i*}\overline{U}_{i*}^\top = \frac{1}{\widetilde{\rho}} \sum_{i:(i,j)\in\mathcal{W}_{2t+1}} H^{(i,j,t)}.$$

We consider an arbitrary $Z \in \mathbb{R}^{n\times k}$ such that $\|Z\|_{\mathrm{F}} = 1$. Let $z = \mathrm{vec}(Z)$ and $v = \mathrm{vec}(V)$. Since

$$\sum_i H^{(i,j,t)} = \overline{U}^{(t)\top}\overline{U}^{(t)}\overline{U}^{(t)\top}\overline{U}^* - \overline{U}^{(t)\top}\overline{U}^* = 0,$$

then by Lemma C.6.1, we have

$$z^\top(S^{(t)}K^{(t)} - J^{(t)})v = \sum_j Z_{*j}^\top(S^{(j,t)}K^{(t)} - J^{(j,t)})V_{j*} \le \frac{1}{\rho}\sum_{p,q}\sqrt{\sum_j Z_{pj}^2(V_{jq})^2}\sqrt{\sum_i [H^{(i,j,t)}]_{pq}^2}.$$

Meanwhile, we have

$$\begin{aligned}
\sum_i [H^{(i,j,t)}]_{pq}^2 &= \sum_i (\overline{U}_{ip}^{(t)})^2(\overline{U}_{i*}^{(t)\top}\overline{U}^{(t)\top}\overline{U}_{*q}^* - \overline{U}_{iq}^*)^2 \le \max_i(\overline{U}_{ip}^{(t)})^2 \sum_i (\overline{U}_{i*}^{(t)\top}\overline{U}^{(t)\top}\overline{U}_{*q}^* - \overline{U}_{iq}^*)^2 \\
&= \max_i(\overline{U}_{ip}^{(t)})^2(1 - \|\overline{U}^{(t)\top}\overline{U}_{*q}^*\|_2^2) \le \max_i \|\overline{U}_{i*}^{(t)}\|_2^2(1 - (\overline{U}_{*q}^{(t)\top}\overline{U}_{*q}^*)^2) \\
&\overset{\text{(i)}}{\le} \frac{4\mu^2 k}{m}(1 - \overline{U}_{*q}^{(t)\top}\overline{U}_{*q}^*)(1 + \overline{U}_{*q}^{(t)\top}\overline{U}_{*q}^*) \\
&\overset{\text{(ii)}}{\le} \frac{4\mu^2 k}{m}\|\overline{U}_{*q} - \overline{U}_{*q}^*\|_2^2 \le \frac{4\sqrt{2}\mu^2 k}{m}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}^2,
\end{aligned}$$

where (i) comes from the incoherence of $\overline{U}^{(t)}$, and (ii) comes from $\overline{U}_{*q}^{(t)\top}\overline{U}_{*q}^{*} \leq \|\overline{U}_{*q}^{(t)}\|_2\|\overline{U}_{*q}^{*}\|_2 \leq 1$.

Combining the above inequalities, by the incoherence of V and Bernstein's inequality, we have

$$z^\top(S^{(t)}K^{(t)} - J^{(t)})v \leq \sum_{p,q}\frac{4\sigma_1\mu^2 k}{m\widetilde{\rho}}\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}\|Z_{*p}\|_2 \leq 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}$$

with probability at least $1 - n^{-3}$, where the last inequality comes from the incoherence of V, $\sum_p \|Z_{*p}\|_2 \leq \sqrt{k}$, and a sufficiently large $\widetilde{\rho}$. Since z is arbitrary, then we have

$$\mathbb{P}\Big(\|(S^{(t)}K^{(t)} - J^{(t)})v\|_2 \leq 3\delta_{2k}k\sigma_1\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}\Big) \geq 1 - n^{-3},$$

which completes the proof. □

C.6.3 Proof of Lemma C.5.7

Proof. Recall that we have $W^{\mathrm{in}} = V^{(t+0.5)}$ and $\overline{V}^{(t+1)} = W^{\mathrm{out}}$ in Algorithm 9. Since $V^{(t+0.5)}$ satisfies (4.4.5) of Lemma 4.4.5, then there exists a factorization of $M^* = U^{*(t+0.5)}\overline{V}^{*(t+0.5)\top}$ such that $\overline{V}^{*(t+0.5)}$ is an orthonormal matrix, and satisfies

$$\|\overline{W}^{\mathrm{in}} - \overline{V}^{*(t+0.5)}\|_{\mathrm{F}} \leq \frac{2}{\sigma_k}\|W^{\mathrm{in}} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{2}{\sigma_k}\cdot\frac{\sigma_k}{8} = \frac{1}{4}. \tag{C.6.3}$$

Since the Frobenius norm projection is contractive, then we have

$$\|\widetilde{W} - \overline{V}^{*(t+0.5)}\|_{\mathrm{F}} \leq \|\overline{W}^{\mathrm{in}} - \overline{V}^{*(t+0.5)}\|_{\mathrm{F}} \leq \frac{1}{4}. \tag{C.6.4}$$

