Provable Algorithms for Scalable and Robust Low-Rank Matrix Recovery

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Yuanxin Li, M.S.

Graduate Program in Electrical and Computer Engineering

The Ohio State University

2018

Dissertation Committee:

Yuejie Chi, Advisor

Yingbin Liang

Wei Zhang

Abstract

Low-rank models are ubiquitous in a wide range of practical applications, and low-rank matrix sensing and recovery has become a problem of great importance. Via leveraging the low-rank structure in data representations, it is possible to faithfully recover the matrix of interest from incomplete observations, in both statistically and computationally efficient manners. This dissertation investigates the fundamental problem of low-rank matrix recovery in different random measurement models, possibly with corruptions.

We first consider recovering low-rank positive semidefinite (PSD) matrices from random rank-one measurements, which spans numerous applications including covariance sketching, phase retrieval, quantum state tomography, and learning shallow polynomial neural networks, among others. Our approach is to directly estimate the low-rank factor by minimizing a nonconvex least-squares loss function via vanilla gradient descent, following a tailored spectral initialization. When the true rank is small, this algorithm is guaranteed to converge to the ground truth (up to global ambiguity) with near-optimal sample and computational complexities with respect to the problem size. To the best of our knowledge, this is the first guarantee that achieves near-optimality in both metrics, without the need of sample splitting.

When the rank-one measurements are possibly corrupted by arbitrary outliers, we propose a convex optimization algorithm that seeks the PSD matrix with the

minimum ℓ_1-norm of the observation residual. The advantage of our algorithm is that it is free of parameters, therefore eliminating the need of tuning and allowing easy implementations. We establish that with high probability, a low-rank PSD matrix can be exactly recovered as soon as the number of measurements is large enough, even when a fraction of the measurements are corrupted by outliers with arbitrary magnitudes. Moreover, the recovery is also stable against bounded noise. With the additional information of an upper bound of the rank of the PSD matrix, we then propose another nonconvex algorithm based on subgradient descent that exhibits excellent empirical performance in terms of computational efficiency and accuracy.

Moreover, we work on the recovery of generic low-rank matrices from random full-rank linear measurements in the presence of outliers, where we employ a median truncation strategy in gradient descent to improve the robustness of recovery procedure against outliers. We demonstrate that, when initialized in a basin of attraction close to the ground truth, the proposed algorithm converges to the ground truth at a linear rate for the Gaussian measurement model with a near-optimal number of measurements, even when a constant fraction of the measurements are arbitrarily corrupted. In addition, we propose a new truncated spectral method that ensures a valid initialization in the basin of attraction at slightly higher requirements.

Dedicated to my parents Gaiming Li and Sumei Chang,

and my brother Changfeng Li

Acknowledgments

First and foremost, I would like to express my deepest and sincerest gratitude to my advisor, Prof. Yuejie Chi, for her unreserved help, support and encouragement during my entire Ph.D. program. Her passion and dedication to work and research has always been inspirational to me. I feel extremely fortunate to work with her. She has always provided excellent environments for research and maintained the highest level of availability for discussion. Without her guidance and persistent help this dissertation would not have been possible.

I would like to thank Prof. Yingbin Liang and Prof. Wei Zhang, who have served on my dissertation committee, for taking the time and providing insightful feedback.

I would like to thank the people I have collaborated during my Ph.D. study, including Prof. Yuxin Chen, Prof. Yingbin Liang, Prof. Louis L. Scharf, Prof. Ali Pezeshki, Prof. Yue M. Lu, Cong Ma, Huishuai Zhang, Yue Sun, Yingsheng He, and others, for their time and input for the research work. Special thanks go to Prof. Louis L. Scharf and Prof. Ali Pezeshki for hosting my visit at Colorado State University. I would also like to thank Dr. Shirin Jalali at Nokia Bell Labs for her mentorship during my internship.

I would like to thank all of the faculty and staff at The Ohio State University that have guided my courses of learning and provided help in different aspects. I would also like to thank my fellow labmates at The Ohio State University and Carnegie

Mellon University, including Jiaqing Huang, Yiran Jiang, Liming Wang, Haoyu Fu, Azer Shikhaliev, Vince Monardo, and Myung Cho, who have been a generous source of support and encouragement.

I would like to thank my friends at and outside The Ohio State University for their friendship. Their company makes my life in Columbus much more colorful. Last but not least, I would like to thank my family, my parents and my brother for their unconditional love and care over the years. To them I dedicate this dissertation.

Vita

2010 ..... B.Eng., Communication Engineering,
Nanjing University of Posts and Telecommunications,
Nanjing, China

2013 ..... M.Eng., Information and Communication Engineering,
Tsinghua University,
Beijing, China

2016 ..... M.S., Electrical and Computer Engineering,
The Ohio State University

2013-present ..... Graduate Research Associate,
Electrical and Computer Engineering,
The Ohio State University

Publications

Research Publications

Journals

Y. Li, Y. Sun and Y. Chi, "Low-Rank Positive Semidefinite Matrix Recovery from Corrupted Rank-One Measurements", *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 397-408, 2017.

