Structured Learning with Parsimony in Measurements and Computations: Theory, Algorithms, and Applications

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xingguo Li

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Professor Jarvis Haupt, Advisor

July, 2018

Acknowledgements

Every achievement is accompanied with supports and assistance from families, friends, colleagues, and even more people. As an old saying goes: one could see further only because he stands higher on what is paved by precursors. In this moment of my personal accomplishment, there are many people that have earned my gratitude for their contribution.

To start with, I want to express my whole-hearted gratefulness to my parents, Xinshu Piao and Zhongxian Li. They have always trusted in me and supported me in their own best ways. More importantly, they are invariably the most patient listeners and life advisors. I also want to convey my gratitude to my parents-in-law, Kuihua Yang and Dawei Lian. They have been so supportive to us unconditionally over the years, which is immensely important to us.

People say that one good partner can prevail against a troop. I am so fortunate to have the perfect life partner, Qinshu Lian, who is more than an army to me. Her thoughtfulness, passion, and caring are among the most important reasons that I can get where I am today. I could not have overcome all challenges along the way without her support and love.

Before I started my Ph.D. study, I heard many people mentioning that how fortuitous one is to have a great Ph.D. supervisor. I cannot declare too much how fortunate I am to have Professor Jarvis Haupt as my supervisor. He not only provided the best professional training to me, but also delivered personal experience that is valuable for both research and life. I became more confident and independent researcher due to numerous support and various opportunities he provided, toward which I would like to send my exceptional appreciation to Jarvis.

In my experience of research, it was extremely important to be able to brainstorm with other active ones for exploring solutions and new ideas. I want to give my special thanks to my friends and colleagues: Tuo Zhao, Junwei Lu, Zhaoran Wang, Mingyi Hong, Weiyang Liu, Bo Dai, just to name a few. It was of great help to communicate with them in advancing my academic path. I also want to thank Professor Raman Arora, Professor David Woodruff, and Professor Robert Nowak, who warmly host my visit to their research labs and opened my eyes to a more diverse academic world.

Many thanks to my office mates: Akshay Soni, Swayambhoo Jain, Sirisha Rambhatla, Mojtaba Kadkhodaie Elyaderani, Di Xiao, Alex Gutierrez, Abhinav Sambasivan, Jineng Ren, and Scott Sievert. It was so much fun working with you and discussing problems regarding research and beyond.

Chinese often say that one counts on parents at home, while on friends outside. I would like to thank all my friends in Pittsburgh, Duluth, Minneapolis, and Atlanta for your help when there was a need, the pleasant time we spent together, and good memories we shared.

Dedication

To my families

Abstract

In modern “Big Data” applications, structured learning is the most widely employed methodology. Within this paradigm, the fundamental challenge lies in developing practical, effective algorithmic inference methods. Often (e.g., deep learning) successful heuristic-based approaches exist but theoretical studies are far behind, limiting understanding and potential improvements. In other settings (e.g., recommender systems) provably effective algorithmic methods exist, but the sheer sizes of datasets can limit their applicability. This twofold challenge motivates this work on developing new analytical and algorithmic methods for structured learning, with a particular focus on parsimony in measurements and computation, i.e., those requiring low storage and computational costs.

Toward this end, we make efforts to investigate the theoretical properties of models and algorithms that present significant improvement in measurement and computation requirement. In particular, we first develop randomized approaches for dimensionality reduction on matrix and tensor data, which allow accurate estimation and inference procedures using significantly smaller data sizes that only depend on the intrinsic dimension (e.g., the rank of matrix/tensor) rather than the ambient ones. Our next effort is to study iterative algorithms for solving high dimensional learning problems, including both convex and nonconvex optimization. Using contemporary analysis techniques, we demonstrate guarantees of iteration complexities that are analogous to the low dimensional cases. In addition, we explore the landscape of nonconvex optimizations that exhibit computational advantages over their convex counterparts and characterize their properties from a general point of view in theory.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In many data science applications the available data can be quite diverse, exhibiting characteristics of large scale (large number of observations) and multiple perspectives (high dimensionality), but also potentially containing uncertainties from various sources. For example, Facebook alone has over 1 billion active users sharing more than 5 billion pieces of data daily (messages, images, posts, comments and etc.), and there exist tens of millions of fake profiles (by Statista[1]). Traditional inference methods often perform relatively poorly in these challenging environments due to limits on computing, storage, and their lack of robustness to the nature of uncertainties present in contemporary datasets. This necessitates new inference methods that can effectively draw inferences from such unwieldy data.

Structured learning can provide a natural advantage when dealing with these challenging tasks. By leveraging the fact that the datasets themselves, or the nature of the data corruptions, adhere to some form of low-complexity model (e.g., natural redundancy in user preference data manifests as low rank structure in large arrays of preferences, while corruptions are often relatively few in number, or sparse), structured learning methods can sometimes cast large scale inference tasks into problems of learning only the parameters of these low-complexity models. This insight can enable potentially significant improvements over more naive, traditional approaches that do not directly exploit such low-complexity models.

Tremendous advances have been made in past decades in developing algorithms for

[1] Available at https://www.statista.com/topics/751/facebook/.

specific learning problems, for example, state-of-the-art results are now obtained using deep learning and neural networks in computer vision and natural language processing. However, other than for a small set of problems, the theoretical understanding of structured learning methods are rather limited and often far behind practitioners achievements. Even in settings where provably effective structured learning algorithmic methods do exist, their utility can be limited by the sheer size of the datasets on which they are to be deployed. For example, in recommendation systems where partially observed arrays of users preferences comprise the available data (as in the so-called Netflix problem), fitting low-rank models to enormous arrays (corresponding to millions of users preferences for potentially tens of thousands of items, or more) can be computationally prohibitive.

The challenges outlined above are pervasive and timely within the structured learning paradigm, and serve as the essential motivation for the work comprising this dissertation. Specifically, motivated by the lack of theoretical understanding for many successful methods in structured learning, we aim to develop new analytical methods to facilitate theoretical comprehension of these techniques. Such exercises are not purely academic; rigorous theoretical understanding of well-performing algorithms not only helps us understand their mechanism(s), but also (more importantly) can provide essential new insights that may be used to improve them. A complementary thrust of our efforts is focused on developing new structured learning methods that are explicitly designed with an eye toward reduced computational cost and storage requirements.

Our approach to this end consists of comprehensive analyses to investigate various measurement reduction, modeling, and computational efficient algorithms for solving the problems described above. In a first series of works, we utilize a careful, judicious reduction in the available data that are ultimately input to the inference engine (Chapter 2 and Chapter 3). In another thrust, we develop efficient iterative algorithms for high dimensional learning problems and study their convergence behavior that achieve same complexities with low dimensional problems (Chapter 4 and Chapter 5). Next, we shred light on some fundamental properties of nonconvex problems that demonstrate superiority over their convex counterparts (Chapter 6). The details of the topics to be discussed are listed as following:

- In Chapter 2, we study a randomized approach for dimension reduction on matrices. In particular, we examine the problem of locating outlier columns in a large, otherwise low-rank, matrix. We propose a simple two-step adaptive sensing and inference approach and establish theoretical guarantees for its performance; our results show that accurate outlier identification is achievable using very few linear summaries of the original data matrix – as few as the squared rank of the low-rank component plus the number of outliers, times constant and logarithmic factors. We demonstrate the performance of our approach experimentally in two stylized applications, one motivated by robust collaborative filtering tasks, and the other by saliency map estimation tasks arising in computer vision and automated surveillance, and also investigate extensions to settings where the data are group-structured, noisy, or possibly incomplete.

- In Chapter 3, we further study randomized approaches for dimensionality reduction on low-rank tensor regression based on least square problems. This is motivated by the fact that the effective number of parameters in a structured tensor is significantly smaller than the its size. We consider the CANDECOMP/PARAFAC decomposition and the Tucker decomposition for tensors. For both models, we show how to apply data dimensionality reduction techniques based on *sparse* random projections to reduce the problem to a much smaller one, for which if a tensor is a near-optimum to the smaller least square problem, then it is also a near-optimum to the original one. Leveraging the randomized sketching techniques, we obtain significant reduction in dimensionality and sparsity in the sketching matrix for ordinary least squares regression. In addition, we provide a number of numerical simulations supporting our theory.

- In Chapter 4, we consider analyzing first order algorithms for solving high dimensional sparse learning problems with guarantees. Many machine learning techniques sacrifice convenient computational structures to gain estimation robustness and modeling flexibility. However, by exploring the modeling structures, we find these "sacrifices" do not always require more computational efforts. To shed light on such a "free-lunch" phenomenon, we study the square-root-Lasso (SQRT-Lasso) type regression problem. Specifically, we show that the nonsmooth loss functions of the SQRT-Lasso type regression ease tuning effort and gain adaptivity to inhomogeneous noise, but is not necessarily more challenging than Lasso to solve. We can directly apply proximal algorithms (e.g. proximal gradient descent, proximal Newton, and proximal quasi-Newton algorithms) without worrying the nonsmoothness of the loss function. Theoretically, we prove that the proximal algorithms combined with the pathwise optimization scheme enjoy fast convergence guarantees with high probability. Numerical evaluations are provided to support our theoretical results.

- In Chapter 5, we study second order algorithms for solving nonconvex high dimensional learning problems. In particular, we propose a difference of convex (DC) proximal Newton algorithm for solving nonconvex regularized sparse learning problems in high dimensions. Our proposed algorithm integrates the proximal Newton algorithm with multi-stage convex relaxation based on DC programming, and enjoys both strong computational and statistical guarantees. Specifically, by leveraging a sophisticated characterization of sparse modeling structures/assumptions (i.e., local restricted strong convexity and Hessian smoothness), we prove that within each stage of convex relaxation, our proposed algorithm achieves (local) quadratic convergence, and eventually obtains a sparse approximate local optimum with optimal statistical properties after only a few convex relaxations. Numerical experiments are provided to support our theory.

- In Chapter 6, we propose a general theory for studying the landscape of nonconvex optimization with underlying symmetric structures for a class of machine learning problems (e.g., low-rank matrix factorization, phase retrieval, and deep linear neural networks). In particular, we characterize the locations of stationary points and the null spaces of Hessian matrices of the objective function through the lens of invariant groups. As a major motivating example, we apply the proposed general theory to characterize the global landscape of the nonconvex optimization in certain low-rank matrix factorization problems. In particular, we illustrate how the rotational symmetry group gives rise to infinitely many nonisolated strict saddle points and equivalent global minima of the objective function. By explicitly identifying all stationary points, we divide the entire parameter space into three regions: the region containing the neighborhoods of all strict saddle points, where the objective has negative curvatures; the region containing neighborhoods of all global minima, where the objective enjoys strong convexity along certain directions; and the complement of the above regions, where the gradient has sufficiently large magnitudes. We further extend our result to the matrix sensing problem.

- In Chapter 7, we provide all detailed discussion of the analyses presented in the thesis.

The authors graciously acknowledge support from the NSF under Award CCF-1217751, DARPA Young Faculty Award N66001-14-1-4047, University of Minnesota Startup Funding, and Doctoral Dissertation Fellowship from University of Minnesota.

Chapter 2

Outlier Identification via Randomized Approaches

2.1 Introduction

We address a matrix *outlier identification* problem. Suppose $M \in \mathbb{R}^{n_1 \times n_2}$ is a data matrix that admits a decomposition of the form

$$M = L + C,$$

where L is a low-rank matrix, and C is a matrix of outliers that is nonzero in only a fraction of its columns. We are ultimately interested in identifying the locations of the nonzero columns of C, with a particular focus on settings where M may be very large. The question we address here is, can we accurately (and efficiently) identify the locations of the outliers from a small number of linear measurements of M?

Our investigation is motivated in part by robust collaborative filtering applications, in which the goal may be to identify the locations (or even quantify the number) of corrupted data points or outliers in a large data array. Such tasks may arise in a number of contemporary applications, for example, when identifying malicious responses in survey data or anomalous patterns in network traffic, to name a few. Depending on the nature of the outliers, conventional low-rank approximation approaches based on principal component analysis (PCA) [1, 2] may be viable options for these tasks, but such approaches become increasingly computationally demanding as the data become very high-dimensional. Here, our aim is to leverage dimensionality reduction ideas along the lines of those utilized in randomized numerical linear algebra, (see, e.g., [3,4] and the references therein) and compressed sensing (see, e.g., [5–7]), in order to reduce the size of the data on which our approach operates. In so doing, we also reduce the computational burden of the inference approach relative to comparable methods that operate on "full data."

We are also motivated by an image processing task that arises in many computer vision and surveillance applications – that of identifying the "saliency map" [8] of a given image, which (ideally) indicates the regions of the image that tend to attract the attention of a human viewer. Saliency map estimation is a well-studied area, and numerous methods have been proposed for obtaining saliency maps for a given image – see, for example, [9–13]. In contrast to these (and other) methods designed to identify saliency map of an image as a "post processing" step, our aim here is to estimate the saliency map *directly from compressive samples* – i.e., without first performing full image reconstruction as an intermediate step. We address this problem here using a linear subspace-based model of saliency, wherein we interpret an image as a collection of distinct (non-overlapping) patches, so that images may be (equivalently) represented as matrices whose columns are *vectorized* versions of the patches. Previous efforts have demonstrated that such local patches extracted from natural images may be well *approximated* as vectors in a union of low-dimensional linear subspaces (see, e.g., [14]). Here, our approach to the saliency map estimation problem is based on an assumption that salient regions in an image may be modeled as outliers from a single common low-dimensional subspace; the efficacy of similar models for visual saliency has been established recently in [15]. Our approach here may find utility in rapid threat detection in security and surveillance applications in high-dimensional imaging tasks, where the goal is not to image the entire scene, but rather to merely identify regions in the image space corresponding to anomalous behavior. Successful identification of salient regions could comprise a first step in an active vision task, where subsequent imaging is restricted to the identified regions.

Innovations and Our Approach. We propose a framework that employs dimensionality reduction techniques within the context of a two-step adaptive sampling and inference procedure [16–18], and our approach is based on a few key insights. First, we exploit the fact that the enabling geometry of our problem (to be formalized in the following section) is approximately preserved if we operate not on M directly, but instead on a "compressed" version ΦM that has potentially many fewer rows. Next, we use the fact that we can learn the (ostensibly, low-dimensional) linear subspace spanned by the columns of the low rank component of ΦM using a small, randomly selected subset of the columns of ΦM. Our algorithmic approach for this step utilizes a recently proposed method called *Outlier Pursuit* (OP) [19] that aims to separate a matrix Y into its low-rank and column-sparse components using the convex optimization

$$\min_{L_{(1)},C_{(1)}} \quad \|L_{(1)}\|_* + \lambda\|C_{(1)}\|_{1,2} \;\; \text{s.t. } Y = L_{(1)} + C_{(1)} \tag{2.1}$$

where $\|L_{(1)}\|_*$ denotes the nuclear norm of $L_{(1)}$ (the sum of its singular values), $\|C_{(1)}\|_{1,2}$ is the sum of the ℓ_2 norms of the columns of $C_{(1)}$, and $\lambda > 0$ is a regularization parameter. Finally, we leverage the fact that correct identification of the subspace spanned by the low-rank component of ΦM facilitates (simple) inference of the column outliers.

We analyze two variants of this overall approach. The first (depicted as Algorithm 1) is based on the notion that, contingent on correct identification of the subspace spanned by the low-rank component of ΦM, we may effectively transform the overall outlier identification problem into a compressed sensing problem, using a carefully-designed linear measurement operator whose net effect is to *(i)* reduce the overall $n_1 \times n_2$ matrix to a $1 \times n_2$ vector whose elements are (nominally) nonzero only at the locations of the outlier columns, and *(ii)* compressively sample the resulting vector. This reduction enables us to employ well-known theoretical results (e.g., [20]) to facilitate our overall analysis. We call this approach Adaptive Compressive Outlier Sensing (ACOS).

The second approach, which we call Simplified ACOS (SACOS) and summarize as Algorithm 2, foregoes the additional dimensionality reduction in the second step and identifies as outliers those columns of ΦM having a nonzero component orthogonal to the subspace spanned by the low-rank component of ΦM. The simplified approach has a (perhaps significantly) higher sample complexity than ACOS, but (as we will see in Section 2.3) benefits from an ability to identify a larger number of outlier columns relative to the ACOS method. In effect, this provides a trade-off between detection

Algorithm 1 Adaptive Compressive Outlier Sensing (ACOS)

```
Assume: $M \in \mathbb{R}^{n_1 \times n_2}$
Input: Column sampling Bernoulli parameter $\gamma \in [0,1]$, regularization parameter $\lambda >$
  $0$, Measurement matrices $\Phi \in \mathbb{R}^{m \times n_1}$, $A \in \mathbb{R}^{p \times n_2}$, measurement vector $\phi \in \mathbb{R}^{1 \times m}$
Initalize: Column sampling matrix $S = I_{:,\mathcal{S}}$, where
  $\mathcal{S} = \{i : s_i = 1\}$ with $\{s_i\}_{i \in [n_2]}$ i.i.d. Bernoulli($\gamma$)
Step 1
  Collect Measurements: $Y_{(1)} = \Phi M S$
  Solve: $\{\widehat{L}_{(1)}, \widehat{C}_{(1)}\} = \text{argmin}_{L_{(1)}, C_{(1)}} \|L_{(1)}\|_* + \lambda \|C_{(1)}\|_{1,2}$   s.t.   $Y_{(1)} = L_{(1)} + C_{(1)}$
  Let: $\widehat{\mathcal{L}}_{(1)}$ be the linear subspace spanned by col's of $\widehat{L}_{(1)}$
Step 2
  Compute: $P_{\widehat{\mathcal{L}}_{(1)}}$, the orthogonal projector onto $\widehat{\mathcal{L}}_{(1)}$
  Set: $P_{\widehat{\mathcal{L}}^{\perp}_{(1)}} \triangleq I - P_{\widehat{\mathcal{L}}_{(1)}}$
  Collect Measurements: $y_{(2)} = \phi \, P_{\widehat{\mathcal{L}}^{\perp}_{(1)}} \Phi M A^T$
  Solve: $\widehat{c} = \text{argmin}_c \ \ \|c\|_1$  s.t. $y_{(2)} = c A^T$
Output: $\widehat{\mathcal{I}}_C = \{i : \widehat{\mathrm{c}}_i \neq 0\}$
```

performance and sample complexity for the two methods. We also investigate extensions to settings where the data are group-structured, noisy, or possibly incomplete [21–24].

Related Work. Our effort here leverages results from Compressive Sensing (CS), where parsimony in the object or signal being acquired, in the form of *sparsity*, is exploited to devise efficient procedures for acquiring and reconstructing high-dimensional objects [5–7, 20]. The sequential and adaptive nature of our proposed approach is inspired by numerous recent works in the burgeoning area of adaptive sensing and adaptive CS (see, for example, [25–42] as well as the summary article [43] and the references therein). The column subsampling inherent in the first step of our approaches is also reminiscent of the data partitioning strategy of the *divide-and-conquer* parallelization approach of [44] (though our approach only utilizes one small partition of the data for the first inference step).

Our efforts here utilize a generalization of the notion of sparsity, formalized in terms of a low-rank plus outlier matrix model. In this sense, our efforts here are related to earlier works in Robust PCA [45, 46] that seek to identify low-rank matrices in the presence of sparse impulsive outliers, and their extensions to settings where the outliers present as entire columns of an otherwise low-rank matrix [19, 47–54]. In fact, the

Algorithm 2 Simplified ACOS (SACOS)

```
Assume: $M \in \mathbb{R}^{n_1 \times n_2}$
Input: Column sampling Bernoulli parameter $\gamma \in [0, 1]$, regularization parameter $\lambda > 0$, Measurement matrix $\Phi \in \mathbb{R}^{m \times n_1}$
Initalize: Column sampling matrix $S = I_{:,\mathcal{S}}$, where
  $\mathcal{S} = \{i : s_i = 1\}$ with $\{s_i\}_{i \in [n_2]}$ i.i.d. Bernoulli($\gamma$)
Step 1
  Collect Measurements: $Y = \Phi M$
  Form: $Y_{(1)} = YS$
  Solve: $\{\widehat{L}_{(1)}, \widehat{C}_{(1)}\} = \text{argmin}_{L_{(1)}, C_{(1)}} \|L_{(1)}\|_* + \lambda \|C_{(1)}\|_{1,2} \quad \text{s.t.} \quad Y_{(1)} = L_{(1)} + C_{(1)}$
  Let: $\widehat{\mathcal{L}}_{(1)}$ be the linear subspace spanned by col's of $\widehat{L}_{(1)}$
Step 2
  Compute: $P_{\widehat{\mathcal{L}}_{(1)}}$, the orthogonal projector onto $\widehat{\mathcal{L}}_{(1)}$
  Set: $P_{\widehat{\mathcal{L}}^{\perp}_{(1)}} \triangleq I - P_{\widehat{\mathcal{L}}_{(1)}}$
  Form: $Y_{(2)} = P_{\widehat{\mathcal{L}}^{\perp}_{(1)}} Y$
  Form: $\widehat{c}$ with $\widehat{c}_i = \|(Y_{(2)})_{:,i}\|_2$ for all $i \in [n_2]$
Output: $\widehat{\mathcal{I}}_C = \{i : \widehat{c}_i \neq 0\}$
```

computational approach and theoretical analysis of the first step of our approach make direct utilization of the results of [19].

We also note a related work [55], which seeks to decompose matrices exhibiting some simple structure (e.g., low-rank plus sparse, etc.) into their constituent components from compressive observations. Our work differs from that approach in both the measurement model and scope. Namely, our linear measurements are formed via simple row and column operations on the matrix and our overall approach is adaptive in nature, in contrast to the non-adaptive "global" compressive measurements acquired in [55], each of which is essentially a linear combination of all of the matrix entries. Further, the goal of [55] was to exactly recover the constituent components, while our aim is only to identify the locations of the outliers. We discuss some further connections with [55] in Section 2.5.

A component of our numerical evaluation here entails assessing the performance of our approach in a stylized image processing task of saliency map estimation. We note that several recent works have utilized techniques from the sparse representation literature in salient region identification, and in compressive imaging scenarios. A seminal effort in this direction was [56], which proposed a model for feature identification via the human visual cortex based on parsimonious (sparse) representations. More recently, [57] applied techniques from *dictionary learning* [56, 58] and low-rank-plus-sparse matrix decomposition [45, 46] in a procedure to identify salient regions of an image from (uncompressed) measurements. Similar sparse representation techniques for salient feature identification were also examined in [59]. An adaptive compressive imaging procedure driven by a saliency "map" obtained via low-resolution discrete cosine transform (DCT) measurements was demonstrated in [60]. Here, unlike in [57,59], we consider salient feature identification based on compressive samples, and while our approach is similar in spirit to the problem examined in [60], here we provide theoretical guarantees for the performance of our approach. Finally, we note several recent works [61,62] that propose methods for identifying salient elements in a data set using compressive samples.

2.2 Main Results

2.2.1 Problem Statement

Our specific problem of interest here may be formalized as follows. We suppose $M \in \mathbb{R}^{n_1 \times n_2}$ admits a decomposition of the form $M = L + C$, where L is a low-rank matrix having rank at most r, and C is a matrix having some $k \leq n_2$ nonzero columns that we will interpret as "outliers" from L, in the sense that they do not lie (entirely) within the span of the columns of L. Formally, let $\mathcal{L}$ denote the linear subspace of $\mathbb{R}^{n_1}$ spanned by the columns of L (and having dimension at most r), denote its orthogonal complement in $\mathbb{R}^{n_1}$ by $\mathcal{L}^{\perp}$, and let $P_{\mathcal{L}}$ and $P_{\mathcal{L}^{\perp}}$ denote the orthogonal projection operators onto $\mathcal{L}$ and $\mathcal{L}^{\perp}$, respectively. We assume that the nonzero columns of C are indexed by a set $\mathcal{I}_C$ of cardinality k, and that $i \in \mathcal{I}_C$ if and only if $\|P_{\mathcal{L}^{\perp}} C_{:,i}\|_2 > 0$. Aside from this assumption, the elements of the nonzero columns of C may be *arbitrary*.

Notice that *without loss of generality*, we may assume that the columns of L are zero at the locations corresponding to the nonzero columns of C (since those columns of L can essentially be aggregated into the nonzero columns of C, and the resulting column will still be an outlier according to our criteria above). We adopt that model here, and assume L has a total of n_L nonzero columns[1] with $n_L \leq n_2 - k$, which allows for the

[1] As we will see, the conditions under which our column subsampling in Step 1 succeeds will depend

case where some $n_2 - (n_L + k) \geq 0$ columns of M itself may be zero.

Given this setup, our problem of interest here may be stated concisely – our aim is to identify the set $\mathcal{I}_C$ containing the locations of the outlier columns.

2.2.2 Assumptions

It is well-known in the robust PCA literature that separation of low-rank and sparse matrices from observations of their sum may not be a well-posed task – for example, matrices having only a single nonzero element are simultaneously low rank, sparse, column-sparse, row-sparse, etc. To overcome these types of identifiability issues, it is common to assume that the linear subspace spanned by the rows and/or columns of the low-rank matrix be "incoherent" with the canonical basis (see, e.g., [19, 45–47]). Incoherence assumptions are also common in matrix completion analyses; see, e.g., [63].

In a similar vein, since our aim here is to identify column outliers from an otherwise low-rank matrix we seek conditions that make the low-rank and outlier components distinguishable. To this end, we assume an incoherence condition on the row space of the low-rank component L. We formalize this notion via the following definition from [19].

Definition 1 (Column Incoherence Property)**.** Let $L \in \mathbb{R}^{n_1 \times n_2}$ be a rank r matrix with at most $n_L \leq n_2$ nonzero columns, and compact singular value decomposition (SVD) $L = U\Sigma V^*$, where U is $n_1 \times r$, Σ is $r \times r$, and V is $n_2 \times r$. The matrix L is said to satisfy the *column incoherence property* with parameter μ_L if

$$\max_i \|V^* e_i\|_2^2 \leq \mu_L \frac{r}{n_L},$$

where $\{e_i\}$ are basis vectors of the canonical basis for $\mathbb{R}^{n_2}$.

Note that $\mu_L \in [1, n_L/r]$. The lower limit is achieved, for instance, when all elements of V^* have the same amplitude, while the upper limit is achieved, for instance, if any one element of V^* is equal to 1. For our purposes here, an undesirable scenario occurs when one of the directions in the span of the columns of L is defined by only a single vector

on the number of *nonzero* columns in the low-rank component, since any all-zero columns are essentially non-informative for learning the low-rank subspace. Thus, we make the distinction between n_2 and n_L explicit throughout.

of L, so that distinguishing that vector from a column outlier becomes ambiguous. In those cases we have that $\max_i \|V^* e_i\|_2^2 = 1$. Thus, assuming that L satisfies the column incoherence property with small μ_L is sufficient to prevent such undesirable scenarios.

With this, we may state our assumptions concisely, as follows: we assume that the components L and C of the matrix $M = L + C$ satisfy the following *structural conditions*:

(**c1**) $\operatorname{rank}(L) = r$,

(**c2**) L has n_L nonzero columns,

(**c3**) L satisfies the *column incoherence property* with parameter μ_L, and

(**c4**) $|\mathcal{I}_C| = k$, where $\mathcal{I}_C = \{i : \|P_{\mathcal{L}^\perp} C_{:,i}\|_2 > 0, L_{:,i} = 0\}$.

2.2.3 Recovery Guarantees and Implications

Our main results identify conditions under which the procedures outlined in Algorithm 1 and Algorithm 2 succeed. Our particular focus is on the case where the measurement matrices are random, and satisfy the following property.

Definition 2 (Distributional Johnson-Lindenstrauss (JL) Property)**.** An $m \times n$ matrix Φ is said to satisfy the *distributional JL property* if for any fixed $v \in \mathbb{R}^n$ and any $\epsilon \in (0, 1)$,

$$\Pr\left(\left| \, \|\Phi v\|_2^2 - \|v\|_2^2 \, \right| \geq \epsilon \|v\|_2^2 \right) \leq 2e^{-m f(\epsilon)}, \tag{2.2}$$

where $f(\epsilon) > 0$ is a constant depending only on ϵ that is specific to the distribution of Φ.

Random matrices satisfying the distributional JL property are those that preserve the length of any fixed vector to within a multiplicative factor of $(1 \pm \epsilon)$ with probability at least $1 - 2e^{-mf(\epsilon)}$. By a simple union bounding argument, such matrices can be shown to approximately preserve the lengths of a finite collection of vectors, all vectors in a linear subspace, all vectors in a union of subspaces, etc., provided the number of rows is sufficiently large. As noted in [64], for many randomly constructed and appropriately normalized Φ, (e.g., such that entries of Φ are i.i.d. zero-mean Gaussian, or are drawn as an ensemble from any subgaussian distribution), $f(\epsilon)$ is quadratic[2] in ϵ as $\epsilon \to 0$.

[2] It was shown in [65], for example, that $f(\epsilon) = \epsilon^2/4 - \epsilon^3/6$ for matrices whose elements are appropriately normalized Gaussian or symmetric Bernoulli random variables.

This general framework also allows us to directly utilize other specially constructed *fast* or *sparse* JL transforms [66, 67].

With this, we are in position to formulate our first main result. We state it here as a theorem; its proof appears in Section 7.1.1.

Theorem 1 (Accurate Recovery via ACOS)**.** Suppose $M = L + C$, where the components L and C satisfy the structural conditions (**c1**)-(**c4**) with

$$k \leq \frac{1}{3(1 + 121\ r\mu_L)}\ n_2. \tag{2.3}$$

For any $\delta \in (0,1)$, if the column subsampling parameter γ satisfies

$$\gamma \geq \max\left\{\frac{200\log(\frac{6}{\delta})}{n_L}, \frac{600(1+121r\mu_L)\log(\frac{6}{\delta})}{n_2}, \frac{10r\mu_L\log(\frac{6r}{\delta})}{n_L}\right\}, \tag{2.4}$$

the measurement matrices are each drawn from any distribution satisfying (2.2) with

$$m \geq \frac{5(r+1) + \log(k) + \log(2/\delta)}{f(1/4)} \tag{2.5}$$

and

$$p \geq \frac{11k + 2k\log(n_2/k) + \log(2/\delta)}{f(1/4)}, \tag{2.6}$$

the elements of ϕ are i.i.d. realizations of any continuous random variable, and for any upper bound k_{ub} of k the regularization parameter is set to $\lambda = \frac{3}{7\sqrt{k_{\text{ub}}}}$, then the following hold simultaneously with probability at least $1 - 3\delta$:

- the ACOS procedure in Algorithm 1 correctly identifies the salient columns of C (i.e., $\widehat{\mathcal{I}}_C = \mathcal{I}_C$), and
- the total number of measurements collected is no greater than $\left(\frac{3}{2}\right)\gamma m n_2 + p$.

It is interesting to compare this result with that of [19], which established that the Outlier Pursuit procedure (2.1) succeeds in recovering the true low-rank subspace and locations of the outlier columns provided M satisfy conditions analogous to (**c1**)-(**c4**) with $k \leq n_2/(1 + (121/9)\ r\mu_L)$. The sufficient condition (2.3) on the number of recoverable outliers that we identify for the ACOS procedure differs from the condition identified in that work by only constant factors. Further, the number of identifiable outliers could be as large as a fixed fraction of n_2 when both the rank r and coherence parameter μ_L are small.

It is also interesting to note the sample complexity improvements that are achievable using the ACOS procedure. Namely, it follows directly from our analysis that for appropriate choice of the parameters γ, m, and p the ACOS algorithm correctly identifies the salient columns of C with high probability from relatively few observations, comprising only a fraction of the measurements required by other comparable (non-compressive) procedures [19] that produce the same correct salient support estimate but operate directly on the full $(n_1 \times n_2)$ matrix M. Specifically, our analysis shows that the ACOS approach succeeds with high probability with an effective sampling rate of $\frac{\text{num obs}}{n_1 n_2} = \mathcal{O}\left(\max\left\{ \frac{(r+\log k)(n_2/n_L)\mu_L r \log r}{n_1 n_2}, \frac{(r+\log k)}{n_1} \right\} + \frac{k \log(n_2/k)}{n_1 n_2} \right)$, which may be small when r and k are each small relative to the problem dimensions (and $n_L \sim n_2$, so that L does not have a large number of zero columns outside of $\mathcal{I}_C$).

Another point of comparison for our result comes from the related work [47], which addresses a different (and in a sense, more difficult) task of identifying both the column space and the set of outlier columns of a matrix $M = L + C$ from observations that take the form of samples of the elements of M. There, to deal with the fact that observations take the form of point samples of the matrix (rather than more general linear measurements as here), the authors of [47] assume that L also satisfy a row incoherence property in addition to a column incoherence property, and show that in this setting that the column space of L and set of nonzero columns of C may be recovered from only $\mathcal{O}\left(n_2 r^2 \mu^2 \log(n_2)\right)$ observations via a convex optimization, where $\mu \in [1, n_1/r]$ is the row incoherence parameter. Normalizing this sample complexity by $n_1 n_2$ facilitates comparison with our result above; we see that the sufficient conditions for the sample complexity of our approach are smaller than for the approach of [47] by a factor of at least $1/r$, and, our approach does not require the row incoherence assumption. We provide some additional, experimental comparisons between our ACOS method and the RMC method in Section 2.3.

We may also obtain performance guarantees for Algorithm 2 (in effect, using a simplified version of the analysis used to establish Theorem 1). This yields the following corollary.

Corollary 1 (Accurate Recovery via SACOS)**.** Suppose $M = L + C$, where the components L and C satisfy the structural conditions (**c1**)-(**c4**) with k as in (2.3). Let the measurement matrix Φ be drawn from a distribution satisfying (2.2), and assume (2.4) and (2.5) hold. If for any upper bound k_{ub} of k the regularization parameter is set to $\lambda = \frac{3}{7\sqrt{k_{\text{ub}}}}$, then the following hold simultaneously with probability at least $1 - 2\delta$:

- the SACOS procedure in Algorithm 2 correctly identifies the salient columns of C (i.e., $\widehat{\mathcal{I}}_C = \mathcal{I}_C$), and
- the total number of measurements collected is no greater than mn_2.

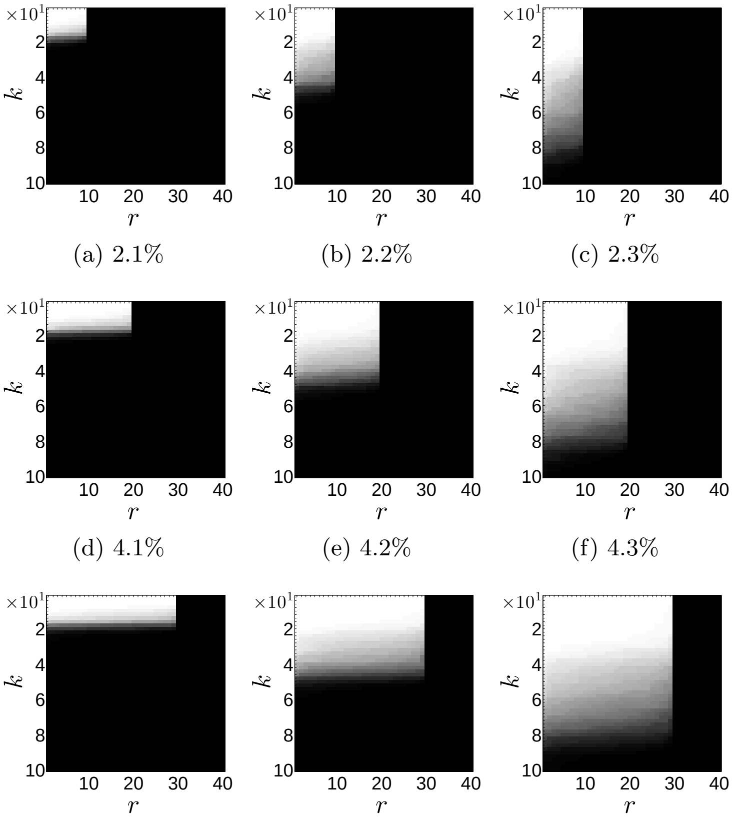
We leave the proof (which is straightforward, using the lemmata in the following section) to the interested reader.

2.3 Experimental Evaluation

In this section we provide a comprehensive experimental evaluation of the performance of our approaches for both synthetically generated and real data, the latter motivated by a stylized application of saliency map estimation in an image processing task. We compare our methods with the Outlier Pursuit (OP) approach of [19] and the Robust Matrix Completion (RMC) approach of [47], each of which employs a convex optimization to identify both the subspace in which the columns of the low rank matrix lie, and the locations of the nonzero columns in the outlier matrix. We implement the RMC method using an accelerated approximate alternating direction method of multipliers (ADMM) method inspired by [68] (as well as [19, 69]). We implement the OP methods (as well as the intermediate execution of the OP-like optimization in Step 1 of our approach) using the procedure in [47]. We implement the ℓ_1-regularized estimation in Step 2 of our procedure by casting it as a LASSO problem and using an accelerated proximal gradient method [69].

2.3.1 Synthetic Data

We experiment on synthetically generated $n_1 \times n_2$ matrices M, with $n_1 = 100$ and $n_2 = 1000$, formed as follows. For a specified rank r and number of outliers k, we let the number of nonzero columns of L be $n_L = n_2 - k$, generate two random matrices



(a) 2.1%

(b) 2.2%

(c) 2.3%

(d) 4.1%

(e) 4.2%

(f) 4.3%

(g) 6.1%

(h) 6.2%

(i) 6.3%

Figure 2.1: Outlier recovery phase transitions plots for ACOS (white regions correspond to successful recovery). Each row of the figure corresponds to a different level of compression of rows of M, where $m = 0.1n_1, 0.2n_1$ and $0.3n_1$, respectively, from top to bottom. Each column corresponds to a different level of compression of rows of M in Step 2 of Algorithm 1, with $p = 0.1n_2, 0.2n_2$ and $0.3n_2$, respectively, from left to right. The fraction of observations obtained (as a percentage, relative to the full dimension) is provided as a caption below each figure. As expected, increasing m (top to bottom) facilitates accurate estimation for increasing rank r of L, while increasing p (left to right) allows for recovery of increasing numbers k of outlier columns.

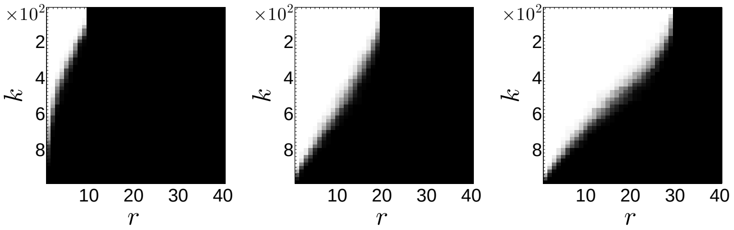
$U \in \mathbb{R}^{n_1 \times r}$ and $V \in \mathbb{R}^{n_L \times r}$ with i.i.d. $\mathcal{N}(0,1)$ entries, and we take $L = [UV^T \; 0_{n_1 \times k}]$. We generate the outlier matrix C as $C = [0_{n_1 \times n_L} \; W]$ where $W \in \mathbb{R}^{n_1 \times k}$ has i.i.d. $\mathcal{N}(0,r)$ entries (which are also independent of entries of U and V). Then, we set $M = L + C$. Notice that the outlier vector elements have been scaled, so that all columns of M have the same squared ℓ_2 norm, in expectation. In all experiments we generate ϕ, Φ, and A

with i.i.d. zero-mean Gaussian entries.

Our first experiment investigates the "phase transition" behavior of our ACOS approach; our experimental setting is as follows. First, we set the average sampling rate by fixing the column downsampling fraction $\gamma = 0.2$, and choosing a row sampling parameter $m \in \{0.1n_1, 0.2n_1, 0.3n_1\}$ and column sampling parameter $p \in \{0.1n_2, 0.2n_2, 0.3n_2\}$. Then, for each (r, k) pair with $r \in \{1, 2, 3, \ldots, 40\}$ and $k \in \{2, 4, 6, \ldots, 100\}$ we generate a synthetic matrix M as above, and for each of 3 different values of the regularization parameter $\lambda \in \{0.3, 0.4, 0.5\}$ we perform 100 trials of Algorithm 1 recording in each whether the recovery approach succeeded[3] in identifying the locations of the true outliers for that value of λ, and associate to each (r, k) pair the (empirical) average success rate. Then, at each (r, k) point examined we identify the point-wise maximum of the average success rates for the 3 different values of λ; in this way, we assess whether recovery for that (r, k) is achievable by our method for the specified sampling regime for *some* choice of regularization parameters. The results in Figure 2.1 depict the outcome of this experiment for the 9 different sampling regimes examined. For easy comparison, we provide the average sampling rate as fraction of observations obtained (relative to the full matrix dimension) in the caption in each figure.

The results of this experiment provide an interesting, and somewhat intuitive, illustration of the efficacy of our approach. Namely, we see that increasing the parameter m of the matrix Φ in Step 1 of our algorithm while keeping the other sampling parameters fixed (i.e., moving from top to bottom in any one column) facilitates accurate recovery for increasing ranks r of the matrix L. Similarly, increasing the parameter p of the matrix A in Step 2 of our algorithm while keeping the other sampling parameters fixed (i.e., moving from left to right in any one row) facilitates accurate recovery for an increasing number k of outlier columns. Overall, our approach can successfully recover the locations of the outliers for non-trivial regimes of r and k using very few measurements – see, for instance, panel (i), where ~ 30 outlier columns can be accurately identified in the presence of a rank ~ 30 background using an effective sampling rate of only $\sim 6.3\%$.

[3] We solve the optimization associated with Step 2 of our approach as a LASSO problem, with 10 different choices of regularization parameter $\mu \in (0, 1)$. We deem any trial a success if for at least one value of μ, there exists a threshold $\tau > 0$ such that $\min_{i \in \mathcal{I}_C} |\widehat{c}_i(\mu)| > \tau > \max_{j \notin \mathcal{I}_C} |\widehat{c}_j(\mu)|$ for the estimate $\widehat{c}(\mu)$ produced in Step 2. An analogous threshold-based methodology was employed to assess the outlier detection performance of the Outlier Pursuit approach in [19].

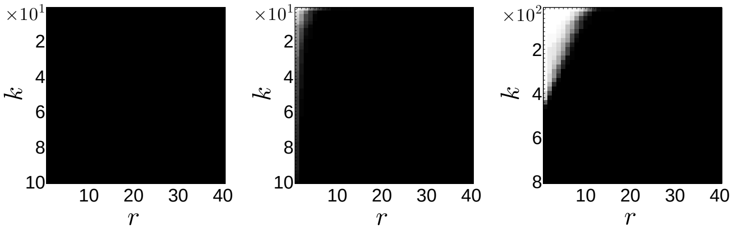


(a) 10%

(b) 20%

(c) 30%

Figure 2.2: Outlier recovery phase transitions plots for SACOS (white regions correspond to successful recovery). The row sampling parameters are $m = 0.1n_1$, $0.2n_1$, and $0.3n_1$ respectively, from left to right. Increasing m in SACOS enables accurate estimation for larger rank and increasing numbers of outlier columns. The sampling rate is provided below each plot.



(a) 5% (RMC)

(b) 10% (RMC)

(c) 20% (RMC)

Figure 2.3: Outlier recovery phase transitions plots for RMC. The average sampling rates are 5%, 10% and 20%, from left to right. Note that the vertical (k) scale in panels (a) and (b) matches that of Figure 2.1, while the scale on panel (c) matches that of Figure 2.2. Further, comparing panels (a) and (b) here with Figure 2.1 shows that ACOS outperforms RMC at low sampling rates, while comparing panels (b) and (c) here with panels (a) and (b) of Figure 2.2 shows that SACOS yields correct outlier identification for a larger portion of the parameter space than RMC for the same average sampling rates.

We adopt a similar methodology to evaluate the Simplified ACOS approach, except that we set $k \in \{20, 40, 60, \ldots, 980\}$ (and the parameter p is no longer applicable, since there is no additional compression in Step 2 for this method). The results are shown in Figure 2.2. As noted above the SACOS approach has a higher average sampling rate than ACOS for the same m, but the results show this facilitates recovery of much larger numbers k of outlier columns (notice the difference in the vertical scales in Figures 2.1 and 2.2). Overall, we may view ACOS and SACOS as complementary; when the number

k of outlier columns is relatively small and low sampling ratio $\frac{\text{num obs}}{n_1 n_2}$ is a primary focus, ACOS may be preferred, while if the number k of outlier columns is relatively large, SACOS is more favorable (at the cost of increased sample complexity).

We also compute phase transition curves for RMC using a similar methodology to that described above. The results are provided in Figure 2.3 . We observe[4] that RMC approach is viable for identifying the outliers from subsampled data provided the sampling rate exceeds about 10%, but even then only for small values of the rank r. As alluded in the discussion in previous sections, the relative difference in performance is likely due in large part to the difference in the observation models between the two approaches – the RMC approach is inherently operating in the presence of "missing data" (a difficult scenario!) while our approach permits us to observe linear combinations of any row or column of the entire matrix (i.e., we are allowed to "see" each entry of the matrix, albeit not necessarily individually, throughout our approach).

2.3.2 Real Data

We also evaluate the performance of our proposed methods on real data in the context of a stylized image processing task that arises in many computer vision and automated surveillance – that of identifying the "saliency map" of an image. For this, we use images from the MSRA Salient Object Database [12][5].

As discussed above, our approach here is based on representing each test image as a collection of (vectorized) non-overlapping image patches. We transform each (color) test image to gray scale, decompose it into non-overlapping 10 × 10-pixel patches, vectorize each patch into a 100×1 column vector, and assemble the column vectors into a matrix. Most of the images in the database are of the size 300 × 400 (or 400 × 300), which here yields matrices of size 100 × 1200, corresponding to 1200 patches. Notice that we only used gray scale values of image as the input feature rather than any high-level images feature – this facilitates the use of our approach, which is based on collecting linear measurements of the data (e.g., using a spatial light modulator, or an architecture like the *single pixel camera* [70]).

[4] Our evaluation of RMC here agrees qualitatively with results in [47], where sampling rates around 10% yielded successful recovery for small r.

[5] Available online at http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm.

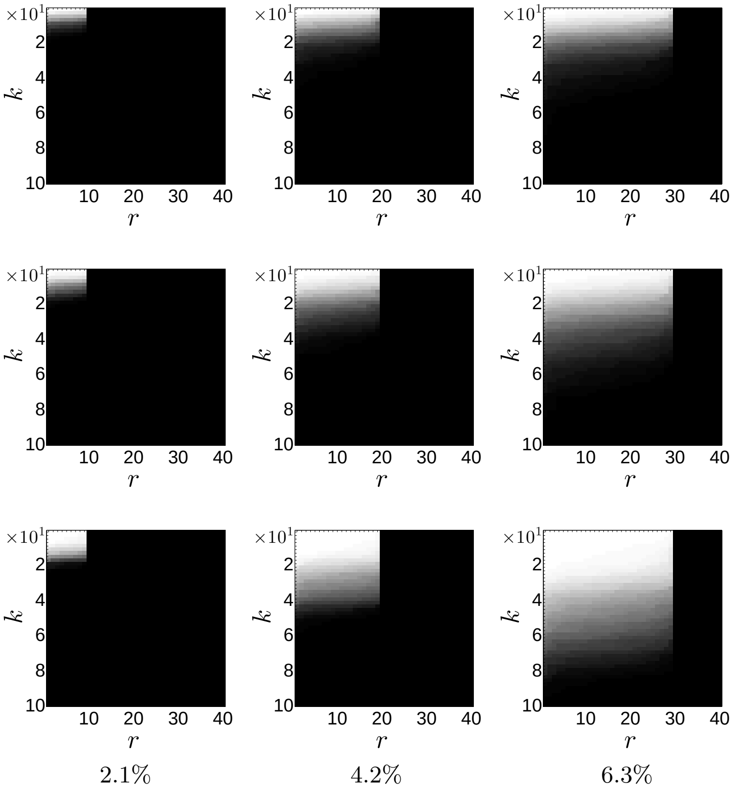


Figure 2.4: Outlier recovery phase transitions plots for ACOS for noisy settings (white regions correspond to successful recovery). Rows correspond to $\sigma = 0.001$, 0.0005 and 0.0001 respectively, from top to bottom; columns correspond to the settings $m = 0.1n_1,\ p = 0.1n_2$; $m = 0.2n_1,\ p = 0.2n_2$; and $m = 0.3n_1,\ p = 0.3n_2$ respectively, from left to right. The fraction of observations obtained is provided below each column. As in Figure 2.1, larger m and p promote accurate recovery for increasing rank r and numbers k of outlier columns. Here, however, increasing noise variance degrades the estimation results, especially with respect to the number k of outliers that can be accurately identified.

Here, our experimental approach is (somewhat necessarily) a bit more heuristic than for the synthetic data experiments above, due in large part to the fact that the data here may not adhere exactly to the low-rank plus outlier model. To compensate for this, we augment Step 1 of Algorithm 1 and Algorithm 2 with an additional "rank reduction" step, where we further reduce the dimension of the subspace spanned by the columns of the learned $\widehat{L}_{(1)}$ by truncating its SVDs to retain the smallest number of leading singular values whose sum is at least $0.95 \times \|\widehat{L}_{(1)}\|_*$. Further, we generalize Step 2 of each procedure by declaring an image patch to be salient when its (residual) column norm is sufficiently large, rather than strictly nonzero. We used visual heuristics to determine the "best" outputs for Step 2 of each method, selecting LASSO parameters (for ACOS) or thresholds (for SACOS) in order to qualitatively trade off false positives with misses.

We implement our ACOS and SACOS methods using three different sampling regimes for each, with the fixed column downsampling parameter $\gamma = 0.2$ throughout. For ACOS, we examine settings where $m = 0.2n_1$, $0.1n_1$ and $0.05n_1$ with $p = 0.5n_2$, which result in average sampling rates of 4.5%, 2.5% and 1.5%, respectively. For SACOS, we examine settings where $m = 0.2n_1$, $0.05n_1$ and $0.03n_1$, resulting in average sampling rates of 20%, 5% and 3%, respectively. As before, we generate the Φ and A matrices to have i.i.d. zero-mean Gaussian entries. We compare our approaches with two "benchmarks" – the Graph-based visual saliency (GBVS) method from the computer vision literature [11] and the OP approach (both of which use the full data) – as well as with the RMC approach at sampling rates of 20% and 5%.

The results of this experiment are provided in Figure 2.5. We note first that the OP approach performs fairly well at identifying the visually salient regions in the image, essentially identifying the same salient regions as the GBVS procedure and providing evidence to validate the use of the low-rank plus outlier model for visual saliency (see also [15]). Next, comparing the results of the individual procedures, we see that the OP approach appears to uniformly give the best detection results, which is reasonable since it is using the full data as input. The RMC approach performs well at the 20% sampling rate, but its performance appears to degrade at the 5% sampling rate. The SACOS approach, on the other hand, still produces reasonably accurate results using only 3% sampling. Moreover, ACOS provides acceptable results even with 1.5-2.5% sampling rate.

We also compare implementation times of the algorithms on this saliency map estimation task. Table 2.1 provides the average execution times (and standard deviations) for each approach, evaluated over 1000 images in the MSRA database[6]. Here, we only

[6] Timing comparisons were done with `MATLAB` R2013a on an iMac with a 3.4 GHz Intel Core i7 processor, 32 GB memory, and running OS X 10.8.5.

execute each procedure for one choice of regularization parameter, and we also include the additional "rank reduction" step discussed above for the ACOS and SACOS methods. Overall, we see the ACOS approach is up to 4× faster than the GBVS method and 15× faster than the OP and RMC methods, while the SACOS approach could result overall in relative speedups of 100× over GBVS and 300× over the OP and RMC methods. Overall, our results suggest a significant improvement obtained via ACOS and SACOS for both detection consistency and timing, which may have a promising impact in a variety of salient signal detection tasks.

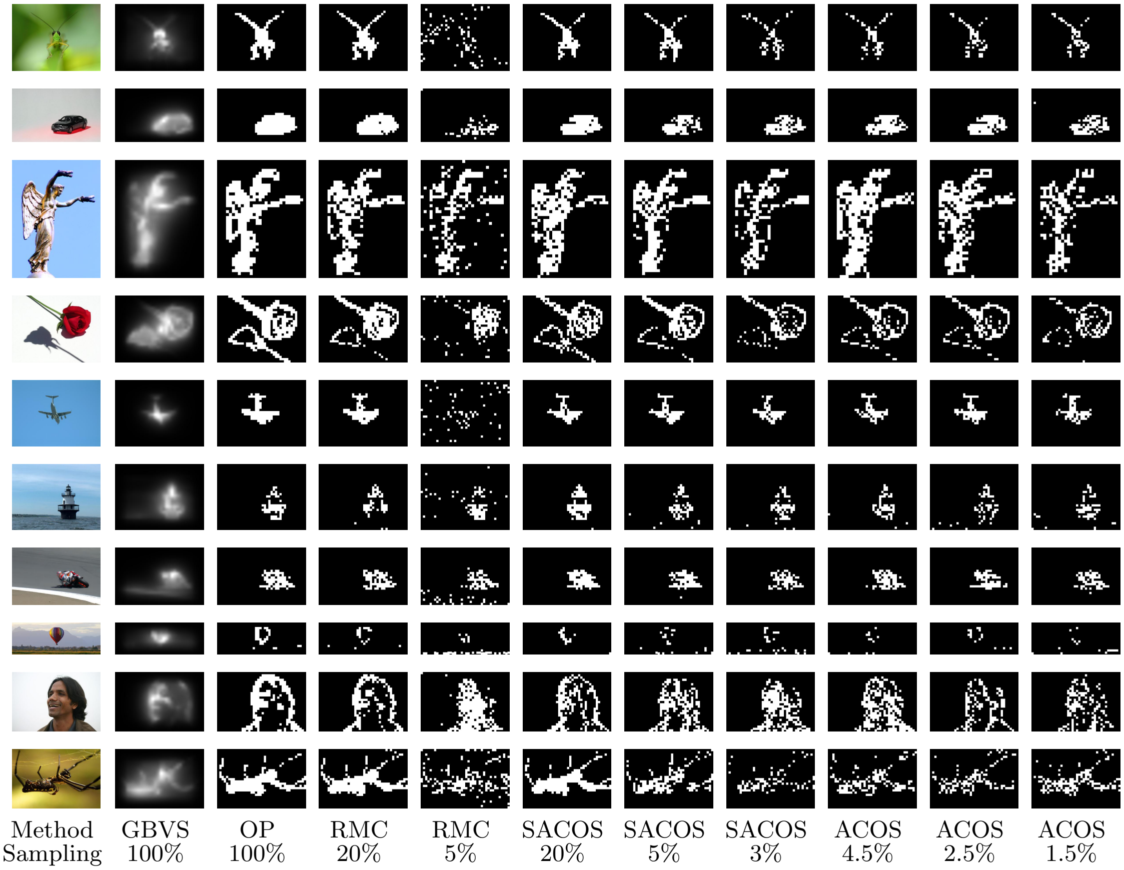


Figure 2.5: Detection results for the MSRA Salient Object Database for various methods. Our ACOS approach produces results comparable to the "full sampling" OP method using an average sampling rate below 5%. The performance of the RMC approach appears to degrade at low sampling rates.

Table 2.1: Timing analysis for detection experiments on 1000 images from MRSA Database. Each entry is the mean execution time in seconds with the standard deviation in parenthesis.

Method Sampling	GBVS 100%	OP 100%	RMC 20%	RMC 5%	SACOS 20%	SACOS 3%	ACOS 4.5%	ACOS 1.5%
Step 1	0.9926	2.9441	2.6324	2.7254	0.0538	0.0074	0.0533	0.0105
	(0.274)	(0.385)	(0.323)	(0.366)	(0.012)	(0.002)	(0.012)	(0.003)
Step 2	–	–	–	–	0.0015	0.0009	0.2010	0.2065
	–	–	–	–	(0.001)	(0.001)	(0.067)	(0.069)

2.4 Extensions

2.4.1 Group Structure

We also consider the task of locating salient group-structured features [21]. Let $G \in \mathbb{R}^{t_1 \times t_2}$ denote the original image panel, and let $F \in \mathbb{R}^{t_1 \times t_2}$ denote the output of a linear operator applied to G. In what follows, the linear operators we will consider correspond to filters that extract specific features from the image G (e.g., vertical or horizontal edge detectors, or Laplacian of Gaussian filters, which detect edges at any orientation). Now, we reinterpret F as a different matrix, by first decomposing it into n_2 patches of size $s_1 \times s_2$, vectorizing each patch into a $n_1 \times 1$ column vector where $n_1 = s_1 s_2$, and assembling the column vectors into a matrix $M \in \mathbb{R}^{n_1 \times n_2}$. We denote the collection of the locations of the outlier columns as a set $\mathcal{I}_C \subset \{1, 2, \ldots, n_2\}$. Here, we assume that $k \triangleq |\mathcal{I}_C| < n_2$; i.e., that the number of outlier columns is (perhaps much) smaller than the number of columns of M. Further, we assume that the elements of $\mathcal{I}_C$ occur in "groups," which may be formalized for our purposes as follows. Suppose that the set $\{1, 2, \ldots, n_2\}$ is partitioned into J disjoint subsets, each of size $B = n_2/J$. Then, we assume that the elements comprising $\mathcal{I}_C$ correspond to only a small number of the J subgroups of column indices.

Our aim here is to estimate $\mathcal{I}_C$ from a small number of linear observations of M. To this end, we adopt the two-step adaptive compressive sensing approach outlined in Algorithm 3. Our analysis in the next section establishes performance guarantees for this approach.

Algorithm 3 Salient Feature Detection via Group Adaptive Compressive Sensing (GACS)

```
Input: $M \in \mathbb{R}^{n_1\times n_2}$, $\gamma \in [0,1]$, $\lambda_1 > 0$, $\Phi \in \mathbb{R}^{m\times n_1}$, $A \in \mathbb{R}^{p\times n_2}$ and $\phi \in \mathbb{R}^{1\times m}$
Initalize: Column sampling matrix $S = \mathbf{I}_{:,\mathcal{S}}$, where
  $\mathcal{S} = \{i : S_i = 1\}$ with $\{S_i\}_{i\in[n_2]}$ i.i.d. Bernoulli($\gamma$)
Step 1
  Collect Measurements: $Y_{(1)} = \Phi M S$
  Solve: $\{\widehat{L}_{(1)}, \widehat{C}_{(1)}\} = \operatorname{argmin}_{L_{(1)},C_{(1)}} \|L_{(1)}\|_* + \lambda_1 \|C_{(1)}\|_{1,2}$ s.t. $Y_{(1)} = L_{(1)} + C_{(1)}$
  Let: $\widehat{\mathcal{L}}_{(1)}$ be the linear subspace spanned by col's of $\widehat{L}_{(1)}$
Step 2
  Collect Measurements: $y_{(2)} = \phi \, P_{\widehat{\mathcal{L}}^{\perp}_{(1)}} \Phi M A^T$
  Solve: $\widehat{c} = \operatorname{argmin}_c \sum_{j=1}^{J} \|c_j\|_2$ s.t. $y_{(2)} = cA^T$
Output: Saliency segmentation $F_{\mathcal{I}} \in \mathbb{R}^{t_1\times t_2}$, where the $i$-th patch is assigned with 1 if $\widehat{c}_i \neq 0$;
  otherwise 0 if $\widehat{c}_i = 0$, $i = 1, 2, \ldots, n_2$
```

The following result quantifies the performance of the GACS approach under the structural assumptions outlined above.

Theorem 2. Given any $\delta \in (0, 1/3)$, suppose $M = L + C$, where the components L and C satisfy the structural assumptions above with

$$k \leq n_2/(c_1 r \mu_L), \tag{2.7}$$

$$\gamma \geq c_2 r \mu_L \log r / n_L, \tag{2.8}$$

$$m \geq c_3 (r + \log k), \tag{2.9}$$

$$p \geq c_4 \left(k + \frac{k}{\sqrt{B}} \log \frac{n_2 - k}{B} \right), \tag{2.10}$$

where $c_1 - c_4$ are some constants depending on δ. Let $\Phi \in \mathbb{R}^{m\times n_1}$ have i.i.d. $\mathcal{N}(0, 1/m)$ entries, $A \in \mathbb{R}^{p\times n_2}$ have i.i.d. $\mathcal{N}(0, 1/p)$ entries, and let elements of $\phi \in \mathbb{R}^{1\times m}$ be i.i.d. realizations of any continuous random variable. For any upper bound k_{ub} of k the regularization parameter is set to $\lambda = \frac{3}{7\sqrt{k_{\text{ub}}}}$, then the following hold simultaneously with probability at least $1 - 3\delta$: (i) the GACS procedure in Algorithm 3 correctly identifies the salient columns of $\boldsymbol{C}$ (i.e., $\widehat{\mathcal{I}}_{\boldsymbol{C}} = \mathcal{I}_{\boldsymbol{C}}$), and (ii) the total number of measurements collected is no greater than $\left(\frac{3}{2}\right) \gamma m n_2 + p$.

Notice that the effective sampling rate here is

$$\frac{\#\text{obs}}{n_1 n_2} = \mathcal{O}\left(\frac{(r + \log k)(n_2/n_L)\mu_L r \log r}{n_1 n_2} + \frac{k + \frac{k}{\sqrt{B}} \log \frac{n_2 - k}{B}}{n_1 n_2} \right),$$

which could be much smaller than 1, indicating that accurate outlier identification may be achieved using significantly downsampled data.

Visual Experimental Results.

In this section, we provide visual experimental results to demonstrate the efficacy of the extensions that we examine here. The image database we used here (and the next section) is MSRA10K [71], where carefully manually labeled ground truth is provided for each image.

Sampling Low-Level Features: We examine several low-level features, including pixel-level features on the color (Red, Green, and Blue) planes, Laplacian of Gaussian filters (LoG), and horizontal and vertical edge-emphasizing filters (referred to here by the acronyms HE and VE, respectively). Since our overall approach allows for incorporation of any linear operator mapping of the original image G to the feature image F, each of these low-level features are valid within our overall framework (provided that the acquisition modality acquires and operates on the conventional RGB color panels).

In addition, we also investigate a variant of our approach that can be used to identify salient features in other color spaces (e.g., HSI, or the hue-saturation-intensity space). Strictly speaking, this transformation would fall outside of our specified observation model, since the mapping from RGB to the HSI space is nonlinear and thus cannot be directly incorporated into our sampling structure. To overcome this limitation here, we consider applying this nonlinear transformation on the *compressed* observed data, when seeking to locate salient features in these color spaces. More formally, suppose that M_r, M_g and M_b denote the reshaped matrices from red, green and blue panels, respectively, of an image. We can consider transforming the *compressed* measurements that result from step 1 of our approach to the HSI color space by applying the RGB to HSI transformation to the stacked color panels $\Phi M_r S$, $\Phi M_g S$ and $\Phi M_b S$. In what follows, we refer to this approach as “Stacked HSI.”

We demonstrate several saliency detection results using these different features; the results appear in Figure 2.6. Overall, we observe (as expected) that different low-level features are receptive to different types of features. What is interesting here is that these features are accurately unveiled from the under sampled data resulting from our approach.

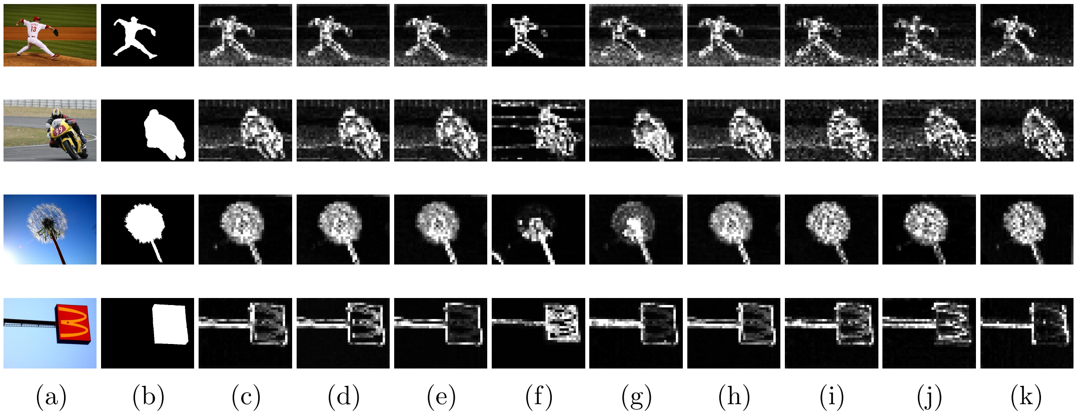
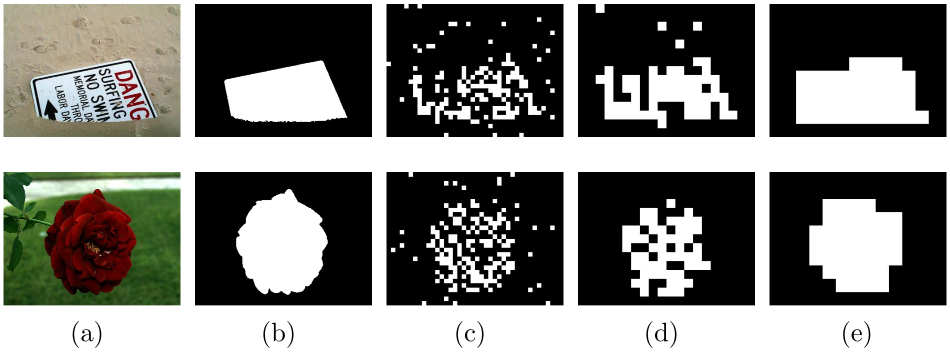


Figure 2.6: Gray scale saliency map estimate via low-level image features, including RGB (c)-(e), stacked HSI (f)-(h), LoG (i), HE (j) and VE (k). Original images the their ground truth are given in (a) and (b) respectively. We set $\gamma = 0.2$, $m = 0.2n_1$ and $p = 0.5n_2$ with the sampling rate of 4.5% for $n_1 = 100$ and $n_2 = 1200$.

Grouping Effect: We provide some visual evidence of the grouping effect in Figure 2.7 using square-shaped groups in the feature space with side lengths $g = 1, 2$, and 3. Notice that $g = 1$ corresponds to no grouping effect. It is evident that grouping adjacent features does show improvement due to its robustness to background noise and lower sample demands ($p = O(k + \frac{k}{\sqrt{B}} \log \frac{n_2 - k}{B})$ may be considerably smaller than $p = O(k \log n_2)$ when B is large) compared with that without grouping.

Comparison with Existing Saliency Detection Methods: We compare our approach with other state-of-art methods, including global based (GB) [11], region contract (RC) [71], self-resemblance (SeR) [72], frequency tuned (FT) [73], low rank (LR) [15], spectral residual (SR) [74] and spatially weighted dissimilarity (SWD) [75] methods, whose saliency maps are provided in the MSRA10K database. For our approach GACS, we set $m = 0.1n_1$, which results in an average sampling rate of 2.5%. A set of selected results are shown in Figure 2.8. We observe that our results are visually comparable, or even better than the results from some state-of-art methods that extract high level features, using only a few (2.5%) linear measurements.



(a) (b) (c) (d) (e)

Figure 2.7: Detection results with the grouping effect. Column (a) - (e) are the original image, ground truth, detection result without grouping effect and with grouping effects ($g = 2$ in (d) and 3 in (e)) respectively. We set $\gamma = 0.2$, $m = 0.1n_1$ and $p = 0.5n_2$, which corresponds to the sampling rate of 2.5%.

Quantitative Experimental Results: We also provide some quantitative experimental evidence to validate our approach. For all evaluations in this section, we use non-overlapping feature patches of size $s_1 = s_2 = 10$, fix $\gamma = 0.2$, $\lambda_1 = 0.4$ and $p = 0.5\ n_2$, and solve a constrained version of the optimization in Step 2 using 100 turning parameters λ_2. For each λ_2, we obtain the saliency segmentation $F_{\mathcal{I}}$ and compare it to the ground truth to calculate $P =$ Precision, $R =$ Recall and F-measure $= \frac{(\beta^2+1)P \cdot R}{(\beta^2 P+R)}$. We follow [15, 71, 73] and set $\beta^2 = 0.3$. For each image, we choose the feature (from R, G, B, H, S, I, LoG, HE and VE) that returns the highest maximum F-measure. For all experiments, the precision and recall curves are averages over the tested set of images, and the maximum average F-measure is provided with the average precision and recall values. In practice, the image background is not exactly low-rank, so a "rank reduction" step is applied in Step 1 that retains the smallest number of leading singular values of $\hat{L}$ whose sum is at least $0.95 \times \|\hat{L}\|_*$.

Grouping: We start with an evaluation of the grouping effect. A random subset of 2,000 images from the MSRA10K database is selected as the test set and the average sampling rate is 4.5% with $m = 0.2n_1$. The group size is chosen from $g \in \{1, 2, 3, 4, 5, 6\}$. Note that $g = i$ here indicate the use of all group sizes $1 \leq g \leq i$ and we choose the g that returns the highest maximum F-measure for each image.

The plot of precision-recall curves and the average precision, recall and F-measure

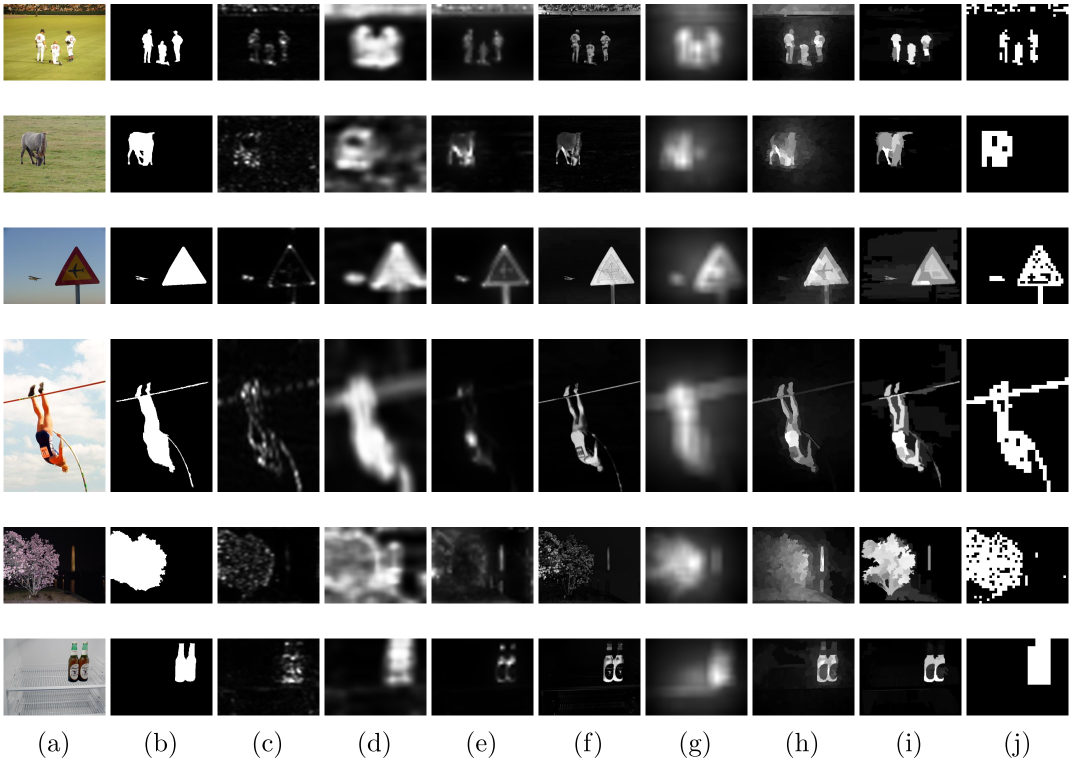
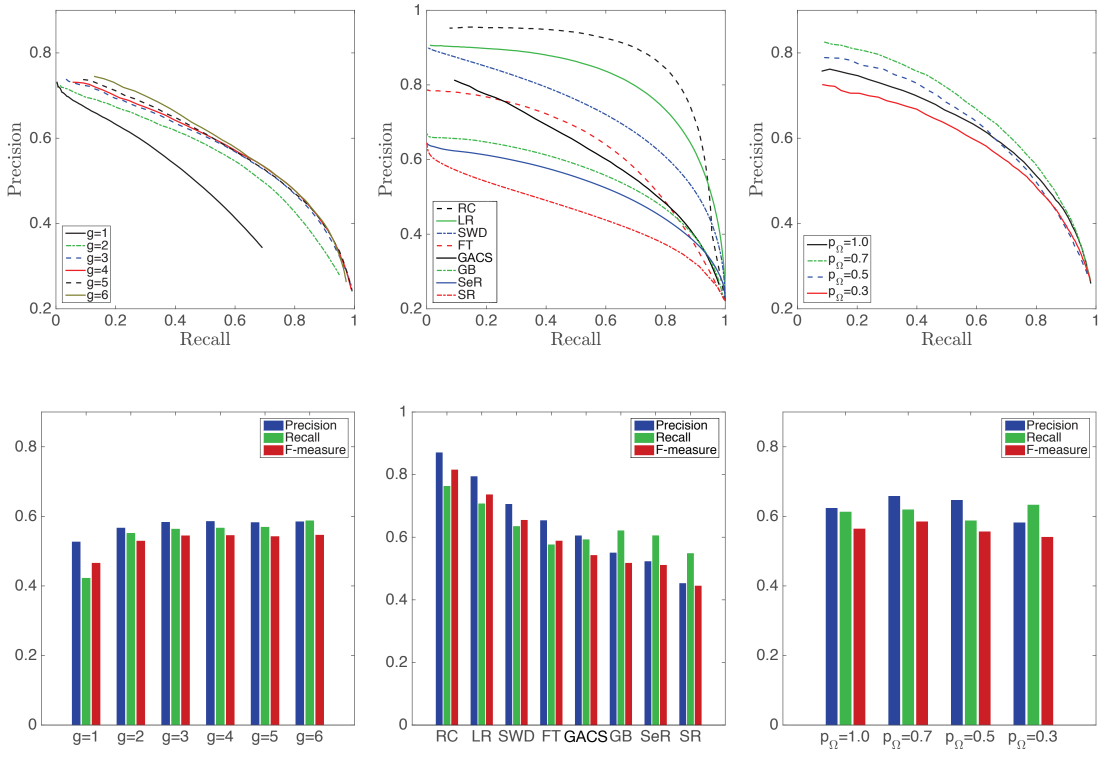


Figure 2.8: Detection results for the MSRA10K Salient Object Database for various methods. From (a) to (j), it corresponds to the original image, ground truth, SR, SeR, GB, FT, SWD, LR, RC and GACS respectively. For GACS, the results corresponds to S, G, LoG, I, R, H respectively from top to bottom.

over the choices of group sizes are provided in Figure 2.9 (a). The observation is that the grouping procedure (*i.e.* $g \geq 2$) does show a considerable enhancement. On the other hand, when $g \geq 4$, the improvement becomes minor, which coincides with our observation that $g = 2$ or 3 gives the best performance most of the time. For computational efficiency, we fix $g = 3$ for the rest our experimental investigations.

Comparisons: We run a thorough test on the entire database (*i.e.*, 10,000 images) for all 8 other approaches and ours, with the precision-recall curves and the average precision, recall and F-measure results demonstrated in Figure 2.9 (b). Though, our approach here is not giving the best performance across all tested methods, it is inspiring to see that GACS is comparable to the stat-of-art using only 2.5% measurements.



(a) (b) (c)

Figure 2.9: Comparison results on the MSRA10K database with (a) different grouping sizes ($g = 1, \ldots, 6$), (b) different methods (GB ,RC, SeR, SR, SWD, FT and LR), and (c) different levels of missing entries ($p_\Omega = 1$, 0.7, 0.5 and 0.3 respectively from left to right). The precision-recall curves (top row) and average precision, recall and F-measure (bottom row) are demonstrated.

2.4.2 Noisy Observations

We demonstrate the outlier detection performance of our approaches under the scenario when M is contaminated by unknown random noise or modeling error. Formally, we consider the setting where L and C are as above, but

$$M = L + C + N, \tag{2.11}$$

where N has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries.

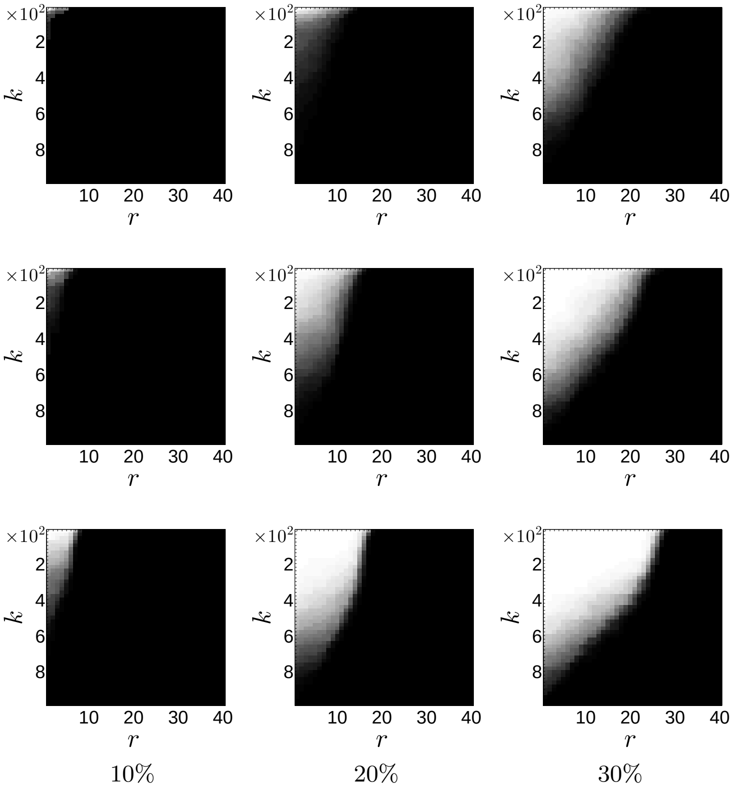


Figure 2.10: Outlier recovery phase transitions plots for SACOS for noisy settings (white regions correspond to successful recovery). Rows of the figure correspond to $\sigma = 0.03$, 0.02 and 0.01 respectively, from top to bottom; columns correspond to $m = 0.1n_1,\ 0.2n_1$, and $0.3n_1$ respectively, from left to right. The fraction of observations obtained is provided below each column. In this case, increasing noise variance results in a decrease in both the rank r as well as the number k of outliers that can be accurately identified.

We first investigate the performance of the ACOS method, following a similar experimental methodology as in Section 2.3 to generate L and C, except that now we renormalize each column of $(L + C)$ to have unit Euclidean norm (essentially to standardize the noise levels). We consider three different noise levels ($\sigma = 0.001$, 0.0005 and 0.0001), three pairs of the row sampling parameter m and the column sampling parameter p ($m = 0.1n_1,\ p = 0.1n_2$; $m = 0.2n_1,\ p = 0.2n_2$; and $m = 0.3n_1,\ p = 0.3n_2$) and for each we fix the column downsampling fraction to be $\gamma = 0.2$; the corresponding sampling ratios are 2.1%, 4.2% and 6.3%, respectively. We again perform 100 trials of Algorithm 1 and record the success frequency for each (using the same criteria for success as in Sect. 2.3.1). The results are given in Figure 2.4.

It can be observed from the results that increasing m and p promote accurate estimation of outlier column indices for increasing rank r and numbers k of outlier columns, which is exactly what we have seen in Figure 2.1 for the noiseless case. However, the presence of noise degrades the estimation performance, albeit gracefully. This is reasonable, since in Step 2 of Algorithm 1, the measurements $\mathbf{y}_2$ might be perturbed more seriously as the energy of noise increases, which results in more difficult recovery of true supports of c. Under this scenario, we will require larger p to enable better recovery of the underlying true supports.

We also evaluate the SACOS procedure in noisy settings for three choices of m ($m = 0.1n_1$, $0.2n_1$ and $0.3n_2$) and fixed column downsampling fraction $\gamma = 0.2$. Here, we again normalize columns of $(L+C)$, but consider three higher noise levels, corresponding to $\sigma = 0.03$, 0.02 and 0.01. The results are presented in Figure 2.10. Here, we again observe a graceful performance degradation with noise. Notice, however, that higher level of variances of noise can be tolerated for SACOS compared with ACOS, which is an artifact of the difference between the inference steps of the two procedures.

2.4.3 Missing Data

We also describe and demonstrate an extension of our SACOS method that is amenable to scenarios characterized by missing data. Suppose that there exists some underlying matrix M that admits a decomposition of the form $M = L + C$ with L and C as above, but we are only able to observe M at a subset of its locations. Formally, we denote by $\Omega \subseteq [n_1] \times [n_2]$ the set of indices corresponding to the available elements of M, and let $P_\Omega(\cdot)$ be the operator that masks its argument at locations not in Ω. Thus, rather than operate on M itself, we consider procedures that operate on the sampled data $P_\Omega(M)$.

In this setting, we can modify our SACOS approach so that the observations obtained in Step 1 are of the form $Y_{(1)} = \Phi P_\Omega(M)S$, where (as before) S is a column selection matrix but Φ is now a *row* subsampling matrix (i.e., it is comprised of a subset of rows of the $n_1 \times n_1$ identity matrix) containing some m rows. The key insight here is that the composite operation of sampling elements of M followed by row subsampling can be expressed in terms of a related operation of subsampling elements of a row-subsampled version of M. Specifically, we have that $\Phi P_\Omega(M) = P_{\Omega_\Phi}(\Phi M)$, where $P_{\Omega_\Phi}(\cdot)$ masks the same elements as $P_\Omega(\cdot)$ in the rows selected by Φ.

Now, given $Y_{(1)}$, we solve a variant of RMC [47]

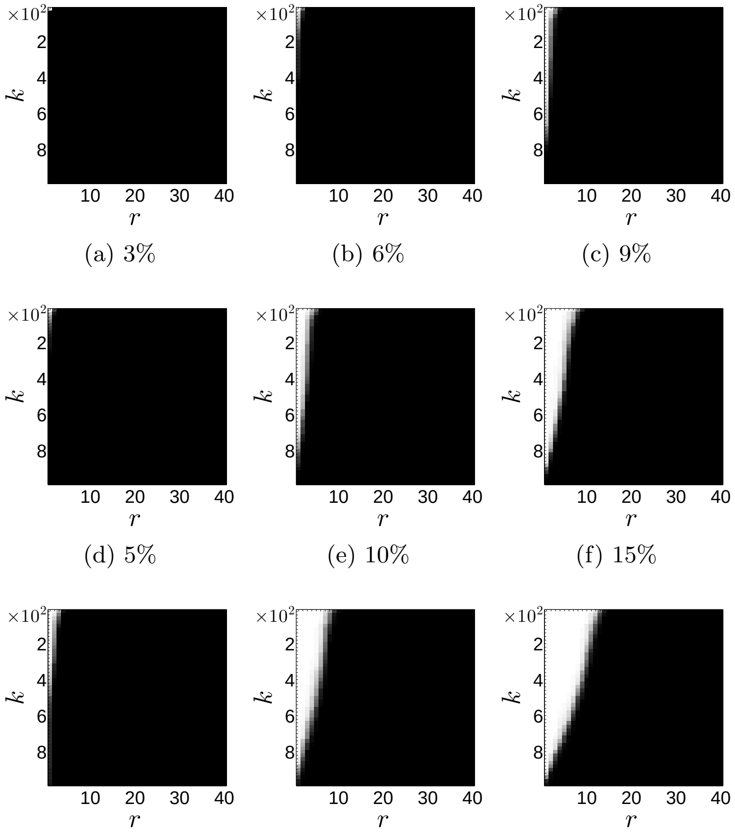
$$\{\widehat{L}_{(1)}, \widehat{C}_{(1)}\} = \underset{L_{(1)}, C_{(1)}}{\operatorname{argmin}} \|L_{(1)}\|_* + \lambda \|C_{(1)}\|_{1,2} \quad \text{s.t. } Y_{(1)} = P_{\Omega_\Phi}(L_{(1)} + C_{(1)})$$

in an initial step, identifying (as before) an estimate $\widehat{L}_{(1)}$ whose column span is an estimate of the subspace spanned by the low-rank component of ΦM.

Then (in a second step) we perform the "missing data" analog of the orthogonal projection operation on every column $j \in [n_2]$ of $\Phi P_\Omega(M)$, as follows. For each $j \in [n_2]$, we let $\mathcal{I}_j \in [m]$ denote the locations at which observations of column j of $\Phi P_\Omega(M)$ are available, and let $(\Phi P_\Omega(M))_{\mathcal{I}_j, j}$ be the sub vector of $(\Phi P_\Omega(M))_{:,j}$ containing only the elements indexed by $\mathcal{I}_j$. Similarly, let $(\widehat{L}_{(1)})_{\mathcal{I}_j,:}$ be the row submatrix of $\widehat{L}_{(1)}$ formed by retaining rows indexed by $\mathcal{I}_j$. Now, let $P_{\widehat{\mathcal{L}}_{(1)_j}}$ denote the orthogonal projection onto the subspace spanned by columns of $(\widehat{L}_{(1)})_{\mathcal{I}_j,:}$ and compute the residual energy of the j-th column as $\|(I - P_{\widehat{\mathcal{L}}_{(1)_j}})(\Phi P_\Omega(M))_{\mathcal{I}_j, j}\|_2$. Overall, the orthogonal projection for the j-th column of $\Phi P_\Omega(M)$ is only computed over the nonzero entries of that column, an approach motivated by recent efforts in subspace detection with missing data [76, 77].

We evaluate this approach empirically using the same data generation methods as above, and using an independent *Bernoulli* model to describe the subsampling operation $P_\Omega(\cdot)$ (so that each $(i, j) \in \Omega$ independently with probability p_Ω). We consider noise-free settings, fix the column subsampling parameter $\gamma = 0.2$, and examine three different row-sampling scenarios ($m = 0.1n_1$, $0.2n_1$ and $0.3n_1$) in each choosing subsets of m rows uniformly at random from the collection of all $\binom{n_1}{m}$ sets of cardinality m. The results are in Figure 2.11. Again, increasing m and p permits accurate estimation of outlier column indices for increasing rank r and numbers k of outlier columns. Further, we do observe the performance degradation as the number of missing entries of M increases.

For noisy observation and missing entry cases, we provide detailed theoretical analyses in [22].



(a) 3%

(b) 6%

(c) 9%

(d) 5%

(e) 10%

(f) 15%

(g) 7%

(h) 14%

(i) 21%

Figure 2.11: Outlier recovery phase transitions plots for a "missing data" variant of the SACOS method (white regions correspond to successful recovery). Rows correspond to available data fractions of p_Ω=0.3, 0.5 and 0.7 respectively, from top to bottom; columns corresponds to row sampling parameters $m = 0.1n_1$, $0.2n_1$, and $0.3n_1$, respectively, from left to right.

2.5 Discussion

It is illustrative here to note a key difference between our approach and more conventional compressive sensing (CS) tasks. Namely, the goal of the original CS works [5–7] and numerous follow-on efforts was to exactly recover or reconstruct a signal from compressive measurements, whereas the nature of our task here is somewhat simpler, amounting to a kind of multidimensional "support recovery" task (albeit in the presence of a low-rank "background"). Exactly recovering the low-rank and column-sparse components would be sufficient for the outlier identification task we consider here, but as our analysis shows it is not strictly necessary. This is the insight that we exploit when operating on the "compressed" data ΦM instead of the original data matrix M. Ultimately, this allows us to successfully identify the locations of the outliers *without first estimating the original (full size) low-rank matrix or the outliers themselves.* For some regimes of μ_L, r and k, we accomplish the outlier identification task using as few as $\mathcal{O}\left((r + \log k)(\mu_L r \log r) + k \log(n_2/k)\right)$ observations.

Along related lines, it is reasonable to conjecture that any procedure would require at least $r^2 + k$ measurements in order to identify k outliers from an r-dimensional linear subspace. Indeed, a necessary condition for the existence of outliers of a rank-r subspace, as we have defined them, is that the number of rows of M be at least $r + 1$. Absent any additional structural conditions on the outliers and the subspace spanned by columns of the low-rank matrix, one would need to identify a collection of r vectors that span the r-dimensional subspace containing the column vectors of the low-rank component (requiring specification of some $\mathcal{O}(r^2)$ parameters) as well as the locations of the k outliers (which would entail specifying another k parameters). In this sense, our approach may be operating near the sample complexity limit for this problem, at least for some regimes of μ_L, r and k.

It would be interesting to see whether the dimensionality reduction insight that we exploit in our approach could be leveraged in the context of the Compressive Principal Component Pursuit (Compressive PCP) of [55] in order to yield a procedure with comparable performance as ours, but which acquires only *non-adaptive* linear measurements of M. Direct implementation of that approach in our experimental setting was somewhat computationally prohibitive (e.g., simulations at a 10% sampling rate would require generation and storage of random matrices having 10^9 elements). Alternatively, it is interesting to consider implementing the Compressive PCP method not on the full data M, but on the a priori compressed data ΦM. Our Lemma 6 establishes that the row compression step preserves rank and column incoherence properties, so it is plausible that the Compressive PCP approach may succeed in recovering the components of the compressed matrix, which would suffice for the outlier identification task. We defer this investigation to a future effort.

Table 2.2: Computational complexities of outlier identification methods. The stated results assume use of an accelerated first order method for all solvers (see text for additional details).

Method	Complexity
OP	$\mathcal{O}\left(\mathrm{IT} \cdot [n_1 n_2 \cdot \min\{n_1, n_2\}]\right)$
RMC	$\mathcal{O}\left(\mathrm{IT} \cdot [n_1 n_2 \cdot \min\{n_1, n_2\}]\right)$
ACOS	$\mathcal{O}\left(\mathrm{IT}_1 \left[m(\gamma n_2) \min\{m, \gamma n_2\}\right] + \mathrm{IT}_2 \left[p n_2\right]\right)$
SACOS	$\mathcal{O}\left(\mathrm{IT}_1 \left[m(\gamma n_2) \min\{m, \gamma n_2\}\right] + m^2 n_2\right)$

We also comment briefly on the computational complexities of the methods we examined. We consider first the OP and RMC approaches, and assume that the solvers for each utilize an iterative accelerated first-order method (like those mentioned in the first part of Section 2.3). In this case, the computational complexity will be dominated by SVD steps in each iteration. Now, for an $n_1 \times n_2$ matrix the computational complexity of the SVD is $\mathcal{O}(n_1 n_2 \cdot \min\{n_1, n_2\})$; with this, and assuming some IT iterations are used, we have that the complexities of both OP and RMC scale as $\mathcal{O}\left(\mathrm{IT} \cdot [n_1 n_2 \cdot \min\{n_1, n_2\}]\right)$. By a similar analysis, we can conclude that the complexity of Step 1 of the ACOS and SACOS methods scales like $\mathcal{O}\left(\mathrm{IT}_1 \cdot [m(\gamma n_2) \cdot \min\{m, \gamma n_2\}]\right)$, where IT_1 denotes the number of iterations for the solver in Step 1. If we further assume an iterative accelerated first-order method for the LASSO in Step 2 of the ACOS approach, and that IT_2 iterations are used, then the second step of the ACOS approach would have overall computational complexity $\mathcal{O}\left(\mathrm{IT}_2 \cdot [p n_2]\right)$. Along similar lines, Step 2 of SACOS would entail $\mathcal{O}(m^2 n_2 + m n_2) = \mathcal{O}(m^2 n_2)$ operations to compute the orthogonal projections and their ℓ_2 norms. We summarize the overall complexity results in Table 2.2. Since we will typically have γ small, $m \ll n_1$, and $p \ll n_2$ in our approaches, the computational complexity of our approaches can be much less than methods that operate on the full data or require intermediate SVD's of matrices of the same size as M.

Note that we have not included here the complexity of acquiring or forming the observations in any of the methods. For the ACOS method, this would comprise up to an additional $\mathcal{O}\left(m n_1 (\gamma n_2)\right)$ operations for Step 1 and $\mathcal{O}(m^2 + m n_1 + n_1 n_2 + n_2 p) = \mathcal{O}(m n_1 + n_1 n_2 + n_2 p)$ operations for Step 2, where the complexity for the second step is achieved by iteratively multiplying together the left-most two factors in the overall product, and using the fact that $m \leq n_1$. Similarly, observations obtained via the SACOS approach could require up $\mathcal{O}(mn_1n_2)$ operations. On the other hand, depending on the implementation platform, forming the observations themselves could also have a negligible computational effect e.g., in our imaging example when linear observations are formed "implicitly" using a spatial light modulator or single pixel camera [70]. Finally, we note that further reductions in the overall complexity of our approach may be achieved using fast or sparse JL embeddings along the lines of [66, 67].

Finally, it is worth noting[7] that the performance in our our visual saliency application could likely be improved using an additional assumption that the salient regions be spatially clustered. This could be implemented here using *group sparse* regularization (e.g. [78]) in Step 2 of ACOS, or by directly identifying groups of nonzero elements in Step 2 of SACOS. We defer investigations along these lines to a future effort.

[7] Thanks to David B. Dunson and Alfred O. Hero for these suggestions.

Chapter 3

Sketching of Low-Rank Tensor Regression

3.1 Introduction

For a sequence of D-way design tensors $A_i \in \mathbb{R}^{p_1 \times \cdots \times p_D}$, $i \in [n] = \{1, \ldots, n\}$, we observe noisy linear measurements of an unknown D-way tensor $\Theta \in \mathbb{R}^{p_1 \times \cdots \times p_D}$, given by

$$b_i = \langle A_i, \Theta \rangle + z_i, \quad \text{for all} \quad i \in [n], \tag{3.1}$$

where $\{z_i\}_{i=1}^n$ corresponds to the noise in each observation, and $\langle A_i, \Theta \rangle = \text{vec}(A_i)^\top \text{vec}(\Theta)$, with $\text{vec}(X)$ denoting the vectorization of a tensor X. Given the design tensors $\{A_i\}_{i=1}^n$ and noisy observations $\{b_i\}_{i=1}^n$, a natural approach for estimating the parameter Θ is to use the *Ordinary Least Square* (OLS) estimation for tensor regression, i.e., to solve

$$\min_{\Theta \in \mathbb{R}^{p_1 \times \cdots \times p_D}} \sum_{i=1}^n \left(b_i - \langle A_i, \Theta \rangle\right)^2 . \tag{3.2}$$

Tensor regression has been widely studied in the literature. Applications include computer vision [79–81], data mining [82], multi-model ensembles [83], neuroimaging analysis [84, 85], multitask learning [86, 87], and multivariate spatial-temporal data analysis [88, 89]. In these applications, modeling the unknown parameters as a tensor is what is needed, as it allows for learning data that has multi-directional relations, such as in climate prediction [90], inherent structure learning with multi-dimensional indices [86], and hand movement trajectory decoding [81].

Due to the high dimensionality of tensor data, structured learning based on low-rank tensor decompositions, such as CANDECOMP/PARAFAC (CP) decomposition and Tucker decomposition models [91, 92] have been proposed in order to obtain tractable tensor regression problems. As discussed more below, requiring the unknown tensor to be low-rank significantly reduces the number of unknown parameters. As natural convex formulations based on the nuclear norm are known to be computationally expensive [93, 94], nonconvex heuristics for low-rank tensor recovery are often used in practice [83, 86, 88].

We consider low-rank tensor regression problems based on the CP decomposition and Tucker decomposition models. For simplicity, we first focus on the CP model, and later extend our analysis to the Tucker model. Suppose that Θ admits a rank-R CP decomposition, that is,

$$\Theta = \sum_{r=1}^{R} \theta_1^{(r)} \circ \cdots \circ \theta_D^{(r)}, \tag{3.3}$$

where $\theta_d^{(r)} \in \mathbb{R}^{p_d}$ for all $r \in [R]$ and $\circ$ is the outer product of vectors. For convenience, we denote the set of factors for low-rank tensors by

$$\mathcal{S}_{D,R} = \Big\{ [[\Theta_1, \ldots, \Theta_D]] \; : \; \Theta_d = [\theta_d^{(1)}, \ldots, \theta_d^{(R)}] \in \mathbb{R}^{p_d \times R}, \text{ for all } d \in [D] \Big\}.$$

Then we can rewrite model (3.1) in a compact form

$$b = A(\Theta_D \odot \cdots \odot \Theta_1) 1_R + z, \tag{3.4}$$

where $b, z \in \mathbb{R}^n$, $A = [\text{vec}(A_1), \cdots, \text{vec}(A_n)]^\top \in \mathbb{R}^{n \times \prod_{d=1}^{D} p_d}$, $1_R = [1, \ldots, 1] \in \mathbb{R}^R$ is a vector of all 1s, $\otimes$ is the Kronecker product, and $\odot$ is the Khatri-Rao product[1]. In addition, the OLS estimation for tensor regression (3.2) can be rewritten as the following

[1] These are defined below in the section of notation.

nonconvex problem in terms of low-rank tensor parameters $[[\Theta_1, \ldots, \Theta_D]]$,

$$\min_{\vartheta \in \mathcal{S}_{\odot D,R}} \|A\vartheta - b\|_2^2, \tag{3.5}$$

where $\mathcal{S}_{\odot D,R} = \left\{ (\Theta_D \odot \cdots \odot \Theta_1) 1_R \in \mathbb{R}^{\prod_D^{d=1} p_d} \ : \ [[\Theta_1, \ldots, \Theta_D]] \in \mathcal{S}_{D,R} \right\}$.

The number of parameters for a general tensor $\Theta \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ is $\prod_{d=1}^D p_d$, which may be prohibitive even for small values of $\{p_d\}_{d=1}^D$. The benefit of the low-rank model (3.3) is that it dramatically reduces the degrees of freedom of the unknown tensor from $\prod_{d=1}^D p_d$ to $R \sum_{d=1}^D p_d$, where we are typically interested in the case when $R \ll p_d$ for each $d \in [D]$. For example, a typical MRI image has size $256^3 \approx 1.7 \times 10^7$, while using the low-rank model with $R = 5$, we reduce the number of unknown parameters to $256 \times 3 \times 5 \approx 4 \times 10^3 \ll 10^7$. This significantly increases the applicability of the tensor regression model in practice.

Nevertheless, solving the tensor regression problem (3.5) is still expensive in terms of both computation and memory requirements, for typical settings, when $n \gg R \cdot \sum_{d=1}^D p_d$, or even $n \gg \prod_{d=1}^D p_d$. In particular, the per iteration complexity is at least linear in n for popular algorithms such as block alternating minimization and block gradient descent [95, 96]. In addition, in order to store A, it takes $n \cdot \prod_{d=1}^D p_d$ words of memory. Both of these aspects are undesirable when n is large. This motivates us to consider data dimensionality reduction techniques, also called *sketching*, for the tensor regression problem.