Since $\overline{V}^{*(t+0.5)}$ is an orthonormal matrix, by Lemma C.1.4, we have

$$\begin{aligned}\|\overline{W}^{\mathrm{out}} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} &\leq \frac{\sqrt{2}\|\overline{V}^{*(t+0.5)\dagger}\|_2\|\widetilde{W} - \overline{V}^{*(t+0.5)}\|_{\mathrm{F}}}{1 - \|\overline{W}^{\mathrm{in}} - \overline{V}^{*(t+0.5)}\|_{\mathrm{F}}\|\overline{V}^{*(t+0.5)\dagger}\|_2} \\ &\leq 2\|\widetilde{W} - \overline{V}^{*(t+0.5)}\|_{\mathrm{F}} \leq \frac{1}{2},\end{aligned} \tag{C.6.5}$$

where $\overline{V}^{*(t+1)} = \overline{V}^{*(t+0.5)}O$ for some unitary matrix $O \in \mathbb{R}^{k\times k}$, and the last inequality comes from (C.6.4). Moreover, since $\overline{V}^{*(t+1)}$ is an orthonormal matrix, then we have

$$\sigma_{\min}(\widetilde{W}) \geq \sigma_{\min}(\overline{V}^{*(t+1)}) - \|\widetilde{W} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \geq 1 - \frac{1}{2} = \frac{1}{2}.$$

where the last inequality comes from (C.6.5). Since $\overline{W}^{\mathrm{out}} = \widetilde{W}(R^{\mathrm{tmp}}_{\widetilde{W}})^{-1}$, then we have

$$\|\overline{W}^{\mathrm{out}}_{i*}\|_2 \leq \|\overline{W}^{\mathrm{out}\top}e_i\|_2 = \|(R_{\widetilde{W}})^{-1}\|_2\|\widetilde{W}^{\top}e_i\|_2 \leq \sigma^{-1}_{\min}(\widetilde{W})\mu\sqrt{\frac{k}{n}} \leq 2\mu\sqrt{\frac{k}{n}}.$$

□

C.6.4 Proof of Lemma C.5.8

Proof. Before we proceed with the main proof, we first introduce the following lemma.

Lemma C.6.2. Suppose that $\widetilde{\rho}$ satisfies (C.5.1). Recall that $\widetilde{U}$, $\widetilde{\Sigma}$, and $\widetilde{V}$ are defined in Algorithm 8. There exists a universal constant C such that

$$\|\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top - M^*\|_2 = C\sqrt{\frac{k}{\widetilde{\rho}\sqrt{mn}}}$$

with high probability.

The proof of Lemma C.6.2 is provided in [77], therefore omitted.

We then proceed with the proof of Lemma C.5.8. Since both $\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top$ and M^* are rank k matrices, then $\widetilde{U}\widetilde{\Sigma}\widetilde{V} - M^*$ has at most rank $2k$. Thus by Lemma C.6.2, we have

$$\begin{aligned}\|\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top - M^*\|_{\mathrm{F}}^2 \le 2k\|\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top - M^*\|_2^2 &\le \frac{2Ck^2}{\rho\sqrt{mn}}\|M^*\|_{\mathrm{F}}^2 \\ &\le \frac{2Ck^3\sigma_1^2}{\widetilde{\rho}\sqrt{mn}} \le \frac{\sigma_k^6(1-\delta_{2k})}{1024(1+\delta_{2k})\sigma_1^4\xi^2} \end{aligned} \tag{C.6.6}$$

with high probability, where the last inequality comes from (C.5.1) with

$$C_7 \ge \frac{2048(1+\delta_{2k})^2\sigma_1^6\xi^2}{\mu^2\sigma_k^6(1-\delta_{2k})^2}.$$

Suppose that M^* has a rank k singular value decomposition $M^* = \widetilde{U}^* \widetilde{D}^* \widetilde{V}^{*\top}$. Then we have

$$\begin{aligned}
\|\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top - M^*\|_{\mathrm{F}}^2 &= \|\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top} - \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top\|_{\mathrm{F}}^2 \\
&= \|\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top} - \widetilde{U}\widetilde{U}^\top\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top} + \widetilde{U}\widetilde{U}^\top\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top} - \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top\|_{\mathrm{F}}^2 \\
&= \|(I_m - \widetilde{U}\widetilde{U}^\top)\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top} + \widetilde{U}(\widetilde{U}^\top\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top} - \widetilde{\Sigma}\widetilde{V}^\top)\|_{\mathrm{F}}^2 \\
&\geq \|(I_m - \widetilde{U}\widetilde{U}^T)\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top}\|_{\mathrm{F}}^2.
\end{aligned}$$

Let $\widetilde{U}_\perp \in \mathbb{R}^{m\times(m-k)}$ denote the orthogonal complement to $\widetilde{U}$. Then we have

$$\|(I_m - \widetilde{U}\widetilde{U}^T)\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top}\|_{\mathrm{F}}^2 = \|(\widetilde{U}_\perp\widetilde{U}_\perp^\top)\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top}\|_2^2 = \|\widetilde{U}_\perp^\top\widetilde{U}^*\widetilde{D}^*\|_{\mathrm{F}}^2 \geq \frac{\sigma_k^2}{2}\|\widetilde{U}_\perp^\top\widetilde{U}^*\|_{\mathrm{F}}^2.$$

Thus Lemma 4.4.2 guarantees that for $\widetilde{O} = \operatorname{argmin}_{O^\top O = I_k}\|\widetilde{U} - \widetilde{U}^* O\|_{\mathrm{F}}$, we have

$$\|\widetilde{U} - \widetilde{U}^*\widetilde{O}\|_{\mathrm{F}} \leq \sqrt{2}\|\widetilde{U}_\perp^\top\widetilde{U}^*\|_{\mathrm{F}} \leq \frac{2}{\sigma_k}\|\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top - M^*\|_{\mathrm{F}}.$$

We define $\widetilde{U}^{*\mathrm{tmp}} = \widetilde{U}^*\widetilde{O}$. Then combining the above inequality with (C.6.6), we have

$$\|\widetilde{U} - \widetilde{U}^{*\mathrm{tmp}}\|_{\mathrm{F}} \leq \frac{2}{\sigma_k}\|\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top - M^*\|_{\mathrm{F}} \leq \frac{\sigma_k^2(1-\delta_{2k})}{16(1+\delta_{2k})\sigma_1^2\xi}.$$