Y. Li and Y. Chi, "Stable Separation and Super-Resolution of Mixture Models", *Applied and Computational Harmonic Analysis*, in press, 2017. [Online]. Available: `https://doi.org/10.1016/j.acha.2017.03.003`

Y. Li and Y. Chi, "Off-the-Grid Line Spectrum Denoising and Estimation with Multiple Measurement Vectors", *IEEE Transactions on Signal Processing*, vol. 64, no. 5, pp. 1257-1269, 2016.

Conference Proceedings

Y. Li, Y. Chi, H. Zhang and Y. Liang, "Non-Convex Low-rank Matrix Recovery from Corrupted Random Linear Measurements", 12*th International Conference on Sampling Theory and Applications*, Tallinn, Estonia, 2017.

Y. Li, A. Pezeshki, L. L. Scharf, and Y. Chi, "Performance Bounds for Modal Analysis using Sparse Linear Arrays", *SPIE Compressive Sensing VI: From Diverse Modalities to Big Data Analytics*, Anaheim, California, USA, 2017.

Y. Sun, Y. Li, and Y. Chi, "Outlier-Robust Recovery of Low-Rank Positive Semidefinite Matrices from Magnitude Measurements", *The* 41*th IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, 2016.

Y. Li, Y. He, Y. Chi and Y. M. Lu, "Blind Calibration of Multi-Channel Samplers using Sparse Recovery", *IEEE* 6*th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, Cancún, Mexico, 2015.

Y. Li and Y. Chi, "Super-Resolution of Mutually Interfering Signals", *IEEE International Symposium on Information Theory*, Hong Kong, China, 2015.

Y. Li and Y. Chi, "Parameter Estimation for Mixture Models via Convex Optimization", 11*th International Conference on Sampling Theory and Applications*, Washington, D.C., USA, 2015.

Y. Li and Y. Chi, "Compressive Parameter Estimation With Multiple Measurement Vectors via Structured Low-Rank Covariance Estimation", *IEEE Workshop on Statistical Signal Processing* , Gold Coast, VIC, Australia, 2014.

Fields of Study

Major Field: Electrical and Computer Engineering

Table of Contents

Page

List of Tables

Table **Page**

List of Figures

Chapter 1: Introduction

1.1 Backgrounds

In applied science and engineering, there are a variety of problems that require learning, extracting and estimating a matrix from the acquired data. One motivating example is the famous Netflix problem in the field of recommender systems [1], where one would like to estimate the whole movie rating matrix, in order to infer the preference of each user and provide personalized recommendations, based on a few known ratings users have submitted. Usually, the matrices of interest can be extremely large, especially in high-dimensional problems, and the potential size will continue growing owing to the availability of vast amounts of data created by modern sensing modalities at an unprecedented rate due to declining cost of data acquisition. As a consequence, it is very difficult and even infeasible to obtain the full observations of the matrix of interest, and modern data applications often have to work with only under-sampled measurements or partial observations (i.e. incomplete observations), the number of which is much smaller than the ambient dimension of the data matrix of interest. Examples struggling with incomplete observations are numerous. In the recommender systems as aforementioned Netflix problem, each user typically rates only very few items, so the entire rating matrix is highly incomplete and needs inference. In sensor

localization, to avoid the prohibitive expense for measuring all the pairwise distances, one may consider to measure only a few connections of sensors close enough with each other and extrapolate the huge sensor map based on the available partial matrix of pairwise distances. Therefore, it is of great practical importance to estimate the full matrix of interest from incomplete observations, which, however, in general is not always possible.

Fortunately, it is promising to exploit the intrinsic low-dimensional geometric structure embedded in most real-world high-dimensional data to cope with the curse of dimensionality, and make the matrix estimation problem solvable. In a wide range of settings, the matrix one wishes to estimate can be assumed to have a low-dimensional geometric structure in the sense that it is low-rank or approximately low-rank, which may result from physical reasons or engineering designs. For instance, in the recommender systems, the entire rating matrix can be approximated by a low-rank matrix because it is commonly believed that the preference of each user is only determined by a few key factors. Low-rank models are also ubiquitous in machine learning such as feature learning [2], collaborative prediction [3] and natural language processing [4]. This low-rank structure, which can be considered as a powerful regularization scheme, opens the door to faithfully estimate the matrix of interest from incomplete observations, in both statistically and computationally efficient manners, even in the sample-starved or resource-starved environments. In fact, a considerable amount of work has been done on low-rank matrix estimation in recent years, where it is shown that low-rank matrices can be estimated accurately and efficiently from much fewer observations than their ambient dimensions in a diverse set of applications [5–10]. Extensive overviews on low-rank matrix estimation can be found in [11, 12].