Instead of solving (3.5), we consider the *Sketched Ordinary Least Square* (SOLS) estimation problem, defined as

$$\min_{\vartheta \in \mathcal{S}_{\odot D,R}} \|\Phi A\vartheta - \Phi b\|_2^2, \tag{3.6}$$

where $\Phi \in \mathbb{R}^{m \times n}$ is a random matrix specified below. Importantly, Φ will satisfy two properties discussed below, namely (1) $m \ll n$ so that we significantly reduce the size of the problem, and (2) Φ will be very sparse so that it can be applied very quickly.

Naïvely applying existing analyses of sketching techniques for least squares regression requires $m = \Omega\left(\prod_{d=1}^D p_d\right)$ (for a survey, see, e.g., [97]), which is prohibitive. Here, we use a sparse Johnson-Lindenstrauss transformation (SJLT) as our sketching matrix, with constant column sparsity and dimension $m = \Omega\left(R \cdot \sum_{d=1}^{D} p_d\right)$, up to logarithmic factors. We show that with high probability, simultaneously for every $\vartheta \in \mathcal{S}_{\odot,D,R}$, we have $\|\Phi A\vartheta - \Phi b\|_2^2 = (1 \pm \epsilon)\|A\vartheta - b\|_2^2$, which implies that any solution to (3.6) has the same cost as in (3.5) up to a $(1+\epsilon)$-factor. In particular, by solving (3.6) we obtain a $(1+\epsilon)$-approximation to (3.5). Our result is the first non-trivial dimensionality reduction for this problem, i.e., dimensionality reduction better than $\left(\prod_{d=1}^{D} p_d\right)$, which is trivial by ignoring the low rank structure of the tensor, and which achieves a relative error $(1+\epsilon)$-approximation.

Our analysis is based on a careful characterization of Talagrand's functional for the parameter space of low-rank tensors. Our sketching dimension m almost meets the intrinsic dimension of low-rank tensors, and is thus nearly optimal. We further provide numerical evaluations on both synthetic and real data to demonstrate the empirical performance of sketching based estimation.

Notation. For scalars $x, y \in \mathbb{R}$, we denote $x = (1 \pm \varepsilon)y$ if $x \in [(1-\varepsilon)y, (1+\varepsilon)y]$, $x \lesssim (\gtrsim) y$ if $x \le (\ge) cy$ for some universal constant $c > 0$, and $x \simeq y$ if both $x \lesssim y$ and $x \gtrsim y$ hold. We also use standard asymptotic notation $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$. Given a positive integer n, let $[n] = \{1, \ldots, n\}$. Given a vector $v \in \mathbb{R}^p$, we denote $\|v\|_1 = \sum_{i=1}^{p} |v_i|$, $\|v\|_2^2 = \sum_{i=1}^{p} v_i^2$, and $\|v\|_\infty = \max_{i \in [p]} |v_i|$. Given d vectors $v_1 \in \mathbb{R}^{p_1}, \ldots, v_d \in \mathbb{R}^{p_d}$, we denote $v_1 \circ \cdots \circ v_d \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ as a tensor formed by the outer product of vectors. Given a matrix $A \in \mathbb{R}^{m \times n}$, we denote its spectral norm by $\|A\|_2$, we let $\mathrm{span}(A) \subseteq \mathbb{R}^m$ be the subspace spanned by the columns of A, we let $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ be the largest and smallest singular values of A, respectively, and $\kappa(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$ be the condition number. We use $\mathrm{nnz}(A)$ to denote the number of nonzero entries of A. We use $\mathcal{P}_A$ as the projection operator onto $\mathrm{span}(A)$. Given two matrices $A = [a_1, \ldots, a_n] \in \mathbb{R}^{m \times n}$ and $B = [b_1, \ldots, b_q] \in \mathbb{R}^{p \times q}$, $A \otimes B = [a_1 \otimes B, \ldots, a_n \otimes B] \in \mathbb{R}^{mp \times nq}$ denotes the Kronecker product, and $A \odot B = [a_1 \otimes b_1, \ldots, a_n \otimes b_n] \in \mathbb{R}^{mp \times n}$ denotes the Khatri-Rao product with $n = q$. We let $\mathcal{B}_n \subset \mathbb{R}^n$ be the unit sphere of $\mathbb{R}^n$, i.e., $\mathcal{B}_n = \{x \in \mathbb{R}^n \ : \ \|x\|_2 = 1\}$. We also let $\mathbb{P}(\cdot)$ be the probability of an event and $\mathbb{E}(\cdot)$ the expectation of a random variable. Without further specification, we let $\prod p_d = \prod_{d=1}^{D} p_d$ and $\sum p_d = \sum_{d=1}^{D} p_d$.

3.2 Dimensionality Reduction for CP Decomposition

3.2.1 Background

We start with a few important definitions.

Definition 3 (Oblivious Subspace Embedding)**.** Suppose Π is a distribution on $m \times n$ matrices Φ, where m is a function of parameters n, d, and ε. Further suppose that with probability at least $1-\delta$, for any fixed $n \times d$ matrix A, a matrix Φ drawn from Π has the property that Φ is a $(1 \pm \varepsilon)$ subspace embedding for A, i.e., $\|\Phi A x\|_2^2 = (1 \pm \varepsilon)\|Ax\|_2^2$ for any $x \in \mathcal{X} \subseteq \mathbb{R}^d$. Then we call Π an (ε, δ) *oblivious subspace embedding* (OSE) of $\mathcal{X}$.

An OSE Φ preserves the norm of vectors in a certain set $\mathcal{X}$ after linear transformation by A. This is widely studied as a key property for sketching based analyses (see [97] and the references therein). We want to show an analogous property when $\mathcal{X}$ is parameterized by a low-rank tensor model.

Definition 4 (Leverage Scores)**.** Given $A \in \mathbb{R}^{n \times d}$, let $Z \in \mathbb{R}^{n \times d}$ have orthonormal columns that span the column space of A. Then $\ell_i^2(A) = \|e_i^\top Z\|_2^2$ is the i-th *leverage score* of A.

Leverage scores play an important role in randomized matrix algorithms [98–100]. Calculating the leverage scores naïvely by orthogonalizing A requires $\mathcal{O}(nd^2)$ time. It is shown in [101] that the leverage scores of A can be approximated individually up to a constant multiplicative factor in $\mathcal{O}(\text{nnz}(A)\log n + \text{poly}(d))$ time using sparse subspace embeddings.

Definition 5 (Sparse Johnson-Lindenstrauss Transforms)**.** Let σ_{ij} be independent Rademacher random variables, i.e., $\mathbb{P}(\sigma_{ij} = 1) = \mathbb{P}(\sigma_{ij} = -1) = 1/2$, and let $\delta_{ij} : \Omega_\delta \to \{0, 1\}$ be random variables, independent of the σ_{ij}, with the following properties:

- δ_{ij} are negatively correlated for fixed j, i.e., for all $1 \le i_1 < \ldots < i_k \le m$,

$$\mathbb{E}\left(\prod_{t=1}^{k} \delta_{i_t,j}\right) \le \prod_{t=1}^{k} \mathbb{E}\left(\delta_{i_t,j}\right) = \left(\frac{s}{m}\right)^k;$$

- There are $s = \sum_{i=1}^{m} \delta_{ij}$ nonzero δ_{ij} for a fixed j;
- The vectors $(\delta_{ij})_{i=1}^{m}$ are independent across $j \in [n]$.

Then $\Phi \in \mathbb{R}^{m \times n}$ is a *sparse Johnson-Lindenstrauss transform* (SJLT) matrix if $\Phi_{ij} = \frac{1}{\sqrt{s}}\sigma_{ij}\delta_{ij}$.

The SJLT has several benefits [97,102,103]. First, the computation of Φx takes only $\mathcal{O}(\mathrm{nnz}(x))$ time when s is a constant. Second, storing Φ takes only sn memory instead of mn, which is significant when $s \ll m$. This can often further be reduced by drawing the entries of Φ from a limited independent family of random variables. We will use an SJLT as the sketching matrix in our analysis and our goal will be to show sufficient conditions on m and s such that the analogue of the OSE property holds for low-rank tensor regression.

Definition 6 (Talagrand's Functional)**.** Given a (semi-)metric ρ on $\mathbb{R}^n$ and a bounded set $\mathcal{S} \subset \mathbb{R}^n$, *Talagrand's γ_2-functional* is

$$\gamma_2(S, \rho) = \inf_{\{\mathcal{S}_r\}_{r=0}^{\infty}} \sup_{x \in S} \sum_{r=0}^{\infty} 2^{r/2} \cdot \rho(x, \mathcal{S}_r), \tag{3.7}$$

where $\rho(x, \mathcal{S}_r)$ is a distance from x to $\mathcal{S}_r$ and the infimum is taken over all collections $\{\mathcal{S}_r\}_{r=0}^{\infty}$ such that $\mathcal{S}_0 \subset \mathcal{S}_1 \subset \ldots \subset \mathcal{S}$ with $|\mathcal{S}_0| = 1$ and $|\mathcal{S}_r| \leq 2^{2^r}$.

A closely related notion of γ_2-functional is the *Gausssian mean width*,

$$\mathcal{G}(\mathcal{S}) = \mathbb{E}_g \sup_{x \in \mathcal{S}} \langle g, x \rangle,$$

where $g \sim \mathcal{N}_n(0, I_n)$. For any bounded $\mathcal{S} \subset \mathbb{R}^n$, $\mathcal{G}(\mathcal{S})$ and $\gamma_2(\mathcal{S}, \|\cdot\|_2)$ differ multiplicatively by at most a universal constant in Euclidean space. Both of these quantities are widely used [104]. Finding a tight upper bound on the γ_2-functional for the parameter space of low-rank tensors is a key part of our analysis.

Definition 7 (Finsler Metric)**.** Let $E, E' \subset \mathbb{R}^n$ be p-dimensional subspaces. The *Finsler metric* of E and E' is

$$\rho_{\mathrm{Fin}}(E, E') = \|\mathcal{P}_E - \mathcal{P}_{E'}\|_2,$$

where $\mathcal{P}_E$ is the projection onto the subspace E.

The Finsler metric is the semi-metric used in the γ_2-functional in our analysis. Note that $\rho_{\text{Fin}}(E, E') \le 1$ always holds for any E and E' [105]. See further discussion in Section 3.2.

For convenience, we introduce the following notation. Given a D-way tensor $\Theta = \sum_{r=1}^{R} \theta_1^{(r)} \circ \cdots \circ \theta_D^{(r)} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$, where $\theta_d^{(r)} \in \mathbb{R}^{p_d}$ for all $d \in [D]$ and $r \in [R]$, we consider fixing all but $\theta_1^{(r)}$ for $r \in [R]$, and denoting

$$A^{\left\{\theta_{\backslash 1}^{(r)}\right\}} = \left[A^{\theta_{\backslash 1}^{(1)}}, \ldots, A^{\theta_{\backslash 1}^{(R)}}\right] \in \mathbb{R}^{n \times R p_1},$$

where

$$A^{\theta_{\backslash 1}^{(i)}} = \sum_{j_D=1}^{p_D} \cdots \sum_{j_2=1}^{p_2} A^{(j_D, \ldots, j_2)} \cdot \theta_{D, j_D}^{(i)} \cdots \theta_{2, j_2}^{(i)}$$
$$A = \left[A^{(1,\ldots,1)}, A^{(1,\ldots,2)}, \ldots, A^{(p_D,\ldots,p_2)}\right] \in \mathbb{R}^{n \times \prod p_d}$$

$\theta_{d,j_d}^{(i)}$ is the j_d-th entry of $\theta_d^{(i)}$, and $A^{(j_D,\ldots,j_2)} \in \mathbb{R}^{n \times p_1}$ for all $j_d \in [p_d]$, $d \in [D] \backslash \{1\}$. The above parameterization allows us to view tensor regression as preserving the norms of vectors in an infinite union of subspaces, described in more detail below.

We provide sufficient conditions for the SJLT matrix $\Phi \in \mathbb{R}^{m \times n}$ to preserve the cost of all solutions for tensor regression, i.e., bounds on the sketching dimension m and the per-column sparsity s for which

$$\mathop{\mathbb{E}}_{\Phi} \sup_{x \in \mathcal{T}} \left| \|\Phi x\|_2^2 - 1 \right| < \varepsilon \tag{3.8}$$

where ε is a given precision, $\mathcal{T} = \bigcup_{E \in \mathcal{V}} \{x \in E \ : \ \|x\|_2 = 1\}$, and

$$\mathcal{V} = \bigcup_{\theta_d^{(r)} \in \mathbb{R}^{p_d}, \forall r \in [R],\ d \in [D] \backslash \{1\}} \text{span}\left[A^{\left\{\theta_{\backslash 1}^{(r)}\right\}}, A^{\left\{\phi_{\backslash 1}^{(r)}\right\}}\right].$$

Note that by linearity, it suffices to consider x with $\|x\|_2 = 1$ in the above, which explains the form of (3.8). Also note that (3.8) implies for all $\vartheta \in \mathcal{S}_{\odot D,R}$, that

$$\|\Phi A\vartheta - \Phi b\|_2^2 = (1 \pm \varepsilon)\|A\vartheta - b\|_2^2, \tag{3.9}$$

which allows us to minimize the much smaller sketched problem to obtain parameters ϑ which, when plugged into the original objective function, provide a multiplicative $(1+\epsilon)$-approximation.

We need the following theorem for embedding an infinite union of subspaces. All proofs can be found in the appendix.

Theorem 3. Let $\mathcal{T} \subset \mathcal{B}_n$ and $\Phi \in \mathbb{R}^{m\times n}$ be an SJLT matrix with column sparsity s, and

$$p_{\mathcal{V}} = \sup_{\substack{\theta_d^{(r)} \in \mathbb{R}^{p_d}, \forall r\in[R],\\ d\in[D]\setminus\{1\}}} \dim\left(\operatorname{span}\left[A^{\left\{\theta_{\setminus 1}^{(r)}\right\}}, A^{\left\{\phi_{\setminus 1}^{(r)}\right\}}\right]\right).$$

Then (3.8) holds if m and s satisfy

$$m \gtrsim \left(\gamma_2^2(\mathcal{V},\rho_{\text{Fin}}) + p_{\mathcal{V}} + \log\mathcal{N}(\mathcal{V},\rho_{\text{Fin}},\varepsilon_0)\right)\cdot(\log^4 m)(\log^5 n)\varepsilon^{-2}, \tag{3.10}$$

$$s \gtrsim \left(\left[\int_0^{\varepsilon_0} (\log\mathcal{N}(\mathcal{V},\rho_{\text{Fin}},t))^{1/2}\,dt\right]^2 + \tilde{\alpha}^2\log^2\mathcal{N}(\mathcal{V},\rho_{\text{Fin}},\varepsilon_0) + \varepsilon_0^2 p_{\mathcal{V}}\log\frac{1}{\varepsilon_0}\right) \cdot (\log^6 m)(\log^5 n)\varepsilon^{-2}, \tag{3.11}$$

where $\tilde{\alpha}^2$ is the largest leverage score of any $\left[A^{\left\{\theta_{\setminus 1}^{(r)}\right\}}, A^{\left\{\phi_{\setminus 1}^{(r)}\right\}}\right] \in \mathcal{V}$ and $\mathcal{N}(\mathcal{V},\rho_{\text{Fin}},t)$ is the covering number of $\mathcal{V}$ with radius t under the Finsler metric.

The proof of Theorem 3 is provided in Appendix 7.2.1. Theorem 3 is based on recent work on a unified theory of dimensionality reduction [106,107]. Note that the parameter space for the tensor regression problem (3.1) is a subspace of $\mathbb{R}^{\prod p_d}$, i.e., $\mathcal{S}_{\odot D,R} \subset \mathbb{R}^{\prod p_d}$. Therefore, a naïve application of sketching requires $m \gtrsim \prod p_d/\varepsilon^2$ in order for (3.8) to hold [108]. However, $\prod p_d$ can be very large and is far larger than the intrinsic number of degrees of freedom of the parameter space $\mathcal{S}_{\odot D,R}$, which is $R\sum p_d$. In the sequel, we provide a careful analysis of dimensionality reduction in terms of $\gamma_2(\mathcal{V},\rho_{\text{Fin}})$, $p_{\mathcal{V}}$, and $\mathcal{N}(\mathcal{V},\rho_{\text{Fin}},\eta_0)$, where sufficient conditions $m = \Omega(R\sum p_d)$ and $s = \Omega(1)$ are achieved, up to logarithmic factors [109].

3.2.2 Base Case: Rank-1 and Two-Way Tensors

We start with the base case when $R = 1$ and $D = 2$, i.e., the parameter space is $\mathcal{S}_{2,1}$. Then the parameter admits the decomposition $\Theta = \theta_1 \circ \theta_2$. For notational convenience, we let $\Theta = u \circ v$, where $u \in \mathbb{R}^{p_1}$ and $v \in \mathbb{R}^{p_2}$, and let $A^v = \sum_{i=1}^{p_2} A^{(i)} v_i$, where $A = [A^{(1)}, \ldots, A^{(p_2)}] \in \mathbb{R}^{n \times p_2 p_1}$ with $A^{(i)} \in \mathbb{R}^{n \times p_1}$ for all $i \in [p_2]$. Consequently, the observation model (3.4) can be written as

$$b = A(v \otimes u) + z = A^v u + z,$$

and the corresponding OLS and SOLS using an SJLT matrix $\Phi \in \mathbb{R}^{m \times n}$ are, respectively,

$$\min_{v \in \mathbb{R}^{p_2}, u \in \mathbb{R}^{p_1}} \|A^v u - b\|_2^2 \quad \text{and} \quad \min_{v \in \mathbb{R}^{p_2}, u \in \mathbb{R}^{p_1}} \|\Phi A^v u - \Phi b\|_2^2.$$

Next, we show the following theorem, which provides sufficient conditions for the base case $\mathcal{S}_{2,1}$.

Theorem 4. Suppose the leverage scores of A are bounded, i.e., $\max_{i \in [n]} \ell_i^2(A) \leq 1/p_2^2$. Let

$$\mathcal{T} = \left\{ \frac{Ax - Ay}{\|Ax - Ay\|_2} \;\middle|\; x = v_1 \otimes u_1, y = v_2 \otimes u_2,\ u_1, u_2 \in \mathbb{R}^{p_1} \right\}$$

and $\Phi \in \mathbb{R}^{m \times n}$ is an SJLT matrix with column sparsity s. Then (3.8) holds if m and s satisfy

$$m \gtrsim \varepsilon^{-2} (p_1 + p_2) \log((p_1 + p_2)\kappa(A))(\log^4 m)(\log^5 n), \tag{3.12}$$

$$s \gtrsim \varepsilon^{-2} \log^2(p_1 + p_2)(\log^6 m)(\log^5 n). \tag{3.13}$$

The proof of Theorem 4 is provided in Appendix 7.2.2. From Theorem 4, when $m = \Omega(p_1 + p_2)$ and $s = \Omega(1)$, (3.9) holds.

3.2.3 Extension to General Ranks

We extend our analysis to the general case of two-way tensors with general rank, i.e., the parameter space is $\mathcal{S}_{2,R}$ for $R \geq 1$. In this case, we have $\Theta = \sum_{r=1}^{R} u^{(r)} \circ v^{(r)}$, where $u^{(r)} \in \mathbb{R}^{p_1}$ and $v^{(r)} \in \mathbb{R}^{p_2}$ for all $r \in [R]$, and $A^{\left\{v^{(r)}\right\}} = \left[\sum_{i=1}^{p_2} A^{(i)} v_i^{(1)}, \ldots, \sum_{i=1}^{p_2} A^{(i)} v_i^{(R)}\right]$, where $A = [A^{(1)}, \ldots, A^{(p_2)}] \in \mathbb{R}^{n \times p_2 p_1}$ and $A^{(i)} \in \mathbb{R}^{n \times p_1}$ for all $i \in [p_2]$. Consequently, the observation model (3.4) can be written as

$$b = A^{\left\{v^{(r)}\right\}} \left[u^{(1)\top} \ldots u^{(R)\top}\right]^\top + z,$$

and the corresponding OLS and SOLS using an SJLT matrix $\Phi \in \mathbb{R}^{m \times n}$ are, respectively,

$$\min_{\substack{v^{(r)} \in \mathbb{R}^{p_2} \\ u^{(r)} \in \mathbb{R}^{p_1}, \forall r \in [R]}} \left\| A^{\left\{v^{(r)}\right\}} \left[u^{(1)\top} \ldots u^{(R)\top}\right]^\top - b \right\|_2^2, \quad \text{and}$$

$$\min_{\substack{v^{(r)} \in \mathbb{R}^{p_2} \\ u^{(r)} \in \mathbb{R}^{p_1}, \forall r \in [R]}} \left\| \Phi A^{\left\{v^{(r)}\right\}} \left[u^{(1)\top} \ldots u^{(R)\top}\right]^\top - \Phi b \right\|_2^2.$$

Our next theorem provides sufficient conditions for $\mathcal{S}_{2,R}$.

Theorem 5. Suppose $R \leq p_2/2$ and the leverage scores of A are bounded, i.e., $\max_{i \in [n]} \ell_i^2(A) \leq 1/(R^2 p_2^2)$. Let

$$\mathcal{T} = \left\{ \frac{Ax - Ay}{\|Ax - Ay\|_2} \,\middle|\, x = \sum_{r=1}^{R} v_1^{(r)} \otimes u_1^{(r)}, y = \sum_{r=1}^{R} v_2^{(r)} \otimes u_2^{(r)},\ u_1^{(r)}, u_1^{(r)} \in \mathbb{R}^{p_1},\ \forall r \in [R] \right\}$$

and $\Phi \in \mathbb{R}^{m \times n}$ is an SJLT matrix with column sparsity s. Then (3.8) holds if m and s satisfy

$$m \gtrsim \varepsilon^{-2} (\log^4 m)(\log^5 n) R\,(p_1 + p_2) \log\left(R(p_1 + p_2)\kappa(A)\right),$$
$$s \gtrsim \varepsilon^{-2} (\log^6 m)(\log^5 n) \log^2\left(R(p_1 + p_2)\kappa(A)\right).$$

The proof of Theorem 5 is provided in Appendix 7.2.3. From Theorem 5, we have that when $m = \Omega(R(p_1 + p_2))$ and $s = \Omega(1)$, (3.9) holds using an SJLT matrix Φ. The extra condition of $R \leq p_2/2$ is not restrictive, as in applications of low-rank tensors, typically $R \ll \min_{d \in [D]} p_d$.

3.2.4 Extension to General Tensors

We first extend our analysis to general tensors with rank 1, i.e., the parameter space is now $\mathcal{S}_{D,1}$ for $D \geq 2$. In this case, we have $\Theta = \theta_1 \circ \cdots \circ \theta_D$, where $\theta_d \in \mathbb{R}^{p_d}$ for all $d \in [D]$. Consequently, the observation model (3.4) can be written as

$$b = A \cdot (\theta_D \otimes \cdots \otimes \theta_1) + z = A^{\{\theta_{\setminus 1}\}} \cdot \theta_1 + z,$$

and the corresponding OLS and SOLS using an SJLT matrix $\Phi \in \mathbb{R}^{m \times n}$ are, respectively,

$$\min_{\substack{\theta_i \in \mathbb{R}^{p_i} \\ \forall i \in [D]}} \left\| A^{\{\theta_{\setminus 1}\}} \theta_1 - b \right\|_2^2 \text{ and } \min_{\substack{\theta_i \in \mathbb{R}^{p_i} \\ \forall i \in [D]}} \left\| \Phi A^{\{\theta_{\setminus 1}\}} \theta_1 - \Phi b \right\|_2^2.$$

Our next theorem provides sufficient conditions for $\mathcal{S}_{D,1}$.

Theorem 6. Suppose the leverage scores of A are bounded, i.e., $\max_{i\in[n]} \ell_i^2(A) \leq 1/\left(\sum_{d=2}^D p_d\right)^2$. For any $\vartheta = \theta_D \otimes \cdots \otimes \theta_1 \in \mathcal{S}_{\odot D,1}$ and $\varphi = \phi_D \otimes \cdots \otimes \phi_1 \in \mathcal{S}_{\odot D,1}$, $\theta_d, \phi_d \in \mathbb{R}^{p_d}$ for all $d \in [D]$, let

$$\mathcal{T} = \left\{ \frac{A\vartheta - A\varphi}{\|A\vartheta - A\varphi\|_2} \;\middle|\; \vartheta = \theta_D \otimes \cdots \otimes \theta_1, \varphi = \phi_D \otimes \cdots \otimes \phi_1,\ \theta_d, \phi_d \in \mathbb{R}^{p_d},\ \forall d \in [D] \right\}$$

and $\Phi \in \mathbb{R}^{m\times n}$ is an SJLT matrix with column sparsity s. Then (3.8) holds if m and s satisfy

$$\begin{aligned} m &\gtrsim \varepsilon^{-2} (\log^4 m)(\log^5 n) \left(\sum_{d=1}^D p_d \log \left(D\kappa(A) \sum_{d=1}^D p_d \right) \right), \\ s &\gtrsim \varepsilon^{-2} (\log^6 m)(\log^5 n) \log^2 \left(\sum_{d=1}^D p_d \right). \end{aligned}$$

The proof of Theorem 6 is provided in Appendix 7.2.4. From Theorem 6, we have that when $m = \Omega\left(\sum_{d=1}^D p_d\right)$ and $s = \Omega(1)$, (3.9) holds using an SJLT matrix Φ.

3.2.5 Extension to General Ranks and Tensors

Finally, we provide our guarantees for general tensors with general ranks, i.e., the parameter space is $\mathcal{S}_{D,R}$ for $D \geq 2$ and $R \geq 1$. We have the observation model (3.4) as

$$b = A \cdot \sum_{r=1}^{R} \theta_D^{(r)} \otimes \cdots \otimes \theta_1^{(r)} + z = \sum_{r=1}^{R} A^{\theta_{\setminus 1}^{(r)}} \cdot \theta_1^{(r)} + z = A^{\left\{\theta_{\setminus 1}^{(r)}\right\}} \cdot \left[\theta_1^{(1)\top} \dots \theta_1^{(R)\top}\right]^{\top} + z,$$

and the corresponding OLS and SOLS using an SJLT matrix $\Phi \in \mathbb{R}^{m \times n}$ are, respectively,

$$\min_{\substack{\theta_i^{(r)} \in \mathbb{R}^{p_i} \\ \forall i \in [D], r \in [R]}} \left\| A^{\left\{\theta_{\setminus 1}^{(r)}\right\}} \cdot \left[\theta_1^{(1)\top} \dots \theta_1^{(R)\top}\right]^{\top} - b \right\|_2^2, \quad \text{and}$$

$$\min_{\substack{\theta_i^{(r)} \in \mathbb{R}^{p_i} \\ \forall i \in [D], r \in [R]}} \left\| \Phi A^{\left\{\theta_{\setminus 1}^{(r)}\right\}} \cdot \left[\theta_1^{(1)\top} \dots \theta_1^{(R)\top}\right]^{\top} - \Phi b \right\|_2^2.$$

Our most general theorem for CP decomposition is the following, providing sufficient conditions for $\mathcal{S}_{D,R}$.

Theorem 7. Suppose $R \leq \max_d p_d/2$ and the leverage scores of A are bounded, i.e., $\max_{i \in [n]} \ell_i^2(A) \leq 1/\left(R^2 \left(\sum_{d=2}^{D} p_d\right)^2\right)$. Let

$$\mathcal{T} = \left\{ \frac{A\vartheta - A\varphi}{\|A\vartheta - A\varphi\|_2} : \vartheta = \sum_{r=1}^{R} \theta_D^{(r)} \otimes \cdots \otimes \theta_1^{(r)}, \varphi = \sum_{r=1}^{R} \phi_D^{(r)} \otimes \cdots \otimes \phi_1^{(r)}, \right.$$
$$\left. \theta_d^{(r)}, \phi_d^{(r)} \in \mathcal{B}_{p_d}, \ \forall r \in [R], d \in [D] \right\}$$

and $\Phi \in \mathbb{R}^{m \times n}$ is an SJLT matrix with column sparsity s. Then (3.8) holds if m and s satisfy

$$m \gtrsim \varepsilon^{-2} (\log^4 m)(\log^5 n) R \sum_{d=1}^{D} p_d \log\left(DR\kappa(A) \sum_{d=1}^{D} p_d \right),$$
$$s \gtrsim \varepsilon^{-2} (\log^6 m)(\log^5 n) \log^2 \left(\sum_{d=1}^{D} p_d \right).$$

The proof of Theorem 7 is provided in Appendix 7.2.5. From Theorem 7, we have that when $m = \Omega\left(R \sum_{d=1}^{D} p_d\right)$ and $s = \Omega(1)$, (3.9) holds using an SJLT matrix Φ. These complexities are optimal, up to logarithmic factors, for the CP decomposition model, since they meet the number of degrees of freedom of the CP model. The extra condition of $R \leq \max_d p_d/2$ is not restrictive, as we are interested in low-rank tensors satisfying $R \ll \min_{d\in[D]} p_d$.

3.3 Dimensionality Reduction for Tucker Decomposition

We start with a formal description of the Tucker model. Suppose Θ admits the following Tucker decomposition:

$$\Theta = \sum_{r_1=1}^{R_1} \cdots \sum_{r_D=1}^{R_D} g(r_1, \ldots, r_D) \cdot \theta_1^{(r_1)} \circ \cdots \circ \theta_D^{(r_D)}, \tag{3.14}$$

where $\theta_d^{(r_d)} \in \mathbb{R}^{p_d}$ for all $r_d \in [R_d]$. Letting

$$\begin{aligned} A^{\theta_{\setminus 1}^{(r_1,\ldots,r_D)}} &= \sum_{j_D=1}^{p_D} \cdots \sum_{j_2=1}^{p_2} A^{(j_D,\ldots,j_2)} \cdot \theta_{D,j_D}^{(r_D)} \cdots \theta_{2,j_2}^{(r_2)}, \\ A^{\left\{\theta_{\setminus 1}^{\{r_d\}}\right\}} &= \Bigg[\sum_{r_2=1}^{R_2} \cdots \sum_{r_D=1}^{R_D} A^{\theta_{\setminus 1}^{(r_1,\ldots,r_D)}} \cdot g(1, r_2, \ldots, r_D), \ldots\ldots, \\ &\qquad \sum_{r_2=1}^{R_2} \cdots \sum_{r_D=1}^{R_D} A^{\theta_{\setminus 1}^{(r_1,\ldots,r_D)}} \cdot g(R_1, r_2, \ldots, r_D) \Bigg], \end{aligned}$$

then the observation model (3.4) can be written as

$$\begin{aligned} b &= A \sum_{r_1=1}^{R_1} \cdots \sum_{r_D=1}^{R_D} g(r_1, \ldots, r_D) \cdot \theta_D^{(r_D)} \otimes \cdots \otimes \theta_1^{(r_1)} + z \\ &= \sum_{r_1=1}^{R_1} \cdots \sum_{r_D=1}^{R_D} A^{\theta_{\setminus 1}^{(r_1,\ldots,r_D)}} \cdot g(r_1, \ldots, r_D) \cdot \theta_1^{(r_1)} + z \\ &= A^{\left\{\theta_{\setminus 1}^{\{r_d\}}\right\}} \cdot \left[\theta_1^{(1)\top} \ldots \theta_1^{(R_1)\top}\right]^\top + z, \end{aligned}$$

and the corresponding OLS and SOLS using an SJLT matrix $\Phi \in \mathbb{R}^{m\times n}$ are, respectively,

$$\min_{\substack{\theta_i^{(r)} \in \mathbb{R}^{p_i} \\ \forall i \in [D], r\in[R]}} \left\| A^{\left\{\theta_{\setminus 1}^{(r)}\right\}} \cdot \left[\theta_1^{(1)\top} \dots \theta_1^{(R)\top}\right]^\top - b \right\|_2^2, \quad \text{and}$$

$$\min_{\substack{\theta_i^{(r)} \in \mathbb{R}^{p_i} \\ \forall i \in [D], r\in[R]}} \left\| \Phi A^{\left\{\theta_{\setminus 1}^{(r)}\right\}} \cdot \left[\theta_1^{(1)\top} \dots \theta_1^{(R)\top}\right]^\top - \Phi b \right\|_2^2.$$

Our next theorem provides sufficient conditions for the general Tucker decomposition model.

Theorem 8. Suppose $\prod_{d=1}^D R_d \le \max_d p_d/2$ and the leverage scores of A are bounded, i.e., $\max_{i\in[n]} \ell_i^2(A) \le 1/\left(\sum_{d=2}^D R_d p_d + \prod_{d=1}^D p_d\right)^2$. Let

$$\begin{aligned}\mathcal{T} = \Bigg\{ \frac{A\vartheta - A\varphi}{\|A\vartheta - A\varphi\|_2} : \vartheta = \sum_{r_1=1}^{R_1} \cdots \sum_{r_D=1}^{R_D} g_1(r_1,\dots,r_D)\cdot \theta_D^{(r_D)} \otimes \cdots \otimes \theta_1^{(r_1)}, \\ \varphi = \sum_{r_1=1}^{R_1} \cdots \sum_{r_D=1}^{R_D} g_2(r_1,\dots,r_D)\cdot \phi_D^{(r_D)} \otimes \cdots \otimes \phi_1^{(r_1)}, \\ \theta_d^{(r_d)}, \phi_d^{(r_d)} \in \mathcal{B}_{p_d},\ \forall r_d \in [R_d], d\in[D] \Bigg\}\end{aligned}$$

and $\Phi \in \mathbb{R}^{m\times n}$ is an SJLT matrix with column sparsity s. Then (3.8) holds if m and s satisfy

$$\begin{aligned} m &\gtrsim \varepsilon^{-2}(\log^4 m)(\log^5 n) C_0 \cdot \log\left(C_0 D \kappa(A)\sqrt{\prod_{d=2}^D R_d}\right), \\ s &\gtrsim \varepsilon^{-2}(\log^6 m)(\log^5 n)\log^2 C_0, \end{aligned}$$

where $C_0 = \sum_{d=1}^D R_d p_d + \prod_{d=1}^D p_d$.

The proof of Theorem 8 is provided in Appendix 7.2.6. From Theorem 8, we have that when $m = \Omega\left(D(\sum_{d=1}^D R_d p_d + \prod_{d=1}^D p_d)\right)$ and $s = \Omega(D)$, then (3.8) holds for the Tucker decomposition model using an SJLT matrix, provided that $\prod R_d$ is not too large compared with $\max_d p_d$, which is typical in applications of low rank tensors in which the goal is to use small values of the R_d when faced with large values of the p_d. Thus, the solution to the SOLS is a $(1+\epsilon)$-approximation to the OLS.

3.4 Flattening Leverage Scores

The analysis above depends on a bound on the leverage scores of the design matrix A. This might be restrictive if we have no control on the design A. In the sequel, we apply a standard idea [110,111] to flatten the leverage scores of a deterministic design A based on the Walsh-Hadamard matrix. An SRHT matrix is defined as

$$\Phi = \sqrt{\frac{n}{m}} P H \Sigma, \tag{3.15}$$

where the components Σ, H and P are generated as:

(G1) Σ is an $n \times n$ diagonal matrix, where $\Sigma_{ii} = 1$ or -1 with equal probabilities 1/2.

(G2) H is an $n \times n$ orthogonal matrix generated from a Walsh-Hadamard matrix scaled by $n^{-1/2}$.

(G3) P is an $m \times n$ SJLT matrix, with column sparsity bounded by s.

Note that computing a matrix-vector product with H takes $\mathcal{O}(n \log n)$ instead of n^2 time. Thus, one can compute $H\Sigma A$ for an $n \times d$ matrix A in $O(nd \log n)$ time, which is well-suited for the case in which A is dense, e.g., $\text{nnz}(A) = \Theta(nd)$. The purpose of the matrix product $H\Sigma$ is to uniformize the leverage scores before applying our SJLT P.

We next give a standard lemma for flattening the leverage scores, included for completeness. Without loss of generality, we assume that $n = 2^q$ for a positive integer q, implying that a Walsh-Hadamard matrix exists.

Lemma 1. Suppose H and Σ are generated as in (G1) and (G2). Given any real value $\delta \in (0,1)$ and an $n \times d$ matrix A with $\text{rank}(A) = r$, we have with probability at least $1 - \delta$,

$$\max_{i \in [n]} \ell_i^2(H\Sigma A) \lesssim \frac{r \cdot \log\left(\frac{nr}{\delta}\right)}{n}.$$

The proof of Lemma 1 is provided in Appendix 7.2.7. Applying this with the bound $\max_{i\in[n]} \ell_i^2(H\Sigma A) \leq 1/\left(R^2\left(\sum_{d=2}^D p_d\right)^2\right)$ of Theorem 7 gives:

Proposition 1. Suppose H and Σ are generated as in (G1) and (G2). For low-rank tensor regression (3.4), where $A \in \mathbb{R}^{n\times\prod_{d=1}^D p_d}$ is the matricization of all tensor designs, if n satisfies

$$n \gtrsim R^2\left(\sum_{d=2}^D p_d\right)^2 \cdot \operatorname{rank}(A)\cdot\log\left(\frac{n\cdot\operatorname{rank}(A)}{\delta}\right),$$

then with probability at least $1-\delta$, we have

$$\max_{i\in[n]} \ell_i^2(H\Sigma A) \leq 1/\left(R^2\left(\sum_{d=2}^D p_d\right)^2\right).$$

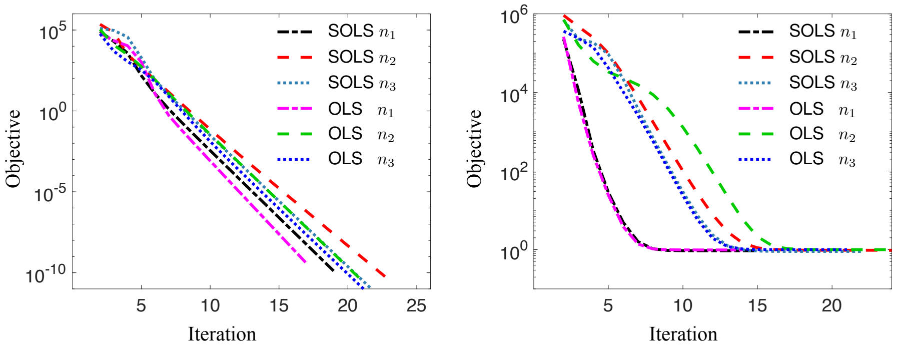
Combining Theorem 7 and Proposition 1, we achieve (3.8), provided n is sufficiently large. Here we use that for all x, $\|H\Sigma Ax\|_2 = \|Ax\|_2$ since $H\Sigma$ is an isometry.

In the worst case, $\operatorname{rank}(A) = \prod_{d=1}^D p_d$, which requires $n = \Omega\left(R^2\left(\sum_{d=2}^D p_d\right)^2\cdot\prod_{d=1}^D p_d\right)$. In overconstrained regression, it is often assumed that the number n of examples is at least a small polynomial in $\operatorname{rank}(A)$ [97], which implies this bound on n. Also, if, for example, A_i is sampled from a distribution with a rank deficient covariance, one may even have $\operatorname{rank}(A) \ll \prod_{d=1}^D p_d$.

One should note that computing $PH\Sigma A$ takes $(n\log n)\prod_{d=1}^D p_d$ time, provided the column sparsity s of P is $O(1)$. This is $O(\operatorname{nnz}(A)\log n)$ time for dense matrices A, i.e., those with $\operatorname{nnz}(A) = \Omega(nd)$, but in general, unlike our earlier results, is not $O(\operatorname{nnz}(A)\log n)$ time for sparse matrices. Analogous results can be obtained for the Tucker decomposition model, which we omit.

3.5 Experiments

We study the performance of sketching for tensor regression through numerical experiments over both synthetic and real data sets. For solving the OLS problem for tensor regression (3.2), we use a cyclic block-coordinate minimization algorithm based on a tensor toolbox [112]. Specifically, in a cyclic manner for all $d \in [D]$, we fix all but one Θ_d of $[[\Theta_1, \ldots, \Theta_D]] \in \mathcal{S}_{D,R}$ and minimize the resulting quadratic loss function (3.2) with respect to Θ_i, until the decrease of the objective is smaller than a predefined threshold τ. For SOLS, we use the same algorithm after multiplying A and b with an SJLT matrix Φ. All results are run on a supercomputer due to the large scale of the data.



(a) $\sigma_z = 0$

(b) $\sigma_z = 1$

Figure 3.1: Comparison of SOLS and OLS for different settings on synthetic data. The vertical axis corresponds to the scaled objectives $\|A\vartheta^t_{\text{SOLS}} - b\|_2^2/n$ for SOLS and $\|A\vartheta^t_{\text{OLS}} - b\|_2^2/n$ for OLS, where ϑ^t is the update in the t-th iteration. The horizontal axis corresponds to the number of iterations (passes of block-coordinate minimization for all blocks). For both the noiseless case $\sigma_z = 0$ and noisy case $\sigma_z = 1$, we set $n_1 = 10^4$, $n_2 = 10^5$, and $n_3 = 10^6$ respectively.

For synthetic data, we generate the low-rank tensor Θ as follows. For every $d \in [D]$, we generate R orthonormal columns to form $\Theta_d = [\theta_d^{(1)}, \ldots, \theta_d^{(R)}]$ of $[[\Theta_1, \ldots, \Theta_D]] \in \mathcal{S}_{D,R}$ independently. We also generate R positive real scalars $\alpha_1, \ldots, \alpha_R$ uniformly and independently from $[1, 10]$. Then Θ is formed by $\Theta = \sum_{r=1}^R \alpha_r \theta_1^{(r)} \circ \cdots \circ \theta_D^{(r)}$. The sequence of n tensor designs are generated independently with i.i.d. $\mathcal{N}(0,1)$ entries for 10% of the entries chosen uniformly at random, and the remaining entries are set to zero. This allows for fast calculation of the leverage scores of matrix A, as well as memory savings. We also generate the noise z to have i.i.d. $\mathcal{N}(0, \sigma_z^2)$ entries, and the generation of the SJLT matrix Φ follows Definition 5. For both OLS and SOLS, we use random initializations for Θ, i.e., Θ_d has i.i.d. $\mathcal{N}(0,1)$ entries for all $d \in [D]$.

We compare OLS and SOLS for low-rank tensor regression under both the noiseless and noisy scenarios. For the noiseless case, i.e., $\sigma_z = 0$, we choose $R = 3$, $p_1 = p_2 = p_3 =$

Table 3.1: Comparison of SOLS and OLS on CPU execution time (in seconds) and the optimal scaled objective over different choices of sample sizes and noise levels on synthetic data. The results are averaged over 50 random trials, with both the mean values and standard deviations (in parentheses) provided. Note that we terminate the program after the running time exceeds 3×10^4 seconds.

Variance of Noise		$\sigma_z = 0$			$\sigma_z = 1$		
Sample Size		$n = 10^4$	$n = 10^5$	$n = 10^6$	$n = 10^4$	$n = 10^5$	$n = 10^6$
Time	OLS	182.96 (72.357)	3536.9 (1627.0)	$> 3 \times 10^4$ (NA)	166.02 (5.6942)	2620.4 (769.81)	$> 3 \times 10^4$ (NA)
	SOLS	123.43 (37.452)	132.81 (38.653)	134.10 (36.406)	122.641 (34.408)	126.09 (35.719)	127.98 (33.339)
Objective	OLS	$< 10^{-10}$ ($< 10^{-10}$)	$< 10^{-10}$ ($< 10^{-10}$)	$< 10^{-10}$ ($< 10^{-10}$)	0.9089 (0.0217)	0.9430 (0.0182)	0.9440 (0.0137)
	SOLS	$< 10^{-10}$ ($< 10^{-10}$)	$< 10^{-10}$ ($< 10^{-10}$)	$< 10^{-10}$ ($< 10^{-10}$)	0.9414 (0.0264)	0.9854 (0.0227)	0.9891 (0.0232)

Table 3.2: Comparison of SOLS and OLS on CPU execution time (in seconds) and the optimal scaled objective over different choices of ranks on the MRI data. The results are averaged over 10 random trials, with both the mean values and standard deviations (in parentheses) provided.

Rank		$R = 3$	$R = 5$	$R = 10$
Time	OLS	2824.4 (768.08)	8137.2 (1616.3)	26851 (8320.1)
	SOLS	196.31 (68.180)	364.09 (145.79)	761.73 (356.76)
Objective	OLS	16.003 (0.1378)	11.164 (0.1152)	6.8679 (0.0471)
	SOLS	17.047 (0.1561)	11.992 (0.1538)	7.3968 (0.0975)

100, $m = 5 \times R(p_1 + p_2 + p_3) = 4500$, and $s = 200$. Different values of $n \in \{10^4, 10^5, 10^6\}$ are chosen to compare both statistical and computational performances of OLS and SOLS. For the noisy case, the settings of all parameters are identical to those in the noiseless case, except that $\sigma_z = 1$. We provide a plot of the scaled objective versus the number of iterations for some random trials in Figure 3.1. The scaled objective is set as $\|A\vartheta^t_{\text{SOLS}} - b\|_2^2/n$ for SOLS and $\|A\vartheta^t_{\text{OLS}} - b\|_2^2/n$ for OLS, where $\vartheta^t_{\text{SOLS}}$ and ϑ^t_{OLS} are the updates in the t-th iterations of SOLS and OLS respectively. Note the we are using $\|\Phi A\vartheta_{\text{SOLS}} - \Phi b\|_2^2/n$ as the objective for solving the SOLS problem, but looking at the original objective $\|A\vartheta_{\text{SOLS}} - b\|_2^2/n$ for the solution of SOLS is ultimately what one is interested in. Moreover, the gap between $\|\Phi A\vartheta_{\text{SOLS}} - \Phi b\|_2^2/n$ and $\|A\vartheta_{\text{SOLS}} - b\|_2^2/n$ is very small in our results ($< 1\%$). The number of iterations is the number of passes of block-coordinate minimization for all blocks. We can see that OLS and SOLS require approximately the same number of iterations for comparable decrease of objective. However, since the SOLS instance has a much smaller size, its per iteration computational cost is much lower than that of OLS.

We further provide numerical results on the running time (CPU execution time) and the optimal scaled objectives in Table 3.1. Using the same stopping criterion, we see that SOLS and OLS achieve comparable objectives (within $< 5\%$ differences), matching our theory. In terms of the running time, SOLS is much faster than OLS, especially when n is large. For example, when $n = 10^6$, SOLS is orders of magnitude faster than OLS while achieving a comparable objective function value. This matches our discussion on the computational cost of OLS and SOLS. Note that here we suppose the rank is known for our simulation, which can be restrictive in practice. We observe that if we choose a moderately larger rank than the true rank of the underlying model,then the result is similar to what we discussed above. Smaller values of the rank result in a much deteriorated statistical performance for both OLS and SOLS.

In addition, we examine sketching for tensor regression on a real dataset of MRI imaging [113]. The dataset consists of 56 frames of a human brain, each of which is of dimension 128×128 pixels, i.e., $p_1 = p_2 = 128$ and $p_3 = 56$. The generation of design tensors $\{A_i\}$ and linear measurements b follows the same settings as for the synthetic data, with $\sigma_z = 0$. We choose three values of $R = 3, 5, 10$, and set $m = 5 \times R(p_1 + p_2 + p_3)$. The sample size is set to $n = 10^4$ for all settings of R. Analogous to the synthetic data, we provide numerical results for SOLS and OLS on the running time (CPU execution time) and the optimal scaled objectives. Again, we have that SOLS is much faster than OLS when they achieve comparable optimal objectives, under all settings of ranks.

Chapter 4

On Fast Convergence of Proximal Algorithms for SQRT-Lasso Optimization

4.1 Introduction

Many statistical machine learning methods can be formulated as optimization problems in the following form

$$\min_{\theta} \mathcal{L}(\theta) + \mathcal{R}(\theta), \tag{4.1}$$

where $\mathcal{L}(\theta)$ is a loss function and $\mathcal{R}(\theta)$ is a regularizer. When the loss function is smooth and has a Lipschitz continuous gradient, (4.1) can be efficiently solved by simple proximal gradient descent and proximal Newton algorithms (also requires a Lipschitz continuous Hessian matrix of $\mathcal{L}(\theta)$). Some statistical machine learning methods, however, sacrifice convenient computational structures to gain estimation robustness and modeling flexibility or the other way round [114–118]. Taking SVM as an example, the hinge loss function gains estimation robustness, but sacrifices the smoothness (compared with the square hinge loss function). However, by exploring the structure of the problem, we find that these "sacrifices" do not always require more computational efforts.

Advantage of SQRT-Lasso over Lasso. To shed a light of such a “free-lunch” phenomenon, we study the high dimensional square-root (SQRT) Lasso regression problem [115, 119]. Specifically, we consider a sparse linear model in high dimensions,

$$y = X\theta^* + \epsilon,$$

where $X \in \mathbb{R}^{n\times d}$ is the design matrix, $y \in \mathbb{R}^n$ is the response vector, $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ is the random noise, and θ^* is the sparse unknown regression coefficient vector. To estimate θ^*, [120] propose the well-known Lasso estimator by solving

$$\overline{\theta}^{\mathsf{Lasso}} = \operatorname*{argmin}_{\theta} \frac{1}{n}\|y - X\theta\|_2^2 + \lambda_{\mathsf{Lasso}}\|\theta\|_1, \tag{4.2}$$

where λ_{Lasso} is the regularization parameter. Existing literature shows that given

$$\lambda_{\mathsf{Lasso}} \asymp \sigma\sqrt{\frac{\log d}{n}}, \tag{4.3}$$

$\overline{\theta}^{\mathsf{Lasso}}$ is minimax optimal for parameter estimation in high dimensions. Note that the optimal regularization parameter for Lasso in (4.3), however, requires the prior knowledge of the unknown parameter σ. This requires the regularization parameter to be carefully tuned over a wide range of potential values to get a good finite-sample performance.

To overcome this drawback, [115] propose the SQRT-Lasso estimator by solving

$$\overline{\theta}^{\mathsf{SQRT}} = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \frac{1}{\sqrt{n}}\|y - X\theta\|_2 + \lambda_{\mathsf{SQRT}}\|\theta\|_1, \tag{4.4}$$

where λ_{SQRT} is the regularization parameter. They further show that $\overline{\theta}^{\mathsf{SQRT}}$ is also minimax optimal in parameter estimation, but the optimal regularization parameter is

$$\lambda_{\mathsf{SQRT}} \asymp \sqrt{\frac{\log d}{n}}. \tag{4.5}$$

Since (4.5) no longer depends on σ, SQRT-Lasso eases tuning effort.

Extensions of SQRT-Lasso. Besides the tuning advantage, the regularization selection for SQRT-Lasso type methods is also adaptive to inhomogeneous noise. For example, [116] propose a multivariate SQRT-Lasso for sparse multitask learning. Given a matrix $A \in \mathbb{R}^{d\times d}$, let A_{*k} denote the k-th column of A, and A_{i*} denote the i-th row of A. Specifically, [116] consider a multitask regression model

$$Y = X\Theta^* + W,$$

where $X \in \mathbb{R}^{n\times d}$ is the design matrix, $Y \in \mathbb{R}^{n\times m}$ is the response matrix, $W_{*k} \sim N(\mathbf{0}, \sigma_k^2 \mathbf{I}_n)$ is the random noise, and $\Theta^* \in \mathbb{R}^{d\times m}$ is the unknown row-wise sparse coefficient matrix, i.e., Θ^* has many rows with all zero entries. To estimate Θ^*, [116] propose a calibrated multivariate regression (CMR) estimator by solving

$$\overline{\theta}^{\mathsf{CMR}} = \mathop{\mathrm{argmin}}_{\theta \in \mathbb{R}^{d\times m}} \frac{1}{\sqrt{n}} \sum_{k=1}^{m} \|Y_{*k} - X\Theta_{*k}\|_2 + \lambda_{\mathsf{CMR}} \|\Theta\|_{1,2},$$

where $\|\Theta\|_{1,2} = \sum_{j=1}^{d} \|\Theta_{j*}\|_2$. [116] further shows that the regularization of CMR approach is adaptive to σ_k's for each regression task, i.e., $Y_{*k} = X\Theta_{*k}^* + W_{*k}$, and therefore CMR achieves better performance in parameter estimation and variable selection than its least square loss based counterpart. With a similar motivation, [121] propose a nodewise SQRT-Lasso approach for sparse precision matrix estimation. Due to space limit, please refer to [121] for more details.

Existing Algorithms for SQRT-Lasso Optimization. Despite of these good properties, in terms of optimization, (4.4) for SQRT-Lasso is computationally more challenging than (4.2) for Lasso. The ℓ_2 loss in (4.4) is not necessarily differentiable, and does not have a Lipschitz continuous gradient, compared with the least square loss in (4.2). A few algorithms have been proposed for solving (4.4) in existing literature, but none of them are satisfactory when n and d are large. [115] reformulate (4.4) as a second order cone program (SOCP) and solve by an interior point method with a computational cost of $\mathcal{O}(nd^{3.5}\log(\epsilon^{-1}))$, where ϵ is a pre-specified optimization accuracy; [118] solve (4.4) by an alternating direction method of multipliers (ADMM) algorithm with a computational cost of $\mathcal{O}(nd^2/\epsilon)$; [119] propose to solve the variational form of (4.4) by an alternating minimization algorithm, and [122] further develop a coordinate descent subroutine to accelerate its computation. However, no iteration complexity is established in [122]. Our numerical study shows that their algorithm only scales to moderate

Table 4.1: Comparison with existing algorithms for solving SQRT-Lasso. SOCP: Second-order Cone Programming; TRM: Trust Region Newton; VAM: Variational Alternating Minimization; ADMM: Alternating Direction Method of Multipliers; VCD: Coordinate Descent; Prox-GD: Proximal Gradient Descent; Prox-Newton: Proximal Newton.

	Algorithm	Theoretical Guarantee	Empirical Performance
[115]	SOCP + TRM	$\mathcal{O}(nd^{3.5}\log(\epsilon^{-1}))$	Very Slow
[119]	VAM	N.A.	Very Slow
[118]	ADMM	$\mathcal{O}(nd^2/\epsilon)$	Slow
[122]	VAM + CD	N.A.	Moderate
Ours	Pathwise Prox-GD	$\mathcal{O}(nd\log(\epsilon^{-1}))$	Fast
Ours	Pathwise Prox-Newton + CD	$\mathcal{O}(snd\log\log(\epsilon^{-1}))$	Very Fast

Remark: [122] requires a good initial guess of σ to achieve moderate performance. Otherwise, its empirical performance is similar to ADMM.

problems. Moreover, [122] require a good initial guess for the lower bound of σ. When the initial guess is inaccurate, the empirical convergence can be slow.

Our Motivations. The major drawback of the aforementioned algorithms is that they do not explore the modeling structure of the problem. The ℓ_2 loss function is not differentiable only when the model are overfitted, i.e., the residuals are zero values $y - X\theta = \mathbf{0}$. Such an extreme scenario rarely happens in practice, especially when SQRT-Lasso is equipped with a sufficiently large regularization parameter λ_{SQRT} to yield a sparse solution and prevent overfitting. Thus, we can treat the ℓ_2 loss as an "almost" smooth function. Moreover, our theoretical investigation indicates that the ℓ_2 loss function also enjoys the restricted strong convexity, smoothness, and Hessian smoothness. In other words, the ℓ_2 loss function behaves as a strongly convex and smooth over a sparse domain. An illustration is provided in Figure 4.1.

Our Contributions. Given these nice geometric properties of the ℓ_2 loss function, we can directly solve (4.4) by proximal gradient descent (Prox-GD), proximal Newton (Prox-Newton), and proximal Quasi-Newton (Prox-Quasi-Newton) algorithms [123,124]. Existing literature only apply these algorithms to solve optimization problems in statistical machine learning when the loss function is smooth. Our theoretical analysis

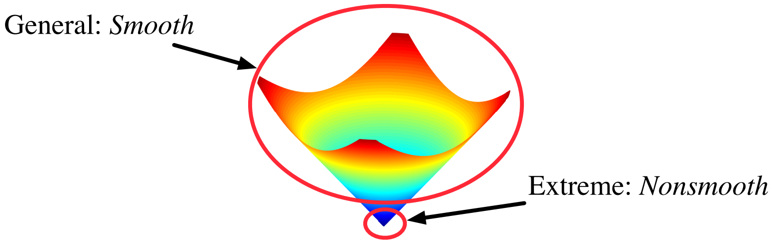


Figure 4.1: The extreme and general cases of the ℓ_2 loss. The nonsmooth region $\{\theta : y - X\theta = 0\}$ is out of our interest, since it corresponds to those overfitted regression models

shows that both algorithms enjoy strong convergence guarantees [125]. Specifically, the Prox-GD algorithm achieves a local linear convergence and the Prox-Newton algorithm achieves a local quadratic convergence. To further ensure global strong convergence, we combine these two algorithms with the pathwise optimization scheme, which solves (4.4) with a decreasing sequence of regularization parameters, $\lambda_0 \geq \ldots \geq \lambda_N$ with $\lambda_N = \lambda_{\mathsf{SQRT}}$. The pathwise optimization scheme helps yield sparse solutions and avoid overfitting throughout all iterations. Besides sparse linear regression, we extend our algorithms and theory to sparse multitask regression and sparse precision matrix estimation. Extensive numerical results show our algorithms uniformly outperform the competing algorithms.

Hardness of Analysis. We highlight that our local analysis with strong convergence guarantees are novel and highly nontrivial for solving the SQRT-Lasso problem using simple and efficient proximal algorithms. First of all, sophisticated analysis is required to demonstrate the restricted strong convexity/smoothness and Hessian smoothness of the ℓ_2 loss function over a neighborhood of the underlying model parameter θ^* in high dimensions. These are key properties for establishing the strong convergence rates of proximal algorithms. Moreover, it is involved to guarantee that the output solution of the proximal algorithms do not fall in the nonsmooth region of the ℓ_2 loss function. This is important in guaranteeing the favored computational and statistical properties. In addition, it is highly technical to show that the pathwise optimization does enter the strong convergence region at certain stage. We defer all detailed analysis to the appendix.

Notations. Given a vector $v \in \mathbb{R}^d$, we define the subvector of v with the j-th entry removed as $v_{\backslash j} \in \mathbb{R}^{d-1}$. Given an index set $\mathcal{I} \subseteq \{1, ..., d\}$, let $\overline{\mathcal{I}}$ be the complementary set to $\mathcal{I}$ and $v_{\mathcal{I}}$ be a subvector of v by extracting all entries of v with indices in $\mathcal{I}$. Given a matrix $A \in \mathbb{R}^{d\times d}$, we denote A_{*j} (A_{k*}) the j-th column (k-th row), $A_{\backslash i \backslash j}$ as a submatrix of A with the i-th row and the j-th column removed and $A_{\backslash ij}$ ($A_{i\backslash j}$) as the j-th column (i-th row) of A with its i-th entry (j-th entry) removed. Let $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ be the largest and smallest eigenvalues of A respectively. Given an index set $\mathcal{I} \subseteq \{1, ..., d\}$, we use $A_{\mathcal{I}\mathcal{I}}$ to denote a submatrix of A by extracting all entries of A with both row and column indices in $\mathcal{I}$. We denote $A \succ 0$ if A is a positive-definite matrix. Given two real sequences $\{A_n\}, \{a_n\}$, we use conventional notations $A_n = \mathcal{O}(a_n)$ (or $A_n = \Omega(a_n)$) denote the limiting behavior, ignoring constant, $\tilde{\mathcal{O}}$ to denote limiting behavior further ignoring logarithmic factors, and $\mathcal{O}_P(\cdot)$ to denote the limiting behavior in probability. $A_n \asymp a_n$ if $A_n = \mathcal{O}(a_n)$ and $A_n = \Omega(a_n)$ simultaneously. Given a vector $\mathbf{x} \in \mathbb{R}^d$ and a real value $\lambda > 0$, we denote the soft thresholding operator $S_\lambda(\mathbf{x}) = [\mathrm{sign}(x_j)\max\{|x_j| - \lambda, 0\}]_{j=1}^d$.

4.2 Algorithm

We review the Prox-GD and Prox-Newton algorithms. For convenience, we denote

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

where $\mathcal{L}(\theta) = \frac{1}{\sqrt{n}}\|y - X\theta\|_2$. Since SQRT-Lasso is equipped with a sufficiently large regularization parameter λ to prevent overfitting, i.e., $y - X\theta \neq 0$, we treat $\mathcal{L}(\theta)$ as a differentiable function in this section. Formal theoretical justifications will be provided in the next section.

4.2.1 Proximal Gradient Desccent Algorithm

Given $\theta^{(t)}$ at t-th iteration, we consider a quadratic approximation of $\mathcal{F}_\lambda(\theta)$ at $\theta = \theta^{(t)}$ as

$$\mathcal{Q}_\lambda(\theta, \theta^{(t)}) = \mathcal{L}(\theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)})^\top(\theta - \theta^{(t)}) + \frac{L^{(t)}}{2}\|\theta - \theta^{(t)}\|_2^2 + \lambda\|\theta\|_1, \tag{4.6}$$

where $L^{(t)}$ is a step size parameter determined by the backtracking line search. We then take

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmin}}\, \mathcal{Q}_\lambda(\theta, \theta^{(t)}) = \mathcal{S}_{\frac{\lambda}{L^{(t)}}}\left(\theta^{(t)} - \frac{\nabla\mathcal{L}(\theta^{(t)})}{L^{(t)}}\right).$$

For simplicity, we denote $\theta^{(t+1)} = \mathcal{T}_{L^{(t+1)},\lambda}(\theta^{(t)})$. Given a pre-specified precision ε, we terminate the iterations when the approximate KKT condition holds:

$$\omega_\lambda(\theta^{(t)}) = \min_{g \in \partial\|\theta^{(t)}\|_1} \|\nabla\mathcal{L}(\theta^{(t)}) + \lambda g\|_\infty \leq \varepsilon. \tag{4.7}$$

4.2.2 Proximal Newton Algorithm

Given $\theta^{(t)}$ at t-th iteration, we denote a quadratic term of θ as

$$\|\theta - \theta^{(t)}\|^2_{\nabla^2\mathcal{L}(\theta^{(t)})} = (\theta - \theta^{(t)})^\top \nabla^2\mathcal{L}(\theta^{(t)})(\theta - \theta^{(t)}),$$

and consider a quadratic approximation of $\mathcal{F}_\lambda(\theta)$ at $\theta = \theta^{(t)}$ is

$$\mathcal{Q}_\lambda(\theta, \theta^{(t)}) = \mathcal{L}(\theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)})^\top(\theta - \theta^{(t)}) + \frac{1}{2}\|\theta - \theta^{(t)}\|^2_{\nabla^2\mathcal{L}(\theta^{(t)})} + \lambda\|\theta\|_1. \tag{4.8}$$

We then take

$$\theta^{(t+0.5)} = \underset{\theta}{\operatorname{argmin}}\, \mathcal{Q}_\lambda(\theta, \theta^{(t)}). \tag{4.9}$$

An additional backtracking line search procedure is required to obtain

$$\theta^{(t+1)} = \theta^{(t)} + \eta_t(\theta^{(t+0.5)} - \theta^{(t)}),$$

which guarantees $\mathcal{F}_\lambda(\theta^{(t+1)}) \leq \mathcal{F}_\lambda(\theta^{(t)})$. The termination criterion for Prox-Newton is same with (4.7).

Remark 1. The ℓ_1 regularized quadratic problem in (7.160) can be solved efficiently by the coordinate descent algorithm combined with the active set strategy. See more details in [126]. The computational cost is $\tilde{\mathcal{O}}(snd)$, where $s \ll d$ is the solution sparsity.

Algorithm 4 Prox-GD algorithm for solving the SQRT-Lasso optimization (4.4). We treat $\mathcal{L}(\theta)$ as a differentiable function.

```
Input: $y$, $X$, $\lambda$, $\varepsilon$, $L_{\max} > 0$
Initialize: $\theta^{(0)}$, $t \leftarrow 0$, $L^{(0)} \leftarrow L_{\max}$, $\tilde{L}^{(0)} \leftarrow L^{(0)}$
Repeat: $t \leftarrow t + 1$
    Repeat: (Line Search)
        $\theta^{(t)} \leftarrow \mathcal{T}_{\tilde{L}^{(t)},\lambda}(\theta^{(t-1)})$
        If $\mathcal{F}_\lambda(\theta^{(t)}) < \mathcal{Q}_\lambda(\theta^{(t)}, \theta^{(t-1)})$
        Then $\tilde{L}^{(t)} \leftarrow \frac{\tilde{L}^{(t)}}{2}$
    Until: $\mathcal{F}_\lambda(\theta^{(t)}) \geq \mathcal{Q}_\lambda(\theta^{(t)}, \theta^{(t-1)})$
    $L^{(t)} \leftarrow \min\{2\tilde{L}^{(t)}, L_{\max}\}$, $\tilde{L}^{(t)} \leftarrow L^{(t)}$
    $\theta^{(t)} \leftarrow \mathcal{T}_{L^{(t)},\lambda}(\theta^{(t-1)})$
Until: $\omega_\lambda(\theta^{(t)}) \leq \varepsilon$
Return: $\hat{\theta} \leftarrow \theta^{(t)}$
```

Details of Prox-GD and Prox-Newton algorithms are summarized in Algorithms 4 and 5 respectively. To facilitate global fast convergence, we further combine the pathwise optimization [127] with the proximal algorithms. See more details in Section 4.4.

Remark 2. We can also apply proximal quasi-Newton method. Accordingly, at each iteration, the Hessian matrix in (4.8) is replaced with an approximation. See [128] for more details.

4.3 Computational Analysis

We start with defining the locally restricted strong convexity/smoothness and Hessian smoothness.

Definition 8. Let $\mathcal{B}_r = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_2^2 \leq r\}$ for some constant $r \in \mathbb{R}^+$. For any $v, w \in \mathcal{B}_r$ satisfying $\|v - w\|_0 \leq s$, $\mathcal{L}$ is *locally restricted strongly convex* (LRSC), *smooth* (LRSS), and *Hessian smooth* (LRHS) respectively on $\mathcal{B}_r$ at sparsity level s, if

Algorithm 5 Prox-Newton algorithm for solving the SQRT-Lasso optimization (4.4). We treat $\mathcal{L}(\theta)$ as a differentiable function.

```
Input: $y$, $X$, $\lambda$, $\varepsilon$
Initialize: $\theta^{(0)}$, $t \leftarrow 0$, $\mu \leftarrow 0.9$, $\alpha \leftarrow \frac{1}{4}$
Repeat: $t \leftarrow t+1$
  $\theta^{(t)} \leftarrow \mathrm{argmin}_\theta\, \mathcal{Q}_\lambda(\theta, \theta^{(t-1)})$
  $\Delta\theta^{(t)} \leftarrow \theta^{(t)} - \theta^{(t-1)}$
  $\gamma_t \leftarrow \nabla\mathcal{L}\left(\theta^{(t-1)}\right)^\top \Delta\theta^{(t)} + \lambda\left(\|\theta^{(t)}\|_1 - \|\theta^{(t-1)}\|_1\right)$
  $\eta_t \leftarrow 1$, $q \leftarrow 0$
  Repeat: $q \leftarrow q+1$ (Line Search)
    $\eta_t \leftarrow \mu^q$
  Until $\mathcal{F}_\lambda\left(\theta^{(t-1)} + \eta_t\Delta\theta^{(t)}\right) \leq \mathcal{F}_\lambda\left(\theta^{(t-1)}\right) + \alpha\eta_t\gamma_t$
  $\theta^{(t)} \leftarrow \theta^{(t)} + \eta_t\Delta\theta^{(t-1)}$
Until: $\omega_\lambda(\theta^{(t)}) \leq \varepsilon$
Return: $\hat{\theta} \leftarrow \theta^{(t)}$
```

there exist universal constants $\rho_s^-, \rho_s^+, L_s \in (0, \infty)$ such that

$$\begin{aligned} &\text{LRSC: } \mathcal{L}(v) - \mathcal{L}(w) - \nabla\mathcal{L}(w)^\top(v-w) \geq \frac{\rho_s^-}{2}\|v-w\|_2^2, \\ &\text{LRSS: } \mathcal{L}(v) - \mathcal{L}(w) - \nabla\mathcal{L}(w)^\top(v-w) \leq \frac{\rho_s^+}{2}\|v-w\|_2^2, \\ &\text{LRHS: } \sup_{\|\mathbf{u}\|_0 \leq s, \|\mathbf{u}\|_2 = 1} v^\top(\nabla^2\mathcal{L}(v) - \nabla^2\mathcal{L}(w))v \leq L_s\|v - w'\|_2^2, \end{aligned} \tag{4.10}$$

We define the locally restricted condition number at sparsity level s as $\kappa_s = \frac{\rho_s^+}{\rho_s^-}$.

LRSC and LRSS are locally constrained variants of restricted strong convexity and smoothness [129, 130], which are keys to establishing the strong convergence guarantees in high dimensions. The LRHS is parallel to the local Hessian smoothness for analyzing the proximal Newton algorithm in low dimensions [124]. This is also close related to the self-concordance [131] in the analysis of Newton method [132].

Next, we prove that the ℓ_2 loss of SQRT-Lasso enjoys the good geometric properties defined in Definition 8 under mild modeling assumptions.

Lemma 2. Suppose $\|\theta^*\|_0 = s^*$ and $\lambda = C_1\sqrt{\frac{\log d}{n}}$, then with high probability, we have

$$\lambda \geq \frac{C_1}{4}\|\nabla\mathcal{L}(\theta^*)\|_\infty.$$

Moreover, given each row of the design matrix X independently sampled from a sub-Gaussian distribution with the positive definite covariance matrix $\mathbf{\Sigma}_X \in \mathbb{R}^{d\times d}$ with bounded eigenvalues. Then for

$$n \geq C_2 s^* \log d,$$

$\mathcal{L}(\theta)$ satisfies LRSC, LRSS, and LRHS properties on $\mathcal{B}_r$ with high probability. Specifically, (4.10) holds with

$$\rho^+_{s^*+2\tilde{s}} \leq \frac{C_3}{\sigma},\ \rho^-_{s^*+2\tilde{s}} \geq \frac{C_4}{\sigma} \text{ and } L_{s^*+2\tilde{s}} \leq \frac{C_5}{\sigma},$$

where $\tilde{s} > C_6(\kappa^2_{s^*+2\tilde{s}} + \kappa_{s^*+2\tilde{s}})s^*$ and $r \geq \frac{C_7 s^* \lambda^2}{(\rho^-_{s^*+2\tilde{s}})^2}$. $C_1, \ldots, C_7 \in \mathbb{R}^+$ are generic constants.

Lemma 2 guarantees that with high probability: (i) λ is sufficiently large to eliminate the irrelevant variables and yields sufficiently sparse solutions [133, 134]; (ii) LRSC, LRSS, and LRHS hold for the ℓ_2 loss of SQRT-Lasso such that fast convergence of the proximal algorithms can be established.

4.3.1 Linear Convergence of Prox-GD

For notational simplicity, we denote

$$\mathcal{S}^* = \{j\ :\ \theta^*_j \neq 0\},\ \overline{\mathcal{S}}^* = \{j\ :\ \theta^*_j = 0\},\ \text{and } \mathcal{B}_r^{s^*+\tilde{s}} = \mathcal{B}_r \cap \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_0 \leq s^* + \tilde{s}\}.$$

To ease the analysis, we provide a local convergence analysis when $\theta \in \mathcal{B}_r^{s^*+\tilde{s}}$. The convergence of Prox-GD is presented as follows.

Theorem 9. Suppose λ, X, and n satisfy conditions in Lemma 2. Given $\|\theta^{(0)} - \theta^*\|_2^2 \leq r$, we have sufficiently sparse solutions throughout all iterations, i.e.,

$$\|[\theta^{(t)}]_{\mathcal{S}_c}\|_0 \leq s^* + \tilde{s}.$$

Moreover, given $\varepsilon > 0$, we need at most

$$T = \mathcal{O}\left(\kappa_{s^*+2\tilde{s}} \log\left(\frac{\kappa^3_{s^*+2\tilde{s}} s^* \lambda^2}{\varepsilon^2}\right)\right)$$

iterations to guarantee that the output solution $\hat{\theta}$ satisfies

$$\|\hat{\theta}-\overline{\theta}\|_2^2 = \mathcal{O}\left(\left(1-\frac{1}{8\kappa_{s^*+2\tilde{s}}}\right)^T \varepsilon\lambda s^*\right) \quad \text{and}$$
$$\mathcal{F}_\lambda(\hat{\theta})-\mathcal{F}_\lambda(\overline{\theta}) = \mathcal{O}\left(\left(1-\frac{1}{8\kappa_{s^*+2\tilde{s}}}\right)^T \varepsilon\lambda s^*\right),$$

where $\overline{\theta}$ is the unique sparse global optimum to (4.4) with

$$\|[\overline{\theta}]_{\overline{\mathcal{S}}^*}\|_0 \leq s^* + \tilde{s}.$$

Theorem 9 guarantees that the Prox-GD algorithm achieves a local linear convergence to the unique sparse global optimum to (4.4).

4.3.2 Quadratic Convergence of Prox-Newton

We then present the convergence analysis of the Prox-Newton algorithm as follows.

Theorem 10. Suppose λ, X, and n satisfy conditions in Lemma 2. Given $\|\theta^{(0)}-\theta^*\|_2^2 \leq r$, we have sufficiently sparse solutions throughout all iterations, i.e.,

$$\|[\theta^{(t)}]_{\mathcal{S}_c}\|_0 \leq s^* + \tilde{s}.$$

Moreover, given $\varepsilon > 0$, we need at most

$$T = \mathcal{O}\left(\log\log\left(\frac{3\rho^+_{s^*+2\tilde{s}}}{\varepsilon}\right)\right)$$

iterations to guarantee that the output solution $\hat{\theta}$ satisfies

$$\|\hat{\theta}-\overline{\theta}\|_2^2 = \mathcal{O}\left(\left(\frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}}\right)^{2^T} \varepsilon\lambda s^*\right) \quad \text{and}$$
$$\mathcal{F}_\lambda(\hat{\theta})-\mathcal{F}_\lambda(\overline{\theta}) = \mathcal{O}\left(\left(\frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}}\right)^{2^T} \varepsilon\lambda s^*\right),$$

where $\overline{\theta}$ is the unique sparse global optimum to (4.4).

Theorem 10 guarantees that the Prox-Newton algorithm achieves a local quadratic convergence to the unique sparse global optimum to (4.4).

Remark 3. Our analysis can be further extended to the proximal quasi-Newton algorithm. The only technical difference is controlling the error of the Hessian approximation under restricted spectral norm.

4.4 Global Fast Convergence via Pathwise Optimization Scheme

In this section, we explain how the pathwise optimization scheme extends the local fast convergence guarantees established in Section 3 to the global setting. The pathwise optimization is essentially a multistage optimization scheme for boosting the computational performance [126, 127, 130].

Specifically, we solve (4.4) using a geometrically decreasing sequence of regularization parameters

$$\lambda_0 > \lambda_1 > \ldots > \lambda_N,$$

where λ_N is the target regularization parameter of SQRT-Lasso. This yields a sequence of output solutions

$$\hat{\theta}_{[0]},\ \hat{\theta}_{[1]}, \ldots,\ \hat{\theta}_{[N]},$$

also known as the solution path. At the K-th optimization stage, we choose $\hat{\theta}_{[K-1]}$ (the output solution of the $(K-1)$-th stage) as the initial solution, and solve (4.4) with $\lambda = \lambda_K$ using the proximal algorithms. This is also referred as the warm start initialization in existing literature [127]. Details of the pathwise optimization is summarized in Algorithm 6.

Before we proceed, we first characterize the statistical properties for the output solutions of the proximal algorithms.

Theorem 11. Suppose λ, X, and n satisfy conditions in Lemma 2. If the output solution $\hat{\theta}$ satisfies

$$\omega_\lambda(\hat{\theta}) \leq \varepsilon = \mathcal{O}(\frac{\sigma s^* \log d}{n}),$$

Algorithm 6 The pathwise optimization scheme for the proximal algorithms. We solve the optimization problem using a geometrically decreasing sequence of regularization parameters.

```
Input: $y$, $X$, $N$, $\lambda_{[N]}$, $\varepsilon_{[N]}$
Initialize: $\hat{\theta}_{[0]} \leftarrow 0$, $\lambda_{[0]} \leftarrow \|\nabla\mathcal{L}(0)\|_\infty$, $\eta_\lambda \leftarrow \left(\frac{\lambda_{[N]}}{\lambda_{[0]}}\right)^{\frac{1}{N}}$
For: $K = 1, \ldots, N$
  $\lambda_{[K]} \leftarrow \eta_\lambda \lambda_{[K-1]}$, $\theta^{(0)}_{[K]} \leftarrow \hat{\theta}_{[K-1]}$, $\varepsilon_{[K]} \leftarrow \varepsilon_{[N]}$
  $\hat{\theta}_{[K]} \leftarrow \text{Prox-Alg}\left(y, X, \lambda_{[K]}, \theta^{(0)}_{[K]}, \varepsilon_{[K]}\right)$
End For
Return: $\hat{\theta}_{[N]}$
```

then we have:

$$\|\hat{\theta} - \theta^*\|_2 = \mathcal{O}_P\left(\sigma\sqrt{\frac{s^* \log d}{n}}\right) \quad \text{and} \quad \|\hat{\theta} - \theta^*\|_1 = \mathcal{O}_P\left(\sigma s^* \sqrt{\frac{\log d}{n}}\right).$$

Moreover, we have

$$|\hat{\sigma} - \sigma| = \mathcal{O}_P\left(\frac{\sigma s^* \log d}{n}\right), \text{ where } \hat{\sigma} = \frac{\|y - X\hat{\theta}\|_2}{\sqrt{n}}.$$

Theorem 11 guarantees that the output solution $\hat{\theta}$ obtained from Algorithm 4 and 5 achieves the minimax optimal rate of convergence in parameter estimation [135, 136]. Moreover, Theorem 11 implies that with sufficiently large regularization parameter, the mean square error satisfies

$$\frac{1}{n}\|y - X\hat{\theta}\|_2^2 = \Omega(\sigma^2) > 0.$$

This guarantees that the obtained model is not overfitted, and the output solution is far from the nonsmooth region of the ℓ_2 loss, i.e., the set $\{\theta : y - X\theta = 0\}$.

As can be seen in Algorithm 3, the pathwise optimization scheme starts with

$$\lambda_0 = \|\nabla\mathcal{L}(0)\|_\infty = \left\|\frac{X^\top y}{\sqrt{n}\|y\|_2}\right\|_\infty,$$

which yields an all zero solution $\hat{\theta}_{[0]} = 0$ (null fit). As the regularization parameter

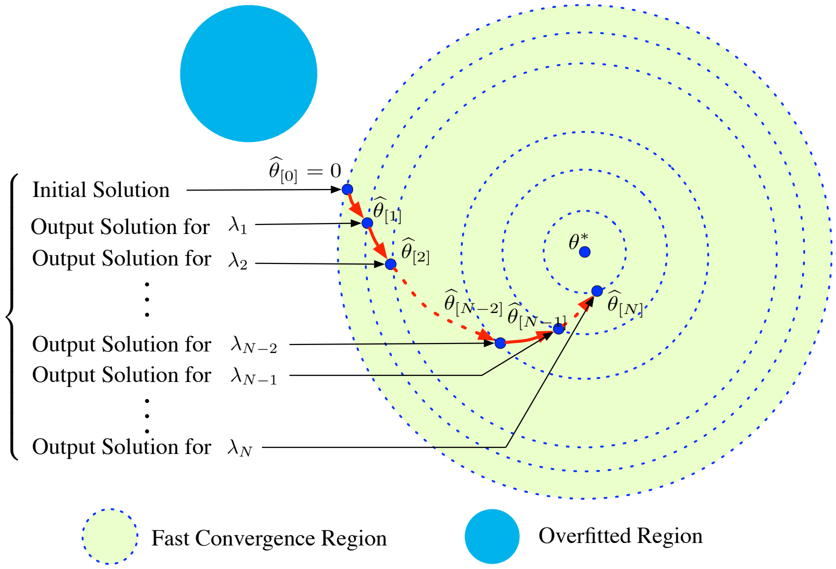

$\widehat{\theta}_{[0]} = 0$
Initial Solution
Output Solution for λ_1
Output Solution for λ_2
Output Solution for λ_{N-2}
Output Solution for λ_{N-1}
Output Solution for λ_N
$\widehat{\theta}_{[1]}$
$\widehat{\theta}_{[2]}$
θ^*
$\widehat{\theta}_{[N-2]}$
$\widehat{\theta}_{[N-1]}$
$\widehat{\theta}_{[N]}$
Fast Convergence Region
Overfitted Region

Figure 4.2: A geometric illustration for the fast convergence of the proximal algorithms. The proximal algorithms combined with the pathwise optimization scheme suppress the overfitting and yield sparse solutions along the solution path. Therefore, the nonsmooth region of the ℓ_2 loss, i.e., the set $\{\theta : y - X\theta = 0\}$, is avoided, and LRSC, LRSS, and LRHS enable the proximal algorithms to achieve fast convergence.

gradually decreases, the number of nonzero coordinates gradually increases. Throughout stages, the regularization parameters are sufficiently large to suppress the overfitting and yield sparse solutions along the solution path. Thus LRSC, LRSS, and LRHS are also expected to hold along the entire solution path, and the proximal algorithms achieve fast convergence. Note that when the design X is normalized, we have $\lambda_0 = \mathcal{O}(d)$, which implies that the total number N of regularization parameter satisfies

$$N = \mathcal{O}(\log d).$$

A geometric illustration of the pathwise optimization is provided in Figure 5.3.

4.5 Extension to CMR and SPME

We extend our algorithm and theory to calibrated multivariate regression (CMR, [116]) and sparse precision matrix estimation (SPME, [121]). Due to space limit, we only provide a brief discussion and omit the detailed theoretical deviation.

Extension to CMR. Recall that CMR solves

$$\overline{\theta}^{\mathsf{CMR}} = \underset{\theta \in \mathbb{R}^{d\times m}}{\operatorname{argmin}} \frac{1}{\sqrt{n}} \sum_{k=1}^{m} \|Y_{*k} - X\Theta_{*k}\|_2 + \lambda_{\mathsf{CMR}} \|\Theta\|_{1,2}.$$

Similar to SQRT-Lasso, we choose a sufficiently large λ_{CMR} to prevent overfitting. Thus, we can expect

$$\|Y_{*k} - X\Theta_{*k}\|_2 \neq 0 \text{ for all } k = 1, ..., m,$$

and treat the nonsmooth loss of CMR as a differentiable function. Accordingly, we can trim our algorithms and theory for the nonsmooth loss of CMR, and establish fast convergence guarantees, as we discussed in §4.4.

Extension to SPME. [121] show that a $d \times d$ sparse precision matrix estimation problem is equivalent to a collection of d sparse linear model estimation problems. For each linear model, we apply SQRT-Lasso to estimate the regression coefficient vector and the standard deviation of the random noise. Since SQRT-Lasso is adaptive to imhomogenous noise, we can use one singular regularization parameter to prevent overfitting for all SQRT-Lasso problems. Accordingly, we treat the nonsmooth loss function in every SQRT-Lasso problem as a differentiable function, and further establish fast convergence guarantees for the proximal algorithms combined with the pathwise optimization scheme.

4.6 Numerical Experiments

We compare the computational performance of the proximal algorithms with other competing algorithms using both synthetic and real data. All algorithms are implemented in `C++` with double precision using a PC with an Intel 2.4GHz Core i5 CPU and 8GB memory. All algorithms are combined with the pathwise optimization scheme to boost the computational performance. Due to space limit, we omit some less important details.

Synthetic Data: For synthetic data, we generate a training dataset of 200 samples, where each row of the design matrix $X_{i*}$is independently from a 2000-dimensional normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}_{jj} = 1$ and $\mathbf{\Sigma}_{jk} = 0.5$ for all $k \neq j$. We set $s^* = 3$ with $\theta_1^* = 3$, $\theta_2^* = -2$, and $\theta_4^* = 1.5$, and $\theta_j^* = 0$ for all $j \neq 1, 2, 4$. The response vector

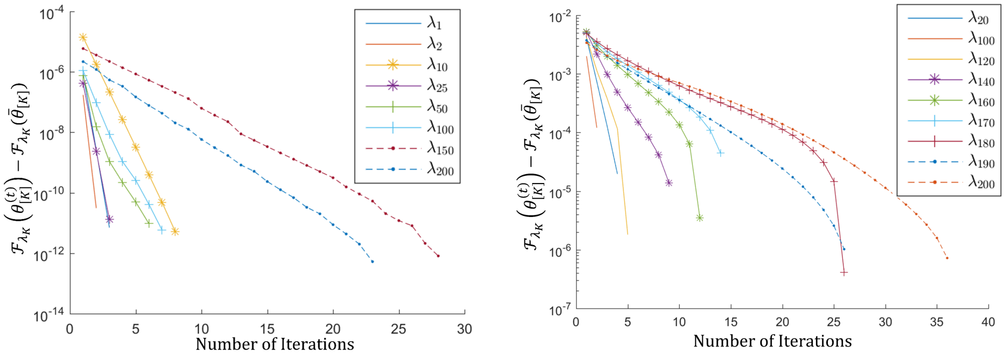


Figure 4.3: The objective gap v.s. the number of iterations. We can see that the Prox-GD (Left) and Prox-Newton (Right) algorithms achieve linear and quadratic convergence at every stage respectively.

Table 4.2: Computational performance of Prox-GD on synthetic data under different choices of variance σ, the number of stages N, and the stopping criterion ε_N. The training time is presented, where each entry is the mean execution time in seconds over 100 random trials. The minimal mean square error (MSE) is $\frac{1}{n}\|y - X\hat{\theta}_{[K]}\|_2^2$, where $\hat{\theta}_{[K]}$ is the optimal solution that attains $\min \mathcal{F}_{\lambda_K}(\theta)$ for all stages $K = 1, \ldots, N$.

σ	N	ε_N			Min. MSE	σ	ε_N			Min. MSE
		10^{-4}	10^{-5}	10^{-6}			10^{-4}	10^{-5}	10^{-6}	
0.1	1	0.3718	0.3721	0.3647	0.0132	0.5	0.2850	0.2951	0.2886	0.3054
	10	**0.2749**	**0.2764**	**0.2804**			**0.1646**	**0.1698**	**0.1753**	
	30	0.3364	0.3452	0.3506			0.2207	0.2247	0.2285	
1	1	0.2347	0.2478	0.2618	1.1833	2	0.4317	0.4697	0.4791	4.2197
	10	**0.1042**	**0.1031**	**0.1091**			**0.1661**	**0.1909**	**0.2110**	
	30	0.2172	0.2221	0.2199			0.2701	0.2955	0.3134	

Table 4.3: Timing comparison between multiple algorithms on real data. Each entry is the execution time in seconds. All experiments are conducted to achieve similar suboptimality.

Data Set	SQRT-Lasso						Lasso
	Prox-GD	Newton	ADMM	ScalReg	CD	Alt.Min	PISTA
Greenhouse	5.812	**1.708**	1027.590	3180.747	14.311	99.814	5.113
DrivFace	**0.421**	0.426	18.879	124.032	3.138	17.691	0.414

is generated by $y = X\theta^* + \epsilon$, where ϵ is sampled from $N(\mathbf{0}, \sigma^2\mathbf{I})$.

We first show the fast convergence of the proximal algorithms at every stage of

Table 4.4: Timing comparison between multiple algorithms for sparse precision matrix estimation on biology data under different levels of sparsity recovery. Each entry is the execution time in seconds. All experiments are conducted to achieve similar suboptimality. Here CD failed to converge and the program aborted before reaching the desired suboptimality. Scalreg failed to terminate in 1 hour for Estrogen.

Sparsity	Arabidopsis					
	Prox-GD	Newton	ADMM	ScalReg	CD	Alt.Min
1%	5.099	**1.264**	292.05	411.7	12.02	183.6
3%	6.201	**2.088**	339.2	426.1	18.18	217.7
5%	7.122	**2.258**	366.7	435.5	28.60	256.9
Sparsity	Estrogen					
1%	108.24	**3.099**	1597	>3600	136.2	634.2
3%	130.93	**7.101**	1846	>3600	332.0	662.2
5%	143.54	**10.12**	2029	>3600	588.4	739.5
Sparsity	Lymph					
1%	3.709	**0.625**	256.4	354.9	7.208	120.2
3%	4.819	**0.905**	289.1	355.3	10.51	130.6
5%	4.891	**1.123**	310.2	358.7	14.95	148.9
Sparsity	Leukemia					
1%	8.542	**2.715**	331.3	610.2	173.3	239.2
3%	10.562	**3.935**	384.7	766.1	174.3	285.1
5%	10.768	**4.712**	442.5	1274	288.9	333.6

Table 4.5: Timing comparison between multiple algorithms for calibrated multivariate regression on synthetic and real data with different values of λ_N. Each entry is the execution time in seconds. All experiments are conducted to achieve similar suboptimality. Here CD failed to converge and the program aborted before reaching the desired suboptimality.

λ_N	Synthetic ($\sigma = 1$)				DrivFace			
	Prox-GD	Newton	ADMM	CD	Prox-GD	Newton	ADMM	CD
$\sqrt{\log d/n}$	0.2964	**0.0320**	14.83	2.409	9.562	**0.2186**	158.9	12.77
$2\sqrt{\log d/n}$	0.1725	**0.0213**	2.230	2.227	8.688	**0.1603**	129.4	20.42
$4\sqrt{\log d/n}$	0.0478	**0.0112**	1.868	1.366	1.824	**0.0924**	94.37	19.17

the pathwise optimization scheme. Here we set $\sigma = 0.5$, $N = 200$, $\lambda_N = \sqrt{\log d/n}$, $\varepsilon_K = 10^{-6}$ for all $K = 1, \ldots, N$. Figure 4.3 presents the objective gap versus the number of iterations. We can see that the proximal algorithms achieves linear (prox-GD) and quadratic (prox-Newton) convergence at every stage. Since the solution sparsity levels are different at each stage, the slopes of these curves are also different.

Next, we show that the computational performance of the pathwise optimization scheme under different settings. Table 4.2 presents the timing performance of Prox-GD combined with the pathwise optimization scheme. We can see that $N = 10$ actually leads to better timing performance than $N = 1$. We can also see that the timing performance of Prox-GD is not sensitive to σ. Moreover, we see that the minimal residual sum of squares along the solution path is much larger than 0, thus the overfitting is prevented and the Prox-GD algorithm enjoys the smoothness of the ℓ_2 loss.

Real Data: We adopt two data sets. The first one is the Greenhouse Gas Observing Network Data Set [137], which contains 2921 samples and 5232 variables. The second one is the DrivFace data set, which contains 606 samples and 6400 variables. We compare our proximal algorithms with ADMM in [118], Coordinate Descent (CD) in [122], Prox-GD (solving Lasso) in [130] and Alternating Minimization (Alt.Min.) [119] and ScalReg (a simple variant of Alt. Min) in [138]. Table 4.3 presents the timing performance of the different algorithms. We can see that Prox-GD for solving SQRT-Lasso significantly outperforms the competitors, and is almost as efficient as Prox-GD for solving Lasso. Prox-Newton is even more efficient than Prox-GD.

Sparse Precision Matrix Estimation. We compare the proximal algorithms with ADMM and CD over real data sets for precision matrix estimation. Particularly, we use four real world biology data sets preprocessed by [139]: Arabidopsis ($d = 834$), Lymph ($d = 587$), Estrogen ($d = 692$), Leukemia ($d = 1,225$). We set three different values for λ_N such that the obtained estimators achieve different levels of sparse recovery. We set $N = 10$, and $\varepsilon_K = 10^{-4}$ for all K's. The timing performance is summarized in Table 4.4. Prox-GD for solving SQRT-Lasso significantly outperforms the competitors, and is almost as efficient as Prox-GD for solving Lasso. Prox-Newton is even more efficient than Prox-GD.

Calibrated Multivariate Regression. We compare the proximal algorithms with ADMM and CD for CMR on both synthetic data and DrivFace data. For synthetic data, the data generating scheme is the same as [116]. Table 4.5 presents the timing performance. Prox-GD for solving SQRT-Lasso significantly outperforms the competitors, and is almost as efficient as Prox-GD for solving Lasso. Prox-Newton is even more efficient than Prox-GD. CD failed to converge and the program aborted before reaching the desired suboptimality.

4.7 Discussion and Conclusions

We show that although the loss function in the SQRT-Lasso optimization problem is nonsmooth, we can directly apply the proximal gradient and Newton algorithms. When further combined with the pathwise optimization scheme, these algorithms enjoy strong guarantees. Our results corroborate that exploiting modeling structures of machine learning problems is of great importance from both computational and statistical perspectives.

Moreover, we remark a gap in our computational theory. In Section 4.3, we prove the restricted strong convexity, smoothness, and Hessian smoothness hold over a neighborhood of θ^*. However, to rigorously establish global fast convergence, we actually need these conditions to hold along the solution path. We highly suspect that this gap is only an artifact of our proof technique, because our empirical results show the proximal algorithms indeed achieve fast convergence along the entire solution path of the pathwise optimization. We will look for more powerful analytic tools and defer a sharper characterization to the future effort.

Chapter 5

On Quadratic Convergence of DC Proximal Newton Algorithm

5.1 Introduction

We consider a high dimensional regression or classification problem: Given n independent observations $\{x_i, y_i\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ sampled from a joint distribution $\mathcal{D}(X, Y)$, we are interested in learning the conditional distribution $\mathbb{P}(Y|X)$ from the data. A popular modeling approach is the Generalized Linear Model (GLM) [140], which assumes

$$\mathbb{P}\left(Y|X;\theta^*\right) \propto \exp\left(\frac{YX^\top\theta^* - \psi(X^\top\theta^*)}{c(\sigma)}\right),$$

where $c(\sigma)$ is a scaling parameter, and ψ is the cumulant function. A natural approach to estimate θ^* is the Maximum Likelihood Estimation (MLE) [141], which essentially minimizes the negative log-likelihood of the data given parameters. However, MLE often performs poorly in parameter estimation in high dimensions due to the curse of dimensionality [142].

To address this issue, machine learning researchers and statisticians follow Occam's razor principle, and propose sparse modeling approaches [115, 120, 143–145]. These sparse modeling approaches assume that θ^* is a sparse vector with only s^* non-zero entries, where $s^* < n \ll d$. This implies that only a few variables in X are essentially relevant to modeling, which is actually very natural to many real world applications, such as genomics and medical imaging [116, 146, 147]. Numerous empirical results have corroborated the success of sparse modeling in high dimensions. Sparse modeling approaches usually obtain a sparse estimator of θ^* by solving the following regularized optimization problem,

$$\overline{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) + \mathcal{R}_{\lambda_{\mathrm{tgt}}}(\theta), \tag{5.1}$$

where $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ is a twice differentiable convex loss function (e.g., negative log-likelihood or pseudo-likelihood), $\mathcal{R}_{\lambda_{\mathrm{tgt}}} : \mathbb{R}^d \to \mathbb{R}$ is a sparsity-inducing decomposable regularizer, i.e., $\mathcal{R}_{\lambda_{\mathrm{tgt}}}(\theta) = \sum_{j=1}^{d} r_{\lambda_{\mathrm{tgt}}}(\theta_j)$ with $r_{\lambda_{\mathrm{tgt}}} : \mathbb{R} \to \mathbb{R}$, and $\lambda_{\mathrm{tgt}} > 0$ is the regularization parameter. Most of the existing sparse modeling approaches can be cast as special examples of (5.1), such as sparse linear regression [120], sparse logistic regression [143], and sparse Poisson regression [144].

For convex regularizers, e.g., $\mathcal{R}_{\mathrm{tgt}}(\theta) = \lambda_{\mathrm{tgt}}\|\theta\|_1$ [120], we can obtain global optima in polynomial time and characterize their statistical properties. However, convex regularizers incur large estimation bias, since they induces too large penalty for the coefficients with large magnitudes. To address this issue, several nonconvex regularizers are proposed, including the minimax concave penalty (MCP, [148]), smooth clipped absolute deviation (SCAD, [149]), and capped ℓ_1-regularization [150]. The obtained estimator (e.g., hypothetical global optima to (5.1)) can achieve faster statistical rates of convergence than their convex counterparts in parameter estimation [134, 151–153].

Related Work: Despite of these superior statistical guarantees, nonconvex regularizers raise greater computational challenge than convex regularizers in high dimensions. Popular iterative algorithms for convex optimization, such as proximal gradient descent [123, 154] and coordinate descent [155–157], no longer have strong global convergence guarantees for nonconvex optimization. Therefore, establishing statistical properties of the estimators obtained by these algorithms becomes very challenging, which explains why existing theoretical studies on computational and statistical guarantees for nonconvex regularized sparse modeling approaches are so limited until recent rise of a new area named "statistical optimization". Specifically, machine learning researchers start to incorporate certain structures of sparse modeling (e.g. restricted strong convexity, large regularization effect) into the algorithmic design and convergence analysis for nonconvex optimization. This further motivates a few recent progresses: [151] propose proximal gradient algorithms for a family of nonconvex regularized estimators with a linear convergence to an approximate local optimum with suboptimal statistical guarantees; [152, 158] further propose homotopy proximal gradient and coordinate gradient descent algorithms with a linear convergence to a local optimum with optimal statistical guarantees; [150, 153] propose a multistage convex relaxation based (also known as Difference of Convex (DC) Programming) proximal gradient algorithm, which can guarantee an approximate local optimum with optimal statistical properties. The computational analysis in [153] further shows that within each stage of the convex relaxation, the proximal gradient algorithm achieves a (local) linear convergence to a unique sparse global optimum for the relaxed convex subproblem.

Motivation: The aforementioned approaches only consider first order algorithms, such as proximal gradient descent and proximal coordinate gradient descent. The second order algorithms with theoretical guarantees are still largely missing for high dimensional nonconvex regularized sparse modeling approaches, but this does not suppress the enthusiasm of applying heuristic second order algorithms to real world problems. Some evidences have already corroborated their superior computational performance over first order algorithms (e.g. `glmnet` [159] and `picasso` [160]). This further motivates our attempt towards understanding the second order algorithms in high dimensions.