Since the Frobenius norm projection is contractive, then we have

$$\|\widetilde{U}^{\text{tmp}} - \widetilde{U}^{*\text{tmp}}\|_{\text{F}} \leq \|\widetilde{U} - \widetilde{U}^{*\text{tmp}}\|_{\text{F}} \leq \frac{\sigma_k^2(1-\delta_{2k})}{16(1+\delta_{2k})\sigma_1^2\xi} \leq \frac{1}{16}, \tag{C.6.7}$$

where the last inequality comes from the definition of ξ and $\sigma_1 \geq \sigma_k$. Since $\widetilde{U}^{*\text{tmp}}$ is an orthonormal matrix, by Lemma C.1.4, we have

$$\begin{aligned}\|\overline{U}^{\text{out}} - \overline{U}^{*(0)}\|_{\text{F}} &\leq \frac{\sqrt{2}\widetilde{U}^{*\text{tmp}\dagger}\|\widetilde{U}^{\text{tmp}} - \widetilde{U}^{*\text{tmp}}\|_{\text{F}}}{1 - \|\widetilde{U}^{\text{tmp}} - \widetilde{U}^{*\text{tmp}}\|_{\text{F}}\|\widetilde{U}^{*\text{tmp}\dagger}\|_2} \\ &\leq 2\|\widetilde{U}^{\text{tmp}} - \widetilde{U}^{*\text{tmp}}\|_{\text{F}} \leq \frac{\sigma_k^2(1-\delta_{2k})}{8(1+\delta_{2k})\sigma_1^2\xi} \leq \frac{1}{8},\end{aligned} \tag{C.6.8}$$

where $\overline{U}^{*(0)} = \widetilde{U}^{*\text{tmp}}\widetilde{O}^{\text{tmp}}$ for some unitary matrix $\widetilde{O}^{\text{tmp}} \in \mathbb{R}^{k\times k}$ such that $\widetilde{O}^{\text{tmp}}\widetilde{O}^{\text{tmp}\top} = I_k$, and the last inequality comes from (C.6.7). Moreover, since $\overline{U}^{*(0)}$ is an orthonormal matrix, then we have

$$\sigma_{\min}(\widetilde{U}^{\text{tmp}}) \geq \sigma_{\min}(\overline{U}^{*(0)}) - \|\widetilde{U}^{\text{tmp}} - \overline{V}^{*(t+1)}\|_{\text{F}} \geq 1 - \frac{1}{8} = \frac{7}{8},$$

where the last inequality comes from (C.6.7). Since $\overline{U}^{\text{out}} = \widetilde{U}^{\text{tmp}}(R_{\overline{U}}^{\text{out}})^{-1}$, then we have

$$\|\overline{U}_{i*}^{\text{out}}\|_2 \leq \|\overline{U}^{\text{out}\top}e_i\|_2 = \|(R_{\overline{U}}^{\text{out}})^{-1}\|_2\|\widetilde{U}^{\top}e_i\|_2 \leq \sigma_{\min}^{-1}(\widetilde{U}^{\text{tmp}})\mu\sqrt{\frac{k}{m}} \leq \frac{8\mu}{7}\sqrt{\frac{k}{m}}.$$

Moreover, we define $V^{*(0)} = M^{*\top}\overline{U}^{*(0)}$. Then we have $\overline{U}^{*(0)}V^{*(0)\top} = \overline{U}^{*(0)}\overline{U}^{*(0)\top}M^* =$

M^*, where the last inequality comes from the fact that $\overline{U}^{*(0)}\overline{U}^{*(0)\top}$ is exactly the projection matrix for the column space of M^*. □

C.6.5 Proof of Corollary C.5.9

Proof. Since $\mathcal{E}_U^{(t)}$ implies that $\mathcal{E}_{U,1}^{(t)}$,..., and $\mathcal{E}_{U,4}^{(t)}$ hold with probability at least $1-4n^{-3}$, then combining Lemmas C.5.5 and C.5.6, we obtain

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{2\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \overset{\text{(i)}}{\leq} \frac{\sigma_k}{2\xi\sigma_1} \cdot \frac{\sigma_k(1-\delta_{2k})}{4(1+\delta_{2k})\sigma_1} = \frac{\sigma_k^2(1-\delta_{2k})}{8(1+\delta_{2k})\sigma_1^2\xi^2} \overset{\text{(ii)}}{\leq} \frac{\sigma_k}{8}$$

with probability at least $1-4n^{-3}$, where (i) comes from the definition of $\mathcal{E}_U^{(t)}$, and (ii) comes from the definition of ξ and $\sigma_1 \geq \sigma_k$. Therefore Lemma C.5.7 implies that $\overline{V}^{(t+1)}$ is incoherent with parameter 2μ, and

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{4}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{2}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k(1-\delta_{2k})}{4(1+\delta_{2k})\sigma_1}$$

with probability at least $1-4n^{-3}$, where the last inequality comes from the definition of ξ and $\mathcal{E}_U^{(t)}$. □

C.7 Lemmas for Theorem 4.5.2 (Alternating Gradient Descent)

C.7.1 Proof of Lemma C.5.13

Proof. Recall that we have $W^{\text{in}} = V^{(t+0.5)}$ and $\overline{V}^{(t+1)} = W^{\text{out}}$ in Algorithm 9. By Lemma C.5.7, we can show

$$\|\overline{W}^{\text{out}} - \overline{V}^{*(t+1)}\|_{\text{F}} \leq \frac{4}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\text{F}}. \tag{C.7.1}$$