1.2 Problem Statements

Mathematically, low-rank matrix estimation refers to estimating a rank-r matrix $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ from a group of measurements of the form

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{M}) + \boldsymbol{\eta} + \boldsymbol{w} \in \mathbb{R}^m \tag{1.1}$$

in different setups, where $r \ll \min\{n_1, n_2\}$. The linear transformation $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ represents an *a priori* known mapping from matrices to vectors, of which the ith entry is defined as $\mathcal{A}_i(\boldsymbol{M}) = \langle \boldsymbol{A}_i, \boldsymbol{M} \rangle$ with the ith sensing matrix given as $\boldsymbol{A}_i \in \mathbb{R}^{n_1 \times n_2}$, for $i = 1, \cdots, m$. The vector $\boldsymbol{\eta} \in \mathbb{R}^m$ and the vector $\boldsymbol{w} \in \mathbb{R}^m$ represent the sparse outlier vector and the dense noise vector, respectively, which are the potential corruptions and contaminations suffered by the measurements. It is natural to ask if it is possible to estimate the low-rank matrix $\boldsymbol{M}$ from an information-theoretically optimal number of measurements $\boldsymbol{y}$ in a computationally efficient manner.

In the literature, low-rank matrix estimation usually can be divided into two categories according to the type of sensing matrices $\boldsymbol{A}_i$'s:

- **Low-rank matrix recovery:** Each sensing matrix $\boldsymbol{A}_i$ is a dense matrix, typically randomly generated following certain distributions, and hence, each measurement $\mathcal{A}_i(\boldsymbol{M})$ is a linear combination of the entries in $\boldsymbol{M}$.

- **Low-rank matrix completion:** Each sensing matrix $\boldsymbol{A}_i$ is a sparse matrix with a single entry equaling 1, which can be interpreted as $\boldsymbol{A}_i = \boldsymbol{e}_{i_1} \boldsymbol{e}_{i_2}^\top$, where $\boldsymbol{e}_i$ represents the ith standard basis vector. Hence, each measurement $\mathcal{A}_i(\boldsymbol{M})$ corresponds to one entry in $\boldsymbol{M}$. Then, matrix completion aims at filling in the missing entries of the partially observed matrix.

There is a plethora of progress in low-rank matrix estimation by searching for the ground truth $\boldsymbol{M}$ directly in the high dimension compatible with measurement models, with the low-rank structure motivated via nuclear norm minimization [5,9,10,13–21], which can be regarded as the convex relaxation counterpart of ℓ_1-minimization in compressed sensing [22, 23]. This convex relaxation strategy guarantees accurate matrix estimation with (near-)optimal sample complexity under mild assumptions, nevertheless, the involved nuclear norm minimization, often cast as the semidefinite programming, is in general computationally expensive with large-scale data. In practice, a widely used alternative, pioneered by Burer and Monteiro [24], is to estimate the low-rank factors $\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}$ as a result of low-rank matrix factorization as $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{Y}^\top \in \mathbb{R}^{n_1 \times n_2}$, if the rank r or its upper bound is approximately known. Since the low-rank factors have a much lower-dimensional representation, this approach admits more computationally and memory efficient algorithms. Even though the bilinear constraint induced by matrix factorization typically leads to a nonconvex loss function that may be difficult to optimize globally, a growing series of recent work is shedding new light on the power of nonconvex optimization approaches for low-rank matrix estimation [25–30].

1.3 Contributions

This dissertation is dedicated to provide algorithms, supported with theoretical performance guarantees, for scalable and robust *low-rank matrix recovery* in different sensing models. Roughly speaking, we focus on the recovery of low-rank matrices from measurements obtained with randomly generated sensing matrices. Furthermore, the measurements, perhaps, are corrupted by outliers and additive noise. Via leveraging

the low-rank structure in data representations, we are capable of reducing the sample complexity as well as computational complexity required in matrix recovery while providing desirable, robust and stable performance, which is not only verified through extensive numerical experiments, but also, more importantly, analytically guaranteed by theories. In particular, our purpose is to design algorithms for low-rank matrix recovery with three principal properties as follows:

- **Provability:** The performance of proposed algorithms could be rigorously analyzed, and the desirable recovery results could be theoretically guaranteed under mild conditions.

- **Robustness:** The proposed algorithms could achieve robust recovery of low-rank matrix even when the measurements are further corrupted by outliers, possibly adversarial. This bears great importance since in real-world applications, outliers are somewhat inevitable which may be caused by sensor failures, malicious attacks, or reading errors [31–33], so it becomes critical to address robust recovery of matrix of interest in the presence of outliers.