Our Contribution: We study a multistage convex relaxation based proximal Newton algorithm for nonconvex regularized sparse learning [161]. This algorithm is not only highly efficient in practice, but also enjoys strong computational and statistical guarantees in theory. Specifically, by leveraging a sophisticated characterization of local restricted strong convexity and Hessian smoothness, we prove that within each stage of convex relaxation, our proposed algorithm maintains the solution sparsity, and achieves a (local) quadratic convergence, which is a significant improvement over the (local) linear convergence of the proximal gradient algorithm in [153] (See more details in later sections). This eventually allows us to obtain an approximate local optimum with optimal statistical properties after only a few number of convex relaxation stages. Numerical experiments are provided to support our theory. To the best of our knowledge, this is the first of second order based approaches for high dimensional sparse learning using convex/nonconvex regularizers with strong statistical and computational guarantees.

Organization: The rest of this paper is as follows: In Section 5.2, we introduce the basic assumptions of the objective function and our algorithm; In Section 5.3, we present both statistical and computational theories that guarantee the convergence of our proposed algorithm; In Section 5.4, we provide numerical experiments to support our theories; In Section 5.5, we detailedly explain why our second order algorithm is superior to the existing first order algorithms in practice, and discuss the extensions of our methodology and theory to proximal sub-sampled Newton and Quasi-Newton algorithms; The proof sketches of our theories are presented in Section 7.4.1; The technical lemmas and supplementary materials are presented in Appendix.

Notations: Given a vector $v \in \mathbb{R}^d$, we denote the set of index for non-zero entries as $\mathrm{supp}(v)$, the number of non-zero entries as $\|v\|_0 = \sum_j \mathbb{1}(v_j \neq 0)$, the p-norm as $\|v\|_p = (\sum_{j=1}^d |v_j|^p)^{1/p}$ for a real $p > 0$, $\|v\|_\infty = \max_j |v_j|$, and the subvector with the j-th entry removed as $v_{\backslash j} = (v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_d)^\top \in \mathbb{R}^{d-1}$. Given an index set $\mathcal{A} \subseteq \{1, ..., d\}$, $\overline{\mathcal{A}} = \{j \mid j \in \{1, ..., d\},\ j \notin \mathcal{A}\}$ is the complementary set to $\mathcal{A}$. We use $v_{\mathcal{A}}$ to denote a subvector of v indexed by $\mathcal{A}$. Given a matrix $A \in \mathbb{R}^{d\times d}$, we use A_{*j} (A_{k*}) to denote the j-th column (k-th row) and $\Lambda_{\max}(A)$ ($\Lambda_{\min}(A)$) as the largest (smallest) eigenvalue of A. We define $\|A\|_{\mathrm{F}}^2 = \sum_j \|A_{*j}\|_2^2$ and $\|A\|_2 = \sqrt{\Lambda_{\max}(A^\top A)}$. We denote $A_{\backslash i \backslash j}$ as the submatrix of A with the i-th row and the j-th column removed, $A_{\backslash ij}$ ($A_{i\backslash j}$) as the j-th column (i-th row) of A with its i-th (j-th) entry removed, and $A_{\mathcal{A}\mathcal{A}}$ as a submatrix of A with both row and column indexed by $\mathcal{A}$. If A is a positive semidefinite matrix, we define $\|v\|_A = \sqrt{v^\top A v}$ as the induced seminorm for vector v. We use conventional notation $\mathcal{O}(\cdot)$, $\Omega(\cdot)$,$\Theta(\cdot)$ to denote the limiting behavior, ignoring constant, and $\mathcal{O}_P(\cdot)$ to denote the limiting behavior in probability. $C_1, C_2, \ldots$ are denoted as generic positive constants.

5.2 DC Proximal Newton Algorithm

Throughout the rest of the paper, we assume: (1) $\mathcal{L}(\theta)$ is nonstrongly convex and twice continuously differentiable, e.g., the negative of log-likelihood function for the generalized linear model (GLM); (2) $\mathcal{L}(\theta)$ takes an additive form, i.e.,

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta),$$

where each $\ell_i(\theta)$ is associated with an observation (x_i, y_i) for $i = 1, ..., n$. Take GLM as an example, we have

$$\ell_i(\theta) = \psi(x_i^\top \theta) - y_i x_i^\top \theta,$$

where ψ is the cumulant function.

For nonconvex regularization, we use the capped ℓ_1 regularizer [150] defined as

$$\mathcal{R}_{\lambda_{\rm tgt}}(\theta) = \sum_{j=1}^{d} r_{\rm tgt}(\theta_j) = \lambda_{\rm tgt} \sum_{j=1}^{d} \min\{|\theta_j|, \beta\lambda_{\rm tgt}\}, \tag{5.2}$$

where $\beta > 0$ is an additional tuning parameter[1]. Our algorithm and theory can also be extended to the SCAD and MCP regularizers in a straightforward manner [148, 149]. As shown in Figure 5.1, $r_{\lambda_{\rm tgt}}(\theta_j)$ can be decomposed as the difference of two convex functions [132],

$$r_\lambda(\theta_j) = \underbrace{\lambda|\theta_j|}_{\text{convex}} - \underbrace{\max\{\lambda|\theta_j| - \beta\lambda^2, 0\}}_{\text{convex}}.$$

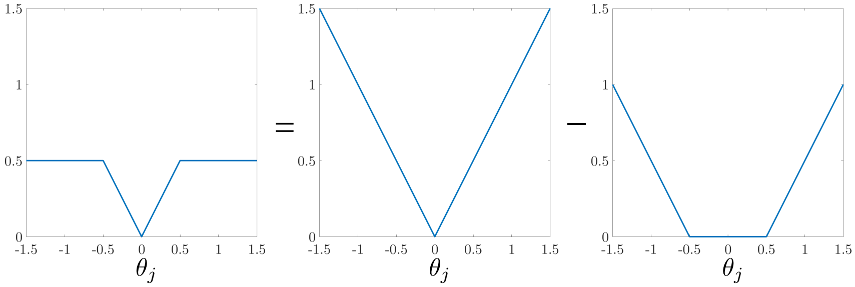


Figure 5.1: The capped ℓ_1 regularizer is the difference of two convex functions. This allows us to relax the nonconvex regularizer based the concave duality.

[1] The capped ℓ_1 regularizer is also independently proposed by [162] with a different name – "Truncated ℓ_1 Regularizer".

This motivates us to apply the difference of convex (DC) programming approach to solve the nonconvex problem. We then introduce the DC proximal Newton algorithm, which contains three components: the multistage convex relaxation, warm initialization, and proximal Newton algorithm.

(I) The multistage convex relaxation is essentially a sequential optimization framework [150][2]. At the $\{K+1\}$-th stage, we have the output solution from the previous stage $\hat{\theta}^{\{K\}}$. For notational simplicity, for all $j = 1, \dots, d$, we define a regularization vector $\lambda^{\{K+1\}} \in \mathbb{R}^d$ as

$$\lambda^{\{K+1\}} = \left(\lambda_1^{\{K+1\}}, ..., \lambda_d^{\{K+1\}}\right)^\top, \quad \text{where} \quad \lambda_j^{\{K+1\}} = \lambda_{\text{tgt}} \cdot \mathbb{1}\left(|\hat{\theta}_j^{\{K\}}| \leq \beta\lambda_{\text{tgt}}\right).$$

Let $\odot$ be the Hadamard (entrywise) product. We solve a convex relaxation of (5.1) at $\theta = \hat{\theta}^{\{K\}}$ as follows,

$$\overline{\theta}^{\{K+1\}} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}}\, \mathcal{F}_{\lambda^{\{K+1\}}}(\theta), \text{ where } \mathcal{F}_{\lambda^{\{K+1\}}}(\theta) = \mathcal{L}(\theta) + \|\lambda^{\{K+1\}} \odot \theta\|_1, \tag{5.3}$$

where $\|\lambda^{\{K+1\}} \odot \theta\|_1 = \sum_{j=1}^d \lambda_j^{\{K+1\}} |\theta_j|$. One can verify that $\|\lambda^{\{K+1\}} \odot \theta\|_1$ is essentially a convex relaxation of $\mathcal{R}_{\lambda_{\text{tgt}}}(\theta)$ at $\theta = \hat{\theta}^{\{K\}}$ based on the concave duality in DC programming.

Remark 4. We emphasize that $\overline{\theta}^{\{K\}}$ denotes the unique sparse global optimum for (5.3) (The uniqueness will be elaborated in later sections), and $\hat{\theta}^{\{K\}}$ denotes the output solution for (5.3) when we terminate the iteration at the K-th convex relaxation stage. The stopping criterion will be explained later.

(II) The warm initialization is the first stage of DC programming, which solves the ℓ_1-regularized counterpart of (5.1),

$$\overline{\theta}^{\{1\}} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}}\, \mathcal{L}(\theta) + \lambda_{\text{tgt}} \|\theta\|_1. \tag{5.4}$$

This is an intuitive choice for sparse statistical recovery, since the ℓ_1-regularized estimator can give us a good initialization, which is sufficiently close to θ^*. Note that (5.4)

[2] The DC programming approach is also independently proposed by [162] as heuristics, and their statistical theory is still based on the hypothetical global optima.

equivalent to (5.3) with $\lambda_j^{\{1\}} = \lambda_{\rm tgt}$ for all $j = 1, \ldots, d$, which can be viewed as the convex relaxation of (5.1) by taking $\hat{\theta}^{\{0\}} = \mathbf{0}$ for the first stage.

(III) The proximal Newton algorithm proposed in [124] is then applied to solve the convex subproblem (5.3) at each stage, including the warm initialization (5.4). For notational simplicity, we omit the stage index $\{K\}$ for all intermediate updates of θ, and only use (t) as the iteration index within the K-th stage for all $K \geq 1$. Specifically, at the K-th stage, given $\theta^{(t)}$ at the t-th iteration of the proximal Newton algorithm, we consider a quadratic approximation of (5.3) at $\theta^{(t)}$ as follows,

$$\mathcal{Q}(\theta; \theta^{(t)}, \lambda^{\{K\}}) = \mathcal{L}(\theta^{(t)}) + (\theta - \theta^{(t)})^\top \nabla \mathcal{L}(\theta^{(t)}) + \frac{1}{2}\|\theta - \theta^{(t)}\|^2_{\nabla^2 \mathcal{L}(\theta^{(t)})} + \|\lambda^{\{K\}} \odot \theta\|_1, \tag{5.5}$$

where $\|\theta - \theta^{(t)}\|^2_{\nabla^2 \mathcal{L}(\theta^{(t)})} = (\theta - \theta^{(t)})^\top \nabla^2 \mathcal{L}(\theta^{(t)})(\theta - \theta^{(t)})$. We then take

$$\theta^{(t+\frac{1}{2})} = \operatorname*{argmin}_{\theta} \ \mathcal{Q}(\theta; \theta^{(t)}, \lambda^{\{K\}}).$$

Since $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_i(\theta)$ takes an additive form, we can avoid directly computing the d by d Hessian matrix in (5.5). Alternatively, in order to reduce the memory usage when d is large, we rewrite (5.5) as a regularized weighted least square problem as follows

$$\mathcal{Q}(\theta; \theta^{(t)}, \lambda^{\{K\}}) = \frac{1}{n}\sum_{i=1}^n w_i(z_i - x_i^\top \theta)^2 + \|\lambda^{\{K\}} \odot \theta\|_1 + \text{constant}, \tag{5.6}$$

where w_i's and z_i's are some easy to compute constants depending on $\theta^{(t)}$, $\ell_i(\theta^{(t)})$'s, x_i's, and y_i's.

Remark 5. Existing literature has shown that the ℓ_1-regularized quadratic problem in (5.6) can be efficiently solved by coordinate descent algorithms in conjunction with the active set strategy [158]. See more details in [159] and Appendix 7.4.2.

For the first stage (i.e., warm initialization), we require an additional backtracking line search procedure to guarantee the descent of the objective value [124]. Specifically, we denote

$$\Delta\theta^{(t)} = \theta^{(t+\frac{1}{2})} - \theta^{(t)}.$$

Then we start from $\eta_t = 1$ and use a backtracking line search procedure to find the optimal $\eta_t \in (0,1]$ such that the Armijo condition [163] holds. Specifically, given a constant $\mu \in (0.9, 1)$, we update $\eta_t = \mu^q$ from $q = 0$ and find the smallest nonnegative integer q such that

$$\mathcal{F}_{\lambda^{\{1\}}}(\theta^{(t)} + \eta_t \Delta\theta^{(t)}) \leq \mathcal{F}_{\lambda^{\{1\}}}(\theta^{(t)}) + \alpha\eta_t\gamma_t,$$

where $\alpha \in (0, \frac{1}{2})$ is a fixed constant and

$$\gamma_t = \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \|\lambda^{\{1\}} \odot \left(\theta^{(t)} + \Delta\theta^{(t)}\right)\|_1 - \|\lambda^{\{1\}} \odot \theta^{(t)}\|_1.$$

We then set $\theta^{(t+1)}$ as $\theta^{(t+1)} = \theta^{(t)} + \eta_t\Delta\theta^{(t)}$ and terminate the iterations for the smallest t when the following approximate KKT condition holds:

$$\omega_{\lambda^{\{1\}}}\left(\theta^{(t)}\right) = \min_{\xi \in \partial\|\theta^{(t)}\|_1} \|\nabla\mathcal{L}(\theta^{(t)}) + \lambda^{\{1\}} \odot \xi\|_\infty \leq \varepsilon,$$

where ε is a predefined precision parameter. Then we set the output solution as $\hat{\theta}^{\{1\}} = \theta^{(t)}$. Note that $\hat{\theta}^{\{1\}}$ is then used as the initial solution for the second stage of convex relaxation (5.3). The proximal Newton algorithm with backtracking line search is summarized in Algorithm 7.

Such a backtracking line search procedure is not necessary at K-th stage for all $K \geq 2$. In other words, we simply take $\eta_t = 1$ and $\theta^{(t+1)} = \theta^{(t)} + \Delta\theta^{(t)} = \theta^{(t+\frac{1}{2})}$ for all $t \geq 0$ when $K \geq 2$. This leads to more efficient updates for the proximal Newton algorithm from the second stage of convex relaxation (5.3). We summarize our proposed DC proximal Newton algorithm in Algorithm 8.

Algorithm 7 Proximal Newton Algorithm (ProxNewton)

```
**Input:** $\theta^{(0)}$, $\lambda_{\mathrm{tgt}}, \varepsilon$
**Initialize:** $t \leftarrow 0$, $\lambda_j^{\{1\}} \leftarrow \lambda_{\mathrm{tgt}}$, $\mu \leftarrow 0.9$, $\alpha \leftarrow \frac{1}{4}$
**Repeat:**
  $\theta^{(t+\frac{1}{2})} \leftarrow \mathrm{argmin}_\theta\, \mathcal{Q}(\theta; \theta^{(t)}, \lambda^{\{1\}})$
  $\Delta\theta^{(t)} \leftarrow \theta^{(t+\frac{1}{2})} - \theta^{(t)}$
  $\gamma_t \leftarrow \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \|\lambda^{\{1\}} \odot \left(\theta^{(t)} + \Delta\theta^{(t)}\right)\|_1 - \|\lambda^{\{1\}} \odot \theta^{(t)}\|_1$
  $\eta_t \leftarrow 1$, $q \leftarrow 0$
  **Repeat:**
    $\eta_t \leftarrow \mu^q$
    $q \leftarrow q + 1$
  **Until** $\mathcal{F}_{\lambda^{\{1\}}}\left(\theta^{(t)} + \eta_t\Delta\theta^{(t)}\right) \leq \mathcal{F}_{\lambda^{\{1\}}}\left(\theta^{(t)}\right) + \alpha\eta_t\gamma_t$
  $\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta_t\Delta\theta^{(t)}$
  $t \leftarrow t + 1$
**Until** $\omega_{\lambda^{\{1\}}}(\theta^{(t)}) \leq \varepsilon$
**Return:** $\theta^{(t)}$.
```

Algorithm 8 DC Proximal Newton Algorithm

```
**Input:** $\hat{\theta}^{\{0\}}$, $\lambda_{\mathrm{tgt}}$, $\beta$, $\varepsilon$
**Warm Initialization:** $\hat{\theta}^{\{1\}} \leftarrow \mathrm{ProxNewton}(\hat{\theta}^{\{0\}}, \lambda_{\mathrm{tgt}}, \varepsilon)$, $K \leftarrow 1$
**Repeat:**
  $\lambda_j^{\{K+1\}} \leftarrow \begin{cases} 0, & \text{if } |\hat{\theta}_j^{\{K\}}| > \beta\lambda_{\mathrm{tgt}} \\ \lambda_{\mathrm{tgt}}, & \text{if } |\hat{\theta}_j^{\{K\}}| \leq \beta\lambda_{\mathrm{tgt}} \end{cases}$
  $t \leftarrow 0$, $\theta^{(0)} = \hat{\theta}^{\{K\}}$
  **Repeat:**
    $\theta^{(t+1)} \leftarrow \mathrm{argmin}_\theta\, \mathcal{Q}(\theta; \theta^{(t)}, \lambda^{\{K+1\}})$
    $t \leftarrow t + 1$
  **Until** $\omega_{\lambda^{\{K+1\}}}(\theta^{(t)}) \leq \varepsilon$
  $\hat{\theta}^{\{K+1\}} \leftarrow \theta^{(t)}$
  $K \leftarrow K + 1$
**Until** Convergence
**Return:** $\hat{\theta}^{\{K\}}$.
```

5.3 Computational and Statistical Theories

Before we present our theoretical analysis, we first introduce some preliminaries, including important definitions and assumptions. We define the largest and smallest s-sparse eigenvalues of the Hessian matrix as follows.

Definition 9. Given any positive integer s, we define the largest and smallest **s-sparse eigenvalues** of $\nabla^2\mathcal{L}(\theta)$ as

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

Moreover, we define $\kappa_s = \rho_s^+/\rho_s^-$ as the s-sparse condition number.

The sparse eigenvalue (SE) properties are widely studied in high dimensional sparse modeling problems, and are closely related to restricted strong convexity/smoothness properties and restricted eigenvalue properties [134, 164–166]. For notational convenience, given a parameter $\theta \in \mathbb{R}^d$ and a real constant $R > 0$, we define a neighborhood of θ with radius R as

$$\mathcal{B}(\theta, R) = \Big\{\phi \in \mathbb{R}^d \mid \|\phi - \theta\|_2 \leq R\Big\}.$$

Our first assumption is for the sparse eigenvalues of the Hessian matrix over a sparse domain.

Assumption 1. Given $\theta \in \mathcal{B}(\theta^*, R)$ for a generic constant R, there exists a generic constant C_0 such that $\nabla^2\mathcal{L}(\theta)$ satisfies the SE properties with parameters $\rho^-_{s^*+2\tilde{s}}$ and $\rho^+_{s^*+2\tilde{s}}$ satisfying

$$0 < \rho^-_{s^*+2\tilde{s}} < \rho^+_{s^*+2\tilde{s}} < +\infty \quad \text{with} \quad \tilde{s} \geq C_0 \kappa^2_{s^*+2\tilde{s}}\, s^* \quad \text{and} \quad \kappa_{s^*+2\tilde{s}} = \rho^+_{s^*+2\tilde{s}}/\rho^-_{s^*+2\tilde{s}}.$$

Assumption 1 requires that $\nabla^2\mathcal{L}(\theta)$ has finite largest and positive smallest sparse eigenvalues, given that θ is sufficiently sparse and close to θ^*. Similar conditions are widely applied in the analyses of efficient algorithms for solving high dimensional learning problems, such as proximal gradient and coordinate gradient descent algorithms [130, 145, 152, 158, 167]. A direct consequence of Assumption 1 is the *restricted strong convexity/smoothness* of $\mathcal{L}(\theta)$ (RSC/RSS, [142]). Given any $\theta, \theta' \in \mathbb{R}^d$, the RSC/RSS parameter can be defined as

$$\delta(\theta', \theta) = \mathcal{L}(\theta') - \mathcal{L}(\theta) - \nabla\mathcal{L}(\theta)^\top(\theta' - \theta).$$

For notational simplicity, we define

$$\mathcal{S} = \{j \mid \theta^*_j \neq 0\} \text{ and } \overline{\mathcal{S}} = \{j \mid \theta^*_j = 0\}.$$

The following proposition connects the SE properties to the RSC/RSS property.

Proposition 2. Given $\theta, \theta' \in \mathcal{B}(\theta^*, R)$ with $\|\theta_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}$ and $\|\theta'_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}$, $\mathcal{L}(\theta)$ satisfies

$$\frac{1}{2}\rho^-_{s^*+2\tilde{s}}\|\theta' - \theta\|_2^2 \leq \delta(\theta', \theta) \leq \frac{1}{2}\rho^+_{s^*+2\tilde{s}}\|\theta' - \theta\|_2^2.$$

The proof of Proposition 2 is provided in [142], and therefore is omitted. Proposition 2 implies that $\mathcal{L}(\theta)$ is essentially strongly convex, but only over a sparse domain (See Figure 5.2).

The second assumption requires $\nabla^2\mathcal{L}(\theta)$ to be smooth over the sparse domain.

Assumption 2 (Local Restricted Hessian Smoothness)**.** Recall that $\tilde{s}$ is defined in Assumption 1. There exist generic constants $L_{s^*+2\tilde{s}}$ and R such that for any $\theta, \theta' \in \mathcal{B}(\theta^*, R)$ with $\|\theta_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}$ and $\|\theta'_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}$, we have

$$\sup_{v \in \Omega,\ \|v\|_2 = 1} v^\top(\nabla^2\mathcal{L}(\theta') - \nabla^2\mathcal{L}(\theta))v \leq L_{s^*+2\tilde{s}}\|\theta - \theta'\|_2^2,$$

where $\Omega = \{v \mid \mathrm{supp}(v) \subseteq (\mathrm{supp}(\theta) \cup \mathrm{supp}(\theta'))\}$.

Assumption 2 guarantees that $\nabla^2\mathcal{L}(\theta)$ is Lipschitz continuous within a neighborhood of θ^* over a sparse domain. The local restricted Hessian smoothness is parallel to the local Hessian smoothness for analyzing the proximal Newton method in low dimensions [124], which is also close related to the self-concordance [131] in the analysis of Newton method [132].

In our analysis, we set the radius R as

$$R = \frac{\rho^-_{s^*+2\tilde{s}}}{2L_{s^*+2\tilde{s}}}. \tag{5.7}$$

Note that $2R = \frac{\rho^-_{s^*+2\tilde{s}}}{L_{s^*+2\tilde{s}}}$ is the radius of the region centered at the unique sparse global minimizer of (5.3) for quadratic convergence of the proximal Newton algorithm, which

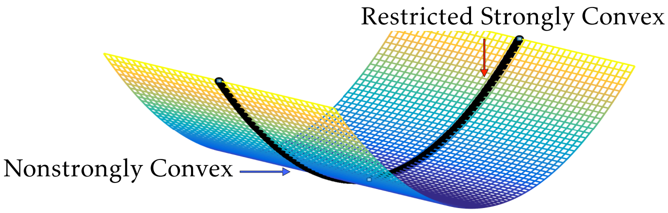


Figure 5.2: An illustrative two dimensional example of the restricted strong convexity. $\mathcal{L}(\theta)$ is not strongly convex. But if we restrict θ to be sparse (Black Curve), $\mathcal{L}(\theta)$ behaves like a strongly convex function.

will be further discussed later. This is parallel to the convergent radius in low dimensions [124], except that we restrict the parameters over the sparse domain.

The third assumption requires λ_{tgt} to be chosen appropriately.

Assumption 3. Given the true modeling parameter θ^*, there exist generic constant C_1 such that

$$\lambda_{\text{tgt}} = C_1\sqrt{\frac{\log d}{n}} \geq 4\|\nabla\mathcal{L}(\theta^*)\|_\infty.$$

Moreover, for large enough n, we have

$$\sqrt{s^*}\lambda_{\text{tgt}} \leq C_2 R \rho^-_{s^*+2\tilde{s}}.$$

Assumption 3 guarantees that the regularization is sufficiently large to eliminate irrelevant coordinates such that the obtained solution is sufficiently sparse [133, 134]. In addition, λ_{tgt} can not be too large, which guarantees that the estimators are close enough to the true model parameter. The above assumptions are deterministic. We will verify these assumptions under GLM in the statistical analysis.

Our last assumption is on the predefined precision parameter ε as follows.

Assumption 4. For each stage of solving the convex relaxed subproblem (5.3) for all $K \geq 1$, we set

$$\varepsilon = \frac{C_3}{\sqrt{n}} \leq \frac{\lambda_{\text{tgt}}}{8} \quad \text{for some generic small constant } C_3.$$

Assumption 4 guarantees that the output solution $\hat{\theta}^{\{K\}}$ at each stage for all $K \geq 1$ has a sufficient precision, which is critical to our convergence analysis of multistage convex relaxation.

5.3.1 Computational Theory

We first characterize the convergence for the first stage of our proposed DC proximal Newton algorithm, i.e., the warm initialization for solving (5.4).

Theorem 12 (Warm Initialization, $K = 1$)**.** Suppose that Assumptions 1 $\sim$ 4 hold with R defined in (5.7). After sufficiently many iterations $T < \infty$, the following results hold for all $t \geq T$:

$$\|\theta^{(t)} - \theta^*\|_2 \leq R \;\text{ and }\; \mathcal{F}_{\lambda^{\{1\}}}(\theta^{(t)}) \leq \mathcal{F}_{\lambda^{\{1\}}}(\theta^*) + \frac{15\lambda_{\mathrm{tgt}}^2 s^*}{4\rho^-_{s^*+2\tilde{s}}},$$

which further guarantee

$$\eta_t = 1,\; \|\theta^{(t)}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}, \;\text{ and }\; \|\theta^{(t+1)} - \overline{\theta}^{\{1\}}\|_2 \leq \frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}}\|\theta^{(t)} - \overline{\theta}^{\{1\}}\|_2^2,$$

where $\overline{\theta}^{\{1\}}$ is the unique sparse global minimizer of (5.4) satisfying $\|\overline{\theta}^{\{1\}}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}$ and $\omega_{\lambda^{\{1\}}}(\overline{\theta}^{\{1\}}) = 0$. Moreover, we need at most

$$T + \log\log\left(3\rho^+_{s^*+2\tilde{s}}/\varepsilon\right)$$

iterations to terminate the proximal Newton algorithm for the warm initialization (5.4), where the output solution $\hat{\theta}^{\{1\}}$ satisfies

$$\|\hat{\theta}^{\{1\}}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s},\; \omega_{\lambda^{\{1\}}}(\hat{\theta}^{\{1\}}) \leq \varepsilon, \;\text{ and }\; \|\hat{\theta}^{\{1\}} - \theta^*\|_2 \leq \frac{18\lambda_{\mathrm{tgt}}\sqrt{s^*}}{\rho^-_{s^*+2\tilde{s}}}.$$

The proof of Theorem 12 is provided in Appendix 7.4.1. Theorem 12 implies:

(1) The objective value is sufficiently small after finite T iterations of the proximal Newton algorithm, which further guarantees solutions to be sparse as well as good computational performance in all follow-up iterations.

(2) The solution enters the ball $\mathcal{B}(\theta^*, R)$ after finite T iterations. Combined with the sparsity of the solution, it further guarantees that the solution enters the region of quadratic convergence. Thus the backtracking line search stops immediately and output $\eta_t = 1$ for all $t \geq T$.

(3) The total number of iterations is at most $\mathcal{O}(T + \log\log(1/\varepsilon))$ to achieve the approximate KKT condition $\omega_{\lambda^{\{1\}}}(\theta^{(t)}) \leq \varepsilon$, which serves as the stopping criterion of the warm initialization (5.4).

Remark 6. To eliminate the notational ambiguity, we emphasis again the difference between $\overline{\theta}^{\{1\}}$ and $\hat{\theta}^{\{1\}}$: $\overline{\theta}^{\{1\}}$ is the *unique sparse global minimizer* of (5.4) that satisfies the KKT condition, i.e., $\omega_{\lambda^{\{1\}}}(\overline{\theta}^{\{1\}}) = 0$; $\hat{\theta}^{\{1\}}$ is the *output solution* of Algorithm 7 that satisfies the approximate KKT condition, i.e., $\omega_{\lambda^{\{1\}}}(\hat{\theta}^{\{1\}}) \leq \varepsilon$ for some predefined $\varepsilon > 0$. Notations $\overline{\theta}^{\{K\}}$ and $\hat{\theta}^{\{K\}}$ with the same interpretations above are also used for later stages $K \geq 2$.

Given these good properties of the output solution $\hat{\theta}^{\{1\}}$ obtained from the warm initialization, we can further show that our proposed DC proximal Newton algorithm for all follow-up stages (i.e., $K \geq 2$) achieves better computational performance than the first stage. This is characterized by the following theorem. For notational simplicity, we omit the iteration index $\{K\}$ for the intermediate updates within each stage for the multistage convex relaxation with $K \geq 2$.

Theorem 13 (Stage K, $K \geq 2$)**.** Suppose Assumptions 1 ∼ 4 hold with R defined in (5.7). Then within each stage $K \geq 2$, for all iterations $t = 1, 2, ...,$ we have

$$\|\theta^{(t)}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s} \quad \text{and} \quad \|\theta^{(t)} - \theta^*\|_2 \leq R,$$

which further guarantee

$$\eta_t = 1,\ \|\theta^{(t+1)} - \overline{\theta}^{\{K\}}\|_2 \leq \frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}}\|\theta^{(t)} - \overline{\theta}^{\{K\}}\|_2^2, \ \text{ and } \ \mathcal{F}_{\lambda^{\{K\}}}(\theta^{(t+1)}) < \mathcal{F}_{\lambda^{\{K\}}}(\theta^{(t)}),$$

where $\overline{\theta}^{\{K\}}$ is the unique sparse global minimizer of (5.3) at the K-th stage satisfying $\|\overline{\theta}^{\{K\}}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}$ and $\omega_{\lambda^{\{K\}}}(\overline{\theta}^{\{K\}}) = 0$. Moreover, we need at most

$$\log\log\left(3\rho^+_{s^*+2\tilde{s}}/\varepsilon\right).$$

iterations to terminate the proximal Newton algorithm for the K-th stage of convex relaxation (5.3), where the output solution $\hat{\theta}^{\{K\}}$ satisfies $\|\hat{\theta}^{\{K\}}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}$, $\omega_{\lambda^{\{K\}}}(\hat{\theta}^{\{K\}}) \leq \varepsilon$, and

$$\|\hat{\theta}^{\{K\}} - \theta^*\|_2 \leq C_2 \left(\|\nabla \mathcal{L}(\theta^*)_{\mathcal{S}}\|_2 + \lambda_{\text{tgt}} \sqrt{\sum_{j \in \mathcal{S}} \mathbb{1}(|\theta^*_j| \leq \beta \lambda_{\text{tgt}})} + \varepsilon\sqrt{s^*} \right) + C_3 0.7^{K-1} \|\hat{\theta}^{\{1\}} - \theta^*\|_2,$$

for some generic constants C_2 and C_3.

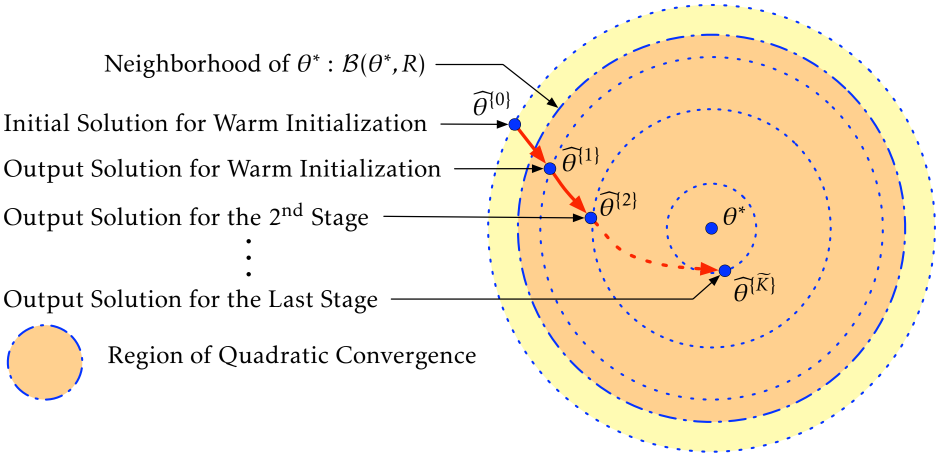

Neighborhood of θ^* : $\mathcal{B}(\theta^*, R)$
$\widehat{\theta}^{\{0\}}$
Initial Solution for Warm Initialization
$\widehat{\theta}^{\{1\}}$
Output Solution for Warm Initialization
$\widehat{\theta}^{\{2\}}$
Output Solution for the 2nd Stage
θ^*
Output Solution for the Last Stage
$\widehat{\theta}^{\{\tilde{K}\}}$
Region of Quadratic Convergence

Figure 5.3: A geometric interpretation of local quadratic convergence: the warm initialization enters the region of quadratic convergence (orange region) after finite iterations and all follow-up stages remain in the region of quadratic convergence. The final estimator $\hat{\theta}^{\{\tilde{K}\}}$ has a better estimation error than the estimator $\hat{\theta}^{\{1\}}$ obtained from the convex warm initialization.

The proof of Theorem 13 is provided in Appendix 7.4.1. A geometric interpretation for the computational theory of local quadratic convergence for our proposed algorithm is provided in Figure 5.3. Within each stage of the convex relaxation (5.3) for all $K \geq 2$, Theorem 13 implies:

(1) The algorithm maintains a sparse solution throughout all iterations $t \geq 1$. The sparsity further guarantees that the SE properties and local restricted Hessian smoothness hold, which are necessary conditions for the fast convergence of the proximal Newton algorithm.

(2) The solution is maintained in the region $\mathcal{B}(\theta^*, R)$ for all $t \geq 1$. Combined with the sparsity of the solution, we have that the solution enters the region of quadratic convergence. This guarantees that we only need to set the step size $\eta_t = 1$ and the objective value is monotone decreasing without the sophisticated backtracking line search procedure. Thus, the proximal Newton algorithm enjoys the same fast convergence as in low dimensional optimization problems [124].

(3) With the quadratic convergence rate, the number of iterations is at most $\mathcal{O}(\log\log(1/\varepsilon))$ to attain the approximate KKT condition $\omega_{\lambda^{\{K\}}}(\theta^{(t)}) \leq \varepsilon$, which is the stopping criteria at each stage.

5.3.2 Statistical Theory

Recall that our computational theory relies on deterministic assumptions (Assumptions 1 $\sim$ 3). However, these assumptions involve data, which are sampled from certain statistical distribution. Therefore, we need to verify that these assumptions hold with high probability under mild data generation process (e.g., GLM) in high dimensions in the following lemma.

Lemma 3 (GLM)**.** Suppose that x_i's are i.i.d. sampled from a zero-mean distribution with covariance matrix $\mathrm{Cov}(x_i) = \Sigma$ such that $\infty > c_{\max} \geq \Lambda_{\max}(\Sigma) \geq \Lambda_{\min}(\Sigma) \geq c_{min} > 0$, and for any $v \in \mathbb{R}^d$, $v^\top x_i$ is sub-Gaussian with parameter at most $a\|v\|_2^2$, where $c_{\max}$, $c_{\min}$, and a are generic constants. Moreover, for some constant $M_\psi > 0$, at least one of the following two conditions holds:

(1) The Hessian of the cumulant function ψ is uniformly bounded: $\|\psi''\|_\infty \leq M_\psi$, or

(2) The covariates are bounded $\|x_i\|_\infty \leq 1$, and

$$\mathbb{E}[\max_{|u|\leq 1}[\psi''(x^\top\theta^*) + u]^p] \leq M_\psi \quad \text{for some} \quad p > 2.$$

Then Assumptions 1 $\sim$ 3 hold with high probability.

The proof of Lemma 3 is provided in Appendix 7.4.5. Given that these assumptions hold with high probability, the computational theory holds, i.e., the proximal Newton algorithm attains quadratic rate convergence within each stage of convex relaxation with high probability. We then further establish the statistical rate of convergence for the obtained estimator in parameter estimation.

Theorem 14. Suppose the observations are generated from GLM satisfying the conditions in Lemma 3 for large enough n such that $n \geq C_4 s^* \log d$ and $\beta = C_5/c_{\min}$ is a constant defined in (5.2) for generic constants C_4 and C_5, then with high probability, the output solution $\hat{\theta}^{\{K\}}$ satisfies

$$\|\hat{\theta}^{\{K\}} - \theta^*\|_2 \leq C_6 \left(\sqrt{\frac{s^*}{n}} + \sqrt{\frac{s' \log d}{n}} \right) + C_7 0.7^K \left(\sqrt{\frac{s^* \log d}{n}} \right)$$

for generic constants C_6 and C_7, where $s' = \sum_{j \in \mathcal{S}} \mathbb{1}(|\theta_j^*| \leq \beta \lambda_{\text{tgt}})$.

Theorem 14 is a direct result combining Theorem 13 and the analyses in [150]. As can be seen, s' is essentially the number of non-zero θ_j's with smaller magnitudes than $\beta\lambda_{\text{tgt}}$, which are often considered as "weak" signals. Theorem 14 essentially implies that by exploiting the multi-stage convex relaxation framework, our DC proximal Newton algorithm gradually reduces the estimation bias for "strong" signals, and eventually obtains an estimator with better statistical properties than the ℓ_1-regularized estimator. Specifically, let $\tilde{K}$ be the smallest integer such that after $\tilde{K}$ stages of convex relaxation we have

$$C_7 0.7^{\tilde{K}} \left(\sqrt{\frac{s^* \log d}{n}} \right) \leq C_6 \max \left\{ \sqrt{\frac{s^*}{n}}, \sqrt{\frac{s' \log d}{n}} \right\},$$

which is equivalent to requiring $\tilde{K} = \mathcal{O}(\log \log d)$. This implies the total number of the proximal Newton updates is at most

$$\mathcal{O}\left(T + \log\log(1/\varepsilon) \cdot (1 + \log\log d)\right).$$

In addition, the obtained estimator attains the optimal statistical properties in parameter estimation:

$$\|\hat{\theta}^{\{\tilde{K}\}} - \theta^*\|_2 \leq \mathcal{O}_P\Big(\sqrt{\frac{s^*}{n}} + \sqrt{\frac{s' \log d}{n}}\Big) \quad \text{v.s.} \quad \|\hat{\theta}^{\{1\}} - \theta^*\|_2 \leq \mathcal{O}_P\Big(\sqrt{\frac{s^* \log d}{n}}\Big). \tag{5.8}$$

Recall that $\hat{\theta}^{\{1\}}$ is obtained by the warm initialization (5.4). As illustrated in Figure 5.3, this implies the statistical rate in (5.8) for $\|\hat{\theta}^{\{\tilde{K}\}} - \theta^*\|_2$ obtained from the multistage convex relaxation for the nonconvex regularized problem (5.1) is a significant improvement over $\|\hat{\theta}^{\{1\}} - \theta^*\|_2$ obtained from the convex problem (5.4). Especially when s' is small, i.e., most of non-zero θ_j's are strong signals, our result approaches the oracle bound[3] $\mathcal{O}_P(\sqrt{s^*/n})$ [149] as illustrated in Figure 5.4.

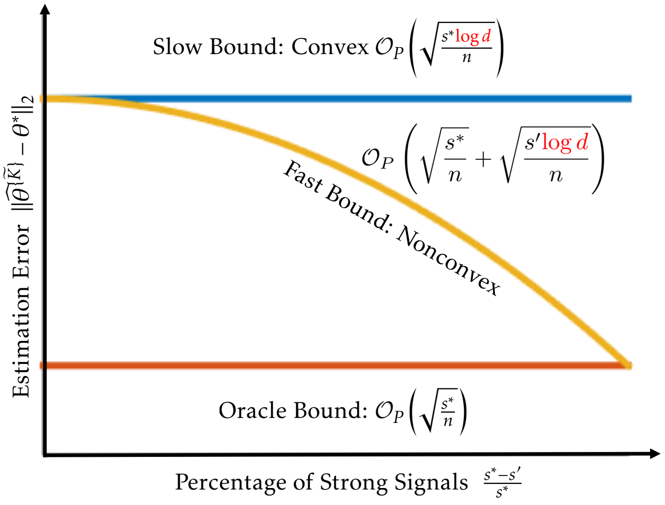

Slow Bound: Convex $\mathcal{O}_P\left(\sqrt{\frac{s^*\log d}{n}}\right)$
$\mathcal{O}_P\left(\sqrt{\frac{s^*}{n}} + \sqrt{\frac{s'\log d}{n}}\right)$
Fast Bound: Nonconvex
Oracle Bound: $\mathcal{O}_P\left(\sqrt{\frac{s^*}{n}}\right)$
Estimation Error $\|\hat{\theta}^{\{\tilde{K}\}} - \theta^*\|_2$
Percentage of Strong Signals $\frac{s^*-s'}{s^*}$

Figure 5.4: An illustration of the statistical rates of convergence in parameter estimation. Our obtained estimator has an error bound between the oracle bound and the slow bound from the convex problem in general. When the percentage of strong signals increases, i.e., s' decreases, then our result approaches the oracle bound.

5.4 Experiments

We compare our DC Proximal Newton algorithm (DC+PN) with two competing algorithms for solving nonconvex regularized sparse logistic regression problems.They are accelerated proximal gradient algorithm (APG) implemented in the SPArse Modeling Software (`SPAMS`, coded in `C++`, [168]), and accelerated coordinate descent (ACD) algorithm implemented in `R` package `gcdnet` (coded in Fortran, [169]). We further optimize the active set strategy in `gcdnet` to boost its computational performance. To integrate

[3] The oracle bound assumes that we know which variables are relevant in advance. It is not a realistic bound, but only for comparison purpose.

these two algorithms with the multistage convex relaxation framework, we revise their source code.

To further boost the computational efficiency at each stage of the convex relaxation, , we apply the pathwise optimization for all algorithms [158, 159]. Specifically, we use a geometrically decreasing sequence of regularization parameters $\{\lambda_{[m]} = \alpha^m \lambda_{[0]}\}_{m=1}^{M}$, where $\alpha \in (0,1)$ is a shrinkage ratio, and $\lambda_{\rm tgt} = \lambda_{[M]}$. For each $\lambda_{[m]}$, we apply the corresponding algorithm (DC+PN, DC+APG, and DC+ACD) to solve the nonconvex regularized problem (5.1). The value of $\lambda_{[0]}$ is chosen to be the smallest value such that the corresponding solution is zero. Moreover, we initialize the solution for a new regularization parameter $\lambda_{[m+1]}$ using the output solution obtained with $\lambda_{[m]}$. Such a pathwise optimization scheme has achieved tremendous success in practice [159,170,171], and we refer [158] for more involved theoretical analysis.

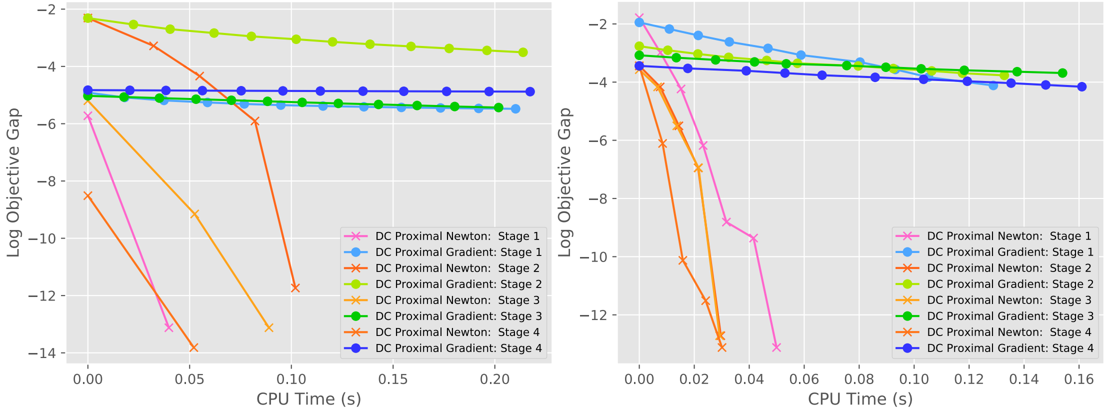
All three algorithms are compared in wall clock time and objective values with $\lambda_{\rm tgt} \approx \frac{1}{4}\sqrt{\log d/n}$. Our DC Proximal Newton algorithm is implemented in `C` with double precisions, and called from `R` by a wrapper. Our comparison contains 3 datasets: "madelon" ($n = 2000, d = 500$, [172]), "gisette" ($n = 2000$,$d = 5000$, [172]), and two simulated datasets: "sim_1k" and "sim_10k". For the simulated data sets, we choose $n = 1000$ and $d = 5000$, and generate each x_i independently from a d-dimensional normal distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma_{jk} = 0.5^{|j-k|}$ for all $j, k = 1, ..., d$. We generate $y \sim \text{Bernoulli}(1/[1+\exp(-x_i^\top \theta^*)])$, where θ^* has all 0 entries except randomly selected 20 entries. The non-zero entries are independently sampled from $\text{U}(0,1)$.

Table 5.1: Quantitive timing comparisons for on nonconvex-regularized sparse logistic regression. DC+PN denotes our proposed DC proximal Newton algorithm; ACD denotes the coordinate descent algorithm combined with the active set strategy; APG denotes the accelerated proximal gradient algorithm. The average values and standard errors (in parentheses) of timing performance (in seconds) are presented.

	madelon	gisette	sim_1k	sim_10k
DC+PN	**1.51**(±0.01)s obj value: 0.52	**5.35**(±0.11)s obj value: 0.01	**1.07**(±0.02)s obj value: 0.01	**8.82**(±0.04)s obj value: 0.01
DC+ACD	**5.83**(±0.03)s obj value: 0.52	**18.92**(±2.25)s obj value: 0.01	**9.46**(±0.09) s obj value: 0.01	**19.1**(±0.56) s obj value: 0.01
DC+APG	**1.60**(±0.03)s obj value: 0.52	**207**(±2.25)s obj value: 0.01	**17.8**(±1.23) s obj value: 0.01	**222**(±5.79) s obj value: 0.01

The experiments are performed on a personal computer with 2.6GHz Intel Core i7 and 16GB RAM. For each algorithm and dataset, we repeat the algorithm 10 times and we report the average values and standard errors of the wall clock time in Table 5.1. The stopping criteria for each algorithms are tuned such that they attains similar optimization errors. As can be seen in Table 5.1, our DC Proximal Newton algorithm significantly outperforms the competing algorithms in terms of the timing performance.

We then illustrate the quadratic convergence of our DC proximal Newton algorithm within each stage of convex relaxation using the "sim" datasets. Specifically, we provide the plots of gap towards the optimal objective of the K-th stage, i.e., $\log(\mathcal{F}_{\lambda^{\{K\}}}(\theta^{(t)}) - \mathcal{F}_{\lambda^{\{K\}}}(\overline{\theta}^{\{K\}}))$, for $K = 1, 2, 3, 4$ in a single simulation in Figure 5.5. We see that our DC proximal Newton algorithm achieves quadratic convergence, which is consistent with our theory.



(a) Simulated Data, $\lambda = 0.036$

(b) Gissete Data, $\lambda = 0.02$

Figure 5.5: Timing comparisons in wall clock time. Our DC proximal Newton algorithm demonstrates superior quadratic convergence (consistent with our theory), and significantly outperforms the DC proximal gradient algorithm.

5.5 Discussions and Future Work

We first provide detailed discussions on the superior performance of our DC proximal Newton in our experiment, and then discuss potential variants – DC proximal subsampled Newton or Quasi-Newton algorithm.

5.5.1 Drawbacks of first order algorithms

There exist two major drawbacks of existing multi-stage convex relaxation based first order algorithms:

(1) The first order algorithms have significant computational overhead in each iteration, e.g., for GLM, computing gradients requires frequently evaluating the cumulant function and its derivatives. This often involves extensive non-arithmetic operations such as `log` and `exp` functions, which naturally appear in the cumulant function and its derivates and are computationally expensive. To the best of our knowledge, even if we use some efficient numerical methods for calculating `exp` in [173, 174], the computation still needs at least $10-30$ times more CPU cycles than basic arithmetic operations, e.g., multiplications. Our proposed DC Proximal Newton algorithm cannot avoid calculating the cumulant function and its derivatives, when computing quadratic approximations. The computation, however, is much less intense, since the convergence is quadratic.

(2) The first order algorithms are computationally expensive with the step size selection. Although for certain GLM, e.g., sparse logistic regression, we can choose the step size parameter as

$$\eta \approx \Lambda_{\max}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right).$$

However, such a step size often leads to very poor performance. In contrast, as our theoretical analysis and experiments suggest, the proposed DC proximal Newton algorithm needs very few line search steps, which saves much computational efforts.

Some recent papers on proximal Newton or inexact proximal Newton also demonstrate local quadratic convergence guarantees, such as [175, 176]. However, their conditions are much more stringent than the SE properties in terms of the dependence on the problem dimensions. Specifically, their quadratic convergence can only be guaranteed on a much smaller ball/neighborhood. For example, the constant nullspace strong convexity in [175], which plays the same role as the smallest sparse eigenvalue $\rho_{s^*+2\tilde{s}}^-$ in our analysis, is as small as $1/d$. Thus, they can only guarantee the quadratic convergence in a region with radius $\mathcal{O}(1/d)$, which is very small in high dimensions. However, in our analysis, $\rho^-_{s^*+2\tilde{s}}$ can be a constant, which is (almost) independent of d [142] and much larger than $\mathcal{O}(1/d)$. A similar issue that the quadratic region is too small exists in [176] as well.

5.5.2 Extension to sub-sampled or Quasi-Newton algorithms

Our methodology can be further extended to proximal sub-sampled Newton or Quasi-Newton algorithms using either BFGS-type or subsampled Hessian matrices. Taking the Proximal sub-sampled Newton algorithm as an example, we replace the Hessian matrix with an approximate Hessian matrix in each proximal Newton iteration. Suppose that at the t-th iteration of the K-th stage, we randomly select a mini-batch $\mathcal{X}^{(t)} \subset \{1, ..., n\}$ of m samples from the data with equal probability (i.e., $|\mathcal{X}^{(t)}| = m$). We then consider an alternative quadratic approximation

$$\begin{aligned}&\hat{\mathcal{Q}}(\theta; \theta^{(t)}, \lambda^{(K)}, \mathcal{X}^{(t)})\\&= \mathcal{L}(\theta^{(t)}) + (\theta - \theta^{(t)})^\top \nabla \mathcal{L}(\theta^{(t)}) + \frac{1}{2}\|\theta - \theta^{(t)}\|^2_{\hat{H}(\theta^{(t)}, \mathcal{X}^{(t)})} + \|\lambda^{\{K\}} \odot \theta\|_1,\end{aligned} \tag{5.9}$$

where $\hat{H}(\theta^{(t)}, \mathcal{X}^{(t)})$ is the subsampled Hessian matrix

$$\hat{H}(\theta^{(t)}, \mathcal{X}^{(t)}) = \frac{1}{m} \sum_{i \in \mathcal{X}^{(t)}} \nabla^2 \ell_i(\theta^{(t)}).$$

By exploiting the additive nature of $\mathcal{L}(\theta)$, we can further rewrite (5.9) as

$$\hat{\mathcal{Q}}(\theta; \theta^{(t)}, \lambda^{(K)}, \mathcal{X}^{(t)}) = \frac{1}{m} \sum_{i \in \mathcal{X}^{(t)}} w_i (x_i^\top \theta)^2 + g^\top \theta + \|\lambda^{\{K\}} \odot \theta\|_1 + \text{constant}, \tag{5.10}$$

where $g \in \mathbb{R}^d$ and $w_i \in \mathbb{R}$ for all $i \in \mathcal{X}^{(t)}$ are some easy to compute constants depending on $\theta^{(t)}$, $\ell_i(\theta^{(t)})$'s, x_i's, and y_i's. Similar to (5.6), (5.10) only requires $O(md)$ memory usage and can be efficiently solved by coordinate descent algorithms in conjunction with the active set strategy, soft thresholding, and residual update. See more details in Appendix 7.4.2. Note that the line search procedure is needed for the proximal sub-sampled Newton algorithm throughout all iterations and stages.

The sub-sampled Hessian matrices preserve the spectral behaviors when the batch size m is large enough (e.g. $m = \Omega(s^* \log d)$). Thus, restricted strong convexity, smoothness, and Hessian smoothness hold, and similar theoretical results are expected to hold. A major difference is that we get slower convergence (e.g. superlinear or linear depending on the batch size m) instead of quadratic convergence. This is a fundamental trade-off between Proximal Newton and proximal sub-sampled Newton (or Quasi-Newton) algorithm in both low and high dimensions. We will leave this for further investigation.

Chapter 6

Global Optimization Landscape of Nonconvex Matrix Factorization

6.1 Introduction

We consider a low-rank matrix estimation problem. Specifically, we want to estimate $M^* \in \mathbb{R}^{n\times m}$ with $\mathrm{rank}(M^*) = r \ll \min\{n, m\}$ by solving the following rank-constrained problem

$$\min_{M} f(M) \quad \text{subject to } \mathrm{rank}(M) \leq r, \tag{6.1}$$

where $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ is usually a convex and smooth loss function. Since solving (6.1) has been known to be NP-hard in general, significant efforts have been also devoted to studying a convex relaxation of (6.1) as follows,

$$\min_{M} f(M) \quad \text{subject to } \|M\|_* \leq \tau, \tag{6.2}$$

where τ is a tuning parameter and $\|M\|_*$ is the sum of all singular values of M, also known as the nuclear norm [177–180].

Although there have been a number of algorithms proposed for solving either (6.1) or (6.2) in existing literature [181–183], all these algorithms are iterative, and each iteration needs to calculate a computationally expensive Singular Value Decomposition (SVD), or an equivalent operation for finding the dominant singular values/vectors. This is very prohibitive for large-scale problems. In practice, most of popular heuristic algorithms resort to factorizing M to a product of smaller matrices, i.e, $M^* = UV^\top$, where $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{m\times r}$, also known as the factorized form. Then instead of solving (6.1) or (6.2), we solve the following nonconvex problem

$$\min_{X\in\mathbb{R}^{n\times r}, Y\in\mathbb{R}^{m\times r}} f(XY^\top), \tag{6.3}$$

where scalable algorithms can iteratively update X and Y very efficiently. The reparametrization of the low rank matrix in (6.3) is closely related to the Burer-Monteiro factorization for semidefinite programing in existing literature. See more details in [184, 185].

Tremendous progress has been made to provide theoretical justifications of the popular nonconvex factorization heuristic algorithms for general classes of functions [186–190]. A wide family of problems can be cast as (6.3). Popular examples include matrix sensing [186, 188, 191–194], matrix completion [195–200], (sparse) principle component analysis (PCA) [201–204], and factorization machine [205, 206]. Recent efforts are also made when the observation is a superposition of low-rank and sparse matrices [207, 208]. Moreover, extensions to low-rank tensor estimation and its related problems, such as independent component analysis (ICA) and topic modeling, are also studied [199, 209–211].

The factorized form $M = XY^\top$ makes (6.3) very challenging to solve. First, it yields infinitely many nonisolated saddle points because of the existence of invariant rotation group. For example, if some (X, Y) pair is a saddle point, then for any orthogonal matrix $\Phi \in \mathbb{R}^{r\times r}$, i.e., $\Phi\Phi^\top = I$, $(X\Phi, Y\Phi)$ is also a saddle point since $XY^\top = X\Phi(Y\Phi)^\top$. For the same reason, there exist infinitely many local/global minima as well for $r > 1$. Second, although $f(M)$ is convex on M, $f(XY^\top)$ is not jointly convex in X and Y (even around a small neighborhood of a global optimum). To address these challenges, various techniques are developed recently. Extensive contemporary works focus on the local convergence rate analysis based on local geometric properties of the optimization problem using generalization of convexity/smoothness of f, such as local regularity condition [192, 198, 212, 213] and local descent condition [188, 207]. However, careful initialization is required in this type of approaches. Another line of works on solving the factorized problem (6.3) focus on the optimality conditions that guarantees global convergence using random initialization [193, 199]. However, since only partial results on the landscape of optimization are discussed, e.g., only stationary points (i.e., saddle points and local minima) are characterized without discussing their neighborhood or the rest region of the parameter space, no explicit global convergence rate can be guaranteed.

In addition to the approaches discussed above, another more clear yet more challenging scheme is to characterize the global landscape of the nonconvex optimization problem, based on which the global convergence analysis becomes possible. Without further distinction, we use "landscape" to denote the the geometry of the objective function in the optimization problem, i.e., *the characterization of all stationary points* and *the explicit geometry of the objective function on the entire parameter domain* (e.g., the characterization of regions $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$ defined below). Nevertheless, there are few works that discuss the global landscape of the nonconvex optimization (6.3) in such an explicit manner. One of the earliest works that study the global landscape of nonconvex optimization in this sense is on the phase retrieval problem [214], which can be viewed as a special case of (6.3). Such global landscape on optimization can further help provide global convergence rate analysis using popular iterative algorithms without careful initialization [187, 214–216]. However, existing works have not discussed the intrinsic reasons of difficulties that present in the nonconvex matrix factorization problems, e.g., the generation of saddle points.

To shed light on the nonconvex matrix factorization problems (6.3), our study here consists of two major parts to answer two questions of our interest [217]: (I) Why are there saddle points and how to identify them effectively? (II) How do the saddle points impact the geometry of the optimization problem? To answer the first question, we study a generic theory for characterizing the landscape of a general class of functions with underlying symmetric structures. Based on a new symmetry principle, we identify stationary points for those functions with invariant groups, which characterizes the underlying principle of generating saddle points in nonconvex matrix factorization problems. Moreover, we characterize the null space of the Hessian matrices of the stationary points via the tangent space. We further provide concrete examples to demonstrate our proposed theory. To the best of our knowledge, this is the first effort to provide a generic framework for characterizing geometric properties of a large class of functions with symmetric structure.

To answer the second question, we establish a comprehensive analysis for global landscape of the low-rank matrix factorization problem based on our proposed generic theory. Specifically, we consider a symmetric positive semidefinite (PSD) matrix $M^* = UU^\top \succeq 0$, and solve the following problem

$$\min_{X \in \mathbb{R}^{n \times r}} \mathcal{F}(X), \text{ where } \mathcal{F}(X) = \frac{1}{4}\|M^* - XX^\top\|_\mathrm{F}^2. \tag{6.4}$$

Here we only consider the PSD matrix for simplicity, and the extension to the general rectangular case is straightforward (see more details in Section 6.2). Though (6.4) has been viewed as an important foundation of many popular matrix factorization problems such as matrix sensing and matrix completion, the global landscape of $\mathcal{F}(X)$ in (6.4) is not very clear yet. Based on our generic theory, we explicitly identify all saddle points and global minima of $\mathcal{F}(X)$. Further, we show that the entire parameter space can be described as one the three regions as follows.

($\mathcal{R}_1$) The region that contains neighborhoods of all saddle points, where any associated Hessian matrix of the objective has negative eigenvalues. This so-called strict saddle property guarantees that many commonly used iterative algorithms cannot not be trapped in those saddle points.

($\mathcal{R}_2$) The region that contains neighborhoods of all global minima, where the objective is only strongly convex along certain trajectories, otherwise is nonconvex, unless $r = 1$. We specify these directions explicitly, along which $\mathcal{F}(X)$ is strongly convex.

($\mathcal{R}_3$) The complement of regions $\mathcal{R}_1$ and $\mathcal{R}_2$ in $\mathbb{R}^{n \times r}$, where the gradient has a sufficiently large norm. Together with $\mathcal{R}_1$ and $\mathcal{R}_2$, a convergence of (6.4) to a global minimum is guaranteed for many commonly used iterative algorithms without special initializations.

Moreover, we further connect our analysis on (6.4) to the matrix sensing problem, which can be considered as a perturbed version of (6.4). Using a suboptimal sample complexity, we establish analogous global geometric properties to (6.4) for the matrix sensing problem. These strong geometric properties imply the convergence to a global minimum of the matrix factorization problem in polynomial time without careful initialization for several popular iterative algorithms, such as the gradient descent algorithm, the noisy stochastic gradient descent algorithm, and the trust-region Newton's algorithm.

After the initial release of our paper, several concurrent and follow-up works have appeared. In specific, [218] extend our analysis to the general rectangular matrices using the lifting formulation and achieve analogous results to ours. Another related work is [219], which provide a unified geometric analysis based on the strict saddle property for several popular nonconvex problems, including matrix sensing, matrix completion, and robust PCA. By partially applying the result in [219], we further demonstrate a sharper result for matrix sensing in terms of the sample complexity, with some sacrifice in the properties of the optimization landscape as a tradeoff. Further discussions will be provided in Section 6.3.3 and 6.5.1.

The rest of the paper is organized as follows. In Section 6.2, we provide a generic theory of identifying stationary points and the null space of their Hessian matrices, along with several concrete examples. In Section 6.3, a global geometric analysis is established for the low-rank matrix factorization problem. In Section 6.4, we extend the analysis to the matrix sensing problem, followed by a further discussion in Section 6.5. All proofs are deferred to Appendix.

Notation. Given an integer $n \geq 1$, we denote $[n] = \{1, \ldots n\}$. Let $\mathfrak{O}_r = \{\Psi \in \mathbb{R}^{r \times r} \ : \ \Psi\Psi^\top = \Psi^\top\Psi = I_r\}$ be the set of all orthogonal matrices in $\mathbb{R}^{r\times r}$. Given a matrix $A \in \mathbb{R}^{n\times m}$ and a subspace $\mathcal{L} \in \mathbb{R}^n$, let $\mathcal{P}_{\mathcal{L}}(A)$ be the orthogonal projection operation of A onto $\mathcal{L}$, and $\mathcal{L}^\perp$ be the complement of $\mathcal{L}$ in $\mathbb{R}^n$. Denote $\mathcal{L}_A$ as the column space of A. We use $A_{(*,k)}$ and $A_{(j,*)}$ to denote the k-th column and the j-th row respectively, $A_{(j,k)}$ to denote the (j,k)-th entry, and $A_{\mathcal{S}}$ to denote a column-wise sub matrix of A indexed by a set $\mathcal{S} \subseteq [m]$. Let $\sigma_i(A)$ be the i-th largest singular value, $\|A\|_2$ be the spectral norm (largest singular value), and $\|A\|_{\mathrm{F}}$ be the Frobenius norm. Given two matrices $A, B \in \mathbb{R}^{n\times m}$, denote $\langle A, B\rangle = \mathrm{Tr}(A^\top B) = \sum_{i,j} A_{(i,j)}B_{(i,j)}$. When $A \in \mathbb{R}^{n\times n}$ is a square matrix, we denote $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ as the largest and smallest eigenvalues respectively. Given a vector $a \in \mathbb{R}^n$, let $a_{(i)}$ be the i-th entry. We use a subscript A_i (a_i) to denote the i-th matrix (vector) in a sequence of matrices (vectors). Denote $\mathbb{E}(X)$ as the expectation of a random variable X and $\mathbb{P}(\mathcal{X})$ as the probability of an event $\mathcal{X}$. We use $\otimes$ as the kronecker product, and preserve $C_1, C_2, \ldots$ and $c_1, c_2, \ldots$ for positive real constants.

6.2 A Generic Theory for Stationary Points

Given a function f, our goal is to find the stationary point. Rigorous mathematical definitions are provided as follows.

Definition 10. Given a smooth function $f : \mathbb{R}^n \to \mathbb{R}$, a point $x \in \mathbb{R}^n$ is called:

(i) a ***stationary point***, if $\nabla f(x) = 0$;

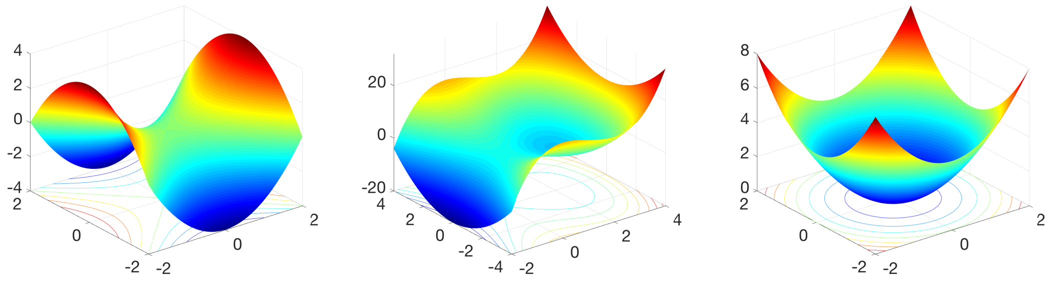
(ii) a ***local minimum (or maximum)***, if x is a stationary point and there exists a neighborhood $\mathcal{B} \subseteq \mathbb{R}^n$ of x such that $f(x) \leq f(y)$ (or $f(x) \geq f(y)$) for any $y \in \mathcal{B}$;

(iii) a ***global minimum (or maximum)***, if x is a stationary point and $f(x) \leq f(y)$ (or $f(x) \geq f(y)$) for any $y \in \mathbb{R}^n$;

(iv) a ***strict saddle point***, if x is a stationary point and for any neighborhood $\mathcal{B} \subseteq \mathbb{R}^n$ of x, there exist $y, z \in \mathcal{B}$ such that $f(z) < f(x) < f(y)$ and $\lambda_{\min}(\nabla^2 f(x)) < 0$.

A visualization of different types of stationary points are provided in Figure 6.1. In general, finding the stationary point requires solving a large system $\nabla f(x) = 0$, which can be computationally challenging. However, when f has special structures, we can develop new principles to find the set of stationary points conveniently.

In this paper, we consider a class of functions with invariant groups, for which we provide a generic theory to determine the stationary point using the symmetry principle. This covers the low-rank matrix factorization problem as a special example. Moreover, we can characterize the null space of the Hessian matrix at the stationary point by leveraging the tangent space. This will further help us to determine the saddle point and local/global minimum (see more details in Section 6.3).



(a) strict saddle (b) local minimum (c) global minimum

Figure 6.1: Examples of a strict saddle point, a local minimum, and a global minimum.

6.2.1 Determine Stationary Points

For self-containedness, we start with a few definitions in group theory [220] as follows.

Definition 11. A ***group*** $\mathcal{G}$ is a set of elements together with a binary operation $\{\cdot\}$ that satisfies the following four properties:

- Closure: for all $a_1, a_2 \in \mathcal{G}$, we have $a_1 \cdot a_2 \in \mathcal{G}$;

- Associativity: for all $a_1, a_2, a_3 \in \mathcal{G}$, we have $(a_1 \cdot a_2) \cdot a_3 = a_1 \cdot (a_2 \cdot a_3)$;

- Identity: there exists an identity element $e \in \mathcal{G}$ such that $e \cdot a = a$ and $a \cdot e = a$ for all $a \in \mathcal{G}$;

- Inverse: for any $a \in \mathcal{G}$, there exists an inverse element $a^{-1} \in \mathcal{G}$ such that $a \cdot a^{-1} = e$ and $a^{-1} \cdot a = e$.

Definition 12. A ***commutative group*** is a group that also satisfies

- Commutativity: for all $a_1, a_2 \in \mathcal{G}$, we have $a_1 \cdot a_2 = a_2 \cdot a_1$.

Definition 13. A ***field*** is a set with two binary operations $\{+, \cdot\}$, addition (denoted $\{+\}$ and multiplication $\{+, \cdot\}$, both of which satisfy associativity, identity ($\{+\}$ is associated with identity 0 and $\{\cdot\}$ is associated with identity 1), inverse, commutativity, and

- Distributivity: for all $a_1, a_2, a_3 \in \mathcal{G}$, we have $a_1 \cdot (a_2 + a_3) = (a_1 \cdot a_2) + (a_1 \cdot a_3)$.

Definition 14. A subset $\mathcal{H}$ of a group $\mathcal{G}$ is a ***subgroup*** if $\mathcal{H}$ is itself a group under the operation induced by $\mathcal{G}$.

Definition 15. The set of all invertible $n \times n$ real matrices with determinant 1, together with the operations of ordinary matrix multiplication and matrix inversion, is a ***special linear group*** of degree n over a field, denoted as $\mathrm{SL}_n(\mathbb{R})$.

Definition 16. Given a function $f : \mathbb{R}^m \to \mathbb{R}$, a subgroup $\mathcal{G}$ of a special linear group $\mathrm{SL}_m(\mathbb{R})$ is an ***invariant group*** if $\mathcal{G}$ satisfies $f(x) = f(g(x))$ for all $x \in \mathbb{R}^m$ and $g \in \mathcal{G}$.

Remark 7. We define the invariant group in terms of the special linear group rather than the general linear group because we want to preserve the volume for linear transformations.

Definition 17. A point $x_{\mathcal{G}}$ is a ***fixed point*** of a group $\mathcal{G}$ if $g(x_{\mathcal{G}}) = x_{\mathcal{G}}$ for all $g \in \mathcal{G}$.

Definition 18. Given a linear space $\mathcal{X}$, let $\mathcal{Y}$ and $\mathcal{Z}$ be subspaces of $\mathcal{X}$. Then $\mathcal{X}$ is the ***direct sum*** of $\mathcal{Y}$ and $\mathcal{Z}$, denoted as $\mathcal{X} = \mathcal{Y} \oplus \mathcal{Z}$, if we have $\mathcal{X} = \{y + z \ : \ y \in \mathcal{Y}, z \in \mathcal{Z}\}$ and $\mathcal{Y} \cap \mathcal{Z} = \{0\}$.

Note that the direct sum we used throughout this paper is the *internal direct sum* since $\mathcal{Y}$ and $\mathcal{Z}$ are subspaces rather than spaces. We then present a generic theory of determining stationary points as follows. The proof is provided in Appendix 7.5.1.

Theorem 15 (Stationary Fixed Point)**.** Suppose f has an invariant group $\mathcal{G}$ and $\mathcal{G}$ has a fixed point $x_{\mathcal{G}}$. If we have

$$\mathcal{G}(\mathbb{R}^m) \overset{\triangle}{=} \mathrm{Span}\{g(x) - x \ : \ g \in \mathcal{G}, x \in \mathbb{R}^m\} = \mathbb{R}^m,$$

then $x_{\mathcal{G}}$ is a stationary point of f.

By Theorem 15, we can find a stationary point of functions with invariant groups given a fixed point. Refined result can be obtained for subspaces when we consider a decomposition $\mathbb{R}^m = \mathcal{Y} \oplus \mathcal{Z}$, where $\mathcal{Y}$ and $\mathcal{Z}$ are orthogonal subspaces of $\mathbb{R}^m$. This naturally induces a subgroup of $\mathcal{G}$ as

$$\mathcal{G}_{\mathcal{Y}} = \{g_{\mathcal{Y}} \ : \ g_{\mathcal{Y}}(y) = g(y \oplus 0), g \in \mathcal{G}, y \in \mathcal{Y}, 0 \in \mathcal{Z}\}.$$

Obviously, $\mathcal{G}_\mathcal{Y}$ is a subgroup of a special linear group on $\mathcal{Y}$. Moreover, $y_{\mathcal{G}_\mathcal{Y}} = \mathcal{P}_\mathcal{Y}(x_\mathcal{G}) \in \mathcal{Y}$ is a fixed point of $\mathcal{G}_\mathcal{Y}$, where $\mathcal{P}_\mathcal{Y}$ is a projection operation onto $\mathcal{Y}$. We then have the following corollary immediately from Theorem 15.

Corollary 2. If $y_{\mathcal{G}_\mathcal{Y}}$ is a fixed point of $\mathcal{G}_\mathcal{Y}$ and

$$z^*(y_{\mathcal{G}_\mathcal{Y}}) \in \operatorname*{arg\,zero}_z \nabla_z f(y_{\mathcal{G}_\mathcal{Y}} \oplus z),$$

where $\operatorname{arg\,zero}_z \nabla_z f(y_{\mathcal{G}_\mathcal{Y}} \oplus z)$ is the set of zero solutions of $\nabla_z f(y \oplus z)$ by fixing $y = y_{\mathcal{G}_\mathcal{Y}}$, then $g(y_{\mathcal{G}_\mathcal{Y}} \oplus z^*)$ is a stationary point for all $g \in \mathcal{G}$.

Given a fixed point in a subspace, we have from Corollary 2 that the direct sum of the fixed point and any zero solution of the partial derivative of the function with respect to the orthogonal subspace is also a stationary point. This allows us to recursively use Theorem 15 and Corollary 2 to find a set of stationary points. We call such a procedure the *symmetry principle* of stationary point. Here, we demonstrate some popular examples with symmetric structures.

Example 1 (Low-rank Matrix Factorization)**.** Recall that given a PSD matrix $M^* = UU^\top$ for some $U \in \mathbb{R}^{n\times r}$, the objective function with respect to variable $X \in \mathbb{R}^{n\times r}$ admits

$$f(X) = \frac{1}{4}\|XX^\top - M^*\|_\mathrm{F}^2. \tag{6.5}$$

Given $g = \Psi_r \in \mathfrak{O}_r$, let $g(X) = X\Psi_r$, then we have $f(X) = f(g(X))$. It is easy to see that the rotation group $\mathcal{G} = \mathfrak{O}_r$ is an invariant group of f and $X_\mathcal{G} = 0$ is a fixed point. Theorem 15 implies that 0 is a stationary point.

The gradient of $f(X)$ is

$$\nabla f(X) = (XX^\top - M^*)X.$$

We consider the subspace $\mathcal{Y} \subseteq \mathcal{L}_U$ of the column space of U and $X_{\mathcal{G}_\mathcal{Y}} = 0_\mathcal{Y}$. Applying Corollary 2 to $\mathcal{Y} = \{0\}$ and $\mathcal{Z} = \mathcal{L}_U$, we have $U\Psi_r$ is a stationary point, where $\Psi_r \in \mathfrak{O}_r$. Analogously, applying Corollary 2 again to $\mathcal{Y} = \mathcal{L}_{U_{r-s}} \subseteq \mathcal{L}_U$ and $\mathcal{Z} = \mathcal{L}_{U_s} \subseteq \mathcal{L}_U$, we have $U_s\Psi_r$ is a stationary point of $f(X)$, where $\Psi_r \in \mathfrak{O}_r$, $U_s = \Phi\Sigma S\Theta^\top$ and $U_{r-s} =$

$\Phi\Sigma(I-S)\Theta^\top$ given the SVD of $U=\Phi\Sigma\Theta^\top$, and S is a diagonal matrix with arbitrary s entries being 1 and the rest being 0 for all $s\in[r]$. This will be discussed in further details in Section 6.3. Note that the degree of freedom of Ψ_r in $U_s\Psi_r$ is in fact $s(s-1)/2$ instead of $r(r-1)/2$, since U_s is of rank s.

The result can be easily extended to general low-rank rectangular matrices. For $X,U\in\mathbb{R}^{n\times r}$ and $Y,V\in\mathbb{R}^{m\times r}$, we consider the function

$$f(X,Y)=\frac{1}{2}\|XY^\top-M^*\|_{\mathrm{F}}^2. \tag{6.6}$$

Using the similar analysis for the symmetric case above, we have $(X,Y)=(0,0)$ and $(X,Y)=(U\Psi_r,V\Psi_r)$ are both stationary points. Moreover, given the SVD of $UV^\top=\Phi\Sigma\Theta^\top$, we have $(X,Y)=(\Phi\Sigma_1S\Psi_r,\Theta\Sigma_2S\Psi_r)$ is a stationary point, where $\Sigma_1\Sigma_2=\Sigma$, and S is a diagonal matrix with arbitrary s entries being 1 and the rest being 0, for all $s\in[r]$. Some early works also quantify the stationary points for the low-rank matrix factorization scenario, e.g., [221]. But as our following examples indicate, our generic theory goes beyond the low-rank matrix factorization, which also covers a wide class of problems.

Example 2 (Phase Retrieval)**.** Given i.i.d. complex Gaussian vectors $\{a_i\}_{i=1}^m$ in $\mathbb{C}^n$ and measurements $y_i=|a_i^{\mathrm{H}}u|$ of complex vector $u\in\mathbb{C}^n$ for $i=1,\ldots,m$, where x^{H} is the Hermitian transpose, a natural square error formulation of the objective of phase retrieval with respect to variable $x\in\mathbb{C}^n$ [213, 214] is

$$h(x)=\frac{1}{2m}\sum_{i=1}^m\left(y_i^2-|a_i^{\mathrm{H}}x|^2\right)^2.$$

For simplicity, we consider the expected objective of h as

$$f(x)=\mathbb{E}(h(x))=\|x\|_2^4+\|u\|_2^4-\|x\|_2^2\|u\|_2^2-|x^{\mathrm{H}}u|^2,$$

It is easy to see that f has an invariant group $\mathcal{G}=\left\{\mathrm{e}^{i\theta}\ :\ \theta\in[0,2\pi)\right\}$ and $x_{\mathcal{G}}=0$ is a fixed point. Then Theorem 15 implies that 0 is a stationary point.

The gradient of $f(x)$ is

$$\nabla f(x) = \begin{bmatrix} (2\|x\|_2^2 I - \|u\|_2^2 I - uu^{\mathrm{H}})x \\ (2\|x\|_2^2 I - \|u\|_2^2 I - uu^{\mathrm{H}})\overline{x} \end{bmatrix},$$

where $\overline{x}$ is the complex conjugate. Consider a coordinate-wise subspace $\mathcal{Y} \subseteq \mathbb{C}^n$ of degree $k \leq n$, where for any $\tilde{y} \in \mathcal{Y}$, $\tilde{y}$ shares identical entire with x in certain k coordinates and has zero entries otherwise. Applying Corollary 2 to $\mathcal{Y} = \{0\}$, i.e., $k = 0$, we have that $ue^{i\theta}$ is a stationary point for any $\theta \in [0, 2\pi)$. For $\mathcal{Y} \neq \{0\}$, i.e., $k > 0$, we have $z^*(0_{\mathcal{Y}}) \in \mathcal{D} = \left\{x \in \mathcal{Z} \ : \ x^{\mathrm{H}}u = 0, \ x_{\mathcal{Y}} = 0, \ \|x\|_2 = \|u\|_2/\sqrt{2}\right\}$. Applying Corollary 2 again, we have $xe^{i\theta}$ is a stationary point for any $x \in \mathcal{D}$ and $\theta \in [0, 2\pi)$.

Example 3 (Deep Linear Neural Networks)**.** Given data $W \in \mathbb{R}^{n_0 \times m}$ and $Y \in \mathbb{R}^{n_L \times m}$, we consider a square error objective of a feedforward deep linear neural network of L layers [222],

$$f(X_1, \ldots, X_L) = \frac{1}{2}\|X_L X_{L-1} \cdots X_1 W - Y\|_{\mathrm{F}}^2,$$

where $X_l \in \mathbb{R}^{n_l \times n_{l-1}}$ is the weight matrix in the l-th layer for all $l \in [L]$. We can see that for any $l \in [L-1]$, f has orthogonal groups $\mathcal{G}_l = \mathfrak{O}_{n_l}$ as the invariant groups and $X_{\mathcal{G}_l} = 0$ is a fixed point. Theorem 15 implies that 0 is a stationary point.

The blockwise structure naturally leads to a derivation of further stationary points by fixing all but one block. Specifically, given some $l \in [L-1]$, we fix all the other blocks $[L-1]\backslash\{l\}$, then the gradient of $f(X_1, \ldots, X_L)$ with respect to X_l is

$$\nabla_{X_l} f(X_1, \ldots, X_L) = A^\top (A X_l B - Y) B^\top,$$

where $A = X_L \cdots X_{l+1}$ and $B = X_{l-1} \cdots X_1 W$. Solving $\nabla_{X_l} f(X_1, \ldots, X_L) = 0$, we have that X_l is a stationary point if X_l satisfies

$$X_l = (A^\top A)^- A^\top Y B^\top (BB^\top)^- + (I - (A^\top A)^- A^\top A) Q_1 + Q_2 (I - BB^\top (BB^\top)^-),$$

where D^- is a generalized inverse of the matrix D and $Q_1, Q_2 \in \mathbb{R}^{n_l \times n_{l-1}}$ are arbitrary matrices. Denote the space $\tilde{\mathcal{L}} = \{(I - (A^\top A)^- A^\top A) Q_1 + Q_2 (I -$

$BB^\top(BB^\top)^-) \ : \ Q_1, Q_2 \in \mathbb{R}^{n_l \times n_{l-1}}\}$. We consider a subspace $\mathcal{Y} \subset \tilde{\mathcal{L}}$, then Corollary 2 implies that $(X_{l+1}\Psi_{n_l}, \Psi_{n_l}^\top X_l)$ is a pair of stationary point, where $X_l = z^*(0_{\mathcal{Y}}) = (A^\top A)^- A^\top Y B^\top (BB^\top)^- + U$ for any $U \in \tilde{\mathcal{L}} \backslash \mathcal{Y}$ and $\Psi_{n_l} \in \mathfrak{O}_{n_l}$. Moreover, for $\mathcal{Y} = \tilde{\mathcal{L}}$, we have $z^*(0_{\mathcal{Y}}) = (A^\top A)^- A^\top Y B^\top (BB^\top)^-$ and Corollary 2 implies that $\big(X_{l+1}\Psi_{n_l}, \Psi_{n_l}^\top (A^\top A)^- A^\top Y B^\top (BB^\top)^-\big)$ is also a pair of stationary point, where $\Psi_{n_l} \in \mathfrak{O}_{n_l}$. Extension to more general deep learning architecture is also studied [223].

6.2.2 Null Space of Hessian Matrix at Stationary Points

We now discuss the null space of the Hessian matrix at a stationary point, which can be used to further distinguish between saddle point and local/global minimum. Our intuition is that the null space of the Hessian matrix should contain the vectors tangent to the invariant group $\mathcal{G}$. We start with a few definitions in manifold [224] as follows.

Definition 19 (Manifold)**.** Given positive integers m and k, we call a subset $\mathcal{M} \subset \mathbb{R}^m$ as a ***smooth k-dimensional manifold*** (or a ***smooth k-submanifold***) if every point $x \in \mathcal{M}$ has an open neighborhood $\mathcal{X} \subset$ such that $\mathcal{X} \cap \mathcal{M}$ is diffeomorphic to an open subset $\mathcal{B} \subset \mathbb{R}^k$, i.e., there exists a function $f : \mathcal{X} \cap \mathcal{M} \to \mathcal{B}$ such that f is bijective, and f and f^{-1} are smooth.

Definition 20 (Tangent Space)**.** Let $\mathcal{M} \subset \mathbb{R}^m$ be a smooth k-dimensional manifold. Given $x \in \mathcal{M}$, we call $v \in \mathbb{R}^m$ as a ***tangent vector*** of $\mathcal{M}$ at x if there exists a smooth curve $\gamma : \mathbb{R} \to \mathcal{M}$ with $\gamma(0) = x$ and $v = \gamma'(0)$. The set of tangent vectors of $\mathcal{M}$ at x is called the ***tangent space*** of $\mathcal{M}$ at x, denoted as

$$T_x\mathcal{M} = \left\{\gamma'(0) \ : \ \gamma : \mathbb{R} \to \mathcal{M} \text{ is smooth }, \ \gamma(0) = v\right\}.$$

A visualization of the manifold and the tangent space is provided in Figure 6.2. The following theorem shows that the null space of the Hessian matrix at a stationary point x contains the tangent space of the set $\mathcal{G}(x) = \{g(x) \ : \ g \in \mathcal{G}\}$. The proof is provided in Appendix 7.5.1.

Theorem 16. If f has an invariant group $\mathcal{G}$ and H_x is the Hessian matrix at a stationary

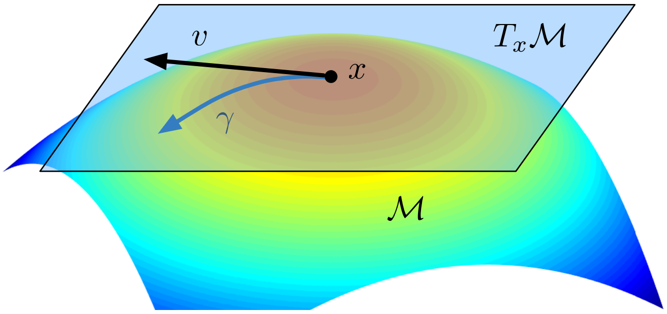


Figure 6.2: A graphical illustration of a manifold $\mathcal{M}$ and a tangent space $T_x\mathcal{M}$ at some point x on the manifold. v is a tangent vector at x and γ is the corresponding smooth curve.

point x, then we have

$$T_x\mathcal{G}(x) \subseteq \mathrm{Null}(H_x).$$

In the following, we demonstrate examples discussed in Section 6.2.1 to instantiate Theorem 16.

Example 4 (Low-rank Matrix Factorization)**.** Remind that for low-rank matrix factorization in Example 1, f has an invariant group $\mathcal{G} = \mathfrak{O}_r$, which is also a smooth submanifold in $\mathbb{R}^{r\times r}$ of dimension $r(r-1)/2$. Given any $X \in \mathbb{R}^{n\times r}$, let $\gamma : \mathbb{R} \to \mathfrak{O}_r(X)$ be a smooth curve, i.e., for every $t \in \mathbb{R}$ there exists $\Psi_r \in \mathfrak{O}_r$ such that $\gamma(t) = g_t(X) = X\Psi_r$ and $\gamma(0) = g_0(X) = X$. By definition, for any $t \in \mathbb{R}$, we have

$$\gamma(t)\gamma(t)^\top = XX^\top.$$

Differentiating both sides, we have $\gamma'(t)\gamma(t)^\top + \gamma(t)\gamma'(t)^\top = 0$. Plugging in $t = 0$, we have

$$\gamma'(0)X^\top + X\gamma'(0)^\top = 0.$$

Then we can see that

$$T_X\mathfrak{O}_r(X) = \{XE \ : \ E \in \mathbb{R}^{r\times r}, E = -E^\top\}.$$

By Example 1, we have that $U_s\Psi_r$ is a stationary point for $\mathcal{Y} = \mathcal{L}_{U_{r-s}} \subseteq \mathcal{L}_U$. Theorem 16 implies that for any skew symmetric matrix $E \in \mathbb{R}^{r\times r}$, we have $U_s\Psi_r E$ belongs to the null space of the Hessian matrix at $U_s\Psi_r$. Similar to Ψ_r, the dimension of $T_X\mathfrak{O}_r(X)$ at $X = U_s\Psi_r E$ depends on s since U_s is of rank s. Specifically, the dimension of the tangent space is at least $s(s-1)/2 + (n-(r-s))(r-s)$, where $s(s-1)/2$ is the degree of freedom of the set of E and $(n-(r-s))(r-s)$ is degree of freedom of $U_s\Psi_r$.

Example 5 (Phase Retrieval)**.** For phase retrieval in Example 2, f has an invariant group $\mathcal{G} = \left\{\mathrm{e}^{i\theta} \ : \ \theta \in [0, 2\pi)\right\}$. Given any $x \in \mathbb{C}^n$, let $\gamma : \mathbb{R} \to \mathcal{G}(x)$ be a smooth curve, i.e., for every $t \in \mathbb{R}$ there exists $\theta \in [0, 2\pi)$ such that $\gamma(t) = x\mathrm{e}^{i\theta}$ and $\gamma(0) = x$. Then for any $t \in \mathbb{R}$, we have

$$\|\gamma(t)\|_2^2 = \|x\|_2^2.$$

Differentiating both sides, we have $\gamma'(t)^{\mathrm{H}}\gamma(t) + \gamma(t)^{\mathrm{H}}\gamma'(t) = 0$. Plugging in $t = 0$, we have

$$\gamma'(0)^{\mathrm{H}}x = -x^{\mathrm{H}}\gamma'(0).$$

Then we can see that

$$T_x\mathcal{G}(x) = ix.$$

By Example 2, we have $u\mathrm{e}^{i\theta}$ is a stationary point for all $\theta \in [0, 2\pi)$. Theorem 16 implies that $iu\mathrm{e}^{i\theta}$ belongs to the null space of Hessian matrix at $u\mathrm{e}^{i\theta}$.

Example 6 (Deep Linear Neural Networks)**.** For the deep linear neural networks in Example 3, f has an invariant group $\mathcal{G}_l = \mathfrak{O}_{n_l}$ for any $l \in [L-1]$. Using the same analysis in Example 4, we have that for any skew symmetric matrix $E \in \mathbb{R}^{r\times r}$, the pair $\left(X_{l+1}\Psi_{n_l}E, E^\top\Psi_{n_l}^\top(A^\top A)^- A^\top Y B^\top(BB^\top)^-\right)$ belongs to the null space of Hessian matrix for a stationary pair $\left(X_{l+1}\Psi_{n_l}, \Psi_{n_l}^\top(A^\top A)^- A^\top Y B^\top(BB^\top)^-\right)$.

6.3 A Geometric Analysis of Low-Rank Matrix Factorization

We apply our generic theories to study the global landscape of the low-rank matrix factorization problem. Our goal is to provide a comprehensive geometric perspective to fully characterize the low-rank matrix factorization problem (6.4). Finding all stationary points is the keystone, based on which we can further identify strict saddle points and global minima. This scheme has been adopted in geometry based convergence rate analyses to guarantee that iterative algorithms do not converge to the strict saddle point [187,214–216]. The landscape of the low-rank matrix factorization problem is also discussed briefly in [225], but no rigorous analysis is provided.

In particular, the zero of the gradient $\nabla\mathcal{F}(X)$ and the eigenspace of the Hessian matrix $\nabla^2\mathcal{F}(X)$ are keys to our analysis. Given $\nabla\mathcal{F}(X)$ and $\nabla^2\mathcal{F}(X)$, our analysis consists of the following major arguments:

(p1) identify all stationary points by finding the solutions of $\nabla\mathcal{F}(X) = 0$, which is further used to identify the strict saddle point and the global minimum,

(p2) identify the strict saddle point and their neighborhood such that $\nabla^2\mathcal{F}(X)$ has a negative eigenvalue, i.e. $\lambda_{\min}(\nabla^2\mathcal{F}(X)) < 0$,

(p3) identify the global minimum, their neighborhood, and the directions such that $\mathcal{F}(X)$ is strongly convex, i.e. $\lambda_{\min}(\nabla^2\mathcal{F}(X)) > 0$, and

(p4) verify that the gradient has a sufficiently large norm outside the regions described in (p2) and (p3).

The analysis can be further extended to other problems, such as matrix sensing and matrix completion, which are considered as perturbed versions of (6.4). For simplicity, we first consider the PSD matrix $M^* = UU^\top$. Then we explain how to extend to a rectangular matrix, which is straightforward.

6.3.1 Warm-up: Rank 1 Case

We start with the basic case of $r = 1$ to obtain some insights. Specifically, suppose $M^* = uu^\top$, where $u \in \mathbb{R}^n$, then we consider

$$\min_{x \in \mathbb{R}^n} \mathcal{F}(x), \text{ where } \mathcal{F}(x) = \frac{1}{4}\|uu^\top - xx^\top\|_{\mathrm{F}}^2. \tag{6.7}$$

The gradient and the Hessian matrix of $\mathcal{F}(x)$, respectively, are

$$\begin{aligned} \nabla \mathcal{F}(x) &= (xx^\top - uu^\top)x \in \mathbb{R}^n \quad \text{and} \\ \nabla^2 \mathcal{F}(x) &= 2xx^\top + \|x\|_2^2 \cdot I_n - uu^\top \in \mathbb{R}^{n\times n}. \end{aligned} \tag{6.8}$$

In the rank 1 case, the invariant group is $\mathcal{G} = \mathfrak{O}_1 = \{1, -1\}$. We then provide the key arguments for the rank 1 setting in the following theorem. The proof is provided in Appendix 7.5.2.

Theorem 17. Consider (6.7) and define the following regions:

$$\begin{aligned} \mathcal{R}_1 &\triangleq \left\{ y \in \mathbb{R}^n \ : \ \|y\|_2 \leq \frac{1}{2}\|u\|_2 \right\}, \\ \mathcal{R}_2 &\triangleq \left\{ y \in \mathbb{R}^n \ : \ \min_{\psi \in \mathfrak{O}_1} \|y - u\psi\|_2 \leq \frac{1}{8}\|u\|_2 \right\}, \text{ and} \\ \mathcal{R}_3 &\triangleq \left\{ y \in \mathbb{R}^d \ : \ \|y\|_2 > \frac{1}{2}\|u\|_2, \ \min_{\psi \in \mathfrak{O}_1} \|y - u\psi\|_2 > \frac{1}{8}\|u\|_2 \right\}. \end{aligned}$$

Then the following properties hold.

(p1) $x = 0$, u and $-u$ are the only stationary points of $\mathcal{F}(x)$.

(p2) $x = 0$ is a strict saddle point, where $\nabla^2 \mathcal{F}(0)$ is negative semi-definite with $\lambda_{\min}(\nabla^2 \mathcal{F}(0)) = -\|u\|_2^2$. Moreover, for any $x \in \mathcal{R}_1$, $\nabla^2 \mathcal{F}(x)$ contains a negative eigenvalue, i.e.

$$\lambda_{\min}(\nabla^2 \mathcal{F}(x)) \leq -\frac{1}{2}\|u\|_2^2.$$

(p3) For $x = \pm u$, x is a global minimum, and $\nabla^2 \mathcal{F}(x)$ is positive definite with $\lambda_{\min}(\mathcal{F}(x)) = \|u\|_2^2$. Moreover, for any $x \in \mathcal{R}_2$, $\mathcal{F}(x)$ is locally strongly convex, i.e.

$$\lambda_{\min}(\nabla^2 \mathcal{F}(x)) \geq \frac{1}{5}\|u\|_2^2.$$

(p4) For any $x \in \mathcal{R}_3$, we have

$$\|\nabla \mathcal{F}(x)\|_2 > \frac{\|u\|_2^3}{8}.$$

The rank 1 setting is intuitive since there is only one strict saddle point and 2 isolated global minima. It is also important to notice that

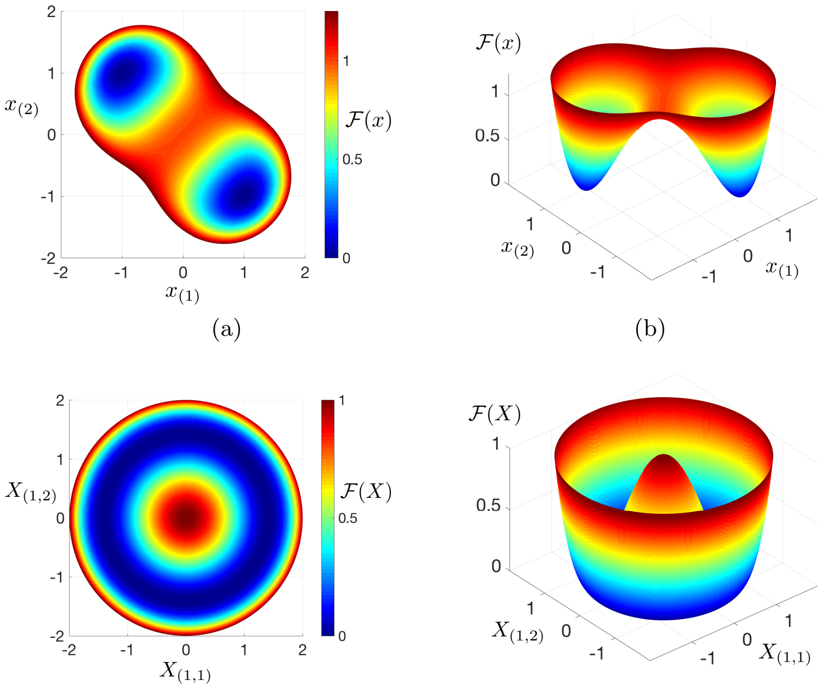
$$\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3 = \mathbb{R}^n.$$

Thus, the entire space $\mathbb{R}^n$ is parameterized by one of the regions: (I) the neighborhood of the strict saddle point, where the Hessian matrix $\nabla^2 \mathcal{F}(x)$ has negative eigenvalues; (II) the neighborhood of the global minima, where $\mathcal{F}(x)$ is strongly convex; and (III) the gradient $\nabla \mathcal{F}(x)$ has a sufficiently large norm. To better understand the landscape, we provide a visualization of the objective function $\mathcal{F}(x)$ in Figure 6.3 (a and b). We set $u = [1 \ -1]^\top$, thus $M^* = uu^\top = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$. It is easy to see that $x = [0 \ 0]^\top$ is a strict saddle point and $x = \pm u$ are global minima, which matches with our analysis.

6.3.2 General Ranks

We then consider the general setting of $r \geq 1$, where $M^* = UU^\top$, $U \in \mathbb{R}^{n \times r}$. Characterizing the global landscape becomes much more involved as neither the strict saddle point nor the global minimum is isolated. Recall that we consider

$$\min_{X \in \mathbb{R}^{n \times r}} \mathcal{F}(X), \text{ where } \mathcal{F}(X) = \frac{1}{4}\|M^* - XX^\top\|_{\mathrm{F}}^2. \tag{6.9}$$



(a) (b)

(c) (d)

Figure 6.3: The visualization of objective functions $\mathcal{F}(X)$ for $r = 1$ (a and b) and $r = 2$ (c and d) using contour plots. In the case $r = 1$, the global minima are $x = [x_{(1)}\ x_{(2)}]^\top = [1\ \ -1]^\top$ and $[-1\ 1]^\top$. In the case $r = 2$, the global minima are $X = [X_{(1,1)}\ X_{(1,2)}]\Psi = [1\ \ -1]\Psi$ for all $\Psi \in \mathfrak{O}_2$, i.e. any X with $\|X\|_2 = \sqrt{2}$ is a global minimum. Note that we can only visualize $X \in \mathbb{R}^{1\times 2}$ when $r = 2$. Here $M^* = UU^\top = [1\ \ -1][1\ \ -1]^\top = 2$ is not low-rank in fact, and $X = [0\ 0]$ is not a strict saddle point but a local maximum.

For notational convenience, for any matrix X, we define:

$$\Psi_X \overset{\triangle}{=} \arg\min_{\Psi \in \mathfrak{O}_r} \|X - U\Psi\|_2 \quad \text{and}$$

$$K_X \overset{\triangle}{=} \begin{bmatrix} X_{(*,1)}X_{(*,1)}^\top & X_{(*,2)}X_{(*,1)}^\top & \cdots & X_{(*,r)}X_{(*,1)}^\top \\ X_{(*,1)}X_{(*,2)}^\top & X_{(*,2)}X_{(*,2)}^\top & \cdots & X_{(*,r)}X_{(*,2)}^\top \\ \vdots & \vdots & \ddots & \vdots \\ X_{(*,1)}X_{(*,r)}^\top & X_{(*,2)}X_{(*,r)}^\top & \cdots & X_{(*,r)}X_{(*,r)}^\top \end{bmatrix}. \tag{6.10}$$

Further, we introduce two sets:

$$\mathcal{X} = \left\{X = \Phi\Sigma_2\Theta_2 \ : \ U \text{ has the SVD } U = \Phi\Sigma_1\Theta_1,\ (\Sigma_2^2 - \Sigma_1^2)\Sigma_2 = 0, \Theta_2 \in \mathfrak{O}_r\right\} \quad \text{and}$$
$$\mathcal{U} = \{X \in \mathcal{X} \ : \ \Sigma_2 = \Sigma_1\}.$$

The set $\mathcal{X}$ contains all strict saddle points, and $\mathcal{U}$ is the set of all global minima, which will be proved in the following theorem. Specifically, for any X that has a strict subset of the column bases of U and identical corresponding singular values, X is a strict saddle point of $\mathcal{F}$. This indicates that the strict saddle points are not isolated, and there are infinite many of them due to rotations (their measures in $\mathbb{R}^{n\times r}$ are zero). On the other hand, when X is different from U only by a rotation, X is also a global minimum of $\mathcal{F}$.