By Lemma C.2.1, we have

$$\begin{aligned}
\|R_{\overline{V}}^{(t+0.5)} - \overline{V}^{*(t+1)\top} V^{*(t)}\|_{\text{F}} &= \|\overline{V}^{(t+1)\top} V^{(t+0.5)} - \overline{V}^{*(t+1)\top} V^{*(t)}\|_{\text{F}} \\
&\leq \|\overline{V}^{(t+1)}\|_2 \|V^{(t+0.5)} - V^{*(t)}\|_{\text{F}} + \|V^{*(t)}\|_2 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\text{F}} \\
&\leq \|V^{(t+0.5)} - V^{*(t)}\|_{\text{F}} + \frac{4\sigma_1}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\text{F}},
\end{aligned} \tag{C.7.2}$$

where the last inequality comes from (C.7.1), $\|\overline{V}^{(t+1)}\|_2 = 1$, and $\|V^{*(t)}\|_2 = \sigma_1$. Moreover, we define $U^{*(t+1)} = \overline{U}^{*(t)}(\overline{V}^{*(t+1)\top} V^{*(t)})^\top$. Then we can verify

$$U^{*(t+1)}\overline{V}^{*(t+1)} = \overline{U}^{*(t)} V^{*(t)\top}\overline{V}^{*(t+1)}\overline{V}^{*(t+1)\top} = M^*,$$

where the last equality holds since $\overline{V}^{*(t+1)}\overline{V}^{*(t+1)}$ is exactly the projection matrix for the row space of M^*. Thus we further have

$$\begin{aligned}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} &= \|\overline{U}^{(t)}(\overline{V}^{(t+1)\top} V^{(t+0.5)})^\top - \overline{U}^{*(t)}(\overline{V}^{*(t+1)\top} V^{*(t)})^\top\|_{\mathrm{F}} \\ &\le \|\overline{U}^{(t)}\|_2 \|R_{\overline{V}}^{(t+0.5)} - \overline{V}^{*(t+1)\top} V^{*(t)}\|_{\mathrm{F}} + \|\overline{V}^{*(t+1)\top} V^{*(t)}\|_2 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\le \frac{5\sigma_1}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}},\end{aligned}$$

where the last inequality comes from (C.7.2), $\|\overline{U}^{(t)}\|_2 = 1$, $\|\overline{V}^{*(t+1)\top} V^{*(t)}\|_2 = \sigma_1$, and $\sigma_1 \ge \sigma_k$. □

C.7.2 Proof of Lemma C.5.14

Proof. Following similar lines to Appendix C.6.4, we can obtain

$$\max_i \|\overline{U}_{i*}^{(0)}\|_2 \le \frac{8\mu}{7}\sqrt{\frac{k}{m}}, \quad \|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{8\xi\sigma_1^2}, \tag{C.7.3}$$

$$\max_j \|\overline{V}_{j*}^{(0)}\|_2 \le \frac{8\mu}{7}\sqrt{\frac{k}{n}}, \quad \|\overline{V}^{(0)} - \overline{V}^{*(0)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{8\xi\sigma_1^2}. \tag{C.7.4}$$

Then by Lemma C.2.1, we have

$$\begin{aligned}\|\overline{U}^{(0)\top}\widetilde{M} - \overline{U}^{*(0)\top}M^*\|_{\mathrm{F}} &\le \|\overline{U}^{(0)}\|_2\|\widetilde{M} - M^*\|_{\mathrm{F}} + \|M^*\|_2\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \\ &\le \frac{\sigma_1^3}{32\xi\sigma_k^2} + \frac{\sigma_k^2}{8\xi\sigma_1} \le \frac{5\sigma_k^2}{32\xi\sigma_1}.\end{aligned} \tag{C.7.5}$$

By Lemma C.2.1 again, we have

$$\begin{aligned}
&\|\overline{V}^{(0)\top}\widetilde{M}^{\top}\overline{U}^{(0)} - \overline{V}^{*(0)\top}M^{*\top}\overline{U}^{*(0)}\|_{\mathrm{F}} \\
&\quad\le \|\overline{V}^{(0)}\|_2\|\widetilde{M}^{\top}\overline{U}^{(0)} - M^{*\top}\overline{U}^{*(0)}\|_{\mathrm{F}} + \|\overline{U}^{*(0)\top}M^{*}\|_2\|\overline{V}^{(0)} - \overline{V}^{*(0)}\|_{\mathrm{F}} \\
&\quad\le \frac{5\sigma_k^2}{32\xi\sigma_1} + \frac{\sigma_k^2}{8\xi\sigma_1} \le \frac{9\sigma_k^2}{32\xi\sigma_1}, \qquad \text{(C.7.6)}
\end{aligned}$$

where the last inequality comes from (C.7.4) and (C.7.5), and $\|M^*\|_2 = \sigma_1$. By Lemma C.2.1 again, we have

$$\begin{aligned}
\|V^{(0)} - V^{*(0)}\|_{\mathrm{F}} &\le \|\overline{V}^{(0)}\overline{V}^{(0)\top}\widetilde{M}^{\top}\overline{U}^{(0)} - \overline{V}^{*(0)}\overline{V}^{*(0)\top}M^{*\top}\overline{U}^{*(0)}\|_{\mathrm{F}} \\
&\le \|\overline{V}^{(0)}\|_2\|\overline{V}^{(0)\top}\widetilde{M}^{\top}\overline{U}^{(0)} - \overline{V}^{*(0)\top}M^{*\top}\overline{U}^{*(0)}\|_{\mathrm{F}} + \|\overline{U}^{*(0)\top}M^{*}\overline{V}^{*(0)}\|_2\|\overline{V}^{(0)} - \overline{V}^{*(0)}\|_{\mathrm{F}} \\
&\le \frac{9\sigma_k^2}{32\xi\sigma_1} + \frac{\sigma_k^2}{8\xi\sigma_1} \le \frac{13\sigma_k^2}{32\xi\sigma_1}, \qquad \text{(C.7.7)}
\end{aligned}$$

where the last two inequalities come from (C.7.4), (C.7.7), and $\|\overline{U}^{*(0)\top}M^{*}\overline{V}^{*(0)}\|_2 \le \sigma_1$, the definition of ξ, and $\sigma_1 \ge \sigma_k$. Moreover, by the incoherence of $V^{(0)}$, we have

$$\begin{aligned}
\|V_{j*}^{(0)}\|_2 &\le \|V^{(0)\top}e_j\|_2 = \|\overline{V}^{(0)\top}\widetilde{M}^{\top}\overline{U}^{(0)}\|_2\|\overline{V}^{(0)\top}e_i\|_2 \\
&\le \left(\|\overline{V}^{*(0)}\overline{V}^{*(0)\top}M^{*\top}\overline{U}^{*(0)}\|_2 + \|\overline{V}^{(0)\top}\widetilde{M}^{\top}\overline{U}^{(0)} - \overline{V}^{*(0)}\overline{V}^{*(0)\top}M^{*\top}\overline{U}^{*(0)}\|_{\mathrm{F}}\right)\frac{6\mu}{5}\sqrt{\frac{k}{m}} \\
&\le \left(1 + \frac{9\sigma_k^2}{32\xi\sigma_1^2}\right)\frac{6\sigma_1\mu}{5}\sqrt{\frac{k}{m}} \le \frac{41\sigma_1\mu}{28}\sqrt{\frac{k}{m}},
\end{aligned}$$

where the last two inequalities come from (C.7.4), (C.7.7), the definition of ξ, and $\sigma_1 \geq \sigma_k$

□

C.7.3 Proof of Corollary C.5.15

Proof. Since $\mathcal{E}_U^{(t)}$ implies that $\mathcal{E}_{U,1}^{(t)}$,...,$\mathcal{E}_{U,6}^{(t)}$ hold with probability $1-6n^{-3}$, then combining Lemmas C.5.11 and C.5.12, we obtain

$$\begin{aligned}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} &\leq \sqrt{\delta_{2k}}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2}{1+\delta_{2k}}\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \\ &\leq \sqrt{\delta_{2k}}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\leq \frac{\sigma_k^3}{12\xi\sigma_1^3}\cdot\frac{\sigma_k^2}{2\xi\sigma_1} + \frac{2\sigma_k^2}{\xi}\frac{\sigma_k}{4\xi\sigma_1^2} = \frac{\sigma_k^5}{24\xi^2\sigma_1^4} + \frac{\sigma_k^3}{2\xi^2\sigma_1^2} \leq \frac{\sigma_k}{8}\end{aligned}$$

with probability $1-6n^{-3}$, where the last inequality comes from the definition of ξ and $\sigma_1 \geq \sigma_k$. Thus by Lemma C.5.13, we have

$$\begin{aligned}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} &\leq \frac{4\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{8}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\leq \frac{4}{\sigma_k}\left(\frac{\sigma_k^5}{24\xi^2\sigma_1^4} + \frac{\sigma_k^3}{2\xi^2\sigma_1^2}\right) = \frac{\sigma_k^4}{6\xi^2\sigma_1^4} + \frac{2\sigma_k^2}{\xi^2\sigma_1^2} \leq \frac{\sigma_k^2}{4\xi\sigma_1^2},\end{aligned}$$

with probability $1 - 6n^{-3}$, where the last inequality comes from the definition of ξ and $\sigma_1 \geq \sigma_k$. Moreover, by Lemma C.5.13 again, we have

$$\begin{aligned}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} &\leq \frac{5\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \left(\frac{10}{\xi} + 1\right)\sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\leq \frac{5\sigma_1}{\sigma_k} \cdot \frac{\sigma_k^3}{12\xi\sigma_1^3} \cdot \frac{\sigma_k^2}{2\xi\sigma_1} + \left(\frac{10}{\xi} + 1\right)\sigma_1 \frac{\sigma_k^2}{4\xi\sigma_1^2} = \frac{5\sigma_k^4}{24\xi^2\sigma_1^3} + \frac{\sigma_k^2}{3\xi\sigma_1} \leq \frac{\sigma_k^2}{2\xi\sigma_1}\end{aligned}$$

with probability $1 - 6n^{-3}$, where the last inequality comes from the definition of ξ and $\sigma_1 \geq \sigma_k$. □

Bibliography

[1] P. Jain, A. Tewari, and P. Kar, "On iterative hard thresholding methods for high-dimensional m-estimation," in *NIPS*, 2014, pp. 685–693.

[2] N. Nguyen, D. Needell, and T. Woolf, "Linear convergence of stochastic iterative greedy algorithms with sparse constraints," *arXiv preprint arXiv:1407.0088*, 2014.

[3] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[4] R. Mazumder, J. H. Friedman, and T. Hastie, "SparseNet: Coordinate descent with nonconvex penalties," *Journal of the American Statistical Association*, vol. 106, pp. 1125–1138, 2011.

[5] T. Zhao, K. Roeder, and H. Liu, "Smooth-projected neighborhood pursuit for high-dimensional nonparanormal graph estimation," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 162–170.

[6] F. Han, T. Zhao, and H. Liu, "Coda: High dimensional copula discriminant analysis," *Journal of Machine Learning Research*, vol. 14, no. Feb, pp. 629–671, 2013.

[7] T. Zhao, K. Roeder, and H. Liu, "Positive semidefinite rank-based correlation matrix estimation with application to semiparametric graph estimation," *Journal of Computational and Graphical Statistics*, vol. 23, no. 4, pp. 895–922, 2014.