By algebraic calculation, the gradient and the Hessian matrix of $\mathcal{F}(X)$, respectively, are

$$\nabla\mathcal{F}(X) = (XX^\top - M^*)X \in \mathbb{R}^{n\times r} \text{ and} \tag{6.11}$$
$$\nabla^2\mathcal{F}(X) = K_X + I_r \otimes XX^\top + X^\top X \otimes I_n - I_r \otimes M^* \in \mathbb{R}^{rn\times rn}. \tag{6.12}$$

The gradient (6.11) and the Hessian matrix (6.12) for the general rank $r \geq 1$ reduce to (6.8) when $r = 1$. We provide the key arguments for the general rank setting in the following theorem. The proof is provided in Appendix 7.5.3.

Theorem 18. Consider (6.9) for the general rank $r \geq 1$ and define the following regions:

$$\mathcal{R}_1 \triangleq \left\{Y \in \mathbb{R}^{n\times r} \ : \ \sigma_r(Y) \leq \frac{1}{2}\sigma_r(U),\ \|YY^\top\|_{\mathrm{F}} \leq 4\|UU^\top\|_{\mathrm{F}}\right\},$$
$$\mathcal{R}_2 \triangleq \left\{Y \in \mathbb{R}^{n\times r} \ : \ \min_{\Psi\in\mathfrak{O}_r} \|Y - U\Psi\|_2 \leq \frac{\sigma_r^2(U)}{8\sigma_1(U)}\right\},$$
$$\mathcal{R}_3' \triangleq \left\{Y \in \mathbb{R}^{n\times r} \ : \ \sigma_r(Y) > \frac{1}{2}\sigma_r(U),\ \min_{\Psi\in\mathfrak{O}_r} \|Y - U\Psi\|_2 > \frac{\sigma_r^2(U)}{8\sigma_1(U)},\right.$$
$$\left.\|YY^\top\|_{\mathrm{F}} \leq 4\|UU^\top\|_{\mathrm{F}}\right\}, \text{ and}$$
$$\mathcal{R}_3'' \triangleq \left\{Y \in \mathbb{R}^{n\times r} \ : \ \|YY^\top\|_{\mathrm{F}} > 4\|UU^\top\|_{\mathrm{F}}\right\}.$$

Then the following properties hold.

(p1) For any $X \in \mathcal{X}$, X is a stationary point of $\mathcal{F}(X)$.

(p2) For any $X \in \mathcal{X} \backslash \mathcal{U}$, X is a strict saddle point with $\lambda_{\min}(\nabla^2 \mathcal{F}(X)) \leq -\lambda_{\max}^2(\Sigma_1 - \Sigma_2)$. Moreover, for any $X \in \mathcal{R}_1$, $\nabla^2 \mathcal{F}(X)$ contains a negative eigenvalue, i.e.

$$\lambda_{\min}(\nabla^2 \mathcal{F}(X)) \leq -\frac{\sigma_r^2(U)}{4}.$$

(p3) For any $X \in \mathcal{U}$, X is a global minimum of $\mathcal{F}(X)$, and $\nabla^2 \mathcal{F}(X)$ is positive semidefinite, which has $r(r-1)/2$ zero eigenvalues with the minimum nonzero eigenvalue at least $\sigma_r^2(U)$. Moreover, for any $X \in \mathcal{R}_2$, we have

$$z^\top \nabla^2 \mathcal{F}(X) z \geq \frac{1}{5} \sigma_r^2(U) \|z\|_2^2$$

for any $z \perp \mathcal{E}$, where $\mathcal{E} \subseteq \mathbb{R}^{n \times r}$ is a subspace spanned by all eigenvectors of $\nabla^2 \mathcal{F}(K_E)$ associated with negative eigenvalues, where $E = X - U\Psi_X$ and Ψ_X and K_E are defined in (6.10).

(p4) Further, we have

$$\|\nabla \mathcal{F}(X)\|_{\mathrm{F}} > \frac{\sigma_r^4(U)}{9\sigma_1(U)} \quad \text{for any } X \in \mathcal{R}_3' \quad \text{and}$$
$$\|\nabla \mathcal{F}(X)\|_{\mathrm{F}} > \frac{3}{4} \sigma_1^3(X) \quad \text{for any } X \in \mathcal{R}_3''.$$

The following proposition shows that any $X \in \mathbb{R}^{n \times r}$ belongs to one of the four regions above. The proof is provided in Appendix 7.5.7.

Proposition 3. Consider the four regions defined in Theorem 18, we have

$$\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3' \cup \mathcal{R}_3'' = \mathbb{R}^{n \times r}.$$

Different from the rank 1 setting, we have one more region $\mathcal{R}_3''$, where the gradient has a sufficiently large norm. When $r = 1$, we have $\mathfrak{O}_1 = \{1, -1\}$. Thus $\mathcal{X}$ reduces to $\{0\}$ and $\mathcal{U}$ reduces to $\{u, -u\}$, which matches with the result in Theorem 17. From (p2) of Theorem 18, we have that X is approximately rank deficient in $\mathcal{R}_1$ since $\sigma_r(X) \leq \frac{1}{2}\sigma_r(U)$. From (p3) of Theorem 18, we have that $\mathcal{F}(X)$ is convex at a global minimum, rather than strongly convex. Moreover, in the neighborhood of a global minimum, $\mathcal{F}(X)$ is only strongly convex along certain directions. Analogous results are also provided in previous literature. For example, [192] (in the analysis of Theorem 3.2) show that for any X that satisfies $\|X - U\Psi_X\|_2 \leq c_1\sigma_r(U)$, we have the **Regularity Property**:

$$\langle \nabla \mathcal{F}(X), X - U\Psi_X \rangle \geq c_2\sigma_r^2(U)\|X - U\Psi_X\|_{\mathrm{F}}^2 + c_3\|UU^\top - XX^\top\|_{\mathrm{F}}^2, \tag{6.13}$$

where c_1, c_2, and c_3 are positive real constants. This indicates that when X is close to a global minimum, $\mathcal{F}(X)$ is only strongly convex along the direction of $E = X - U\Psi_X$ (Procrustes difference). But our results are much more general. Specifically, we guarantee in (p3) of Theorem 18 that $\mathcal{F}(X)$ is strongly convex along all directions that are orthogonal to the subspace spanned by eigenvectors associated with negative eigenvalues of $\nabla^2\mathcal{F}(K_E)$ for $K_E = X - U\Psi_X$. As we have shown in the analysis, there are at most $r(r-1)/2$ such directions potentially associated with the negative eigenvalues of $\nabla^2\mathcal{F}(K_E)$. In other words, there are at least $nr - r(r-1)/2$ such directions, where $\mathcal{F}(X)$ is strongly convex. In the following lemma, we further show that $\mathcal{F}(X)$ is nonconvex in any neighborhood of a global minimum. The proof is provided in Appendix 7.5.7.

Proposition 4. Let $\mathcal{B}_\varepsilon(U) = \{X \ : \ \|X - U\|_2 \leq \varepsilon\}$ be a neighborhood of U with radius $\varepsilon > 0$. Then $\mathcal{F}(X)$ is a nonconvex function in $\mathcal{B}_\varepsilon(U)$.

We provide a visualization of the objective function $\mathcal{F}(X)$ in Figure 6.3 (c and d) by setting $r = 2$ and $U = [1 \ \ -1]$. The observation is that any X satisfying $X = U\Psi_2$ is a global minimum, where $\Psi_2 \in \mathfrak{O}_2$. Moreover, if we restrict X to be a convex combination of any two distinct global minima, then $\mathcal{F}(X)$ is nonconvex, as we have shown in Proposition 4. Note that we can only visualize the case of $X \in \mathbb{R}^{1\times 2}$, which results in a full rank $M^* = UU^\top = 2$ here. Thus $X = [0 \ 0]$ is a not strict saddle point in this degenerated example.

6.3.3 General Rectangular Matrices

We further discuss briefly on the scenario where the low-rank matrix is a general rectangular matrix. Recall that for $M^* = UV^\top \in \mathbb{R}^{n\times m}$ for some $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{m\times r}$, we consider

$$\min_{X \in \mathbb{R}^{n\times r}, Y \in \mathbb{R}^{m\times r}} \mathcal{F}(X,Y), \text{ where } \mathcal{F}(X,Y) = \frac{1}{2}\|M^* - XY^\top\|_\mathrm{F}^2. \tag{6.14}$$

Compared with the PSD matrix scenario (6.9) with $M^* \succeq 0$, it has one more issue of scaling invariance for the general rectangular matrix (6.14). Specifically, in addition to the rotation invariance as in the PSD case, when we multiply X and divide Y by an identical (nonzero) constant, $\mathcal{F}(X,Y)$ is also invariant. This results in a significantly increasing complexity of the structure for both strict saddle points and global minima. Moreover, the scaling issue also leads to a badly conditioned problem, e.g., when $\|X\|_F^2$ is very small and $\|Y\|_F^2$ is very large with $XY^\top$ fixed.

For ease of discussion, we provide an example when $n = m = r = 1$. Suppose $M^* = 1$, then the objective in (6.14) is $\mathcal{F}(x,y) = \frac{1}{2}(1-xy)^2$. The corresponding Hessian matrix is

$$\nabla^2\mathcal{F}(x,y) = \begin{bmatrix} y^2 & 2xy-1 \\ 2xy-1 & x^2 \end{bmatrix}.$$

It is easy to see that any (x,y) satisfying $xy = 1$ is a global minimum, which makes the structure of the global minimum much more complicated than the PSD matrix case with rank $r = 1$ (only two global minima points in Figure 6.3). A visualization of $\mathcal{F}(x,y)$ is provided in Figure 6.4 (panel a and b). On the other hand, the problem becomes poorly conditioned, i.e., $\lambda_{\max}(\nabla^2\mathcal{F}(x,y))/\lambda_{\min}(\nabla^2\mathcal{F}(x,y)) \to \infty$ when $\|x\|_2 \to 0$ and $\|y\|_2 \to \infty$ with $xy = 1$.

To avoid such a scaling issue, we consider a regularized form as follows,

$$\min_{X \in \mathbb{R}^{n\times r}, Y \in \mathbb{R}^{m\times r}} \mathcal{F}_\lambda(X,Y), \text{ where } \mathcal{F}_\lambda(X,Y) = \frac{1}{2}\|M^* - XY^\top\|_\mathrm{F}^2 + \frac{\lambda}{4}\|X^\top X - Y^\top Y\|_F^2. \tag{6.15}$$

where $\lambda > 0$ is a regularization parameter. Such a regularization has been considered in related problems of low-rank matrix factorization [192, 207], which enforces positive curvature when X and Y have similar spectrum to avoid the scaling issue discussed above.

Taking the example discussed above again, we have the regularized objective as $\mathcal{F}_\lambda(x,y) = \frac{1}{2}(1-xy)^2 + \frac{\lambda}{4}(x^2-y^2)^2$ and the corresponding Hessian matrix as

$$\nabla^2 \mathcal{F}_\lambda(x,y) = \begin{bmatrix} (1-\lambda)y^2 + 3\lambda x^2 & 2(1-\lambda)xy - 1 \\ 2(1-\lambda)xy - 1 & (1-\lambda)x^2 + 3\lambda y^2 \end{bmatrix}.$$

With a proper value of λ, $\mathcal{F}_\lambda(x,y)$ has strong convexity in the neighborhood of $x = y = 1$ and $x = y = -1$, resulting in a much simplified structure of global minima, analogous to the PSD rank $r = 1$ case. A visualization of of $\mathcal{F}_\lambda(x,y)$ with $\lambda = 0.5$ is provided in Figure 6.4 (panel c and d). Compared with the objective $\mathcal{F}$ without a regularization, the regularized objective $\mathcal{F}_\lambda$ is much better conditioned even when one of $\|x\|_2$ and $\|y\|_2$ is very small and the other is very large.

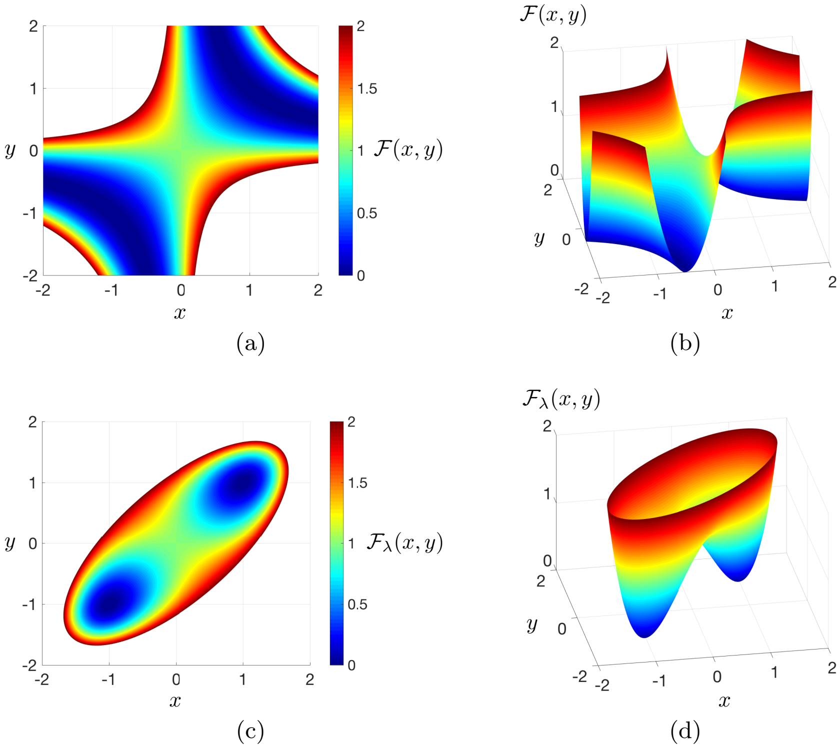
We remark that after the initial release of our paper, [218] provide an extension of our analysis to the case of general rectangular matrices using the lifting formulation. Specifically, they show $U^\top U = V^\top V$ (Lemma 3 therein) at stationary points in the noiseless case, which implies that the stationary points are not affected by the regularization function in (6.15). Beyond stationary points, careful characterization is required to deal with the regularization, which is a fourth order polynomial on the factors (similar to the loss function). Consequently, they achieve analogous geometric result to our Theorem 18 for the asymmetric case.

6.4 Matrix Sensing via Factorization

We extend our geometric analysis to the matrix sensing problem, which can be considered as a perturbed version of the low-rank matrix factorization problem. For simplicity, we first introduce the noiseless scenario and the noisy setting is discussed later, both of which preserve the entire landscape of optimization in the matrix factorization problem.

6.4.1 Matrix Sensing as a Perturbed Matrix Factorization Problem

We start with a formal description of the matrix sensing problem. For all $i \in [d]$, suppose $A_i \in \mathbb{R}^{n\times n}$ has i.i.d. zero mean sub-Gaussian entries with variance 1, then we



(a)

(b)

(c)

(d)

Figure 6.4: The visualization of objective functions $\mathcal{F}(x, y)$ with $u = v = 1$ (a, b) and $\mathcal{F}_\lambda(x, y)$ with $u = v = 1$ and $\lambda = 0.5$ (c, d). For $f(x, y)$, any (x, y) that satisfies $xy = 1$ is a global minimum. For $\mathcal{F}_\lambda(x, y)$, $x = y = 1$ and $x = y = -1$ are the only global minima.

observe

$$y_{(i)} = \langle A_i, M^* \rangle,$$

where $M^* \in \mathbb{R}^{n \times n}$ is a low-rank PSD matrix with $\text{Rank}(M^*) = r$. Denote $M^* = UU^\top$, where $U \in \mathbb{R}^{n \times r}$, then $y_{(i)} = \langle A_i, UU^\top \rangle$ and we recover U by solving

$$\min_X \; F(X), \text{ where } F(X) = \frac{1}{4d} \sum_{i=1}^{d} \langle A_i, XX^\top - M^* \rangle^2. \tag{6.16}$$

The gradient and the Hessian matrix of $F(X)$, respectively, are

$$\nabla F(X) = \frac{1}{2d}\sum_{i=1}^{d}\langle A_i, XX^\top - M^*\rangle \cdot (A_i + A_i^\top)X \quad \text{and} \tag{6.17}$$

$$\begin{aligned}\nabla^2 F(X) = \frac{1}{2d}\sum_{i=1}^{d} I_r \otimes \langle A_i, XX^\top - M^*\rangle \cdot (A_i + A_i^\top)\\ + \operatorname{vec}\Big((A_i + A_i^\top)X\Big) \cdot \operatorname{vec}\Big((A_i + A_i^\top)X\Big)^\top.\end{aligned} \tag{6.18}$$

We first show the connection between the matrix sensing problem and the low-rank matrix factorization problem in the following lemma. The proof is provided in Appendix 7.5.7.

Lemma 4. We have $\mathbb{E}(F(X)) = \mathcal{F}(X)$, $\mathbb{E}(\nabla F(X)) = \nabla\mathcal{F}(X)$, and $\mathbb{E}(\nabla^2 F(X)) = \nabla^2\mathcal{F}(X)$.

From Lemma 4, we have that the objective (6.16), the gradient (6.17), and the Hessian matrix (6.18) of the matrix sensing problem are unbiased estimators of the counterparts of the low-rank matrix factorization problem in (6.9), (6.11), and (6.12) respectively. We then provide a finite sample perturbation bound for the gradient and the Hessian matrix of the matrix sensing problem. The proof is provided in Appendix 7.5.7.

Lemma 5. Suppose $N \geq \max\{\|XX^\top - M^*\|_{\mathrm{F}}^2, \|X\|_{\mathrm{F}}^2, 1\}$. Given $\delta > 0$, if d satisfies

$$d = \Omega(N \max\{nr, \sqrt{nr}\log(nr)\}/\delta),$$

then with high probability, we have

$$\|\nabla^2 F(X) - \nabla^2\mathcal{F}(X)\|_2 \leq \delta \quad \text{and} \quad \|\nabla F(X) - \nabla\mathcal{F}(X)\|_{\mathrm{F}} \leq \delta.$$

From Lemma 5, we have that the landscape of the gradient and the Hessian matrix of low-rank matrix factorization is preserved for matrix sensing with high probability based on the concentrations of sub-Gaussian designs $\{A_i\}_{i=1}^d$, as long as the sample size d is sufficiently large. These further allow us to derive the key properties (p1) – (p4) for matrix sensing directly from the counterparts of low-rank matrix factorization in

Theorem 18. We formalize the result in the following Theorem. The proof is provided in Appendix 7.5.4.

Theorem 19. Consider (6.16) for the general rank $r \geq 1$. If d satisfies

$$d \geq C \cdot \max\left\{nr^2,\ n\sqrt{r}\log(n),\ r\sqrt{nr}\log(nr)\right\},$$

where $C > 0$ is a generic real constant, then with high probability, we have the following properties.

(p1) For any $X \in \mathcal{U} \cup \{0\}$, X is a stationary point of $F(X)$.

(p2) $X = 0$ is a strict saddle point with $\lambda_{\min}(F(0)) \leq -\frac{7}{8}\|U\|_2^2$. Moreover, for any $X \in \mathcal{R}_1$, $\nabla^2 F(X)$ contains a negative eigenvalue, i.e.

$$\lambda_{\min}(\nabla^2 F(X)) \leq -\frac{\sigma_r^2(U)}{8}.$$

(p3) For any $X \in \mathcal{U}$, X is a global minimum, and $\nabla^2 F(X)$ is positive semidefinite. Moreover, for any $X \in \mathcal{R}_2$, we have

$$z^\top \nabla^2 F(X) z \geq \frac{1}{10}\sigma_r^2(U)\|z\|_2^2$$

for any $z \perp \mathcal{E}$, where $\mathcal{E} \subseteq \mathbb{R}^{n\times r}$ is a subspace is spanned by all eigenvectors of $\nabla^2\mathcal{F}(K_E)$ associated with negative eigenvalues, where $E = X - U\Psi_X$ and Ψ_X and K_E are defined in (6.10).

(p4) Further, we have

$$\|\nabla F(X)\|_{\mathrm{F}} > \frac{\sigma_r^4(U)}{18\sigma_1(U)} \quad \text{for any } X \in \mathcal{R}_3' \quad \text{and}$$
$$\|\nabla F(X)\|_{\mathrm{F}} > \frac{1}{4}\sigma_1^3(X) \quad \text{for any } X \in \mathcal{R}_3''.$$

From Theorem 19, we have that the landscape of the low-rank matrix factorization problem is preserved for the matrix sensing problem given a sufficiently large sample size d. This is to say, $F(X)$ has a negative curvature in the neighborhoods of strict saddle points, strong convexity along certain directions in the neighborhoods of global minima, and a sufficiently large norm for the gradient in the rest of domain. On the other hand, due to random perturbations by sensing matrices $\{A_i\}_{i=1}^d$, the set of strict saddle points in $\mathcal{X}\backslash\mathcal{U}$ reduces to $\{0\}$, while the rest of the points in $\mathcal{X}\backslash\mathcal{U}$ are nearly strict saddle.

6.4.2 Noisy Observation

We further consider a noisy scenario of the matrix sensing problem. Specifically, suppose $\{A_i\}_{i=1}^d$ are random matrices described above, then we observe

$$y_{(i)} = \langle A_i, M^* \rangle + z_{(i)} \quad \text{for all} \quad i \in [d],$$

where $\{z_{(i)}\}_{i=1}^d$ are independent zero mean sub-Gaussian random noise with variance σ_z^2. Consequently, denoting $M^* = UU^\top$, we recover U by solving

$$\min_X \ F(X), \ \text{where } F(X) = \frac{1}{4d}\sum_{i=1}^d \left(\langle A_i, XX^\top - M^* \rangle - z_{(i)}\right)^2. \tag{6.19}$$

We then provide the key properties (p1) – (p4) for the noisy version of the matrix sensing problem in the following corollary. The proof is provided in Appendix 7.5.5.

Corollary 3. Consider (6.19) for the general rank $r \geq 1$. Given $\varepsilon > 0$, if d satisfies

$$d \geq \frac{C\sigma_z^2 \cdot \max\left\{nr^2, \ n\sqrt{r}\log(n), \ r\sqrt{nr}\log(nr)\right\}}{\varepsilon^2},$$

where $C > 0$ is a generic real constant, then with high probability, we have that properties (p1) – (p4) in Theorem 19 hold, as well as the following estimation error

$$\|\widehat{M} - M^*\|_{\mathrm{F}}^2 = \mathcal{O}\left(\varepsilon^2\right),$$

where $\widehat{M} = \widehat{X}\widehat{X}^\top$ for $\widehat{X} = \arg\min_X F(X)$ in (6.19).

Compared with Theorem 19, the sufficient sample complexity for preserving the key properties (p1) – (p4) of the landscape in Corollary 3 has one more dependence on the variance of noise, which is a natural result for noisy measurements. We remark that preserving the global landscape is more challenging than guaranteeing the convergence to a local minimum within the optimal distance to the true model parameter, which only requires a local analysis in a neighborhood of the true model parameter. Existing results only discuss some local geometry instead of the global one as we do, such as the strict saddle points and the neighborhood of true model parameter [188, 193].

6.5 Discussion

We provide further discussion on extending our analysis for matrix sensing to achieve the optimal sample complexity by relaxing the geometric properties as a tradeoff. In addition, we make some comments on how the geometric analysis in this paper can imply strong convergence guarantees for several popular iterative algorithms.

6.5.1 From Suboptimal to Optimal Sampling Complexity for Matrix Sensing

The sampling complexity is $\mathcal{O}(nr^2)$ for matrix sensing when we preserve the entire landscape of the matrix factorization problem (6.9). If we relax the properties of optimization landscape to be preserved, the optimal complexity $\mathcal{O}(nr)$ can be attained. In specific, consider the noiseless scenario by solving (6.16). Then we have the following geometric properties for matrix sensing. The proof is provided in Appendix 7.5.6.

Theorem 20. Consider (6.9) for the general rank $r \geq 1$ and define the following regions:

$$
\begin{aligned}
\mathcal{R}_1 &\triangleq \left\{ Y \in \mathbb{R}^{n\times r} \ : \ \min_{\Psi \in \mathfrak{O}_r} \|Y - U\Psi\|_2 > \frac{\sigma_r(U)}{4}, \|\nabla F(Y)\|_{\mathrm{F}} \leq \frac{\sigma_r^3(U)}{96} \right\}, \\
\mathcal{R}_2 &\triangleq \left\{ Y \in \mathbb{R}^{n\times r} \ : \ \min_{\Psi \in \mathfrak{O}_r} \|Y - U\Psi\|_2 \leq \frac{\sigma_r(U)}{4} \right\}, \quad \text{and} \\
\mathcal{R}_3 &\triangleq \left\{ Y \in \mathbb{R}^{n\times r} \ : \ \min_{\Psi \in \mathfrak{O}_r} \|Y - U\Psi\|_2 > \frac{\sigma_r(U)}{4}, \|\nabla F(Y)\|_{\mathrm{F}} > \frac{\sigma_r^3(U)}{96} \right\},
\end{aligned}
$$

If d satisfies

$$
d \geq C \cdot nr,
$$

where C is a generic real constant, then with high probability, we have the following properties.

(p1) For any $X \in \mathcal{U} \cup \{0\}$, X is a stationary point of $F(X)$.

(p2) [Direct result from [219]] For any $X \in \mathcal{R}_1$, including the strict saddle point $X = 0$, $\nabla^2 F(X)$ contains a negative eigenvalue, i.e.

$$\lambda_{\min}(\nabla^2 F(X)) \leq -\frac{1}{6}\sigma_r^2(U).$$

(p3) For any $X \in \mathcal{U}$, X is a global minimum, and $\nabla^2 F(X)$ is positive semidefinite. Moreover, for any $X \in \mathcal{R}_2$ with Ψ_X defined in (6.10), we have

$$\langle \nabla F(X), X - U\Psi_X \rangle \geq \frac{\sigma_r^2(U)}{4} \left\| X - U\Psi_X \right\|_{\mathrm{F}}^2 + \frac{1}{20 \left\| U \right\|_2} \left\| \nabla F(X) \right\|_{\mathrm{F}}^2.$$

(p4) Further, for any $X \in \mathcal{R}_3$, we have

$$\left\| \nabla F(X) \right\|_{\mathrm{F}} > \frac{\sigma_r^3(U)}{96}.$$

It is immediate from Theorem 20 that we have

$$\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3 = \mathbb{R}^n.$$

When $d = \Omega(nr)$, weaker properties of optimization landscape can be obtained. First of all, unlike $\mathcal{R}_1$ of Theorem 18, it is not clear whether there is (approximate) rank deficiency in $\mathcal{R}_1$ from Theorem 20. Since the rank deficiency is a key reason for generating strict saddle points, we face a gap in the geometric interpretation. Moreover, in the neighborhood of global minima in (p3), we have the regularity property (6.13). As we have discussed after Theorem 18, this is a weaker result than (p3) therein, which can guarantee the strong convexity in a larger number of directions. We suspect that this is a tradeoff between the optimal sample complexity and strong geometric properties (though this may be a proof artifact). In addition, the characterization of both regions $\mathcal{R}_1$ and $\mathcal{R}_3$ in Theorem 20 depend on both problem parameter X and sensing matrices

$\{A_i\}_{i=1}^d$ (embedded in $\nabla F(X)$). This makes the regions less explicit than $\mathcal{R}_1$ and $\mathcal{R}_3$ in Theorem 19, which only on X.

We further address a brief comparison with [192, 219]. Our Theorem 20 has slightly stronger geometric guarantees than [192, 219] under the same conditions. In specific, due to a refined analysis, our neighborhood of global minima $\mathcal{R}_2$ characterized via the spectral norm of the Procrustes difference is larger than the corresponding region in [192, 219] characterized via the Frobenius norm, i.e., for all $\operatorname{rank}(U) > 1$, we have

$$\left\{ Y \in \mathbb{R}^{n\times r} : \min_{\Psi\in\mathfrak{O}_r} \|Y - U\Psi\|_{\mathrm{F}} \leq \frac{\sigma_r(U)}{4} \right\} \subset \left\{ Y \in \mathbb{R}^{n\times r} : \min_{\Psi\in\mathfrak{O}_r} \|Y - U\Psi\|_2 \leq \frac{\sigma_r(U)}{4} \right\}.$$

Moreover, [192] only provide a local geometric property in the neighborhood of global minima $\mathcal{R}_2$ using the regularity property. In contrast, we provide a global one in Theorem 20.

6.5.2 Convergence of Iterative Algorithms

Here are some comments on the convergence guarantees. With the explicit geometry of the objective function, it is straightforward to provide convergence guarantees using many popular iterative algorithms, even without special initializations. A few examples of recent progress on related nonconvex problems are listed as follows.

- A trust-region type of algorithm is proposed in [214] to solve a specific type of nonconvex problem, i.,e., phase retrieval. Similar to our analysis, the authors explicitly divide the whole domain into three overlapping regions $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$, based on which they show a sufficient decrease of objective in $\mathcal{R}_1$ and $\mathcal{R}_3$ and an overall R-quadratic convergence to a global minimum. Another closely related algorithm is the second-order majorization type of algorithm proposed in [226], which finds an ε-second-order stationary point x_ε for a predefined precision $\varepsilon > 0$, i.e.,

  $$\|\nabla f(x_\varepsilon)\|_2 \leq \varepsilon \quad \text{and} \quad \nabla^2 f(x_\varepsilon) \succeq -\sqrt{\beta\varepsilon} I$$

  for general lower bounded objective f that has a Lipschitz gradient and a 2β-Lipschitz Hessian. The algorithm is based on iteratively solving a cubic-regularized quadratic approximation of the objective function using gradient descent steps, and an overall sublinear convergence guarantee is provided.

- A gradient descent algorithm is analyzed in [215,216] for twice-continuously differentiable functions with a Lipschitz gradient. The authors provide an asymptotic convergence guarantee of Q-linear convergence to a local minimum if all saddle points are strict saddle.

- A noisy stochastic gradient descent algorithm is proposed in [187] for so-called strict saddle problems, i.e., any point the given objective function is in $\mathcal{R}_1$ (negative curvature in neighborhood of strict saddle points), $\mathcal{R}_3$ (the gradient has a sufficiently large norm), or a strongly convex neighborhood containing a local minimum. The authors show a sufficient decrease of objective for each noisy stochastic gradient step in $\mathcal{R}_1$ and $\mathcal{R}_3$, and an overall R-sublinear convergence to a local minimum.

The algorithms discussed above can be extended to solve the matrix factorization type of problems considered in this paper, with convergence guarantees. Note that for those requiring a local strong convexity, such as [187], the analysis does not apply directly here for the matrix factorization type of problems in general. This can be settled by applying the Polyak-Lojasiewicz condition instead [227,228].

6.5.3 Extension to Matrix Completion

Finally, we comment on a closely related problem – matrix completion, where we expect similar global geometric properties to hold. Specifically, given a entry-wise observed matrix $\mathcal{P}_\Omega(M^*) \in \mathbb{R}^{n\times n}$ for $M^* \succeq 0$, where $\mathcal{P}_\Omega(M^*_{i,j}) = 0$ if $(i,j) \notin \Omega$ and $\mathcal{P}_\Omega(M^*_{i,j}) = M^*_{i,j}$ if $(i,j) \in \Omega$ for some subset $\Omega \subseteq [n] \times [n]$, we solve

$$\min_{X\in\mathbb{R}^{n\times r}} H(X) + R(X), \text{ where } H(X) = \frac{1}{p}\|\mathcal{P}_\Omega(M^* - XX^\top)\|_{\mathrm{F}}^2. \tag{6.20}$$

where $p = |\Omega|/n^2$ is the sampling rate and $R(X)$ is a regularization function to enforce low coherence of X (see more details in [197, 199]). Similar to the matrix sensing problem, (6.20) can be also considered as a perturbed version of the low-rank matrix factorization problem (6.4). It is easy to see that if Ω is uniformly sampled over all subsets of $[n] \times [n]$ for a given cardinality, then we have

$$\mathbb{E}(H(X)) = \|M^* - XX^\top\|_{\mathrm{F}}^2.$$

However, because the entry-wise sampling model is more challenging than the random linear measurement model and the incoherence of the low-rank matrix is generally required, the extra regularization term is inevitable for the matrix completion problem. This leads to a much more involved perturbation analysis for (6.20) than that of matrix sensing. For example, [197] establish the geometric analysis around the global minimizers; [199] show that there exists no spurious local optima.

Chapter 7

Proofs for All Analyses

7.1 Proofs for Chapter 2

7.1.1 Proof of Theorem 1

First, we note that in both of the steps of Algorithm 1 the prescribed observations are functions of M only through ΦM; stated another way, M never appears in the algorithm in isolation from the measurement matrix Φ. Motivated by this, we introduce

$$\widetilde{M} \triangleq \Phi M = \Phi L + \Phi C = \widetilde{L} + \widetilde{C}, \tag{7.1}$$

to effectively subsume the action of Φ into $\widetilde{M}$. Now, our proof is a straightforward consequence of assembling three intermediate probabilistic results via a union bounding argument. The first intermediate result establishes that for $M = L + C$ with components L and C satisfying the structural conditions **(*c*1)**-**(*c*4)**, the components $\widetilde{L}$ and $\widetilde{C}$ of $\widetilde{M}$ as defined in (7.1) satisfy analogous structural conditions provided that m, the number of rows of Φ, be sufficiently large. We state this result here as a lemma; its proof appears in Appendix 7.1.2.

Lemma 6. Suppose $M = L + C$, where L and C satisfy the structural conditions (**c1**)-(**c4**). Fix any $\delta \in (0, 1)$, suppose Φ is an $m \times n_1$ matrix drawn from a distribution satisfying the distributional JL property (2.2) with m satisfying (2.5) and let $\widetilde{M} = \widetilde{L} + \widetilde{C}$ be as defined in (7.1). Then, the components $\widetilde{L}$ and $\widetilde{C}$ satisfy the following conditions simultaneously with probability at least $1-\delta$:

($\widetilde{\boldsymbol{c}}$1) $\mathrm{rank}(\widetilde{L}) = r$,

($\widetilde{\boldsymbol{c}}$2) $\widetilde{L}$ has n_L nonzero columns,

($\widetilde{\boldsymbol{c}}$3) $\widetilde{L}$ satisfies the column incoherence property with parameter μ_L, and

($\widetilde{\boldsymbol{c}}$4) $\mathcal{I}_{\widetilde{C}} \triangleq \{i : \|P_{\widetilde{\mathcal{L}}^\perp} \widetilde{C}_{:,i}\|_2 > 0, \widetilde{L}_{:,i} = 0\} = \mathcal{I}_C$, where $\widetilde{\mathcal{L}}$ is the linear subspace of $\mathbb{R}^m$ spanned by the columns of $\widetilde{L}$, and $P_{\widetilde{\mathcal{L}}^\perp}$ denotes the orthogonal projection onto the orthogonal complement of $\widetilde{\mathcal{L}}$ in $\mathbb{R}^m$.

The second intermediate result guarantees two outcomes – first, that Step 1 of Algorithm 1 succeeds in identifying the correct column space of $\widetilde{\mathcal{L}}$ (i.e., that $\widehat{\mathcal{L}}_{(1)} = \widetilde{\mathcal{L}}$) with high probability provided the components $\widetilde{L}$ and $\widetilde{C}$ of $\widetilde{M}$ as specified in (7.1) satisfy the structural conditions **($\widetilde{\boldsymbol{c}}$1)**-**($\widetilde{\boldsymbol{c}}$4)** and the column sampling probability parameter γ be sufficiently large, and second, that the number of columns of the randomly generated sampling matrix S be close to γn_2. We also provide this result as a lemma; its proof appears in Appendix 7.1.3.

Lemma 7. Let $\widetilde{M} = \widetilde{L} + \widetilde{C}$ be an $m \times n_2$ matrix, where the components $\widetilde{L}$ and $\widetilde{C}$ satisfy the conditions **($\widetilde{\boldsymbol{c}}$1)**-**($\widetilde{\boldsymbol{c}}$4)** with k satisfying (2.3). Fix $\delta \in (0,1)$ and suppose the column sampling parameter γ satisfies (2.4). When $\lambda = \frac{3}{7\sqrt{k_{\mathrm{ub}}}}$ for any $k_{\mathrm{ub}} \geq |\mathcal{I}_{\widetilde{C}}|$, the following hold simultaneously with probability at least $1-\delta$: S has $|\mathcal{S}| \leq (3/2)\gamma n_2$ columns, and the subspace $\widehat{\mathcal{L}}_{(1)}$ resulting from Step 1 of Algorithm 1 satisfies $\widehat{\mathcal{L}}_{(1)} = \widetilde{\mathcal{L}}$.

Our third intermediate result shows that the support set of the vector $\widehat{\mathbf{c}}$ produced in Step 2 of Algorithm 1 is the same as the set of salient columns of $\widetilde{C}$, provided that $\widehat{\mathcal{L}}_{(1)} = \widetilde{\mathcal{L}}$ and that p, the number of rows of A, is sufficiently large. We state this result here as a lemma; its proof appears in Appendix 7.1.4.

Lemma 8. $\widetilde{M} = \widetilde{L} + \widetilde{C}$ be an $m \times n_2$ matrix, where the components $\widetilde{L}$ and $\widetilde{C}$ satisfy the conditions **($\widetilde{\boldsymbol{c}}$1)**-**($\widetilde{\boldsymbol{c}}$4)** for any $k \leq n_2$, and suppose $\widehat{\mathcal{L}}_{(1)} = \widetilde{\mathcal{L}}$, the subspace spanned by the columns of $\widetilde{L}$. Let $\Phi M = \widetilde{M}$ in Step 2 of Algorithm 1. Fix $\delta \in (0,1)$, suppose A is a $p \times n_2$ matrix drawn from a distribution satisfying the distributional JL property (2.2) with p satisfying (2.6), and suppose the elements of $\boldsymbol{\phi}$ are i.i.d. realizations of any continuous random variable. Then with probability at least $1 - \delta$ the support $\mathcal{I}_{\widehat{\mathbf{c}}} \triangleq \{i : \widehat{c}_i \neq 0\}$ of the vector $\widehat{\mathbf{c}}$ produced by Step 2 of Algorithm 1 satisfies $\mathcal{I}_{\widehat{\mathbf{c}}} = \mathcal{I}_{\widetilde{C}}$.

Our overall result follows from assembling these intermediate results via union bound. In the event that the conclusion of Lemma 6 holds, then so do the requisite conditions of Lemma 7. Thus, with probability at least $1 - 2\delta$ the conclusions of Lemmata 6 and 7 both hold. This implies that the requisite conditions of Lemma 8 hold also with probability at least $1 - 2\delta$, and so it follows that the conclusions of all three Lemmata hold with probability at least $1 - 3\delta$.

7.1.2 Proof of Lemma 6

We proceed using the formalism of *stable embeddings* that has emerged from the dimensionality reduction and compressive sensing literature (see, e.g., [229]).

Definition 21 (Stable Embedding)**.** For $\epsilon \in [0, 1]$ and $\mathcal{U}, \mathcal{V} \subseteq \mathbb{R}^n$, we say Φ is an *ϵ-stable embedding* of $(\mathcal{U}, \mathcal{V})$ if

$$(1 - \epsilon)\|u - v\|_2^2 \leq \|\Phi u - \Phi v\|_2^2 \leq (1 + \epsilon)\|u - v\|_2^2 \tag{7.2}$$

for all $u \in \mathcal{U}$ and $v \in \mathcal{V}$.

Our proof approach is comprised of two parts. First, we show that each of the four claims in the lemma follow when Φ is an ϵ-stable embedding of

$$(\mathcal{L}, \cup_{i \in \mathcal{I}_C} \{C_{:,i}\} \cup \{0\}) \tag{7.3}$$

for any choice of $\epsilon < 1/2$. Second, we show that for any $\delta \in (0, 1)$, generating Φ as a random matrix as specified in the lemma ensures it will be a $\sqrt{2}/4$-stable embedding of (7.3) with probability at least $1 - \delta$. The choice of $\sqrt{2}/4$ in the last step is somewhat arbitrary – we choose this fixed value for concreteness here, but note that the structural conclusions of the lemma follow using any choice of $\epsilon < 1/2$ (albeit with slightly different conditions on m).

Part 1

Throughout this portion of the proof we assume that Φ is an ϵ-stable embedding of (7.3) for some $\epsilon < 1/2$, and establish each of the four claims in turn. First, to establish that $\text{rank}(\Phi L) = r = \text{rank}(L)$, we utilize an intermediate result of [64], stated here as a lemma (without proof) and formulated in the language of stable embeddings.

Lemma 9 (Adapted from [64], Theorem 1)**.** Let L be an $n_1 \times n_2$ matrix of rank r, and let $\mathcal{L}$ denote the column space of L, which is an r-dimensional linear subspace of $\mathbb{R}^{n_1}$. If for some $\epsilon \in (0, 1)$, Φ is an ϵ-stable embedding of $(\mathcal{L}, \{0\})$ then $\text{rank}(\Phi L) = r = \text{rank}(L)$.

Here, since Φ being an ϵ-stable embedding of (7.3) implies it is also an ϵ-stable embedding of $(\mathcal{L}, \{0\})$, the first claim (of Lemma 6) follows from Lemma 9.

Next we show that ΦL has n_L nonzero columns. Since Φ is a stable embedding of $(\mathcal{L}, \{0\})$, it follows that for each of the n_L nonzero columns $L_{:,i}$ of L we have $\|\Phi L_{:,i}\|_2^2 > (1 - \epsilon)\|L_{:,i}\|_2^2 > 0$, while for each of the remaining $n_2 - n_L$ columns $L_{:,j}$ of L that are identically zero we have $\|\Phi L_{:,j}\|_2^2 = 0$ so that $\Phi L_{:,j} = 0$.

Continuing, we show next that ΦL satisfies the column incoherence property with parameter μ_L. Recall from above that we write the compact SVD of L as $L = U\Sigma V^*$, where U is $n_1 \times r$, V is $n_2 \times r$, and Σ is an $r \times r$ nonnegative diagonal matrix of singular values (all of which are strictly positive). The incoherence condition on L is stated in terms of column norms of the matrix V^* whose rows form an orthonormal basis for the row space of L. Now, when the rank of ΦL is the same as that of L, which is true here on account of Lemma 9, the row space of ΦL is *identical* to that of L, since each are r-dimensional subspaces of $\mathbb{R}^{n_2}$ spanned by linear combinations of the columns of the V^*. Thus since the rank and number of nonzero columns of ΦL are the same as for L, the coherence parameter of ΦL is just μ_L, and the third claim is established.

Finally, we establish the last claim, that the set of salient columns of ΦC is the same as for C. Recall that the condition that a column $C_{:,i}$ be salient was equivalent to the condition that $\|P_{\mathcal{L}^\perp} C_{:,i}\|_2 > 0$, where $P_{\mathcal{L}^\perp}$ is the orthogonal projection operator onto the orthogonal complement of $\mathcal{L}$ in $\mathbb{R}^{n_1}$. Here, our aim is to show that an analogous result holds in the "projected" space – that for all $i \in \mathcal{I}_C$ we have $\|P_{(\Phi\mathcal{L})^\perp} \Phi C_{:,i}\|_2 > 0$, where $\Phi\mathcal{L}$ is the linear subspace spanned by the columns of ΦL. For this we utilize an intermediate result of [229] formulated there in terms of a "compressive interference cancellation" method. We state an adapted version of that result here as a lemma (without proof).

Lemma 10 (Adapted from [229], Theorem 5)**.** Let $\mathcal{V}_1$ be an r-dimensional linear subspace of $\mathbb{R}^n$ with $r < n$, let $\mathcal{V}_2$ be any subset of $\mathbb{R}^n$, and let $\check{\mathcal{V}}_2 = \{P_{\mathcal{V}_1^\perp} v : v \in \mathcal{V}_2\}$, where $P_{\mathcal{V}_1^\perp}$ is the orthogonal projection operator onto the orthogonal complement of $\mathcal{V}_1$ in $\mathbb{R}^n$. If Φ is an ϵ-stable embedding of $(\mathcal{V}_1, \check{\mathcal{V}}_2 \cup \{0\})$, then for all $\check{v} \in \check{\mathcal{V}}_2$

$$\|P_{(\Phi\mathcal{V}_1)^\perp}(\Phi\check{v})\|_2^2 \geq \left(1 - \frac{\epsilon}{1-\epsilon}\right) \|\check{v}\|_2^2, \tag{7.4}$$

where $P_{(\Phi\mathcal{V}_1)^\perp}$ is the orthogonal projection operator onto the orthogonal complement of the subspace of $\mathbb{R}^n$ spanned by the elements of $\Phi\mathcal{V}_1 = \{\Phi v : v \in \mathcal{V}_1\}$.

Before applying this result we first note a useful fact, that Φ being an ϵ-stable embedding of $(\mathcal{V}_1, \check{\mathcal{V}}_2 \cup \{0\})$ is equivalent to Φ being an ϵ-stable embedding of $(\mathcal{V}_1, \mathcal{V}_2 \cup \{0\})$, which follows directly from the definition of stable embeddings and the (easy to verify) fact that $\left\{v_1 - \check{v}_2 : v_1 \in \mathcal{V}_1, \check{v}_2 \in \check{\mathcal{V}}_2 \cup \{0\}\right\} = \{v_1 - v_2 : v_1 \in \mathcal{V}_1, v_2 \in \mathcal{V}_2 \cup \{0\}\}$. Now, to apply Lemma 10 here, we let $\mathcal{V}_1 = \mathcal{L}$, $\mathcal{V}_2 = \cup_{i \in \mathcal{I}_C}\{C_{:,i}\}$, and $\check{\mathcal{V}}_2 = \cup_{i \in \mathcal{I}_C}\{P_{\mathcal{L}^\perp} C_{:,i}\}$. Since Φ is an ϵ-stable embedding of (7.3), we have that for all $i \in \mathcal{I}_{C_{:,i}}$, $\|P_{(\Phi\mathcal{L})^\perp}(\Phi C)_{:,i}\|_2^2 \geq \left(1 - \frac{\epsilon}{1-\epsilon}\right) \|P_{\mathcal{L}^\perp} C_{:,i}\|_2^2$. Since $\epsilon < 1/2$, the above result implies $\|P_{(\Phi\mathcal{L})^\perp}\Phi C_{:,i}\|_2 > 0$ for all $i \in \mathcal{I}_C$, while for all $j \notin \mathcal{I}_C$ we have $C_{:,j} = 0$, implying that $\Phi C_{:,j} = 0$ and hence $\|P_{(\Phi\mathcal{L})^\perp}\Phi C_{:,j}\|_2 = 0$. Using this, and the fact that the nonzero columns of ΦL coincide with the nonzero columns of L, we conclude that $\mathcal{I}_{\Phi C} = \{i : \|P_{(\Phi\mathcal{L})^\perp}\Phi C_{:,j}\|_2 > 0, (\Phi L)_{:,i} = 0\}$ is the same as $\mathcal{I}_C$.

Part 2

Given the structural result established in the previous step, the last part of the proof entails establishing that a random matrix Φ generated as specified in the statement of Lemma 6 is an $\sqrt{2}/4$-stable embedding of (7.3). Our approach here begins with a brief geometric discussion, and a bit of "stable embedding algebra." Appealing to the definition of stable embeddings, we see that Φ being an ϵ-stable embedding of (7.3) is equivalent to Φ being such that

$$(1-\epsilon)\|v\|_2^2 \leq \|\Phi v\|_2^2 \leq (1+\epsilon)\|v\|_2^2 \tag{7.5}$$

holds for all $v \in \mathcal{L} \cup \bigcup_{i \in \mathcal{I}_C} \mathcal{L} - C_{:,i}$, where $\mathcal{L} - C_{:,i}$ denotes the r-dimensional *affine* subspace of $\mathbb{R}^{n_1}$ comprised of all elements taking the form of a vector in $\mathcal{L}$ minus the fixed vector $C_{:,i}$. Thus, in words, establishing our claim here entails showing that a random Φ (generated as specified in the lemma, with appropriate dimensions) approximately preserves the lengths of all vectors in a *union of subspaces* comprised of one r-dimensional linear subspace and some $|\mathcal{I}_C| = k$, r-dimensional affine subspaces.

Stable embeddings of linear subspaces using random matrices is, by now, well-studied (see, e.g., [64, 229, 230], as well as a slightly weaker result [231, Lemma 10]), though stable embeddings of *affine* subspaces has received less attention in the literature. Fortunately, using a straightforward argument we may leverage results for the former in order to establish the latter. Recall the discussion above, and suppose that rather than establishing that (7.5) holds for all $v \in \mathcal{L} \cup \bigcup_{i \in \mathcal{I}_C} \mathcal{L} - C_{:,i}$ we instead establish a slightly stronger result, that (7.5) holds for all $v \in \mathcal{L} \cup \bigcup_{i \in \mathcal{I}_C} \mathcal{L}^i$, where for each $i \in \mathcal{I}_C$, $\mathcal{L}^i$ denotes the $(r+1)$-dimensional *linear* subspace of $\mathbb{R}^{n_1}$ spanned by the columns of the matrix $[L\ C_{:,i}]$. (That the dimension of each $\mathcal{L}^i$ be $r+1$ follows from the assumption that columns $C_{:,i}$ for $i \in \mathcal{I}_C$ be outliers.) Clearly, if for some $i \in \mathcal{I}_C$ the condition (7.5) holds for all $v \in \mathcal{L}^i$, then it holds for all vectors formed as linear combinations of $[L\ C_{:,i}]$, so it holds in particular for all vectors in the r dimensional affine subspace denoted by $\mathcal{L} - C_{:,i}$. Further, that (7.5) holds for all $v \in \mathcal{L}^i$ for any $i \in \mathcal{I}_C$ implies it holds for linear combinations that use a weight of zero on the component $C_{:,i}$, so in this case (7.5) holds also for all $v \in \mathcal{L}$.

Based on the above discussion, we see that a sufficient condition to establish that Φ be an ϵ-stable embedding of (7.3) is that (7.5) hold for all $v \in \bigcup_{i \in \mathcal{I}_C} \mathcal{L}^i$; in other words, that Φ preserve (up to multiplicative $(1 \pm \epsilon)$ factors) the squared lengths of all vectors in a union of (up to) k unique $(r+1)$-dimensional linear subspaces of $\mathbb{R}^{n_1}$. To this end we make use of another result adapted from [64], and based on the union of subspaces embedding approach utilized in [230].

Lemma 11 (Adapted from [64], Lemma 1)**.** Let $\bigcup_{i=1}^{k} \mathcal{V}^i$ denote a union of k linear subspaces of $\mathbb{R}^n$, each of dimension at most d. For fixed $\epsilon \in (0,1)$ and $\delta \in (0,1)$, suppose Φ is an $m \times n$ matrix satisfying the distributional JL property with

$$m \geq \frac{d \log(42/\epsilon) + \log(k) + \log(2/\delta)}{f(\epsilon/\sqrt{2})} \tag{7.6}$$

Then $(1-\epsilon)\|v\|_2^2 \leq \|\Phi v\|_2^2 \leq (1+\epsilon)\|v\|_2^2$ holds simultaneously for all $v \in \bigcup_{i=1}^{k} \mathcal{V}^i$ with probability at least $1-\delta$.

Applying this lemma here with $d = r+1$ and $\epsilon = \sqrt{2}/4$, and using the fact that $\log(84\sqrt{2}) < 5$ yields the final result.

7.1.3 Proof of Lemma 7

Our approach is comprised of two parts. In the first, we show that the two claims of Lemma 7 follow directly when the following five conditions are satisfied

(**a1**) S has $(1/2)\gamma n_2 \leq |\mathcal{S}| \leq (3/2)\gamma n_2$ columns,

(**a2**) $\widetilde{L}S$ has at most $(3/2)\gamma n_L$ nonzero columns,

(**a3**) $\widetilde{C}S$ has at most $(3/2)\gamma k$ nonzero columns,

(**a4**) $\sigma_1^2(\widetilde{V}^* S) \leq (3/2)\gamma$, and

(**a5**) $\sigma_r^2(\widetilde{V}^* S) \geq (1/2)\gamma$,

where the matrix $\widetilde{V}^*$ that arises in (**a4**)-(**a5**) is the matrix of right singular vectors from the compact SVD $\widetilde{L} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^*$ of $\widetilde{L}$, and $\sigma_i(\widetilde{V}^* S)$ denotes the i-th largest singular value of $\widetilde{V}^* S$. Then, in the second part of the proof we show that (**a1**)-(**a5**) hold with high probability when S is a random subsampling matrix generated with parameter γ in the specified range.

We briefly note that parameters $(1/2)$ and $(3/2)$ arising in the conditions (**a1**)-(**a5**) are somewhat arbitrary, and are fixed to these values here for ease of exposition. Analogous results to that of Lemma 7 could be established by replacing $(1/2)$ with any constant in $(0,1)$ and $(3/2)$ with any constant larger than 1, albeit with slightly different conditions on γ.

Part 1

Throughout this portion of the proof, we assume that conditions (**a1**)-(**a5**) hold. Central to our analysis is a main result of [19], which we state as a lemma (without proof).

Lemma 12 (Outlier Pursuit, adapted from [19])**.** Let $\check{M} = \check{L} + \check{C}$ be an $\check{n}_1 \times \check{n}_2$ matrix whose components $\check{L}$ and $\check{C}$ satisfy the structural conditions

(**č1**) $\operatorname{rank}(\check{L}) = \check{r}$,

(**č2**) $\check{L}$ has $n_{\check{L}}$ nonzero columns,

(**č3**) $\check{L}$ satisfies the *column incoherence property* with parameter $\mu_{\check{L}}$, and

(**č4**) $|\mathcal{I}_{\check{C}}| = \{i : \|P_{\check{\mathcal{L}}^\perp} \check{C}_{:,i}\|_2 > 0, \check{L}_{:,i} = 0\} = \check{k}$, where $\check{\mathcal{L}}$ denotes the linear subspace spanned by columns of $\check{L}$ and $P_{\check{\mathcal{L}}^\perp}$ is the orthogonal projection operator onto the orthogonal complement of $\check{\mathcal{L}}$ in $\mathbb{R}^{\check{n}_1}$,

with

$$\check{k} \leq \left(\frac{1}{1 + (121/9)\ \check{r}\mu_{\check{L}}} \right) \check{n}_2. \tag{7.7}$$

For any upper bound $\check{k}_{\text{ub}} \geq \check{k}$ and $\lambda = \frac{3}{7\sqrt{\check{k}_{\text{ub}}}}$ any solutions of the *outlier pursuit* procedure

$$\{\widehat{\check{L}}, \widehat{\check{C}}\} = \operatorname*{argmin}_{L_{(1)}, C_{(1)}} \|L_{(1)}\|_* + \lambda \|C_{(1)}\|_{1,2} \text{ s.t. } \check{M} = L_{(1)} + C_{(1)}, \tag{7.8}$$

are such that the columns of $\widehat{\check{L}}$ span the same linear subspace as the columns of $\check{L}$, and the set of nonzero columns of $\widehat{\check{C}}$ is the same as the set of locations of the nonzero columns of $\check{C}$.

Introducing the shorthand notation $\check{L} = \widetilde{L}S$, $\check{C} = \widetilde{C}S$, and $\check{n}_2 = |\mathcal{S}|$, our approach will be to show that conditions (**a1**)-(**a5**) along with the assumptions on $\widetilde{M}$ ensure that (**č1**)-(**č4**) in Lemma 12 are satisfied for some appropriate parameters $\check{r}$, $n_{\check{L}}$, $\mu_{\check{L}}$, and $\check{k}$ that depend on analogous parameters of $\widetilde{M}$.

First, note that (**a5**) implies that the matrix $\widetilde{V}^* S$ has rank r, which in turn implies that $\check{L}$ has rank r. Thus, (**č1**) is satisfied with $\check{r} = r$. The condition (**č2**) is also satisfied here for $n_{\check{L}}$ no larger than $(3/2)\gamma n_L$; this is a restatement of (**a2**).

We next establish (**č3**). To this end, note that since $\check{L}$ has rank r, it follows that the r-dimensional linear subspace spanned by the rows of $\check{L} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^* S$ is the same as that spanned by the rows of $\widetilde{V}^* S$. Now, let $S^T \widetilde{\mathcal{V}}$ denote the r-dimensional linear subspace of $\mathbb{R}^{\check{n}_2}$ spanned by the columns of $S^T\widetilde{V}$ and let $P_{S^T\widetilde{\mathcal{V}}}$ denote the orthogonal projection operator onto $S^T\widetilde{\mathcal{V}}$. Then, bounding the column incoherence parameter of $\check{L}$ entails establishing an upper bound on $\max_{i\in[\check{n}_2]}\|P_{S^T\widetilde{\mathcal{V}}}e_i\|_2^2$, where e_i is the i-th canonical basis vector of $\mathbb{R}^{\check{n}_2}$. Directly constructing the orthogonal projection operator (and using that $\widetilde{V}^*S$ is a rank r matrix) we have that

$$\begin{aligned}
\max_{i\in[\check{n}_2]}\|P_{S^T\widetilde{\mathcal{V}}}e_i\|_2^2 &= \max_{i\in[\check{n}_2]}\left\|S^T\widetilde{V}\left(\widetilde{V}^*SS^T\widetilde{V}\right)^{-1}\widetilde{V}^*Se_i\right\|_2^2\\
&\overset{(a)}{\leq} \max_{j\in[n_2]}\left\|S^T\widetilde{V}\left(\widetilde{V}^*SS^T\widetilde{V}\right)^{-1}\widetilde{V}^*e_j\right\|_2^2\\
&\overset{(b)}{\leq} \left(\frac{\sigma_1(\widetilde{V}^*S)}{\sigma_r^2(\widetilde{V}^*S)}\right)^2\mu_L\frac{r}{n_L}\overset{(c)}{\leq}\left(\frac{6}{\gamma}\right)\mu_L\frac{r}{n_L},
\end{aligned}$$

where (a) follows from the fact that for any $i\in[\check{n}_2]$ the vector Se_j is either the zero vector or one of the canonical basis vectors for $\mathbb{R}^{n_2}$, (b) follows from straightforward linear algebraic bounding ideas and the column incoherence assumption on $\widetilde{L}$, and (c) follows from (**a4**)-(**a5**). Now, we let $n_{\check{L}}$ denote the number of nonzero columns of $\check{L}$, and write

$$\max_{i\in[\check{n}_2]}\|P_{S^T\widetilde{\mathcal{V}}}e_i\|_2^2\leq\left(\frac{6}{\gamma}\right)\mu_L\frac{r}{n_L}\left(\frac{n_{\check{L}}}{n_{\check{L}}}\right)\leq 9\mu_L\frac{r}{n_{\check{L}}}, \tag{7.9}$$

where the last inequality uses (**a2**). Thus (**č3**) holds with

$$\mu_{\check{L}} = 9\mu_L. \tag{7.10}$$

Next, we establish (**č4**). Recall from above that $\check{L}$ has rank r, and is comprised of columns of $\widetilde{L}$; it follows that the subspace $\check{\mathcal{L}}$ spanned by columns of $\check{L}$ is the same as the subspace $\widetilde{\mathcal{L}}$ spanned by columns of $\widetilde{L}$. Thus, $\|P_{\check{\mathcal{L}}^\perp}\check{C}_{:,i}\|_2 = \|P_{\widetilde{\mathcal{L}}^\perp}\check{C}_{:,i}\|_2$, so to obtain an upper bound on $\check{k}$ we can simply count the number $\check{k}$ of nonzero columns of $\check{C}=\widetilde{C}S$. By (**a1**), (**a3**), (2.3), and the fact that $\check{r}=r$, we have

$$\check{k}\leq\frac{3}{2}\gamma k_{\mathrm{u}}\leq\frac{3}{2}\cdot\frac{2\check{n}_2}{n_2}k_{\mathrm{u}}=\frac{3k_{\mathrm{u}}\check{n}_2}{n_2},$$

which, combined with (7.10), implies

$$\check{k} \leq \left(\frac{1}{1 + (121/9)\ \check{r}\mu_{\check{L}}} \right) \check{n}_2 = \frac{3k_{\mathrm{u}}\check{n}_2}{n_2},$$

Finally, we show that the two claims of Lemma 7 hold. The first follows directly from (**a1**). For the second, note that for any $k_{\mathrm{ub}} \geq k$ we have that $\check{k}_{\mathrm{ub}} \triangleq k_{\mathrm{ub}} \geq \check{k}$. Thus, since $\lambda = \frac{3}{7\sqrt{k_{\mathrm{ub}}}} = \frac{3}{7\sqrt{\check{k}_{\mathrm{ub}}}}$ and (**č1**)-(**č4**) hold, it follows from Lemma 12 that the optimization (7.8) produces an estimate $\widehat{\check{L}}$ whose columns span the same linear subspace as that of $\check{L}$. But, since $\check{L}$ has rank r and its columns are just a subset of columns of the rank-r matrix $\widetilde{L}$, the subspace spanned by the columns of $\check{L}$ is the same as that spanned by columns of $\widetilde{L}$.

Part 2

The last part of our proof entails showing (**a1**)-(**a5**) hold with high probability when S is randomly generated as specified. Let $\mathcal{E}_1, \ldots, \mathcal{E}_5$ denote the events that conditions (**a1**)-(**a5**), respectively, hold. Then $\Pr\left(\left\{ \bigcap_{i=1}^{5} \mathcal{E}_i \right\}^c \right) \leq \sum_{i=1}^{5} \Pr(\mathcal{E}_i^c)$, and we consider each term in the sum in turn.

First, since $|\mathcal{S}|$ is a Binomial(n_2, γ) random variable, we may bound its tails using [232, Theorem 2.3 (b-c)]. This gives that $\Pr\left(|\mathcal{S}| > 3\gamma n_2/2\right) \leq \exp\left(-3\gamma n_2/28\right)$ and $\Pr\left(|\mathcal{S}| < \gamma n_2/2\right) \leq \exp\left(-\gamma n_2/8\right)$. By union bound, we obtain that $\Pr(\mathcal{E}_1^c) \leq \exp\left(-3\gamma n_2/28\right) + \exp\left(-\gamma n_2/8\right)$.

Next, observe that conditionally on $|\mathcal{S}| = s$, the number of nonzero columns present in the matrix $\widetilde{L}S$ is a hypergeometric random variable parameterized by a population of size n_2 with n_L positive elements and s draws. Denoting this hypergeometric distribution here by hyp(n_2, n_L, s) and letting $H_{|\mathcal{S}|} \sim$ hyp$(n_2, n_L, |\mathcal{S}|)$, we have that $\Pr(\mathcal{E}_2^c) = \Pr\left(H_{|\mathcal{S}|} > \left(\frac{3}{2}\right)\gamma n_L\right)$. Using a simple conditioning argument, $\Pr(\mathcal{E}_2^c) \leq \sum_{s=\lceil (2/3)\gamma n_2 \rceil}^{\lfloor (4/3)\gamma n_2 \rfloor} \Pr\left(H_s > \left(\frac{3}{2}\right)\gamma n_L\right) \Pr(|\mathcal{S}| = s) + \Pr\left(||\mathcal{S}| - \gamma n_2| > \left(\frac{1}{3}\right)\gamma n_2\right)$, and our next step is to simplify the terms in the sum. Note that for any s in the range of summation, we have $\Pr\left(H_s > \left(\frac{3}{2}\right)\gamma n_L\right) = \Pr\left(H_s > \left(\frac{3}{2}\right)\gamma n_L\left(\frac{sn_2}{sn_2}\right)\right)$, and thus

$$\Pr\left(H_s > \left(\frac{3}{2}\right)\gamma n_L\right) \overset{(a)}{\leq} \Pr\left(H_s > \left(\frac{9}{8}\right) s \left(\frac{n_L}{n_2}\right)\right) \overset{(b)}{\leq} \exp\left(-\frac{3s(n_L/n_2)}{400}\right)$$
$$\overset{(c)}{\leq} \exp\left(-\frac{\gamma n_L}{200}\right),$$

where (a) utilizes the largest value of s to bound the term $\gamma n_2/s$, (b) follows from an application of Lemma 15 in Appendix 7.1.5, and (c) results from using the smallest value of s (within the range of summation) to bound the error term. Assembling these results, we have that $\Pr(\mathcal{E}_2^c) \leq \exp(-\gamma n_L/200) + \exp(-\gamma n_2/24) + \exp(-\gamma n_2/18)$, where we use the fact that the probability mass function of $|\mathcal{S}|$ sums to one, and another application of [232, Theorem 2.3(b,c)].

To bound $\Pr(\mathcal{E}_3^c)$, we discuss the following two cases: **Case 1**. By construction, $\check{k}$ is a Hypergeometric random variable, parameterized by the population size n_2, the total number of draws $\check{n}_2$ and the total positive elements k, denoted here by $\text{Hyp}(n_2, \check{n}_2, k)$. Then we have that

$$\Pr(\mathcal{E}_3^c) = \Pr\left(\check{k} > \frac{3}{2}\gamma k_{\text{u}}\right) \leq \Pr\left(\check{k} > \frac{3}{2}\gamma k\right) \leq \exp\left(-\frac{\gamma k}{200}\right).$$

When γ satisfies (2.4), we have $\Pr(\mathcal{E}_3^c) < \frac{\delta}{6}$ provided k satisfies

$$\frac{200}{\gamma}\log\left(\frac{6}{\delta}\right) \leq k \leq k_{\text{u}} = \frac{n_2}{3(1 + 121 r\mu_{\mathbf{V}})}.$$

Case 2. Now consider $k < \frac{200}{\gamma}\log(\frac{6}{\delta})$. Let $\check{k}_1$ and $\check{k}_2$ be Hypergeometric random variables with distributions $\text{Hyp}(n_2, \check{n}_2, k_1)$ and $\text{Hyp}(n_2, \check{n}_2, k_2)$ respectively, where $k_1 > k_2$. Our analysis relied upon a stochastic ordering property of Hypergeometric random variables; we establish that result here as a lemma.

Lemma 13 (Adapted from Theorem 1 of [233] for Hypergeometric distribution)**.** Let $X_1 \sim \text{Hyp}(n_2, \check{n}_2, k_1)$ and $X_2 \sim \text{Hyp}(n_2, \check{n}_2, k_2)$ be Hypergeometric random variables, whose distributions are parameterized by identical population n_2 and draws $\check{n}_2$ with k_1 and k_2 positive elements respectively, where $k_1 > k_2$. Then for any $x \in [0, \infty)$, we have

$$\Pr(X_2 \le x) \ge \Pr(X_1 \le x). \tag{7.11}$$

Proof of Lemma 13. Theorem 1 of [233] provides a general result of stochastic ordering for Hypergeometric distributions. Specifically, for $X_1 \sim \mathrm{Hyp}(n_2, \check{n}_2, k_1)$ and $X_2 \sim \mathrm{Hyp}(n_2, \check{n}_2, k_2)$, (7.11) holds for any $x \in [0, \infty)$ if and only if the left tail condition

$$\binom{n_2 - \check{n}_2}{k_2 - k_*}\binom{n_2}{k_1} \ge \binom{n_2 - \check{n}_2}{k_1 - k_*}\binom{n_2}{k_2} \tag{7.12}$$

and the right tail condition

$$\binom{n_2 - \check{n}_2}{k_2 - k^*}\binom{n_2}{k_1} \le \binom{n_2 - \check{n}_2}{k_1 - k^*}\binom{n_2}{k_2} \tag{7.13}$$

hold simultaneously, where k_* and k^* are the minimum and maximum supports of $\mathrm{Hyp}(n_2, \check{n}_2, k_1)$ and $\mathrm{Hyp}(n_2, \check{n}_2, k_2)$ respectively, defined as

$$\begin{aligned} k_* &= \min\{k : (\mathrm{Hyp}(n_2, \check{n}_2, k_1) + \mathrm{Hyp}(n_2, \check{n}_2, k_2))(k) > 0\} \\ &= \min\{(\check{n}_2 - (n_2 - k_1))_+, (\check{n}_2 - (n_2 - k_2))_+\}, \\ k^* &= \max\{k : (\mathrm{Hyp}(n_2, \check{n}_2, k_1) + \mathrm{Hyp}(n_2, \check{n}_2, k_2))(k) > 0\} \\ &= \max\{\min\{\check{n}_2, k_1\}, \min\{\check{n}_2, k_2\}\}. \end{aligned}$$

Here we only need to verify that if $k_1 > k_2$, then (7.12) and (7.13) hold simultaneously, which then implies (7.11). Note that all arguments below are for the case when x is a non-negative integer. But same results hold for any real x because a Hypergeometric random variable is discrete and for any non-negative integer y and a real $x \in [y, y+1)$, a Hypergeometric random variable X satisfies $\Pr(X \le x) = \Pr(X \le y)$.

We first verify (7.12) when $k_1 > k_2$. Let $(x)_+ = \max\{x, 0\}$. By definition we have,

$$\begin{aligned} k_* &= \min\{(\check{n}_2 - (n_2 - k_1))_+, (\check{n}_2 - (n_2 - k_2))_+\} \\ &= \begin{cases} \check{n}_2 + k_2 - n_2, & \text{if } \check{n}_2 + k_2 > n_2 \\ 0, & \text{o.w.} \end{cases}. \end{aligned}$$

When $k_* = \check{n}_2 + k_2 - n_2$, we have

$$\begin{pmatrix} n_2 - \check{n}_2 \\ k_2 - k_* \end{pmatrix} \begin{pmatrix} n_2 \\ k_1 \end{pmatrix} = \begin{pmatrix} n_2 - \check{n}_2 \\ n_2 - \check{n}_2 \end{pmatrix} \begin{pmatrix} n_2 \\ k_1 \end{pmatrix}$$

$$= \begin{pmatrix} n_2 \\ k_1 \end{pmatrix} > 1 > \begin{pmatrix} n_2 - \check{n}_2 \\ n_2 - \check{n}_2 + k_1 - k_2 \end{pmatrix} \begin{pmatrix} n_2 \\ k_2 \end{pmatrix} = 0,$$

where the last equality holds since $n_2 - \check{n}_2 < n_2 - \check{n}_2 + k_1 - k_2$ and x choose y is 0, if $x < y$. When $k_* = 0$, we have

$$\frac{\begin{pmatrix} n_2 - \check{n}_2 \\ k_2 - k_* \end{pmatrix} \begin{pmatrix} n_2 \\ k_1 \end{pmatrix}}{\begin{pmatrix} n_2 - \check{n}_2 \\ k_1 - k_* \end{pmatrix} \begin{pmatrix} n_2 \\ k_2 \end{pmatrix}} = \frac{\begin{pmatrix} n_2 - \check{n}_2 \\ k_2 \end{pmatrix} \begin{pmatrix} n_2 \\ k_1 \end{pmatrix}}{\begin{pmatrix} n_2 - \check{n}_2 \\ k_1 \end{pmatrix} \begin{pmatrix} n_2 \\ k_2 \end{pmatrix}}$$

$$= \frac{(n_2 - \check{n}_2) \cdots (n_2 - \check{n}_2 - k_2 + 1)}{(n_2 - \check{n}_2) \cdots (n_2 - \check{n}_2 - k_1 + 1)} \times \frac{n_2 \cdots \ (n_2 - k_1 + 1)}{n_2 \cdots \ (n_2 - k_2 + 1)}$$

$$= \frac{(n_2 - k_2)}{(n_2 - \check{n}_2 - k_2)} \times \cdots \times \frac{(n_2 - k_1 + 1)}{(n_2 - \check{n}_2 - k_1 + 1)} > 1$$

Therefore, (7.12) holds by combining the two scenarios.

To verify (7.13), we use analogous arguments. Specifically, we have from definition,

$$k^* = \max\{\min\{\check{n}_2, k_1\}, \min\{\check{n}_2, k_2\}\} = \begin{cases} k_1, & \text{if } k_1 \leq \check{n}_2 \\ \check{n}_2, & \text{o.w.} \end{cases}.$$

When $k^* = k_1$, we have

$$\begin{pmatrix} n_2 - \check{n}_2 \\ k_1 - k^* \end{pmatrix} \begin{pmatrix} n_2 \\ k_2 \end{pmatrix} = \begin{pmatrix} n_2 - \check{n}_2 \\ k_1 - k_1 \end{pmatrix} \begin{pmatrix} n_2 \\ k_2 \end{pmatrix}$$

$$= \begin{pmatrix} n_2 \\ k_2 \end{pmatrix} > 1 > \begin{pmatrix} n_2 - \check{n}_2 \\ k_2 - k_1 \end{pmatrix} \begin{pmatrix} n_2 \\ k_1 \end{pmatrix} = 0,$$

where the last equality holds since $k_1 > k_2$ and x choose y is 0, if $y < 0$.

When $k^* = \check{n}_2$, we have two different cases: (i) $k_2 < \check{n}_2 < k_1$ and (ii) $\check{n}_2 \leq k_2 < k_1$. In Case (i), we have similar argument as when $k^* = k_1$. In Case (ii), we have

$$
\begin{aligned}
&\frac{\begin{pmatrix} n_2-\check{n}_2 \\ k_1-k^* \end{pmatrix}\begin{pmatrix} n_2 \\ k_2 \end{pmatrix}}{\begin{pmatrix} n_2-\check{n}_2 \\ k_2-k^* \end{pmatrix}\begin{pmatrix} n_2 \\ k_1 \end{pmatrix}} = \frac{\begin{pmatrix} n_2-\check{n}_2 \\ k_1-\check{n}_2 \end{pmatrix}\begin{pmatrix} n_2 \\ k_2 \end{pmatrix}}{\begin{pmatrix} n_2-\check{n}_2 \\ k_2-\check{n}_2 \end{pmatrix}\begin{pmatrix} n_2 \\ k_1 \end{pmatrix}} \\
&= \frac{(n_2-\check{n}_2)\cdots(n_2-\check{n}_2-(k_1-\check{n}_2)+1)}{(n_2-\check{n}_2)\cdots(n_2-\check{n}_2-(k_2-\check{n}_2)+1)} \times \frac{n_2\cdots(n_2-k_2+1)}{n_2\cdots(n_2-k_1+1)} \\
&= \frac{(n_2-k_2)\cdots(n_2-k_1+1)}{(n_2-\check{n}_2-(k_2-\check{n}_2))\cdots(n_2-\check{n}_2-(k_2-\check{n}_2)+1)} \\
&= \frac{(n_2-k_2)\cdots(n_2-k_1+1)}{(n_2-k_2)\cdots(n_2-k_1+1)} = 1.
\end{aligned}
$$

Therefore, (7.13) holds by combining the two scenarios.

Using the general result, since (7.12) and (7.13) hold, we have $\Pr(X_2 \leq x) \geq \Pr(X_1 \leq x)$ for any $x \in [0, \infty)$. □

Then by Lemma 13, which is based on the stochastic ordering ideas from [233], we have $\Pr(\check{k}_2 \leq x) \geq \Pr(\check{k}_1 \leq x)$ for any $x \in [0, \infty)$. This, coupled with the analysis from Case 1 above, implies that $\Pr(\mathcal{E}_3^c) < \frac{\delta}{6}$ in this case as well.

Finally, we can obtain bounds on the largest and smallest singular values of $\widetilde{V}^* S$ using the Matrix Chernoff inequalities of [234]. Namely, letting $Z = \widetilde{V}^* S$ we note that the matrix ZZ^* may be expressed as a sum of independent positive semidefinite rank-one $r \times r$ Hermitian matrices, as $ZZ^* = \widetilde{V}^* S S^T \widetilde{V} = \sum_{i=1}^{n_2} s_i (\widetilde{V}^*_{:,i})(\widetilde{V}^*_{:,i})^*$, where the $\{s_i\}_{i=1}^{n_2}$ are i.i.d. Bernoulli(γ) random variables as in the statement of Algorithm 1 (and, $s_i^2 = s_i$). To instantiate the result of [234], we note that $\lambda_{\max}(s_i(\widetilde{V}^*_{:,i})(\widetilde{V}^*_{:,i})^*) \leq \|\widetilde{V}^*_{:,i}\|_2^2 \leq \mu_L r/n_L \triangleq R$ almost surely for all i, where the last inequality follows from the incoherence assumption ($\widetilde{\mathbf{c}}\mathbf{3}$) (as well as ($\widetilde{\mathbf{c}}\mathbf{1}$)-($\widetilde{\mathbf{c}}\mathbf{2}$)). Further, direct calculation yields $\mu_{\min} \triangleq \lambda_{\min}(\mathbb{E}[ZZ^*]) = \lambda_{\min}(\gamma I) = \gamma$ and $\mu_{\max} \triangleq \lambda_{\max}(\mathbb{E}[ZZ^*]) = \lambda_{\max}(\gamma I) = \gamma$, where the identity matrices in each case are of size $r \times r$. Thus, applying [234, Corollary 5.2] (with $\delta = 1/2$ in that formulation) we obtain that $\Pr(\mathcal{E}_4^c) = \Pr\left(\sigma_1^2\left(\widetilde{V}^* S\right) \geq 3\gamma/2\right) \leq r \cdot (9/10)^{\frac{\gamma n_L}{r \mu_L}}$, and $\Pr(\mathcal{E}_5^c) = \Pr\left(\sigma_r^2\left(\widetilde{V}^* S\right) \leq \gamma/2\right) \leq r \cdot (9/10)^{\frac{\gamma n_L}{r \mu_L}}$.

Putting the results together, and using a further bound on $\Pr(\mathcal{E}_1^c)$, we have $\Pr\left(\left\{\bigcap_{i=1}^5 \mathcal{E}_i\right\}^c\right) \leq \exp\left(-\frac{\gamma n_L}{200}\right) + 2\exp\left(-\frac{\gamma n_2}{24}\right) + 2\exp\left(-\frac{\gamma n_2}{18}\right) + r \cdot \left(\frac{9}{10}\right)^{\frac{\gamma n_L}{r \mu_L}} + r \cdot \left(\frac{9}{10}\right)^{\frac{\gamma n_L}{r \mu_L}}$, which is no larger than δ given that γ satisfies (2.4) (in particular, this ensures each term in the sum is no larger than $\delta/5$).

7.1.4 Proof of Lemma 8

First, note that since $\widehat{\mathcal{L}}_{(1)} = \widetilde{\mathcal{L}}$, we have that $\|P_{\widehat{\mathcal{L}}_{(1)}^{\perp}} \widetilde{M}_{:,i}\|_2 > 0$ for all $i \in \mathcal{I}_{\widetilde{C}}$, and $\|P_{\widehat{\mathcal{L}}_{(1)}^{\perp}} \widetilde{M}_{:,i}\|_2 = 0$ otherwise. This, along with the fact that the entries of $\boldsymbol{\phi}$ be i.i.d. realizations of a *continuous* random variable, imply that with probability one the $1 \times n_2$ vector $x^T \triangleq \boldsymbol{\phi} P_{\widehat{\mathcal{L}}_{(1)}^{\perp}} \widetilde{M}$ is nonzero at every $i \in \mathcal{I}_{\widetilde{C}}$ and zero otherwise. Indeed, since for each $i \in \mathcal{I}_{\widetilde{C}}$ the distribution of $\mathrm{x}_i = \boldsymbol{\phi} P_{\widehat{\mathcal{L}}_{(1)}^{\perp}} \widetilde{M}_{:,i}$ is a continuous random variable with nonzero variance, it takes the value zero with probability zero. On the other hand, for $j \notin \mathcal{I}_{\widetilde{C}}$, $\mathrm{x}_j = \boldsymbol{\phi} P_{\widehat{\mathcal{L}}_{(1)}^{\perp}} \widetilde{M}_{:,j} = 0$ with probability one. With this, we see that exact identification of $\mathcal{I}_{\widetilde{C}}$ can be accomplished if we can identify the support of x from linear measurements of the form $\mathbf{y} = (\mathbf{y}_{(2)})^T = Ax$.

To proceed, we appeal to (now, well-known) results from the compressive sensing literature. We recall one representative result of [20] that is germane to our effort below. Here, we cast the result in the context of the stable embedding formalism introduced above, and state it as a lemma without proof.

Lemma 14 (Adapted from Theorem 1.2 of [20])**.** Let $x \in \mathbb{R}^n$ and $z = Ax$. If A is an ϵ-stable embedding of $(\mathcal{U}_{\binom{n}{2k}}, \{0\})$ for some $\epsilon < \sqrt{2} - 1$ where $\mathcal{U}_{\binom{n}{2k}}$ denotes the union of all $\binom{n}{2k}$ unique $2k$-dimensional linear subspaces of $\mathbb{R}^n$ spanned by canonical basis vectors, and x has at most k nonzero elements, then the solution $\widehat{x}$ of

$$\operatorname*{argmin}_{x} \|x\|_1 \quad \text{s.t. } z = Ax. \tag{7.14}$$

is equal to x.

Now, a straightforward application of Lemma 11 above provides that for any $\delta \in (0, 1)$, if

$$p \geq \frac{2k \log(42/\epsilon) + \log \binom{n}{2k} + \log(2/\delta)}{f(\epsilon/\sqrt{2})} \tag{7.15}$$

then the randomly generated $p \times n_2$ matrix A will be an ϵ-stable embedding of $(\mathcal{U}_{\binom{n}{2k}}, \{0\})$ with probability at least $1 - \delta$. This, along with the well-known bound $\binom{n}{2k} \leq \left(\frac{en}{2k}\right)^{2k}$ and some straightforward simplifications, imply that the condition that p satisfy (2.6) is sufficient to ensure that with probability at least $1-\delta$, A is a $(\sqrt{2}/4)$-stable embedding of $(\mathcal{U}_{\binom{n}{2k}}, \{0\})$. Since $\sqrt{2}/4 < \sqrt{2}-1$, the result follows.

7.1.5 An Upper Tail Bound for the Hypergeometric Distribution

Let $\mathrm{hyp}(N, M, n)$ denote the hypergeometric distribution parameterized by a population of size N with M positive elements and n draws, so $H \sim \mathrm{hyp}(N, M, n)$ is a random variable whose value corresponds to the number of positive elements acquired from n draws (without replacement). The probability mass function of $H \sim \mathrm{hyp}(N, M, n)$ is $\Pr(H = k) = \binom{M}{k}\binom{N-M}{n-k}/\binom{N}{n}$ for $k \in \{\max\{0, n+M-N\}, \ldots, \min\{M, n\}\}$, and its mean value is $\mathbb{E}[H] = nM/N$.

It is well-known that the tails of the hypergeometric distribution are similar to those of the binomial distribution for n trials and success probability $p = M/N$. For example, [235] established that for all $t \geq 0$, $\Pr(H - np \geq nt) \leq e^{-2t^2 n}$, a result that follows directly from Hoeffding's work [236], and exhibits the same tail behavior as predicted by the Hoeffding Inequality for a $\mathrm{Binomial}(n, p)$ random variable (see, e.g., [232]). Below we provide a lemma that yields tighter bounds on the upper tail of H when the fraction of positive elements in the population is near 0 or 1. Our result is somewhat analogous to [232, Theorem 2.3(b)] for the Binomial case.

Lemma 15. Let $H \sim \mathrm{hyp}(N, M, n)$, and set $p = M/N$. For any $\epsilon \geq 0$,

$$\Pr(H \geq (1+\epsilon)np) \leq e^{-\frac{\epsilon^2 np}{2(1+\epsilon/3)}}. \tag{7.16}$$

Proof. We begin with an intermediate result of [235], that for any $t \geq 0$ and $h \geq 1$,

$$\Pr(H - pn \geq tn) \leq \Big(h^{-(p+t)}(1-p+hp)\Big)^n. \tag{7.17}$$

Now, for the specific choices $t = \epsilon p$ and $h = 1+\epsilon$ we have

$$\Pr(H - np \geq \epsilon np) \leq \Big((1+\epsilon)^{-(1+\epsilon)p}(1+\epsilon p)\Big)^n \overset{(a)}{\leq} \Big((1+\epsilon)^{-(1+\epsilon)}e^{\epsilon}\Big)^{np} \overset{(b)}{\leq} e^{-\frac{\epsilon^2 np}{2(1+\epsilon/3)}},$$

where (a) follows from the inequality $1 + x \leq e^x$ (with $x = \epsilon p$), and (b) follows directly from [232, Lemma 2.4]. □

7.2 Proofs for Chapter 3

7.2.1 Proof of Theorem 3

From the main result in [107], we have that (3.8) holds if m and s satisfy

$$\begin{aligned} m &\gtrsim \varepsilon^{-2}(\log^3 m)(\log^5 n)\gamma_2^2(\mathcal{V}, \rho_{\text{Fin}}) + \varepsilon^{-2}(\log^4 m)(\log^5 n)\left(p_{\mathcal{V}} + \log \mathcal{N}(\mathcal{V}, \rho_{\text{Fin}}, \varepsilon_0)\right), \\ s &\gtrsim \left(\tilde{\alpha}^2 \log^2 \mathcal{N}(\mathcal{V}, \rho_{\text{Fin}}, \varepsilon_0) + \varepsilon_0^2 p_{\mathcal{V}} \log \frac{1}{\varepsilon_0} + \left[\int_0^{\varepsilon_0} (\log \mathcal{N}(\mathcal{V}, \rho_{\text{Fin}}, t))^{1/2}\, dt\right]^2\right) \\ &\qquad \cdot (\log^4 m)(\log^5 n)\varepsilon^{-2} + \varepsilon^{-2}(\log^6 m)(\log^4 n), \end{aligned}$$

which can be obtained from (3.10) and (3.11).

7.2.2 Proof of Theorem 4

We start with an illustration that the set $\mathcal{T}$ can be reparameterized to the following set with respect to tensors with orthogonal factors:

$$\mathcal{T} = \bigcup_{E \in \mathcal{V}} \{x \in E \;:\; \|x\|_2 = 1\}, \quad \text{where} \;\; \mathcal{V} = \bigcup_{\widetilde{\mathcal{W}}} \{\operatorname{span}[A^{v_1}, A^{v_2}]\} \quad \text{and}$$
$$\widetilde{\mathcal{W}} = \{v_1, v_2 \in \mathcal{B}_{p_2} \text{ with } \langle v_1, v_2 \rangle = 0\}.$$

Suppose $\langle v_1, v_2 \rangle \neq 0$, then let $v_2 = \alpha v_1 + \beta z$ for some $\alpha, \beta \in \mathbb{R}$ and a unit vector $z \in \mathbb{R}^{p_2}$, where $\langle v_1, z \rangle = 0$. Then we have

$$\begin{aligned} \frac{Ax - Ay}{\|Ax - Ay\|_2} &= \frac{A^{v_1}u_1 - A^{v_2}u_2}{\|A^{v_1}u_1 - A^{v_2}u_2\|_2} = \frac{A^{v_1}u_1 - A^{\alpha v_1 + \beta z}u_2}{\|A^{v_1}u_1 - A^{\alpha v_1 + \beta z}u_2\|_2} \\ &= \frac{A^{v_1}u_1 - A^{\alpha v_1}u_2 - A^{\beta z}u_2}{\|A^{v_1}u_1 - A^{\alpha v_1}u_2 - A^{\beta z}u_2\|_2} = \frac{A^{v_1}(u_1 - \alpha u_2) - A^{z}(\beta u_2)}{\|A^{v_1}(u_1 - \alpha u_2) - A^{z}(\beta u_2)\|_2}, \end{aligned}$$

which is equivalent to $\langle v_1, v_2 \rangle = 0$ by reparameterizing z as v_2.

Next, by Theorem 3, we need to upper bound $\rho_{\mathcal{V}}$, $\gamma_2^2(\mathcal{V}, \rho_{\text{Fin}})$, and $\mathcal{N}(\mathcal{V}, \rho_{\text{Fin}}, \varepsilon_0)$. These will be addressed separately as follows.

Part 1: Bound $p_{\mathcal{V}}$. For notational convenience, we denote $A^{v_1,v_2} = [A^{v_1}, A^{v_2}]$. It is straightforward that

$$p_{\mathcal{V}} = \sup_{v_1, v_2 \in \mathcal{B}_{p_2}, \langle v_1, v_2 \rangle = 0} \dim \{\text{span}\,(A^{v_1,v_2})\} \leq 2p_1. \tag{7.18}$$

Part 2: Bound $\gamma_2^2(\mathcal{V}, \rho_{\mathbf{Fin}})$. By the definition of γ_2-functional in (3.7) for the Finsler metric, we have

$$\gamma_2(\mathcal{V}, \rho_{\text{Fin}}) = \inf_{\{\mathcal{V}_k\}_{k=0}^{\infty}} \sup_{A^{v_1,v_2} \in \mathcal{V}} \sum_{k=0}^{\infty} 2^{k/2} \cdot \rho_{\text{Fin}}(A^{v_1,v_2}, \overline{\mathcal{V}}_k),$$

where $\overline{\mathcal{V}}_k$ is an ε_k-net of $\mathcal{V}_k$, i,e., for any $A^{v_1,v_2} \in \mathcal{V}$ there exist $\overline{v}_1, \overline{v}_2 \in \mathcal{B}_{p_2}$ with $\langle \overline{v}_1, \overline{v}_2 \rangle = 0$, $\|v_1 - \overline{v}_1\|_2 \leq \eta_k$, and $\|v_2 - \overline{v}_2\|_2 \leq \eta_k$, such that $A^{\overline{v}_1,\overline{v}_2} \in \overline{\mathcal{V}}_k$ and $\rho_{\text{Fin}}(A^{v_1,v_2}, A^{\overline{v}_1,\overline{v}_2}) \leq \varepsilon_k$.

From Lemma 20, we have $\rho_{\text{Fin}}(A^{v_1,v_2}, \overline{\mathcal{V}}_k) \leq 2\kappa(A)\eta_k$ for $\|v_1 - \overline{v}_1\|_2 \leq \eta_k$ and $\|v_2 - \overline{v}_2\|_2 \leq \eta_k$. On the other hand, we have that $\rho_{\text{Fin}}(A^{v_1,v_2}, \overline{\mathcal{V}}_k) \leq 1$ always holds. Therefore, we have

$$\rho_{\text{Fin}}(A^{v_1,v_2}, \overline{\mathcal{V}}_k) \leq \min\{2\kappa(A)\eta_k, 1\}.$$

Let k' be the smallest integer such that $2\kappa(A)\eta_{k'} \leq 1$. Then we have

$$\gamma_2(\mathcal{V}, \rho_{\text{Fin}}) \leq \sum_{k=0}^{\infty} 2^{k/2} \rho_{\text{Fin}}(A^{v_1,v_2}, \overline{\mathcal{V}}_k) \leq \sum_{k=0}^{k'} 2^{k/2} + \sum_{k=k'+1}^{\infty} 2^{k/2} \rho_{\text{Fin}}(A^{v_1,v_2}, \overline{\mathcal{V}}_k). \tag{7.19}$$

Suppose that $\eta_0 = 1$. Then we have $|\overline{\mathcal{V}}_0| = 1$. For $k \geq 1$, we have $\eta_k < 1$ and $|\overline{\mathcal{V}}_k| \leq (3/\eta_k)^{p_2}$ [237]. By the definition of admissible sequences in the γ_2-functional, we require $|\overline{\mathcal{V}}_k| \leq 2^{2^k}$. Without loss of generality, suppose that for all $k \leq k'$, we have $|\overline{\mathcal{V}}_k| \leq 2^{2^k} \leq (3/\eta_k)^{p_2}$. Then we have $2^{k/2} \leq \sqrt{p_2 \log \frac{3}{\eta_k}}$, which implies

$$\sum_{k=0}^{k'} 2^{k/2} = \frac{2^{k'/2}}{\sqrt{2} - 1} \lesssim \sqrt{p_2 \log \frac{1}{\eta_{k'}}}. \tag{7.20}$$

For $k > k'$, suppose we choose $\eta_{k+1} = \eta_k^2$. Then we have

$$\left(\frac{3}{\eta_{k+1}}\right)^{p_2} \le \left(\frac{3}{\eta_k}\right)^{2p_2} \le \left(2^{2^k}\right)^2 = 2^{2^{k+1}}, \tag{7.21}$$

which implies $|\overline{\mathcal{V}}_{k+1}| \le 2^{2^{k+1}}$ as long as $|\overline{\mathcal{V}}_{k+1}| \le (3/\eta_{k+1})^{p_2}$ holds. In other words, we have $|\overline{\mathcal{V}}_k| \le 2^{2^k}$ if we choose $\eta_{k+1} = \eta_k^2$ for all $k > k'$. Suppose k' is the smallest integer such that when we choose $\eta_{k'+1} = \frac{1}{4\kappa(A)}$, then $\left(\frac{3}{\eta_{k'+1}}\right)^{p_2} \le 2^{2^{k'+1}}$ holds. This implies (7.21) holds and $\rho_{\mathrm{Fin}}(A^{v_1,v_2}, \overline{\mathcal{V}}_k) \le (1/2)^{2^{k-k'}}$ for all $k > k'$. Then we have

$$\sum_{k=k'+1}^{\infty} 2^{k/2} \cdot \rho_{\mathrm{Fin}}(A^{v_1,v_2}, \overline{\mathcal{V}}_k) = 2^{k'/2} \cdot \sum_{t=1}^{\infty} 2^{t/2} \cdot \left(\frac{1}{2}\right)^{2^t} \le 2^{k'/2} \lesssim \sqrt{p_2 \log \frac{1}{\eta_{k'}}}, \tag{7.22}$$

where the first inequality is from the Cauchy condensation test $\sum_{t=0}^{\infty} 2^{t/2} \cdot \left(\frac{1}{2}\right)^{2^t} \le 2 \cdot \sum_{t=0}^{\infty} \left(\frac{1}{2}\right)^t = 1$ and the second inequality is from (7.20).

Combining (7.19), (7.20), and (7.22), we have

$$\gamma_2^2(\mathcal{V}, \rho_{\mathrm{Fin}}) \lesssim p_2 \log \frac{1}{\eta_{k'}}. \tag{7.23}$$

From Lemma 20, suppose we choose a small enough ε_0 such that $\varepsilon_0 \le 2\kappa(A)\eta_{k'}$. Then (7.23) implies

$$\gamma_2^2(\mathcal{V}, \rho_{\mathrm{Fin}}) \lesssim p_2 \log \frac{\kappa(A)}{\varepsilon_0}. \tag{7.24}$$

Part 3: Bound $\mathcal{N}(\mathcal{V}, \rho_{\mathbf{Fin}}, \varepsilon_0)$. From our choice from Part 2, $\varepsilon_0 \in (0,1)$ is a constant. Then it is straightforward that

$$\mathcal{N}(\mathcal{V}, \rho_{\mathrm{Fin}}, \varepsilon_0) \le \left(\frac{3}{\varepsilon_0}\right)^{2p_2}. \tag{7.25}$$

This implies

$$\int_0^{\varepsilon_0} [\log \mathcal{N}(\mathcal{V}, \rho_{\text{Fin}}, t)]^{1/2} dt \leq \int_0^{\varepsilon_0} \left(\log (3/t)^{p_2}\right)^{1/2} dt \tag{7.26}$$

$$\lesssim \sqrt{p_2} \int_0^{\varepsilon_0} (-\log t)^{1/2}\, dt \quad \left(\text{Let } w = (-\log t)^{1/2}\right)$$

$$= \sqrt{p_2} \int_{-\infty}^{(-\log \varepsilon_0)^{1/2}} 2w^2 e^{-w^2} dw = \sqrt{p_2} \left(\left[w \cdot e^{-w^2} \right]_{-\infty}^{(-\log \varepsilon_0)^{1/2}} - \int_{-\infty}^{(-\log \varepsilon_0)^{1/2}} e^{-w^2} dw \right)$$

$$\leq \sqrt{p_2} \left[w \cdot e^{-w^2} \right]_{-\infty}^{(-\log \varepsilon_0)^{1/2}} = \varepsilon_0 \sqrt{p_2 \log \frac{1}{\varepsilon_0}}. \tag{7.27}$$

From Lemma 18, we have

$$\tilde{\alpha}^2 = \max_{i \in [n]} \ell_i^2(A^{v_1, v_2}) \leq \max_{i \in [n]} \ell_i^2(A) \leq 1/p_2^2. \tag{7.28}$$

Combining (7.18), (7.24)–(7.28), and Theorem 3, we have that the claim holds if

$$m \gtrsim \varepsilon^{-2} \left(p_2 \log \frac{\kappa(A)}{\varepsilon_0} + p_1 + p_2 \log \frac{1}{\varepsilon_0} \right) (\log^4 m)(\log^5 n),$$
$$s \gtrsim \varepsilon^{-2} \left(\log^2 \frac{1}{\varepsilon_0} + \varepsilon_0^2 (p_1 + p_2) \log \frac{1}{\varepsilon_0} \right) (\log^6 m)(\log^5 n).$$

Taking $\varepsilon_0 = 1/(p_1 + p_2)$, we finish the proof. Note that since $2\kappa(A)\eta_{k'} \geq 1/2$, we only require $\rho_{\text{Fin}}(A^{v_1, v_2}, \overline{\mathcal{V}}_{k'}) \leq 1/2$ in Part 2. Thus the choice $\varepsilon_0 = 1/(p_1 + p_2)$ is valid here.