[8] T. Zhao and H. Liu, "Calibrated precision matrix estimation for high-dimensional elliptical distributions," *IEEE transactions on information theory/Professional Technical Group on Information Theory*, vol. 60, no. 12, pp. 7874–7887, 2014.

[9] T. Zhao, H. Liu, and T. Zhang, "A general theory of pathwise coordinate optimization," *arXiv preprint arXiv:1412.7477*, 2014.

[10] T. Zhao and H. Liu, "Accelerated path-following iterative shrinkage thresholding algorithm," *Journal of Computational and Graphical Statistics*, 2015, to appear.

[11] T. Zhao, Z. Wang, and H. Liu, "A nonconvex optimization framework for low rank matrix factorizaiton," *Advances in Neural Information Processing Systems*, 2015, to appear.

[12] ——, "Nonconvex low rank matrix factorization via inexact first order oracle," 2016, submitted.

[13] X. Li, T. Zhao, R. Arora, H. Liu, and J. Haupt, "Stochastic variance reduced optimization for nonconvex sparse learning," in *International Conference on Machine Learning*, 2016.

[14] B. M. Neale, Y. Kou, L. Liu, A. Ma'Ayan, K. E. Samocha, A. Sabo, C.-F. Lin, C. Stevens, L.-S. Wang, V. Makarov *et al.*, "Patterns and rates of exonic de novo mutations in autism spectrum disorders," *Nature*, vol. 485, no. 7397, pp. 242–245, 2012.

[15] A. Eloyan, J. Muschelli, M. B. Nebel, H. Liu, F. Han, T. Zhao, A. D. Barber, S. Joel, J. J. Pekar, S. H. Mostofsky *et al.*, "Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging," *Frontiers in systems neuroscience*, vol. 6, p. 61, 2012.

[16] H. Liu, L. Wang, and T. Zhao, "Calibrated multivariate regression with application to neural semantic basis discovery," *Journal of Machine Learning Research*, vol. 16, no. 8, pp. 1579–1606, 2015.

[17] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[18] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[19] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[20] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, pp. 894–942, 2010.

[21] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[22] N. Meinshausen and P. Bühlmann, "High dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[23] Y. Nesterov, "Gradient methods for minimizing composite objective function," *Mathematical Programming Series B*, vol. 140, pp. 125–161, 2013.

[24] Z.-Q. Luo and P. Tseng, "On the convergence of the coordinate descent method for convex differentiable minimization," *Journal of Optimization Theory and Applications*, vol. 72, no. 1, pp. 7–35, 1992.

[25] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of

block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.

[26] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 2, pp. 245–266, 2012.

[27] W. J. Fu, "Penalized regressions: the bridge versus the lasso," *Journal of Computational and Graphical Statistics*, vol. 7, no. 3, pp. 397–416, 1998.

[28] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, pp. 1–13, 2010.

[29] C.-H. Zhang and J. Huang, "The sparsity and bias of the lasso selection in high-dimensional linear regression," *The Annals of Statistics*, vol. 36, no. 4, pp. 1567–1594, 2008.

[30] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," *The Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.

[31] Z. Wang, H. Liu, and T. Zhang, "Optimal computational and statistical

rates of convergence for sparse nonconvex learning problems," *The Annals of Statistics*, vol. 42, no. 6, pp. 2164–2201, 2014.

[32] P.-L. Loh and M. J. Wainwright, "Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima," *Journal of Machine Learning Research*, 2015, to appear.

[33] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2012.

[34] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated gaussian designs," *The Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, 2010.

[35] C.-H. Zhang and T. Zhang, "A general theory of concave regularization for high-dimensional sparse estimation problems," *Statistical Science*, vol. 27, no. 4, pp. 576–593, 2012.

[36] T. Zhang *et al.*, "Multi-stage convex relaxation for feature selection," *Bernoulli*, vol. 19, no. 5B, pp. 2277–2293, 2013.

[37] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *The Annals of Statistics*, vol. 36, no. 4, pp. 1509–1533, 2008.

[38] T. E. Scheetz, K.-Y. A. Kim, R. E. Swiderski, A. R. Philp, T. A. Braun, K. L.

Knudtson, A. M. Dorrance, G. F. DiBona, J. Huang, T. L. Casavant *et al.*, "Regulation of gene expression in the mammalian eye and its relevance to eye disease," *Proceedings of the National Academy of Sciences*, vol. 103, no. 39, pp. 14 429–14 434, 2006.

[39] J. Huang, S. Ma, and C.-H. Zhang, "Adaptive lasso for sparse high-dimensional regression models," *Statistica Sinica*, vol. 18, no. 4, p. 1603, 2008.

[40] L. Wang, Y. Kim, and R. Li, "Calibrating nonconvex penalized regression in ultra-high dimension," *The Annals of Statistics*, vol. 41, no. 5, pp. 2505–2536, 2013.

[41] J. Fan, L. Xue, H. Zou *et al.*, "Strong oracle optimality of folded concave penalized estimation," *The Annals of Statistics*, vol. 42, no. 3, pp. 819–849, 2014.

[42] J. Duchi, "Lecture notes for statistics and information theory," 2015, http://stanford.edu/class/stats311/Lectures/full_notes.pdf.

[43] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over-balls," *Information Theory, IEEE Transactions on*, vol. 57, no. 10, pp. 6976–6994, 2011.

[44] S. A. van de Geer, "High-dimensional generalized linear models and the lasso," *The Annals of Statistics*, vol. 36, no. 2, pp. 614–645, 2008.

[45] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[46] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *The Journal of Machine Learning Research*, vol. 9, pp. 485–516, 2008.

[47] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[48] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.

[49] X.-T. Yuan, P. Li, and T. Zhang, "Gradient hard thresholding pursuit for sparsity-constrained optimization," in *ICML*, 2013, pp. 71–79.

[50] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comp. Harm. Anal.*, vol. 27, no. 3, pp. 594–607, 2009.

[51] S. Foucart, "Hard thresholding pursuit: An algorithm for compressive sensing," *SIAM J. Numer. Anal.*, vol. 49, no. 6, pp. 2543–2563, 2011.

[52] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *NIPS*, 2013, pp. 315–323.

[53] J. Konečnỳ and P. Richtárik, "Semi-stochastic gradient descent methods," *arXiv preprint arXiv:1312.1666*, 2013.

[54] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2011, pp. 693–701.

[55] S. J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. J. Smola, "On variance reduction in stochastic gradient descent and its asynchronous variants," in *Advances in Neural Information Processing Systems*, 2015, pp. 2629–2637.

[56] J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar, "An asynchronous parallel stochastic coordinate descent algorithm," *Journal of Machine Learning Research*, vol. 16, no. 285-322, pp. 1–5, 2015.

[57] L. Xiao and T. Zhang, "A proximal stochastic gradient method with progressive variance reduction," *arXiv preprint arXiv:1403.4699*, 2014.

[58] N. L. Roux, M. Schmidt, and F. R. Bach, "A stochastic gradient method with an exponential convergence rate for finite training sets," in *NIPS*, 2012, pp. 2663–2671.

[59] S. Shalev-Shwartz and T. Zhang, "Stochastic dual coordinate ascent methods for regularized loss," *JMLR*, vol. 14, no. 1, pp. 567–599, 2013.

[60] X. Shen, W. Pan, and Y. Zhu, "Likelihood-based selection and sharp parameter estimation," *Journal of the American Statistical Association*, vol. 107, no. 497, pp. 223–232, 2012.

[61] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Advances in Neural Information Processing Systems*, 2014, pp. 1646–1654.

[62] A. Agarwal, S. Negahban, and M. J. Wainwright, "Fast global convergence of gradient methods for high-dimensional statistical recovery," *Ann. Statist.*, vol. 40, no. 5, pp. 2452–2482, 2012.

[63] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[64] E. J. Candes and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[65] E. L. Lehmann, G. Casella, and G. Casella, *Theory of point estimation*. Wadsworth & Brooks/Cole Advanced Books & Software, 1991.

[66] S. Negahban and M. J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, vol. 39, no. 2, pp. 1069–1097, 2011.

[67] ——, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *Journal of Machine Learning Research*, vol. 13, no. May, pp. 1665–1697, 2012.

[68] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan, "Perturbed iterate analysis for asynchronous stochastic optimization," *arXiv preprint arXiv:1507.06970*, 2015.

[69] D. Cai and X. He, "Manifold adaptive experimental design for text categorization," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 4, pp. 707–719, 2012.

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

[71] Y. Nesterov, *Introductory lectures on convex optimization : a basic course*, ser. Applied optimization. Kluwer Academic Publishers, 2004.

[72] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[73] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[74] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[75] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[76] K. Lee and Y. Bresler, "Admira: Atomic decomposition for minimum rank approximation," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4402–4416, 2010.

[77] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[78] ——, "Matrix completion from noisy entries," *Journal of Machine Learning Research*, vol. 11, pp. 2057–2078, 2010.

[79] P. Jain, R. Meka, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Advances in Neural Information Processing Systems*, 2010, pp. 937–945.

[80] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[81] B. Recht, "A simpler approach to matrix completion," *Journal of Machine Learning Research*, vol. 12, pp. 3413–3430, 2011.

[82] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[83] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[84] D. Hsu, S. M. Kakade, and T. Zhang, "Robust matrix decomposition with sparse corruptions," *IEEE Transactions on Information Theory*, vol. 57, no. 11, pp. 7221–7234, 2011.

[85] A. Rohde and A. B. Tsybakov, "Estimation of high-dimensional low-rank matrices," *The Annals of Statistics*, vol. 39, no. 2, pp. 887–930, 2011.

[86] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, vol. 39, no. 5, pp. 2302–2329, 2011.

[87] J. Chen, J. Zhou, and J. Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 42–50.

[88] S. Xiang, Y. Zhu, X. Shen, and J. Ye, "Optimal exact least squares rank minimization," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 480–488.

[89] A. Agarwal, S. Negahban, and M. J. Wainwright, "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions," *The Annals of Statistics*, vol. 40, no. 2, pp. 1171–1197, 2012.

[90] B. Recht and C. Ré, "Parallel stochastic gradient algorithms for large-scale matrix completion," *Mathematical Programming Computation*, vol. 5, no. 2, pp. 201–226, 2013.

[91] Y. Chen, "Incoherence-optimal matrix completion," *arXiv preprint arXiv:1310.0154*, 2013.

[92] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward, "Coherent matrix completion," *arXiv preprint arXiv:1306.2979*, 2013.

[93] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis, "Low-rank matrix re-

covery from errors and erasures," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4324–4337, 2013.

[94] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Symposium on Theory of Computing*, 2013, pp. 665–674.

[95] P. Jain and P. Netrapalli, "Fast exact matrix completion with finite samples," *arXiv preprint arXiv:1411.1087*, 2014.

[96] M. Hardt, "Understanding alternating minimization for matrix completion," in *Symposium on Foundations of Computer Science*, 2014, pp. 651–660.

[97] M. Hardt, R. Meka, P. Raghavendra, and B. Weitz, "Computational limits for matrix completion," *arXiv preprint arXiv:1402.2331*, 2014.

[98] M. Hardt and M. Wootters, "Fast matrix completion without the condition number," *arXiv preprint arXiv:1407.4070*, 2014.