7.2.3 Proof of Theorem 5

Denote $A^{\left\{v_i^{(r)}\right\}} = \left[A^{\left\{v_1^{(r)}\right\}}, A^{\left\{v_2^{(r)}\right\}} \right] \in \mathbb{R}^{n \times 2Rp_1}$. We illustrate that the set $\mathcal{T}$ can be reparameterized to the following set with respect to tensors with partial orthogonal factors:

$$\mathcal{T} = \bigcup_{E \in \mathcal{V}} \left\{ x \in E \ : \ \|x\|_2 = 1 \right\}, \quad \text{where} \ \ \mathcal{V} = \bigcup_{\widetilde{\mathcal{W}}} \text{span} \left(A^{\left\{v_i^{(r)}\right\}} \right) \quad \text{and}$$
$$\widetilde{\mathcal{W}} = \left\{ \forall i \in [2],\ r, q \in [R],\ q \neq r, v_i^{(r)} \in \mathcal{B}_{p_2}, \langle v_1^{(r)}, v_2^{(r)} \rangle = \langle v_i^{(r)}, v_i^{(q)} \rangle = 0 \right\}.$$

Suppose for all $r \in [R]$, $v_2^{(r)} = \alpha^{(r)} v_1^{(r)} + \beta^{(r)} z^{(r)}$ for some $\alpha^{(r)}, \beta^{(r)} \in \mathbb{R}$ and unit vectors $z^{(r)} \in \mathbb{R}^{p_2}$, where $\langle v_1^{(r)}, z^{(r)} \rangle = 0$. Then we have

$$\begin{aligned} Ax - Ay &= \sum_{r=1}^{R} \left(A^{v_1^{(r)}} \cdot u_1^{(r)} - A^{v_2^{(r)}} \cdot u_2^{(r)} \right) = \sum_{r=1}^{R} \left(A^{v_1^{(r)}} \cdot u_1^{(r)} - A^{\alpha^{(r)} v_1^{(r)} + \beta^{(r)} z^{(r)}} \cdot u_2^{(r)} \right) \\ &= \sum_{r=1}^{R} \left(A^{v_1^{(r)}} \cdot u_1^{(r)} - A^{\alpha^{(r)} v_1^{(r)}} \cdot u_2^{(r)} - A^{\beta^{(r)} z^{(r)}} \cdot u_2^{(r)} \right) \\ &= \sum_{r=1}^{R} \left(A^{v_1^{(r)}} \cdot \left(u_1^{(r)} - \alpha^{(r)} u_2^{(r)} \right) - A^{z^{(r)}} \cdot \left(\beta^{(r)} u_2^{(r)} \right) \right). \end{aligned}$$

which is equivalent to $\langle v_1^{(r)}, v_2^{(r)} \rangle = 0$ by reparameterizing $z^{(r)}$ as $v_2^{(r)}$.

Using a similar argument, we show the general scenario. For any $r \in [R]$, $r \geq 2$, w.l.o.g., suppose

$$v_1^{(r)} = \alpha_1^{(r,1)} v_1^{(1)} + \sum_{i=2}^{r} \alpha_1^{(r,i)} z_1^{(i)} \quad \text{and} \quad v_2^{(r)} = \beta_1^{(r,1)} v_1^{(1)} + \sum_{i=2}^{r} \beta_1^{(r,i)} z_1^{(i)} + \sum_{j=1}^{r} \beta_2^{(r,j)} z_2^{(j)}.$$

where $\alpha_1^{(r,i)}, \beta_1^{(r,i)}, \beta_2^{(r,j)} \in \mathbb{R}$ are real coefficients and $\langle v_1^{(1)}, z_1^{(i)} \rangle = \langle v_1^{(1)}, z_2^{(i)} \rangle = \langle z_1^{(i)}, z_2^{(j)} \rangle = 0$ for any $i, j \in [r]$. For $R = 1$, the argument is identical to the one above. For $2 \leq R \leq p_2/2$, we have

$$\begin{aligned} Ax - Ay &= \sum_{r=1}^{R} \left(A^{v_1^{(r)}} \cdot u_1^{(r)} - A^{v_2^{(r)}} \cdot u_2^{(r)} \right) \\ &= \sum_{r=2}^{R} \left(A^{\alpha_1^{(r,1)} v_1^{(1)} + \sum_{i=2}^{r} \alpha_1^{(r,i)} z_1^{(i)}} \cdot u_1^{(r)} - A^{\beta_1^{(r,1)} v_1^{(1)} + \sum_{i=2}^{r} \beta_1^{(r,i)} z_1^{(i)} + \sum_{j=1}^{r} \beta_2^{(r,j)} z_2^{(j)}} \cdot u_2^{(r)} \right) \\ &\qquad + A^{v_1^{(1)}} \cdot u_1^{(r)} - A^{\left(\beta_1^{(1,1)} v_1^{(1)} + \beta_2^{(1,1)} z_2^{(1)} \right)} \cdot u_2^{(r)} \\ &= \sum_{r=2}^{R} \Bigg(A^{v_1^{(1)}} \cdot \left(\alpha_1^{(r,1)} u_1^{(1)} - \beta_1^{(r,1)} u_2^{(1)} \right) + \sum_{i=2}^{r} A^{z_1^{(i)}} \cdot \left(\alpha_1^{(r,i)} u_1^{(i)} - \beta_1^{(r,i)} u_2^{(i)} \right) \\ &\qquad - \sum_{j=1}^{r} A^{z_2^{(j)}} \cdot \left(\beta_2^{(r,j)} u_2^{(j)} \right) \Bigg) + A^{v_1^{(1)}} \cdot u_1^{(r)} - A^{\left(\beta_1^{(1,1)} v_1^{(1)} + \beta_2^{(1,1)} z_2^{(1)} \right)} \cdot u_2^{(r)}, \end{aligned}$$

which is equivalent to $\langle v_i^{(r)}, v_i^{(q)} \rangle = 0$ and $\langle v_1^{(r)}, v_2^{(r)} \rangle = 0$ for all $i \in [2]$, $r \in [R]$, and

$q \neq r$ by reparameterizing $z_1^{(i)}$ as $v_1^{(i)}$ and $z_2^{(j)}$ as $v_2^{(j)}$.

Next, analogous to Theorem 4, we analyze upper bounds on $\rho_{\mathcal{V}}$, $\gamma_2^2(\mathcal{V}, \rho_{\text{Fin}})$, and $\mathcal{N}(\mathcal{V}, \rho_{\text{Fin}}, \varepsilon_0)$, and obtain the result from Theorem 3.

Part 1: Bound $p_{\mathcal{V}}$. It is straightforward that

$$p_{\mathcal{V}} = \sup_{\widetilde{\mathcal{W}}} \ \dim \left\{ \operatorname{span} \left(A^{\left\{ v_i^{(r)} \right\}} \right) \right\} \leq 2Rp_1. \tag{7.29}$$

Part 2: Bound $\gamma_2^2(\mathcal{V}, \rho_{\mathbf{Fin}})$. The γ_2-functional in this case is

$$\gamma_2^2(\mathcal{V}, \rho_{\text{Fin}}) = \inf_{\{\mathcal{V}_k\}_{k=0}^{\infty}} \sup_{A^{\left\{ v_i^{(r)} \right\}} \in \mathcal{V}} \sum_{k=0}^{\infty} 2^{r/2} \cdot \rho_{\text{Fin}} \left(A^{\left\{ v_i^{(r)} \right\}}, \overline{\mathcal{V}}_k \right),$$

where $\overline{\mathcal{V}}_k$ is an ε_k-net of $\mathcal{V}_k$.

Following the same argument in Part 2 of the proof for Theorem 4, we have from Lemma 21 that if k' is the smallest integer such that $2R\kappa(A)\eta_{k'} \leq 1$ and we choose $\eta_{k'+1} = \frac{1}{4R\kappa(A)}$, then we choose a small enough ε_0 such that $\varepsilon_0 \leq 2R\kappa(A)\eta_{k'}$,

$$\gamma_2^2(\mathcal{V}, \rho_{\text{Fin}}) \lesssim Rp_2 \log \frac{R\kappa(A)}{\varepsilon_0}. \tag{7.30}$$

Part 3: Bound $\mathcal{N}(\mathcal{V}, \rho_{\mathbf{Fin}}, \varepsilon_0)$. It is straightforward that

$$\mathcal{N}(\mathcal{V}, \rho_{\text{Fin}}, \varepsilon_0) \leq \left(\frac{3}{\varepsilon_0} \right)^{2Rp_2}.$$

Following the same argument in Part 3 of the proof for Theorem 4, we have

$$\int_0^{\varepsilon_0} [\log \mathcal{N}(\mathcal{V}, \rho_{\text{Fin}}, t)]^{1/2} dt \lesssim \varepsilon_0 \sqrt{Rp_2 \log \frac{1}{\varepsilon_0}}. \tag{7.31}$$

From Lemma 19, we have

$$\tilde{\alpha}^2 = \max_{i \in [n]} \ell_i^2 \left(A^{\left\{ v_i^{(r)} \right\}} \right) \leq \max_{i \in [n]} \ell_i^2(A) \leq 1/(R^2 p_2^2). \tag{7.32}$$

Combining (7.29) – (7.32) and Theorem 3, we have that the claim holds if

$$m \gtrsim \varepsilon^{-2} R \left(p_2 \log \frac{R\kappa(A)}{\varepsilon_0} + p_1 + p_2 \log \frac{1}{\varepsilon_0} \right) (\log^4 m)(\log^5 n),$$
$$s \gtrsim \varepsilon^{-2} \left(\log^2 \frac{1}{\varepsilon_0} + \varepsilon_0^2 R(p_1+p_2) \log \frac{1}{\varepsilon_0} \right) (\log^6 m)(\log^5 n).$$

We finish the proof by taking $\varepsilon_0 = \frac{1}{R(p_1+p_2)}$. Note that this choice of ε satisfies the requirement in Part 2.

7.2.4 Proof of Theorem 6

Denote $\vartheta_{\setminus 1} = \theta_D \otimes \cdots \otimes \theta_2$, $\varphi_{\setminus 1} = \phi_D \otimes \cdots \otimes \phi_2$ and $A^{\vartheta_{\setminus 1}, \varphi_{\setminus 1}} = \left[A^{\{\theta_{\setminus 1}\}}, A^{\{\phi_{\setminus 1}\}} \right] \in \mathbb{R}^{n \times 2p_1}$. We illustrate that the set $\mathcal{T}$ can be reparameterized to the following set with respect to tensors with partial orthogonal factors:

$$\mathcal{T} = \bigcup_{E \in \mathcal{V}} \{ x \in E \ : \ \|x\|_2 = 1 \}, \quad \text{where} \ \ \mathcal{V} = \bigcup_{\widetilde{\mathcal{W}}} \text{span} \left(A^{\vartheta_{\setminus 1}, \varphi_{\setminus 1}} \right) \quad \text{and}$$
$$\widetilde{\mathcal{W}} = \{ \forall d \in [D] \backslash \{1\}, \ \theta_d, \phi_d \in \mathcal{B}_{p_d}, \exists i \in [D] \backslash \{1\} \text{ s.t. } \langle \theta_i, \phi_i \rangle = 0 \},$$

W.l.o.g., suppose $\phi_D = \alpha \theta_D + \beta z$ for some $\alpha, \beta \in \mathbb{R}$ and a unit vector $z \in \mathbb{R}^{p_D}$, where $\langle \theta_D, z \rangle = 0$. Then we have

$$\begin{aligned}
&A\vartheta - A\varphi \qquad = A^{\{\theta_{\setminus 1}\}}\theta_1 - A^{\{\phi_{\setminus 1}}\\
&= A(\theta_D \otimes \cdots \otimes \theta_2 \otimes I_{p_1})\theta_1 - A((\alpha\theta_D + \beta z) \otimes \phi_{D-1} \otimes \cdots \otimes \phi_2 \otimes I_{p_1})\phi_1 \\
&= A(\theta_D \otimes \cdots \otimes \theta_2 \otimes I_{p_1})\theta_1 - A(\alpha\theta_D \otimes \cdots \otimes \phi_2 \otimes I_{p_1})\phi_1 - A(\beta z \otimes \cdots \otimes \phi_2 \otimes I_{p_1})\phi_1 \\
&= A^{\theta_D} (\theta_{D-1} \otimes \cdots \otimes \theta_1 - \alpha \phi_{D-1} \otimes \cdots \otimes \phi_1) - A^z (\phi_{D-1} \otimes \cdots \otimes \phi_1),
\end{aligned}$$

This is equivalent to $\langle \theta_D, \phi_D \rangle = 0$ by reparameterizing z as ϕ_D.

Next, analogous to Theorem 4, we analyze upper bounds on $\rho_{\mathcal{V}}$, $\gamma_2^2(\mathcal{V}, \rho_{\text{Fin}})$, and $\mathcal{N}(\mathcal{V}, \rho_{\text{Fin}}, \varepsilon_0)$, and obtain the result from Theorem 3.

Part 1: Bound $p_{\mathcal{V}}$. It is straightforward that

$$p_{\mathcal{V}} = \sup_{\widetilde{\mathcal{W}}} \ \dim \left\{ \text{span} \left(A^{\vartheta_{\setminus 1}, \varphi_{\setminus 1}} \right) \right\} \leq 2p_1. \tag{7.33}$$

Part 2: Bound $\gamma_2^2(\mathcal{V}, \rho_{\mathbf{Fin}})$. The γ_2-functional in this case is

$$\gamma_2^2(\mathcal{V}, \rho_{\mathrm{Fin}}) = \inf_{\{\mathcal{V}_k\}_{k=0}^{\infty}} \sup_{A^{\vartheta_{\backslash 1}, \varphi_{\backslash 1}} \in \mathcal{V}} \sum_{k=0}^{\infty} 2^{r/2} \cdot \rho_{\mathrm{Fin}} \left(A^{\vartheta_{\backslash 1}, \varphi_{\backslash 1}}, \overline{\mathcal{V}}_k \right),$$

where $\overline{\mathcal{V}}_k$ is an ε_k-net of $\mathcal{V}_k$.

Following the same argument in Part 2 of the proof of Theorem 4, we have from Lemma 22 that if k' is the smallest integer such that $2\kappa(A)\left((1+\eta_{k'})^D - 1\right) \leq 1$, then we choose ε_0 small enough such that

$$\varepsilon \leq 2\kappa(A) D \eta_{k'} \leq 2\kappa(A)\left((1+\eta_{k'})^D - 1\right).$$

where the second inequality is from the binomial expansion. Then we have

$$\gamma_2^2(\mathcal{V}, \rho_{\mathrm{Fin}}) \lesssim \sum_{d=2}^{D} p_d \cdot \log \frac{D\kappa(A)}{\varepsilon_0}. \tag{7.34}$$

Part 3: Bound $\mathcal{N}(\mathcal{V}, \rho_{\mathbf{Fin}}, \varepsilon_0)$. It is straightforward that

$$\mathcal{N}(\mathcal{V}, \rho_{\mathrm{Fin}}, \varepsilon_0) \leq \left(\frac{3}{\varepsilon_0}\right)^{2\sum_{d=2}^{D} p_d}.$$

Following the same argument in Part 3 of the proof for Theorem 4, we have

$$\int_0^{\varepsilon_0} [\log \mathcal{N}(\mathcal{V}, \rho_{\mathrm{Fin}}, t)]^{1/2} dt \lesssim \varepsilon_0 \sqrt{\sum_{d=2}^{D} p_d \log \frac{1}{\varepsilon_0}}. \tag{7.35}$$

From Lemma 18, we have

$$\tilde{\alpha}^2 = \max_{i \in [n]} \ell_i^2 \left(A^{\vartheta_{\backslash 1}, \varphi_{\backslash 1}}\right) \leq \max_{i \in [n]} \ell_i^2(A) \leq 1 / \left(\sum_{d=2}^{D} p_d\right)^2. \tag{7.36}$$

Combining (7.33) – (7.36) and Theorem 3, we have that the claim holds if

$$m \gtrsim \varepsilon^{-2}\left(p_1 + \sum_{d=2}^{D} p_d \cdot \log \frac{D\kappa(A)}{\varepsilon_0}\right)(\log^4 m)(\log^5 n),$$
$$s \gtrsim \varepsilon^{-2}\left(\log^2 \frac{1}{\varepsilon_0} + \varepsilon_0^2 \sum_{d=1}^{D} p_d \log \frac{1}{\varepsilon_0}\right)(\log^6 m)(\log^5 n).$$

We finish the proof by taking $\varepsilon_0 = \frac{1}{\sum_{d=1}^{D} p_d}$. Note that this choice of ε satisfies the requirement in Part 2.

7.2.5 Proof of Theorem 7

Denote $A^{\left\{\vartheta_{\backslash 1}^{(r)}, \varphi_{\backslash 1}^{(r)}\right\}} = \left[A^{\left\{\theta_{\backslash 1}^{(r)}\right\}}, A^{\left\{\phi_{\backslash 1}^{(r)}\right\}}\right]$. We illustrate that the set $\mathcal{T}$ can be reparameterized to the following set with respect to tensors with partial orthogonal factors:

$$\mathcal{T} = \bigcup_{E \in \mathcal{V}} \left\{x \in E \ : \ \|x\|_2 = 1\right\}, \quad \text{where} \ \ \mathcal{V} = \bigcup_{\widehat{\mathcal{W}}} \text{span}\left(A^{\left\{\vartheta_{\backslash 1}^{(r)}, \varphi_{\backslash 1}^{(r)}\right\}}\right),$$
$$\widetilde{\mathcal{W}} = \Big\{\forall r \in [R], d \in [D]\backslash\{1\}, \theta_d^{(r)}, \phi_d^{(r)} \in \mathcal{B}_{p_d}; \forall r, q \in [R], \exists i \in [D]\backslash\{1\} \text{ s.t.} \langle \theta_i^{(r)}, \phi_i^{(q)} \rangle = 0;$$
$$\forall r \in [R-1], q \in [R]\backslash[r], \exists j, k \in [D]\backslash\{1\} \text{ s.t. } \langle \theta_j^{(r)}, \theta_j^{(q)} \rangle = \langle \phi_k^{(r)}, \phi_k^{(q)} \rangle = 0\Big\}.$$

For $R = 1$, the argument is identical to the analysis in Theorem 6. For any $r \in [R]$, $r \geq 2$, w.l.o.g., suppose

$$\theta_D^{(r)} = \alpha_1^{(r,1)} \theta_D^{(1)} + \sum_{i=2}^{r} \alpha_1^{(r,i)} z_1^{(i)} \quad \text{and} \quad \phi_D^{(r)} = \beta_1^{(r,1)} \theta_D^{(1)} + \sum_{i=2}^{r} \beta_1^{(r,i)} z_1^{(i)} + \sum_{j=1}^{r} \beta_2^{(r,j)} z_2^{(j)},$$

where $\alpha_1^{(r,i)}, \beta_1^{(r,i)}, \beta_2^{(r,j)} \in \mathbb{R}$ are real coefficients and $\langle \theta_D^{(1)}, z_1^{(i)} \rangle = \langle \theta_D^{(1)}, z_2^{(i)} \rangle =$

$\langle z_1^{(i)}, z_2^{(j)} \rangle = 0$ for any $i, j \in [r]$. Then for $2 \leq R \leq p_2/2$, we have

$$\begin{aligned}
A\vartheta - A\varphi &= A \cdot \sum_{r=1}^{R} \left(\theta_D^{(r)} \otimes \cdots \otimes \theta_2^{(r)} \otimes I_{p_1}\right) \theta_1^{(r)} - A \cdot \sum_{r=1}^{R} \left(\phi_D^{(r)} \otimes \cdots \otimes \phi_2^{(r)} \otimes I_{p_1}\right) \phi_1^{(r)} \\
&= A \cdot \sum_{r=2}^{R} \left(\left(\alpha_1^{(r,1)} \theta_D^{(1)} + \sum_{i=2}^{r} \alpha_1^{(r,i)} z_1^{(i)} \right) \otimes \cdots \otimes \theta_1^{(r)} \right) + A \cdot \left(\theta_D^{(1)} \otimes \cdots \otimes \theta_1^{(1)} \right) \\
&\quad - A \cdot \sum_{r=2}^{R} \left(\left(\beta_1^{(r,1)} \theta_D^{(1)} + \sum_{i=2}^{r} \beta_1^{(r,i)} z_1^{(i)} + \sum_{j=1}^{r} \beta_2^{(r,j)} z_2^{(j)} \right) \otimes \cdots \otimes \phi_1^{(r)} \right) \\
&\quad - A \cdot \left(\left(\beta_1^{(1,1)} \theta_D^{(1)} + \beta_2^{(1,1)} z_2^{(1)} \right) \otimes \cdots \otimes \phi_1^{(1)} \right) \\
&= \sum_{r=r}^{R} A^{\theta_D^{(1)}} \left(\alpha_1^{(r,1)} \theta_{D-1}^{(r)} \otimes \cdots \otimes \theta_1^{(r)} - \beta_1^{(r,1)} \phi_{D-1}^{(r)} \otimes \cdots \otimes \phi_1^{(r)} \right) \\
&\quad + \sum_{r=2}^{R} \sum_{i=2}^{r} A^{z_1^{(1)}} \left(\alpha_1^{(r,i)} \theta_{D-1}^{(r)} \otimes \cdots \otimes \theta_1^{(r)} - \beta_1^{(r,i)} \phi_{D-1}^{(r)} \otimes \cdots \otimes \phi_1^{(r)} \right) \\
&\quad - \sum_{r=1}^{R} \sum_{j=1}^{r} A^{z_2^{(j)}} \left(\beta_2^{(r,j)} \phi_{D-1}^{(r)} \otimes \cdots \otimes \phi_1^{(r)} \right)
\end{aligned}$$

where $\alpha_1^{(,1)} = 1$. This is equivalent to $\langle \theta_D^{(r)}, \phi_D^{(r)} \rangle = 0$, $\langle \theta_D^{(r)}, \theta_D^{(q)} \rangle = 0$, and $\langle \phi_D^{(r)}, \phi_D^{(q)} \rangle = 0$ for all $r \in [R]$ and $q \neq [R] \backslash [r]$, by reparameterizing $z_1^{(i)}$ and $z_2^{(j)}$ as $\theta_D^{(i)}$ and $\phi_D^{(j)}$ properly. The remaining pairs of orthogonality in $\widetilde{\mathcal{W}}$ can be checked analogously by repeating the argument above.

Part 1: Bound $p_{\mathcal{V}}$. It is straightforward that

$$p_{\mathcal{V}} = \sup_{\widetilde{\mathcal{W}}} \dim \left\{ \operatorname{span} \left(A^{\left\{\vartheta_{\backslash 1}^{(r)}, \varphi_{\backslash 1}^{(r)}\right\}} \right) \right\} \leq 2Rp_1. \tag{7.37}$$

Part 2: Bound $\gamma_2^2(\mathcal{V}, \rho_{\mathbf{Fin}})$. The γ_2-functional in this case is

$$\gamma_2^2(\mathcal{V}, \rho_{\text{Fin}}) = \inf_{\{\mathcal{V}_k\}_{k=0}^{\infty}} \sup_{A^{\left\{\vartheta_{\backslash 1}^{(r)}, \varphi_{\backslash 1}^{(r)}\right\}} \in \mathcal{V}} \sum_{k=0}^{\infty} 2^{r/2} \cdot \rho_{\text{Fin}} \left(A^{\left\{\vartheta_{\backslash 1}^{(r)}, \varphi_{\backslash 1}^{(r)}\right\}}, \overline{\mathcal{V}}_k \right),$$

where $\overline{\mathcal{V}}_k$ is an ε_k-net of $\mathcal{V}_k$.

Following the same argument in Part 2 of the proof for Theorem 4, we have from

Lemma 23 that if k' is the smallest integer such that $2R\kappa(A)\left((1+\eta_{k'})^D - 1\right) \le 1$, then we choose ε_0 small enough such that

$$\varepsilon \le 2RD\kappa(A)\eta_{k'} \le 2R\kappa(A)\left((1+\eta_{k'})^D - 1\right),$$

where the second inequality follows from the binomial expansion. Then we have

$$\gamma_2^2(\mathcal{V}, \rho_{\text{Fin}}) \lesssim \sum_{d=2}^{D} p_d \cdot \log \frac{RD\kappa(A)}{\varepsilon_0}. \tag{7.38}$$

Part 3: Bound $\mathcal{N}(\mathcal{V}, \rho_{\mathbf{Fin}}, \varepsilon_0)$. It is straightforward that

$$\mathcal{N}(\mathcal{V}, \rho_{\text{Fin}}, \varepsilon_0) \le \left(\frac{3}{\varepsilon_0}\right)^{2R\sum_{d=2}^{D} p_d}.$$

Following the same argument in Part 3 of the proof for Theorem 4, we have

$$\int_0^{\varepsilon_0} [\log \mathcal{N}(\mathcal{V}, \rho_{\text{Fin}}, t)]^{1/2} dt \lesssim \varepsilon_0 \sqrt{R \sum_{d=2}^{D} p_d \log \frac{1}{\varepsilon_0}}. \tag{7.39}$$

From Lemma 18, we have

$$\tilde{\alpha}^2 = \max_{i\in[n]} \ell_i^2\left(A^{\vartheta_{\setminus 1}, \varphi_{\setminus 1}}\right) \le \max_{i\in[n]} \ell_i^2(A) \le 1/\left(R\sum_{d=2}^{D} p_d\right)^2. \tag{7.40}$$

Combining (7.37) – (7.40) and Theorem 3, we have that the claim holds if

$$m \gtrsim \varepsilon^{-2} R \left(p_1 + \sum_{d=2}^{D} p_d \cdot \log \frac{RD\kappa(A)}{\varepsilon_0}\right) (\log^4 m)(\log^5 n),$$
$$s \gtrsim \varepsilon^{-2} \left(\log^2 \frac{1}{\varepsilon_0} + \varepsilon_0^2 R \sum_{d=1}^{D} p_d \log \frac{1}{\varepsilon_0}\right) (\log^6 m)(\log^5 n).$$

We finish the proof by taking $\varepsilon_0 = \frac{1}{R\sum_{d=1}^{D} p_d}$. Note that this choice of ε satisfies the requirement in Part 2.

7.2.6 Proof of Theorem 8

Denote $A^{\left\{\vartheta_{\backslash 1}^{\{r_d\}},\varphi_{\backslash 1}^{\{r_d\}}\right\}} = \left[A^{\left\{\theta_{\backslash 1}^{\{r_d\}}\right\}}, A^{\left\{\phi_{\backslash 1}^{\{r_d\}}\right\}}\right]$. We illustrate that the set $\mathcal{T}$ can be reparameterized to the following set with respect to tensors with partial orthogonal factors:

$$
\begin{aligned}
\mathcal{T} = \bigcup_{E\in\mathcal{V}} \left\{x \in E \ : \ \|x\|_2 = 1\right\}, \quad \text{where} \ \ \mathcal{V} = \bigcup_{\widehat{\mathcal{W}}} \operatorname{span}\left(A^{\left\{\vartheta_{\backslash 1}^{\{r_d\}},\varphi_{\backslash 1}^{\{r_d\}}\right\}}\right) \quad \text{and}& \\
\widetilde{\mathcal{W}} = \Big\{\forall r_d \in [R_d], d \in [D]\backslash\{1\}, \theta_d^{(r_d)}, \phi_d^{(r_d)} \in \mathcal{B}_{p_d}; \forall r_d, q_d \in [R_d], \exists d \in [D]\backslash\{1\}& \\
\text{s.t. } \langle\theta_d^{(r_d)}, \phi_d^{(q_d)}\rangle = 0; \forall r_d \in [R_d - 1], q_d \in [R_d]\backslash[r_d], \exists d, t \in [D]\backslash\{1\}& \\
\text{s.t. } \langle\theta_d^{(r_d)}, \theta_d^{(q_d)}\rangle = \langle\phi_t^{(r_d)}, \phi_t^{(q_d)}\rangle = 0\Big\}.&
\end{aligned}
$$

Repeating the argument in the proof of Theorem 7, we have the equivalence of $\mathcal{T}$ and the set above.

Part 1: Bound $p_{\mathcal{V}}$. It is straightforward that

$$
p_{\mathcal{V}} = \sup_{\widetilde{\mathcal{W}}} \dim\left\{\operatorname{span}\left(A^{\left\{\vartheta_{\backslash 1}^{\{r_d\}},\varphi_{\backslash 1}^{\{r_d\}}\right\}}\right)\right\} \leq 2R_1 p_1. \tag{7.41}
$$

Part 2: Bound $\gamma_2^2(\mathcal{V}, \rho_{\mathbf{Fin}})$. The γ_2-functional in this case is

$$
\gamma_2^2(\mathcal{V}, \rho_{\text{Fin}}) = \inf_{\{\mathcal{V}_k\}_{k=0}^{\infty}} \sup_{A^{\left\{\vartheta_{\backslash 1}^{\{r_d\}},\varphi_{\backslash 1}^{\{r_d\}}\right\}}\in\mathcal{V}} \sum_{k=0}^{\infty} 2^{r/2} \cdot \rho_{\text{Fin}}\left(A^{\left\{\vartheta_{\backslash 1}^{\{r_d\}},\varphi_{\backslash 1}^{\{r_d\}}\right\}}, \overline{\mathcal{V}}_k\right),
$$

where $\overline{\mathcal{V}}_k$ is an ε_k-net of $\mathcal{V}_k$.

Following the same argument as in Part 2 of the proof for Theorem 4, we have from Lemma 24 that if k' is the smallest integer such that $2\kappa(A)\left((1+\eta_{k'})^D - 1\right)\sqrt{\prod_{d=2}^{D} R_d} \leq 1$, then we choose ε_0 small enough such that

$$
\varepsilon \leq 2D\kappa(A)\eta_{k'}\sqrt{\prod_{d=2}^{D} R_d} \leq 2\kappa(A)\left((1+\eta_{k'})^D - 1\right)\sqrt{\prod_{d=2}^{D} R_d},
$$

where the second inequality follows from the binomial theorem. Then we have

$$\gamma_2^2(\mathcal{V},\rho_{\mathrm{Fin}}) \lesssim \left(\sum_{d=2}^{D} R_d p_d + \prod_{d=1}^{D} p_d\right) \cdot \log \frac{D\kappa(A)\sqrt{\prod_{d=2}^{D} R_d}}{\varepsilon_0}. \tag{7.42}$$

Part 3: Bound $\mathcal{N}(\mathcal{V},\rho_{\mathbf{Fin}},\varepsilon_0)$. It is straightforward that

$$\mathcal{N}(\mathcal{V},\rho_{\mathrm{Fin}},\varepsilon_0) \le \left(\frac{3}{\varepsilon_0}\right)^{2\left(\sum_{d=2}^{D} R_d p_d + \prod_{d=1}^{D} p_d\right)}.$$

Following the same argument in Part 3 of the proof for Theorem 4, we have

$$\int_0^{\varepsilon_0} [\log \mathcal{N}(\mathcal{V},\rho_{\mathrm{Fin}},t)]^{1/2} dt \lesssim \varepsilon_0 \sqrt{\left(\sum_{d=2}^{D} R_d p_d + \prod_{d=1}^{D} p_d\right) \log \frac{1}{\varepsilon_0}}. \tag{7.43}$$

From Lemma 18, we have

$$\tilde{\alpha}^2 = \max_{i\in[n]} \ell_i^2\left(A^{\vartheta_{\setminus 1},\varphi_{\setminus 1}}\right) \le \max_{i\in[n]} \ell_i^2(A) \le 1/\left(\sum_{d=2}^{D} R_d p_d + \prod_{d=1}^{D} p_d\right)^2. \tag{7.44}$$

Combining (7.37) – (7.40) and Theorem 3, we have that the claim holds if

$$\begin{aligned} m &\gtrsim \varepsilon^{-2}\left(R_1 p_1 + \left(\sum_{d=2}^{D} R_d p_d + \prod_{d=1}^{D} p_d\right) \cdot \log \frac{D\kappa(A)\sqrt{\prod_{d=2}^{D} R_d}}{\varepsilon_0}\right)(\log^4 m)(\log^5 n), \\ s &\gtrsim \varepsilon^{-2}\left(\log^2 \frac{1}{\varepsilon_0} + \varepsilon_0^2\left(\sum_{d=1}^{D} R_d p_d + \prod_{d=1}^{D} p_d\right)\log\frac{1}{\varepsilon_0}\right)(\log^6 m)(\log^5 n). \end{aligned}$$

We finish the proof by taking $\varepsilon_0 = \frac{1}{\sum_{d=1}^{D} R_d p_d + \prod_{d=1}^{D} p_d}$. Note that this choice of ε satisfies the requirement in Part 2.

7.2.7 Proof of Lemma 1

Given a unit vector $y \in \mathbb{R}^n$, let $Z_{jk} = H_{jk}\Sigma_{kk}y_k$ for all $j \in [n]$. Then from the independence of H_{jk} and Σ_{kk}, we have

$$\begin{aligned}\mathbb{E}(Z_{jk}) &= \mathbb{E}(H_{jk}\Sigma_{kk}y_k) = \mathbb{E}(H_{jk}) \cdot \mathbb{E}(\Sigma_{kk}) \cdot y_k = 0,\\ \mathrm{Var}(Z_{jk}) &\le \mathbb{E}(H_{jk}^2\Sigma_{kk}^2y_k^2) = \mathbb{E}(H_{jk}^2) \cdot \mathbb{E}(\Sigma_{kk}^2) \cdot y_k^2 = \frac{y_k^2}{n}.\end{aligned}$$

From the Azuma-Hoeffding inequality, for any $t > 0$ we have

$$\mathbb{P}\left(\left|\sum_{k=1}^n Z_{jk}\right| > t\right) \le 2\exp\left(-\frac{nt^2}{2\sum_{k=1}^n y_k^2}\right) = 2\exp\left(-\frac{nt^2}{2}\right).$$

By taking $t = \sqrt{\frac{2\log\left(\frac{2nr}{\delta}\right)}{n}}$, we have

$$\mathbb{P}\left(\left|\sum_{k=1}^n Z_{jk}\right| > \sqrt{\frac{2\log\left(\frac{2nr}{\delta}\right)}{n}}\right) \le 2\exp\left(\log\left(\frac{\delta}{2nr}\right)\right) = \frac{\delta}{nr}.$$

By a union bound, we have

$$\mathbb{P}\left(\|H\Sigma y\|_\infty > \sqrt{\frac{2\log\left(\frac{2nr}{\delta}\right)}{n}}\right) = \mathbb{P}\left(\max_{j\in[n]}\left|\sum_{k=1}^n Z_{jk}\right| > \sqrt{\frac{2\log\left(\frac{2nr}{\delta}\right)}{n}}\right) \le \frac{\delta}{r}.$$

Suppose $A = UQ$, where $U \in \mathbb{R}^{n\times r}$ has orthonormal columns. Then we have for all $i \in [n]$ and $k \in [r]$,

$$\ell_i^2(H\Sigma A) = \ell_i^2(H\Sigma U) \le r \cdot \left(e_i^\top H\Sigma U e_k\right)^2.$$

Using a union bound again, we finish the proof by

$$\mathbb{P}\left(\max_{i\in[n]}\ell_i^2(H\Sigma A) > \frac{2r\log\left(\frac{2nr}{\delta}\right)}{n}\right) \le \mathbb{P}\left(\max_{i\in[n]} r \cdot \left\|e_i^\top H\Sigma U e_k\right\|_\infty^2 > \frac{2r\log\left(\frac{2nr}{\delta}\right)}{n}\right) \le \delta.$$

7.2.8 Intermediate Results

Here we introduce all intermediate results applied in our main analysis.

Lemma 16. Suppose for $A = [A^{(1)}, A^{(2)}, \ldots, A^{(m)}] \in \mathbb{R}^{n \times mp}$, each $A^{(i)} \in \mathbb{R}^{n \times p}$ is a column-wise sub-matrix of A. Given a vector $v \in \mathbb{R}^m$, we have

$$\left\| \sum_{i=1}^{m} A^{(i)} v_i \right\|_2 \leq \|A\|_2 \|v\|_2.$$

Proof. This is an extension of the Cauchy-Schwartz inequality. We have $\sum_{i=1}^{m} A^{(i)} v_i = A(v \otimes I_p)$, where $\otimes$ is the Kronecker product. This implies

$$\left\| \sum_{i=1}^{m} A^{(i)} v_i \right\|_2 = \|A(v \otimes I_p)\|_2 \leq \|A\|_2 \|v \otimes I_p\|_2 = \|A\|_2 \|v\|_2.$$

□

Lemma 17. Given two sequences of unit vectors $\{\phi_i\}_{i=1}^n$ and $\{\psi_i\}_{i=1}^n$, where $\phi_i, \psi_i \in \mathbb{R}^{p_i}$ with $\|\phi_i - \psi_i\|_2 \leq \varepsilon$ for all $i \in [n]$, we have

$$\|\phi_1 \otimes \phi_2 \otimes \cdots \otimes \phi_n - \psi_1 \otimes \psi_2 \otimes \cdots \otimes \psi_n\|_2 \leq (1+\varepsilon)^n - 1.$$

Proof. Suppose for all $i \in [n]$, we have $\psi_i = \phi_1 + x_i$ for some vector $x_i \in \mathbb{R}^{p_i}$. Then we have

$$\begin{aligned}
&\|\phi_1 \otimes \cdots \otimes \phi_n - \psi_1 \otimes \cdots \otimes \psi_n\|_2 = \|\phi_1 \otimes \cdots \otimes \phi_n - (\phi_1 + x_i) \otimes \cdots \otimes (\psi_n + x_n)\|_2 \\
&\leq \sum_{i=1}^{n} \|\phi_1 \otimes \cdots \otimes x_i \otimes \cdots \otimes \phi_n\|_2 + \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \|\phi_1 \otimes \cdots \otimes x_i \otimes \cdots \otimes x_j \otimes \cdots \otimes \phi_n\|_2 \\
&\qquad + \cdots + \|x_1 \otimes \cdots \otimes x_n\|_2 \\
&\leq \begin{pmatrix} n \\ 1 \end{pmatrix} \varepsilon + \begin{pmatrix} n \\ 2 \end{pmatrix} \varepsilon^2 + \cdots + \begin{pmatrix} n \\ n \end{pmatrix} \varepsilon^n = (1+\varepsilon)^n - 1,
\end{aligned}$$

where the last inequality is from the fact that $\|v \otimes u\|_2 = \|v\|_2 \|u\|_2$ for any vectors v and u. □

Lemma 18. Suppose that $A \in \mathbb{R}^{n\times\prod_{d=1}^{2}p_d}$ has leverage scores $\ell_i^2(A)$ for all $i \in [n]$. Then for any $v_1, v_2 \in \mathbb{R}^{p_2}$, the leverage scores of $A^{v_1,v_2} = [A^{v_1}, A^{v_2}] \in \mathbb{R}^{n\times 2p_1}$ are bounded by $\ell_i^2(A^{v_1,v_2}) \le \ell_i^2(A)$.

Proof. Let Z have orthonormal columns and have the same span as the column space of A. Then we have $\ell_i^2(A) = \|e_i^\top Z\|_2^2$ for all $i \in [n]$. Since the column space of A^{v_1,v_2} is a subspace of the column space of A, we can always find a column sub-matrix $Z_1 \in \mathbb{R}^{n\times 2p_1}$ of Z such that Z_1 spans the column space of A^{v_1,v_2}. Therefore, for each $i \in [n]$, we have

$$\ell_i^2(A^{v_1,v_2}) = \|e_i^\top Z_1\|_2^2 \le \|e_i^\top Z\|_2^2 = \ell_i^2(A).$$

□

Lemma 19. Suppose $A \in \mathbb{R}^{n\times\prod_{d=1}^{2}p_d}$ has leverage scores $\ell_i^2(A)$ for all $i \in [n]$. Then for any $v_i^{(r)} \in \mathbb{R}^{p_2}$, $i \in [2]$, $r \in [R]$ with $R \le p_2/2$, the leverage scores of $A^{\left\{v_i^{(r)}\right\}} = \left[A^{v_1^{(1)}}, \ldots, A^{v_1^{(R)}}, A^{v_2^{(1)}}, \ldots, A^{v_2^{(R)}}\right] \in \mathbb{R}^{n\times 2Rp_1}$ are bounded by $\ell_i^2\left(A^{\left\{v_i^{(r)}\right\}}\right) \le \ell_i^2(A)$.

Proof. Let Z have orthonormal columns and have the same span as the column space of A. Then we have $\ell_i^2(A) = \|e_i^\top Z\|_2^2$ for all $i \in [n]$. Since the column space of $A^{\left\{v_i^{(r)}\right\}}$ is a subspace of the column space of A, as the column space of each $A^{v_i^{(r)}}$ is a subspace of the column space of A, we can always find a column sub-matrix $Z_1 \in \mathbb{R}^{n\times 2Rp_1}$ of Z such that Z_1 spans the column space of $A^{\left\{v_i^{(r)}\right\}}$. Therefore, for each $i \in [n]$, we have

$$\ell_i^2\left(A^{\left\{v_i^{(r)}\right\}}\right) = \|e_i^\top Z_1\|_2^2 \le \|e_i^\top Z\|_2^2 = \ell_i^2(A).$$

□

Lemma 20. For any $v_1, v_2 \in \mathcal{B}_{p_2}$, suppose $\langle v_1, v_2\rangle = 0$, and $\overline{v}_1, \overline{v}_2 \in \mathcal{B}_{p_2}$ are vectors such that $\|v_1 - \overline{v}_1\|_2 \le \eta_0$ and $\|v_2 - \overline{v}_2\|_2 \le \eta_0$. Then we have

$$\rho_{\mathrm{Fin}}([A^{v_1}, A^{v_2}], [A^{\overline{v}_1}, A^{\overline{v}_2}]) \le 2\kappa(A)\eta_0.$$

Proof. Denote $A^{v_1,v_2} = [A^{v_1}, A^{v_2}]$. From a perturbation bound for orthogonal projections given in [238], we have

$$\rho_{\text{Fin}}(A^{v_1,v_2}, A^{\overline{v}_1,\overline{v}_2}) \leq \frac{\|A^{v_1,v_2} - A^{\overline{v}_1,\overline{v}_2}\|_2}{\sigma_{\min}(A^{v_1,v_2})}. \tag{7.45}$$

We first provide an upper bound on the numerator as

$$\begin{aligned}\|A^{v_1,v_2} - A^{\overline{v}_1,\overline{v}_2}\|_2 &= \left\| \left[\sum_{i=1}^{p_2} A^{(i)}(v_{1,i} - \overline{v}_{1,i}), \sum_{i=1}^{p_2} A^{(i)}(v_{2,i} - \overline{v}_{2,i}) \right] \right\|_2 \\ &\leq \left\| \sum_{i=1}^{p_2} A^{(i)}(v_{1,i} - \overline{v}_{1,i}) \right\|_2 + \left\| \sum_{i=1}^{p_2} A^{(i)}(v_{2,i} - \overline{v}_{2,i}) \right\|_2 \leq 2\sigma_{\max}(A)\eta_0,\end{aligned} \tag{7.46}$$

where the last inequality is from Lemma 16.

Next, we provide a lower bound on the denominator. Let $[u_1^\top, u_2^\top]^\top$ be a unit vector corresponding to the smallest singular value of A^{v_1,v_2}, where $u_1, u_2 \in \mathbb{R}^{p_1}$. Then we have

$$\begin{aligned}\sigma_{\min}(A^{v_1,v_2}) &= \left\| A^{v_1,v_2} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \right\|_2 = \|A(v_1 \otimes u_1 + v_2 \otimes u_2)\|_2 \\ &\geq \sigma_{\min}(A)\|v_1 \otimes u_1 + v_2 \otimes u_2\|_2 \\ &= \sigma_{\min}(A)\sqrt{\|v_1 \otimes u_1\|_2^2 + \|v_2 \otimes u_2\|_2^2 + 2\langle v_1 \otimes u_1, v_2 \otimes u_2\rangle} \\ &= \sigma_{\min}(A)\sqrt{\|u_1\|_2^2 + \|u_2\|_2^2 + 2\sum_{i=1}^{p_2}\sum_{j=1}^{p_1} v_{1,i}u_{1,j}v_{2,i}u_{2,j}} \\ &= \sigma_{\min}(A)\sqrt{1 + 2\langle v_1, v_2\rangle\langle u_1, u_2\rangle} = \sigma_{\min}(A),\end{aligned} \tag{7.47}$$

where the last equality is from the condition $\langle v_1, v_2\rangle = 0$. We finish the proof by combining (7.45), (7.46), and (7.47).

□

Lemma 21. For all $i \in [2]$ and $r \in [R]$, $v_i^{(r)} \in \mathcal{B}_{p_2}$. Suppose for all $i \in [2]$, $r \in [R]$, $q \in [R]\backslash\{r\}$, we have $\langle v_i^{(r)}, v_i^{(q)}\rangle = \langle v_1^{(r)}, v_2^{(r)}\rangle = 0$. Further suppose for all $i \in [2]$ and $r \in [R]$, $\overline{v}_i^{(r)} \in \mathcal{B}_{p_2}$ is a vector such that $\|v_i^{(r)} - \overline{v}_i^{(r)}\|_2 \leq \eta_0$. Denote $A^{\left\{v_i^{(r)}\right\}} =$

$\left[A^{v_1^{(1)}}, \ldots, A^{v_1^{(R)}}, A^{v_2^{(1)}}, \ldots, A^{v_2^{(R)}}\right]$. Then we have

$$\rho_{\text{Fin}}\left(A^{\left\{v_i^{(r)}\right\}}, A^{\left\{\overline{v}_i^{(r)}\right\}}\right) \leq 2R\kappa(A)\eta_0.$$

Proof. From the perturbation bound for orthogonal projection given in [238], we have

$$\rho_{\text{Fin}}\left(A^{\left\{v_i^{(r)}\right\}}, A^{\left\{\overline{v}_i^{(r)}\right\}}\right) \leq \frac{\left\|A^{\left\{v_i^{(r)}\right\}} - A^{\left\{\overline{v}_i^{(r)}\right\}}\right\|_2}{\sigma_{\min}\left(A^{\left\{v_i^{(r)}\right\}}\right)}. \tag{7.48}$$

We first upper bound the numerator as

$$\begin{aligned}
&\left\|A^{\left\{v_i^{(r)}\right\}} - A^{\left\{\overline{v}_i^{(r)}\right\}}\right\|_2 \\
&= \left\|\left[\sum_{j=1}^{p_2} A_j\left(v_{1,j}^{(1)} - \overline{v}_{1,j}^{(1)}\right), \ldots, \sum_{j=1}^{p_2} A_j\left(v_{1,j}^{(R)} - \overline{v}_{1,j}^{(R)}\right), \sum_{j=1}^{p_2} A_j\left(v_{2,j}^{(1)} - \overline{v}_{2,j}^{(1)}\right), \ldots, \right.\right. \\
&\qquad\qquad\qquad \left.\left. \sum_{j=1}^{p_2} A_j\left(v_{2,j}^{(R} - \overline{v}_{2,j}^{(R)}\right)\right]\right\|_2 \\
&\leq \sum_{r=1}^{R}\left\|\sum_{j=1}^{p_2} A_j\left(v_{1,j}^{(r)} - \overline{v}_{1,j}^{(r)}\right)\right\|_2 + \sum_{r=1}^{R}\left\|\sum_{j=1}^{p_2} A_j\left(v_{2,j}^{(r)} - \overline{v}_{2,j}^{(r)}\right)\right\|_2 \leq 2R\sigma_{\max}(A)\eta_0,
\end{aligned} \tag{7.49}$$

where the last inequality is from Lemma 16.

Next, we provide a lower bound on the denominator. Let $\left[u_1^{(1)\top}, \ldots, u_1^{(R)\top}, u_2^{(1)\top}, \ldots, u_2^{(R)\top}\right]^\top \in \mathbb{R}^{2Rp_1}$ be a unit vector corresponding to the smallest singular value of $A^{\left\{v_i^{(r)}\right\}}$, where $u_i^{(r)} \in \mathbb{R}^{p_1}$ for all $i \in [2]$ and $r \in [R]$. Then we have

$$\begin{aligned}
\sigma_{\min}\left(A^{\left\{v_i^{(r)}\right\}}\right) &= \left\| A^{\left\{v_i^{(r)}\right\}} \left[u_1^{(1)\top}, \dots, u_1^{(R)\top}, u_2^{(1)\top}, \dots, u_2^{(R)\top}\right]^\top \right\|_2 \\
&= \left\| A \cdot \left(\sum_{r=1}^{R} v_1^{(r)} \otimes u_1^{(r)} + v_2^{(r)} \otimes u_2^{(r)} \right) \right\|_2 \\
&\geq \sigma_{\min}(A) \left\| \sum_{r=1}^{R} \left(v_1^{(r)} \otimes u_1^{(r)} + v_2^{(r)} \otimes u_2^{(r)} \right) \right\|_2 \\
&= \sigma_{\min}(A) \sqrt{\begin{array}{l} \sum_{r=1}^{R} \left(\left\| u_1^{(r)} \right\|_2^2 + \left\| u_2^{(r)} \right\|_2^2 \right) + 2 \sum_{r=1}^{R} \sum_{j=1}^{p_2} \sum_{k=1}^{p_1} v_{1,j}^{(r)} u_{1,k}^{(r)} v_{2,j}^{(r)} u_{2,k}^{(r)} \\ \quad + 2 \sum_{i=1}^{2} \sum_{r=1}^{R-1} \sum_{q=r+1}^{R} \sum_{j=1}^{p_2} \sum_{k=1}^{p_1} v_{i,j}^{(r)} u_{i,k}^{(r)} v_{i,j}^{(q)} u_{i,k}^{(q)} \end{array}} \\
&= \sigma_{\min}(A) \sqrt{1 + 2 \sum_{r=1}^{R} \langle v_1^{(r)}, v_2^{(r)} \rangle \langle u_1^{(r)}, u_2^{(r)} \rangle + 2 \sum_{i=1}^{2} \sum_{r=1}^{R-1} \sum_{q=r+1}^{R} \langle v_i^{(r)}, v_i^{(q)} \rangle \langle u_i^{(r)}, u_i^{(q)} \rangle} \\
&= \sigma_{\min}(A), \qquad (7.50)
\end{aligned}$$

where the last equality uses the conditions that for all $i \in [2]$ and $r \in [R]$, $\langle v_i^{(r)}, v_i^{(q)} \rangle = \langle v_1^{(r)}, v_2^{(r)} \rangle = 0$ for $q \in [R] \backslash \{r\}$. We finish the proof by combining (7.48), (7.49), and (7.50).

□

Lemma 22. For all $d \in [D] \backslash \{1\}$, $\theta_d, \phi_d \in \mathcal{B}_{p_d}$. Suppose there exists an $i \in [D] \backslash \{1\}$ such that $\langle \theta_i, \phi_i \rangle = 0$. Further suppose for all $d \in [D] \backslash \{1\}$, $\overline{\theta}_d, \overline{\phi}_d \in \mathcal{B}_{p_d}$ are vectors such that $\|\theta_d - \overline{\theta}_d\|_2 \leq \eta_0$ and $\|\phi_d - \overline{\phi}_d\|_2 \leq \eta_0$. Then we have

$$\rho_{\text{Fin}}\left(\left[A^{\{\theta_{\backslash 1}\}}, A^{\{\phi_{\backslash 1}\}} \right], \left[A^{\{\overline{\theta}_{\backslash 1}\}}, A^{\{\overline{\phi}_{\backslash 1}\}} \right] \right) \leq 2\kappa(A) \left((1 + \eta_0)^{D-1} - 1 \right).$$

Proof. Let $A^{\vartheta_{\backslash 1}, \varphi_{\backslash 1}} = \left[A^{\{\theta_{\backslash 1}\}}, A^{\{\phi_{\backslash 1}\}} \right] \in \mathbb{R}^{n \times 2p_1}$. From the perturbation bound for orthogonal projection given in [238], we have

$$\rho_{\text{Fin}}\left(A^{\vartheta_{\backslash 1}, \varphi_{\backslash 1}}, A^{\overline{\vartheta}, \overline{\varphi}} \right) \leq \frac{\left\| A^{\vartheta_{\backslash 1}, \varphi_{\backslash 1}} - A^{\overline{\vartheta}, \overline{\varphi}} \right\|_2}{\sigma_{\min}(A^{\vartheta_{\backslash 1}, \varphi_{\backslash 1}})}. \qquad (7.51)$$

We denote $\sum_{j_2 \cdots j_D} = \sum_{j_D=1}^{p_D} \cdots \sum_{j_2=1}^{p_2}$. We first provide an upper bound on the numerator:

$$
\begin{aligned}
&\left\| A^{\vartheta_{\backslash 1}, \varphi_{\backslash 1}} - A^{\overline{\vartheta}, \overline{\varphi}} \right\|_2 \\
&= \left\| \left[\sum_{j_2 \cdots j_D} A^{(j_D, \ldots, j_2)} \cdot \left(\theta_{D,j_D} \cdots \theta_{2,j_2} - \overline{\theta}_{D,j_D} \cdots \overline{\theta}_{2,j_2} \right), \right.\right. \\
&\qquad \left.\left. \sum_{j_2 \cdots j_D} A^{(j_D, \ldots, j_2)} \cdot \left(\phi_{D,j_D} \cdots \phi_{2,j_2} - \overline{\phi}_{D,j_D} \cdots \overline{\phi}_{2,j_2} \right) \right] \right\|_2 \\
&\leq \left\| \sum_{j_2 \cdots j_D} A^{(j_D, \ldots, j_2)} \cdot \left(\theta_{D,j_D} \cdots \theta_{2,j_2} - \overline{\theta}_{D,j_D} \cdots \overline{\theta}_{2,j_2} \right) \right\|_2 \\
&\qquad + \left\| \sum_{j_2 \cdots j_D} A^{(j_D, \ldots, j_2)} \cdot \left(\phi_{D,j_D} \cdots \phi_{2,j_2} - \overline{\phi}_{D,j_D} \cdots \overline{\phi}_{2,j_2} \right) \right\|_2 \\
&\leq \sigma_{\max}(A) \cdot \left(\left\| \theta_D \otimes \cdots \otimes \theta_2 - \overline{\theta}_D \otimes \cdots \otimes \overline{\theta}_2 \right\|_2 + \left\| \phi_D \otimes \cdots \otimes \phi_2 - \overline{\phi}_D \otimes \cdots \otimes \overline{\phi}_2 \right\|_2 \right) \\
&\leq 2\sigma_{\max}(A) \left((1+\eta_0)^{D-1} - 1 \right), \qquad (7.52)
\end{aligned}
$$

where the second inequality is from Lemma 16 and the last inequality is from Lemma 17.

Next, we provide a lower bound on the denominator. Let $[u_1^\top, u_2^\top]^\top$ be a unit vector corresponding to the smallest singular value of $A^{\vartheta_{\backslash 1}, \varphi_{\backslash 1}}$, where $u_1, u_2 \in \mathbb{R}^{p_1}$. Then we have

$$
\begin{aligned}
&\sigma_{\min}\left(A^{\vartheta_{\backslash 1}, \varphi_{\backslash 1}} \right) = \left\| A^{\vartheta_{\backslash 1}, \varphi_{\backslash 1}} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \right\|_2 = \left\| A \left(\theta_D \otimes \cdots \otimes \theta_2 \otimes u_1 + \phi_D \otimes \cdots \otimes \phi_2 \otimes u_2 \right) \right\|_2 \\
&\geq \sigma_{\min}(A) \| \theta_D \otimes \cdots \otimes \theta_2 \otimes u_1 + \phi_D \otimes \cdots \otimes \phi_2 \otimes u_2 \|_2 \\
&= \sigma_{\min}(A) \sqrt{ \| \theta_D \otimes \cdots \otimes \theta_2 \otimes u_1 \|_2^2 + \| \phi_D \otimes \cdots \otimes \phi_2 \otimes u_2 \|_2^2 } \\
&\qquad \overline{+2 \langle \theta_D \otimes \cdots \otimes \theta_2 \otimes u_1, \phi_D \otimes \cdots \otimes \phi_2 \otimes u_2 \rangle} \\
&= \sigma_{\min}(A) \sqrt{ \| u_1 \|_2^2 + \| u_2 \|_2^2 + 2 \sum_{j_2 \cdots j_D} \sum_{j_1=1}^{p_1} \theta_{D,j_D} \cdots \theta_{2,j_2} u_{1,j_1} \cdot \phi_{D,j_D} \cdots \phi_{2,j_2} u_{2,j_1} } \\
&= \sigma_{\min}(A) \sqrt{ 1 + 2 \langle \theta_D, \phi_D \rangle \cdots \langle \theta_2, \phi_2 \rangle \langle u_1, u_2 \rangle } = \sigma_{\min}(A), \qquad (7.53)
\end{aligned}
$$

where the last inequality is from $\langle \theta_i, \phi_i \rangle = 0$ for some $i \in \{2, \ldots, D\}$. We finish the proof by combining (7.51), (7.52) and (7.53).

□

Lemma 23. For all $d \in [D]\backslash\{1\}$ and $r \in [R]$, $\theta_d^{(r)}, \phi_d^{(r)} \in \mathcal{B}_{p_d}$. Suppose that for any $r, q \in [R]$, there exists an $i \in [D]\backslash\{1\}$ such that $\langle \theta_i^{(r)}, \phi_i^{(q)} \rangle = 0$, and further, for all $r \in [R-1]$, $q \in [R]\backslash[r]$, there exist $j, k \in [D]\backslash\{1\}$ such that $\langle \theta_j^{(r)}, \theta_j^{(q)} \rangle = 0$ and $\langle \phi_k^{(r)}, \phi_k^{(q)} \rangle = 0$. Further suppose for all $d \in [D]\backslash\{1\}$ and $r \in [R]$, $\overline{\theta}_d^{(r)}, \overline{\phi}_d^{(r)} \in \mathcal{B}_{p_d}$ are vectors such that $\|\theta_d^{(r)} - \overline{\theta}_d^{(r)}\|_2 \leq \eta_0$ and $\|\phi_d^{(r)} - \overline{\phi}_d^{(r)}\|_2 \leq \eta_0$. Then we have

$$\rho_{\text{Fin}}\left(\left[A^{\left\{\theta_{\backslash 1}^{(r)}\right\}}, A^{\left\{\phi_{\backslash 1}^{(r)}\right\}}\right], \left[A^{\left\{\overline{\theta}_{\backslash 1}^{(r)}\right\}}, A^{\left\{\overline{\phi}_{\backslash 1}^{(r)}\right\}}\right]\right) \leq 2R\kappa(A)\left((1+\eta_0)^{D-1} - 1\right).$$

Proof. Denote $A^{\left\{\vartheta_{\backslash 1}^{(r)}, \varphi_{\backslash 1}^{(r)}\right\}} = \left[A^{\left\{\theta_{\backslash 1}^{(r)}\right\}}, A^{\left\{\phi_{\backslash 1}^{(r)}\right\}}\right] \in \mathbb{R}^{n \times 2Rp_1}$. From the perturbation bound on orthogonal projection given in [238], we have

$$\rho_{\text{Fin}}\left(A^{\left\{\vartheta_{\backslash 1}^{(r)}, \varphi_{\backslash 1}^{(r)}\right\}}, A^{\left\{\overline{\vartheta}_{\backslash 1}^{(r)}, \overline{\varphi}_{\backslash 1}^{(r)}\right\}}\right) \leq \frac{\left\|A^{\left\{\vartheta_{\backslash 1}^{(r)}, \varphi_{\backslash 1}^{(r)}\right\}} - A^{\left\{\overline{\vartheta}_{\backslash 1}^{(r)}, \overline{\varphi}_{\backslash 1}^{(r)}\right\}}\right\|_2}{\sigma_{\min}\left(A^{\left\{\vartheta_{\backslash 1}^{(r)}, \varphi_{\backslash 1}^{(r)}\right\}}\right)}. \tag{7.54}$$

We denote $\sum_{j_2\cdots j_D} = \sum_{j_D=1}^{p_D}\cdots\sum_{j_2=1}^{p_2}$. We first upper bound the numerator as

$$
\begin{aligned}
&\left\| A^{\left\{\vartheta_{\setminus 1}^{(r)}, \varphi_{\setminus 1}^{(r)}\right\}} - A^{\left\{\overline{\vartheta}_{\setminus 1}^{(r)}, \overline{\varphi}_{\setminus 1}^{(r)}\right\}} \right\|_2 \\
&= \left\| \left[\sum_{j_2\cdots j_D} A^{(j_D,\ldots,j_2)} \cdot \left(\theta_{D,j_D}^{(1)} \cdots \theta_{2,j_2}^{(1)} - \overline{\theta}_{D,j_D}^{(1)} \cdots \overline{\theta}_{2,j_2}^{(1)}\right), \ldots, \right.\right. \\
&\qquad \sum_{j_2\cdots j_D} A^{(j_D,\ldots,j_2)} \cdot \left(\theta_{D,j_D}^{(R)} \cdots \theta_{2,j_2}^{(R)} - \overline{\theta}_{D,j_D}^{(R)} \cdots \overline{\theta}_{2,j_2}^{(R)}\right), \\
&\qquad \sum_{j_2\cdots j_D} A^{(j_D,\ldots,j_2)} \cdot \left(\phi_{D,j_D}^{(1)} \cdots \phi_{2,j_2}^{(1)} - \overline{\phi}_{D,j_D}^{(1)} \cdots \overline{\phi}_{2,j_2}^{(1)}\right), \ldots, \\
&\qquad \left.\left. \sum_{j_2\cdots j_D} A^{(j_D,\ldots,j_2)} \cdot \left(\phi_{D,j_D}^{(R)} \cdots \phi_{2,j_2}^{(R)} - \overline{\phi}_{D,j_D}^{(R)} \cdots \overline{\phi}_{2,j_2}^{(R)}\right) \right] \right\|_2 \\
&\leq \sum_{r=1}^{R} \left\| \sum_{j_2\cdots j_D} A^{(j_D,\ldots,j_2)} \cdot \left(\theta_{D,j_D}^{(r)} \cdots \theta_{2,j_2}^{(r)} - \overline{\theta}_{D,j_D}^{(r)} \cdots \overline{\theta}_{2,j_2}^{(r)}\right) \right\|_2 \\
&\qquad + \left\| \sum_{j_2\cdots j_D} A^{(j_D,\ldots,j_2)} \cdot \left(\phi_{D,j_D}^{(r)} \cdots \phi_{2,j_2}^{(r)} - \overline{\phi}_{D,j_D}^{(r)} \cdots \overline{\phi}_{2,j_2}^{(r)}\right) \right\|_2 \\
&\leq \sigma_{\max}(A) \cdot \left(\sum_{r=1}^{R} \left\| \theta_D^{(r)} \otimes \cdots \otimes \theta_2^{(r)} - \overline{\theta}_D^{(r)} \otimes \cdots \otimes \overline{\theta}_2^{(r)} \right\|_2 \right. \\
&\qquad \left. + \left\| \phi_D^{(r)} \otimes \cdots \otimes \phi_2^{(r)} - \overline{\phi}_D^{(r)} \otimes \cdots \otimes \overline{\phi}_2^{(r)} \right\|_2 \right) \\
&\leq 2R\sigma_{\max}(A)\left((1+\eta_0)^{D-1} - 1\right), \qquad (7.55)
\end{aligned}
$$

where the second inequality is from Lemma 16 and the last inequality is from Lemma 17.

Next, we lower bound the denominator. Let $\left[u_1^{(1)\top}, \ldots, u_1^{(R)\top}, u_2^{(1)\top}, \ldots, u_2^{(R)\top}\right]^\top \in \mathbb{R}^{2Rp_1}$ be a unit vector corresponding to the smallest singular value of $A^{\left\{\vartheta_{\setminus 1}^{(r)}, \varphi_{\setminus 1}^{(r)}\right\}}$, where

$u_i^{(r)} \in \mathbb{R}^{p_1}$ for all $i \in [2]$ and $r \in [R]$. Then we have

$$
\begin{aligned}
&\sigma_{\min}\left(A^{\left\{\vartheta_{\backslash 1}^{(r)}, \varphi_{\backslash 1}^{(r)}\right\}}\right) = \left\| A^{\left\{\vartheta_{\backslash 1}^{(r)}, \varphi_{\backslash 1}^{(r)}\right\}} \left[u_1^{(1)\top}, \dots, u_1^{(R)\top}, u_2^{(1)\top}, \dots, u_2^{(R)\top}\right]^\top \right\|_2 \\
&= \left\| A \cdot \left(\sum_{r=1}^{R} \theta_D^{(r)} \otimes \cdots \otimes \theta_2^{(r)} \otimes u_1^{(r)} + \phi_D^{(r)} \otimes \cdots \otimes \phi_2^{(r)} \otimes u_2^{(r)} \right) \right\|_2 \\
&\geq \sigma_{\min}(A) \left\| \sum_{r=1}^{R} \theta_D^{(r)} \otimes \cdots \otimes \theta_2^{(r)} \otimes u_1^{(r)} + \phi_D^{(r)} \otimes \cdots \otimes \phi_2^{(r)} \otimes u_2^{(r)} \right\|_2 \\
&= \sigma_{\min}(A) \sqrt{\sum_{r=1}^{R} \left(\left\| u_1^{(r)} \right\|_2^2 + \left\| u_2^{(r)} \right\|_2^2 \right) + 2 \sum_{r=1}^{R} \sum_{q=1}^{R} \sum_{j_1 \cdots j_D} \theta_{D,j_D}^{(r)} \cdots \theta_{2,j_2}^{(r)} u_{1,j_1}^{(r)} \cdot \phi_{D,j_D}^{(q)} \cdots \phi_{2,j_2}^{(q)} u_{2,j_1}^{(q)} + 2 \sum_{r=1}^{R-1} \sum_{q=r+1}^{R} \sum_{j_1 \cdots j_D} \left(\theta_{D,j_D}^{(r)} \cdots \theta_{2,j_2}^{(r)} u_{1,j_1}^{(r)} \cdot \theta_{D,j_D}^{(q)} \cdots \theta_{2,j_2}^{(q)} u_{1,j_1}^{(q)} + \phi_{D,j_D}^{(r)} \cdots \phi_{2,j_2}^{(r)} u_{2,j_1}^{(r)} \cdot \phi_{D,j_D}^{(q)} \cdots \phi_{2,j_2}^{(q)} u_{2,j_1}^{(q)} \right)} \\
&= \sigma_{\min}(A) \sqrt{1 + 2 \sum_{r=1}^{R} \sum_{q=1}^{R} \langle \theta_D^{(r)}, \phi_D^{(q)} \rangle \cdots \langle \theta_2^{(r)}, \phi_2^{(q)} \rangle \langle u_1^{(r)}, u_2^{(q)} \rangle + 2 \sum_{r=1}^{R-1} \sum_{q=r+1}^{R} \left(\langle \theta_D^{(r)}, \theta_D^{(q)} \rangle \cdots \langle \theta_2^{(r)}, \theta_2^{(q)} \rangle \langle u_1^{(r)}, u_1^{(q)} \rangle + \langle \phi_D^{(r)}, \phi_D^{(q)} \rangle \cdots \langle \phi_2^{(r)}, \phi_2^{(q)} \rangle \langle u_2^{(r)}, u_2^{(q)} \rangle \right)} \\
&= \sigma_{\min}(A), \qquad (7.56)
\end{aligned}
$$

where the last inequality is from the conditions on $\theta_d^{(r)}$ and $\phi_d^{(r)}$. We finish the proof by combining (7.54), (7.55), and (7.56). □

Lemma 24. For all $d \in [D]\backslash\{1\}$ and $r \in [R]$, $\theta_d^{(r_d)}, \phi_d^{(r_d)} \in \mathcal{B}_{p_d}$. Suppose that for any $r_d, q_d \in [R_d]$, $d \in [R]\backslash\{1\}$, there exists an $i \in [D]\backslash\{1\}$ such that $\langle \theta_i^{(r)}, \phi_i^{(q)} \rangle = 0$, and for all $r \in [R-1]$, $q \in [R]\backslash[r]$, there exist $j, k \in [D]\backslash\{1\}$ such that $\langle \theta_j^{(r)}, \theta_j^{(q)} \rangle = 0$ and $\langle \phi_k^{(r)}, \phi_k^{(q)} \rangle = 0$. Further suppose for all $d \in [D]\backslash\{1\}$ and $r \in [R]$, $\overline{\theta}_d^{(r)}, \overline{\phi}_d^{(r)} \in \mathcal{B}_{p_d}$ are vectors such that $\|\theta_d^{(r)} - \overline{\theta}_d^{(r)}\|_2 \leq \eta_0$ and $\|\phi_d^{(r)} - \overline{\phi}_d^{(r)}\|_2 \leq \eta_0$. Then we have

$$
\rho_{\text{Fin}}\left(\left[A^{\left\{\theta_{\backslash 1}^{(r)}\right\}}, A^{\left\{\phi_{\backslash 1}^{(r)}\right\}} \right], \left[A^{\left\{\overline{\theta}_{\backslash 1}^{(r)}\right\}}, A^{\left\{\overline{\phi}_{\backslash 1}^{(r)}\right\}} \right] \right) \leq 2\kappa(A) \left((1+\eta_0)^{D-1} - 1 \right) \sqrt{\prod_{d=2}^{D} R_d}.
$$

Proof. Denote $A^{\left\{\vartheta_{\setminus 1}^{\{r_d\}},\varphi_{\setminus 1}^{\{r_d\}}\right\}} = \left[A^{\left\{\theta_{\setminus 1}^{\{r_d\}}\right\}}, A^{\left\{\phi_{\setminus 1}^{\{r_d\}}\right\}}\right] \in \mathbb{R}^{n\times 2R_1p_1}$. From the perturbation bound for orthogonal projection given in [238], we have

$$\rho_{\text{Fin}}\left(A^{\left\{\vartheta_{\setminus 1}^{\{r_d\}},\varphi_{\setminus 1}^{\{r_d\}}\right\}}, A^{\left\{\overline{\vartheta}_{\setminus 1}^{\{r_d\}},\overline{\varphi}_{\setminus 1}^{\{r_d\}}\right\}}\right) \leq \frac{\left\|A^{\left\{\vartheta_{\setminus 1}^{\{r_d\}},\varphi_{\setminus 1}^{\{r_d\}}\right\}} - A^{\left\{\overline{\vartheta}_{\setminus 1}^{\{r_d\}},\overline{\varphi}_{\setminus 1}^{\{r_d\}}\right\}}\right\|_2}{\sigma_{\min}\left(A^{\left\{\vartheta_{\setminus 1}^{\{r_d\}},\varphi_{\setminus 1}^{\{r_d\}}\right\}}\right)}. \quad (7.57)$$

We denote $\sum_{j_2 \cdots j_D} = \sum_{j_D=1}^{p_D} \cdots \sum_{j_2=1}^{p_2}$. We first upper bound the numerator as

$$\begin{aligned}
&\left\| A^{\left\{\vartheta_{\setminus 1}^{\{r_d\}}, \varphi_{\setminus 1}^{\{r_d\}}\right\}} - A^{\left\{\overline{\vartheta}_{\setminus 1}^{\{r_d\}}, \overline{\varphi}_{\setminus 1}^{\{r_d\}}\right\}} \right\|_2 \\
&= \Bigg\| \Bigg[\sum_{j_2 \cdots j_D} A^{(j_D, \ldots, j_2)} \cdot \left(\theta_{D,j_D}^{(1)} \cdots \theta_{2,j_2}^{(1)} - \overline{\theta}_{D,j_D}^{(1)} \cdots \overline{\theta}_{2,j_2}^{(1)} \right), \ldots, \\
&\qquad \sum_{j_2 \cdots j_D} A^{(j_D, \ldots, j_2)} \cdot \left(\theta_{D,j_D}^{(1)} \cdots \theta_{2,j_2}^{(R_2)} - \overline{\theta}_{D,j_D}^{(1)} \cdots \overline{\theta}_{2,j_2}^{(R_2)} \right), \ldots, \\
&\qquad \sum_{j_2 \cdots j_D} A^{(j_D, \ldots, j_2)} \cdot \left(\phi_{D,j_D}^{(R_D)} \cdots \phi_{2,j_2}^{(1)} - \overline{\phi}_{D,j_D}^{(R_D)} \cdots \overline{\phi}_{2,j_2}^{(1)} \right), \ldots, \\
&\qquad \sum_{j_2 \cdots j_D} A^{(j_D, \ldots, j_2)} \cdot \left(\phi_{D,j_D}^{(R_D)} \cdots \phi_{2,j_2}^{(R_2)} - \overline{\phi}_{D,j_D}^{(R_D)} \cdots \overline{\phi}_{2,j_2}^{(R_2)} \right) \Bigg] \Bigg\|_2 \\
&\leq \sum_{r_2=1}^{R_2} \cdots \sum_{r_D=1}^{R_D} \left\| \sum_{j_2 \cdots j_D} A^{(j_D, \ldots, j_2)} \cdot \left(\theta_{D,j_D}^{(r_D)} \cdots \theta_{2,j_2}^{(r_2)} - \overline{\theta}_{D,j_D}^{(r_D)} \cdots \overline{\theta}_{2,j_2}^{(r_2)} \right) \right\|_2 \\
&\qquad\qquad + \left\| \sum_{j_2 \cdots j_D} A^{(j_D, \ldots, j_2)} \cdot \left(\phi_{D,j_D}^{(r_D)} \cdots \phi_{2,j_2}^{(r_2)} - \overline{\phi}_{D,j_D}^{(r_D)} \cdots \overline{\phi}_{2,j_2}^{(r_2)} \right) \right\|_2 \\
&\leq \sigma_{\max}(A) \cdot \Bigg(\sum_{r_2=1}^{R_2} \cdots \sum_{r_D=1}^{R_D} \left\| \theta_D^{(r_D)} \otimes \cdots \otimes \theta_2^{(r_2)} - \overline{\theta}_D^{(r_D)} \otimes \cdots \otimes \overline{\theta}_2^{(r_2)} \right\|_2 \\
&\qquad\qquad + \left\| \phi_D^{(r_D)} \otimes \cdots \otimes \phi_2^{(r_2)} - \overline{\phi}_D^{(r_D)} \otimes \cdots \otimes \overline{\phi}_2^{(r_2)} \right\|_2 \Bigg) \\
&\leq 2 \prod_{d=2}^{D} R_d \cdot \sigma_{\max}(A) \left((1 + \eta_0)^{D-1} - 1 \right), \qquad (7.58)
\end{aligned}$$

where the second inequality is from Lemma 16 and the last inequality is from Lemma 17.

Next, we provide a lower bound on the denominator. Let $\left[u_1^{(1)\top}, \ldots, u_1^{(R_1)\top}, u_2^{(1)\top}, \ldots, u_2^{(R_1)\top} \right]^\top \in \mathbb{R}^{2R_1 p_1}$ be a unit vector corresponding to the smallest singular value of $A^{\left\{\vartheta_{\setminus 1}^{\{r_d\}}, \varphi_{\setminus 1}^{\{r_d\}}\right\}}$, where $u_i^{(r_1)} \in \mathbb{R}^{p_1}$ for all $i \in [2]$ and

$r_1 \in [R_1]$. Denote $\sum_{r_1,\ldots,r_D} = \sum_{r_1=1}^{R_1} \cdots \sum_{r_D=1}^{R_D}$. Then we have

$$
\begin{aligned}
&\sigma_{\min}\left(A^{\left\{\vartheta_{\backslash 1}^{\{r_d\}},\varphi_{\backslash 1}^{\{r_d\}}\right\}}\right) = \left\|A^{\left\{\vartheta_{\backslash 1}^{\{r_d\}},\varphi_{\backslash 1}^{\{r_d\}}\right\}}\left[u_1^{(1)\top},\ldots,u_1^{(R_1)\top},u_2^{(1)\top},\ldots,u_2^{(R_1)\top}\right]^\top\right\|_2 \\
&= \left\|A\cdot\left(\sum_{r_1,\ldots,r_D}\theta_D^{(r_D)}\otimes\cdots\otimes\theta_2^{(r_2)}\otimes u_1^{(r_1)} + \phi_D^{(r_D)}\otimes\cdots\otimes\phi_2^{(r_2)}\otimes u_2^{(r_1)}\right)\right\|_2 \\
&\geq \sigma_{\min}(A)\left\|\sum_{r_1,\ldots,r_D}\theta_D^{(r_D)}\otimes\cdots\otimes\theta_2^{(r_2)}\otimes u_1^{(r_1)} + \phi_D^{(r_D)}\otimes\cdots\otimes\phi_2^{(r_2)}\otimes u_2^{(r_1)}\right\|_2 \\
&= \sigma_{\min}(A)\sqrt{\sum_{r_1,\ldots,r_D}\left(\left\|u_1^{(r_1)}\right\|_2^2+\left\|u_2^{(r_1)}\right\|_2^2\right)+2\sum_{r_1,\ldots,r_D}\sum_{q_1,\ldots,q_D}\sum_{j_1\cdots j_D}\theta_{D,j_D}^{(r_D)}\cdots\theta_{2,j_2}^{(r_2)}u_{1,j_1}^{(r_1)}} \\
&\overline{\cdot\phi_{D,j_D}^{(q_D)}\cdots\phi_{2,j_2}^{(q_2)}u_{2,j_1}^{(q_1)}+\sum_{r_1,\ldots,r_D}\sum_{q_1,\ldots,q_D}\sum_{j_1\cdots j_D}\left(\theta_{D,j_D}^{(r_D)}\cdots\theta_{2,j_2}^{(r_2)}u_{1,j_1}^{(r_1)}\cdot\theta_{D,j_D}^{(q_D)}\cdots\theta_{2,j_2}^{(q_2)}u_{1,j_1}^{(q_1)}\right.} \\
&\overline{\left.+\phi_{D,j_D}^{(r_D)}\cdots\phi_{2,j_2}^{(r_2)}u_{2,j_1}^{(r_1)}\cdot\phi_{D,j_D}^{(q_D)}\cdots\phi_{2,j_2}^{(q_2)}u_{2,j_1}^{(q_1)}\right)} \\
&= \sigma_{\min}(A)\sqrt{\prod_{d=2}^{D}R_d+2\sum_{r_1,\ldots,r_D}\sum_{q_1,\ldots,q_D}\langle\theta_D^{(r_D)},\phi_D^{(q_D)}\rangle\cdots\langle\theta_2^{(r_2)},\phi_2^{(q_2)}\rangle\langle u_1^{(r_1)},u_2^{(q_1)}\rangle} \\
&\overline{+\sum_{r_1,\ldots,r_D}\sum_{q_1,\ldots,q_D}\left(\langle\theta_D^{(r_D)},\theta_D^{(q_D)}\rangle\cdots\langle\theta_2^{(r_2)},\theta_2^{(q_2)}\rangle\langle u_1^{(r_1)},u_1^{(q_1)}\rangle+\langle\phi_D^{(r_D)},\phi_D^{(q_D)}\rangle\cdots\right.} \\
&\overline{\left(\langle\phi_2^{(r_2)},\phi_2^{(q_2)}\rangle\langle u_1^{(r_1)},u_2^{(q_1)}\rangle\right.} \\
&= \sigma_{\min}(A)\sqrt{\prod_{d=2}^{D}R_d}, \qquad (7.59)
\end{aligned}
$$

where the last inequality is from the conditions on $\theta_d^{(r)}$ and $\phi_d^{(r)}$. We finish the proof by combining (7.57), (7.58), and (7.59).

□

7.3 Proofs for Chapter 4

7.3.1 Proof of Lemma 2

Part 1. We first show the claim on λ. By $y = X\theta^* + \epsilon$ and (7.64), we have

$$\nabla \mathcal{L}(\theta^*) = \frac{X^\top (X\theta^* - y)}{\sqrt{n}\|y - X\theta^*\|_2} = -\frac{X^\top \epsilon}{\sqrt{n}\|\epsilon\|_2}. \tag{7.60}$$

Since ϵ has i.i.d. sub-Gaussian entries with $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i^2] = \sigma^2$ for all $i = 1, \ldots, n$, then from [239] we have

$$\mathbb{P}\left[\|\epsilon\|_2^2 \leq \frac{1}{4}n\sigma^2\right] \leq \exp\left(-\frac{n}{32}\right), \tag{7.61}$$

By [134], we have the following result.

Lemma 25. Assume X satisfies $\|\mathbf{x}_j\|_2 \leq \sqrt{n}$ for all $j \in \{1, \ldots, d\}$ and ϵ has i.i.d. zero-mean sub-Gaussian entries with $\mathbb{E}[w_i^2] = \sigma^2$ for all $i = 1, \ldots, n$, then we have $\mathbb{P}\left[\frac{1}{n}\|X^\top \epsilon\|_\infty \geq 2\sigma\sqrt{\frac{\log d}{n}}\right] \leq 2d^{-1}$.

Combining (7.60), (7.61) and Lemma 25, we have $\|\nabla \mathcal{L}(\theta^*)\|_\infty \leq 4\sqrt{\log d/n}$ with probability at least $1 - 2d^{-1} - \exp\left(-\frac{n}{32}\right)$.

Part 2. Next, we show that LRSC, LRSS, and LRHS holds. First, for correlated sub-Gaussian random design with the covariance satisfying the bounded eigenvalues, we have from [240] that the design matrix X satisfies the RE condition with high probability given $n \geq cs^* \log d$, i.e.,

$$\begin{aligned} \psi_{\min}\|v\|_2^2 - \varphi_{\min}\frac{\log d}{n}\|v\|_1^2 &\leq \frac{\|Xv\|_2^2}{n} \\ \psi_{\max}\|v\|_2^2 + \varphi_{\max}\frac{\log d}{n}\|v\|_1^2 &\geq \frac{\|Xv\|_2^2}{n}, \end{aligned} \tag{7.62}$$

where $\psi_{\min}, \psi_{\max}, \varphi_{\min}, \varphi_{\max} \in (0, \infty)$ are generic constants. The RE condition has been extensively studied for sparse recovery [133, 166, 241].

We divide the proof into three steps.

Step 1. When X satisfies the RE condition, i.e.

$$\psi_{\min}\|v\|_2^2 - \varphi_{\min}\frac{\log d}{n}\|v\|_1^2 \leq \frac{\|Xv\|_2^2}{n} \leq \psi_{\max}\|v\|_2^2 + \varphi_{\max}\frac{\log d}{n}\|\boldsymbol{v}\|_1^2,$$

Denote $s = s^* + 2\tilde{s}$. Since $\|v\|_0 \leq s$, which implies $\|v\|_1^2 \leq s\|v\|_2^2$, then we have

$$\left(\psi_{\min} - \varphi_{\min}\frac{s\log d}{n}\right)\|v\|_2^2 \leq \frac{\|Xv\|_2^2}{n} \leq \left(\psi_{\max} + \varphi_{\max}\frac{s\log d}{n}\right)\|v\|_2^2,$$

Then there exists a universal constant c_1 such that if $n \geq c_1 s^* \log d$, we have

$$\frac{1}{2}\psi_{\min}\|v\|_2^2 \leq \frac{\|Xv\|_2^2}{n} \leq 2\psi_{\max}\|v\|_2^2. \tag{7.63}$$

Step 2. Conditioning on (7.63), we show that $\mathcal{L}$ satisfies LRSC and LRSS with high probability. The gradient of $\mathcal{L}(\theta)$ is

$$\nabla\mathcal{L}(\theta) = \frac{1}{\sqrt{n}}\left(\left(\frac{\partial\|y - X\theta\|_2}{\partial(y - X\theta)}\right)^\top \left(\frac{\partial(y - X\theta)}{\partial\theta}\right)^\top\right)^\top = \frac{X^\top(X\theta - y)}{\sqrt{n}\|y - X\theta\|_2}. \tag{7.64}$$

The Hessian of $\mathcal{L}(\theta)$ is

$$\nabla^2\mathcal{L}(\theta) = \frac{1}{n}\frac{\partial(-X^\top\tilde{z})}{\partial\theta} = \frac{1}{\sqrt{n}\|y - X\theta\|_2}X^\top\left(\mathbf{I} - \frac{(y - X\theta)(y - X\theta)^\top}{\|y - X\theta\|_2^2}\right)X. \tag{7.65}$$

For notational convenience, we define $\Delta = v - w$ for any $v, w \in \mathcal{B}_s^*$. Also denote the residual of the first order Taylor expansion as $\delta\mathcal{L}(w + \Delta, w) = \mathcal{L}(w + \Delta) - \mathcal{L}(w) - \nabla\mathcal{L}(w)^\top\Delta$. Using the first order Taylor expansion of $\mathcal{L}(\theta)$ at w and the Hessian of $\mathcal{L}(\theta)$ in (7.65), we have from mean value theorem that there exists some $\alpha \in [0, 1]$ such that $\delta\mathcal{L}(w+\Delta, w) = \frac{1}{\sqrt{n}\|\xi\|_2}\Delta^\top X^\top\left(\mathbf{I} - \frac{\xi\xi^\top}{\|\xi\|_2^2}\right)X\Delta$, where $\xi = y - X(w+\alpha\Delta)$. For notational simplicity, let's denote $\dot{z} = X(v - \theta^*)$ and $\ddot{z} = X(w - \theta^*)$, which can be considered as two fixed vectors in $\mathbb{R}^n$. Without loss of generality, assume $\|\dot{z}\|_2 \leq \|\ddot{z}\|_2$. Then we have

$$\|\dot{z}\|_2^2 \leq \|\ddot{z}\|_2^2 \leq 2\psi_{\max}n\|w - \theta^*\|_2^2 \leq \frac{n\sigma^2}{4}.$$

Further, we have

$$\xi = y - X(w + \alpha\Delta) = \epsilon - X(w + \alpha\Delta - \theta^*) = \epsilon - \alpha\dot{z} - (1-\alpha)\ddot{z}, \quad \text{and} \quad X\Delta = \dot{z} - \ddot{z}.$$

We have from [239] that

$$\mathbb{P}\left[\|\epsilon\|_2^2 \leq n\sigma^2(1-\delta)\right] \leq \exp\left(-\frac{n\delta^2}{16}\right), \tag{7.66}$$

Then by taking $\delta = 1/3$ in (7.66), we have with probability $1 - \exp\left(-\frac{n}{144}\right)$,

$$\|\xi\|_2 \geq \|\epsilon\|_2 - \alpha\|\dot{z}\|_2 - (1-\alpha)\|\ddot{z}\|_2 \geq \|\epsilon\|_2 - \|\ddot{z}\|_2 \geq \frac{4}{5}\sqrt{n}\sigma - \frac{1}{2}\sqrt{n}\sigma \geq \frac{1}{4}\sqrt{n}\sigma. \tag{7.67}$$

We first discuss the RSS property. From (7.67), we have

$$\begin{aligned} \delta\mathcal{L}(w+\Delta, w) &= \frac{\Delta^\top X^\top \left(\mathbf{I} - \frac{\xi\xi^\top}{\|\xi\|_2^2}\right) X\Delta}{\sqrt{n}\|\xi\|_2} = \frac{\left(\|X\Delta\|_2^2 - \frac{(\xi^\top X\Delta)^2}{\|\xi\|_2^2}\right)}{\sqrt{n}\|\xi\|_2} \\ &\leq \frac{\|X\Delta\|_2^2}{\sqrt{n}\|\xi\|_2} \leq \frac{8\psi_{\max}}{\sigma}\|\Delta\|_2^2 \end{aligned}$$

Next, we verify the RSC property. We want to show that with high probability, for any constant $a \in (0,1)$

$$\left|\frac{\xi^\top}{\|\xi\|_2} X\Delta\right| \leq \sqrt{1-a}\|X\Delta\|_2. \tag{7.68}$$

Consequently, we have

$$\Delta^\top X^\top \left(\mathbf{I} - \frac{\xi\xi^\top}{\|\xi\|_2^2}\right) X\Delta = \|X\Delta\|_2^2 - \left(\frac{\xi^\top}{\|\xi\|_2} X\Delta\right)^2 \geq a\|X\Delta\|_2^2.$$

This further implies

$$\delta\mathcal{L}(w+\Delta, w) = \frac{1}{\sqrt{n}\|\xi\|_2}\Delta^\top X^\top \left(\mathbf{I} - \frac{\xi\xi^\top}{\|\xi\|_2^2}\right) X\Delta \geq \frac{a\psi_{\min}}{2\|\xi\|_2/\sqrt{n}}\|\Delta\|_2^2. \tag{7.69}$$

Since $\|\dot{z}\|_2 \leq \|\ddot{z}\|_2$, then for any real constant $a \in (0,1)$,

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{\xi^\top}{\|\xi\|_2}X\Delta\right| \leq \sqrt{1-a}\|X\Delta\|_2\right] = \mathbb{P}\left[\left|\frac{(\epsilon - \alpha\dot{z} - (1-\alpha)\ddot{z})^\top}{\|\epsilon - \alpha\dot{z} - (1-\alpha)\ddot{z}\|_2}(\dot{z}-\ddot{z})\right| \leq \sqrt{1-a}\|\dot{z}-\ddot{z}\|_2\right] \\
&\overset{(i)}{\geq} \mathbb{P}\left[\left|\frac{(\epsilon - \dot{z})^\top(\dot{z}-\ddot{z})}{\|\epsilon - \dot{z}\|_2}\right| \leq \sqrt{1-a}\|\dot{z}-\ddot{z}\|_2\right] \\
&= \mathbb{P}\left[\Big(\epsilon^\top(\dot{z}-\ddot{z}) - \dot{z}^\top(\dot{z}-\ddot{z})\Big)^2 \leq (1-a)\|\epsilon-\dot{z}\|_2^2\|\dot{z}-\ddot{z}\|_2^2\right] \\
&\overset{(ii)}{=} \mathbb{P}\left[\left|\left(\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}\right)^2 + \|\dot{z}\|_2^2 - 2\epsilon^\top\dot{z}\right| \leq (1-a)(\|\epsilon\|_2^2 + \|\dot{z}\|_2^2 - 2\epsilon^\top\dot{z})\right], \qquad (7.70)
\end{aligned}$$

where (ii) is from dividing both sides by $\|v\|_2^2$, and (i) is from a geometric inspection and the randomness of ϵ, i.e., for any $\alpha \in [0,1]$ and $\|\dot{z}\|_2 \leq \|\ddot{z}\|_2$, we have $\left|\frac{-\dot{z}^\top}{\|-\dot{z}\|_2}(\dot{z}-\ddot{z})\right| \leq \left|\frac{(-\alpha\dot{z}-(1-\alpha)\ddot{z})^\top}{\|-\alpha\dot{z}-(1-\alpha)\ddot{z}\|_2}(\dot{z}-\ddot{z})\right|$. The random vector ϵ with i.i.d. entries does not affect the inequality above. Let's first discuss one side of the probability in (7.70),

$$\begin{aligned}
&\mathbb{P}\left[\left(\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}\right)^2 + \|\dot{z}\|_2^2 - 2\epsilon^\top\dot{z} \leq (1-a)(\|\epsilon\|_2^2 + \|\dot{z}\|_2^2 - 2\epsilon^\top\dot{z})\right] \\
&= \mathbb{P}\left[(1-a)\|\epsilon\|_2^2 \geq \left(\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}\right)^2 + a(\|\dot{z}\|_2^2 - 2\epsilon^\top\dot{z})\right]. \qquad (7.71)
\end{aligned}$$

Since ϵ has i.i.d. sub-Gaussian entries with $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i^2] = \sigma^2$ for all $i = 1,\ldots,n$, then $\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}$ and $\epsilon^\top\dot{z}$ are also zero-mean sub-Gaussians with variances σ^2 and $\sigma^2\|\dot{z}\|_2^2$ respectively. We have from [239] that

$$\mathbb{P}\left[\|\epsilon\|_2^2 \leq n\sigma^2(1-\delta)\right] \leq \exp\left(-\frac{n\delta^2}{16}\right), \qquad (7.72)$$

$$\mathbb{P}\left[\left(\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}\right)^2 \geq n\sigma^2\delta^2\right] \leq \exp\left(-\frac{n\delta^2}{2}\right), \qquad (7.73)$$

$$\mathbb{P}\left[\epsilon^\top\dot{z} \leq -n\sigma^2\delta\right] \leq \exp\left(-\frac{n^2\sigma^2\delta^2}{2\|\dot{z}\|_2^2}\right). \qquad (7.74)$$

Combining (7.72) – (7.74) with $\|\dot{z}\|_2^2 \leq n\sigma^2/4$, we have from union bound that with probability at least $1-\exp\left(-\frac{n}{144}\right)-\exp\left(-\frac{n}{128}\right)-\exp\left(-\frac{n}{128}\right)\geq 1-3\exp\left(-\frac{n}{144}\right)$,

$$\|\epsilon\|_2^2 \geq \frac{2}{3}n\sigma^2, \quad \left(\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}\right)^2 \leq \frac{1}{64}n\sigma^2, \quad -\epsilon^\top\dot{z} \leq \frac{1}{16}n\sigma^2.$$

This implies for $a \leq 3/5$, we have $\frac{\xi^\top}{\|\xi\|_2}X\Delta \leq \sqrt{1-a}\|X\Delta\|_2$. For the other side of (7.70), we have

$$\begin{aligned}&\mathbb{P}\left[\left(\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}\right)^2+\|\dot{z}\|_2^2-2\epsilon^\top\dot{z}\geq -(1-a)(\|\epsilon\|_2^2+\|\dot{z}\|_2^2-2\epsilon^\top\dot{z})\right]\\&=\mathbb{P}\left[(1-a)\|\epsilon\|_2^2\geq -\left(\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}\right)^2-(2-a)(\|\dot{z}\|_2^2-2\epsilon^\top\dot{z})\right]\\&\geq \mathbb{P}\left[(1-a)\|\epsilon\|_2^2\geq \left(\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}\right)^2+a(\|\dot{z}\|_2^2-2\epsilon^\top\dot{z})\right]. \qquad (7.75)\end{aligned}$$

Combining (7.70), (7.71) and (7.75), we have (7.68) holds with high probability, i.e., for any $r>0$,

$$\mathbb{P}\left[\left|\frac{\xi^\top}{\|\xi\|_2}X\Delta\right|\leq\sqrt{1-a}\|X\Delta\|_2\right]\geq 1-6\exp\left(-\frac{n}{144}\right).$$

Now wo bound $\|\xi\|_2$ to obtain the desired result. From [239], we have

$$\mathbb{P}\left[\|\epsilon\|_2^2\geq n\sigma^2(1+\delta)\right]\leq\exp\left(-\frac{n\delta^2}{18}\right)=\exp\left(-\frac{n}{72}\right), \qquad (7.76)$$

where we take $\delta=1/2$. From $\xi=\epsilon-\alpha\dot{z}-(1-\alpha)\ddot{z}$, we have

$$\|\xi\|_2\leq\|\epsilon\|_2+\alpha\|\dot{z}\|_2+(1-\alpha)\|\ddot{z}\|_2\overset{(i)}{\leq}\|\epsilon\|_2+\|\ddot{z}\|_2\overset{(ii)}{\leq}\sqrt{\frac{3n}{2}}\sigma+\frac{1}{2}\sqrt{n}\sigma. \qquad (7.77)$$

where (i) is from $\|\dot{z}\|_2\leq\|\ddot{z}\|_2$ and (ii) is from (7.76) and $\|\dot{z}\|_2^2\leq n\sigma^2/4$. Then by the union bound setting $a=1/2$, with probability at least $1-7\exp\left(-\frac{n}{144}\right)$, we have $\delta\mathcal{L}(w+\Delta,w)\geq\frac{\psi_{\min}}{8\sigma}\|\Delta\|_2^2$. Moreover, we also have $r=\frac{\sigma^2}{8\psi_{\max}}>s^*\left(64\sigma\lambda/\psi_{\min}\right)^2\geq s^*\left(8\lambda/\rho^-_{s^*+\tilde{s}}\right)^2$ for large enough $n\geq c_1s^*\log d$, where $\lambda=24\sqrt{\log d/n}$.

Step 3. Given the proposed conditions, we have that $\mathcal{L}$ satisfies the LRHS property by combining the analysis in [242].

7.3.2 Intermediate Results of Theorem 9

We introduce some important implications of the proposed assumptions. Recall that $\mathcal{S}^* = \{j : \theta_j^* \neq 0\}$ be the index set of non-zero entries of θ^* with $s^* = |\mathcal{S}^*|$ and $\overline{\mathcal{S}}^* = \{j : \theta_j^* = 0\}$ be the complement set. Lemma 2 implies RSC and RSS hold with parameter $\rho^-_{s^*+2\tilde{s}}$ and $\rho^+_{s^*+2\tilde{s}}$ respectively. By [243], the following conditions are equivalent to RSC and RSS, i.e., for any $v, w \in \mathbb{R}^d$ satisfying $\|v - w\|_0 \leq s^* + 2\tilde{s}$,

$$\rho^-_{s^*+2\tilde{s}}\|v - w\|_2^2 \leq (v - w)^\top \nabla \mathcal{L}(w) \leq \rho^+_{s^*+2\tilde{s}}\|v - w\|_2^2, \tag{7.78}$$

$$\frac{1}{\rho^+_{s^*+2\tilde{s}}}\|\nabla \mathcal{L}(v) - \nabla \mathcal{L}(w)\|_2^2 \leq (v - w)^\top \nabla \mathcal{L}(w) \leq \frac{1}{\rho^-_{s^*+2\tilde{s}}}\|\nabla \mathcal{L}(v) - \nabla \mathcal{L}(w)\|_2^2. \tag{7.79}$$

From the convexity of ℓ_1 norm, we have

$$\|v\|_1 - \|w\|_1 \geq (v - w)^\top g, \tag{7.80}$$

where $g \in \partial\|w\|_1$. Combining and (7.78) and (7.80), we have for any $v, w \in \mathbb{R}^d$ satisfying $\|v - w\|_0 \leq s^* + 2\tilde{s}$,

$$\mathcal{F}_\lambda(v) - \mathcal{F}_\lambda(w) - (v - w)^\top \nabla \mathcal{F}_\lambda(w) \geq \rho^-_{s^*+2\tilde{s}}\|v - w\|_2^2, \tag{7.81}$$

Remark 8. For any t and k, the line search satisfies

$$\tilde{L}^{(t)} \leq L^{(t)} \leq L_{\max},\ L \leq \tilde{L}^{(t)} \leq L^{(t)} \leq 2L \ \text{ and } \ \rho^+_{s^*+2\tilde{s}} \leq \tilde{L}^{(t)} \leq L^{(t)} \leq 2\rho^+_{s^*+2\tilde{s}}, \tag{7.82}$$

where $L = \min\{L : \|\nabla \mathcal{L}(v) - \nabla \mathcal{L}(w)\|_2 \leq L\|\boldsymbol{x} - y\|_2, \forall v, w \in \mathbb{R}^d\}$.

We first show that when θ is sparse and the approximate KKT condition is satisfied, then both estimation error and objective error, w.r.t. the true model parameter, are bounded. This is formalized in Lemma 26, and its proof is deferred to Appendix 7.3.8.

Lemma 26. Suppose conditions in Lemma 2 hold. If θ satisfies $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$ and the approximate KKT condition $\min_{g\in\partial\|\theta\|_1}\|\nabla\mathcal{L}(\theta)+\lambda g\|_\infty\le\lambda/2$, then we have

$$\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1 \le 5\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1, \tag{7.83}$$

$$\|\theta-\theta^*\|_2 \le \frac{2\lambda\sqrt{s^*}}{\rho^-_{s^*+2\tilde{s}}}, \tag{7.84}$$

$$\|\theta-\theta^*\|_1 \le \frac{12\lambda s^*}{\rho^-_{s^*+2\tilde{s}}}, \tag{7.85}$$

$$\mathcal{F}_\lambda(\theta)-\mathcal{F}_\lambda(\theta^*) \le \frac{6\lambda^2 s^*}{\rho^-_{s^*+2\tilde{s}}}. \tag{7.86}$$

Next, we show that if θ is sparse and the objective error is bounded, then the estimation error is also bounded. This is formalized in Lemma 27, and its proof is deferred to Appendix 7.3.8.