[99] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *arXiv preprint arXiv:1411.8003*, 2014.

[100] T. Hastie, R. Mazumder, J. Lee, and R. Zadeh, "Matrix completion and low-rank svd via fast alternating least squares," *arXiv preprint arXiv:1410.2596*, 2014.

[101] T. T. Cai and A. Zhang, "ROP: Matrix recovery via rank-one projections," *The Annals of Statistics*, vol. 43, no. 1, pp. 102–138, 2015.

[102] Q. Yan, J. Ye, and X. Shen, "Simultaneous pursuit of sparseness and rank structures for matrix decomposition," *Journal of Machine Learning Research*, vol. 16, pp. 47–75, 2015.

[103] Y. Zhu, X. Shen, and C. Ye, "Personalized prediction and sparsity pursuit in latent factor models," *Journal of the American Statistical Association*, 2015, to appear.

[104] Z. Wang, M.-J. Lai, Z. Lu, W. Fan, H. Davulcu, and J. Ye, "Orthogonal rank-one matrix pursuit for low rank matrix completion," *SIAM Journal on Scientific Computing*, vol. 37, no. 1, pp. A488–A514, 2015.

[105] C.-J. Hsieh and P. Olsen, "Nuclear norm minimization via active subspace selection," in *International Conference on Machine Learning*, 2014, pp. 575–583.

[106] Y. Koren, "The bellkor solution to the netflix grand prize," *Netflix Prize Documentation*, vol. 81, 2009.

[107] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Computer*, vol. 18, pp. 30–37, 2009.

[108] G. Takács, I. Pilászy, B. Németh, and D. Tikk, "Major components of the

gravity recommendation system," *ACM SIGKDD Explorations Newsletter*, vol. 9, no. 2, pp. 80–83, 2007.

[109] A. Paterek, "Improving regularized singular value decomposition for collaborative filtering," in *Proceedings of KDD Cup and workshop*, vol. 2007, 2007, pp. 5–8.

[110] R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis, "Large-scale matrix factorization with distributed stochastic gradient descent," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 69–77.

[111] Y. Zhuang, W.-S. Chin, Y.-C. Juan, and C.-J. Lin, "A fast parallel sgd for matrix factorization in shared memory systems," in *ACM Conference on Recommender Systems*, 2013, pp. 249–256.

[112] O. Güler, "New proximal point algorithms for convex minimization," *SIAM Journal on Optimization*, vol. 2, no. 4, pp. 649–664, 1992.

[113] Z.-Q. Luo and P. Tseng, "Error bounds and convergence analysis of feasible descent methods: a general approach," *Annals of Operations Research*, vol. 46, no. 1, pp. 157–178, 1993.

[114] A. Nedić and D. Bertsekas, "Convergence rate of incremental subgradient al-

gorithms," in *Stochastic optimization: algorithms and applications*. Springer, 2001, pp. 223–264.

[115] A. d'Aspremont, "Smooth optimization with approximate gradient," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1171–1183, 2008.

[116] M. Baes, "Estimate sequence methods: extensions and approximations," *Institute for Operations Research, ETH, Zürich, Switzerland*, 2009.

[117] M. P. Friedlander and M. Schmidt, "Hybrid deterministic-stochastic methods for data fitting," *SIAM Journal on Scientific Computing*, vol. 34, no. 3, pp. A1380–A1405, 2012.

[118] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Mathematical Programming*, vol. 146, no. 1-2, pp. 37–75, 2014.

[119] E. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *arXiv preprint arXiv:1407.1065*, 2014.

[120] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *arXiv preprint arXiv:1408.2156*, 2014.

[121] S. Arora, R. Ge, T. Ma, and A. Moitra, "Simple, efficient, and neural algorithms for sparse coding," *arXiv preprint arXiv:1503.00778*, 2015.

[122] Q. Zheng and J. Lafferty, "A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements," *arXiv preprint arXiv:1506.06081*, 2015.

[123] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi, "Dropping convexity for faster semi-definite optimization," *arXiv preprint arXiv:1509.03917*, 2015.

[124] S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," *arXiv preprint arXiv:1507.03566*, 2015.

[125] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," *arXiv preprint arXiv:1509.03025*, 2015.

[126] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer, 2004, vol. 87.

[127] G. W. Stewart, J.-g. Sun, and H. B. Jovanovich, *Matrix perturbation theory*. Academic press New York, 1990, vol. 175.

[128] P. Bühlmann and S. van de Geer, *Statistics for high-dimensional data: Methods, theory and applications*. Springer, 2011.

[129] M. Rudelson and R. Vershynin, "Hanson-wright inequality and sub-

gaussian concentration," *Electronic Communications in Probability*, vol. 18, no. 82, pp. 1–9, 2013.

Vita

Tuo Zhao is finishing his Ph.D. degree in Computer Science at Johns Hopkins University under supervision of Prof. Han Liu and Prof. Raman Arora. He was a visiting student in Department of Biostatistics at Johns Hopkins School of Public Health from 2011 to 2012, and Department of Operations Research and Financial Engineering at Princeton University from 2014 to 2016. He was the core member of the JHU team winning the INDI ADHD 200 global competition on fMRI imaging-based diagnosis classification in 2011. He received Google Summer of Code awards from 2011 to 2014. He received Siebel scholarship in 2014, Baidu Fellowship in 2015-2016, and China Development Bank Scholarship for Outstanding Graduates Abroad in 2016. He received 2016 ASA Best Student Paper Award on Statistical Computing, and 2016 INFORMS SAS Best Paper Award on Data Mining. He will join H. Milton Stewart School of Industrial and Systems Engineering at Georgia Institute of Technology as an assistant professor in 2017 Spring.