Lemma 27. Suppose conditions in Lemma 2 hold. If θ satisfies $\|\theta_{\overline{\mathcal{S}}^*}\|_0\le\tilde{s}$ and the objective satisfies $\mathcal{F}_\lambda(\theta)-\mathcal{F}_\lambda(\theta^*)\le\frac{6\lambda^2s^*}{\rho^-_{s^*+2\tilde{s}}}$, then we have

$$\|\theta-\theta^*\|_2 \le \frac{4\lambda\sqrt{3s^*}}{\rho^-_{s^*+2\tilde{s}}}, \tag{7.87}$$

$$\|\theta-\theta^*\|_1 \le \frac{24\lambda s^*}{\rho^-_{s^*+2\tilde{s}}}. \tag{7.88}$$

We then show that if θ is sparse and the objective error is bounded, then each proximal-gradient update preserves solution to be sparse. This is formalized in Lemma 28, and its proof is deferred to Appendix 7.3.8.

Lemma 28. Suppose conditions in Lemma 2 hold. If θ satisfies $\|\theta_{\overline{\mathcal{S}}^*}\|_0\le\tilde{s}$, L satisfies $L<2\rho^+_{s^*+2\tilde{s}}$, and the objective satisfies $\mathcal{F}_\lambda(\theta)-\mathcal{F}_\lambda(\theta^*)\le\frac{6\lambda^2s^*}{\rho^-_{s^*+2\tilde{s}}}$, then we have $\|\left(\mathcal{T}_{L,\lambda}(\theta)\right)_{\overline{\mathcal{S}}^*}\|_0\le\tilde{s}$.

Moreover, we show that if θ satisfies the approximate KKT condition, then the objective has a bounded error w.r.t. the regularizaSuppose conditions in Lemma 2 holdtion parameter λ. This characterizes the geometric decrease of the objective error when we choose a geometrically decreasing sequence of regularization parameters. This is formalized in Lemma 29, and its proof is deferred to Appendix 7.3.8.

Lemma 29. . If θ satisfies $\omega_\lambda(\theta)\le\lambda/2$, then for $\overline{\theta}=\operatorname{argmin}_\theta\mathcal{F}_\lambda(\theta)$, we have $\mathcal{F}_\lambda(\theta)-\mathcal{F}_\lambda(\overline{\theta})\le\frac{24\lambda\omega_\lambda(\theta)s^*}{\rho^-_{s^*+2\tilde{s}}}$.

Furthermore, we show a local linear convergence rate if the initial value $\theta^{(0)}$ is sparse and satisfies the approximate KKT condition with adequate precision. Besides, the estimation after each proximal gradient update is also sparse. This is the key result in demonstrating the overall geometric convergence rate of the algorithm. This is formalized in Lemma 30, and its proof is deferred to Appendix 7.3.8.

Lemma 30. Suppose conditions in Lemma 2 holds. If the initialization $\theta^{(0)}$ satisfies $\|\theta^{(0)}\|_0 \leq \tilde{s}$. Then with $\overline{\theta} = \operatorname{argmin}_\theta \mathcal{F}_\lambda(\theta)$, for any $t = 1, 2, \ldots$, we have $\|\theta^{(t)}\|_0 \leq \tilde{s}$ and $\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta}) \leq \left(1 - \frac{1}{8\kappa_{s^*+2\tilde{s}}}\right)^t \left(\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\overline{\theta})\right)$.

Finally, we introduce two results characterizing the proximal gradient mapping operation, adapted from [123] and [130] without proof. The first lemma describes sufficient descent of the objective by proximal gradient method.

Lemma 31 (Adapted from Theorem 2 in [123])**.** For any $L > 0$,

$$\mathcal{Q}_\lambda\left(\mathcal{T}_{L,\lambda}(\theta), \theta\right) \leq \mathcal{F}_\lambda\left(\theta\right) - \frac{L}{2}\|\mathcal{T}_{L,\lambda}(\theta) - \theta\|_2^2.$$

Besides, if $\mathcal{L}(\theta)$ is convex, we have

$$\mathcal{Q}_\lambda\left(\mathcal{T}_{L,\lambda}(\theta), \theta\right) \leq \min_{\mathbf{x}} \mathcal{F}_\lambda\left(\mathbf{x}\right) + \frac{L}{2}\|\mathbf{x} - \theta\|_2^2. \tag{7.89}$$

Further, we have for any $L \geq L$,

$$\mathcal{F}_\lambda\left(\mathcal{T}_{L,\lambda}(\theta)\right) \leq \mathcal{Q}_\lambda\left(\mathcal{T}_{L,\lambda}(\theta), \theta\right) \leq \mathcal{F}_\lambda\left(\theta\right) - \frac{L}{2}\|\mathcal{T}_{L,\lambda}(\theta) - \theta\|_2^2. \tag{7.90}$$

The next lemma provides an upper bound of the optimal residue $\omega(\cdot)$.

Lemma 32 (Adapted from Lemma 2 in [130])**.** For any $L > 0$, if L is the Lipschitz constant of $\nabla\mathcal{L}$, then

$$\omega_\lambda\left(\mathcal{T}_{L,\lambda}(\theta)\right) \leq \left(L + S_L(\theta)\right)\|\mathcal{T}_{L,\lambda}(\theta) - \theta\|_2 \leq 2L\|\mathcal{T}_{L,\lambda}(\theta) - \theta\|_2,$$

where $S_L(\theta) = \frac{\|\nabla\mathcal{L}(\mathcal{T}_{L,\lambda}(\theta)) - \nabla\mathcal{L}(\theta)\|_2}{\|\mathcal{T}_{L,\lambda}(\theta) - \theta\|_2}$ is a local Lipschitz constant, which satisfies $S_L(\theta) \leq L$.

7.3.3 Proof of Theorem 9

We demonstrate the linear rate when the initial value $\theta^{(0)}$ satisfies $\omega_\lambda(\theta^{(0)}) \leq \frac{\lambda}{2}$ with $\|(\theta^{(0)})_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$. The proof is provided in Appendix 7.3.6.

Theorem 21. Suppose conditions in Lemma 2 hold. Let $\overline{\theta} = \operatorname{argmin}_\theta \mathcal{F}_\lambda(\theta)$ be the optimal solution with regularization parameter λ. If the initial value $\theta^{(0)}$ satisfies $\omega_\lambda(\theta^{(0)}) \leq \frac{\lambda}{2}$ with $\|(\theta^{(0)})_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$, then for any $t = 1, 2, \ldots$, we have $\|(\theta^{(t)})_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$,

$$\begin{aligned} \|\theta^{(t)} - \overline{\theta}\|_2^2 &\leq \left(1 - \frac{1}{8\kappa_{s^*+2\tilde{s}}}\right)^t \frac{24\lambda s^* \omega_\lambda(\theta^{(t)})}{(\rho^-_{s^*+2\tilde{s}})^2} \quad \text{and} \\ \mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta}) &\leq \left(1 - \frac{1}{8\kappa_{s^*+2\tilde{s}}}\right)^t \frac{24\lambda s^* \omega_\lambda(\theta^{(t)})}{\rho^-_{s^*+2\tilde{s}}}, \end{aligned} \tag{7.91}$$

In addition, to achieve the approximate KKT condition $\omega_\lambda(\theta^{(t)}) \leq \varepsilon$, the number of proximal gradient steps is no more than

$$\frac{\log\left(96\,(1 + \kappa_{s^*+2\tilde{s}})^2\, \lambda^2 s^* \kappa_{s^*+2\tilde{s}}/\varepsilon^2\right)}{\log\left(8\kappa_{s^*+2\tilde{s}}/(8\kappa_{s^*+2\tilde{s}} - 1)\right)}. \tag{7.92}$$

From basic inequalities, since $\kappa_{s^*+2\tilde{s}} \geq 1$, we have $\log\left(\frac{8\kappa_{s^*+2\tilde{s}}}{8\kappa_{s^*+2\tilde{s}}-1}\right) \geq \log\left(1 + \frac{1}{8\kappa_{s^*+2\tilde{s}}-1}\right) \geq \frac{1}{8\kappa_{s^*+2\tilde{s}}}$. Then (7.92) can be simplified as $\mathcal{O}\left(\kappa_{s^*+2\tilde{s}}\left(\log\left(\kappa^3_{s^*+2\tilde{s}}\lambda^2 s^*/\varepsilon^2\right)\right)\right)$.

As can be seen from Theorem 21, when the initial value $\theta^{(0)}$ satisfies $\omega_\lambda(\theta^{(0)}) \leq \frac{\lambda}{2}$ with $\|(\theta^{(0)})_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$, then we can guarantee the geometric convergence rate of the estimated objective value towards the minimal objective.

Next, we need to show that when $\theta_{(0)} \in \mathcal{B}_r$, the approximate KKT holds for $\theta_{(1)}$, which is also sparse. We demonstrate this result in Lemma 33 and provide its proof in Appendix 7.3.7.

Lemma 33. Suppose conditions in Lemma 2 hold.s. If $\frac{\rho^-_{s^*+\tilde{s}}}{8}\sqrt{\frac{r}{s^*}} > \lambda$ and $\|\theta - \theta^*\|_2^2 \leq r$ holds, then we have $\omega_\lambda(\theta) \leq 4\sqrt{r}$ and $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$.

Combining the results above, we finish the proof.

7.3.4 Proof of Theorem 10

We present a few important intermediate results that are key components of our main proof. The first result shows that in a neighborhood of the true model parameter θ^*, the sparsity of the solution is preserved when we use a sparse initialization. The proof is provided in Appendix 7.4.3.

Lemma 34 (Sparsity Preserving Lemma)**.** Suppose conditions in Lemma 2 hold with $\varepsilon \leq \frac{\lambda}{8}$. Given $\theta^{(t)} \in \mathcal{B}\left(\theta^*, R\right)$ and $\|\theta_{\overline{\mathcal{S}}}^{(t)}\|_0 \leq \tilde{s}$, there exists a generic constant C_1 such that

$$\|\theta_{\overline{\mathcal{S}}}^{(t+1)}\|_0 \leq \tilde{s}, \quad \|\theta^{(t+1)} - \theta^*\|_2 \leq \frac{C_1 \lambda \sqrt{s^*}}{\rho^-_{s^*+2\tilde{s}}} \quad \text{and} \quad \mathcal{F}_\lambda(\theta^{(t)}) \leq \mathcal{F}_\lambda(\theta^*) + \frac{15\lambda^2 s^*}{4\rho^-_{s^*+2\tilde{s}}}..$$

Denote $\mathcal{B}(\theta, r) = \left\{\boldsymbol{\phi} \in \mathbb{R}^d \ : \ \|\boldsymbol{\phi} - \theta\|_2 \leq r\right\}$. We then show that every step of proximal Newton updates within each stage has a quadratic convergence rate to a local minimizer, if we start with a sparse solution in the refined region. The proof is provided in Appendix 7.4.3.

Lemma 35. Suppose conditions in Lemma 2 hold. If $\theta^{(t)} \in \mathcal{B}\left(\theta^*, r\right)$ and $\left\|\theta^{(t)}{}_{\overline{\mathcal{S}}}\right\|_0 \leq \tilde{s}$, then for each stage $K \geq 2$, we have

$$\|\theta^{(t+1)} - \overline{\theta}\|_2 \leq \frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}} \|\theta^{(t)} - \overline{\theta}\|_2^2.$$

In the following, we need to use the property that the iterates $\theta^{(t)} \in \mathcal{B}(\overline{\theta}, 2r)$ instead of $\theta^{(t)} \in \mathcal{B}\left(\theta^*, r\right)$ for convergence analysis of the proximal Newton method. This property holds since we have $\theta^{(t)} \in \mathcal{B}\left(\theta^*, r\right)$ and $\overline{\theta} \in \mathcal{B}\left(\theta^*, r\right)$ simultaneously. Thus $\theta^{(t)} \in \mathcal{B}\left(\overline{\theta}, 2r\right)$, where $2r = \frac{\rho^-_{s^*+2\tilde{s}}}{L_{s^*+2\tilde{s}}}$ is the radius for quadratic convergence region of the proximal Newton algorithm.

The following lemma demonstrates that the step size parameter is simply 1 if the the sparse solution is in the refined region. The proof is provided in Appendix 7.4.3.

Lemma 36. Suppose conditions in Lemma 2 hold. If $\theta^{(t)} \in \mathcal{B}(\overline{\theta}, 2r)$ and $\|\theta^{(t)}{}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}$ at each stage $K \geq 2$ with $\frac{1}{4} \leq \alpha < \frac{1}{2}$, then $\eta_t = 1$. Further, we have

$$\mathcal{F}_\lambda(\theta^{(t+1)}) \leq \mathcal{F}_\lambda(\theta^{(t)}) + \frac{1}{4}\gamma_t.$$

Moreover, we present a critical property of γ_t. The proof is provided in Appendix 7.4.3.

Lemma 37. Denote $\Delta\theta^{(t)} = \theta^{(t)} - \theta^{(t+1)}$ and

$$\gamma_t = \nabla \mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \|\lambda\left(\theta^{(t)} + \Delta\theta^{(t)}\right)\|_1 - \|\lambda\left(\theta^{(t)}\right)\|_1.$$

Then we have $\gamma_t \leq -\|\Delta\theta^{(t)}\|^2_{\nabla^2\mathcal{L}(\theta^{(t)})}$.

In addition, we present the sufficient number of iterations for each convex relaxation stage to achieve the approximate KKT condition. The proof is provided in Appendix 7.4.3.

Lemma 38. Suppose conditions in Lemma 2 hold. To achieve the approximate KKT condition $\omega_\lambda\left(\theta^{(t)}\right) \leq \varepsilon$ for any $\varepsilon > 0$ at each stage $K \geq 2$, the number of iteration for proximal Newton updates is at most

$$\log\log\left(\frac{3\rho^+_{s^*+2\tilde{s}}}{\varepsilon}\right).$$

Combining the results above, we have desired results in Theorem 10.

7.3.5 Proof of Theorem 11

Part 1. We first show that estimation errors are as claimed. We have that $\omega_\lambda(\hat{\theta}^{(0)}) \leq \lambda/2$. By Theorem 21, we have for any $t = 1, 2, \ldots$, $\|(\theta^{(t)}_{[K+1]})_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$. Applying Lemma 26 recursively, we have

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{2\lambda\sqrt{s^*}}{\rho^-_{s^*+2\tilde{s}}} \quad \text{and} \quad \|\hat{\theta} - \theta^*\|_1 \leq \frac{12\lambda s^*}{\rho^-_{s^*+2\tilde{s}}}.$$

Applying Lemma 2 with $\lambda = 24\sqrt{\log d/n}$ and $\rho^-_{s^*+2\tilde{s}} = \frac{\psi_{\min}}{8\sigma}$, then by union bound, with probability at least $1 - 8\exp\left(-\frac{n}{144}\right) - 2d^{-1}$, we have

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{384\sigma\sqrt{s^*\log d/n}}{\psi_{\min}} \quad \text{and} \quad \|\hat{\theta} - \theta^*\|_1 \leq \frac{2304\sigma s^*\sqrt{\log d/n}}{\psi_{\min}}.$$

Part 2. Next, we demonstrate the result of the estimation of variance. Let $\overline{\theta} = \operatorname{argmin}_\theta \mathcal{F}_\lambda(\theta)$ be the optimal solution. Apply the argument in Part recursively, we have

$$\|\overline{\theta} - \theta^*\|_1 \leq \frac{2304\sigma s^* \sqrt{\log d/n}}{\psi_{\min}}. \tag{7.93}$$

Denote $c_1, c_2, \ldots$ as positive universal constants. Then we have

$$\begin{aligned} \mathcal{L}(\overline{\theta}) - \mathcal{L}(\theta^*) &\leq \lambda(\|\theta^*\|_1 - \|\overline{\theta}\|_1) \leq \lambda(\|\theta^*_{\mathcal{S}^*}\|_1 - \|(\overline{\theta})_{\mathcal{S}^*}\|_1 - \|(\overline{\theta})_{\overline{\mathcal{S}}^*}\|_1) \\ &\leq \lambda\|(\overline{\theta} - \theta^*)_{\mathcal{S}^*}\|_1 \leq \lambda\|\overline{\theta} - \theta^*\|_1 \overset{(ii)}{\leq} c_1 \frac{\sigma s^* \log d}{n}, \end{aligned} \tag{7.94}$$

where (i) is from the value of λ and ℓ_1 error bound in (7.93).

On the other hand, from the convexity of $\mathcal{L}(\theta)$, we have

$$\begin{aligned} \mathcal{L}(\overline{\theta}) - \mathcal{L}(\theta^*) &\geq (\overline{\theta} - \theta^*)^\top \nabla \mathcal{L}(\theta^*) \geq -\|\nabla \mathcal{L}(\theta^*)\|_\infty \|\hat{\theta} - \theta\|_1 \\ &\overset{(i)}{\geq} -c_2 \lambda \|\overline{\theta} - \theta\|_1 \overset{(ii)}{\geq} -c_3 \frac{\sigma s^* \log d}{n}, \end{aligned} \tag{7.95}$$

where (i) is from Lemma 2 and (ii) value of λ and ℓ_1 error bound in (7.93). By definition, we have

$$\mathcal{L}(\overline{\theta}) - \mathcal{L}(\theta^*) = \frac{\|y - X\overline{\theta}\|_2}{\sqrt{n}} - \frac{\|\epsilon\|_2}{\sqrt{n}}. \tag{7.96}$$

From [239], we have for any $\delta > 0$,

$$\mathbb{P}\left[\left|\frac{\|\epsilon\|_2^2}{n} - \sigma^2\right| \geq \sigma^2 \delta\right] \leq 2 \exp\left(-\frac{n\delta^2}{18}\right). \tag{7.97}$$

Combining (7.94), (7.95), (7.96) and (7.97) with $\delta^2 = \frac{c_3 s^* \log d}{n}$, we have with high probability,

$$\left|\frac{\|y - X\overline{\theta}\|_2}{\sqrt{n}} - \sigma\right| = \mathcal{O}\left(\frac{\sigma s^* \log d}{n}\right). \tag{7.98}$$

From Part 1, for $n \geq c_4 s^* \log d$, with high probability, we have $\|\overline{\theta} - \theta^*\|_2 \leq \frac{384\sigma\sqrt{s^* \log d/n}}{\psi_{\min}} \leq \frac{\sigma}{2\sqrt{2\psi_{\max}}}$, then $\overline{\theta} \in \mathcal{B}_r^{s^*+\tilde{s}}$ and $\|\hat{\theta} - \overline{\theta}\|_0 \leq s^* + 2\tilde{s}$. Then from the analysis of Theorem 21, we have

$$\omega_\lambda(\theta^{(t+1)}) \leq (1+\kappa_{s^*+2\tilde{s}})\sqrt{4\rho^+_{s^*+2\tilde{s}}\left(\mathcal{F}_\lambda(\theta^{(t)})-\mathcal{F}_\lambda(\overline{\theta})\right)} \leq \varepsilon.$$

This implies

$$\mathcal{F}_\lambda(\theta^{(t)})-\mathcal{F}_\lambda(\overline{\theta}) \leq \frac{\epsilon^2}{4\rho^+_{s^*+2\tilde{s}}\left(1+\kappa_{s^*+2\tilde{s}}\right)^2}. \tag{7.99}$$

On the other hand, from the LRSC property of $\mathcal{L}$, convexity of ℓ_1 norm and optimality of $\overline{\theta}$, we have

$$\mathcal{F}_\lambda(\theta^{(t)})-\mathcal{F}_\lambda(\overline{\theta}) \geq \rho^-_{s^*+2\tilde{s}}\|\hat{\theta}-\overline{\theta}\|_2^2. \tag{7.100}$$

Combining (7.99), (7.100) and Lemma 2, we have

$$\frac{\|X(\hat{\theta}-\overline{\theta})\|_2}{\sqrt{n}} \leq \sqrt{\frac{8\rho^+_{s^*+2\tilde{s}}}{\sigma}}\|\hat{\theta}-\theta^*\|_2 \leq \sqrt{\frac{2}{\sigma\rho^-_{s^*+2\tilde{s}}}}\frac{\epsilon}{\left(1+\kappa_{s^*+2\tilde{s}}\right)} \leq \frac{4\epsilon}{\left(1+\kappa_{s^*+2\tilde{s}}\right)\sqrt{\psi_{\min}}}. \tag{7.101}$$

Combining (7.98) and (7.101), we have

$$\left|\frac{\|y-X\hat{\theta}\|_2}{\sqrt{n}}\right| \leq \left|\frac{\|y-X\overline{\theta}\|_2}{\sqrt{n}}\right| + \frac{\|X(\hat{\theta}-\overline{\theta})\|_2}{\sqrt{n}} \leq \left|\frac{\|y-X\overline{\theta}\|_2}{\sqrt{n}}\right| + \frac{4\epsilon}{\left(1+\kappa_{s^*+2\tilde{s}}\right)\sqrt{\psi_{\min}}}.$$

If $\epsilon \leq c_5\frac{\sigma s^*\log d}{n}$ for some constant c_5, then we have the desired result.

7.3.6 Proof of Theorem 21

Note that the RSS property implies that line search terminate when $\tilde{L}^{(t)}$ satisfies

$$\rho^+_{s^*+2\tilde{s}} \leq \tilde{L}^{(t)} \leq 2\rho^+_{s^*+2\tilde{s}}. \tag{7.102}$$

Since the initialization $\theta^{(0)}$ satisfies $\omega_\lambda(\theta^{(0)}) \leq \frac{\lambda}{2}$ with $\|(\theta^{(0)})_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$, then by Lemma 26, we have $\mathcal{F}_\lambda(\theta^{(0)})-\mathcal{F}_\lambda(\theta^*) \leq \frac{6\lambda^2 s^*}{\rho^-_{s^*+2\tilde{s}}}$. Then by Lemma 28, we have $\|(\theta^{(1)})_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$.

By monotone decrease of $\mathcal{F}_\lambda(\theta^{(t)})$ from (7.90) in Lemma 31 and recursively applying Lemma 28, $\|(\theta^{(t)})_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$ holds in (7.91) for all $t = 1, 2, \ldots$.

For the objective error, we have

$$\begin{aligned}\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta}) &\overset{(i)}{\leq} \left(1 - \frac{1}{8\kappa_{s^*+2\tilde{s}}}\right)^t \left(\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\overline{\theta})\right) \\ &\overset{(ii)}{\leq} \left(1 - \frac{1}{8\kappa_{s^*+2\tilde{s}}}\right)^t \frac{24\lambda s^* \omega_\lambda(\theta^{(t)}}{\rho^-_{s^*+2\tilde{s}}},\end{aligned} \tag{7.103}$$

where (i) is from Lemma 30, and (ii) is from Lemma 29 and $\omega_\lambda(\theta^{(t+1)}) \leq \lambda/2 \leq \lambda$, which results in (7.91).

Combining (7.103), (7.81) with $\nabla\mathcal{F}_\lambda(\overline{\theta}) = 0$, we have

$$\|\theta^{(t)} - \overline{\theta}\|_2^2 \leq \frac{1}{\rho^-_{s^*+2\tilde{s}}} \left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta}) - \nabla\mathcal{F}_\lambda(\overline{\theta})\right) \leq \left(1 - \frac{1}{8\kappa_{s^*+2\tilde{s}}}\right)^t \frac{24\lambda s^* \omega_\lambda(\theta^{(t)})}{(\rho^-_{s^*+2\tilde{s}})^2}$$

For $\omega_\lambda(\theta^{(t+1)})$ of $(t+1)$-th iteration, we have

$$\begin{aligned}&\omega_\lambda(\theta^{(t+1)}) \\ &\overset{(i)}{\leq} \left(\tilde{L}^{(t)} + S_{\tilde{L}^{(t)}}(\theta^{(t)})\right)\|\theta^{(t+1)} - \theta^{(t)}\|_2 \overset{(ii)}{\leq} \left(\tilde{L}^{(t)} + \rho^+_{s^*+2\tilde{s}}\right)\|\theta^{(t+1)} - \theta^{(t)}\|_2 \\ &\overset{(iii)}{\leq} \tilde{L}^{(t)}\left(1 + \frac{\rho^+_{s^*+2\tilde{s}}}{\rho^-_{s^*+2\tilde{s}}}\right)\|\theta^{(t+1)} - \theta^{(t)}\|_2 \overset{(iv)}{\leq} \tilde{L}^{(t)}\left(1 + \frac{\rho^+_{s^*+2\tilde{s}}}{\rho^-_{s^*+2\tilde{s}}}\right)\sqrt{\frac{2\left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\theta^{(t+1)})\right)}{\tilde{L}^{(t)}}} \\ &\overset{(v)}{\leq} (1 + \kappa_{s^*+2\tilde{s}})\sqrt{4\rho^+_{s^*+2\tilde{s}}\left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\right)} \\ &\overset{(vi)}{\leq} (1 + \kappa_{s^*+2\tilde{s}})\sqrt{96\lambda^2 s^* \kappa_{s^*+2\tilde{s}}\left(1 - \frac{1}{8\kappa_{s^*+2\tilde{s}}}\right)^t},\end{aligned} \tag{7.104}$$

where (i) is from Lemma 32, (ii) is from $S_{\tilde{L}^{(t)}}(\theta^{(t)}) \leq \rho^+_{s^*+2\tilde{s}}$, (iii) is from $\rho^-_{s^*+2\tilde{s}} \leq \tilde{L}^{(t)}$ in (7.102), (iv) is from (7.90) in Lemma 31, (v) is from $\tilde{L}^{(t)} \leq 2\rho^+_{s^*+2\tilde{s}}$ in (7.102) and monotone decrease of $\mathcal{F}_\lambda(\theta^{(t)})$ from (7.90) in Lemma 31, and (vi) is from (7.103) and $\kappa_{s^*+2\tilde{s}} = \frac{\rho^+_{s^*+2\tilde{s}}}{\rho^-_{s^*+2\tilde{s}}}$.

Then we need $\omega_\lambda(\hat{\theta}) \leq \varepsilon \leq \lambda/4$. Set the R.H.S. of (7.104) to be no greater than ε, which is equivalent to require the number of iterations k to be an upper bound of (7.92).

7.3.7 Proof of Lemma 33

Part 1. We first show that given $\|\theta^{(0)} - \theta^*\|_2^2 \le r$, $\omega_\lambda(\theta^{(1)}) \le 4\sqrt{r}$ holds. From Lemma 32, we have

$$\omega_\lambda(\theta^{(1)}) \le 2L\|\theta^{(1)} - \theta^{(0)}\|_2 \le 4\|\theta^{(1)} - \theta^*\|_2 \le 4\sqrt{r}.$$

Part 2. We next demonstrate the sparsity of θ. From $\lambda \ge 6\|\nabla\mathcal{L}(\theta^*)\|_\infty$, then we have

$$\left|\left\{i \in \overline{\mathcal{S}}^* : |\nabla_i\mathcal{L}(\theta^*)| \ge \frac{\lambda}{6}\right\}\right| = 0. \tag{7.105}$$

Denote $\check{\mathcal{S}}_1 = \left\{i \in \overline{\mathcal{S}}^* : |\nabla_i\mathcal{L}(\theta) - \nabla_i\mathcal{L}(\theta^*)| \ge \frac{2\lambda}{3}\right\}$ and $\check{s}_1 = |\check{\mathcal{S}}_1|$. Then there exists some $\boldsymbol{b} \in \mathbb{R}^d$ such that $\|\boldsymbol{b}\|_\infty = 1$, $\|\boldsymbol{b}\|_0 \le \check{s}_1$ and $\boldsymbol{b}^\top(\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*)) \ge \frac{2\lambda\check{s}_1}{3}$. Then by the mean value theorem, we have for some $\check{\theta} = (1-\alpha)\theta + \alpha\theta^*$ with $\alpha \in [0,1]$, $\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*) = \nabla^2\mathcal{L}(\check{\theta})\Delta$, where $\Delta = \theta - \theta^*$. Then we have

$$\begin{aligned}\frac{2\lambda\check{s}_1}{3} &\le \boldsymbol{b}^\top\nabla^2\mathcal{L}(\check{\theta})\Delta \overset{(i)}{\le} \sqrt{\boldsymbol{b}^\top\nabla^2\mathcal{L}(\check{\theta})\boldsymbol{b}}\sqrt{\Delta^\top\nabla^2\mathcal{L}(\check{\theta})\Delta} \\ &\overset{(ii)}{\le} \sqrt{\check{s}_1\rho^+_{\check{s}_1}}\sqrt{\Delta^\top(\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*))},\end{aligned} \tag{7.106}$$

where (i) is from the generalized Cauchy-Schwarz inequality, (ii) is from the definition of RSS and the fact that $\|\boldsymbol{b}\|_2 \le \sqrt{\check{s}_1}\|\boldsymbol{b}\|_\infty = \sqrt{\check{s}_1}$. Let g achieve $\min_{g\in\partial\|\theta\|_1}\mathcal{F}_\lambda(\theta)$. Further, we have

$$\begin{aligned}&\Delta^\top(\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*)) \le \|\Delta\|_1\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*)\|_\infty \le \|\Delta\|_1(\|\nabla\mathcal{L}(\theta^*)\|_\infty + \|\nabla\mathcal{L}(\theta)\|_\infty) \\ &\le \|\Delta\|_1(\|\nabla\mathcal{L}(\theta^*)\|_\infty + \|\nabla\mathcal{L}(\theta) + \lambda g\|_\infty + \lambda\|g\|_\infty) \overset{(i)}{\le} \frac{28\lambda s^*}{3\rho^-_{s^*+\tilde{s}}}(\frac{\lambda}{6} + \frac{\lambda}{4} + \lambda) \le \frac{14\lambda^2 s^*}{\rho^-_{s^*+\tilde{s}}},\end{aligned} \tag{7.107}$$

where (i) is from $\|\tilde{\Delta}_{\overline{\mathcal{S}}^*}\|_1 \le \frac{5}{2}\|\tilde{\Delta}_{\mathcal{S}^*}\|_1$ and $\|\tilde{\Delta}_{\mathcal{S}^*}\|_1 \le \frac{8\lambda s^*}{3\rho^-_{s^*+\tilde{s}}}$, condition on λ, approximate KKT condition and $\|g\|_\infty \le 1$. Combining (7.106) and (7.107), we have

$\frac{2\sqrt{\check{s}_1}}{3} \leq \sqrt{\frac{14\rho^+_{\check{s}_1} s^*}{\rho^-_{s^*+\tilde{s}}}}$, which further implies

$$\check{s}_1 \leq \frac{32\rho^+_{\check{s}_1} s^*}{\rho^-_{s^*+\tilde{s}}} \leq 32\kappa_{s^*+2\tilde{s}} s^* \leq \tilde{s}. \tag{7.108}$$

For any $v \in \mathbb{R}^d$ that satisfies $\|v\|_0 \leq 1$, we have

$$\check{\mathcal{S}}_2 = \left\{ i \in \overline{\mathcal{S}}^* : \left| \nabla_i \mathcal{L}(\theta) + \frac{\lambda}{4} v_i \right| \geq \frac{5\lambda}{6} \right\} \subseteq \left\{ i \in \overline{\mathcal{S}}^* : |\nabla_i \mathcal{L}(\theta^*)| \geq \frac{\lambda}{6} \right\} \bigcup \check{\mathcal{S}}_1.$$

Then we have $|\check{\mathcal{S}}_2| \leq |\check{\mathcal{S}}_1| \leq \tilde{s}$. Since for any $i \in \overline{\mathcal{S}}^*$ and $\left|\nabla_i \mathcal{L}(\theta) + \frac{\lambda}{4} v_i\right| < \frac{5\lambda}{6}$, we can find g_i that satisfies $|g_i| \leq 1$ such that $\nabla_i \mathcal{L}(\theta) + \frac{\lambda}{4} v_i + \lambda g_i = 0$ which implies $\theta_i = 0$, then we have

$$\left| \left\{ i \in \overline{\mathcal{S}}^* : \left| \nabla_i \mathcal{L}(\theta) + \frac{\lambda}{4} v_i \right| < \frac{5\lambda}{6} \right\} \right| = 0.$$

This implies $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \leq |\check{\mathcal{S}}_2| \leq \tilde{s}$.

7.3.8 Proofs of Intermediate Lemmas in Appendix 7.3.2

Proof of Lemma 26

We first bound the estimation error. From Lemma 2, we have the RSC property, which indicates

$$\mathcal{L}(\theta) \geq \mathcal{L}(\theta^*) + (\theta - \theta^*)^\top \nabla \mathcal{L}(\theta^*) + (\rho^-_{s^*+2\tilde{s}}/2)\|\theta - \theta^*\|_2^2, \tag{7.109}$$

$$\mathcal{L}(\theta^*) \geq \mathcal{L}(\theta) + (\theta^* - \theta)^\top \nabla \mathcal{L}(\theta) + (\rho^-_{s^*+2\tilde{s}}/2)\|\theta - \theta^*\|_2^2, \tag{7.110}$$

Adding (7.110) and (7.109), we have

$$(\theta - \theta^*)^\top \nabla \mathcal{L}(\theta) \geq (\theta - \theta^*)^\top \nabla \mathcal{L}(\theta^*) + \rho^-_{s^*+2\tilde{s}}\|\theta - \theta^*\|_2^2. \tag{7.111}$$

Let $g \in \partial\|\theta\|_1$ be the subgradient that achieves the approximate KKT condition, then

$$(\theta-\theta^*)^\top\left(\nabla\mathcal{L}(\theta)+\lambda g\right) \le \|\theta-\theta^*\|_1\left\|\nabla\mathcal{L}(\theta)+\lambda g\right\|_\infty \le \frac{1}{2}\lambda\|\theta-\theta^*\|_1. \tag{7.112}$$

On the other hand, we have from (7.111)

$$(\theta-\theta^*)^\top\left(\nabla\mathcal{L}(\theta)+\lambda g\right) \ge (\theta-\theta^*)^\top\nabla\mathcal{L}(\theta^*)+\rho^-_{s^*+2\tilde{s}}\|\theta-\theta^*\|_2^2+\lambda g^\top(\theta-\theta^*), \tag{7.113}$$

Since $\|\theta-\theta^*\|_1=\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1+\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1$, then

$$(\theta-\theta^*)^\top\nabla\mathcal{L}(\theta^*) \ge -\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1\|\mathcal{L}(\theta^*)\|_\infty-\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1\|\mathcal{L}(\theta^*)\|_\infty. \tag{7.114}$$

Besides, we have

$$\begin{aligned}(\theta-\theta^*)^\top g &= g_{\mathcal{S}^*}^\top(\theta-\theta^*)_{\mathcal{S}^*}+g_{\overline{\mathcal{S}}^*}^\top(\theta-\theta^*)_{\overline{\mathcal{S}}^*} \overset{(i)}{\ge} -\|g_{\mathcal{S}^*}\|_\infty\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1+g_{\overline{\mathcal{S}}^*}^\top\theta_{\overline{\mathcal{S}}^*}\\ &\overset{(ii)}{\ge} -\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1+\|g_{\overline{\mathcal{S}}^*}\|_1 \overset{(iii)}{=} -\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1+\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1, \end{aligned} \tag{7.115}$$

where (i) and (iii) is from $\theta^*_{\overline{\mathcal{S}}^*}=0$, (ii) is from $\|g_{\mathcal{S}^*}\|_\infty\le 1$ and $g\in\partial\|\theta\|_1$.

Combining (7.112), (7.113), (7.114) and (7.115), we have

$$\begin{aligned}\frac{1}{2}\lambda\|\theta-\theta^*\|_1 \ge \rho^-_{s^*+2\tilde{s}}\|\theta-\theta^*\|_2^2-(\lambda+\|\mathcal{L}(\theta^*)\|_\infty)\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1\\ +(\lambda-\|\mathcal{L}(\theta^*)\|_\infty)\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1.\end{aligned}$$

This implies

$$\begin{aligned}&\rho^-_{s^*+2\tilde{s}}\|\theta-\theta^*\|_2^2+(\frac{1}{2}\lambda-\|\mathcal{L}(\theta^*)\|_\infty)\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1\\ &\le\left(\frac{3}{2}\lambda+\|\mathcal{L}(\theta^*)\|_\infty\right)\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1,\end{aligned} \tag{7.116}$$

which results in (7.83) from $\rho^-_{s^*+2\tilde{s}}>0$ and Lemma 2 as

$$\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1 \le \frac{\frac{3}{2}\lambda+\|\mathcal{L}(\theta^*)\|_\infty}{\frac{1}{2}\lambda-\|\mathcal{L}(\theta^*)\|_\infty}\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1.$$

Combining $\frac{1}{2}\lambda - \|\mathcal{L}(\theta^*)\|_\infty \geq 0$, $\frac{3}{2}\lambda + \|\mathcal{L}(\theta^*)\|_\infty \leq 2\lambda$ and (7.116), we have estimation errors in (7.84) and (7.85) as

$$\rho^-_{s^*+2\tilde{s}}\|\theta - \theta^*\|_2^2 \leq 2\lambda\|(\theta - \theta^*)_{\mathcal{S}^*}\|_1 \leq 2\lambda\sqrt{s^*}\|\theta - \theta^*\|_2 \text{ and}$$
$$\|\theta - \theta^*\|_1 \leq 6\|(\theta - \theta^*)_{\mathcal{S}^*}\|_1 \leq 6\sqrt{s^*}\|\theta - \theta^*\|_2.$$

Next, we bound the objective error in (7.86). We have

$$\begin{aligned}\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(\theta^*) &\overset{(i)}{\leq} -(\nabla\mathcal{L}(\theta) + \lambda g)^\top(\theta^* - \theta) \leq \|\nabla\mathcal{L}(\theta) + \lambda g\|_\infty\|\theta^* - \theta\|_1 \leq \frac{1}{2}\lambda\|\theta^* - \theta\|_1\\ &= \frac{1}{2}\lambda(\|(\theta^* - \theta)_{\mathcal{S}^*}\|_1 + \|(\theta^* - \theta)_{\overline{\mathcal{S}}^*}\|_1) \overset{(ii)}{\leq} 3\lambda\|(\theta^* - \theta)_{\mathcal{S}^*}\|_1\\ &\leq 3\lambda\sqrt{s^*}\|(\theta^* - \theta)_{\mathcal{S}^*}\|_2 \overset{(iii)}{\leq} \frac{6\lambda^2 s^*}{\rho^-_{s^*+2\tilde{s}}},\end{aligned}$$

where (i) is from the convexity of $\mathcal{F}_\lambda(\theta)$ with subgradient $\nabla\mathcal{L}(\theta)+\lambda g$, (ii) is from (7.83), and (iii) is from (7.84).

Proof of Lemma 27

Assumption $\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(\theta^*) \leq 6\lambda^2 s^*/\rho^-_{s^*+2\tilde{s}}$ implies

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) + \lambda(\|\theta\|_1 - \|\theta^*\|_1) \leq \frac{6\lambda^2 s^*}{\rho^-_{s^*+2\tilde{s}}}. \tag{7.117}$$

We have from the RSC property that

$$\mathcal{L}(\theta) \geq \mathcal{L}(\theta^*) + (\theta - \theta^*)^\top\nabla\mathcal{L}(\theta^*) + \frac{\rho^-_{s^*+2\tilde{s}}}{2}\|\theta - \theta^*\|_2^2, \tag{7.118}$$

Then we have (7.117) and (7.118),

$$\frac{\rho^-_{s^*+2\tilde{s}}}{2}\|\theta - \theta^*\|_2^2 \leq \frac{6\lambda^2 s^*}{\rho^-_{s^*+2\tilde{s}}} - (\theta - \theta^*)^\top\nabla\mathcal{L}(\theta^*) + \lambda(\|\theta^*\|_1 - \|\theta\|_1). \tag{7.119}$$

Besides, we have

$$(\theta-\theta^*)^\top \nabla\mathcal{L}(\theta^*) \geq -\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1\|\mathcal{L}(\theta^*)\|_\infty - \|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1\|\mathcal{L}(\theta^*)\|_\infty, \text{ and} \tag{7.120}$$

$$\begin{aligned}\|\theta^*\|_1-\|\theta\|_1 &= \|\theta^*_{\mathcal{S}^*}\|_1-\|\theta_{\mathcal{S}^*}\|_1-\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1\\ &\leq \|(\theta-\theta^*)_{\mathcal{S}^*}\|_1-\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1. \end{aligned} \tag{7.121}$$

Combining (7.119), (7.120) and (7.121), we have

$$\begin{aligned}\frac{\rho^{-}_{s^*+2\tilde{s}}}{2}\|\theta-\theta^*\|_2^2 \leq \frac{6\lambda^2 s^*}{\rho^{-}_{s^*+2\tilde{s}}} &+ (\|\nabla\mathcal{L}(\theta^*)\|_\infty+\lambda)\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1\\ &+ (\|\nabla\mathcal{L}(\theta^*)\|_\infty-\lambda)\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1. \end{aligned} \tag{7.122}$$

We discuss two cases as following:

Case 1. We first assume $\|\theta-\theta^*\|_1 \leq \frac{12\lambda s^*}{\rho^{-}_{s^*+2\tilde{s}}}$. Then (7.122) implies

$$\begin{aligned}\frac{\rho^{-}_{s^*+2\tilde{s}}}{2}\|\theta-\theta^*\|_2^2 &\overset{(i)}{\leq} \frac{6\lambda^2 s^*}{\rho^{-}_{s^*+2\tilde{s}}} + (\|\nabla\mathcal{L}(\theta^*)\|_\infty+\lambda)\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1\\ &\overset{(ii)}{\leq} \frac{6\lambda^2 s^*}{\rho^{-}_{s^*+2\tilde{s}}} + \frac{3}{2}\lambda\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1 \leq \frac{24\lambda^2 s^*}{\rho^{-}_{s^*+2\tilde{s}}}. \end{aligned}$$

where (i) is from $\|\nabla\mathcal{L}(\theta^*)\|_\infty-\lambda \leq 0$ and (ii) is from $\|\nabla\mathcal{L}(\theta^*)\|_\infty+\lambda \leq \frac{3}{2}\lambda$. This indicates

$$\|\theta-\theta^*\|_2 \leq \frac{4\sqrt{3s^*}\lambda}{\rho^{-}_{s^*+2\tilde{s}}}. \tag{7.123}$$

Case 2. Next, we assume $\|\theta-\theta^*\|_1 > \frac{12\lambda s^*}{\rho^{-}_{s^*+2\tilde{s}}}$. Then (7.122) implies

$$\begin{aligned}&\frac{\rho^{-}_{s^*+2\tilde{s}}}{2}\|\theta-\theta^*\|_2^2\\ &\leq (\|\nabla\mathcal{L}(\theta^*)\|_\infty+\lambda)\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1 + (\|\nabla\mathcal{L}(\theta^*)\|_\infty-\lambda)\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1 + \frac{1}{2}\lambda\|\theta-\theta^*\|_1\\ &= (\|\nabla\mathcal{L}(\theta^*)\|_\infty+\frac{3}{2}\lambda)\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1 + (\|\nabla\mathcal{L}(\theta^*)\|_\infty-\frac{1}{2}\lambda)\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1\\ &\overset{(i)}{\leq} 2\lambda\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1 \leq 2\sqrt{s^*}\lambda\|(\theta-\theta^*)_{\mathcal{S}^*}\|_2, \end{aligned} \tag{7.124}$$

where (i) is from $\|\nabla\mathcal{L}(\theta^*)\|_\infty + \frac{3}{2}\lambda \leq 2\lambda$ and $\|\nabla\mathcal{L}(\theta^*)\|_\infty - \frac{1}{2}\lambda \leq 0$. This indicates

$$\|\theta - \theta^*\|_2 \leq \frac{4\sqrt{s^*}\lambda}{\rho^-_{s^*+2\tilde{s}}}. \tag{7.125}$$

Besides, we have

$$\|\theta - \theta^*\|_1 \overset{(i)}{\leq} 6\|(\theta - \theta^*)_{\mathcal{S}^*}\|_1 \leq 6\sqrt{s^*}\|(\theta - \theta^*)_{\mathcal{S}^*}\|_2 \leq \frac{24\lambda s^*}{\rho^-_{s^*+2\tilde{s}}}, \tag{7.126}$$

where (i) is from $\|\nabla\mathcal{L}(\theta^*)\|_\infty + \frac{3}{2}\lambda \leq 2\lambda$ and (7.124).

Combining (7.123) and (7.125), we have desired result (7.87). Combining the assumption in Case 1 and (7.126), we have desired result (7.88).

Proof of Lemma 28

Recall that the proximal-gradient update can be computed by the soft-thresholding operation,

$$\left(\mathcal{T}_{L,\lambda}(\theta)\right)_i = \operatorname{sign}(\check{\theta}_i) \max\left\{|\check{\theta}_i| - \lambda/L, 0\right\} \;\; \forall i = 1, \ldots, d,$$

where $\check{\theta} = \theta - \nabla\mathcal{L}(\theta)/L$. To bound $\|\left(\mathcal{T}_{L,\lambda}(\theta)\right)_{\overline{\mathcal{S}}^*}\|_0$, we consider

$$\check{\theta} = \theta - \frac{1}{L}\nabla\mathcal{L}(\theta) = \theta - \frac{1}{L}\nabla\mathcal{L}(\theta^*) + \frac{1}{L}\left(\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta)\right).$$

We then consider the following three events:

$$A_1 = \left\{i \in \overline{\mathcal{S}}^* : |\theta_i| \geq \lambda/(3L)\right\}, \tag{7.127}$$

$$A_2 = \left\{i \in \overline{\mathcal{S}}^* : |(\nabla\mathcal{L}(\theta^*)/L)_i| > \lambda/(6L)\right\}, \tag{7.128}$$

$$A_3 = \left\{i \in \overline{\mathcal{S}}^* : \left|\left(\nabla\mathcal{L}(\theta^*)/L - \nabla\mathcal{L}(\theta)/L\right)_i\right| \geq \lambda/(2L)\right\}, \tag{7.129}$$

Event A_1. Note that for any $i \in \overline{\mathcal{S}}^*$, $|\theta_i| = |\theta_i - \theta_i^*|$, then we have

$$\begin{aligned}|A_1| &\le \sum_{i\in\overline{\mathcal{S}}^*} \frac{3L}{\lambda}|\theta_i - \theta_i^*| \cdot \mathbb{1}(|\theta_i - \theta_i^*| \ge \lambda/(3L)) \le \frac{3L}{\lambda}\sum_{i\in\overline{\mathcal{S}}^*}|\theta_i - \theta_i^*| \le \frac{3L}{\lambda}\|\theta - \theta^*\|_1 \\ &\overset{(i)}{\le} \frac{72Ls^*}{\rho^-_{s^*+2\tilde{s}}}, \qquad (7.130)\end{aligned}$$

where (i) is from (7.88) in Lemma 27.

Event A_2. By Lemma 2, we have

$$\begin{aligned}0 \le |A_2| &\le \sum_{i\in\overline{\mathcal{S}}^*}\frac{6L}{\lambda}|(\nabla\mathcal{L}(\theta^*)/L)_i| \cdot \mathbb{1}(|(\nabla\mathcal{L}(\theta^*)/L)_i| > \lambda/(6L)) \\ &= \sum_{i\in\overline{\mathcal{S}}^*}\frac{6L}{\lambda}|(\nabla\mathcal{L}(\theta^*)/L)_i| \cdot 0 = 0, \qquad (7.131)\end{aligned}$$

which indicates that $|A_2| = 0$.

Event A_3. Consider the event $\tilde{A} = \{i : |(\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta))_i| \ge \lambda/2\}$, which satisfies $A_3 \subseteq \tilde{A}$. We will provide an upper bound of $|\tilde{A}|$, which is also an upper bound of $|A_3|$. Let $v \in \mathbb{R}^d$ be chosen such that, $v_i = \operatorname{sign}\{(\nabla\mathcal{L}(\theta^*)/L - \nabla\mathcal{L}(\theta)/L)_i\}$ for any $i \in \tilde{A}$, and $v_i = 0$ for any $i \notin \tilde{A}$. Then we have

$$\begin{aligned}v^\top(\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta)) &= \sum_{i\in\tilde{A}} v_i\left(\nabla\mathcal{L}(\theta^*)/L - \nabla\mathcal{L}(\theta)/L\right)_i = \sum_{i\in\tilde{A}}|(\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta))_i| \\ &\ge \lambda|\tilde{A}|/2. \qquad (7.132)\end{aligned}$$

On the other hand, we have

$$\begin{aligned}v^\top(\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta)) &\le \|v\|_2\|\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta)\|_2 \overset{(i)}{\le} \sqrt{|\tilde{A}|}\cdot\|\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta)\|_2 \\ &\overset{(ii)}{\le} \rho^+_{s^*+2\tilde{s}}\sqrt{|\tilde{A}|}\cdot\|\theta - \theta^*\|_2, \qquad (7.133)\end{aligned}$$

where (i) is from $\|v\|_2 \le \sqrt{|\tilde{A}|}\max\{i : |A_i|\} \le \sqrt{|\tilde{A}|}$, and (ii) is from (7.78) and (7.79).

Combining (7.165) and (7.166), we have

$$\lambda|\tilde{A}| \le 2\rho^+_{s^*+2\tilde{s}}\sqrt{|\tilde{A}|}\cdot\|\theta - \theta^*\|_2 \overset{(i)}{\le} 8\lambda\kappa_{s^*+2\tilde{s}}\sqrt{3s^*|\tilde{A}|}$$

where (i) is from (7.87) in Lemma 27 and definition of $\kappa_{s^*+2\tilde{s}} = \frac{\rho^+_{s^*+2\tilde{s}}}{\rho^-_{s^*+2\tilde{s}}}$. Considering $A_3 \subseteq \tilde{A}$, this implies

$$|A_3| \leq |\tilde{A}| \leq 196\kappa^2_{s^*+2\tilde{s}} s^*. \tag{7.134}$$

Now combining Even A_1, A_2, A_3 and $L \leq 2\rho^+_{s^*+2\tilde{s}}$ in assumption, we close the proof as

$$\begin{aligned} \| \left(\mathcal{T}_{L,\lambda}(\theta)\right)_{\overline{\mathcal{S}}^*} \|_0 &\leq |A_1| + |A_2| + |A_3| \leq \frac{72Ls^*}{\rho^-_{s^*+2\tilde{s}}} + 196\kappa^2_{s^*+2\tilde{s}} s^* \\ &\leq (144\kappa_{s^*+2\tilde{s}} + 196\kappa^2_{s^*+2\tilde{s}})s^* \leq \tilde{s}. \end{aligned}$$

Proof of Lemma 29

Let $g = \operatorname{argmin}_{g\in\partial\|\theta\|_1} \mathcal{L} + \lambda\|\theta\|_1$, then $\omega_\lambda = \|\nabla\mathcal{L} + \lambda g\|_\infty$. By the optimality of $\overline{\theta}$ and convexity of $\mathcal{F}_\lambda$, we have

$$\begin{aligned} \mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(\overline{\theta}) &\leq \left(\nabla\mathcal{L} + \lambda g\right)^\top (\theta - \overline{\theta}) \leq \|\nabla\mathcal{L} + \lambda g\|_\infty \|\theta - \overline{\theta}\|_1 \\ &\leq \left(\omega_\lambda(\theta)\right) \|\theta - \overline{\theta}\|_1. \end{aligned} \tag{7.135}$$

Besides, we have

$$\begin{aligned} \|\theta - \overline{\theta}\|_1 &\leq \|\theta - \theta^*\|_1 + \|\overline{\theta} - \theta^*\|_1 \overset{(i)}{\leq} 6\left(\|(\theta - \theta^*)_{\mathcal{S}^*}\|_1 + \|(\overline{\theta} - \theta^*)_{\mathcal{S}^*}\|_1\right) \\ &\leq 6\sqrt{s^*}\left(\|(\theta - \theta^*)_{\mathcal{S}^*}\|_2 + \|(\overline{\theta} - \theta^*)_{\mathcal{S}^*}\|_2\right) \overset{(ii)}{\leq} \frac{24\lambda s^*}{\rho^-_{s^*+2\tilde{s}}}. \end{aligned} \tag{7.136}$$

where (i) and (ii) are from (7.83) and (7.84) in Lemma 26 respectively. Combining (7.135) and (7.136), we have desired result.

Proof of Lemma 30

Our analysis has two steps. In the first step, we show that $\{\theta^{(t)}\}_{t=0}^\infty$ converges to the unique limit point $\overline{\theta}$. In the second step, we show that the proximal gradient method has linear convergence rate.

Step 1. Note that $\theta^{(t+1)} = \mathcal{T}_{L,\lambda}(\theta^{(t)})$. Since $\mathcal{F}_\lambda(\theta)$ is convex in θ (but not strongly convex), the sub-level set $\{\theta : \mathcal{F}_\lambda(\theta) \le \mathcal{F}_\lambda(\theta^{(0)})\}$ is bounded. By the monotone decrease of $\mathcal{F}_\lambda(\theta^{(t)})$ from (7.90) in Lemma 31, $\{\theta^{(t)}\}_{t=0}^{\infty}$ is also bounded. By BolzanoWeierstrass theorem, it has a convergent subsequence and we will show that $\overline{\theta}$ is the unique accumulation point.

Since $\mathcal{F}_{\lambda(\theta)}$ is bounded below,

$$\lim_{k\to\infty} \|\theta^{(t+1)} - \theta^{(t)}\|_2 \le \frac{2}{L^{(t)}} \cdot \lim_{k\to\infty} \left[\mathcal{F}_\lambda\left(\theta^{(t+1)}\right) - \mathcal{F}_\lambda\left(\theta^{(t)}\right)\right] = 0.$$

By Lemma 32, we have $\lim_{k\to\infty} \omega_\lambda(\theta^{(t)}) = 0$. This implies $\lim_{k\to\infty} \theta^{(t)}$ satisfies the KKT condition, hence is an optimal solution.

Let $\overline{\theta}$ be an accumulation point. Since $\overline{\theta} = \operatorname{argmin}_\theta \mathcal{F}_\lambda(\theta)$, then there exists some $g \in \partial\|\overline{\theta}\|_1$ such that

$$\nabla\mathcal{F}_\lambda(\overline{\theta}) = \mathcal{L}_\lambda(\overline{\theta}) + \lambda g = 0. \tag{7.137}$$

By Lemma 28, every proximal update is sparse, hence $\|\overline{\theta}_{\overline{\mathcal{S}}^*}\|_0 \le \tilde{s}$. By RSC property in (4.10), if $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \le \tilde{s}$, i.e.,$\|(\theta - \overline{\theta})_{\overline{\mathcal{S}}^*}\|_0 \le \tilde{s}$, then we have

$$\mathcal{L}(\theta) - \mathcal{L}(\overline{\theta}) \ge (\theta - \overline{\theta})^\top \nabla\mathcal{L}(\overline{\theta}) + \frac{\rho^-_{s^*+2\tilde{s}}}{2}\|\theta - \overline{\theta}\|_2^2, \tag{7.138}$$

From the convexity of $\|\theta\|_1$ and $g \in \partial\|\overline{\theta}\|_1$, we have

$$\|\theta\|_1 - \|\overline{\theta}\|_1 \ge (\theta - \overline{\theta})^\top g. \tag{7.139}$$

Combining (7.138) and (7.139), we have for any $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \le \tilde{s}$,

$$\begin{aligned}\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(\overline{\theta}) &= \mathcal{L}(\theta) + \lambda\|\theta\|_1 - \left(\mathcal{L}(\overline{\theta}) - \lambda\|\overline{\theta}\|_1\right)\\ &\ge (\theta - \overline{\theta})^\top \left(\mathcal{L}(\overline{\theta}) + \lambda g\right) + \frac{\rho^-_{s^*+2\tilde{s}}}{2}\|\theta - \overline{\theta}\|_2^2 \overset{(i)}{=} \frac{\rho^-_{s^*+2\tilde{s}}}{2}\|\theta - \overline{\theta}\|_2^2 \ge 0,\end{aligned} \tag{7.140}$$

where (i) is from (7.137). Therefore, $\overline{\theta}$ is the unique accumulation point, i.e. $\lim_{k\to\infty} \theta^{(t)} = \overline{\theta}$.

Step 2. The objective $\mathcal{F}_\lambda(\theta^{(t+1)})$ satisfies

$$\begin{aligned}\mathcal{F}_\lambda(\theta^{(t+1)}) &\overset{(i)}{\leq} \mathcal{Q}_\lambda\left(\theta^{(t+1)}, \theta^{(t)}\right)\\ &\overset{(ii)}{=} \min_\theta \mathcal{L}(\theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)})^\top(\theta - \theta^{(t)}) + \frac{\tilde{L}_\lambda^{(t)}}{2}\|\theta - \theta^{(t)}\|_2^2 + \lambda\|\theta\|_1. \qquad (7.141)\end{aligned}$$

where (i) is from (7.90) in Lemma 31, (ii) is from the definition of $\mathcal{O}_\lambda$ in (4.6). To further bound R.H.S. of (7.141), we consider the line segment $S(\overline{\theta}, \theta^{(t)}) = \{\theta : \theta = \alpha\overline{\theta} + (1-\alpha)\theta^{(t)}, \alpha \in [0,1]\}$. Then we restrict the minimization over the line segment $S(\overline{\theta}, \theta^{(t)})$,

$$\mathcal{F}_\lambda(\theta^{(t+1)}) \leq \min_{\theta \in S(\overline{\theta}, \theta^{(t)})} \mathcal{L}(\theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)})^\top(\theta - \theta^{(t)}) + \frac{\tilde{L}_\lambda^{(t)}}{2}\|\theta - \theta^{(t)}\|_2^2 + \lambda\|\theta\|_1. \quad (7.142)$$

Since $\|\overline{\theta}_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$ and $\|\theta^{(t)}_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$, then for any $\theta \in S(\overline{\theta}, \theta^{(t)})$, we have $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \leq \tilde{s}$ and $\|(\theta - \theta^{(t)})_{\overline{\mathcal{S}}^*}\|_0 \leq 2\tilde{s}$. By RSC property, we have

$$\begin{aligned}\mathcal{L}(\theta) &\geq \mathcal{L}(\theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)})^\top(\theta - \theta^{(t)}) + \frac{\rho^-_{s^*+2\tilde{s}}}{2}\|\theta - \theta^{(t)}\|_2^2\\ &\geq \mathcal{L}(\theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)})^\top(\theta - \theta^{(t)}). \qquad (7.143)\end{aligned}$$

Combining (7.142) and (7.143), we have

$$\begin{aligned}\mathcal{F}_\lambda(\theta^{(t+1)}) &\leq \min_{\theta \in S(\overline{\theta}, \theta^{(t)})} \mathcal{L}(\theta) + \frac{\tilde{L}_\lambda^{(t)}}{2}\|\theta - \theta^{(t)}\|_2^2 + \lambda\|\theta\|_1\\ &= \min_{\alpha \in [0,1]} \mathcal{F}_\lambda(\alpha\overline{\theta} + (1-\alpha)\theta^{(t)}) + \frac{\alpha^2\tilde{L}_\lambda^{(t)}}{2}\|\overline{\theta} - \theta^{(t)}\|_2^2\\ &\overset{(i)}{\leq} \min_{\alpha \in [0,1]} \alpha\mathcal{F}_\lambda(\overline{\theta}) + (1-\alpha)\mathcal{F}_\lambda(\theta^{(t)}) + \frac{\alpha^2\tilde{L}_\lambda^{(t)}}{2}\|\overline{\theta} - \theta^{(t)}\|_2^2\\ &\overset{(ii)}{\leq} \min_{\alpha \in [0,1]} \mathcal{F}_\lambda(\theta^{(t)}) - \alpha\left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\right) + \frac{\alpha^2\tilde{L}_\lambda^{(t)}}{\rho^-_{s^*+2\tilde{s}}}\left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\right)\\ &= \min_{\alpha \in [0,1]} \mathcal{F}_\lambda(\theta^{(t)}) - \alpha\left(1 - \frac{\alpha\tilde{L}_\lambda^{(t)}}{\rho^-_{s^*+2\tilde{s}}}\right)\left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\right), \qquad (7.144)\end{aligned}$$

where (i) is from the convexity of $\mathcal{F}_\lambda$ and (ii) is from (7.140).

Minimize the R.H.S. of (7.144) w.r.t. α, the optimal value $\alpha = \frac{\rho^-_{s^*+2\tilde{s}}}{2\tilde{L}^{(t)}_\lambda}$ results in

$$\mathcal{F}_\lambda(\theta^{(t+1)}) \le \mathcal{F}_\lambda(\theta^{(t)}) - \frac{\rho^-_{s^*+2\tilde{s}}}{4\tilde{L}^{(t)}_\lambda}\left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\right). \tag{7.145}$$

Subtracting both sides of (7.145) by $\mathcal{F}_\lambda(\overline{\theta})$, we have

$$\begin{aligned}\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(\overline{\theta}) &\le \left(1 - \frac{\rho^-_{s^*+2\tilde{s}}}{4\tilde{L}^{(t)}_\lambda}\right)\left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\right)\\ &\overset{(i)}{\le} \left(1 - \frac{\rho^-_{s^*+2\tilde{s}}}{8\rho^+_{s^*+2\tilde{s}}}\right)\left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\right),\end{aligned} \tag{7.146}$$

where (i) is from Remark 8. Apply (7.146) recursively, we have the desired result.

7.3.9 Proof of Intermediate Results for Theorem 10

We also introduce an important notion as follows, which is closely related with the SE properties.

Definition 22. We denote the local ℓ_1 cone as

$$\mathcal{C}(s,\vartheta,r){=}\big\{v,\theta : \mathcal{S}\subseteq\mathcal{M}, |\mathcal{M}|\le s, \|v_{\mathcal{M}_\perp}\|_1 \le \vartheta\|v\mathcal{M}\|_1, \|\theta-\theta^*\|_2 \le r\big\}.$$

Then we define the largest and smallest **localized restricted eigenvalues** (LRE) as

$$\begin{aligned}\psi^+_{s,\vartheta,r} &= \sup_{u,\theta}\left\{\frac{v^\top\nabla^2\mathcal{L}(\theta)v}{v^\top v} : (v,\theta)\in\mathcal{C}(s,\vartheta,r)\right\},\\ \psi^-_{s,\vartheta,r} &= \inf_{u,\theta}\left\{\frac{v^\top\nabla^2\mathcal{L}(\theta)v}{v^\top v} : (v,\theta)\in\mathcal{C}(s,\vartheta,r)\right\}.\end{aligned}$$

The following proposition demonstrates the relationships between SE and LRE. The proof can be found in [244], thus is omitted here.

Proposition 5. Given any $\theta,\theta'\in\mathcal{C}(s,\vartheta,r)\cap\mathcal{B}(\theta^*,r)$, we have

$$c_1\psi^-_{s,\vartheta,r} \le \rho^-_s \le c_2\psi^-_{s,\vartheta,r}, \quad\text{and}\quad c_3\psi^+_{s,\vartheta,r} \le \rho^+_s \le c_4\psi^+_{s,\vartheta,r}.$$

where c_1, c_2, c_3, and c_4 are constants.

Proof of Lemma 44

We first demonstrate the sparsity of the update. Since $\theta^{(t+1)}$ is the minimizer to the proximal Newton problem, we have

$$\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)}) + \lambda\xi^{(t+1)} = 0,$$

where $\xi^{(t+1)} \in \partial\|\theta^{(t+1)}\|_1$.

It follows from [153] that if conditions in Lemma 2 holds, then we have $\min_{j\in\overline{\mathcal{S}}'}\{\lambda_j\} \geq \lambda/2$ for some set $\mathcal{S}' \supset \mathcal{S}$ with $|\mathcal{S}'| \leq 2s^*$. Then the analysis of sparsity of can be performed through λ directly.

We then consider the following decomposition

$$\begin{aligned}&\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)})\\ &= \underbrace{\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*)}_{V_1} + \underbrace{\nabla^2\mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)})}_{V_2} + \underbrace{\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)}_{V_3} + \underbrace{\nabla\mathcal{L}(\theta^*)}_{V_4}.\end{aligned}$$

Consider the following sets: $\mathcal{A}_i = \left\{j \in \overline{\mathcal{S}}' \ : \ |(V_i)_j| \geq \lambda/4\right\}$, for all $i \in \{1,2,3,4\}$.
Set $\mathcal{A}_2$. Suppose we choose a vector $v \in \mathbb{R}^d$ such that $v_j = \text{sign}\left\{(\nabla^2\mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)}))_j\right\}$ for all $j \in \mathcal{A}_2$ and $v_j = 0$ for $j \notin \mathcal{A}_2$. Then we have

$$\begin{aligned}v^\top\nabla^2\mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)}) &= \sum_{j\in\mathcal{A}_2} v_j(\nabla^2\mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)}))_j\\ &= \sum_{j\in\mathcal{A}_2} |(\nabla^2\mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)}))_j| \geq \lambda|\mathcal{A}_2|/4. \qquad (7.147)\end{aligned}$$

On the other hand, we have

$$\begin{aligned}v^\top\nabla^2\mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)}) &\leq \|v(\nabla^2\mathcal{L}(\theta^{(t)}))^{1/2}\|_2\|(\nabla^2\mathcal{L}(\theta^{(t)}))^{1/2}(\theta^* - \theta^{(t)})\|_2\\ &\overset{(i)}{\leq} \rho^+_{s^*+2\tilde{s}}\|v\|_2\|\theta^* - \theta^{(t)}\|_2 \overset{(ii)}{\leq} \sqrt{|\mathcal{A}_2|}\rho^+_{s^*+2\tilde{s}}\|\theta^* - \theta^{(t)}\|_2\\ &\overset{(iii)}{\leq} C'\sqrt{|\mathcal{A}_2|}\kappa_{s^*+2\tilde{s}}\lambda\sqrt{s^*}, \qquad (7.148)\end{aligned}$$

where (i) is from the SE properties, (ii) is from the definition of v, and (iii) is from $\|\theta^{(t)} - \theta^*\|_2 \leq C'\lambda\sqrt{s^*}/\rho^-_{s^*+2\tilde{s}}$. Combining (7.147) and (7.163), we have $|\mathcal{A}_2| \leq C_2\kappa^2_{s^*+2\tilde{s}}s^*$.

Set $\mathcal{A}_3$. Consider the event $\tilde{A} = \left\{i : \left|\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right)_i\right| \geq \lambda/4\right\}$, which satisfies $\mathcal{A}_3 \subseteq \tilde{A}$. We will provide an upper bound of $|\tilde{A}|$, which is also an upper bound of $|\mathcal{A}_3|$. Let $v \in \mathbb{R}^d$ be chosen such that $v_i = \text{sign}\left\{\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right)_i\right\}$ for any $i \in \tilde{A}$, and $v_i = 0$ for any $i \notin \tilde{A}$. Then we have

$$\begin{aligned} v^\top\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right) &= \sum_{i\in\tilde{A}} v_i\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right)_i = \sum_{i\in\tilde{A}}\left|\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right)_i\right| \\ &\geq \lambda|\tilde{A}|/4. \end{aligned} \tag{7.149}$$

On the other hand, we have

$$\begin{aligned} v^\top\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right) &\leq \|v\|_2\|\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\|_2 \overset{(i)}{\leq} \sqrt{|\tilde{A}|}\cdot\|\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\|_2 \\ &\overset{(ii)}{\leq} \rho^+_{s^*+2\tilde{s}}\sqrt{|\tilde{A}|}\cdot\|\theta^{(t)} - \theta^*\|_2, \end{aligned} \tag{7.150}$$

where (i) is from $\|v\|_2 \leq \sqrt{|\tilde{A}|}\max\{i : |\mathcal{A}_i|\} \leq \sqrt{|\tilde{A}|}$, and (ii) is from the mean value theorem and the SE properties.

Combining (7.149) and (7.150), we have

$$\lambda|\tilde{A}| \leq 4\rho^+_{s^*+2\tilde{s}}\sqrt{|\tilde{A}|}\cdot\|\theta - \theta^*\|_2 \overset{(i)}{\leq} 8\lambda\kappa_{s^*+2\tilde{s}}\sqrt{3s^*|\tilde{A}|}$$

where (i) is from $\|\theta^{(t)} - \theta^*\|_2 \leq C'\lambda\sqrt{s^*}/\rho^-_{s^*+2\tilde{s}}$ and definition of $\kappa_{s^*+2\tilde{s}} = \rho^+_{s^*+2\tilde{s}}/\rho^-_{s^*+2\tilde{s}}$. Considering $\mathcal{A}_3 \subseteq \tilde{A}$, this implies $|\mathcal{A}_3| \leq |\tilde{A}| \leq C_3\kappa^2_{s^*+2\tilde{s}}s^*$.

Set $\mathcal{A}_4$. By conditions in Lemma 2 and $\lambda \geq 4\|\nabla\mathcal{L}(\theta^*)\|_\infty$, we have

$$0 \leq |V_4| \leq \sum_{i\in\overline{\mathcal{S}}^*}\frac{4}{\lambda}|(\nabla\mathcal{L}(\theta^*))_i|\cdot\mathbb{1}(|(\nabla\mathcal{L}(\theta^*))_i| > \lambda/(4)) = \sum_{i\in\overline{\mathcal{S}}^*}\frac{4}{\lambda}|(\nabla\mathcal{L}(\theta^*))_i|\cdot 0 = 0,$$

Set $\mathcal{A}_1$. From Lemma 50, we have $\mathcal{F}_\lambda(\theta^{(t+1)}) \leq \mathcal{F}_\lambda(\theta^*) + \frac{\lambda}{4}\|\theta^{(t+1)} - \theta^*\|_1$. This implies

$$\begin{aligned}
\mathcal{L}(\theta^{(t+1)}) - \mathcal{L}(\theta^*) &\leq \lambda(\|\theta^*\|_1 - \|\theta^{(t+1)}\|_1) + \frac{\lambda}{4}\|\theta^{(t+1)} - \theta^*\|_1 \\
&= \lambda(\|\theta^*_{\mathcal{S}'}\|_1 - \|\theta^{(t+1)}_{\mathcal{S}'}\|_1 - \|\theta^{(t+1)}_{\mathcal{S}'_\perp}\|_1) + \frac{\lambda}{4}\|\theta^{(t+1)} - \theta^*\|_1 \\
&\leq \frac{5\lambda}{4}\|\theta^{(t+1)}_{\mathcal{S}'} - \theta^*_{\mathcal{S}'}\|_1 - \frac{3\lambda}{4}\|\theta^{(t+1)}_{\mathcal{S}'_\perp} - \theta^*_{\mathcal{S}'_\perp}\|_1. \qquad (7.151)
\end{aligned}$$

where the equality holds since $\theta^*_{\mathcal{S}'_\perp} = 0$. On the other hand, we have

$$\begin{aligned}
\mathcal{L}(\theta^{(t+1)}) - \mathcal{L}(\theta^*) &\overset{(i)}{\geq} \nabla\mathcal{L}(\theta^*)(\theta^{(t+1)} - \theta^*) \geq -\|cL(\theta^*)\|_\infty\|\theta^{(t+1)} - \theta^*\|_1 \\
&\overset{(ii)}{\geq} -\frac{\lambda}{4}\|\theta^{(t+1)} - \theta^*\|_1 = -\frac{\lambda}{4}\|\theta^{(t+1)}_{\mathcal{S}'} - \theta^*_{\mathcal{S}'}\|_1 - \frac{\lambda}{4}\|\theta^{(t+1)}_{\mathcal{S}'_\perp} - \theta^*_{\mathcal{S}'_\perp}\|_1, \\
&\qquad (7.152)
\end{aligned}$$

where (i) is from the convexity of $\mathcal{L}$ and (ii) is from conditions of Lemma 2. Combining (7.168) and (7.169), we have

$$\|\theta^{(t+1)}_{\mathcal{S}'_\perp} - \theta^*_{\mathcal{S}'_\perp}\|_1 \leq 3\|\theta^{(t+1)}_{\mathcal{S}'} - \theta^*_{\mathcal{S}'}\|_1,$$

which implies that $(\theta^{(t+1)} - \theta^*, \theta^{(t+1)}) \in \mathcal{C}(s^*, 3, r)$ with respect to the set $\mathcal{S}'$. Then we choose a vector $v \in \mathbb{R}^d$ such that $v_j = \text{sign}\left\{(\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*))_j\right\}$ for all $j \in \mathcal{A}_1$ and $v_j = 0$ for $j \notin \mathcal{A}_1$. Then we have

$$\begin{aligned}
v^\top\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*) &= \sum_{j \in \mathcal{A}_2} v_j(\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*))_j \\
&= \sum_{j \in \mathcal{A}_2} |(\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*))_j| \geq \lambda|\mathcal{A}_1|/4. \qquad (7.153)
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
v^\top\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*) &\leq \|v(\nabla^2\mathcal{L}(\theta^{(t)}))^{1/2}\|_2\|(\nabla^2\mathcal{L}(\theta^{(t)}))^{1/2}(\theta^{(t+1)} - \theta^*)\|_2 \\
&\overset{(i)}{\leq} c_1\rho^+_{s^*+2\tilde{s}}\|v\|_2\|\theta^{(t+1)} - \theta^*\|_2 \overset{(ii)}{\leq} c_1\sqrt{|\mathcal{A}_2|}\rho^+_{s^*+2\tilde{s}}\|\theta^{(t+1)} - \theta^*\|_2 \\
&\overset{(iii)}{\leq} c_2\sqrt{|\mathcal{A}_1|}\kappa_{s^*+2\tilde{s}}\lambda\sqrt{s^*}, \qquad (7.154)
\end{aligned}$$

where (i) is from the SE properties and Proposition 6, (ii) is from the definition of v, and (iii) is from $\|\theta^{(t+1)} - \theta^*\|_2 \leq C'\lambda\sqrt{s^*}/\rho^{-}_{s^*+2\tilde{s}}$. Combining (7.153) and (7.170), we have $|\mathcal{A}_1| \leq C_1\kappa^2_{s^*+2\tilde{s}}s^*$.

Combining the results for Set $\mathcal{A}_1 \sim \mathcal{A}_4$, we have that there exists some constant C_0 such that

$$\|\theta^{(t+1)}_{\overline{\mathcal{S}}}\|_0 \leq C_0\kappa^2_{s^*+2\tilde{s}}s^* \leq \tilde{s}.$$

This finishes the first part. The estimation error follows directly from Lemma 51.

Proof of Lemma 45

For notational simplicity, we introduce the following proximal operator,

$$\text{prox}_r^{H,g}(\theta) = \text{argmin}_{\theta'} r(\theta') + g^\top(\theta' - \theta) + \frac{1}{2}\|\theta' - \theta\|_H^2.$$

Then we have

$$\theta^{(t+1)} = \text{prox}_{\mathcal{R}_\lambda^{\ell_1}(\theta^{(t)})}^{\nabla^2\mathcal{L}\left(\theta^{(t)}\right),\nabla\mathcal{L}\left(\theta^{(t)}\right)}\left(\theta^{(t)}\right).$$

By Lemma 44, we have

$$\|\theta^{(t+1)}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}.$$

By the KKT condition of function $\min \mathcal{F}_\lambda$, i.e., $-\nabla\mathcal{L}(\overline{\theta}) \in \partial\mathcal{R}_\lambda^{\ell_1}(\overline{\theta})$, we also have

$$\overline{\theta} = \text{prox}_{\mathcal{R}_\lambda^{\ell_1}(\overline{\theta})}^{\nabla^2\mathcal{L}\left(\theta^{(t)}\right),\nabla\mathcal{L}(\overline{\theta})}\left(\overline{\theta}\right).$$

By monotonicity of sub-gradient of a convex function, we have the *strictly non-expansive* property: for any $\theta, \theta' \in \mathbb{R}$, let $u = \text{prox}_r^{H,g}(\theta)$ and $v = \text{prox}_r^{H,g'}(\theta')$, then

$$(u - v)^\top H(\theta - \theta') - (u - v)^\top \left(g - g'\right) \geq \|u - v\|_H^2.$$

Thus by the strictly non-expansive property of the proximal operator, we obtain

$$\begin{aligned}\|\theta^{(t+1)}-\overline{\theta}\|^2_{\nabla^2\mathcal{L}(\overline{\theta})} &\leq \left(\theta^{(t+1)}-\overline{\theta}\right)^\top\left[\nabla^2\mathcal{L}(\theta^{(t)})\left(\theta^{(t)}-\overline{\theta}\right)+\left(\nabla\mathcal{L}(\overline{\theta})-\nabla\mathcal{L}(\theta^{(t)})\right)\right]\\ &\leq \|\theta^{(t+1)}-\overline{\theta}\|_2\left\|\nabla^2\mathcal{L}(\theta^{(t)})\left(\theta^{(t)}-\overline{\theta}\right)+\left(\nabla\mathcal{L}(\overline{\theta})-\nabla\mathcal{L}(\theta^{(t)})\right)\right\|_2. \end{aligned}\tag{7.155}$$

Note that both $\|\theta^{(t+1)}\|_0 \leq \tilde{s}$ and $\|\overline{\theta}\|_0 \leq \tilde{s}$. On the other hand, from the SE properties, we have

$$\|\theta^{(t+1)}-\overline{\theta}\|^2_{\nabla^2\mathcal{L}(\overline{\theta})} = (\theta^{(t+1)}-\overline{\theta})^\top\nabla^2\mathcal{L}(\overline{\theta})(\theta^{(t+1)}-\overline{\theta}) \geq \rho^-_{s^*+2\tilde{s}}\|\theta^{(t+1)}-\overline{\theta}\|_2^2. \tag{7.156}$$

Combining (7.171) and (7.172), we have

$$\begin{aligned}\left\|\theta^{(t+1)}-\overline{\theta}\right\|_2 &\leq \frac{1}{\rho^-_{s^*+2\tilde{s}}}\left\|\nabla^2\mathcal{L}(\theta^{(t)})\left(\theta^{(t)}-\overline{\theta}\right)+\left(\nabla\mathcal{L}(\overline{\theta})-\nabla\mathcal{L}(\theta^{(t)})\right)\right\|_2\\ &= \frac{1}{\rho^-_{s^*+2\tilde{s}}}\left\|\int_0^1\left[\nabla^2\mathcal{L}\left(\theta^{(t)}+\tau\left(\overline{\theta}-\theta^{(t)}\right)\right)-\nabla^2\mathcal{L}\left(\theta^{(t)}\right)\right]\cdot\left(\overline{\theta}-\theta^{(t)}\right)d\tau\right\|_2\\ &\leq \frac{1}{\rho^-_{s^*+2\tilde{s}}}\int_0^1\left\|\left[\nabla^2\mathcal{L}\left(\theta^{(t)}+\tau\left(\overline{\theta}-\theta^{(t)}\right)\right)-\nabla^2\mathcal{L}\left(\theta^{(t)}\right)\right]\cdot\left(\overline{\theta}-\theta^{(t)}\right)\right\|_2 d\tau\\ &\leq \frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}}\left\|\theta^{(t)}-\overline{\theta}\right\|_2^2,\end{aligned}$$

where the last inequality is from the local restricted Hessian smoothness of $\mathcal{L}$. Then we finish the proof by the definition of r.

Proof of Lemma 46

Suppose the step size $\eta_t < 1$. Note that we do not need the step size to be $\eta_t = 1$ in Lemma 44 and Lemma 45. We denote $\Delta\theta^{(t)} = \theta^{(t+1/2)} - \theta^{(t)}$. Then we have

$$\begin{aligned}\left\|\Delta\theta^{(t)}\right\|_2 &\overset{(i)}{\leq} \left\|\theta^{(t)}-\overline{\theta}\right\|_2+\left\|\theta^{(t+1/2)}-\overline{\theta}\right\|_2 \overset{(ii)}{\leq} \left\|\theta^{(t)}-\overline{\theta}\right\|_2+\frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}}\left\|\theta^{(t)}-\overline{\theta}\right\|_2^2\\ &\overset{(iii)}{\leq} \frac{3}{2}\left\|\theta^{(t)}-\overline{\theta}\right\|_2,\end{aligned}\tag{7.157}$$

where (i) is from triangle inequality, (ii) is from Lemma 45, and (iii) is from $\left\|\theta^{(t)}-\overline{\theta}\right\|_2 \leq r \leq \frac{\rho^-_{s^*+2\tilde{s}}}{L_{s^*+2\tilde{s}}}$.

By Lemma 44, we have $\left\|\Delta\theta^{(t)}{}_{\overline{S}}\right\|_0 \leq 2\tilde{s}$. To show $\eta_t = 1$, it is now suffice to demonstrate that

$$\mathcal{F}_\lambda(\theta^{(t+1/2)}) - \mathcal{F}_\lambda(\theta^{(t)}) \leq \frac{1}{4}\gamma_t.$$

By expanding $\mathcal{F}_\lambda$, we have

$$\begin{aligned}
&\mathcal{F}_\lambda(\theta^{(t)} + \Delta\theta^{(t)}) - \mathcal{F}_\lambda(\theta^{(t)}) = \mathcal{L}(\theta^{(t)} + \Delta\theta^{(t)}) - \mathcal{L}(\theta^{(t)}) + \mathcal{R}_\lambda^{\ell_1}(\theta^{(t)} + \Delta\theta^{(t)}) - \mathcal{R}_\lambda^{\ell_1}(\theta^{(t)}) \\
&\overset{(i)}{\leq} \nabla\mathcal{L}(\theta^{(t)})^\top\Delta\theta^{(t)} + \frac{1}{2}\Delta(\theta^{(t)})^\top\nabla^2\mathcal{L}(\theta)\Delta\theta^{(t)} + \frac{L_{s^*+2\tilde{s}}}{6}\left\|\Delta\theta^{(t)}\right\|_2^3 \\
&\qquad\quad + \mathcal{R}_\lambda^{\ell_1}(\theta^{(t)} + \Delta\theta^{(t)}) - \mathcal{R}_\lambda^{\ell_1}(\theta^{(t)}) \\
&\overset{(ii)}{\leq} \gamma_t - \frac{1}{2}\gamma_t + \frac{L_{s^*+2\tilde{s}}}{6}\left\|\Delta\theta^{(t)}\right\|_2^3 \overset{(iii)}{\leq} \frac{1}{2}\gamma_t + \frac{L_{s^*+2\tilde{s}}}{6\rho^{-}_{s^*+2\tilde{s}}}\left\|\Delta\theta^{(t)}\right\|_{\nabla^2\mathcal{L}(\theta)}\left\|\Delta\theta^{(t)}\right\|_2 \\
&\overset{(iv)}{\leq} \left(\frac{1}{2} - \frac{L_{s^*+2\tilde{s}}}{6\rho^{-}_{s^*+2\tilde{s}}}\left\|\Delta\theta^{(t)}\right\|_2\right)\gamma_t \overset{(v)}{\leq} \frac{1}{4}\gamma_t,
\end{aligned}$$

where (i) is from the restricted Hessian smooth condition, (ii) and (iv) are from Lemma 47, (iii) is from the same argument of (7.172), and (v) is from (7.173), $\gamma_t < 0$, and $\left\|\theta^{(t)} - \overline{\theta}\right\|_2 \leq r \leq \frac{\rho^{-}_{s^*+2\tilde{s}}}{L_{s^*+2\tilde{s}}}$. This implies $\theta^{(t+1)} = \theta^{(t+1/2)}$.

Proof of Lemma 47

We denote $H = \nabla^2\mathcal{L}(\theta^{(t)})$. Since $\Delta\theta^{(t)}$ is the solution for

$$\min_{\Delta\theta^{(t)}} \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{1}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \mathcal{R}_\lambda^{\ell_1}\left(\theta^{(t)} + \Delta\theta^{(t)}\right)$$

then for any $\eta_t \in (0, 1]$, we have

$$\begin{aligned}
&\eta_t\nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{\eta_t^2}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \mathcal{R}_\lambda^{\ell_1}\left(\theta^{(t)} + \eta_t\Delta\theta^{(t)}\right) \\
&\geq \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{1}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \mathcal{R}_\lambda^{\ell_1}\left(\theta^{(t)} + \Delta\theta^{(t)}\right)
\end{aligned}$$

By the convexity of $\mathcal{R}_\lambda^{\ell_1}$, we have

$$\begin{aligned}
&\eta_t \nabla \mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{\eta_t^2}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \eta_t \mathcal{R}_\lambda^{\ell_1}\left(\theta^{(t)} + \Delta\theta^{(t)}\right) + (1-\eta_t)\mathcal{R}_\lambda^{\ell_1}(\theta^{(t)}) \\
&\geq \nabla \mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{1}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \mathcal{R}_\lambda^{\ell_1}\left(\theta^{(t)} + \Delta\theta^{(t)}\right).
\end{aligned}$$

Rearranging the terms, we obtain

$$(1-\eta_t)\left(\nabla \mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \mathcal{R}_\lambda^{\ell_1}\left(\theta^{(t)} - \Delta\theta^{(t)}\right) - \mathcal{R}_\lambda^{\ell_1}(\theta^{(t)})\right) + \frac{1-\eta_t^2}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 \leq 0$$

Canceling the $(1-\eta_t)$ factor from both sides and let $\eta_t \to 1$, we obtain the desired inequality,

$$\gamma_t \leq -\left\|\Delta\theta^{(t)}\right\|_H^2.$$

Proof of Lemma 48

We first demonstrate an upper bound of the approximate KKT parameter ω_λ. Given the solution $\theta^{(t-1)}$ from the $(t-1)$-th iteration, the optimal solution at t-th iteration satisfies the KKT condition:

$$\nabla^2 \mathcal{L}(\theta^{(t-1)})(\theta^{(t)} - \theta^{(t-1)}) + \nabla \mathcal{L}(\theta^{(t-1)}) + \lambda \xi^{(t)} = 0,$$

where $\xi^{(t)} \in \partial\|\theta^{(t)}\|_1$. Then for any vector v with $\|v\|_2 \leq \|v\|_1 = 1$ and $\|v\|_0 \leq s^* + 2\tilde{s}$, we have

$$\begin{aligned}
(\nabla \mathcal{L}(\theta^{(t)}) + \lambda\xi^{(t)})^\top v &= (\nabla \mathcal{L}(\theta^{(t)}))^\top v - (\nabla^2 \mathcal{L}(\theta^{(t-1)})(\theta^{(t)} - \theta^{(t-1)}) + \nabla \mathcal{L}(\theta^{(t-1)}))^\top v \\
&= (\nabla \mathcal{L}(\theta^{(t)}) - \nabla \mathcal{L}(\theta^{(t-1)}))^\top v - (\nabla^2 \mathcal{L}(\theta^{(t-1)})(\theta^{(t)} - \theta^{(t-1)}))^\top v \\
&\overset{(i)}{\leq} \left\|(\nabla^2 \mathcal{L}(\tilde{\theta}))^{1/2}(\theta^{(t)} - \theta^{(t-1)})\right\|_2 \cdot \left\|v^\top (\nabla^2 \mathcal{L}(\tilde{\theta}))^{1/2}\right\|_2 \\
&\quad + \left\|(\nabla^2 \mathcal{L}(\theta^{(t-1)}))^{1/2}(\theta^{(t)} - \theta^{(t-1)})\right\|_2 \cdot \left\|v^\top (\nabla^2 \mathcal{L}(\theta^{(t-1)}))^{1/2}\right\|_2 \\
&\overset{(ii)}{\leq} 2\rho_{s^*+2\tilde{s}}^+ \left\|\theta^{(t)} - \theta^{(t-1)}\right\|_2,
\end{aligned} \tag{7.158}$$

where (i) is from mean value theorem with some $\tilde{\theta} = (1-a)\theta^{(t-1)} + a\theta^{(t)}$ for some $a \in [0,1]$ and Cauchy-Schwarz inequality, and (ii) is from the SE properties. Take the supremum of the L.H.S. of (7.174) with respect to v, we have

$$\left\|\nabla \mathcal{L}(\theta^{(t)}) + \lambda \xi^{(t)}\right\|_\infty \leq 2\rho^+_{s^*+2\tilde{s}} \left\|\theta^{(t)} - \theta^{(t-1)}\right\|_2. \tag{7.159}$$

Then from Lemma 45, we have

$$\left\|\theta^{(t+1)} - \overline{\theta}\right\|_2 \leq \left(\frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}}\right)^{1+2+4+\ldots+2^{t-1}} \left\|\theta^{(0)} - \overline{\theta}\right\|_2^{2^{\top}} \leq \left(\frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}} \left\|\theta^{(0)} - \overline{\theta}\right\|_2\right)^{2^t}.$$

By (7.175) and (7.173) by taking $\Delta\theta^{(t-1)} = \theta^{(t)} - \theta^{(t-1)}$, we obtain

$$\begin{aligned} \omega_\lambda\left(\theta^{(t)}\right) &\leq 2\rho^+_{s^*+2\tilde{s}} \left\|\theta^{(t)} - \theta^{(t-1)}\right\|_2 \leq 3\rho^+_{s^*+2\tilde{s}} \left\|\theta^{(t-1)} - \overline{\theta}\right\|_2 \\ &\leq 3\rho^+_{s^*+2\tilde{s}} \left(\frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}} \left\|\theta^{(0)} - \overline{\theta}\right\|_2\right)^{2^t}. \end{aligned}$$

By requiring the R.H.S. equal to ε we obtain

$$\begin{aligned} t &= \log \frac{\log\left(\frac{3\rho^+_{s^*+2\tilde{s}}}{\varepsilon}\right)}{\log\left(\frac{2\rho^-_{s^*+2\tilde{s}}}{L_{s^*+2\tilde{s}}\left\|\theta^{(0)} - \overline{\theta}\right\|_2}\right)} = \log\log\left(\frac{3\rho^+_{s^*+2\tilde{s}}}{\varepsilon}\right) - \log\log\left(\frac{2\rho^-_{s^*+2\tilde{s}}}{L_{s^*+2\tilde{s}}\left\|\theta^{(0)} - \overline{\theta}\right\|_2}\right) \\ &\overset{(i)}{\leq} \log\log\left(\frac{3\rho^+_{s^*+2\tilde{s}}}{\varepsilon}\right) - \log\log 4 \leq \log\log\left(\frac{3\rho^+_{s^*+2\tilde{s}}}{\varepsilon}\right), \end{aligned}$$

where (i) is from the fact that $\left\|\theta^{(0)} - \overline{\theta}\right\|_2 \leq r = \frac{\rho^-_{s^*+2\tilde{s}}}{2L_{s^*+2\tilde{s}}}$.

Lemma 39. Given $\omega_\lambda(\theta^{(t)}) \leq \frac{\lambda}{4}$, we have

$$\mathcal{F}_\lambda(\theta^{(t)}) \leq \mathcal{F}_\lambda(\theta^*) + \frac{\lambda}{4}\|\theta^{(t)} - \theta^*\|_1.$$

Proof. For some $\xi^{(t)} = \operatorname{argmin}_{\xi \in \partial\|\theta^{(t)}\|_1} \|\nabla\mathcal{L}(\theta^{(t)}) + \lambda\xi\|_\infty$, we have

$$\begin{aligned}
\mathcal{F}_\lambda(\theta^*) &\overset{(i)}{\geq} \mathcal{F}_\lambda(\theta^{(t)}) - (\nabla\mathcal{L}(\theta^{(t)}) + \lambda\xi^{(t)})^\top(\theta^{(t)} - \theta^*) \\
&\geq \mathcal{F}_\lambda(\theta^{(t)}) - \|\nabla\mathcal{L}(\theta^{(t)}) + \lambda\xi^{(t)}\|_\infty\|\theta^{(t)} - \theta^*\|_1 \\
&\overset{(ii)}{\geq} \mathcal{F}_\lambda(\theta^{(t)}) - \frac{\lambda}{4}\|\theta^{(t)} - \theta^*\|_1
\end{aligned}$$

where (i) is from the convexity of $\mathcal{F}_\lambda$ and (ii) is from the fact that for all $t \geq 0$, $\mathcal{F}_\lambda(\theta^{(t)}) \leq \mathcal{F}_\lambda(\theta^{(t-1)})$ and $\omega_\lambda(\theta^{(t)}) \leq \frac{\lambda}{4}$. This finishes the proof. □

Lemma 40 (Adapted from [153])**.** Suppose $\|\theta^{(t)}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}$ and $\omega_\lambda(\theta^{(t)}) \leq \frac{\lambda}{4}$. Then there exists a generic constant c_1 such that $\|\theta^{(t)} - \theta^*\|_2 \leq \frac{c_1\lambda\sqrt{s^*}}{\rho^-_{s^*+2\tilde{s}}}$.

7.4 Proofs for Chapter 5

7.4.1 Proofs of Main Results

We provide proof sketches for the main results of Theorem 12 and 13 in this section.

Proof of Theorem 12

We provide a few important intermediate results. The first result characterizes the sparsity of the solution and an upper bound of the objective after sufficiently many iterations as follows. The proof is provided in Appendix 7.4.4.

Lemma 41. Suppose that Assumptions 1 ∼ 4 hold. After sufficiently many iterations $T < \infty$, the following results hold for all $t \geq T$:

$$\|\theta^{(t)}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s} \quad \text{and} \quad \mathcal{F}_{\lambda\{1\}}(\theta^{(t)}) \leq \mathcal{F}_{\lambda\{1\}}(\theta^*) + \frac{15\lambda^2_{\text{tgt}}s^*}{4\rho^-_{s^*+2\tilde{s}}}.$$

We then demonstrate the parameter estimation and quadratic convergence conditioning on the sparse solution and bounded objective. The proof is provided in Appendix 7.4.4.

Lemma 42. Suppose that Assumptions 1 $\sim$ 4 hold. If $\|\theta_{\overline{\mathcal{S}}}^{(t)}\|_0 \leq \tilde{s}$, and $\mathcal{F}_{\lambda^{\{1\}}}(\theta^{(t)}) \leq \mathcal{F}_{\lambda^{\{1\}}}(\theta^*) + \frac{15\lambda_{\mathrm{tgt}}^2 s^*}{4\rho_{s^*+2\tilde{s}}^-}$, we have

$$\|\theta^{(t)} - \theta^*\|_2 \leq \frac{18\lambda_{\mathrm{tgt}}\sqrt{s^*}}{\rho_{s^*+2\tilde{s}}^-} \quad \text{and} \quad \|\theta^{(t+1)} - \overline{\theta}^{\{1\}}\|_2 \leq \frac{L_{s^*+2\tilde{s}}}{2\rho_{s^*+2\tilde{s}}^-}\|\theta^{(t)} - \overline{\theta}^{\{1\}}\|_2^2$$

Moreover, we characterize the sufficient number of iterations for the proximal Newton updates to achieve the approximate KKT condition. The proof is provided in Appendix 7.4.4.

Lemma 43. Suppose that Assumptions 1 $\sim$ 4 hold. If $\|\theta_{\overline{\mathcal{S}}}^{(T)}\|_0 \leq \tilde{s}$, and $\mathcal{F}_{\lambda^{\{1\}}}(\theta^{(T)}) \leq \mathcal{F}_{\lambda^{\{1\}}}(\theta^*) + \frac{15\lambda_{\mathrm{tgt}}^2 s^*}{4\rho_{s^*+2\tilde{s}}^-}$ at some iteration T, we need at most

$$T_1 \leq \log\log\left(\frac{3\rho_{s^*+2\tilde{s}}^+}{\varepsilon}\right)$$

extra iterations of the proximal Newton updates such that $\omega_{\lambda^{\{1\}}}(\theta^{(T+T_1)}) \leq \frac{\lambda_{\mathrm{tgt}}}{8}$.

Combining Lemma 41 $\sim$ 43, we have desired results in Theorem 12.

Proof of Theorem 13

We present a few important intermediate results that are key components of our main proof. The first result shows that in a neighborhood of the true model parameter θ^*, the sparsity of the solution is preserved when we use a sparse initialization. The proof is provided in Appendix 7.4.3.

Lemma 44 (Sparsity Preserving Lemma)**.** Suppose that Assumptions 1 and 2 hold with $\varepsilon \leq \frac{\lambda_{\mathrm{tgt}}}{8}$. Given $\theta^{(t)} \in \mathcal{B}(\theta^*, R)$ and $\|\theta_{\overline{\mathcal{S}}}^{(t)}\|_0 \leq \tilde{s}$, there exists a generic constant C_1 such that

$$\|\theta_{\overline{\mathcal{S}}}^{(t+1)}\|_0 \leq \tilde{s} \quad \text{and} \quad \|\theta^{(t+1)} - \theta^*\|_2 \leq \frac{C_1\lambda_{\mathrm{tgt}}\sqrt{s^*}}{\rho_{s^*+2\tilde{s}}^-}.$$

We then show that every step of proximal Newton updates within each stage has a quadratic convergence rate to a local minimizer, if we start with a sparse solution in the refined region. The proof is provided in Appendix 7.4.3.

Lemma 45. Suppose that Assumptions 1 $\sim$ 4 hold. If $\theta^{(t)} \in \mathcal{B}(\theta^*, R)$ and $\left\|\theta^{(t)}{}_{\overline{\mathcal{S}}}\right\|_0 \leq \tilde{s}$, then for each stage $K \geq 2$, we have

$$\|\theta^{(t+1)} - \overline{\theta}^{\{K\}}\|_2 \leq \frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}}\|\theta^{(t)} - \overline{\theta}^{\{K\}}\|_2^2.$$

In the following, we need to use the property that the iterates $\theta^{(t)} \in \mathcal{B}(\overline{\theta}^{\{K\}}, 2R)$ instead of $\theta^{(t)} \in \mathcal{B}(\theta^*, R)$ for convergence analysis of the proximal Newton method. This property holds since we have $\theta^{(t)} \in \mathcal{B}(\theta^*, R)$ and $\overline{\theta}^{\{K\}} \in \mathcal{B}(\theta^*, R)$ simultaneously. Thus $\theta^{(t)} \in \mathcal{B}\left(\overline{\theta}^{\{K\}}, 2R\right)$, where $2R = \frac{\rho^-_{s^*+2\tilde{s}}}{L_{s^*+2\tilde{s}}}$ is the radius for quadratic convergence region of the proximal Newton algorithm.

The following lemma demonstrates that the step size parameter is simply 1 if the the sparse solution is in the refined region. The proof is provided in Appendix 7.4.3.

Lemma 46. Suppose that Assumptions 1 $\sim$ 4 hold. If $\theta^{(t)} \in \mathcal{B}(\overline{\theta}^{\{K\}}, 2R)$ and $\|\theta^{(t)}{}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}$ at each stage $K \geq 2$ with $\frac{1}{4} \leq \alpha < \frac{1}{2}$, then $\eta_t = 1$. Further, we have

$$\mathcal{F}_{\lambda^{\{K\}}}(\theta^{(t+1)}) \leq \mathcal{F}_{\lambda^{\{K\}}}(\theta^{(t)}) + \frac{1}{4}\gamma_t.$$

Moreover, we present a critical property of γ_t. The proof is provided in Appendix 7.4.3.

Lemma 47. Denote $\Delta\theta^{(t)} = \theta^{(t)} - \theta^{(t+1)}$ and

$$\gamma_t = \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \|\lambda^{\{K\}} \odot \left(\theta^{(t)} + \Delta\theta^{(t)}\right)\|_1 - \|\lambda^{\{K\}} \odot \left(\theta^{(t)}\right)\|_1.$$

Then we have $\gamma_t \leq -\|\Delta\theta^{(t)}\|^2_{\nabla^2\mathcal{L}(\theta^{(t)})}$.

In addition, we present the sufficient number of iterations for each convex relaxation stage to achieve the approximate KKT condition. The proof is provided in Appendix 7.4.3.

Lemma 48. Suppose that Assumptions 1 $\sim$ 4 hold. To achieve the approximate KKT condition $\omega_{\lambda^{\{K\}}}\left(\theta^{(t)}\right) \leq \varepsilon$ for any $\varepsilon > 0$ at each stage $K \geq 2$, the number of iteration for proximal Newton updates is at most

$$\log\log\left(\frac{3\rho^{+}_{s^*+2\tilde{s}}}{\varepsilon}\right).$$

We further present the contraction of the estimation error along consecutive stages, which is a direct result from oracle statistical rate in [153].

Lemma 49. Suppose that Assumptions 1 $\sim$ 4 hold. Then there exists a generic constant c_1 such that the output solutions for all $K \geq 2$ satisfy

$$\|\hat{\theta}^{\{K\}} - \theta^*\|_2 \leq c_1 \left(\|\nabla \mathcal{L}(\theta^*)_{\mathcal{S}}\|_2 + \lambda_{\text{tgt}} \sqrt{\sum_{j \in \mathcal{S}} \mathbb{1}(|\theta^*_j| \leq \beta\lambda)} + \varepsilon\sqrt{s^*} \right) + 0.7\|\hat{\theta}^{\{K-1\}} - \theta^*\|_2.$$

Combining Lemma 44 $\sim$ Lemma 47, we have the quadratic convergence of the proximal Newton algorithm within each convex relaxation stage. The rest of the results in Theorem 13 hold by further combining Lemma 48 and recursively applying Lemma 49.

7.4.2 Coordinate Descent Algorithms with the Active Set Strategy

We first provide a brief derivation of the quadratic approximation (5.5) into a weighted least square problem. For notational convenience, we omit the indices $\{K\}$ and (t) for a particular iteration of a stage. Recall that we want to minimize the following ℓ_1-regularized quadratic problem

$$\Delta\hat{\theta} = \underset{\Delta\theta}{\text{argmin}} \;\; \Delta\theta^\top \nabla \mathcal{L}(\theta) + \frac{1}{2}\Delta\theta^\top \nabla^2 \mathcal{L}(\theta)\Delta\theta + \|\lambda \odot (\theta + \Delta\theta)\|_1. \tag{7.160}$$

For GLM, we have

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i^\top \theta) - y_i x_i^\top \theta,$$

where ψ is the cumulant function. Then we can rewrite the quadratic function $\Delta\theta^\top \nabla \mathcal{L}(\theta) + \frac{1}{2}\Delta\theta^\top \nabla^2 \mathcal{L}(\theta)\Delta\theta$ in subproblem (7.160) as a *weighted least squares* form

[159]:

$$\frac{1}{2n}\sum_{i=1}^{n}\Big(2\Big(y_i-\psi'(x_i^\top\theta)\Big)x_i^\top\Delta\theta+\psi''(x_i^\top\theta)(x_i^\top\Delta\theta)^2\Big)$$
$$=\frac{1}{2n}\sum_{i=1}^{n}w_i(z_i-x_i^\top\Delta\theta)^2+\text{constant},$$

where $w_i=\psi''(x_i^\top\theta)$, $z_i=\frac{y_i-\psi'(x_i^\top\theta)}{\psi''(x_i^\top\theta)}$, and the constant term does not depend on $\Delta\theta$. This indicates that (7.160) is equivalent to a Lasso problem with reweighted least square loss function:

$$\Delta\hat{\theta}=\underset{\Delta\theta}{\operatorname{argmin}}\ \frac{1}{2n}\sum_{i=1}^{n}w_i(z_i-x_i^\top\Delta\theta)^2+\|\lambda\odot(\theta+\Delta\theta)\|_1. \tag{7.161}$$

By solving (7.161), we can avoid directly computing the $d\times d$ Hessian matrix $\nabla^2\mathcal{L}(\theta)$ in (7.160) and significantly reduce the memory usage when d is large.

We then introduce an algorithm for solving (7.161) leveraging the idea of active set update. The active set update scheme is very efficient in practice [159] with rigid theoretical justifications [158]. The algorithm contains two nested loops. In the *outer loop*, we separate all coordinates into two sets: active set and inactive set. Such a partition is based on some heuristic greedy scheme, such as gradient thresholding (also called strong rule, [245]). Then within each iteration of the middle loop, the *inner loop* only updates coordinates in the active set in a cyclic manner until convergence, where the coordinates in the inactive set remain to be zero. After the inner loop converges, we update the active set based on a greedy selection rule that further decreases the objective value, and repeat the inner loop. Such a procedure continues until the active set no longer changes in the outer loop. We provide the algorithm description as follows and refer [158] for further details of active set based coordinate minimization. We use (p) to index the p-th iteration of the outer loop, and (p,l) to index the l-th iteration of the inner loop at the p-th iteration of the outer loop.

Inner Loop. The active set $\mathcal{A}$ and inactive set $\mathcal{A}_\perp$ are respectively set as

$$\mathcal{A}\leftarrow\{j\ :\ \theta_j\neq 0\}=\{j_1,j_2,\ldots,j_s\}\ \text{ and }\ \mathcal{A}_\perp\leftarrow\{j\ :\ j\notin\mathcal{A}\},$$

where $j_1 < j_2 < \ldots < j_s$. A coordinate-wise minimization of (7.161) is performed throughout the inner loop. Specifically, given $\theta^{(p,l)}$ at the l-th iteration of the inner loop, we solve (7.161) by only considering the j-th coordinate in the active set and fix the rest coordinates in a cyclic manner for all $j = j_1, j_2, \ldots, j_s$, i.e.,

$$\Delta\hat{\theta}_j = \underset{\Delta\theta_j}{\operatorname{argmin}} \ \frac{1}{2n}\sum_{i=1}^{n} w_i(z_i - \sum_{k\in\mathcal{A},k\neq j} x_{ik}^\top \Delta\theta_k - x_{ij}^\top \Delta\theta_j)^2 + |\lambda_j(\theta_j + \Delta\theta_j)|. \tag{7.162}$$

Then we update $\theta_j^{(p,l+1)} = \theta_j^{(p,l)} + \Delta\hat{\theta}_j$. Solving (7.162) has a simple closed form solution by soft thresholding, i.e.,

$$\Delta\hat{\theta}_j \leftarrow \frac{S(\frac{1}{n}\sum_{i=1}^n w_i\delta_{ij}, \lambda_j)}{\frac{1}{n}\sum_{i=1}^n w_i x_{ij}^2},$$

where $\delta_{ij} = z_i - \sum_{k\in\mathcal{A},k\neq j} x_{ik}\Delta\theta_k$ and $S(a,b) = \operatorname{sign}(a)\max\{|a| - b, 0\}$ for real values a and b. Moreover, the residual δ_{ij} can be updated efficiently. Specifically, after the update of $\Delta\hat{\theta}_j$ for the j-th coordinate, then for the next non-zero coordinate, e.g., $j' \in \mathcal{A}$, we update the residual as

$$\delta_{ij'} = \delta_{ij} - x_{ij}^\top \Delta\hat{\theta}_j + x_{ij'}\Delta\theta_{j'}.$$

This reduces the computational cost of updating each coordinate from $\mathcal{O}(s)$ to $\mathcal{O}(1)$, only with an increase of the memory cost $\mathcal{O}(s)$ for maintaining the previous updates of $\Delta\theta_j$.

Given a convergence parameter $a \in (0,1)$, we terminate the inner loop when

$$\|\theta^{(p,l+1)} - \theta^{(p,l)}\|_2 \leq a\lambda.$$

Outer Loop. At the beginning of the outer loop, we initialize the active set $\mathcal{A}^{(0)}$ as follows

$$\mathcal{A}^{(0)} \leftarrow \{j \ : \ |\nabla_j\mathcal{L}(\theta^{(0)})| \geq (1-\nu)\lambda\} \cup \{j \ : \ \theta_j^{(0)} \neq 0\},$$

where $\nabla_j\mathcal{L}(\theta^{(0)})$ is the j-th entry of $\nabla\mathcal{L}(\theta^{(0)})$, $\nu \in (0, 0.1)$ is a thresholding parameter, and the inactive set is $\mathcal{A}_\perp^{(0)} = \{j \ : \ j \notin \mathcal{A}^{(0)}\}$.

Suppose at the p-th iteration of the outer loop, the active set is $\mathcal{A}^{(p)}$. We then perform the inner loop introduced above using $\mathcal{A}^{(p)}$ until the convergence of the inner loop and denote $\theta^{(p+1)} = \theta^{(p,l)}$, which is the output of the inner loop. Next, we describe how to update the active set $\mathcal{A}^{(p)}$ using the following greedy selection rule.

- We first shrink the active set as follows. The active coordinate minimization (inner loop) may yield zero solutions on $\mathcal{A}^{(p)}$. We eliminate the zero coordinates of $\theta^{(p+1)}$ from $\mathcal{A}^{(p)}$, and update the intermediate active set and inactive set respectively as

$$\mathcal{A}^{(p+\frac{1}{2})} \leftarrow \{j \in \mathcal{A}^{(p)} \ : \ \theta_j^{(p+1)} \neq 0\} \ \text{ and } \ \mathcal{A}_\perp^{(p+\frac{1}{2})} \leftarrow \{j \ : \ j \notin \mathcal{A}^{(p+\frac{1}{2})}\}.$$

- We then expand the active set as follows. Denote

$$j^{(p)} = \operatorname*{argmax}_{j \in \mathcal{A}_\perp^{(p+\frac{1}{2})}} |\nabla_j \mathcal{L}(\theta^{(p+1)})|.$$

  The outer loop is terminated if

$$|\nabla_{j^{(p)}} \mathcal{L}(\theta^{(p+1)})| \leq (1+\delta)\lambda,$$

  where $\delta \ll 1$ is a real positive convergence parameter, e.g., $\delta = 10^{-5}$. Otherwise, we update the sets as

$$\mathcal{A}^{(p+1)} \leftarrow \mathcal{A}^{(p+\frac{1}{2})} \cup \{j^{(p)}\} \ \text{ and } \ \mathcal{A}_\perp^{(p+1)} \leftarrow \mathcal{A}_\perp^{(p+\frac{1}{2})} \backslash \{j^{(p)}\},$$

7.4.3 Proof of Intermediate Results for Theorem 13

For notational convenience, we denote

$$\mathcal{R}_\lambda^{\ell_1}(\theta) = \|\lambda \odot \theta\|_1.$$

We also introduce an important notion as follows, which is closely related with the SE properties.

Definition 23. We denote the local ℓ_1 cone as

$$\mathcal{C}(s,\vartheta,R)=\big\{v,\theta:\mathcal{S}\subseteq\mathcal{M},|\mathcal{M}|\leq s,\|v_{\mathcal{M}_\perp}\|_1\leq\vartheta\|v_{\mathcal{M}}\|_1,\|\theta-\theta^*\|_2\leq R\big\}.$$

Then we define the largest and smallest **localized restricted eigenvalues** (LRE) as

$$\psi^+_{s,\vartheta,R}=\sup_{u,\theta}\left\{\frac{v^\top\nabla^2\mathcal{L}(\theta)v}{v^\top v}:(v,\theta)\in\mathcal{C}(s,\vartheta,R)\right\},$$
$$\psi^-_{s,\vartheta,R}=\inf_{u,\theta}\left\{\frac{v^\top\nabla^2\mathcal{L}(\theta)v}{v^\top v}:(v,\theta)\in\mathcal{C}(s,\vartheta,R)\right\}.$$

The following proposition demonstrates the relationships between SE and LRE. The proof can be found in [142], thus is omitted here.

Proposition 6. Given any $\theta,\theta'\in\mathcal{C}(s,\vartheta,R)\cap\mathcal{B}(\theta^*,R)$, we have

$$c_1\psi^-_{s,\vartheta,R}\leq\rho^-_s\leq c_2\psi^-_{s,\vartheta,R},\quad\text{and}\quad c_3\psi^+_{s,\vartheta,R}\leq\rho^+_s\leq c_4\psi^+_{s,\vartheta,R}.$$

where c_1, c_2, c_3, and c_4 are constants.

Proof of Lemma 44

We first demonstrate the sparsity of the update. For notational convenience, we omit the stage index $\{K\}$. Since $\theta^{(t+1)}$ is the minimizer to the proximal Newton problem, we have

$$\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)}-\theta^{(t)})+\nabla\mathcal{L}(\theta^{(t)})+\lambda\odot\xi^{(t+1)}=0,$$

where $\xi^{(t+1)}\in\partial\|\theta^{(t+1)}\|_1$.

It follows from [153] that if Assumption 3 holds, then we have $\min_{j\in\overline{\mathcal{S}}'}\{\lambda_j\}\geq\lambda_{\rm tgt}/2$ for some set $\mathcal{S}'\supset\mathcal{S}$ with $|\mathcal{S}'|\leq 2s^*$. Then the analysis of sparsity of can be performed through $\lambda_{\rm tgt}$ directly.

We then consider the following decomposition

$$\begin{aligned}
&\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)}) \\
&= \underbrace{\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*)}_{V_1} + \underbrace{\nabla^2\mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)})}_{V_2} + \underbrace{\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)}_{V_3} + \underbrace{\nabla\mathcal{L}(\theta^*)}_{V_4}.
\end{aligned}$$

Consider the following sets:

$$\mathcal{A}_i = \left\{ j \in \overline{\mathcal{S}}' \;:\; |(V_i)_j| \geq \lambda_{\text{tgt}}/4 \right\}, \text{ for all } i \in \{1,2,3,4\}.$$

Set $\mathcal{A}_2$. We have $\mathcal{A}_2 = \left\{ j \in \overline{\mathcal{S}}' : |(\nabla^2\mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)}))_j| \geq \lambda_{\text{tgt}}/4 \right\}$. Consider a subset $\mathcal{S}' \subset \mathcal{A}_2$ with $|\mathcal{S}'| = s' \leq \tilde{s}$. Suppose we choose a vector $v \in \mathbb{R}^d$ such that $\|v\|_\infty = 1$ and $\|v\|_0 = s'$ with $s'\lambda_{\text{tgt}}/4 \leq v^\top \nabla^2\mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)})$. Then we have

$$\begin{aligned}
s'\lambda_{\text{tgt}}/4 &\leq v^\top \nabla^2\mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)}) \leq \|v(\nabla^2\mathcal{L}(\theta^{(t)}))^{\frac{1}{2}}\|_2 \|(\nabla^2\mathcal{L}(\theta^{(t)}))^{\frac{1}{2}}(\theta^* - \theta^{(t)})\|_2 \\
&\overset{(i)}{\leq} \sqrt{\rho^+_{s^*+2\tilde{s}}\rho^+_{s'}}\|v\|_2\|\theta^* - \theta^{(t)}\|_2 \overset{(ii)}{\leq} \sqrt{s'\rho^+_{s^*+2\tilde{s}}\rho^+_{s'}}\|\theta^* - \theta^{(t)}\|_2 \\
&\overset{(iii)}{\leq} \frac{C'\sqrt{s'\rho^+_{s^*+2\tilde{s}}\rho^+_{s'}}\lambda_{\text{tgt}}\sqrt{s^*}}{\rho^-_{s^*+2\tilde{s}}},
\end{aligned} \tag{7.163}$$

where (i) is from the SE properties, (ii) is from the definition of v, and (iii) is from $\|\theta^{(t)} - \theta^*\|_2 \leq C'\lambda_{\text{tgt}}\sqrt{s^*}/\rho^-_{s^*+2\tilde{s}}$. Then (7.163) implies

$$s' \leq \frac{C_2\rho^+_{s^*+2\tilde{s}}\rho^+_{s'}s^*}{(\rho^-_{s^*+2\tilde{s}})^2} \leq C_2\kappa^2_{s^*+2\tilde{s}}s^*, \tag{7.164}$$

where the last inequality is from the fact that $s' = |\mathcal{S}'|$ achieves the maximum possible value such that $s' \leq \tilde{s}$ for any subset $\mathcal{S}'$ of $\mathcal{A}_2$. (7.164) implies that $s' < \tilde{s}$, so wo must have $\mathcal{S}' = \mathcal{A}_2$ to attain the maximum. Then we have

$$|\mathcal{A}_2| = s' \leq C_2\kappa^2_{s^*+2\tilde{s}}s^*.$$

Set $\mathcal{A}_3$. We have $\mathcal{A}_3 = \left\{ j \in \overline{\mathcal{S}}' : \left|\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right)_i\right| \geq \lambda_{\text{tgt}}/4 \right\}$. Suppose we choose a vector $v \in \mathbb{R}^d$ such that $\|v\|_\infty = 1$, $\|v\|_0 = |\mathcal{A}_3|$ and

$$\begin{aligned} v^\top \left(\nabla \mathcal{L}(\theta^{(t)}) - \nabla \mathcal{L}(\theta^*)\right) &= \sum_{i \in \mathcal{A}_3} v_i \left(\nabla \mathcal{L}(\theta^{(t)}) - \nabla \mathcal{L}(\theta^*)\right)_i \\ &= \sum_{i \in \mathcal{A}_3} \left|\left(\nabla \mathcal{L}(\theta^{(t)}) - \nabla \mathcal{L}(\theta^*)\right)_i\right| \geq \lambda_{\mathrm{tgt}} |\mathcal{A}_3|/4. \end{aligned} \tag{7.165}$$

Then we have

$$\begin{aligned} v^\top \left(\nabla \mathcal{L}(\theta^{(t)}) - \nabla \mathcal{L}(\theta^*)\right) \leq \|v\|_2 \|\nabla \mathcal{L}(\theta^{(t)}) - \nabla \mathcal{L}(\theta^*)\|_2 &\overset{(i)}{\leq} \sqrt{|\mathcal{A}_3|} \|\nabla \mathcal{L}(\theta^{(t)}) - \nabla \mathcal{L}(\theta^*)\|_2 \\ &\overset{(ii)}{\leq} \rho^+_{s^*+2\tilde{s}} \sqrt{|\mathcal{A}_3|} \cdot \|\theta^{(t)} - \theta^*\|_2, \end{aligned} \tag{7.166}$$

where (i) is from the definition of v, and (ii) is from the mean value theorem and analogous argument for $\mathcal{A}_2$.

Combining (7.165) and (7.166), we have

$$\lambda_{\mathrm{tgt}} |\mathcal{A}_3| \leq 4\rho^+_{s^*+2\tilde{s}} \sqrt{|\mathcal{A}_3|} \cdot \|\theta - \theta^*\|_2 \overset{(i)}{\leq} 8\lambda_{\mathrm{tgt}} \kappa_{s^*+2\tilde{s}} \sqrt{3s^* |\mathcal{A}_3|}$$

where (i) is from $\|\theta^{(t)} - \theta^*\|_2 \leq C' \lambda_{\mathrm{tgt}} \sqrt{s^*}/\rho^-_{s^*+2\tilde{s}}$ and definition of $\kappa_{s^*+2\tilde{s}} = \rho^+_{s^*+2\tilde{s}}/\rho^-_{s^*+2\tilde{s}}$. This implies

$$|\mathcal{A}_3| \leq C_3 \kappa^2_{s^*+2\tilde{s}} s^*.$$

Set $\mathcal{A}_4$. By Assumption 3 and $\lambda_{\mathrm{tgt}} \geq 4\|\nabla \mathcal{L}(\theta^*)\|_\infty$, we have

$$\begin{aligned} 0 \leq |V_4| &\leq \sum_{i \in \overline{\mathcal{S}}^*} \frac{4}{\lambda_{\mathrm{tgt}}} |(\nabla \mathcal{L}(\theta^*))_i| \cdot \mathbb{1}(|(\nabla \mathcal{L}(\theta^*))_i| > \lambda_{\mathrm{tgt}}/(4)) \\ &= \sum_{i \in \overline{\mathcal{S}}^*} \frac{4}{\lambda_{\mathrm{tgt}}} |(\nabla \mathcal{L}(\theta^*))_i| \cdot 0 = 0, \end{aligned} \tag{7.167}$$

Set A_1. From Lemma 50, we have $\mathcal{F}_\lambda(\theta^{(t+1)}) \leq \mathcal{F}_\lambda(\theta^*) + \frac{\lambda_{\text{tgt}}}{4}\|\theta^{(t+1)} - \theta^*\|_1$. This implies

$$\begin{aligned}\mathcal{L}(\theta^{(t+1)}) - \mathcal{L}(\theta^*) &\leq \lambda_{\text{tgt}}(\|\theta^*\|_1 - \|\theta^{(t+1)}\|_1) + \frac{\lambda_{\text{tgt}}}{4}\|\theta^{(t+1)} - \theta^*\|_1 \\ &= \lambda_{\text{tgt}}(\|\theta^*_{\mathcal{S}'}\|_1 - \|\theta^{(t+1)}_{\mathcal{S}'}\|_1 - \|\theta^{(t+1)}_{\mathcal{S}'_\perp}\|_1) + \frac{\lambda_{\text{tgt}}}{4}\|\theta^{(t+1)} - \theta^*\|_1 \\ &\leq \frac{5\lambda_{\text{tgt}}}{4}\|\theta^{(t+1)}_{\mathcal{S}'} - \theta^*_{\mathcal{S}'}\|_1 - \frac{3\lambda_{\text{tgt}}}{4}\|\theta^{(t+1)}_{\mathcal{S}'_\perp} - \theta^*_{\mathcal{S}'_\perp}\|_1. \end{aligned} \tag{7.168}$$

where the equality holds since $\theta^*_{\mathcal{S}'_\perp} = 0$. On the other hand, we have

$$\begin{aligned}&\mathcal{L}(\theta^{(t+1)}) - \mathcal{L}(\theta^*) \overset{(i)}{\geq} \nabla\mathcal{L}(\theta^*)(\theta^{(t+1)} - \theta^*) \geq -\|cL(\theta^*)\|_\infty\|\theta^{(t+1)} - \theta^*\|_1 \\ &\overset{(ii)}{\geq} -\frac{\lambda_{\text{tgt}}}{4}\|\theta^{(t+1)} - \theta^*\|_1 = -\frac{\lambda_{\text{tgt}}}{4}\|\theta^{(t+1)}_{\mathcal{S}'} - \theta^*_{\mathcal{S}'}\|_1 - \frac{\lambda_{\text{tgt}}}{4}\|\theta^{(t+1)}_{\mathcal{S}'_\perp} - \theta^*_{\mathcal{S}'_\perp}\|_1, \end{aligned} \tag{7.169}$$

where (i) is from the convexity of $\mathcal{L}$ and (ii) is from Assumption 3. Combining (7.168) and (7.169), we have

$$\|\theta^{(t+1)}_{\mathcal{S}'_\perp} - \theta^*_{\mathcal{S}'_\perp}\|_1 \leq 3\|\theta^{(t+1)}_{\mathcal{S}'} - \theta^*_{\mathcal{S}'}\|_1,$$

which implies that $\theta^{(t+1)} - \theta^* \in \mathcal{C}(s^*, 3, R)$ with respect to the set $\mathcal{S}'$.

We have $\mathcal{A}_4 = \left\{j \in \overline{\mathcal{S}}' : |(\nabla^2\mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t+1)}))_j| \geq \lambda_{\text{tgt}}/4\right\}$. Consider a subset $\mathcal{S}' \subset \mathcal{A}_2$ with $|\mathcal{S}'| = s' \leq \tilde{s}$ and a vector $v \in \mathbb{R}^d$ similar to that in $\mathcal{A}_2$. Then we have

$$\begin{aligned}s'\lambda_{\text{tgt}}/4 &\leq v^\top\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*) \leq \|v(\nabla^2\mathcal{L}(\theta^{(t)}))^{\frac{1}{2}}\|_2\|(\nabla^2\mathcal{L}(\theta^{(t)}))^{\frac{1}{2}}(\theta^{(t+1)} - \theta^*)\|_2 \\ &\overset{(i)}{\leq} c_1\sqrt{\rho^+_{s^*+2\tilde{s}}\rho^+_{s'}}\|v\|_2\|\theta^* - \theta^{(t+1)}\|_2 \overset{(ii)}{\leq} c_1\sqrt{s'\rho^+_{s^*+2\tilde{s}}\rho^+_{s'}}\|\theta^* - \theta^{(t+1)}\|_2 \\ &\overset{(iii)}{\leq} \frac{c_2\sqrt{s'\rho^+_{s^*+2\tilde{s}}\rho^+_{s'}}\lambda_{\text{tgt}}\sqrt{s^*}}{\rho^-_{s^*+2\tilde{s}}},, \end{aligned} \tag{7.170}$$

where (i) is from SE condition and Proposition 6, (ii) is from the definition of v, and (iii) is from $\|\theta^{(t+1)} - \theta^*\|_2 \leq C'\lambda_{\text{tgt}}\sqrt{s^*}/\rho^-_{s^*+2\tilde{s}}$. Following analogous argument in for A_2, we have

$$|A_1| \leq C_1\kappa^2_{s^*+2\tilde{s}}s^*.$$

Combining the results for Set $A_1 \sim A_4$, we have that there exists some constant C_0 such that

$$\|\theta_{\overline{\mathcal{S}}}^{(t+\frac{1}{2})}\|_0 \leq C_0 \kappa_{s^*+2\tilde{s}}^2 s^* \leq \tilde{s}.$$

From Lemma 46, we further have that the step size satisfies $\eta_t = 1$, then we have $\theta^{(t+1)} = \theta^{(t+\frac{1}{2})}$. The estimation error follows directly from Lemma 51.

Proof of Lemma 45

For notational simplicity, we introduce the following proximal operator,

$$\operatorname{prox}_r^{H,g}(\theta) = \operatorname{argmin}_{\theta'} r(\theta') + g^\top(\theta' - \theta) + \frac{1}{2}\|\theta' - \theta\|_H^2.$$

Then we have

$$\theta^{(t+1)} = \operatorname{prox}_{\mathcal{R}_{\lambda\{K\}}^{\ell_1}(\theta^{(t)})}^{\nabla^2\mathcal{L}\left(\theta^{(t)}\right),\nabla\mathcal{L}\left(\theta^{(t)}\right)}\left(\theta^{(t)}\right).$$

By Lemma 44, we have

$$\|\theta_{\overline{\mathcal{S}}}^{(t+1)}\|_0 \leq \tilde{s}.$$

By the KKT condition of function $\min \mathcal{F}_{\lambda\{K\}}$, i.e., $-\nabla\mathcal{L}(\overline{\theta}^{\{K\}}) \in \partial\mathcal{R}_{\lambda\{K\}}^{\ell_1}(\overline{\theta}^{\{K\}})$, we also have

$$\overline{\theta}^{\{K\}} = \operatorname{prox}_{\mathcal{R}_{\lambda\{K\}}^{\ell_1}(\overline{\theta}^{\{K\}})}^{\nabla^2\mathcal{L}\left(\theta^{(t)}\right),\nabla\mathcal{L}(\overline{\theta}^{\{K\}})}\left(\overline{\theta}^{\{K\}}\right).$$

By monotonicity of sub-gradient of a convex function, we have the *strictly non-expansive* property: for any $\theta, \theta' \in \mathbb{R}$, let $u = \operatorname{prox}_r^{H,g}(\theta)$ and $v = \operatorname{prox}_r^{H,g'}(\theta')$, then

$$(u-v)^\top H(\theta - \theta') - (u-v)^\top \left(g - g'\right) \geq \|u - v\|_H^2 .$$

Thus by the strictly non-expansive property of the proximal operator, we obtain

$$\begin{aligned}
&\|\theta^{(t+1)} - \overline{\theta}^{\{K\}}\|^2_{\nabla^2\mathcal{L}(\overline{\theta}^{\{K\}})} \\
&\leq \left(\theta^{(t+1)} - \overline{\theta}^{\{K\}}\right)^\top \left[\nabla^2\mathcal{L}(\theta^{(t)}) \left(\theta^{(t)} - \overline{\theta}^{\{K\}}\right) + \left(\nabla\mathcal{L}(\overline{\theta}^{\{K\}}) - \nabla\mathcal{L}(\theta^{(t)})\right)\right] \\
&\leq \|\theta^{(t+1)} - \overline{\theta}^{\{K\}}\|_2 \left\|\nabla^2\mathcal{L}(\theta^{(t)}) \left(\theta^{(t)} - \overline{\theta}^{\{K\}}\right) + \left(\nabla\mathcal{L}(\overline{\theta}^{\{K\}}) - \nabla\mathcal{L}(\theta^{(t)})\right)\right\|_2. \qquad (7.171)
\end{aligned}$$

Note that both $\|\theta^{(t+1)}\|_0 \leq \tilde{s}$ and $\|\overline{\theta}^{\{K\}}\|_0 \leq \tilde{s}$. On the other hand, from the SE properties, we have

$$\begin{aligned}
\|\theta^{(t+1)} - \overline{\theta}^{\{K\}}\|^2_{\nabla^2\mathcal{L}(\overline{\theta}^{\{K\}})} &= (\theta^{(t+1)} - \overline{\theta}^{\{K\}})^\top \nabla^2\mathcal{L}(\overline{\theta}^{\{K\}})(\theta^{(t+1)} - \overline{\theta}^{\{K\}}) \\
&\geq \rho^-_{s^*+2\tilde{s}} \|\theta^{(t+1)} - \overline{\theta}^{\{K\}}\|^2_2. \qquad (7.172)
\end{aligned}$$

Combining (7.171) and (7.172), we have

$$\begin{aligned}
&\left\|\theta^{(t+1)} - \overline{\theta}^{\{K\}}\right\|_2 \\
&\leq \frac{1}{\rho^-_{s^*+2\tilde{s}}} \left\|\nabla^2\mathcal{L}(\theta^{(t)}) \left(\theta^{(t)} - \overline{\theta}^{\{K\}}\right) + \left(\nabla\mathcal{L}(\overline{\theta}^{\{K\}}) - \nabla\mathcal{L}(\theta^{(t)})\right)\right\|_2 \\
&= \frac{1}{\rho^-_{s^*+2\tilde{s}}} \left\|\int_0^1 \left[\nabla^2\mathcal{L}\left(\theta^{(t)} + \tau\left(\overline{\theta}^{\{K\}} - \theta^{(t)}\right)\right) - \nabla^2\mathcal{L}\left(\theta^{(t)}\right)\right] \cdot \left(\overline{\theta}^{\{K\}} - \theta^{(t)}\right) d\tau\right\|_2 \\
&\leq \frac{1}{\rho^-_{s^*+2\tilde{s}}} \int_0^1 \left\|\left[\nabla^2\mathcal{L}\left(\theta^{(t)} + \tau\left(\overline{\theta}^{\{K\}} - \theta^{(t)}\right)\right) - \nabla^2\mathcal{L}\left(\theta^{(t)}\right)\right] \cdot \left(\overline{\theta}^{\{K\}} - \theta^{(t)}\right)\right\|_2 d\tau \\
&\leq \frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}} \left\|\theta^{(t)} - \overline{\theta}^{\{K\}}\right\|^2_2,
\end{aligned}$$

where the last inequality is from the local restricted Hessian smoothness of $\mathcal{L}$. Then we finish the proof by the definition of R.

Proof of Lemma 46

Suppose the step size $\eta_t < 1$. Note that we do not need the step size to be $\eta_t = 1$ in Lemma 44 and Lemma 45. We denote $\Delta\theta^{(t)} = \theta^{(t+\frac{1}{2})} - \theta^{(t)}$. Then we have

$$\begin{aligned}
\left\|\Delta\theta^{(t)}\right\|_2 &\overset{(i)}{\leq} \left\|\theta^{(t)} - \overline{\theta}^{\{K\}}\right\|_2 + \left\|\theta^{(t+\frac{1}{2})} - \overline{\theta}^{\{K\}}\right\|_2 \\
&\overset{(ii)}{\leq} \left\|\theta^{(t)} - \overline{\theta}^{\{K\}}\right\|_2 + \frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}} \left\|\theta^{(t)} - \overline{\theta}^{\{K\}}\right\|_2^2 \\
&\overset{(iii)}{\leq} \frac{3}{2} \left\|\theta^{(t)} - \overline{\theta}^{\{K\}}\right\|_2, \qquad (7.173)
\end{aligned}$$

where (i) is from triangle inequality, (ii) is from Lemma 45, and (iii) is from $\left\|\theta^{(t)} - \overline{\theta}^{\{K\}}\right\|_2 \leq R \leq \frac{\rho^-_{s^*+2\tilde{s}}}{L_{s^*+2\tilde{s}}}$.

By Lemma 44, we have

$$\left\|\Delta\theta^{(t)}_{\overline{\mathcal{S}}}\right\|_0 \leq 2\tilde{s}.$$

To show $\eta_t = 1$, it is now suffice to demonstrate that

$$\mathcal{F}_{\lambda^{\{K\}}}(\theta^{(t+\frac{1}{2})}) - \mathcal{F}_{\lambda^{\{K\}}}(\theta^{(t)}) \leq \frac{1}{4}\gamma_t.$$

By expanding $\mathcal{F}_{\lambda^{\{K\}}}$, we have

$$\begin{aligned}
&\mathcal{F}_{\lambda^{\{K\}}}(\theta^{(t)} + \Delta\theta^{(t)}) - \mathcal{F}_{\lambda^{\{K\}}}(\theta^{(t)}) \\
&= \mathcal{L}(\theta^{(t)} + \Delta\theta^{(t)}) - \mathcal{L}(\theta^{(t)}) + \mathcal{R}^{\ell_1}_{\lambda^{\{K\}}}(\theta^{(t)} + \Delta\theta^{(t)}) - \mathcal{R}^{\ell_1}_{\lambda^{\{K\}}}(\theta^{(t)}) \\
&\overset{(i)}{\leq} \nabla\mathcal{L}(\theta^{(t)})^\top\Delta\theta^{(t)} + \frac{1}{2}\Delta(\theta^{(t)})^\top\nabla^2\mathcal{L}(\theta)\Delta\theta^{(t)} + \frac{L_{s^*+2\tilde{s}}}{6}\left\|\Delta\theta^{(t)}\right\|_2^3 + \mathcal{R}^{\ell_1}_{\lambda^{\{K\}}}(\theta^{(t)} + \Delta\theta^{(t)}) \\
&\qquad - \mathcal{R}^{\ell_1}_{\lambda^{\{K\}}}(\theta^{(t)}) \\
&\overset{(ii)}{\leq} \gamma_t - \frac{1}{2}\gamma_t + \frac{L_{s^*+2\tilde{s}}}{6}\left\|\Delta\theta^{(t)}\right\|_2^3 \overset{(iii)}{\leq} \frac{1}{2}\gamma_t + \frac{L_{s^*+2\tilde{s}}}{6\rho^-_{s^*+2\tilde{s}}}\left\|\Delta\theta^{(t)}\right\|_{\nabla^2\mathcal{L}(\theta)}\left\|\Delta\theta^{(t)}\right\|_2 \\
&\overset{(iv)}{\leq} \left(\frac{1}{2} - \frac{L_{s^*+2\tilde{s}}}{6\rho^-_{s^*+2\tilde{s}}}\left\|\Delta\theta^{(t)}\right\|_2\right)\gamma_t \overset{(v)}{\leq} \frac{1}{4}\gamma_t,
\end{aligned}$$

where (i) is from the restricted Hessian smooth condition, (ii) and (iv) are from Lemma 47, (iii) is from the same argument of (7.172), and (v) is from (7.173), $\gamma_t < 0$, and $\left\|\theta^{(t)} - \overline{\theta}^{\{K\}}\right\|_2 \leq R \leq \frac{\rho^{-}_{s^*+2\tilde{s}}}{L_{s^*+2\tilde{s}}}$. This implies $\theta^{(t+1)} = \theta^{(t+\frac{1}{2})}$.

Proof of Lemma 47

We denote $H = \nabla^2 \mathcal{L}(\theta^{(t)})$. Since $\Delta\theta^{(t)}$ is the solution for

$$\min_{\Delta\theta^{(t)}} \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{1}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \mathcal{R}^{\ell_1}_{\lambda\{K\}}\left(\theta^{(t)} + \Delta\theta^{(t)}\right)$$

then for any $\eta_t \in (0,1]$, we have

$$\begin{aligned}
&\eta_t \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{\eta_t^2}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \mathcal{R}^{\ell_1}_{\lambda\{K\}}\left(\theta^{(t)} + \eta_t\Delta\theta^{(t)}\right) \\
&\geq \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{1}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \mathcal{R}^{\ell_1}_{\lambda\{K\}}\left(\theta^{(t)} + \Delta\theta^{(t)}\right)
\end{aligned}$$

By the convexity of $\mathcal{R}^{\ell_1}_{\lambda\{K\}}$, we have

$$\begin{aligned}
&\eta_t \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{\eta_t^2}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \eta_t\mathcal{R}^{\ell_1}_{\lambda\{K\}}\left(\theta^{(t)} + \Delta\theta^{(t)}\right) + (1-\eta_t)\mathcal{R}^{\ell_1}_{\lambda\{K\}}(\theta^{(t)}) \\
&\geq \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{1}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \mathcal{R}^{\ell_1}_{\lambda\{K\}}\left(\theta^{(t)} + \Delta\theta^{(t)}\right).
\end{aligned}$$

Rearranging the terms, we obtain

$$\begin{aligned}
&(1-\eta_t)\left(\nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \mathcal{R}^{\ell_1}_{\lambda\{K\}}\left(\theta^{(t)} - \Delta\theta^{(t)}\right) - \mathcal{R}^{\ell_1}_{\lambda\{K\}}(\theta^{(t)})\right) + \frac{1-\eta_t^2}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 \\
&\leq 0
\end{aligned}$$

Canceling the $(1-\eta_t)$ factor from both sides and let $\eta_t \to 1$, we obtain the desired inequality,

$$\gamma_t \leq -\left\|\Delta\theta^{(t)}\right\|_H^2.$$

Proof of Lemma 48

We first demonstrate an upper bound of the approximate KKT parameter $\omega_{\lambda\{K\}}$. Given the solution $\theta^{(t-1)}$ from the $(t-1)$-th iteration, the optimal solution at t-th iteration satisfies the KKT condition:

$$\nabla^2\mathcal{L}(\theta^{(t-1)})(\theta^{(t)}-\theta^{(t-1)})+\nabla\mathcal{L}(\theta^{(t-1)})+\lambda^{\{K\}}\odot\xi^{(t)}=0,$$

where $\xi^{(t)}\in\partial\|\theta^{(t)}\|_1$. Then for any vector v with $\|v\|_2\le\|v\|_1=1$ and $\|v\|_0\le s^*+2\tilde{s}$, we have

$$\begin{aligned}
&(\nabla\mathcal{L}(\theta^{(t)})+\lambda^{\{K\}}\odot\xi^{(t)})^\top v\\
&=(\nabla\mathcal{L}(\theta^{(t)}))^\top v-(\nabla^2\mathcal{L}(\theta^{(t-1)})(\theta^{(t)}-\theta^{(t-1)})+\nabla\mathcal{L}(\theta^{(t-1)}))^\top v\\
&=(\nabla\mathcal{L}(\theta^{(t)})-\nabla\mathcal{L}(\theta^{(t-1)}))^\top v-(\nabla^2\mathcal{L}(\theta^{(t-1)})(\theta^{(t)}-\theta^{(t-1)}))^\top v\\
&\overset{(i)}{\le}\left\|(\nabla^2\mathcal{L}(\tilde{\theta}))^{\frac{1}{2}}(\theta^{(t)}-\theta^{(t-1)})\right\|_2\cdot\left\|v^\top(\nabla^2\mathcal{L}(\tilde{\theta}))^{\frac{1}{2}}\right\|_2\\
&\qquad\qquad+\left\|(\nabla^2\mathcal{L}(\theta^{(t-1)}))^{\frac{1}{2}}(\theta^{(t)}-\theta^{(t-1)})\right\|_2\cdot\left\|v^\top(\nabla^2\mathcal{L}(\theta^{(t-1)}))^{\frac{1}{2}}\right\|_2\\
&\overset{(ii)}{\le}2\rho^+_{s^*+2\tilde{s}}\left\|\theta^{(t)}-\theta^{(t-1)}\right\|_2,
\end{aligned}\tag{7.174}$$

where (i) is from mean value theorem with some $\tilde{\theta}=(1-a)\theta^{(t-1)}+a\theta^{(t)}$ for some $a\in[0,1]$ and Cauchy-Schwarz inequality, and (ii) is from the SE properties. Take the supremum of the L.H.S. of (7.174) with respect to v, we have

$$\left\|\nabla\mathcal{L}(\theta^{(t)})+\lambda^{\{K\}}\odot\xi^{(t)}\right\|_\infty\le2\rho^+_{s^*+2\tilde{s}}\left\|\theta^{(t)}-\theta^{(t-1)}\right\|_2.\tag{7.175}$$

Then from Lemma 45, we have

$$\left\|\theta^{(t+1)}-\overline{\theta}^{\{K\}}\right\|_2\le\left(\frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}}\right)^{1+2+4+\ldots+2^{t-1}}\left\|\theta^{(0)}-\overline{\theta}^{\{K\}}\right\|_2^{2^\top}\le\left(\frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}}\left\|\theta^{(0)}-\overline{\theta}^{\{K\}}\right\|_2\right)^{2^t}.$$

By (7.175) and (7.173) by taking $\Delta\theta^{(t-1)}=\theta^{(t)}-\theta^{(t-1)}$, we obtain

$$\begin{aligned}
\omega_{\lambda^{\{K\}}}\left(\theta^{(t)}\right)&\le2\rho^+_{s^*+2\tilde{s}}\left\|\theta^{(t)}-\theta^{(t-1)}\right\|_2\le3\rho^+_{s^*+2\tilde{s}}\left\|\theta^{(t-1)}-\overline{\theta}^{\{K\}}\right\|_2\\
&\le3\rho^+_{s^*+2\tilde{s}}\left(\frac{L_{s^*+2\tilde{s}}}{2\rho^-_{s^*+2\tilde{s}}}\left\|\theta^{(0)}-\overline{\theta}^{\{K\}}\right\|_2\right)^{2^t}.
\end{aligned}$$

By requiring the R.H.S. equal to ε we obtain

$$t = \log \frac{\log \left(\frac{3\rho^{+}_{s^*+2\tilde{s}}}{\varepsilon} \right)}{\log \left(\frac{2\rho^{-}_{s^*+2\tilde{s}}}{L_{s^*+2\tilde{s}} \left\| \theta^{(0)} - \overline{\theta}^{\{K\}} \right\|_2} \right)} = \log \log \left(\frac{3\rho^{+}_{s^*+2\tilde{s}}}{\varepsilon} \right) - \log \log \left(\frac{2\rho^{-}_{s^*+2\tilde{s}}}{L_{s^*+2\tilde{s}} \left\| \theta^{(0)} - \overline{\theta}^{\{K\}} \right\|_2} \right)$$
$$\overset{(i)}{\leq} \log \log \left(\frac{3\rho^{+}_{s^*+2\tilde{s}}}{\varepsilon} \right) - \log \log 4 \leq \log \log \left(\frac{3\rho^{+}_{s^*+2\tilde{s}}}{\varepsilon} \right),$$

where (i) is from the fact that $\left\| \theta^{(0)} - \overline{\theta}^{\{K\}} \right\|_2 \leq R = \frac{\rho^{-}_{s^*+2\tilde{s}}}{2L_{s^*+2\tilde{s}}}$.

7.4.4 Proof of Intermediate Results for Theorem 12

Proof of Lemma 41

Given the assumptions, we will show that for all large enough t, we have

$$\|\theta^{(t+1)}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}.$$

Following the analysis of Lemma 46, Lemma 47, and Appendix 7.4.7, we have that the objective $\mathcal{F}_{\lambda^{\{1\}}}$ has sufficient descendant in each iteration of proximal Newton step, which is also discussed in [175]. Then there exists a constant T such that for all $t \geq T$, we have

$$\mathcal{F}_{\lambda^{\{1\}}}(\theta^{(t)}) \leq \mathcal{F}_{\lambda^{\{1\}}}(\theta^*) + \frac{\lambda_{\text{tgt}}}{4} \|\theta^{(t)} - \theta^*\|_1,$$

where $\|\theta^{(t)} - \theta^*\|_1 \leq c\lambda_{\text{tgt}}\sqrt{s^*}/\rho^{-}_{s^*+\tilde{s}}$ from similar analysis in [153]. The rest of the analysis is analogous to that of Lemma 44, from which we have $\|\theta^{(t)}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}$.

Proof of Lemma 42

The estimation error is derived analogously from [153], thus we omit it here. The claim of the quadratic convergence follows directly from Lemma 45 given sparse solutions.

Proof of Lemma 43

The upper bound of the number of iterations for proximal Newton update is obtained by combining Lemma 41 and Lemma 48. Note that

$$T_1 \leq \log \frac{\log\left(\frac{3\rho^+_{s^*+2\tilde{s}}}{\varepsilon}\right)}{\log\left(\frac{2\rho^-_{s^*+2\tilde{s}}}{L_{s^*+2\tilde{s}}\left\|\theta^{(T+1)}-\overline{\theta}^{\{1\}}\right\|_2}\right)}.$$

Then we obtain the result from $\left\|\theta^{(T+1)} - \overline{\theta}^{\{1\}}\right\|_2 \leq R = \frac{\rho^-_{s^*+2\tilde{s}}}{2L_{s^*+2\tilde{s}}}$.

7.4.5 Proof of Theorem 14

It is demonstrated in [242] that Assumptions 1 $\sim$ 3 hold given the LRE properties defined in Definition 23. Thus, combining the analyses in [242] and Proposition 6, we have that Assumptions 1 $\sim$ 3 hold with high probability. Assumption 4 also holds trivially by choosing $\varepsilon = \frac{c}{\sqrt{n}}$ for some generic constant c. The rest of the results follow directly from Theorem 13 and the analyses in [150].

7.4.6 Further Intermediate Results

Lemma 50. Given $\omega_{\lambda^{\{K\}}}(\hat{\theta}^{\{K\}}) \leq \frac{\lambda_{\text{tgt}}}{8}$, we have that for all $t \geq 1$ at the $\{K+1\}$-th stage,

$$\omega_{\lambda^{\{K+1\}}}(\theta^{(t)}) \leq \frac{\lambda_{\text{tgt}}}{4} \quad \text{and} \quad \mathcal{F}_{\lambda^{\{K+1\}}}(\theta^{(t)}) \leq \mathcal{F}_{\lambda^{\{K+1\}}}(\theta^*) + \frac{\lambda_{\text{tgt}}}{4}\|\theta^{(t)} - \theta^*\|_1.$$

Proof. Note that at the $\{K+1\}$-th stage, $\theta^{(0)} = \hat{\theta}^{\{K\}}$. Then we have

$$\begin{aligned}
\omega_{\lambda^{\{K+1\}}}(\theta^{(0)}) &= \min_{\xi \in \partial\|\theta^{(0)}\|_1} \|\nabla\mathcal{L}(\theta^{(0)}) + \lambda^{\{K+1\}} \odot \xi\|_\infty \\
&\overset{(i)}{\leq} \min_{\xi \in \|\theta^{(0)}\|_1} \|\nabla\mathcal{L}(\theta^{(0)}) + \lambda^{\{K\}} \odot \xi\|_\infty + \|(\lambda^{\{K+1\}} - \lambda^{\{K\}}) \odot \xi\|_\infty \\
&\overset{(ii)}{\leq} \omega_{\lambda^{\{K\}}}(\theta^{(0)}) + \|\lambda^{\{K+1\}} - \lambda^{\{K\}}\|_\infty \overset{(iii)}{\leq} \frac{\lambda_{\text{tgt}}}{8} + \frac{\lambda_{\text{tgt}}}{8} \leq \frac{\lambda_{\text{tgt}}}{4},
\end{aligned}$$

where (i) is from triangle inequality, (ii) is from the definition of the approximate KKT condition and ξ, and (iii) is from $\omega_{\lambda^{\{K\}}}(\theta^{(0)}) = \omega_{\lambda^{\{K\}}}(\hat{\theta}^{\{K\}}) \leq \frac{\lambda_{\rm tgt}}{8}$ and $\|\lambda^{\{K+1\}} - \lambda^{\{K\}}\|_\infty \leq \frac{\lambda_{\rm tgt}}{8}$.

For some $\xi^{(t)} = \operatorname{argmin}_{\xi \in \partial\|\theta^{(t)}\|_1} \|\nabla\mathcal{L}(\theta^{(t)}) + \lambda^{\{K+1\}} \odot \xi\|_\infty$, we have

$$\begin{aligned}\mathcal{F}_{\lambda^{\{K+1\}}}(\theta^*) &\overset{(i)}{\geq} \mathcal{F}_{\lambda^{\{K+1\}}}(\theta^{(t)}) - (\nabla\mathcal{L}(\theta^{(t)}) + \lambda^{\{K+1\}} \odot \xi^{(t)})^\top(\theta^{(t)} - \theta^*)\\ &\geq \mathcal{F}_{\lambda^{\{K+1\}}}(\theta^{(t)}) - \|\nabla\mathcal{L}(\theta^{(t)}) + \lambda^{\{K+1\}} \odot \xi^{(t)}\|_\infty \|\theta^{(t)} - \theta^*\|_1\\ &\overset{(ii)}{\geq} \mathcal{F}_{\lambda^{\{K+1\}}}(\theta^{(t)}) - \frac{\lambda_{\rm tgt}}{4}\|\theta^{(t)} - \theta^*\|_1\end{aligned}$$

where (i) is from the convexity of $\mathcal{F}_{\lambda^{\{K+1\}}}$ and (ii) is from the fact that for all $t \geq 0$, $\|\nabla\mathcal{L}(\theta^{(t)}) + \lambda^{\{K+1\}} \odot \xi^{(t)}\|_\infty \leq \frac{\lambda_{\rm tgt}}{4}$. This finishes the proof. □

Lemma 51 (Adapted from [153])**.** Suppose $\|\theta^{(t)}_{\overline{\mathcal{S}}}\|_0 \leq \tilde{s}$ and $\omega_{\lambda^{\{K\}}}(\theta^{(t)}) \leq \frac{\lambda_{\rm tgt}}{4}$. Then there exists a generic constant c_1 such that

$$\|\theta^{(t)} - \theta^*\|_2 \leq \frac{c_1 \lambda_{\rm tgt}\sqrt{s^*}}{\rho^-_{s^*+2\tilde{s}}}.$$

7.4.7 Global Convergence Analysis

For notational convenience, we denote $\mathcal{F} = \mathcal{F}_\lambda$ and $\mathcal{R} = \mathcal{R}^{\ell_1}_\lambda$ in the sequel. We first provide an upper bound of the objective gap.

Lemma 52. Suppose the $\mathcal{F}(\theta) = \mathcal{R}(\theta) + \mathcal{L}(\theta)$ and $\mathcal{L}(\theta)$ satisfies the restricted Hessian smoothness property, namely, for any $\theta, h \in \mathbb{R}^d$

$$\frac{d}{d\tau}\nabla^2\mathcal{L}(\theta + \tau h)|_{\tau=0} \preceq C\sqrt{h^\top \nabla^2\mathcal{L}(\theta) h} \cdot \nabla^2\mathcal{L}(\theta),$$

for some constant C. Let $\Delta\theta$ be the search direction and let $\theta_+ = \theta + \tau\Delta\theta$ for some $\tau \in (0, 1]$. Then

$$\mathcal{F}(\theta_+) \leq \mathcal{F}(\theta) + \left[-\tau + \mathcal{O}(\tau^2)\right] \|\Delta\theta\|^2_H .$$

Proof. From the convexity of $\mathcal{R}$, we have

$$\begin{aligned}
&\mathcal{F}(\theta_+) - \mathcal{F}(\theta) \\
&= \mathcal{L}(\theta_+) - \mathcal{L}(\theta) + \mathcal{R}(\theta_+) - \mathcal{R}(\theta) \\
&\leq \mathcal{L}(\theta_+) - \mathcal{L}(\theta) + \tau\mathcal{R}(\theta + \Delta\theta) + (1-\tau)\mathcal{R}(\theta) - \mathcal{R}(\theta) \\
&= \mathcal{L}(\theta_+) - \mathcal{L}(\theta) + \tau\left(\mathcal{R}(\theta + \Delta\theta) - \mathcal{R}(\theta)\right) \\
&= \nabla\mathcal{L}(\theta)^\top \cdot (\tau\Delta\theta) + \tau\left(\mathcal{R}(\theta + \Delta\theta) - \mathcal{R}(\theta)\right) + \tau\int_0^\tau (\Delta\theta)^\top \nabla^2\mathcal{L}(\theta + \alpha\Delta\theta)\Delta\theta d\alpha.
\end{aligned}$$

By Lemma 47 and the restricted Hessian smoothness property, we obtain

$$\begin{aligned}
&\mathcal{F}(\theta_+) - \mathcal{F}(\theta) \\
&\leq -\tau\left\|\Delta\theta\right\|_{\nabla^2\mathcal{L}(\theta)} + \tau\int_0^\tau (\Delta\theta)^\top\nabla^2\mathcal{L}(\theta + \alpha\Delta\theta)\Delta\theta d\alpha \\
&= -\tau\left\|\Delta\theta\right\|_{\nabla^2\mathcal{L}(\theta)} + \tau\int_0^\tau d\alpha\int_0^\alpha dz\frac{d}{dz}(\Delta\theta)^\top\nabla^2\mathcal{L}(\theta + z\Delta\theta)\Delta\theta \\
&\quad + \tau\int_0^\tau d\alpha(\Delta\theta)^\top\nabla^2\mathcal{L}(\theta)\Delta\theta \\
&= \left(-\tau + \mathcal{O}(\tau^2)\right)\left\|\Delta\theta\right\|^2_{\nabla^2\mathcal{L}(\theta)}.
\end{aligned}$$

□

Next, we show that $\Delta\theta \neq 0$ when θ have not attained the optimum.

Lemma 53. Suppose the $\mathcal{F}(\theta) = \mathcal{R}(\theta) + \mathcal{L}(\theta)$ has a unique minimizer, and $\mathcal{L}(\theta)$ satisfies the restricted Hessian smoothness property. Then $\Delta\theta^{(t)} = 0$ if and only if $\theta^{(t)} = \overline{\theta}$.

Proof. Suppose $\Delta\theta$ is non-zero at $\overline{\theta}$. Lemma 52 implies that for sufficiently small $0 < \tau \leq 1$,

$$\mathcal{F}(\overline{\theta} + \tau\Delta\theta^{(t)}) - \mathcal{F}(\overline{\theta}) \leq 0.$$

However $\mathcal{F}(\theta)$ is uniquely minimized at $\overline{\theta}$, which is a contradiction. Thus $\Delta\theta = 0$ at $\overline{\theta}$.

Now we consider the other direction. Suppose $\Delta\theta = 0$, then θ is a minimizer of $\mathcal{F}$. Thus for any direction h and $\tau \in (0, 1]$, we obtain

$$\nabla\mathcal{L}(\theta)^\top(\tau h) + \frac{1}{2}\tau^2 h^\top H h + \mathcal{R}(\theta + \tau h) - \mathcal{R}(\theta) \geq 0.$$

Rearrange, we obtain

$$\mathcal{R}(\theta + \tau h) - \mathcal{R}(\theta) \geq -\tau \nabla \mathcal{L}(\theta)^\top h - \frac{1}{2}\tau^2 h^\top H h$$

Let $D\mathcal{F}(\theta, h)$ be the directional derivative of $\mathcal{F}$ at θ in the direction h, thus

$$\begin{aligned} D\mathcal{F}(\theta, h) &= \lim_{\tau \to 0} \frac{\mathcal{F}(\theta + \tau h) - \mathcal{F}(\theta)}{\tau} \\ &= \lim_{\tau \to 0} \frac{\tau \nabla \mathcal{L}(\theta)^\top h + \mathcal{O}(\tau^2) + \mathcal{R}(\theta + \tau h) - \mathcal{R}(\theta)}{\tau} \\ &\geq \lim_{\tau \to 0} \frac{\tau \nabla \mathcal{L}(\theta)^\top h + \mathcal{O}(\tau^2) - \tau \nabla \mathcal{L}(\theta)^\top h - \frac{1}{2}\tau^2 h^\top H h}{\tau} = 0. \end{aligned}$$

Since $\mathcal{F}$ is convex, then θ is the minimizer of $\mathcal{F}$. □

Then, we show the behavior of $\|\Delta\theta\|_H$ and $\mathcal{R}(\theta + \Delta\theta)$ when $\Delta\theta \neq 0$.

Lemma 54. Suppose at any point $\theta \in \mathbb{R}^d$, we have $\nabla \mathcal{L}(\theta) \in \operatorname{span}\left(\nabla^2 \mathcal{L}(\theta)\right)$. If $\Delta\theta \neq 0$ then either

$$\|\Delta\theta\|_H > 0 \quad \text{or} \quad \mathcal{R}\left(\theta + \Delta\theta\right) < \mathcal{R}\left(\theta\right).$$

Proof. Recall that $\Delta\theta$ is obtained by solving the following sub-problem,

$$\Delta\theta = \operatorname*{argmin}_{\Delta\theta} \mathcal{R}(\theta + \Delta\theta) + \nabla \mathcal{L}(\theta)^\top \Delta\theta + \|\Delta\theta\|_H^2.$$

If $\|\Delta\theta\|_H = 0$ and $\Delta\theta \neq 0$, then

$$\Delta\theta \perp \operatorname{span}(H) \quad \text{and} \quad \nabla \mathcal{L}(\theta)^\top \Delta\theta = 0.$$

Thus

$$\mathcal{R}\left(\theta + \Delta\theta\right) < \mathcal{R}\left(\theta\right).$$

Notice that $\mathcal{R}\left(\theta + \Delta\theta\right) \neq \mathcal{R}\left(\theta\right)$, since otherwise $\Delta\theta = 0$ is a solution. □

Finally, we demonstrate the strict decrease of the objective in each proximal Newton step.

Lemma 55. Suppose at any point $\theta \in \mathbb{R}^d$, we have $\nabla \mathcal{L}(\theta) \in \operatorname{span}\left(\nabla^2 \mathcal{L}(\theta)\right)$. If $\Delta\theta \neq 0$ then

$$\mathcal{F}(\theta + \tau\Delta\theta) < \mathcal{F}(\theta),$$

for small enough $\tau > 0$.

Proof. By Lemma 54, if $\Delta\theta \neq 0$, then either $\|\Delta\theta\|_H > 0$ or $\mathcal{R}(\theta + \Delta\theta) - \mathcal{R}(\theta) < 0$. If it is the first case, then by Lemma 47,

$$\gamma = \nabla \mathcal{L}(\theta)^\top \Delta\theta + \mathcal{R}(\theta + \Delta\theta) - \mathcal{R}(\theta) < -\|\Delta\theta\|_H < 0.$$

It is the second case, then $\nabla \mathcal{L}(\theta)^\top \Delta\theta = 0$ and

$$\gamma = \mathcal{R}(\theta + \Delta\theta) - \mathcal{R}(\theta) < 0.$$

Moreover, we have

$$\begin{aligned}
&\mathcal{F}(\theta + \tau\Delta\theta) - \mathcal{F}(\theta) \\
&= \mathcal{L}(\theta + \tau\Delta\theta) - \mathcal{L}(\theta) + \mathcal{R}(\theta + \tau\Delta\theta) - \mathcal{R}(\theta) \\
&\leq \tau \nabla \mathcal{L}(\theta)^\top \Delta\theta + \frac{\tau^2}{2}\Delta\theta^\top H \Delta\theta + \mathcal{O}(\tau^3) + \mathcal{R}(\theta + \tau\Delta\theta) - \mathcal{R}(\theta) \\
&\leq \tau \nabla \mathcal{L}(\theta)^\top \Delta\theta + \tau\mathcal{R}(\theta + \Delta\theta) + (1-\tau)\mathcal{R}(\theta) - \mathcal{R}(\theta) + \frac{\tau^2}{2}\Delta\theta^\top H \Delta\theta + \mathcal{O}(\tau^3) \\
&= \tau(\gamma + \mathcal{O}(\tau)).
\end{aligned}$$

where the first inequality is from the restricted Hessian smoothness property. Thus $\mathcal{F}(\theta + \tau\Delta\theta) - \mathcal{F}(\theta) < 0$ for sufficiently small $\tau > 0$. □

Since each step, the objective is strictly decreasing, thus the algorithm will eventually reach the minimum.

7.5 Proofs for Chapter 6

7.5.1 Proofs of Results in Section 6.2

Proof of Theorem 15

From the directional derivative of f at $x_{\mathcal{G}}$, for any x, we have

$$0 = \lim_{t \to 0} \frac{f(x_{\mathcal{G}} + tg(x)) - f(x_{\mathcal{G}} + tx)}{t} = \nabla f(x_{\mathcal{G}})^\top (g(x) - x),$$

which implies $x_{\mathcal{G}}$ is a stationary point.

Proof of Theorem 16

Given any $v \in T_x G(x)$, there exists a smooth path $\gamma : (-1, 1) \to G(x)$ with $\gamma(0) = x$ and $v = \gamma'(0)$. We consider the function $\ell(t) = f(\gamma(t))$. By chain rule, we have

$$\ell'(t) = \nabla f(\gamma(t))^\top \gamma'(t) \text{ and } \ell''(t) = \gamma'(t)^\top \nabla^2 f(\gamma(t))\gamma'(t) + \nabla f(\gamma(t))^\top \gamma''(t). \tag{7.176}$$

Furthermore, since G is the invariant group, we have $\ell(t) = f(\gamma(t)) = \text{const}$ and $\ell'(t) = \ell''(t) = 0$ for any $t \in (-1, 1)$. Since x is stationary, $\nabla f(\gamma(0)) = \nabla f(x) = 0$ and we plug it into (7.176) to have

$$0 = \ell''(0) = \gamma'(0)^\top \nabla^2 f(\gamma(0))\gamma'(0) = v^\top H_x v,$$

which implies that $v \in \text{Null}(H_x)$. This completes our proof.

7.5.2 Proof of Theorem 17

We separate the analysis into four intermediate components, one for each claim. We first identifies the stationary point of $\mathcal{F}(x)$ in the following lemma. The proof is provided in Appendix 7.5.2.

Lemma 56. 0, u and $-u$ are the only stationary points of $\mathcal{F}(x)$, i.e., $\nabla \mathcal{F}(x) = 0$.

Next, we characterize two types of stationary points. We show a stronger result in the following lemma that $x = 0$ is a strict saddle point, and $\nabla^2 \mathcal{F}(x)$ has both positive and negative eigenvalue in the neighborhood of $x = 0$. The proof is provided in Appendix 7.5.2.

Lemma 57. $x = 0$ is a strict saddle point, where $\nabla^2 \mathcal{F}(0)$ is negative semi-definite with $\lambda_{\min}(\mathcal{F}(0)) = -\|u\|_2^2$. Moreover, for any $x \in \mathcal{R}_1$, $\nabla^2 \mathcal{F}(x)$ contains positive eigenvalues and negative eigenvalues, i.e.

$$\lambda_{\max}(\nabla^2 \mathcal{F}(x)) \geq \|x\|_2^2 \quad \text{and} \quad \lambda_{\min}(\nabla^2 \mathcal{F}(x)) \leq -\frac{1}{2}\|u\|_2^2.$$

Moreover, we identify that $x = \pm u$ are global minima, and $\mathcal{F}(x)$ is strongly convex in a neighborhood of $x = \pm u$. The proof is provided in Appendix 7.5.2.

Lemma 58. For $x = \pm u$, x is a global minimum, and $\nabla^2 \mathcal{F}(x)$ is positive definite with $\lambda_{\min}(\nabla^2 \mathcal{F}(x)) = \|u\|_2^2$. Moreover, for any $x \in \mathcal{R}_2$, $\mathcal{F}(x)$ is locally strongly convex, i.e.

$$\lambda_{\min}(\nabla^2 \mathcal{F}(x)) \geq \frac{1}{5}\|u\|_2^2.$$

Finally, we show that outside the regions $\mathcal{R}_1$ and $\mathcal{R}_2$, the gradient $\nabla \mathcal{F}(s)$ has a sufficiently large norm. The proof is provided in Appendix 7.5.2.

Lemma 59. For any $x \in \mathcal{R}_3$, we have

$$\|\nabla \mathcal{F}(x)\|_2 > \frac{\|u\|_2^3}{8}.$$

Combining Lemma 56 – Lemma 59, we finish the proof.

Proof of Lemma 56

We provide an algebraic approach to determine stationary points here. Without loss of generality, we assume $\|u\|_2 = 1$. Then we write

$$x = \alpha u + w,$$

where $\alpha \in \mathbb{R}$ is a constant and $w^\top u = 0$. Accordingly, we solve

$$\begin{aligned}(xx^\top - uu^\top)x &= [(\alpha u + w)(\alpha u + w)^\top - uu^\top](\alpha u + w)\\ &= [\alpha^2 uu^\top + \alpha wu^\top + \alpha uw^\top + ww^\top - uu^\top](\alpha u + w)\\ &= \alpha^3 u + \alpha^2 w + \alpha u\|w\|_2^2 + w\|w\|_2^2 - \alpha u\\ &= u(\alpha^3 + \alpha\|w\|_2^2 - \alpha) + w(\alpha^2 + \|w\|_2^2) = 0.\end{aligned}$$

1. Suppose $\alpha = 0$, which implies $u^\top x = 0$. Thus we must have $(xx^\top - uu^\top)x = x\|x\|_2^2 = 0$, which further implies $x = 0$ is a stationary point.

2. Suppose $\|w\|_2 = 0$, which implies $w = 0$. Thus we must have $(xx^\top - uu^\top)x = (\alpha^3 - \alpha)u = 0$, which further implies $\alpha = -1$ or 1, i.e., $x = -u$ and $x = u$ are stationary points.

3. Suppose $\alpha \neq 0$ and $w \neq 0$. We then require

$$\alpha^2 + \|w\|_2^2 - 1 = 0 \quad \text{and} \quad \alpha^2 + \|w\|_2^2 = 0.$$

   This conflict with each other, which implies there is no stationary point when $\alpha \neq 0$ and $w \neq 0$.

The results are identical to those by applying generic theories in Section 6.2 directly.

Proof of Lemma 57

We first show that $x = 0$ is a strict saddle point, by verifying that $\lambda_{\min}(\nabla^2\mathcal{F}(0)) < 0$ and for any neighborhood $\mathcal{B}$ of $x = 0$, there exist $y_1, y_2 \in \mathcal{B}$ such that $\mathcal{F}(y_1) \leq \mathcal{F}(0) \leq \mathcal{F}(y_2)$.

From (6.8) we have $\nabla^2\mathcal{F}(0) = -uu^\top$. For any $z \in \mathbb{R}^n$ with $\|z\|_2 = 1$, we have

$$z^\top \nabla^2\mathcal{F}(0) z = -(z^\top u)^2 \geq -\|u\|_2^2,$$

where the last inequality is from Cauchy-Schwarz. Then we have $\nabla^2\mathcal{F}(0)$ is negative semi-definite. The minimal eigenvalue is $\lambda_{\min}(\nabla^2\mathcal{F}(0)) = -\|u\|_2^2$ with the corresponding eigenvector $u/\|u\|_2$ and the maximal eigenvalue is $\lambda_{\max}(\nabla^2\mathcal{F}(0)) = 0$ with the corresponding eigenvector z that satisfies $u^\top z = 0$.

Let $y_1 = \alpha u$, where $\alpha \in [0,1]$, and y_2 be any vector that satisfies $y_2^\top u = 0$. Then we have

$$\mathcal{F}(y_1) = \frac{1}{4}\|uu^\top - \alpha^2 uu^\top\|_2^2 = \frac{(1-\alpha^2)}{4}\|uu^\top\|_2^2 \le \frac{1}{4}\|uu^\top\|_2^2 = \mathcal{F}(0) \quad \text{and}$$
$$\mathcal{F}(y_2) = \frac{1}{4}\left(\|uu^\top\|_2^2 + \|y_2\|_2^2\right) \ge \mathcal{F}(0).$$

Therefore, we have $\mathcal{F}(y_1) \le \mathcal{F}(0) \le \mathcal{F}(y_2)$, which implies $x = 0$ is a strict saddle point.

Next, we show that for any $\|x\|_2 \le \frac{1}{2}\|u\|_2$, $\nabla^2\mathcal{F}(x)$ has both positive and negative eigenvalues. Given a point x, let $z_{\max}(x)$ and $z_{\min}(x)$ denote the eigenvectors of $\lambda_{\max}(\nabla^2\mathcal{F}(x))$ corresponding to the largest and smallest eigenvalues respectively. Then for any $x \in \mathcal{R}_1$, $\nabla^2\mathcal{F}(x)$ has at least a positive eigenvalue since

$$z_{\max}^\top(x)\nabla^2\mathcal{F}(x)z_{\max}(x) \ge z_{\max}^\top(0)\nabla^2\mathcal{F}(x)z_{\max}(0) = 2(z_{\max}^\top(0)x)^2 + \|x\|_2^2 \ge \|x\|_2^2.$$

On the other hand, we have $z_{\min}(0) = u/\|u\|_2$ and $\lambda_{\min}(\nabla^2\mathcal{F}(0)) = -\|u\|_2^2$ from the previous discussion. Then for any $x \in \mathcal{R}_1$, $\nabla^2\mathcal{F}(x)$ has at least a negative eigenvalue since

$$\begin{aligned} z_{\min}^\top(x)\nabla^2\mathcal{F}(x)z_{\min}(x) \le z_{\min}^\top(0)\nabla^2\mathcal{F}(x)z_{\min}(0) &= 2(z_{\min}^\top(0)x)^2 + \|x\|_2^2 - \|u\|_2^2 \\ &\le 3\|x\|_2^2 - \|u\|_2^2 \le -\frac{1}{4}\|u\|_2^2. \end{aligned}$$

Proof of Lemma 58

We only discuss the scenario when $x = u$. The argument for $x = -u$ is similar. From the Hessian matrix $\nabla^2\mathcal{F}(x)$ in (6.8), we have $\nabla^2\mathcal{F}(u) = uu^\top + \|u\|_2^2 \cdot I_n$. For any $z \in \mathbb{R}^n$ with $\|z\|_2 = 1$, we have

$$z^\top\nabla^2\mathcal{F}(u)z = (z^\top u)^2 + \|u\|_2^2 \ge \|u\|^2,$$

then $\lambda_{\min}(\nabla^2\mathcal{F}(u)) = \|u\|_2^2$ with the corresponding eigenvector z satisfying $u^\top z = 0$. Therefore, $\nabla^2\mathcal{F}(u)$ is positive definite and $x = u$ is a local minimum of $\mathcal{F}(x)$. Moreover,

$x = u$ is a also a global minimum since

$$\mathcal{F}(u) = \min_{x \in \mathbb{R}^n} \mathcal{F}(x) = 0.$$

On the other hand, let $x = u + e$. For any $x \in \mathcal{R}_2$, we have

$$\begin{aligned}\left| z^\top \left(\nabla^2 \mathcal{F}(x) - \nabla^2 \mathcal{F}(u)\right) z \right| &= \left| z^\top \left(2(u+e)(u+e)^\top + \|x\|_2^2 \cdot I_n - 2uu^\top - \|u\|_2^2 \cdot I_n \right) z \right| \\ &= \left| z^\top \left(2(ee^\top + eu^\top + ue^\top) + (\|e\|_2^2 + 2e^\top u) \cdot I_n \right) z \right| \\ &\le \left(3\|e\|_2^2 + 6\|e\|_2\|u\|_2\right) \cdot \|z\|_2^2 \le \frac{51}{64}\|u\|_2^2,\end{aligned}$$

which further implies

$$z^\top \nabla^2 \mathcal{F}(x) z \ge z^\top \nabla^2 \mathcal{F}(u) z - \left| z^\top \left(\nabla^2 \mathcal{F}(x) - \nabla^2 \mathcal{F}(u)\right) z \right| \ge \frac{1}{5}\|u\|_2^2.$$

Proof of Lemma 59

Let $x = \alpha u + \beta w \|u\|_2$, where $\alpha, \beta \in \mathbb{R}$, $w^\top u = 0$ and $\|w\|_2 = 1$. Then we have

$$\begin{aligned}\|\nabla \mathcal{F}(x)\|_2^2 = \|(xx^\top - uu^\top)x\|_2^2 &= \|(\alpha^2 + \beta^2)\|u\|_2^2 \cdot (\alpha u + \beta w \|u\|_2) - \alpha \|u\|_2^2 \cdot u\|_2^2 \\ &= \|(\alpha^3 + \alpha\beta^2 - \alpha)\|u\|_2^2 \cdot u + \beta(\alpha^2 + \beta^2)\|u\|_2^3 \cdot w\|_2^2 \\ &= [(\alpha^3 + \alpha\beta^2 - \alpha)^2 + \beta^2(\alpha^2 + \beta^2)^2]\|u\|_2^6\end{aligned}$$

Then region $\mathcal{R}_3$ is equivalent to the following set

$$\mathcal{X}_u = \left\{ x = \alpha u + \beta w \|u\|_2 \mid \alpha^2 + \beta^2 > \frac{1}{4}, (\alpha - 1)^2 + \beta^2 > \frac{1}{64} \right\}.$$

By direct calculation, the infimum of $\|\nabla \mathcal{F}(x)\|_2$ subject to $x \in \mathcal{X}_u$ is obtained when $\alpha \to 0$ and $\beta \to \frac{1}{2}$, i.e., $\|\nabla \mathcal{F}(x)\|_2 > \frac{\|u\|_2^3}{8}$.

7.5.3 Proof of Theorem 18

The proof scheme is identical to that of the rank 1 case in Theorem 17. However, the analysis is much more challenging due to the nonisolated strict saddle points and minimum points.

First, we identify the stationary points of $\mathcal{F}(X)$ in the following lemma. The proof is provided in Appendix 7.5.3.

Lemma 60. For any $X \in \mathcal{X}$, X is a stationary point of $\mathcal{F}(X)$.

Next, we characterize two types of stationary points. We show a stronger result in the following lemma that for any $X \in \mathcal{X}$, it is a strict saddle point, where the Hessian matrix has both positive and negative eigenvalues. Further, the Hessian matrix has a negative eigenvalue in the neighborhood of $X \in \mathcal{X}$. The proof is provided in Appendix 7.5.3.

Lemma 61. For any $X \in \mathcal{X} \backslash \mathcal{U}$, X is a strict saddle point with

$$\lambda_{\min}(\nabla^2 \mathcal{F}(X)) \leq -\lambda^2_{\max}(\Sigma_1 - \Sigma_2) \quad \text{and} \quad \lambda_{\max}(\nabla^2 \mathcal{F}(X)) \geq 2\lambda^2_{\max}(\Sigma_2).$$

Moreover, for any $X \in \mathcal{R}_1$, $\nabla^2 \mathcal{F}(X)$ contains a negative eigenvalue, i.e.

$$\lambda_{\min}(\nabla^2 \mathcal{F}(X)) \leq -\frac{\sigma_r^2(U)}{4}.$$

Moreover, we show in the following lemma that for any $X \in \mathcal{U}$, it is a global minimum, and $\mathcal{F}(X)$ is only strongly convex along certain directions in the neighborhood of $X \in \mathcal{U}$. The proof is provided in Appendix 7.5.3.

Lemma 62. For any $X \in \mathcal{U}$, X is a global minimum of $\mathcal{F}(X)$, and $\nabla^2 \mathcal{F}(X)$ is positive semidefinite, which has exactly $r(r-1)/2$ zero eigenvalues with the minimum nonzero eigenvalue at least $\sigma_r^2(U)$. Moreover, for any $X \in \mathcal{R}_2$, we have

$$z^\top \nabla^2 \mathcal{F}(X) z \geq \frac{1}{5}\sigma_r^2(U)\|z\|_2^2$$

for any $z \perp \mathcal{E}$, where $\mathcal{E} \subseteq \mathbb{R}^{n \times r}$ is a subspace is spanned by all eigenvectors of $\nabla^2 \mathcal{F}(K_E)$ associated with the negative eigenvalues, where $E = X - U\Psi_X$ and Ψ_X and K_E are defined in (6.10).

Finally, we show in the following lemma that the gradient $\nabla \mathcal{F}(X)$ has a sufficiently large norm outside the neighborhood of stationary points. The proof is provided in Appendix 7.5.3

Lemma 63. The gradient $\nabla \mathcal{F}(X)$ has sufficiently large norm in $\mathcal{R}_3'$ and $\mathcal{R}_3''$, i.e.,

$$\|\nabla \mathcal{F}(X)\|_{\mathrm{F}} > \frac{\sigma_r^4(U)}{9\sigma_1(U)} \quad \text{for any } X \in \mathcal{R}_3', \quad \text{and}$$
$$\|\nabla \mathcal{F}(X)\|_{\mathrm{F}} > \frac{3}{4}\sigma_1^3(X) \quad \text{for any } X \in \mathcal{R}_3''.$$

Combining Lemma 60 – Lemma 63, we finish the proof.

Proof of Lemma 60

We provide an algebraic approach to determine stationary points here. We denote $X = \Phi\Sigma_2\Theta_2 + W$, where $W^\top\Phi = 0$. Accordingly, we solve

$$\begin{aligned}(XX^\top - UU^\top)X &= [(\Phi\Sigma_2\Theta_2 + W)(\Phi\Sigma_2\Theta_2 + W)^\top - UU^\top](\Phi\Sigma_2\Theta_2 + W)\\ &= \Phi\Sigma_2\Theta_2^\top(\Theta_2\Sigma_2^2\Theta_2 + W^\top W) + W(\Theta_2\Sigma_2^2\Theta_2 + W^\top W) - \Phi\Sigma_1^2\Sigma_2\Theta_2^\top = 0.\end{aligned}$$

1. Suppose $\Sigma_2 = 0$, which implies

   $$WW^\top W = 0.$$

   The solution to the equation above is $W = 0$, which indicates that $X = 0$ is a stationary point.

2. Suppose $W = 0$, which implies

   $$\Phi(\Sigma_2^2 - \Sigma_1^2)\Sigma_2\Theta_2^\top = 0.$$

   The solution to the equation above is $(\Sigma_2^2 - \Sigma_1^2)\Sigma_2$, which indicates that $\Phi\Sigma_2\Theta_2$ is a stationary point for any $\Theta_2 \in \mathfrak{O}_r$ and $\Sigma_2 = \Sigma_1 I_{\text{Mask}}$, where I_{Mask} is setting arbitrary number of diagonal elements of the identity matrix as 0 at arbitrary locations (include 2^r combinations). This includes $X = 0$ and $X = U\Psi$ for any $\Psi \in \mathfrak{O}_r$ as special examples.

3. Suppose $\Sigma_2 \neq 0$ and $W \neq 0$. Since Φ and W have orthogonal column spaces, we

then require

$$\Theta_2\Sigma_2^2\Theta_2 + W^\top W = 0,$$

which further implies

$$\Phi\Sigma_1^2\Sigma_2\Theta_2^\top = 0.$$

The solution to the equation above is $\Sigma_2 = 0$, which conflicts with the assumption. This finishes the proof.

The results are identical to those by applying generic theories in Section 6.2 directly.

Proof of Lemma 61

For notational convenience, denote $\widetilde{\mathcal{X}} = \mathcal{X}\backslash\mathcal{U}$. Associate each $X \in \widetilde{\mathcal{X}}$ with a rank deficient set $\mathcal{S} \subseteq [r]$, $\mathcal{S} \neq \emptyset$, which is equivalent with saying that $\Sigma_2 = \Sigma_1 D$, where D is a diagonal matrix with $D_{ii} = 0$ for all $i \in \mathcal{S}$, and $D_{jj} = 1$ for all $j \in \overline{\mathcal{S}} = [r]\backslash\mathcal{S}$. Let $s \in \mathcal{S}$ be the smallest index value in $\mathcal{S}$ and $\overline{s} \in \overline{\mathcal{S}}$ be the smallest index value in $\overline{\mathcal{S}}$.

Part 1. We first show that the rank deficient stationary points are strict saddle points, i.e., their eigenvalue satisfies

$$\lambda_{\min}(\nabla^2\mathcal{F}(X)) \leq -\sigma_s^2(U)$$
$$\lambda_{\max}(\nabla^2\mathcal{F}(X)) \geq 2\sigma_{\overline{s}}^2(U).$$

If $\overline{\mathcal{S}} = \emptyset$, i.e., $X = 0$, then $\lambda_{\max}(\nabla^2\mathcal{F}(X)) \geq 0$.

We start with the proof of $\lambda_{\min}(\nabla^2\mathcal{F}(X))$. Remind that

$$K_X = \begin{bmatrix} X_{(*,1)}X_{(*,1)}^\top & X_{(*,2)}X_{(*,1)}^\top & \cdots & X_{(*,r)}X_{(*,1)}^\top \\ X_{(*,1)}X_{(*,2)}^\top & X_{(*,2)}X_{(*,2)}^\top & \cdots & X_{(*,r)}X_{(*,2)}^\top \\ \vdots & \vdots & \ddots & \vdots \\ X_{(*,1)}X_{(*,r)}^\top & X_{(*,2)}X_{(*,r)}^\top & \cdots & X_{(*,r)}X_{(*,r)}^\top \end{bmatrix}.$$

Let $X_{(*,1)}, \ldots, X_{(*,r)}$ be the columns of X. Since X is rank deficient, then there exists a unit vector $w = [w_1, \ldots, w_r]^\top \in \mathbb{R}^r$, $\|w\|_2 = 1$, such that $w^\top X^\top X w = 0$. Let ϕ_s be the s-th column of Φ, which satisfies $\phi_s^\top X_{(*,i)} = 0$ for any $i \in [r]$ from the construction of X, and $z = [z_1^\top, \ldots, z_r^\top]^\top \in \mathbb{R}^{nr}$ be a vector by taking the i-th subvector as $z_i = w_{(i)}\phi_s \in \mathbb{R}^n$ for all $i \in [r]$, then

$$\begin{aligned}\lambda_{\min}(\nabla^2\mathcal{F}(X)) &\le z^\top \nabla^2 \mathcal{F}(X) z = z^\top (K_X + I_r \otimes XX^\top + X^\top X \otimes I_n - I_r \otimes UU^\top) z \\ &= \sum_{i,j}^{r} w_{(i)} w_{(j)} \phi_s^\top X_{(*,j)} X_{(*,i)}^\top \phi_s + \phi_s^\top XX^\top \phi_s + w^\top X^\top X w - \phi_s^\top UU^\top \phi_s \\ &= 0 + 0 + 0 - \sigma_s^2(U) = -\sigma_s^2(U).\end{aligned}$$

The proof of $\lambda_{\max}(\nabla^2\mathcal{F}(X))$ follows analogous analysis. Let a unit vector $w = [w_1, \ldots, w_r]^\top \in \mathbb{R}^r$ be the singular vector of $X^\top X$ corresponding to the largest singular value $\sigma_{\overline{s}}^2(U)$, and $\phi_{\overline{s}}$ be the $\overline{s}$-th column of Φ, i.e., $\phi_{\overline{s}}^\top XX^\top \phi_{\overline{s}} = \phi_{\overline{s}}^\top UU^\top \phi_{\overline{s}} = \sigma_{\overline{s}}^2(U)$. Let $z = [z_1^\top, \ldots, z_r^\top]^\top \in \mathbb{R}^{nr}$ be a vector by taking the i-th subvector as $z_i = w_{(i)}\phi_{\overline{s}} \in \mathbb{R}^n$ for all $i \in [r]$, then

$$\begin{aligned}\lambda_{\max}(\nabla^2\mathcal{F}(X)) &\ge z^\top \nabla^2 \mathcal{F}(X) z = z^\top (K_X + I_r \otimes XX^\top + X^\top X \otimes I_n - I_r \otimes UU^\top) z \\ &= \sum_{i,j}^{r} w_{(i)} w_{(j)} \phi_{\overline{s}}^\top X_{(*,j)} X_{(*,i)}^\top \phi_{\overline{s}} + \phi_{\overline{s}}^\top XX^\top \phi_{\overline{s}} + w^\top X^\top X w - \phi_{\overline{s}}^\top UU^\top \phi_{\overline{s}} \\ &= \sigma_{\overline{s}}^2(U) + \sigma_{\overline{s}}^2(U) + \sigma_{\overline{s}}^2(U) - \sigma_{\overline{s}}^2(U) = 2\sigma_{\overline{s}}^2(U).\end{aligned}$$

When $X = 0$, let $w \in \mathbb{R}^r$ be any unit vector and $\phi \in \mathbb{R}^n$ be a unit vector that satisfies $\phi^\top \Phi = 0$. Construct $z \in \mathbb{R}^{nr}$ as the same way above, then

$$\begin{aligned}\lambda_{\max}(\nabla^2\mathcal{F}(X)) &\ge z^\top \nabla^2 \mathcal{F}(X) z = z^\top (K_X + I_r \otimes XX^\top + X^\top X \otimes I_n - I_r \otimes UU^\top) z \\ &= \sum_{i,j}^{r} w_{(i)} w_{(j)} \phi^\top X_{(*,j)} X_{(*,i)}^\top \phi + \phi^\top XX^\top \phi + w^\top X^\top X w - \phi^\top UU^\top \phi \\ &= 0 + 0 + 0 - 0 = 0.\end{aligned}$$

Next, we show that for any neighborhood $\mathcal{B}$ of $X \in \widetilde{\mathcal{X}}$, there exist $Y_1, Y_2 \in \mathcal{B}$ such that $\mathcal{F}(Y_1) \le \mathcal{F}(X) \le \mathcal{F}(Y_2)$. Suppose $X = \Phi\Sigma_1 D\Theta_2$ and $E_1 = \Phi\Sigma_1 D_1\Theta_2$, where

$D + D_1 = I$, then $\langle E_1, X\rangle = 0$. Given $\alpha \in [0, \sqrt{2}]$, let $Y_1 = X + \alpha E_1$, then we have

$$\begin{aligned}
\mathcal{F}(Y_1) &= \frac{1}{4}\|Y_1 Y_1^\top - UU^\top\|_{\mathrm{F}}^2 \\
&= \frac{1}{4}\left(\|XX^\top - UU^\top\|_{\mathrm{F}}^2 + \alpha^4\|E_1E_1^\top\|_{\mathrm{F}}^2 + 2\langle \alpha^2 E_1 E_1^\top, XX^\top - UU^\top\rangle\right) \\
&= \mathcal{F}(X) + \frac{1}{4}\langle \alpha^2 \Phi\Sigma_1^2 D_1 \Phi^\top, \alpha^2\Phi\Sigma_1^2 D_1\Phi^\top - 2\Phi\Sigma_1^2 D_1 \Phi^\top\rangle \\
&= \mathcal{F}(X) + \frac{(\alpha^4 - 2\alpha^2)}{4}\Phi\Sigma_1^4 D_1 \Phi^\top \\
&\leq \mathcal{F}(X).
\end{aligned}$$

Similarly, let $E_2 = \widetilde{\Phi}\widetilde{\Sigma}\widetilde{\Theta}$, where $\widetilde{\Phi} \in \mathbb{R}^{n\times r}$ has orthogonal columns satisfying $\widetilde{\Phi}^\top \Phi = 0$, $\widetilde{\Sigma} \in \mathbb{R}^{r\times r}$ is any diagonal matrix with nonnegative entries, and $\widetilde{\Theta} \in \mathbb{R}^{r\times r}$ is any orthogonal matrix. Given $\alpha \geq 0$, let $Y_2 = X + \alpha E_2$, then we have

$$\mathcal{F}(Y_2) = \frac{1}{4}\|Y_2Y_2^\top - UU^\top\|_{\mathrm{F}}^2 = \frac{1}{4}\left(\|XX^\top - UU^\top\|_{\mathrm{F}}^2 + \alpha^4\|E_2E_2^\top\|_{\mathrm{F}}^2\right) \geq \mathcal{F}(X).$$

Part 2. Next, we show that for any X in a neighborhood of saddle points, the Hessian matrix $\nabla^2\mathcal{F}(X)$ has a negative eigenvalue. Given any $X^* \in \widetilde{\mathcal{X}}$ with the associated rank deficient set $\mathcal{S}^* \subseteq [r]$, $\mathcal{S} \neq \emptyset$, let $X = X^* + E$. For any $s \in \mathcal{S}^*$, let ϕ_s be the corresponding singular vector of U, i.e., the s-th column of Φ, $w \in \mathbb{R}^r$ be the singular vector of $X^\top X$ associated with the smallest singular value, and $z \in \mathbb{R}^{nr}$ be a unit vector with the i-th subvector as $z_i = w_{(i)}\phi_s$ for all $i \in [r]$, then

$$\begin{aligned}
\lambda_{\min}(\nabla^2\mathcal{F}(X)) &\leq z^\top \nabla^2 \mathcal{F}(X) z = z^\top(K_X + I_r \otimes XX^\top + X^\top X \otimes I_n - I_r \otimes UU^\top)z \\
&= \sum_{i,j}^{r} w_{(i)}w_{(j)}\phi_s^\top X_{(*,j)}X_{(*,i)}^\top \phi_s + \phi_s^\top XX^\top\phi_s + w^\top X^\top X w - \phi_s^\top UU^\top \phi_s \\
&= (\phi_s^\top E w^\top)^2 + \phi_s^\top EE^\top \phi_s + \sigma_r^2(X) - \sigma_s^2(U) \\
&\leq 2\|\phi_s^\top E\|_2^2 + \sigma_r^2(X) - \sigma_s^2(U). \qquad (7.177)
\end{aligned}$$

We claim that from (7.177), if $\sigma_r(X) \leq \frac{1}{2}\sigma_r(U)$, we have

$$\lambda_{\min}(\nabla^2\mathcal{F}(X)) \leq -\frac{1}{4}\sigma_r^2(U).$$

The discussion is addressed by the following cases. Let $\mathcal{L}_{\Phi}$ denote the column space of Φ and $\mathcal{L}_{\Phi_{\mathcal{S}^*}}$ be the column space of $\Phi_{\mathcal{S}^*}$.

Case 1: Suppose X is rank deficient, i.e., $\sigma_r(X) = 0$. Without loss of generality, we can argue that E is also rank deficient. Otherwise, if E is full rank, then there exist some subspace in columns of X^* eliminated by the corresponding subspace in columns of E. Therefore, we can consider the rank is deficient in both X^* and E in that particular subspace. Then there exists a subspace $\mathcal{L}_1 \subset \mathcal{L}_{\Phi}$ such that $E = \mathcal{P}_{\mathcal{L}_1}(E) + (I - \mathcal{P}_{\mathcal{L}_{\Phi}})(E)$. We can always find a $s \in \mathcal{S}^*$ such that $\phi_s \in \mathcal{L}_{\Phi} \backslash \mathcal{L}_1$, i.e., $\phi_s^\top \mathcal{P}_{\mathcal{L}_1} x = 0$ for any $x \in \mathbb{R}^n$, such that

$$\phi_s^\top E = \phi_s^\top (\mathcal{P}_{\mathcal{L}_1}(E) + (I - \mathcal{P}_{\mathcal{L}_{\Phi}})(E)) = 0.$$

This further implies

$$\lambda_{\min}(\nabla^2 \mathcal{F}(X)) \le -\sigma_s^2(U) \le -\sigma_r^2(U).$$

Case 2: Suppose X has full column rank, and the singular vector y associated with the smallest singular value $\sigma_r(X)$ satisfies $\|\mathcal{P}_{\mathcal{L}_{\Phi}}(y)\|_2 = 0$ without loss of generality. This implies that for any singular vector $\tilde{y}$ of X, there exists $s \in \mathcal{S}^*$ such that $\phi_s^\top(\tilde{y}) = 0$. This further implies $\phi_s^\top E = 0$, then combining with (7.177) we have

$$\lambda_{\min}(\nabla^2 \mathcal{F}(X)) \le \sigma_r^2(X) - \sigma_s^2(U) \le -\frac{3}{4}\sigma_r^2(U).$$

Case 3: Suppose X has full column rank, and the singular vector y associated with the smallest singular value $\sigma_r(X)$ satisfies $\|\mathcal{P}_{\mathcal{L}_{\Phi}}(y)\|_2 \in (0, 1]$. This implies that there exists $s \in \mathcal{S}^*$ such that $\|\phi_s^\top E\|_2 \le \sigma_r(X)$ without loss of generality because there exists a potential subspace of E that is orthogonal to ϕ_s. If the singular vector associated with smallest singular value of X is not closest to ϕ_s for any $s \in \mathcal{S}^* \subset [r]$, then it must be closest to some other $s' \in [r] \backslash \mathcal{S}^*$. Then we can always consider the rank is deficient for s' without loss of generality and the same argument above holds. This further results in

$$\lambda_{\min}(\nabla^2 \mathcal{F}(X)) \le 3\sigma_r^2(X) - \sigma_s^2(U) \le -\frac{1}{4}\sigma_r^2(U).$$

Proof of Lemma 62

It is obvious that for any $X \in \mathcal{U}$, $\mathcal{F}(X) = 0$, thus it is a global minimum since $\mathcal{F}(Y) \geq 0$ for any $Y \in \mathbb{R}^{n \times r}$. Without loss of generality, let $X = U$, i.e., $\Psi = I$, then we have

$$\nabla^2 \mathcal{F}(U) = K_U + \begin{bmatrix} U_{(*,1)}^\top U_{(*,1)} \cdot I_n & U_{(*,1)}^\top U_{(*,2)} \cdot I_n & \cdots & U_{(*,1)}^\top U_{(*,r)} \cdot I_n \\ U_{(*,2)}^\top U_{(*,1)} \cdot I_n & U_{(*,2)}^\top U_{(*,2)} \cdot I_n & \cdots & U_{(*,2)}^\top U_{(*,r)} \cdot I_n \\ \vdots & \ddots & \vdots & \vdots \\ U_{(*,r)}^\top U_{(*,1)} \cdot I_n & U_{(*,r)}^\top U_{(*,2)} \cdot I_n & \cdots & U_{(*,r)}^\top U_{(*,r)} \cdot I_n \end{bmatrix}.$$

Part 1. We first characterize the eigenvectors associated with zero eigenvalues of $\nabla^2 \mathcal{F}(U)$. For any i and j chosen from $1, ..., r$, where $i < j$, we define a vector $v^{(i,j)} \in \mathbb{R}^{nr}$ as

$$v^{(i,j)} = [0^\top, ..., \underbrace{-U_{(*,j)}^\top}_{i\text{-th block}}, ..., \underbrace{U_{(*,i)}^\top}_{j\text{-th block}}, ..., 0^\top]^\top / \sqrt{\|U_{(*,j)}\|_2^2 + \|U_{(*,i)}\|_2^2},$$

where $-U_{(*,j)}$ is the i-th block of $v^{(i,j)}$, and $U_{(*,i)}$ is the j-th block of $v^{(i,j)}$. Then we can verify

$$\begin{aligned} & v^{(i,j)\top} \nabla^2 \mathcal{F}(U) v^{(i,j)} \cdot \left(\|U_{(*,j)}\|_2^2 + \|U_{(*,i)}\|_2^2\right) \\ &= U_{(*,j)}^\top U_{(*,j)} \cdot U_{(*,i)}^\top U_{(*,i)} - U_{(*,j)}^\top U_{(*,i)} \cdot U_{(*,i)}^\top U_{(*,j)} - U_{(*,i)}^\top U_{(*,i)} \cdot U_{(*,j)}^\top U_{(*,i)} \\ &\quad + U_{(*,i)}^\top U_{(*,i)} \cdot U_{(*,j)}^\top U_{(*,j)}^\top + U_{(*,j)}^\top U_{(*,i)} \cdot U_{(*,i)}^\top U_{(*,j)} - U_{(*,j)}^\top U_{(*,j)} \cdot U_{(*,i)}^\top U_{(*,i)} \\ &\quad - U_{(*,i)}^\top U_{(*,i)} \cdot U_{(*,j)}^\top U_{(*,j)} + U_{(*,i)}^\top U_{(*,j)} \cdot U_{(*,j)}^\top U_{(*,i)} = 0, \end{aligned}$$

which implies that $v^{(i,j)}$ is an eigenvector of $\nabla^2 \mathcal{F}(U)$ and the associated eigenvalue is 0.

We then prove the linear independence among all $v^{(i,j)}$'s by contradiction. Assume that all $v^{(i,j)}$'s are linearly dependent. Then there exist $\alpha_{(i,j)}$'s with at least two nonzero $\alpha_{(i,j)}$'s such that

$$\sum_{i<j} \alpha_{(i,j)} v^{(i,j)} = 0.$$

This further implies that for any $i < k < j$, we have

$$\alpha_{(i,k)} U_{(*,i)} - \alpha_{(k,j)} U_{(*,j)} = 0.$$

Since $U_{(*,j)}$ and $U_{(*,i)}$ are linearly independent, we must have $\alpha_{(i,k)} = \alpha_{(k,j)} = 0$. This is contradicted by our assumption. Thus, all $v^{(i,j)}$'s are linearly independent, i.e., we can obtain all $r(r-1)/2$ eigenvectors associated with zero eigenvalues of $\nabla^2 \mathcal{F}(U)$ by conducting the orthogonalization over all $v^{(i,j)}$. Meanwhile, this also implies that $\mathcal{F}(X)$ is not strongly convex at $X = U$.

We then show that the minimum nonzero eigenvalue of $\nabla^2 \mathcal{F}(U)$ is lower bounded by $\sigma_r^2(U)$. We consider a vector

$$z = [z_1^\top, ..., z_r^\top]^\top \in \mathbb{R}^{nr},$$

which is orthogonal to all $v^{(i,j)}$, i.e., for any $i < j$, we have

$$z_i^\top U_{(*,j)} = z_j^\top U_{(*,i)}.$$

Meanwhile, we also have

$$\begin{aligned} z^\top \nabla^2 \mathcal{F}(U) z &= z^\top (U^\top U \otimes I) z + \sum_{i=1}^r (z_i^\top U_{(*,i)})^2 + 2 \sum_{j<k} (z_j^\top U_{(*,k)})(z_k^\top U_{(*,j)}) \\ &= z^\top (U^\top U \otimes I) z + \sum_{i=1}^r (z_i^\top U_{(*,i)})^2 + 2 \sum_{j<k} (z_j^\top U_{(*,k)})^2 \\ &= z^\top (U^\top U \otimes I) z + z^\top (I \otimes U U^\top) z. \end{aligned}$$

We can construct a valid z as follows: let $w = [w_1, ..., w_r]^\top \in \mathbb{R}^r$ be the eigenvector associated with the smallest eigenvalue of $U^\top U$, and y be a vector, which is orthogonal to all $U_{(*,i)}$'s. Then we take

$$z_i = w_{(i)} y.$$

It can be further verified that $z^\top v^{(i,j)} = 0$ for any (i,j), and $z^\top (I \otimes U U^\top) z = 0$. Since both $U^\top U \otimes I$ and $I \otimes U U^\top$ are PSD matrices, then we have from the Weyl's inequality that the minimum nonzero eigenvalue $\lambda_{\min}^{+}(\nabla^2\mathcal{F}(U))$ of $\nabla^2\mathcal{F}(U)$ satisfies

$$\lambda_{\min}^{+}(\nabla^2\mathcal{F}(U)) \geq \lambda_{\min}^{+}(U^\top U \otimes I) = z^\top(U^\top U \otimes I)z = \lambda_{\min}(U^\top U) = \sigma_r^2(U).$$

Part 2. Next, we characterize the neighborhood of the global minima. Let $E = X - U$. We then have

$$\begin{aligned} & z^\top(\nabla^2\mathcal{F}(X) - \nabla^2\mathcal{F}(U))z \\ & = z^\top(I_r \otimes (UE^\top + EU^\top + EE^\top) + (U^\top E + E^\top U + E^\top E) \otimes I_n + K_E + \tilde{E}_1 + \tilde{E}_2)z, \end{aligned}$$

where $\tilde{E}_1$ and $\tilde{E}_2$ are defined as

$$\tilde{E}_1 = \begin{bmatrix} E_{(*,1)}U_{(*,1)}^\top & E_{(*,2)}U_{(*,1)}^\top & \cdots & E_{(*,r)}U_{(*,1)}^\top \\ E_{(*,1)}U_{(*,2)}^\top & E_{(*,2)}U_{(*,2)}^\top & \cdots & E_{(*,r)}U_{(*,2)}^\top \\ \vdots & \ddots & \vdots & \vdots \\ E_{(*,1)}U_{(*,r)}^\top & E_{(*,2)}U_{(*,r)}^\top & \cdots & E_{(*,r)}U_{(*,r)}^\top \end{bmatrix},$$

$$\tilde{E}_2 = \begin{bmatrix} U_{(*,1)}E_{(*,1)}^\top & U_{(*,2)}E_{(*,1)}^\top & \cdots & U_{(*,r)}E_{(*,1)}^\top \\ U_{(*,1)}E_{(*,2)}^\top & U_{(*,2)}E_{(*,2)}^\top & \cdots & U_{(*,r)}E_{(*,2)}^\top \\ \vdots & \ddots & \vdots & \vdots \\ U_{(*,1)}E_{(*,r)}^\top & U_{(*,2)}E_{(*,r)}^\top & \cdots & U_{(*,r)}E_{(*,r)}^\top \end{bmatrix}.$$

Meanwhile, we have

$$
\begin{aligned}
|z^\top(I_r \otimes (UE^\top + EU^\top + EE^\top))z| &\leq \|UE^\top + EU^\top + EE^\top\|_2 \|z\|_2^2 \\
&\leq (2\sigma_1(U)\|E\|_2 + \|E\|_2^2)\|z\|_2^2, \\
|z^\top((U^\top E + E^\top U + E^\top E) \otimes I_n)z| &\leq \|U^\top E + E^\top U + E^\top E\|_2 \|z\|_2^2 \\
&\leq (2\sigma_1(U)\|E\|_2 + \|E\|_2^2)\|z\|_2^2, \\
\left|z^\top \left(\tilde{E}_1 + \tilde{E}_2\right) z\right| &= \left|\sum_{i,j} z_i^\top E_{(*,j)} U_{(*,i)}^\top z_j + \sum_{i,j} z_i^\top U_{(*,j)} E_{(*,i)}^\top z_j\right| \\
&= 2\left|\sum_{i,j} z_i^\top E_{(*,j)} U_{(*,j)}^\top z_i\right| = 2\left|\sum_{i} z_i^\top E U^\top z_i\right| \\
&\leq 2\sigma_1(U)\|E\|_2 \sum_{j=1}^{r} \|z_i\|_2^2 = 2\sigma_1(U)\|E\|_2\|z\|_2^2.
\end{aligned}
$$

where the second equality comes from $z_i^\top U_{(*,j)} = z_j^\top U_{(*,i)}$ for all i, j's by constructing z as in Part 1.

We then characterize the eigenvectors associated with negative eigenvalues of K_E. For any i and j chosen from $1, ..., r$, where $i < j$, we define

$$
w^{(i,j)} = [0^\top, ..., \underbrace{-E_{(*,j)}^\top}_{i\text{-th block}}, ..., \underbrace{E_{(*,i)}^\top}_{j\text{-th block}}, ..., 0^\top]^\top / \sqrt{\|E_{(*,j)}\|_2^2 + \|E_{(*,i)}\|_2^2},
$$

where the i-th block of $w^{(i,j)}$ is $-E_{(*,j)}$, and the j-th block of $w^{(i,j)}$ is $E_{(*,i)}$. Then we have

$$
K_E w^{(i,j)} = \underbrace{\frac{2\left(E_{(*,i)}^\top E_{(*,j)}\right)^2 - 2\|E_{(*,i)}\|_2^2 \|E_{(*,j)}\|_2^2}{\|E_{(*,i)}\|_2^2 + \|E_{(*,j)}\|_2^2}}_{\widetilde{\lambda}} w^{(i,j)},
$$

which implies that $w^{(i,j)}$ is an eigenvector of K_E and the associated eigenvalue $\widetilde{\lambda}$ is nonpositive by the Cauchy-Schwarz inequality.

We then prove the linear independence among all $w^{(i,j)}$'s by contradiction. Assume that all $w^{(i,j)}$'s are linearly dependent. Then there exist $\alpha_{(i,j)}$'s with at least two nonzero

$\alpha_{(i,j)}$'s such that

$$\sum_{i<j} \alpha_{(i,j)} w^{(i,j)} = 0.$$

This further implies that for any $i < k < j$, we have

$$\alpha_{(i,k)} E_{(*,i)} - \alpha_{(k,j)} E_{(*,j)} = 0.$$

Since $E_{(*,j)}$ and $E_{(*,i)}$ are linearly independent, we must have $\alpha_{(i,k)} = \alpha_{(k,j)} = 0$. This is contradicted by our assumption. Thus, all $w^{(i,j)}$'s are linearly independent, i.e., we can obtain all $r(r-1)/2$ eigenvectors associated with negative eigenvalues of K_E by conducting the orthogonalization over all $w^{(i,j)}$'s.

We consider to construct z analogous to that in Part 1, which is orthogonal to all $w^{(i,j)}$'s. Then we have

$$z^\top w^{(i,j)} = z_i^\top E_{(*,j)} - z_j^\top E_{(*,i)} = 0 \quad \text{for any } i \text{ and } j.$$

This further implies

$$\begin{aligned} & z^\top \begin{bmatrix} E_{(*,1)} E_{(*,1)}^\top & E_{(*,2)} E_{(*,1)}^\top & \cdots & E_{(*,r)} E_{(*,1)}^\top \\ E_{(*,1)} E_{(*,2)}^\top & E_{(*,2)} E_{(*,2)}^\top & \cdots & E_{(*,r)} E_{(*,2)}^\top \\ \vdots & \ddots & \vdots & \vdots \\ E_{(*,1)} E_{(*,r)}^\top & E_{(*,2)} E_{(*,r)}^\top & \cdots & E_{(*,r)} E_{(*,r)}^\top \end{bmatrix} z \\ &= \sum_{i,j} z_i^\top E_{(*,j)} E_{(*,i)}^\top z_j = \sum_{i,j} z_i^\top E_{(*,j)} E_{(*,j)}^\top z_i = z^\top (I_r \otimes EE^\top) z. \end{aligned}$$

Note that $0 \leq z^\top (I_r \otimes EE^\top) z \leq \sigma_1^2(E) \|z\|_2^2$, which implies $\|K_E\|_2 \leq \sigma_1^2(E)$. Thus, there exists no other eigenvector associated with negative eigenvalues of K_E besides all $w^{(i,j)}$'s. Meanwhile, we also have

$$\begin{aligned} \lambda_{\min}(K_E) &= \min_{i,j} \frac{2(E_{(*,i)}^\top E_{(*,j)})^2 - 2\|E_{(*,i)}\|_2^2 \|E_{(*,j)}\|_2^2}{\|E_{(*,i)}\|_2^2 + \|E_{(*,j)}\|_2^2} \geq -\max_{i,j} \frac{2\|E_{(*,i)}\|_2^2 \|E_{(*,j)}\|_2^2}{\|E_{(*,i)}\|_2^2 + \|E_{(*,j)}\|_2^2} \\ &\geq -\sigma_1^2(E). \end{aligned}$$

Combining all results above, we need

$$\|E\|_2 \leq \frac{\sigma_r^2(U)}{8\sigma_1(U)}$$

such that

$$|z^\top(\nabla^2\mathcal{F}(X) - \nabla^2\mathcal{F}(U))z| \leq (6\sigma_1(U)\|E\|_2 + 3\|E\|_2^2)\|z\|_2^2 < \frac{4\sigma_r^2(U)}{5}\|z\|_2^2.$$

This implies that

$$z^\top\nabla^2\mathcal{F}(X)z \geq z^\top\nabla^2\mathcal{F}(U)z - |z^\top(\nabla^2\mathcal{F}(X) - \nabla^2\mathcal{F}(U))z| > \frac{\sigma_r^2(U)}{5}\|z\|_2^2,$$

since z is orthogonal to the eigenvectors corresponding to the zero eigenvalues of $\nabla^2\mathcal{F}(U)$ by the way of its construction.

Proof of Lemma 63

Part 1. We first discuss $X \in \mathcal{R}_3'$. Recall that $\nabla\mathcal{F}(X) = (XX^\top - UU^\top)X$. For notational simplicity, let $U = U\Psi_X$, where $\Psi_X = \arg\min_{\Psi\in\mathfrak{O}_r} \|X - U\Psi\|_2$.

Let the compact SVD be $X = \Phi_1\Sigma_1\Theta_1^\top$, $\Phi_1, \in \mathbb{R}^{n\times r}$, $\Sigma_1, \Sigma_2 \in \mathbb{R}^{r\times r}$. Then we have

$$\|(XX^\top - UU^\top)X\|_\mathrm{F}^2 \geq \|(XX^\top - UU^\top)X\|_2^2 \geq \|(XX^\top - UU^\top)\|_2^2 \cdot \sigma_r^2(X). \quad (7.178)$$

Moreover, we claim that

$$\|XX^\top - UU^\top\|_2^2 \geq 2(\sqrt{2}-1)\sigma_r^2(U) \cdot \min_{\Psi\in\mathfrak{O}_r} \|X - U\Psi\|_2^2. \quad (7.179)$$

We then demonstrate (7.179). Let $E = X - U\Psi_X$ with $\Psi_X = \mathrm{argmin}_{\Psi\in\mathfrak{O}_r} \|X - U\Psi\|_2^2$ and the SVD of $U^\top X$ be $U^\top X = A\Sigma B^\top$, then we have $\Psi_X = AB^\top$. This implies

$$X^\top U\Psi_X = B\Sigma B^\top = \Psi_X^\top U^\top X \succeq 0.$$

Further, we have $E^\top U\Psi_X$ is symmetric since

$$E^\top U\Psi_X = X^\top U\Psi_X - \Psi_X^\top U^\top U\Psi_X = \Psi_X^\top U^\top X - \Psi_X^\top U^\top U\Psi_X = \Psi_X^\top U^\top E.$$

Without loss of generality, we assume $\Psi_X = I$, then we have $X^\top U \succeq 0$ and $E^\top U = U^\top E$. Substituting $X = U + E$ and denoting $\alpha = 2(\sqrt{2}-1)\sigma_r^2(U)$, we have

$$\begin{aligned}
0 &\le \lambda_{\max}\left(\left(XX^\top - UU^\top\right)^\top \left(XX^\top - UU^\top\right)\right) - \alpha\lambda_{\max}\left((X-U)^\top (X-U)\right) \\
&\le \lambda_{\max}\left(\left(XX^\top - UU^\top\right)^\top \left(XX^\top - UU^\top\right) - \alpha\,(X-U)^\top (X-U)\right) \\
&= \lambda_{\max}\left(\left(E^\top E\right)^2 + 4E^\top EE^\top U + 2(E^\top U)^2 + 2U^\top UE^\top E - \alpha E^\top E\right) \\
&= \lambda_{\max}\left(\left(E^\top E + \sqrt{2}E^\top U\right)^2 + (4-2\sqrt{2})E^\top EE^\top U + 2U^\top UE^\top E - \alpha E^\top E\right).
\end{aligned}$$

This implies we only need to show that

$$\lambda_{\max}\left((E^\top E + \sqrt{2}E^\top U)^2 + (4-2\sqrt{2})E^\top EE^\top U + 2U^\top UE^\top E - \alpha E^\top E\right) \ge 0.$$

It is sufficient to show that $(4-2\sqrt{2})E^\top U + 2U^\top U - \alpha I_r \succeq 0$. From $E = X - U$ and $X^\top U \succeq 0$, we have

$$(4-2\sqrt{2})E^\top U + 2U^\top U - \alpha I_r = (4-2\sqrt{2})X^\top U + 2(\sqrt{2}-1)U^\top U - \alpha I_r \succeq 0,$$

provided $2(\sqrt{2}-1)U^\top U - \alpha I_r \succeq 0$, which is satisfied by the choice of α.

Combining (7.178), (7.179), and $\min_{\Psi\in\mathfrak{O}_r}\|X - U\Psi\|_2 > \frac{\sigma_r^2(U)}{8\sigma_1(U)}$, we have

$$\|(XX^\top - UU^\top)X\|_{\mathrm{F}}^2 \ge 2(\sqrt{2}-1)\sigma_r^4(U)\cdot\frac{\sigma_r^4(U)}{64\sigma_1^2(U)} \ge \frac{\sigma_r^8(U)}{81\sigma_1^2(U)}.$$

Part 2. Next, we discuss $X \in \mathcal{R}_3''$.

Let $U = \Phi_1\Sigma_1\Theta_1^\top$ and $X = \Phi_2\Sigma_2\Theta_2^\top$ be the SVDs, then we have a lower bound of

$\|\nabla\mathcal{F}(X)X^\top\|_{\mathrm{F}}$ when X and U has the same column space, i.e,

$$\begin{aligned}
&\|\nabla\mathcal{F}(X)X^\top\|_{\mathrm{F}} = \|XX^\top XX^\top - UU^\top XX^\top\|_{\mathrm{F}} = \|\Phi_2\Sigma_2^4\Phi_2^\top - \Phi_1\Sigma_1^2\Phi_1\Phi_2\Sigma_2^2\Phi_2\|_{\mathrm{F}} \\
&= \sqrt{\|\Sigma_2^4\|_{\mathrm{F}}^2 + \|\Sigma_1^2\Phi_1^\top\Phi_2\Sigma_2^2\|_{\mathrm{F}}^2 - 2\mathrm{Tr}(\Phi_1\Sigma_1^2\Phi_1^\top\Phi_2\Sigma_2^6\Phi_2^\top)} \\
&\geq \sqrt{\|\Sigma_2^4\|_{\mathrm{F}}^2 + \|\Sigma_1^2\Sigma_2^2\|_{\mathrm{F}}^2 - 2\mathrm{Tr}(\Sigma_1^2\Sigma_2^6)} \geq \frac{3}{4}\sqrt{\|\Sigma_2^4\|_{\mathrm{F}}^2} = \frac{3}{4}\|XX^\top XX^\top\|_{\mathrm{F}},
\end{aligned} \tag{7.180}$$

where the last inequality is from the definition of $\mathcal{R}''$ that $\|\Sigma_2^2\|_{\mathrm{F}}^2 \geq 16\|\Sigma_1^2\|_{\mathrm{F}}^2$ and the minimum is achieved when $(\Sigma_1)_{ii} = \frac{1}{2}(\Sigma_2)_{ii}$ for all $i \in [r]$. Further, we have

$$\|\nabla\mathcal{F}(X)X^\top\|_{\mathrm{F}} \leq \sigma_1(X)\|\nabla\mathcal{F}(X)\|_{\mathrm{F}} \;\text{ and }\; \|XX^\top XX^\top\|_{\mathrm{F}} \geq \sigma_1^4(X). \tag{7.181}$$

Combining (7.180) and (7.181), we have the desired result.

7.5.4 Proof of Theorem 19

The proofs are based on the analysis of the general rank $r \geq 1$ case in Theorem 18, combined with the concentration properties of sub-Gaussian matrices $\{A_i\}_{i=1}^d$.

First, we identify the stationary points of $F(X)$ in the following lemma. The proof is provided in Appendix 7.5.4.

Lemma 64. For any $X \in \mathcal{U} \cup \{0\}$, X is a stationary point of $F(X)$.

Next, we characterize two types of stationary points. We show in the following lemma that $X = 0$ is the only the strict saddle point, and the Hessian matrix has negative eigenvalues in the neighborhood of $\mathcal{X}$ with high probability if d is large enough. The proof is provided in Appendix 7.5.4

Lemma 65. For any $X \in \mathcal{R}_1$, if $\max\left\{\|XX^\top - UU^\top\|_{\mathrm{F}}^2,\ \|X\|_{\mathrm{F}}^2,\ 1\right\} \leq N_1$ holds for some constant N_1 and the number of linear measurements d satisfies $d = \Omega\left(N_1 nr/\sigma_r^2(U)\right)$, then with probability at least $1 - \exp\left(-C_1 nr\right)$ for some constant C_1, $\nabla^2 F(X)$ contains a negative eigenvalue, i.e.

$$\lambda_{\min}(\nabla^2 F(X)) \leq -\frac{\sigma_r^2(U)}{8}.$$

Moreover, $X = 0$ is a strict saddle point with $\lambda_{\min}(F(0)) \leq -\frac{7}{8}\|U\|_2^2$.

Moreover, we show in the following lemma that any $X \in \mathcal{U}$ is a global minimum, and $F(X)$ is only strongly convex along certain directions in the neighborhood of $X \in \mathcal{U}$ with high probability if d is large enough. The proof is provided in Appendix 7.5.4.

Lemma 66. For any $X \in \mathcal{U}$, X is a global minimum, and $\nabla^2 F(X)$ is positive semidefinite. Moreover, for any $X \in \mathcal{R}_2$, if $\max\left\{\|XX^\top - UU^\top\|_{\mathrm{F}}^2,\ 4\|U\|_{\mathrm{F}}^2,\ 1\right\} \leq N_2$ holds for some constant N_2 and d satisfies $d = \Omega\left(N_2 nr/\sigma_r^2(U)\right)$, then with probability at least $1 - \exp\left(-C_2 nr\right)$ for some constant C_2, we have

$$z^\top \nabla^2 F(X) z \geq \frac{1}{10}\sigma_r^2(U)\|z\|_2^2$$

for any $z \perp \mathcal{E}$, where $\mathcal{E} \subseteq \mathbb{R}^{n\times r}$ is a subspace is spanned by all eigenvectors of $\nabla^2 \mathcal{F}(K_E)$ associated with the negative eigenvalues, where $E = X - U\Psi_X$ and Ψ_X and K_E are defined in (6.10).

Finally, we show in the following lemma that the gradient $\nabla F(X)$ is sufficiently large norm outside the neighborhood of $\mathcal{X}$ with high probability if d is large enough. The proof is provided in Appendix 7.5.4.

Lemma 67. For any $X \in \mathcal{R}_3'$, if $\max\left\{\|XX^\top - UU^\top\|_{\mathrm{F}}^2,\ \max_k \|X_{(*,k)}\|_{\mathrm{F}}^2\right\} \leq N_3$ holds for some constant N_3 and d satisfies $d = \Omega\left(N_3\sqrt{nr}\log(nr)\sigma_1(U)/\sigma_r^4(U)\right)$, then with probability at least $1 - (C_3 nr)^{-1}$ for some constant C_3, we have

$$\|\nabla F(X)\|_{\mathrm{F}} > \frac{\sigma_r^4(U)}{18\sigma_1(U)}.$$

Moreover, for any $X \in \mathcal{R}_3''$, if $d = \Omega\left(n\sqrt{r}\log(n)\right)$, then with probability at least $1 - (C_4 n)^{-2}$ for some constant C_4, we have

$$\|\nabla F(X)\|_{\mathrm{F}} > \frac{1}{4}\sigma_1^3(X).$$

For $X \in \mathcal{R}_1$, $N_1 \leq \left(\|XX^\top\|_{\mathrm{F}} + \|UU^\top\|_{\mathrm{F}}\right)^2 \leq 25\|UU^\top\|_{\mathrm{F}}^2$. Similarly, we have $N_2 \leq 25\|UU^\top\|_{\mathrm{F}}^2$ and $N_3 \leq 25\|UU^\top\|_{\mathrm{F}}^2$. Then combining $\|UU^\top\|_{\mathrm{F}}^2 \leq r\sigma_1^4(U)$ and Lemma 64 – Lemma 67, if d satisfies

$$d = \Omega\left(\max\left\{\frac{\sigma_1^4(U)nr^2}{\sigma_r^2(U)}, \frac{\sigma_1^5(U)r\sqrt{nr}\log(nr)}{\sigma_r^4(U)}, n\sqrt{r}\log(n)\right\}\right),$$

with probability at least $1 - 2\exp\left(-C_5 nr\right) - \left(C_3 nr\right)^{-1} - \left(C_4 n\right)^{-2}$, we have the desired results.

Proof of Lemma 64

Recall that the gradient $F(X)$ is

$$\nabla F(X) = \frac{1}{2d}\sum_{i=1}^{d}\langle A_i, XX^\top - UU^\top\rangle \cdot (A_i + A_i^\top)X.$$

It is easy to see that $X \in \mathcal{U} \cup \{0\}$ is a stationary point of $F(X)$. Note that due to the perturbation of the linear mapping $\mathcal{A}$, $X \in \mathcal{X}\backslash\mathcal{U}$ is not a strict saddle point.

Proof of Lemma 65

We only need to verify

$$\left|\lambda_{\min}(\nabla^2 F(X)) - \lambda_{\min}(\nabla^2 \mathcal{F}(X))\right| \leq \|\nabla^2 F(X) - \nabla^2 \mathcal{F}(X)\|_2 \leq \frac{\sigma_r^2(U)}{8},$$

where the first inequality is from Weyl's inequality and the second inequality holds with high probability at least $1 - \exp\left(-cnr\right)$ if $d = \Omega(N_1 nr/\sigma_r^2(U))$ by taking $\delta = \sigma_r^2(U)/8$ in Lemma 5. Similarly, we have $\lambda_{\min}(\nabla^2 F(0)) \leq -\frac{7}{8}\|U\|_2^2$ with high probability, which finishes the proof.

Proof of Lemma 66

First of all, it is easy to see that for any $X \in \mathcal{U}$, $F(X) = 0$ attains the minimal objective value of F, thus X is a global minimum. From (6.18), we have $\nabla^2 F(U) = \mathrm{vec}((A_i + A_i^\top)U) \cdot \mathrm{vec}((A_i + A_i^\top)U)^\top$, which is positive semidefinite.

The rest of the analysis is analogous to the proof of Lemma 65, where we only need to verify

$$\left|\lambda_{\max}(\nabla^2 F(X)) - \lambda_{\max}(\nabla^2 \mathcal{F}(X))\right| \leq \|\nabla^2 F(X) - \nabla^2 \mathcal{F}(X)\|_2 \leq \frac{\sigma_r^2(U)}{10}.$$

Now we only need to verify the bound of N_2. Let $\tilde{\Psi} = \arg\min_{\Psi \in \mathfrak{O}_r} \|X - U\Psi\|_2$ and $\tilde{U} = U\tilde{\Psi}$, then $\|\tilde{U}\|_{\mathrm{F}} = \|U\|_{\mathrm{F}}$ and $\sigma_i(\tilde{U}) = \sigma_i(U)$ for all $i = 1\ldots, r$. From $\min_{\Psi \in \mathfrak{O}_r} \|X -$

$U\Psi\|_2 \leq \frac{\sigma_r^2(U)}{8\sigma_1(U)}$, we have

$$\|X - \tilde{U}\|_{\mathrm{F}} \leq \sqrt{r}\|X - \tilde{U}\|_2 \leq \sqrt{r}\sigma_r(U) \leq \sqrt{\sum_{i=1}^{r} \sigma_i^2(U)} = \|U\|_{\mathrm{F}}.$$

This implies

$$\|X\|_{\mathrm{F}} \leq \|X - \tilde{U}\|_{\mathrm{F}} + \|U\|_{\mathrm{F}} \leq 2\|U\|_{\mathrm{F}}.$$

Following the analysis of Lemma 65, we finish the proof.

Proof of Lemma 67

Part 1. We first discuss $X \in \mathcal{R}_3'$. By taking $\delta = \frac{\sigma_r^4(U)}{18\sigma_1(U)}$ in the analysis of Lemma 5, we have that if $d = \Omega\left(N_3\sqrt{nr}\log(nr)\sigma_1(U)/\sigma_r^4(U)\right)$, then with probability at least $1 - (c_2nr)^{-1}$,

$$\|\nabla F(X)\|_{\mathrm{F}} \geq \|\nabla \mathcal{F}(X)\|_{\mathrm{F}} - \|\nabla F(X) - \nabla \mathcal{F}(X)\|_{\mathrm{F}} \geq \frac{\sigma_r^4(U)}{18\sigma_1(U)}.$$

Part 2. Next, we discuss $X \in \mathcal{R}_3''$. Remind that from (7.180) we have

$$\|\nabla \mathcal{F}(X)X^\top\|_{\mathrm{F}} \geq \frac{3}{4}\|XX^\top XX^\top\|_{\mathrm{F}}. \tag{7.182}$$

Moreover, we have

$$\nabla F(X)X^\top = \frac{1}{d}\sum_{i=1}^{d} \underbrace{\langle A_i, XX^\top - UU^\top\rangle \cdot (A_i + A_i^\top)XX^\top/2}_{\hat{\mathrm{I}}\mathrm{I}}.$$

Ignore the index i for $\hat{\mathrm{I}}\mathrm{I}$ for convenience. Consider the (j,k)-th entry of $\hat{\mathrm{I}}\mathrm{I}$, i.e. $\langle A, XX^\top - UU^\top\rangle \cdot (A_{(j,*)} + A_{(*,j)}^\top)XX_{(*,k)}^\top/2$. Analogous to the analysis in Part 1, since A has i.i.d. zero mean sub-Gaussian entries with variance 1, we have $\langle A, XX^\top - UU^\top\rangle$ and $(A_{(j,*)} + A_{(*,j)}^\top)XX_{(*,k)}^\top$ are also zero mean sub-Gaussian entries with variance bounded by $\|XX^\top - UU^\top\|_{\mathrm{F}}^2$ and $\|XX_{(*,k)}^\top\|_{\mathrm{F}}^2$ respectively.

It is easy to check $\mathbb{E}(\nabla F(X)X^\top) = \nabla \mathcal{F}(X)X^\top$. By Lemma 77, we have $\hat{\text{II}}$ is sub-exponential with variance proxy upper bounded by $N_4 = \max\left\{\|XX^\top - UU^\top\|_{\mathrm{F}}^2,\ \|XX_{(*,k)}^\top\|_{\mathrm{F}}^2\right\}$. Then by the concentration of sub-exponential random variables,

$$\mathbb{P}\left(|(\nabla F(X)X^\top)_{(j,k)} - (\nabla \mathcal{F}(X)X^\top)_{(j,k)}| > t\right) \leq \exp\left(-\frac{c_1 dt}{N_4}\right).$$

This implies

$$\begin{aligned}\mathbb{P}\left(\|\nabla F(X)X^\top - \nabla \mathcal{F}(X)X^\top\|_{\mathrm{F}} > t\right) &\leq n^2 \exp\left(-\frac{c_3 dt}{N_4 n}\right) \\ &= \exp\left(-\frac{c_3 dt}{N_4 n} + 2\log n\right). \end{aligned} \tag{7.183}$$

On the other hand, we have

$$\|XX_{(*,k)}^\top\|_{\mathrm{F}} \leq \|XX^\top\|_{\mathrm{F}},$$

and

$$\|XX^\top - UU^\top\|_{\mathrm{F}} \leq \|XX^\top\|_{\mathrm{F}} + \|UU^\top\|_{\mathrm{F}} \leq 2\|XX^\top\|_{\mathrm{F}},$$

which implies

$$N_4 = \max\left\{\|XX^\top - UU^\top\|_{\mathrm{F}}^2,\ \|XX_{(*,k)}^\top\|_{\mathrm{F}}^2\right\} \leq 4\|XX^\top\|_{\mathrm{F}}^2. \tag{7.184}$$

Let $X = \Psi_X \Sigma_X \Theta_X$ be the SVD of X, then

$$\|XX^\top\|_{\mathrm{F}}^2 = \|\Sigma_X^2\|_{\mathrm{F}}^2 = \sum_{i=1}^r \sigma_i^4(X) \leq \sqrt{r\sum_{i=1}^r \sigma_i^8(X)} = \sqrt{r}\|XX^\top XX^\top\|_{\mathrm{F}}. \tag{7.185}$$

Combining (7.183), (7.184), and (7.185), then if $t = \frac{1}{2}\|XX^\top XX^\top\|_{\mathrm{F}}$ and $d = \Omega\left(n\sqrt{r}\log(n)\right)$, with probability at least $1 - (c_4 n)^{-2}$, we have

$$\|\nabla F(X)X^\top\|_{\mathrm{F}} \geq \|\nabla \mathcal{F}(X)X^\top\|_{\mathrm{F}} - \|\nabla F(X)X^\top - \nabla \mathcal{F}(X)X^\top\|_{\mathrm{F}} \geq \frac{1}{4}\|XX^\top XX^\top\|_{\mathrm{F}}.$$

Combining with

$$\|\nabla F(X)X^\top\|_{\mathrm{F}} \leq \sigma_1(X)\|\nabla F(X)\|_{\mathrm{F}} \;\text{ and }\; \|XX^\top XX^\top\|_{\mathrm{F}} \geq \sigma_1^4(X),$$

we have the desired result.

7.5.5 Proof of Corollary 3

For completeness of the analysis, we provide the intermediate results for Corollary 3 as in the analysis for Theorem 19. Recall that for the noisy scenario, we observe

$$y_{(i)} = \langle A_i, M^* \rangle + z_{(i)},$$

where $\{z_{(i)}\}_{i=1}^d$ are independent zero mean sub-Gaussian random noise with variance σ_z^2. Denoting $M^* = UU^\top$, we have the corresponding objective, gradient, and Hessian matrix as

$$F(X) = \frac{1}{4d}\sum_{i=1}^d \Big(\langle A_i, XX^\top - UU^\top \rangle - z_{(i)}\Big)^2, \tag{7.186}$$

$$\nabla F(X) = \frac{1}{2d}\sum_{i=1}^d \Big(\langle A_i, XX^\top - UU^\top \rangle - z_{(i)}\Big) \cdot (A_i + A_i^\top)X, \text{ and} \tag{7.187}$$

$$\begin{aligned}\nabla^2 F(X) = \frac{1}{2d}\sum_{i=1}^d I_r \otimes \Big(\langle A_i, XX^\top - UU^\top \rangle - z_{(i)}\Big) \cdot (A_i + A_i^\top) \\ + \operatorname{vec}\Big((A_i + A_i^\top)X\Big) \cdot \operatorname{vec}\Big((A_i + A_i^\top)X\Big)^\top.\end{aligned} \tag{7.188}$$

We first show the connection between the noisy model and low-rank matrix factorization in the following lemma.

Lemma 68. We have $\mathbb{E}(F(X)) = \mathcal{F}(X) + \frac{\sigma_z^2}{4}$, $\mathbb{E}(\nabla F(X)) = \nabla \mathcal{F}(X)$, and $\mathbb{E}(\nabla^2 F(X)) = \nabla^2 \mathcal{F}(X)$.

We have from Lemma 68 that the objective $F(X)$ for noisy model (7.186) differs from the unbiased estimator of the objective $\mathcal{F}(X)$ for low-rank matrix factorization (6.9) only by a quantity depending on σ_z. Moreover, the gradient (7.187) and the Hessian matrix (7.188) of the noisy model are unbiased estimators of the counterparts of the low-rank matrix factorization problem in (6.11) and (6.12) respectively. These further allow us to derive the lemmas below directly from the counterparts of the low-rank matrix factorization problem in Theorem 18, using the concentrations of sub-Gaussian quantities $\{A_i\}_{i=1}^d$ and $\{z_{(i)}\}_{i=1}^d$. The proofs of the lemmas below are analogous to those of Lemma 64 – Lemma 67, thus we omit them here.

First, we identify the stationary points of $F(X)$ in the following lemma.

Lemma 69. For any $X \in \mathcal{U} \cup \{0\}$, X is a stationary point of $F(X)$.

Next, we show in the following lemma that $X = 0$ is the only the strict saddle point, and the Hessian matrix has negative eigenvalues in the neighborhood of $\mathcal{X}$ with high probability if d is large enough.

Lemma 70. For any $X \in \mathcal{R}_1$, if $\max\left\{\|XX^\top - UU^\top\|_\mathrm{F}^2 + \sigma_z^2,\ \|X\|_\mathrm{F}^2,\ 1\right\} \leq N_1$ holds for some constant N_1 and the number of linear measurements d satisfies $d = \Omega\left(N_1 nr/\sigma_r^2(U)\right)$, then with probability at least $1 - \exp\left(-C_1 nr\right)$ for some constant C_1, $\nabla^2 F(X)$ contains a negative eigenvalue, i.e.

$$\lambda_{\min}(\nabla^2 F(X)) \leq -\frac{\sigma_r^2(U)}{8}.$$

Moreover, $X = 0$ is a strict saddle point with $\lambda_{\min}(F(0)) \leq -\frac{7}{8}\|U\|_2^2$.

Moreover, we show in the following lemma that any $X \in \mathcal{U}$ is a global minimum, and $F(X)$ is only strongly convex along certain directions in the neighborhood of $X \in \mathcal{U}$ with high probability if d is large enough.

Lemma 71. For any $X \in \mathcal{U}$, X is a global minimum, and $\nabla^2 F(X)$ is positive semidefinite. Moreover, for any $X \in \mathcal{R}_2$, if $\max\left\{\|XX^\top - UU^\top\|_\mathrm{F}^2 + \sigma_z^2,\ \|U\|_\mathrm{F}^2,\ 1\right\} \leq N_2$ holds for some constant N_2 and d satisfies $d = \Omega\left(N_2 nr/\sigma_r^2(U)\right)$, then with probability at least $1 - \exp\left(-C_2 nr\right)$ for some constant C_2, we have

$$z^\top \nabla^2 F(X) z \geq \frac{1}{10}\sigma_r^2(U)\|z\|_2^2$$

for any $z \perp \mathcal{E}$, where $\mathcal{E} \subseteq \mathbb{R}^{n\times r}$ is a subspace is spanned by all eigenvectors of $\nabla^2 \mathcal{F}(K_E)$ associated with the negative eigenvalues, where $E = X - U\Psi_X$.

Finally, we show in the following lemma that the gradient $\nabla F(X)$ is sufficiently large norm outside the neighborhood of $\mathcal{X}$ with high probability if d is large enough.

Lemma 72. For any $X \in \mathcal{R}_3'$, if

$$\max\left\{\|XX^\top - UU^\top\|_\mathrm{F}^2 + \sigma_z^2,\ \max_k \|X_{(*,k)}\|_\mathrm{F}^2,\ \sigma_1(U)/\sigma_r^2(U)\right\} \leq N_3$$

holds for some constant N_3 and d satisfies $d = \Omega\left(N_3\sqrt{nr}\log(nr)\sigma_1(U)/\sigma_r^4(U)\right)$, then with probability at least $1-(C_3nr)^{-1}$ for some constant C_3, we have

$$\|\nabla F(X)\|_\mathrm{F} > \frac{\sigma_r^4(U)}{18\sigma_1(U)}.$$

Moreover, for any $X \in \mathcal{R}_3''$, if $d = \Omega\left(n\sqrt{r}\log(n)\right)$, then with probability at least $1-(C_4n)^{-2}$ for some constant C_4, we have

$$\|\nabla F(X)\|_\mathrm{F} > \frac{1}{4}\sigma_1^3(X).$$

In terms of the estimation error, the result follows directly from combining [192] (Lemma 5.3) and [188] (Corollary 2) for the sub-Gaussian case. Note that $\widehat{X}$ is the optimal solution here. Note that the statistical rate here is consistent with the result for general noisy setting [246].

7.5.6 Proof of Theorem 20

First, (p1) follows directly from Lemma 64. It is also immediate that for any $X \in \mathcal{U} \cup \{0\}$, we have $\nabla F(X) = 0$, which implies X is a stationary point of $F(X)$. Moreover, for any $X \in \mathcal{U}$, we have $F(X) = 0$, which implies X is a global minimum.

Then, we have from [247,248] that when A_i has i.i.d. zero mean sub-Gaussian entries with variance 1 and $d \geq cnr$, then with high probability, we have that for any matrices M_1, M_2 of rank at most $6r$,

$$\left|\frac{1}{d}\sum_{i=1}^{d}\langle A_i, M_1\rangle^2 - \|M_1\|_\mathrm{F}^2\right| \leq \rho_1 \|M_1\|_\mathrm{F}^2. \tag{7.189}$$

Note that $\Psi_X = \mathrm{argmin}_{\Psi\in\mathfrak{O}_r}\|X - U\Psi_X\|_\mathrm{F}^2 = \mathrm{argmin}_{\Psi\in\mathfrak{O}_r}\|X - U\Psi_X\|_2^2 = AB^\top$, where the SVD of $U^\top X = A\Sigma B^\top$.

Then we demonstrate (p2). Here we state an intermediate result to be used later.

Lemma 73 (Lemma 6 in [219])**.** Given $X, U \in \mathbb{R}^{n\times r}$, and $E = X - U\Psi_X$, where Ψ_X is defined in (6.10), we have $\left\|EE^\top\right\|_{\mathrm{F}}^2 \leq 2\left\|XX^\top - UU^\top\right\|_{\mathrm{F}}^2$ and $\|E\|_{\mathrm{F}}^2 \leq \frac{1}{2(\sqrt{2}-1)\sigma_r^2(U)}\left\|XX^\top - UU^\top\right\|_{\mathrm{F}}^2$.

Let $z = [E_{(*,1)}^\top, \ldots, E_{(*,r)}^\top,] \in \mathbb{R}^{nr}$, $E = X - U\Psi_X$, and Ψ_X is defined in (6.10), then

$$\begin{aligned}
&z^\top \nabla^2 F(X) z \\
&= z^\top \Bigg(\frac{1}{2d}\sum_{i=1}^{d} I_r \otimes \langle A_i, XX^\top - UU^\top\rangle \cdot (A_i + A_i^\top) \\
&\qquad\quad + \mathrm{vec}((A_i + A_i^\top)X) \cdot \mathrm{vec}((A_i + A_i^\top)X)^\top \Bigg) z \\
&= \frac{1}{d}\sum_{i=1}^{d} \left(\left\langle A_i, EE^\top \right\rangle^2 - 3\left\langle A_i, XX^\top - UU^\top \right\rangle^2 \right) \\
&\qquad\quad + \frac{2}{d}\sum_{i=1}^{d} \langle A_i, XX^\top - UU^\top\rangle \cdot \left\langle (A_i + A_i^\top)X, E \right\rangle \qquad (7.190) \\
&\overset{(i)}{\leq} (1+\rho_1)\left\|EE^\top\right\|_{\mathrm{F}}^2 - 3(1-\rho_1)\left\|XX^\top - UU^\top\right\|_{\mathrm{F}}^2 + 4\|\nabla F(X)\|_* \|E\|_2 \\
&\overset{(ii)}{\leq} -\frac{1}{3}\sigma_r^2(U)\|E\|_2^2 + 4\|\nabla F(X)\|_* \|E\|_2
\end{aligned}$$

where (i) is from (7.189) and Fenchel's duality theorem, and (ii) is from Lemma 73 by taking $\rho_1 \leq \frac{1}{10}$ and $\|E\|_2 \leq \|E\|_{\mathrm{F}}$.

On the other hand, we have from (7.190) and Lemma 73 by taking $\rho_1 \leq \frac{1}{10}$ that

$$\begin{aligned}
z^\top \nabla^2 F(X) z &\leq (1+\rho_1)\left\|EE^\top\right\|_{\mathrm{F}}^2 - 3(1-\rho_1)\left\|XX^\top - UU^\top\right\|_{\mathrm{F}}^2 + 4\|\nabla F(X)\|_{\mathrm{F}}\|E\|_{\mathrm{F}} \\
&\leq -\frac{1}{3}\sigma_r^2(U)\|E\|_{\mathrm{F}}^2 + 4\|\nabla F(X)\|_{\mathrm{F}}\|E\|_{\mathrm{F}}. \qquad (7.191)
\end{aligned}$$

For $\|\nabla F(X)\|_{\mathrm{F}} \leq \frac{\sigma_r^3(U)}{96}$ and $\|E\|_{\mathrm{F}} \geq \frac{\sigma_r(U)}{4}$, we have from (7.191) that

$$z^\top \nabla^2 F(X) z \leq -\frac{1}{6}\sigma_r^2(U)\|E\|_{\mathrm{F}}^2,$$

which implies $\lambda_{\min}(\nabla^2 F(X)) \le -\frac{1}{6}\sigma_r^2(U)$. Since we have

$$\left\{X \mid \|E\|_2 \ge \frac{\sigma_r(U)}{4}\right\} \subseteq \left\{X \mid \|E\|_{\mathrm{F}} \ge \frac{\sigma_r(U)}{4}\right\},$$

then it follows that $\lambda_{\min}(\nabla^2 F(X)) \le -\frac{1}{6}\sigma_r^2(U)$ also holds in $\mathcal{R}_1$.

To demonstrate (p3), we have the following intermediate results from [192].

Lemma 74 (Lemma 5.7 in [192])**.** Given $X, U \in \mathbb{R}^{n\times r}$, and $E = X - U\Psi_X$, where Ψ_X is defined in (6.10), with $\|E\|_{\mathrm{F}} \le \frac{\sigma_r(U)}{4}$, then with high probability, we have

$$\langle \nabla \mathcal{F}(X), E\rangle - \frac{1}{20}\left(\left\|XX^\top - UU^\top\right\|_{\mathrm{F}}^2 + \left\|EX^\top\right\|_{\mathrm{F}}^2\right) \ge \frac{\sigma_r^2(U)}{4}\|E\|_{\mathrm{F}}^2 + \frac{1}{5}\left\|XX^\top - UU^\top\right\|_{\mathrm{F}}^2.$$

Lemma 75 (Lemma 5.8 in [192])**.** Given $X, U \in \mathbb{R}^{n\times r}$, $E = X - U\Psi_X$, where Ψ_X is defined in (6.10), with $\|E\|_{\mathrm{F}} \le \frac{\|U\|_2}{4}$, and any $V \in \mathbb{R}^{n\times r}$, then with high probability, we have

$$|\langle \nabla \mathcal{F}(X) - \nabla F(X), V\rangle| \le \frac{1}{10}\left\|XX^\top - UU^\top\right\|_{\mathrm{F}}\left\|VX^\top\right\|_{\mathrm{F}}.$$

Lemma 76 (Lemma 5.9 in [192])**.** Given any $X \in \mathbb{R}^{n\times r}$, with high probability, we have

$$\left\|XX^\top - UU^\top\right\|_{\mathrm{F}}^2 \ge \frac{1}{2\|X\|_2}\|\nabla F(X)\|_{\mathrm{F}}^2.$$

Then we have

$$\begin{aligned}\langle \nabla \mathcal{F}(X), E\rangle &= \langle \nabla F(X), E\rangle + \langle \nabla \mathcal{F}(X) - \nabla F(X), E\rangle\\ &\overset{(i)}{\le} \langle \nabla F(X), E\rangle + \frac{1}{10}\left\|XX^\top - UU^\top\right\|_{\mathrm{F}}\left\|EX^\top\right\|_{\mathrm{F}}\\ &\overset{(ii)}{\le} \langle \nabla F(X), E\rangle + \frac{1}{20}\left(\left\|XX^\top - UU^\top\right\|_{\mathrm{F}}^1 + \left\|EX^\top\right\|_{\mathrm{F}}^2\right) \qquad (7.192)\end{aligned}$$

where (i) is from Lemma 75 and (ii) is from the inequality of arithmetic and geometric means. Then we have

$$\begin{aligned}\langle \nabla F(X), E\rangle &\overset{(i)}{\geq} \frac{\sigma_r^2(U)}{4}\|E\|_{\mathrm{F}}^2 + \frac{1}{5}\left\|XX^\top - UU^\top\right\|_{\mathrm{F}}^2 \\ &\overset{(ii)}{\geq} \frac{\sigma_r^2(U)}{4}\|E\|_{\mathrm{F}}^2 + \frac{1}{10\|X\|_2}\|\nabla F(X)\|_{\mathrm{F}}^2 \\ &\overset{(iii)}{\geq} \frac{\sigma_r^2(U)}{4}\|E\|_{\mathrm{F}}^2 + \frac{1}{20\|U\|_2}\|\nabla F(X)\|_{\mathrm{F}}^2,\end{aligned}$$

where (i) is from Lemma 74 and (7.192), (ii) is from Lemma 76, and (iii) is from $\|X\|_2 \leq \frac{5}{4}\|U\|_2$ given $\|E\|_2 \leq \frac{1}{4}\|U\|_2$.

7.5.7 Further Intermediate Results

Proof of Proposition 3

Consider the following regions:

$$\begin{aligned}\widetilde{\mathcal{R}}_1 &\triangleq \left\{Y \in \mathbb{R}^{n\times r} \mid \sigma_r(Y) \leq \frac{1}{2}\sigma_r(U)\right\}, \\ \widetilde{\mathcal{R}}_2 &\triangleq \left\{Y \in \mathbb{R}^{n\times r} \mid \min_{\Psi\in\mathfrak{O}_r}\|Y - U\Psi\|_2 \leq \frac{\sigma_r^2(U)}{8\sigma_1(U)}\right\}, \text{ and} \\ \widetilde{\mathcal{R}}_3 &\triangleq \left\{Y \in \mathbb{R}^{n\times r} \mid \sigma_r(Y) > \frac{1}{2}\sigma_r(U),\ \min_{\Psi\in\mathfrak{O}_r}\|Y - U\Psi\|_2 > \frac{\sigma_r^2(U)}{8\sigma_1(U)}\right\}.\end{aligned}$$

Then it is obvious to see that $\widetilde{\mathcal{R}}_1 \cup \widetilde{\mathcal{R}}_2 \cup \widetilde{\mathcal{R}}_3 = \mathbb{R}^{n\times r}$. Moreover, we immediately have $\mathcal{R}_1 = \widetilde{\mathcal{R}}_1 \cap \mathcal{R}_3''^{\perp}$ and $\mathcal{R}_3' = \widetilde{\mathcal{R}}_3 \cap \mathcal{R}_3''^{\perp}$. Since for $X \in \mathcal{R}_2$, we have for any $i \in [r]$,

$$|\sigma_i(X) - \sigma_i(U)| \leq \frac{\sigma_r(U)}{8},$$

$\|XX^\top\|_{\mathrm{F}} \leq 2\|UU^\top\|_{\mathrm{F}}$ always holds, i.e., $\mathcal{R}_2 \subseteq \mathcal{R}_3''^{\perp}$, thus $\mathcal{R}_2 = \widetilde{\mathcal{R}}_2 \cap \mathcal{R}_3''^{\perp}$ also holds. Then we have

$$\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3' = \left(\widetilde{\mathcal{R}}_1 \cup \widetilde{\mathcal{R}}_2 \cup \widetilde{\mathcal{R}}_3\right) \cap \mathcal{R}_3''^{\perp} = \mathcal{R}_3''^{\perp}.$$

Proof of Proposition 4

For any $\alpha \in (0,1)$ and $\Psi \in \mathfrak{O}_r$, $\Psi \neq I_r$, we have

$$\begin{aligned}
\mathcal{F}(\alpha U + (1-\alpha)U\Psi) &= \frac{1}{4}\|(\alpha U + (1-\alpha)U\Psi)(\alpha U + (1-\alpha)U\Psi)^\top - UU^\top\|_{\mathrm{F}}^2 \\
&= \frac{\alpha^2(1-\alpha)^2}{4}\|U(\Psi + \Psi^\top - 2I_r)U^\top\|_{\mathrm{F}}^2 \\
&> 0 = \alpha\mathcal{F}(U) + (1-\alpha)\mathcal{F}(U\Psi).
\end{aligned}$$

Proof of Lemma 4

We first demonstrate the objective function. By the definition of $F(X)$, we have

$$\begin{aligned}
\mathcal{F}(X) = \mathbb{E}(F(X)) &= \mathbb{E}\left(\frac{1}{4d}\sum_{i=1}^{d}(y_{(i)} - \langle A_i, XX^\top\rangle)^2\right) \\
&= \frac{1}{4d}\sum_{i=1}^{d}\mathbb{E}\left(\langle A_i, UU^\top\rangle - \langle A_i, XX^\top\rangle\right)^2 \\
&= \frac{1}{4d}\sum_{i=1}^{d}\mathbb{E}\langle A_i, UU^\top - XX^\top\rangle^2 = \frac{1}{4d}\sum_{i=1}^{d}\mathbb{E}\left(\mathrm{vec}(A_i)^\top\mathrm{vec}(UU^\top - XX^\top)\right)^2 \\
&= \frac{1}{4d}\sum_{i=1}^{d}\mathbb{E}\left(\mathrm{vec}(UU^\top - XX^\top)^\top\mathrm{vec}(A_i)\mathrm{vec}(A_i)^\top\mathrm{vec}(UU^\top - XX^\top)\right) \\
&= \frac{1}{4}\mathrm{vec}(UU^\top - XX^\top)^\top\cdot\frac{1}{d}\sum_{i=1}^{d}\mathbb{E}(\mathrm{vec}(A_i)\mathrm{vec}(A_i)^\top)\cdot\mathrm{vec}(UU^\top - XX^\top) \\
&= \frac{1}{4}\|\mathrm{vec}(UU^\top - XX^\top)\|_2^2 = \frac{1}{4}\|UU^\top - XX^\top\|_{\mathrm{F}}^2,
\end{aligned}$$

Next, we demonstrate the gradient and the Hessian matrix. From the independence of A_i's, we have

$$\begin{aligned}
\mathbb{E}(\nabla F(X)) &= \frac{1}{2}\mathbb{E}\left(\langle A_i, XX^\top - UU^\top\rangle\cdot(A_i + A_i^\top)X\right) \\
\mathbb{E}(\nabla^2 F(X)) &= \frac{1}{2}\mathbb{E}\Bigg(I_r\otimes\langle A_i, XX^\top - UU^\top\rangle\cdot(A_i + A_i^\top) \\
&\qquad\qquad + \mathrm{vec}\left((A_i + A_i^\top)X\right)\cdot\mathrm{vec}\left((A_i + A_i^\top)X\right)^\top\Bigg).
\end{aligned}$$

We ignore the index i and denote A_i as A for the convenience of notation. The proof is analyzed by entry-wise agreement.

For the (j,k)-th entry of gradient $\nabla F(X)$, we have

$$\begin{aligned}
&\mathbb{E}(\nabla F(X)_{(j,k)}) = \frac{1}{2}\mathbb{E}\left(\langle A, XX^\top - UU^\top\rangle \cdot (A + A^\top)_{(j,*)}X_{(*,k)}\right)\\
&= \frac{1}{2}\mathbb{E}\left(\sum_{s,t} A_{(s,t)}(XX^\top - UU^\top)_{(s,t)} \cdot \sum_{l}(A_{(j,l)} + A_{(l,j)})X_{(l,k)}\right)\\
&\overset{(i)}{=} \frac{1}{2}\mathbb{E}\left(\sum_{l} A^2_{(j,l)}(XX^\top - UU^\top)_{(j,l)}X_{(l,k)} + A^2_{(l,j)}(XX^\top - UU^\top)_{(l,j)}X_{(l,k)}\right)\\
&= \frac{1}{2}\left(\sum_{l} \mathbb{E}(A^2_{(j,l)})(XX^\top - UU^\top)_{(j,l)}X_{(l,k)} + \mathbb{E}(A^2_{(l,j)})(XX^\top - UU^\top)_{(l,j)}X_{(l,k)}\right)\\
&\overset{(ii)}{=} \frac{1}{2}\left(\sum_{l}(XX^\top - UU^\top)_{(j,l)}X_{(l,k)} + (XX^\top - UU^\top)_{(l,j)}X_{(l,k)}\right)\\
&= (XX^\top - UU^\top)X_{(j,k)}, \qquad (7.193)
\end{aligned}$$

where (i) is from the independence and zero mean of entries of A, and (ii) is from $\sigma^2 = 1$.

We use double index for the Hessian matrix, i.e., denote (jk, st) as the $((k-1)n + j, (t-1)n + s)$-th entry of $\nabla^2 F(X)$. We discuss by separating the two components of $\nabla^2 F(X)$. For the first component,

$$\begin{aligned}
&\mathbb{E}\left(\langle A, XX^\top - UU^\top\rangle \cdot (A + A^\top)_{(j,k)}\right)\\
&= \mathbb{E}\left(\sum_{s,t} A_{(s,t)}(XX^\top - UU^\top)_{(s,t)} \cdot (A_{(j,k)} + A_{(k,j)})\right)\\
&= \mathbb{E}\left(A^2_{(j,k)}(XX^\top - UU^\top)_{(j,k)} + A^2_{(k,j)}(XX^\top - UU^\top)_{(k,j)}\right)\\
&= 2(XX^\top - UU^\top)_{(j,k)}.
\end{aligned}$$

Therefore, we have

$$\frac{1}{2}\mathbb{E}\left(I_r \otimes \langle A_i, XX^\top - UU^\top\rangle \cdot (A_i + A_i^\top)\right) = I_r \otimes (XX^\top - UU^\top). \qquad (7.194)$$

For the second component

$$
\begin{aligned}
&\mathbb{E}\left(\operatorname{vec}((A+A^\top)X)\cdot\operatorname{vec}((A+A^\top)X)^\top_{(jk,st)}\right)\\
&=\mathbb{E}\left(\operatorname{vec}((A+A^\top)_{(j,*)}X_{(*,k)})\cdot\operatorname{vec}((A+A^\top)_{(s,*)}X_{(*,t)})^\top\right)\\
&=\mathbb{E}\left(\left(\sum_l A_{(j,l)}X_{(l,k)}+\sum_m A_{(m,j)}X_{(m,k)}\right)\cdot\left(\sum_l A_{(s,l)}X_{(l,t)}+\sum_m A_{(m,s)}X_{(m,t)}\right)\right).
\end{aligned}
$$

Remind that $K_X=\begin{bmatrix}K_{11}&K_{21}&\cdots&K_{r1}\\K_{12}&K_{22}&\cdots&K_{r2}\\\vdots&\vdots&\ddots&\vdots\\K_{1r}&K_{2r}&\cdots&K_{rr}\end{bmatrix}$, where $K_{kt}=X_{(*,k)}X^\top_{(*,t)}$. If $j\neq s$, we have

$$
\begin{aligned}
&\mathbb{E}\left(\operatorname{vec}((A+A^\top)X)\cdot\operatorname{vec}((A+A^\top)X)^\top_{(jk,st)}\right)\\
&=\mathbb{E}(2A^2_{(j,s)}X_{(s,k)}X_{(j,t)})=2X_{(s,k)}X_{(j,t)}=\left(X^\top X\otimes I_n+K_X\right)_{(jk,st)}.
\end{aligned}
\tag{7.195}
$$

If $j=s$, we have

$$
\begin{aligned}
&\mathbb{E}\left(\operatorname{vec}((A+A^\top)X)\cdot\operatorname{vec}((A+A^\top)X)^\top_{(jk,jt)}\right)\\
&=\mathbb{E}\left(\sum_l A^2_{(j,l)}X_{(l,k)}X_{(l,t)}+\sum_m A^2_{(m,j)}X_{(m,k)}X_{(m,t)}+2A^2_{(j,j)}X_{(j,k)}X_{(j,t)}\right)\\
&=2(X^\top_{(*,k)}X_{(*,t)}+X_{(j,k)}X_{(j,t)})=\left(X^\top X\otimes I_n+K_X\right)_{(jk,jt)}.
\end{aligned}
\tag{7.196}
$$

Combining (7.194), (7.195), and (7.196), we have

$$\mathbb{E}(\nabla^2F(X)_{(jk,st)})=\nabla^2\mathcal{F}(X)_{(jk,st)}.$$

Proof of Lemma 5

From Lemma 4, we have that $\mathbb{E}(\nabla F(X))=\nabla\mathcal{F}(X)$ and $\mathbb{E}(\nabla^2F(X))=\nabla^2\mathcal{F}(X)$. We start with an intermediate result to show that the product of two sub-Gaussian random variables is a sub-exponential random variable.

Lemma 77. Suppose X and Y are two zero mean sub-Gaussian random variables with variance proxies σ_1^2 and σ_2^2 respectively. Let $\sigma^2 = \max\{\sigma_1^2, \sigma_2^2\}$, then XY is a sub-exponential random variable with variance proxy σ^2, i.e. there exist some constant c such that for all $t > 0$,

$$\mathbb{P}(|XY - \mathbb{E}(XY)| > t) \leq \exp\left(-ct/\sigma^2\right). \tag{7.197}$$

Proof. By the definition of sub-exponential random variables [237], we have that if Z is a centered sub-exponential random variable, we have

$$\|Z\|_{\psi_1} = \sup_{p\geq 1} \frac{1}{p}(\mathbb{E}|Z|^p)^{1/p} = c_1\sigma_Z^2,$$

where $\|Z\|_{\psi_1}$ is the sub-exponential norm of Z and σ_Z^2 is the proxy of the variance of Z. Using basic inequalities, we have

$$|\|XY\||_{\psi_1} = \sup_{p\geq 1} \frac{1}{p}(\mathbb{E}(XY)^p)^{\frac{1}{p}} \leq \sup_{p\geq 1} p^{-\frac{1}{2}}(\mathbb{E}X^p)^{\frac{1}{p}} p^{-\frac{1}{2}}(\mathbb{E}Y^p)^{\frac{1}{p}} = c_2\sigma_1\sigma_2 \leq c_2\sigma^2.$$

where c_2 is a constant and the last equality holds since X and Y are sub-Gaussian random variables. Thus, XY is a sub-exponential random variable with variance proxy σ^2. Then for general uncentered sub-exponential XY, we have that (7.197) holds for all $t > 0$ for some constant c. □

Part 1: The perturbation result of the Hessian matrix is discussed first. To bound $\|\nabla^2 F(X) - \nabla^2\mathcal{F}(X)\|_2$, we first bound $|z^\top\nabla^2 F(X)z - z^\top\nabla^2\mathcal{F}(X)z|$ for any unit vector $z \in \mathbb{R}^{nr}$, and apply ε-Net argument. Let $z = [z_1^\top, \ldots, z_r^\top,] \in \mathbb{R}^{nr}$ be a unit vector, where $z_i \in \mathbb{R}^n$ for all $i = 1, \dots, r$, then

$$\begin{aligned}
&z^\top \nabla^2 F(X) z \\
&= z^\top \Bigg(\frac{1}{2d} \sum_{i=1}^{d} I_r \otimes \langle A_i, XX^\top - UU^\top \rangle \cdot (A_i + A_i^\top) \\
&\qquad + \operatorname{vec}((A_i + A_i^\top)X) \cdot \operatorname{vec}((A_i + A_i^\top)X)^\top \Bigg) z \\
&= \frac{1}{d} \sum_{i=1}^{d} \underbrace{\frac{1}{2} \sum_{t=1}^{r} z_t^\top (A_i + A_i^\top) z_t \cdot \langle A_i, XX^\top - UU^\top \rangle}_{\hat{\mathrm{I}}_i} \\
&\qquad + \frac{1}{d} \sum_{i=1}^{d} \underbrace{\frac{1}{2} \left(\sum_{t=1}^{r} z_t (A_i + A_i^\top) X_{(*,t)} \right)^2}_{\hat{\mathrm{II}}_i} .
\end{aligned} \tag{7.198}$$

On the other hand,

$$z^\top \nabla^2 \mathcal{F}(X) z = \underbrace{z^\top \left(I_r \otimes (XX^\top - UU^\top) \right) z}_{\mathrm{I}} + \underbrace{z^\top \left(X^\top X \otimes I_n + K_X \right) z}_{\mathrm{II}} .$$

From the analysis of Lemma 4, we have $\mathbb{E}(\hat{\mathrm{I}}_i) = \mathrm{I}$ and $\mathbb{E}(\hat{\mathrm{II}}_i) = \mathrm{II}$.

We ignore the index i of A_i for convenience. To bound $\hat{\mathrm{I}}_i$, we have

$$\begin{aligned}
&\underbrace{\frac{1}{2} \sum_{t=1}^{r} z_t^\top (A + A^\top) z_t \cdot \langle A, XX^\top - UU^\top \rangle}_{\hat{\mathrm{I}}} \\
&= \underbrace{\sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{t=1}^{r} z_{t(j)} z_{t(k)} A_{(j,k)}}_{\mathrm{I\hat{I}I}} \cdot \underbrace{\sum_{j=1}^{n} \sum_{k=1}^{n} \left(XX^\top - UU^\top \right)_{(j,k)} A_{(j,k)}}_{\hat{\mathrm{V}}\mathrm{I}} .
\end{aligned}$$

Since A has i.i.d. zero mean sub-Gaussian entires with variance 1, then $\mathrm{I\hat{I}I}$ is also a zero mean sub-Gaussian with variance upper bounded by 1 since $\|z\|_2 = 1$, and $\hat{\mathrm{V}}\mathrm{I}$ is also a zero mean sub-Gaussian with variance upper bounded by $\|XX^\top - UU^\top\|_{\mathrm{F}}^2$. By Lemma 77, we have each $\hat{\mathrm{I}}_i$ is sub-exponential with proxy $\sigma_1^2 = \max\{1, \|XX^\top - UU^\top\|_{\mathrm{F}}^2\}$.

Then, from the concentration of sum of sub-exponential random variables, there exist some constant c_1 such that

$$\mathbb{P}\left(\left|\frac{1}{d}\sum_{i=1}^{d}\hat{\mathrm{I}}_i - \mathrm{I}\right| > t_1\right) \leq \exp\left(-\frac{c_1 d t_1}{\sigma_1^2}\right). \tag{7.199}$$

On the other hand, $\hat{\mathrm{II}}_i$ is sub-exponential with variance proxy upper bounded by $\sigma_2^2 = \|X\|_{\mathrm{F}}^2$ since $\sum_{t=1}^{r} z_t(A_i + A_i^\top)X_{(*,t)}$ is a zero mean sub-Gaussian, then from the concentration of sum of sub-exponential random variables, there exist some constant c_2 such that

$$\mathbb{P}\left(\left|\frac{1}{d}\sum_{i=1}^{d}\hat{\mathrm{II}}_i - \mathrm{II}\right| > t_2\right) \leq \exp\left(-\frac{c_2 d t_2}{\sigma_2^2}\right). \tag{7.200}$$

Let $t_1 = t_2 = \delta/4$, then combining (7.198), (7.199), and (7.200), for $N_1 \geq \max\left\{\sigma_1^2, \sigma_2^2\right\}$, we have

$$\mathbb{P}\left(\left|z^\top(\nabla^2 F(X) - \nabla^2 \mathcal{F}(X))z\right| > \frac{\delta}{2}\right) \leq \exp\left(-\frac{c_3 d\delta}{\sigma_1^2}\right) + \exp\left(-\frac{c_4 d\delta}{\sigma_2^2}\right) \leq 2\exp\left(-\frac{c_5 d\delta}{N_1}\right), \tag{7.201}$$

Using the ε-Net, we have

$$\begin{aligned}\|\nabla^2 F(X) - \nabla^2 \mathcal{F}(X)\|_2 &= \sup_{z\in\mathbb{R}^{nr}}\left|z^\top(\nabla^2 F(X) - \nabla^2 \mathcal{F}(X))z\right| \\ &\leq (1-2\varepsilon)^{-1}\sup_{z\in\mathcal{N}_\varepsilon}\left|z^\top(\nabla^2 F(X) - \nabla^2 \mathcal{F}(X))z\right|.\end{aligned} \tag{7.202}$$

Combining (7.201) and (7.202), if we take $\varepsilon = 1/4$, then the covering number of a unit sphere of $\mathbb{R}^{nr}$ can be bounded as $|\mathcal{N}_\varepsilon| \leq 10^{nr} \leq \exp(3nr)$, we have

$$\begin{aligned}\mathbb{P}\left(\|\nabla^2 F(X) - \nabla^2 \mathcal{F}(X)\|_2 > \delta\right) &\leq \mathbb{P}\left(\sup_{z\in\mathcal{N}_{1/4}}\left|z^\top(\nabla^2 F(X) - \nabla^2 \mathcal{F}(X))z\right| > \delta\right) \\ &\leq 2|\mathcal{N}_{1/4}|\exp\left(-\frac{c_5 d\delta}{N_1}\right) \leq 2\exp\left(3nr - \frac{c_5 d\delta}{N_1}\right).\end{aligned}$$

If $d = \Omega(N_1 nr/\delta)$, then with probability at least $1 - \exp\left(-c_6 nr\right)$, we have

$$\|\nabla^2 F(X) - \nabla^2 \mathcal{F}(X)\|_2 \leq \delta.$$

Part 2: The perturbation result of the gradient is discussed then. Remind that

$$\nabla F(X) = \frac{1}{d}\sum_{i=1}^{d} \underbrace{\frac{1}{2}\langle A_i, XX^\top - UU^\top\rangle \cdot (A_i + A_i^\top)X}_{\hat{\mathrm{I}}}.$$

Ignore the index i for $\hat{\mathrm{I}}$ for convenience. Consider the (j,k)-th entry of $\hat{\mathrm{I}}$, i.e.,

$$\frac{1}{2}\langle A, XX^\top - UU^\top\rangle \cdot (A_{(j,*)} + A_{(*,j)}^\top)X_{(*,k)}.$$

Analogous to the analysis of Part 1, since A has i.i.d. zero mean sub-Gaussian entries with variance 1, we have $\langle A, XX^\top - UU^\top\rangle$ and $(A_{(j,*)} + A_{(*,j)}^\top)X_{(*,k)}$ are also zero mean sub-Gaussian entries with variance bounded by $\|XX^\top - UU^\top\|_\mathrm{F}^2$ and $\|X_{(*,k)}\|_\mathrm{F}^2$ respectively.

By Lemma 77, we have that $\hat{\mathrm{I}}$ is sub-exponential with variance proxy upper bounded by

$$N_2 \geq \max\left\{\|XX^\top - UU^\top\|_\mathrm{F}^2,\ \|X_{(*,k)}\|_\mathrm{F}^2\right\}.$$

Then by the concentration of sub-exponential random variables,

$$\mathbb{P}\left(|\nabla F(X)_{(j,k)} - \nabla \mathcal{F}(X)_{(j,k)}| > t\right) \leq \exp\left(-\frac{c_1 dt}{N_2}\right).$$

This implies

$$\mathbb{P}\left(\|\nabla F(X) - \nabla \mathcal{F}(X)\|_\mathrm{F} > t\right) \leq nr\exp\left(-\frac{c_1 dt}{N_2\sqrt{nr}}\right) = \exp\left(-\frac{c_1 dt}{N_2\sqrt{nr}} + \log(nr)\right).$$

Let $\delta = t$, then if $d = \Omega\left(N_2\sqrt{nr}\log(nr)/\delta\right)$, with probability at least $1 - (c_2 nr)^{-1}$, we have

$$\|\nabla F(X) - \nabla \mathcal{F}(X)\|_{\mathrm{F}} \leq \delta.$$

Combining Part 1 and Part 2, we have the desired result.

Proof of Lemma 68

We first demonstrate the objective function. By the definition of $F(X)$, we have

$$\begin{aligned}
\mathcal{F}(X) = \mathbb{E}(F(X)) &= \mathbb{E}\left(\frac{1}{4d}\sum_{i=1}^{d}(y_{(i)} - \langle A_i, XX^\top\rangle)^2\right) \\
&= \frac{1}{4d}\sum_{i=1}^{d}\mathbb{E}\left(\langle A_i, UU^\top\rangle + z_{(i)} - \langle A_i, XX^\top\rangle\right)^2 \\
&= \frac{1}{4d}\sum_{i=1}^{d}\mathbb{E}\left(\langle A_i, UU^\top - XX^\top\rangle + z_{(i)}\right)^2 \\
&= \frac{1}{4d}\sum_{i=1}^{d}\mathbb{E}\left(\mathrm{vec}(A_i)^\top \mathrm{vec}(UU^\top - XX^\top) + z_{(i)}\right)^2 \\
&\overset{(i)}{=} \frac{1}{4d}\sum_{i=1}^{d}\mathbb{E}\left(\mathrm{vec}(UU^\top - XX^\top)^\top \mathrm{vec}(A_i)\mathrm{vec}(A_i)^\top \mathrm{vec}(UU^\top - XX^\top) + z_{(i)}^2\right) \\
&= \frac{1}{4}\mathrm{vec}(UU^\top - XX^\top)^\top \cdot \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}(\mathrm{vec}(A_i)\mathrm{vec}(A_i)^\top)\cdot \mathrm{vec}(UU^\top - XX^\top) + \frac{\sigma_z^2}{4} \\
&= \frac{1}{4}\|\mathrm{vec}(UU^\top - XX^\top)\|_2^2 + \frac{\sigma_z^2}{4} = \frac{1}{4}\|UU^\top - XX^\top\|_{\mathrm{F}}^2 + \frac{\sigma_z^2}{4},
\end{aligned}$$

where (i) from the fact that $z_{(i)}$ has zero mean and is independent of A_i.

Next, we demonstrate the gradient and the Hessian matrix. From the independence of A_i's, we have

$$\mathbb{E}(\nabla F(X)) = \frac{1}{2}\mathbb{E}\left(\left(\langle A_i, XX^\top - UU^\top\rangle - z_{(i)}\right)\cdot (A_i + A_i^\top)X\right)$$
$$\mathbb{E}(\nabla^2 F(X)) = \frac{1}{2}\mathbb{E}\Bigg(I_r \otimes \left(\langle A_i, XX^\top - UU^\top\rangle - z_{(i)}\right)\cdot (A_i + A_i^\top)$$
$$+ \operatorname{vec}\left((A_i + A_i^\top)X\right)\cdot \operatorname{vec}\left((A_i + A_i^\top)X\right)^\top\Bigg).$$

We ignore the index i and denote A_i $(z_{(i)})$ as A (z) for the convenience of notation. The proof is analyzed by entry-wise agreement.

For the (j,k)-th entry of gradient $\nabla F(X)$, we have

$$\begin{aligned}\mathbb{E}(\nabla F(X)_{(j,k)}) &= \frac{1}{2}\mathbb{E}\left(\left(\langle A, XX^\top - UU^\top\rangle - z\right)\cdot (A + A^\top)_{(j,*)}X_{(*,k)}\right)\\ &\overset{(i)}{=} \frac{1}{2}\mathbb{E}\left(\sum_{s,t} A_{(s,t)}(XX^\top - UU^\top)_{(s,t)}\cdot \sum_{l}(A_{(j,l)} + A_{(l,j)})X_{(l,k)}\right)\\ &\overset{(ii)}{=} (XX^\top - UU^\top)X_{(j,k)}\end{aligned}$$

where (i) is from the zero mean of z and (ii) is from (7.193) in the proof of Lemma 4.

We use double index again for the Hessian matrix, i.e., denote (jk, st) as the $((k-1)n+j, (t-1)n+s)$-th entry of $\nabla^2 F(X)$. We discuss by separating the two components of $\nabla^2 F(X)$. For the first component,

$$\begin{aligned}&\mathbb{E}\left(\left(\langle A, XX^\top - UU^\top\rangle - z\right)\cdot (A + A^\top)_{(j,k)}\right)\\ &= \mathbb{E}\left(\sum_{s,t} A_{(s,t)}(XX^\top - UU^\top)_{(s,t)}\cdot (A_{(j,k)} + A_{(k,j)})\right)\\ &= \mathbb{E}\left(A_{(j,k)}^2(XX^\top - UU^\top)_{(j,k)} + A_{(k,j)}^2(XX^\top - UU^\top)_{(k,j)}\right)\\ &= 2(XX^\top - UU^\top)_{(j,k)}.\end{aligned}$$

The rest of the analysis is identical to that of Lemma 4.

References

[1] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[2] Ian Jolliffe. *Principal component analysis*. Springer, 2011.

[3] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[4] Michael W Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

[5] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[6] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[7] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

[8] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[9] John K Tsotsos, Scan M Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1-2):507–545, 1995.

[10] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[11] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

[12] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[13] Nikhil Rao, Joseph Harrison, Tyler Karrels, Robert Nowak, and Timothy T Rogers. Using machines to improve human saliency detection. In *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*, pages 80–84. IEEE, 2010.

[14] Guoshen Yu and Guillermo Sapiro. Statistical compressive sensing of gaussian mixture models. pages 3728–3731, 2011.

[15] Xiaohui Shen and Ying Wu. A unified approach to salient object detection via low rank matrix recovery. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 853–860. IEEE, 2012.

[16] Xingguo Li and Jarvis Haupt. Identifying outliers in large matrices via randomized adaptive compressive sampling. *IEEE Transactions on Signal Processing*, 63(7):1792–1807, 2015.

[17] Xingguo Li and Jarvis Haupt. A refined analysis for the sample complexity of adaptive compressive outlier sensing. In *Statistical Signal Processing Workshop (SSP), 2016 IEEE*, pages 1–5. IEEE, 2016.

[18] Xingguo Li and Jarvis Haupt. Outlier identification via randomized adaptive compressive sampling. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 3302–3306. IEEE, 2015.

[19] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. pages 2496–2504, 2010.

[20] Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

[21] Xingguo Li and Jarvis Haupt. Locating salient group-structured image features via adaptive compressive sensing. In *GlobalSIP*, pages 393–397, 2015.

[22] Xingguo Li and Jarvis Haupt. Robust low-complexity randomized methods for locating outliers in large matrices. *arXiv preprint arXiv:1612.02334*, 2016.

[23] Xingguo Li and Jarvis Haupt. Robust low-complexity methods for matrix column outlier identification. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2017 IEEE 7th International Workshop on*, pages 1–5. IEEE, 2017.

[24] Xingguo Li and Jarvis Haupt. Robust outlier identification for noisy data via randomized adaptive compressive sampling. In *The Signal Processing with Adaptive Sparse Structured Representations Workshop (SPARS)*, 2017.

[25] Shihao Ji, Ya Xue, Lawrence Carin, et al. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346, 2008.

[26] Eran Bashan, Raviv Raich, and Alfred O Hero. Optimal two-stage search for sparse targets using convex criteria. *IEEE Transactions on Signal Processing*, 56(11):5389–5402, 2008.

[27] Jarvis Haupt, Robert Nowak, and Rui Castro. Adaptive sensing for sparse signal recovery. In *Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009. DSP/SPE 2009. IEEE 13th*, pages 702–707. IEEE, 2009.

[28] Jarvis Haupt, Rui M Castro, and Robert Nowak. Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Transactions on Information Theory*, 57(9):6222–6235, 2011.

[29] Eran Bashan, Gregory Newstadt, and Alfred O Hero. Two-stage multiscale search for sparse targets. *IEEE Transactions on Signal Processing*, 59(5):2331–2341, 2011.

[30] Piotr Indyk, Eric Price, and David P Woodruff. On the power of adaptivity in sparse recovery. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 285–294. IEEE, 2011.

[31] MA Iwen and AH Tewfik. Adaptive group testing strategies for target detection and localization in noisy environments. 2010.

[32] Matthew L Malloy and Robert D Nowak. Sequential testing for sparse recovery. *IEEE Transactions on Information Theory*, 60(12):7862–7873, 2014.

[33] Sivaraman Balakrishnan, Mladen Kolar, Alessandro Rinaldo, and Aarti Singh. Recovering block-structured activations using compressive measurements. *arXiv preprint arXiv:1209.3431*, 2012.

[34] Rui M Castro et al. Adaptive sensing performance lower bounds for sparse signal detection and support estimation. *Bernoulli*, 20(4):2217–2246, 2014.

[35] Eric Price and David P Woodruff. Lower bounds for adaptive sparse recovery. pages 652–663, 2013.

[36] Matthew L Malloy and Robert D Nowak. Near-optimal adaptive compressed sensing. volume 60, pages 4001–4012. IEEE, 2014.

[37] Mark A Davenport and Ery Arias-Castro. Compressive binary search. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 1827–1831. IEEE, 2012.

[38] Akshay Krishnamurthy, James Sharpnack, and Aarti Singh. Recovering graph-structured activations using adaptive compressive measurements. *arXiv preprint arXiv:1305.0213*, 2013.

[39] Ery Arias-Castro, Emmanuel J Candes, and Mark A Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2013.

[40] Dennis Wei and Alfred O Hero. Multistage adaptive estimation of sparse signals. *IEEE Journal of Selected Topics in Signal Processing*, 7(5):783–796, 2013.

[41] Akshay Krishnamurthy and Aarti Singh. Low-rank matrix and tensor completion via adaptive sampling. In *Advances in Neural Information Processing Systems*, pages 836–844, 2013.

[42] Akshay Soni and Jarvis Haupt. On the fundamental limits of recovering tree sparse vectors from noisy linear measurements. *IEEE Transactions on Information Theory*, 60(1):133–149, 2014.

[43] Jarvis Haupt and Robert Nowak. Adaptive sensing for sparse recovery. In Yonina C Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and applications*. Cambridge University Press, 2011.

[44] Lester W Mackey, Michael I Jordan, and Ameet Talwalkar. Divide-and-conquer matrix factorization. In *Advances in neural information processing systems*, pages 1134–1142, 2011.

[45] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[46] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[47] Yudong Chen, Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust matrix completion and corrupted columns. pages 873–880, 2011.

[48] Michael McCoy, Joel A Tropp, et al. Two proposals for robust PCA using semidefinite programming. *Electronic Journal of Statistics*, 5:1123–1160, 2011.

[49] Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Conference on Learning Theory*, pages 354–375, 2013.

[50] Gilad Lerman, Michael B McCoy, Joel A Tropp, and Teng Zhang. Robust computation of linear models by convex relaxation. *Foundations of Computational Mathematics*, 15(2):363–410, 2015.

[51] Sirisha Rambhatla, Xingguo Li, and Jarvis D Haupt. A dictionary based generalization of robust pca. In *GlobalSIP*, pages 1315–1319, 2016.

[52] Jineng Ren, Xingguo Li, and Jarvis Haupt. Robust pca via tensor outlier pursuit. In *Signals, Systems and Computers, 2016 50th Asilomar Conference on*, pages 1744–1749. IEEE, 2016.

[53] Xingguo Li, Jineng Ren, Sirisha Rambhatla, YangYang Xu, and Jarvis Haupt. Robust pca via dictionary based outlier pursuit. In *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[54] Sirisha Rambhatla, Xingguo Li, and Jarvis Haupt. Target based hyperspectral demixing via generalized robust pca. In *Asilomar Conference on Signals, Systems, and Computers (Asilomar)*, 2017.

[55] John Wright, Arvind Ganesh, Kerui Min, and Yi Ma. Compressive principal component pursuit. *Information and Inference: A Journal of the IMA*, 2(1):32–68, 2013.

[56] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[57] Junchi Yan, Mengyuan Zhu, Huanxi Liu, and Yuncai Liu. Visual saliency detection via sparsity pursuit. *IEEE Signal Processing Letters*, 17(8):739–742, 2010.

[58] Michal Aharon, Michael Elad, Alfred Bruckstein, et al. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311, 2006.

[59] Yin Li, Yue Zhou, Lei Xu, Xiaochao Yang, and Jie Yang. Incremental sparse saliency detection. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 3093–3096. IEEE, 2009.

[60] Ying Yu, Bin Wang, and Liming Zhang. Saliency-based compressive sampling for image signals. *IEEE signal processing letters*, 17(11):973–976, 2010.

[61] Cem Aksoylar, George K Atia, and Venkatesh Saligrama. Sparse signal processing with linear and nonlinear observations: A unified shannon-theoretic approach. *IEEE Transactions on Information Theory*, 63(2):749–776, 2017.

[62] Jarvis Haupt. Locating salient items in large data collections with compressive linear measurements. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2013 IEEE 5th International Workshop on*, pages 9–12. Citeseer, 2013.

[63] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[64] Anna C Gilbert, Jae Young Park, and Michael B Wakin. Sketched SVD: Recovering spectral features from compressive measurements. *arXiv preprint arXiv:1211.0361*, 2012.

[65] Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281. ACM, 2001.

[66] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563. ACM, 2006.

[67] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 341–350. ACM, 2010.

[68] Tom Goldstein, Brendan O'Donoghue, Simon Setzer, and Richard Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.

[69] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[70] Marco F Duarte, Mark A Davenport, Dharmpal Takhar, Jason N Laska, Ting Sun, Kevin F Kelly, and Richard G Baraniuk. Single-pixel imaging via compressive sampling. *IEEE signal processing magazine*, 25(2):83–91, 2008.

[71] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[72] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15–15, 2009.

[73] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*, pages 1597–1604. IEEE, 2009.

[74] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[75] Lijuan Duan, Chunpeng Wu, Jun Miao, Laiyun Qing, and Yu Fu. Visual saliency detection by spatially weighted dissimilarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 473–480. IEEE, 2011.

[76] Laura Balzano, Benjamin Recht, and Robert Nowak. High-dimensional matched subspace detection when data are missing. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1638–1642. IEEE, 2010.

[77] Akshay Krishnamurthy and Aarti Singh. On the power of adaptivity in matrix completion and approximation. *arXiv preprint arXiv:1407.3619*, 2014.

[78] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[79] Sung Won Park and Marios Savvides. Individual kernel tensor-subspaces for robust face recognition: A computationally efficient tensor framework without requiring mode factorization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(5):1156–1166, 2007.

[80] Weiwei Guo, Irene Kotsia, and Ioannis Patras. Tensor learning for regression. *IEEE Transactions on Image Processing*, 21(2):816–827, 2012.

[81] Qibin Zhao, Cesar F Caiafa, Danilo P Mandic, Zenas C Chao, Yasuo Nagasaka, Naotaka Fujii, Liqing Zhang, and Andrzej Cichocki. Higher order partial least squares (hopls): a generalized multilinear regression method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1660–1673, 2013.

[82] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[83] Rose Yu, Dehua Cheng, and Yan Liu. Accelerated online low-rank tensor learning for multivariate spatio-temporal streams. In *International Conference on Machine Learning*, 2015.

[84] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

[85] Xiaoshan Li, Hua Zhou, and Lexin Li. Tucker tensor regression and neuroimaging analysis. *arXiv preprint arXiv:1304.5637*, 2013.

[86] Bernardino Romera-Paredes, Hane Aung, Nadia Bianchi-Berthouze, and Massimiliano Pontil. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1444–1452, 2013.

[87] Yongxin Yang and Timothy Hospedales. Deep multi-task representation learning: A tensor factorisation approach. *arXiv preprint arXiv:1605.06391*, 2016.

[88] Mohammad Taha Bahadori, Qi Rose Yu, and Yan Liu. Fast multivariate spatio-temporal analysis via low-rank tensor learning. In *Advances in Neural Information Processing Systems*, pages 3491–3499, 2014.

[89] Peter D Hoff. Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics*, 9(3):1169, 2015.

[90] Rose Yu and Yan Liu. Learning from multiway data: Simple and efficient tensor regression. In *International Conference on Machine Learning*, pages 373–381, 2016.

[91] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[92] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *arXiv preprint arXiv:1607.01668*, 2016.

[93] Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[94] Ryota Tomioka and Taiji Suzuki. Convex tensor decomposition via structured schatten norm regularization. In *Advances in Neural Information Processing Systems*, pages 1331–1339, 2013.

[95] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

[96] Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009.

[97] David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[98] Michael W Mahoney and Petros Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[99] Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

[100] Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506, 2012.

[101] Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, pages 81–90. ACM, 2013.

[102] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson–Lindenstrauss transform. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, pages 341–350. ACM, 2010.

[103] Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. *Journal of the ACM*, 61(1):4:1–4:23, 2014.

[104] Michel Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science &amp; Business Media, 2006.

[105] Zhongmin Shen. *Lectures on Finsler geometry*, volume 2001. World Scientific, 2001.

[106] Sjoerd Dirksen. Dimensionality reduction with subgaussian matrices: A unified theory. *Foundations of Computational Mathematics*, pages 1–30, 2015.

[107] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in Euclidean space. *Geometric and Functional Analysis*, 25(4):1009–1088, 2015.

[108] Jelani Nelson and Huy L Nguyen. Lower bounds for oblivious subspace embeddings. In *International Colloquium on Automata, Languages, and Programming*, pages 883–894. Springer, 2014.

[109] Jarvis Haupt, Xingguo Li, and David P Woodruff. Near optimal sketching of low-rank tensor regression. *arXiv preprint arXiv:1709.07093*, 2017.

[110] Joel A Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.

[111] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[112] Hua Zhou. Matlab TensorReg toolbox. http://hua-zhou.github.io/softwares/tensorreg/, 2013.

[113] Antoine Rosset, Luca Spadola, and Osman Ratib. Osirix: an open-source software for navigating in multidimensional DICOM images. *Journal of Digital Imaging*, 17(3):205–216, 2004.

[114] Lie Wang. The ℓ_1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.

[115] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[116] Han Liu, Lie Wang, and Tuo Zhao. Calibrated multivariate regression with application to neural semantic basis discovery. *Journal of Machine Learning Research*, 16:1579–1606, 2015.

[117] Xingguo Li, Yanbo Xu, Tuo Zhao, and Han Liu. Statistical and computational tradeoff of regularized dantzig-type estimators. 2016.

[118] Xingguo Li, Tuo Zhao, Xiaoming Yuan, and Han Liu. The flare package for high dimensional linear regression and precision matrix estimation in r. *The Journal of Machine Learning Research*, 16(1):553–557, 2015.

[119] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[120] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[121] Han Liu and Lie Wang. Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. *Electronic Journal of Statistics*, 11(1):241–294, 2017.

[122] Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, Vincent Leclère, and Joseph Salmon. Efficient smoothed concomitant lasso estimation for high dimensional regression. *arXiv preprint arXiv:1606.02702*, 2016.

[123] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

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

[125] Xingguo Li, Haoming Jiang, Jarvis Haupt, Raman Arora, Han Liu, Mingyi Hong, and Tuo Zhao. On fast convergence of proximal algorithms for sqrt-lasso optimization: Don't worry about its nonsmooth loss function. *arXiv preprint arXiv:1605.07950*, 2016.

[126] Tuo Zhao, Han Liu, and Tong Zhang. Pathwise coordinate optimization for sparse learning: Algorithm and theory. *arXiv preprint arXiv:1412.7477*, 2014.

[127] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[128] Dimitri P Bertsekas. *Nonlinear programming.* Athena scientific Belmont, 1999.

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

[130] Lin Xiao and Tong Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

[131] Arkadi Nemirovski. Interior point polynomial time methods in convex programming. *Lecture Notes*, 2004.

[132] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[133] Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[134] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[135] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.

[136] Fei Ye and Cun-Hui Zhang. Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *The Journal of Machine Learning Research*, 11:3519–3540, 2010.

[137] D. D. Lucas, C. Yver Kwok, P. Cameron-Smith, H. Graven, D. Bergmann, T. P. Guilderson, R. Weiss, and R. Keeling. Designing optimal greenhouse gas observing networks that consider performance and cost. *Geoscientific Instrumentation, Methods and Data Systems*, 4(1):121–137, 2015.

[138] Tingni Sun. Package 'scalreg'. 2013.

[139] Lu Li and Kim-Chuan Toh. An inexact interior point method for ℓ_1-regularized sparse covariance selection. *Mathematical Programming Computation*, 2(3-4):291–315, 2010.

[140] Peter McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

[141] Johann Pfanzagl. *Parametric statistical theory.* Walter de Gruyter, 1994.

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

[143] Sara A van de Geer. High-dimensional generalized linear models and the Lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

[144] Maxim Raginsky, Rebecca M Willett, Zachary T Harmany, and Roummel F Marcia. Compressed sensing performance bounds under poisson noise. *IEEE Transactions on Signal Processing*, 58(8):3990–4002, 2010.

[145] Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Jarvis Haupt. Stochastic variance reduced optimization for nonconvex sparse learning. In *International Conference on Machine Learning*, pages 917–925, 2016.

[146] Benjamin M Neale, Yan Kou, Li Liu, Avi Ma'Ayan, Kaitlin E Samocha, Aniko Sabo, Chiao-Feng Lin, Christine Stevens, Li-San Wang, Vladimir Makarov, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397):242–245, 2012.

[147] Ani Eloyan, John Muschelli, Mary Beth Nebel, Han Liu, Fang Han, Tuo Zhao, Anita D Barber, Suresh Joel, James J Pekar, Stewart H Mostofsky, et al. Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in systems neuroscience*, 6, 2012.

[148] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[149] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[150] Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11:1081–1107, 2010.

[151] Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 2015. to appear.

[152] Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics*, 42(6):2164–2201, 2014.

[153] Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. Tac for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *arXiv preprint arXiv:1507.01037*, 2015.

[154] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *Image Processing, IEEE Transactions on*, 18(11):2419–2434, 2009.

[155] Zhi-Quan Luo and Paul Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2):408–425, 1992.

[156] Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for ℓ_1-regularized loss minimization. *The Journal of Machine Learning Research*, 12:1865–1892, 2011.

[157] Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Mingyi Hong. On faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization. *Journal of Machine Learning Research*, 18(184):1–24, 2018.

[158] Tuo Zhao, Han Liu, and Tong Zhang. A general theory of pathwise coordinate optimization. *arXiv preprint arXiv:1412.7477*, 2014.

[159] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1–13, 2010.

[160] Xingguo Li, Tuo Zhao, Tong Zhang, and Han Liu. The picasso package for nonconvex regularized m-estimation in high dimensions in r. Technical report, Technical Report, 2015.

[161] Xingguo Li, Lin Yang, Jason Ge, Jarvis Haupt, Tong Zhang, and Tuo Zhao. On quadratic convergence of dc proximal newton algorithm in nonconvex sparse

learning. In *Advances in Neural Information Processing Systems*, pages 2742–2752, 2017.

[162] X. Shen, W. Pan, and Y. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

[163] Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

[164] Shuheng Zhou. Restricted eigenvalue conditions on sub-Gaussian random matrices. Technical report, University of Michigan Ann Arbor, 2009.

[165] Sara A van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[166] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research*, 11(8):2241–2259, 2010.

[167] Xingguo Li, Jarvis Haupt, Raman Arora, Han Liu, Mingyi Hong, and Tuo Zhao. A first order free lunch for sqrt-lasso. *arXiv preprint arXiv:1605.07950*, 2016.

[168] Julien Mairal, Francis Bach, Jean Ponce, et al. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

[169] Yi Yang and Hui Zou. An efficient algorithm for computing the hhsvm and its generalizations. *Journal of Computational and Graphical Statistics*, 22(2):396–415, 2013.

[170] Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. The huge package for high-dimensional undirected graph estimation in R. *The Journal of Machine Learning Research*, 13:1059–1062, 2012.

[171] Xingguo Li, Tuo Zhao, Tong Zhang, and Han Liu. The picasso package for nonconvex regularized m-estimation in high dimensions in R. *Technical Report*, 2015.

[172] Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the nips 2003 feature selection challenge. In *Advances in neural information processing systems*, pages 545–552, 2005.

[173] Nicol N Schraudolph. A fast, compact approximation of the exponential function. *Neural Computation*, 11(4):853–862, 1999.

[174] A Cristiano I Malossi, Yves Ineichen, Costas Bekas, and Alessandro Curioni. Fast exponential computation on simd architectures. *Proc. of HIPEAC-WAPCO, Amsterdam NL*, 2015.

[175] Ian En-Hsu Yen, Cho-Jui Hsieh, Pradeep K Ravikumar, and Inderjit S Dhillon. Constant nullspace strong convexity and fast convergence of proximal methods under high-dimensional settings. In *Advances in Neural Information Processing Systems*, pages 1008–1016, 2014.

[176] Man-Chung Yue, Zirui Zhou, and Anthony Man-Cho So. Inexact regularized proximal newton method: provable convergence guarantees for non-smooth convex minimization without strong convexity. *arXiv preprint arXiv:1605.07522*, 2016.

[177] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[178] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[179] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[180] Stephen R Becker, Emmanuel J Candès, and Michael C Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.

[181] Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010.

[182] Kiryung Lee and Yoram Bresler. Admira: Atomic decomposition for minimum rank approximation. *IEEE Transactions on Information Theory*, 56(9):4402–4416, 2010.

[183] Shai Shalev-Shwartz, Alon Gonen, and Ohad Shamir. Large-scale convex minimization with a low-rank constraint. *arXiv preprint arXiv:1106.1622*, 2011.

[184] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[185] Samuel Burer and Renato DC Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

[186] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582, 2016.

[187] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

[188] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[189] Anima Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. *arXiv preprint arXiv:1602.05908*, 2016.

[190] Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Finding low-rank solutions via non-convex matrix factorization, efficiently and provably. *arXiv preprint arXiv:1606.03168*, 2016.

[191] Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low-rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.

[192] Stephen Tu, Ross Boczar, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

[193] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.

[194] Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. *arXiv preprint arXiv:1609.03240*, 2016.

[195] Raghunandan H Keshavan, Sewoong Oh, and Andrea Montanari. Matrix completion from a few entries. In *2009 IEEE International Symposium on Information Theory*, pages 324–328. IEEE, 2009.

[196] Moritz Hardt. Understanding alternating minimization for matrix completion. In *Foundations of Computer Science, 2014 IEEE 55th Annual Symposium on*, pages 651–660. IEEE, 2014.

[197] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via nonconvex factorization. In *Foundations of Computer Science, 2015 IEEE 56th Annual Symposium on*, pages 270–289. IEEE, 2015.

[198] Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

[199] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.

[200] Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *arXiv preprint arXiv:1605.08370*, 2016.

[201] T Tony Cai, Zongming Ma, Yihong Wu, et al. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

[202] Aharon Birnbaum, Iain M Johnstone, Boaz Nadler, and Debashis Paul. Minimax bounds for sparse PCA with noisy high-dimensional data. *The Annals of Statistics*, 41(3):1055, 2013.

[203] Vincent Q Vu, Jing Lei, et al. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

[204] Zhehui Chen, Xingguo Li, Lin F Yang, Jarvis Haupt, and Tuo Zhao. On landscape of lagrangian functions and stochastic search for constrained nonconvex optimization. *arXiv preprint arXiv:1806.05151*, 2018.

[205] Ming Lin and Jieping Ye. A non-convex one-pass framework for generalized factorization machines and rank-one matrix sensing. *arXiv preprint arXiv:1608.05995*, 2016.

[206] Mathieu Blondel, Masakazu Ishihata, Akinori Fujino, and Naonori Ueda. Polynomial networks and factorization machines: New insights and efficient training algorithms. *arXiv preprint arXiv:1607.08810*, 2016.

[207] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. *arXiv preprint arXiv:1605.07784*, 2016.

[208] Quanquan Gu, Zhaoran Wang, and Han Liu. Low-rank and sparse structure pursuit via alternating minimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 600–609, 2016.

[209] Ery Arias-Castro and Nicolas Verzelen. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 06 2014.

[210] Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ICA with unknown Gaussian noise, with implications for Gaussian mixtures and autoencoders. In *Advances in Neural Information Processing Systems*, pages 2375–2383, 2012.

[211] James Y Zou, Daniel J Hsu, David C Parkes, and Ryan P Adams. Contrastive learning using spectral methods. In *Advances in Neural Information Processing Systems*, pages 2238–2246, 2013.

[212] Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015.

[213] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[214] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*, 2016.

[215] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *University of California, Berkeley*, 1050:16, 2016.

[216] Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *arXiv preprint arXiv:1605.00405*, 2016.

[217] Xingguo Li, Zhaoran Wang, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, and Tuo Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint*, 2016.

[218] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. The global optimization geometry of nonsymmetric matrix factorization and sensing. *arXiv preprint arXiv:1703.01256*, 2017.

[219] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

[220] David Steven Dummit and Richard M Foote. *Abstract algebra*, volume 3. Wiley Hoboken, 2004.

[221] Bin Yang. Projection approximation subspace tracking. *IEEE Transactions on Signal processing*, 43(1):95–107, 1995.

[222] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning.* MIT press, 2016.

[223] Weiyang Liu, Yan-Ming Zhang, Xingguo Li, Zhiding Yu, Bo Dai, Tuo Zhao, and Le Song. Deep hyperspherical learning. In *Advances in Neural Information Processing Systems*, pages 3950–3960, 2017.

[224] Joel W Robbin and Dietmar A Salamon. Introduction to differential geometry. *ETH, Lecture Notes, preliminary version, January*, 2011.

[225] Nathan Srebro and Tommi Jaakkola. Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning*, pages 720–727, 2003.

[226] Yair Carmon and John C Duchi. Gradient descent efficiently finds the cubic-regularized non-convex newton step. *arXiv preprint arXiv:1612.00547*, 2016.

[227] Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[228] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[229] Mark A Davenport, Petros T Boufounos, Michael B Wakin, and Richard G Baraniuk. Signal processing with compressive measurements. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):445–460, 2010.

[230] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[231] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 143–152. IEEE, 2006.

[232] Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

[233] Achim Klenke and Lutz Mattner. Stochastic ordering of classical discrete distributions. *Advances in Applied probability*, 42(2):392–410, 2010.

[234] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[235] Vasek Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25(3):285–287, 1979.

[236] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[237] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[238] Bingxiang Li, Wen Li, and Lubin Cui. New bounds for perturbation of the orthogonal projection. *Calcolo*, 50(1):69–78, 2013.

[239] Martin Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. *preparation. University of California, Berkeley*, 2015.

[240] Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on*, 59(6):3434–3447, 2013.

[241] Emmanuel J Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[242] Yang Ning, Tianqi Zhao, and Han Liu. A likelihood ratio framework for high dimensional semiparametric regression. *arXiv preprint arXiv:1412.2295*, 2014.

[243] Yu Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

[244] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science &amp; Business Media, 2011.

[245] Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J Tibshirani. Strong rules for discarding predictors in Lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.

[246] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

[247] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[248] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.