- **Scalability:** The proposed algorithms could handle a growing amount of data in a computational efficient manner whose complexity scales nearly linearly with the problem dimension.

Specifically, we first consider the low-rank matrix recovery in the rank-one sensing model, of which each sensing matrix is generated as $\boldsymbol{A}_i = \boldsymbol{a}_i\boldsymbol{a}_i^\top$, where $\boldsymbol{a}_i$ is a random vector with entries independently drawn from standard Gaussian distribution, for $i = 1, \cdots, m$. In Chapter 2, we provide a refined analysis on low-rank positive semidefinite (PSD) matrix recovery from clean rank-one measurements via gradient

descent, and our theoretical results significantly improve upon existing results both statistically and computationally. To the best of our knowledge, this work is the first nonconvex algorithm (without resampling) that achieves both near-optimal statistical and computational guarantees with respect to the problem size.

For the rank-one sensing model, when the available measurements are further corrupted by arbitrary outliers and additive bounded noise, in Chapter 3 we present an algorithm based on convex optimization and establish the theoretical guarantees to demonstrate the robust and stable performance of the proposed algorithm against outliers and noise. The proposed convex program is free of tuning parameters and, consequently, easy to implement. Moreover, to further reduce the computational burden when facing large-scale problems, we also design a nonconvex algorithm based on subgradient descent for the same corruption scenario, which exhibits excellent empirical performance in terms of computational efficiency and accuracy.

We finally study low-rank matrix recovery in the full-rank linear sensing model, where each sensing matrix $\boldsymbol{A}_i$ is a random matrix composed of independent and identically distributed (i.i.d.) standard Gaussian entries, for $i = 1, \cdots, m$. Furthermore, the measurements may suffer from adversarial outliers with arbitrary amplitudes. In Chapter 4, benefiting from an adaptive, iteration-varying truncation strategy in gradient descent to mitigate the effects of outliers, we develop a fast and robust nonconvex algorithm consisting of spectral initialization and gradient descent update for robust generic low-rank matrix recovery, which subsumes the low-rank PSD matrix recovery as a special case. In particular, the proposed algorithm does not assume a priori information regarding the outliers in terms of their fraction, distribution nor values. The effectiveness of the developed algorithm is provably guaranteed by theoretical

analysis, and numerical examples are provided to validate the favorable performance of the proposed algorithm as well.

1.4 Notations

We collect the notations that are frequently used throughout this dissertation here. We use boldface lowercase (resp. uppercase) letters to represent vectors (resp. matrices). In particular, we use $\boldsymbol{I}_n$ to represent an n dimensional identity matrix. We denote by $\boldsymbol{x}^\top$ and $\|\boldsymbol{x}\|_p$ the transpose and the ℓ_p-norm of a vector $\boldsymbol{x}$, respectively, and $\boldsymbol{X}^\top$, $\|\boldsymbol{X}\|$, $\|\boldsymbol{X}\|_{\mathsf{F}}$ and $\|\boldsymbol{X}\|_1$ the transpose, the spectral norm, the Frobenius norm and the nuclear norm of a matrix $\boldsymbol{X}$, respectively. We denote the kth singular value of $\boldsymbol{X}$ by $\sigma_k(\boldsymbol{X})$, and the kth eigenvalue by $\lambda_k(\boldsymbol{X})$. Moreover, the inner product between two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \mathrm{Tr}\left(\boldsymbol{Y}^\top \boldsymbol{X}\right)$, where $\mathrm{Tr}\left(\cdot\right)$ is the trace. We also use $\mathrm{vec}(\boldsymbol{X})$ to denote vectorization of a matrix $\boldsymbol{X}$ in a column-major order. The (k,t)th entry of a matrix $\boldsymbol{X}$ is denoted by $\boldsymbol{X}_{k,t}$. For a vector $\boldsymbol{x}$, $\mathrm{med}(\boldsymbol{x})$ denotes the median of the entries in $\boldsymbol{x}$, and $|\boldsymbol{x}|$ denotes the vector that contains its entry-wise absolute values. $\mathbb{E}\left[\cdot\right]$ denotes the expectation operation with respect to an appropriate probability distribution. The indicator function of an event $\mathcal{E}$ is denoted by $\mathbb{I}_{\mathcal{E}}$, which equals 1 if $\mathcal{E}$ is true and 0 otherwise. The notation $a \ll b$ means the scalar a is much smaller than b, and the notation $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ means that there is a universal constant $c > 0$ such that $|f(n)| \leq c|g(n)|$. We use $:=$ for making definitions. In addition, we use c and C with different superscripts and subscripts to represent positive numerical constants, whose values may change from line to line.

Chapter 2: Nonconvex Matrix Recovery from Rank-One Measurements

This chapter is concerned with recovering a low-rank PSD matrix from random rank-one measurements. The results of this chapter are summarized in the paper submission [34].

2.1 Problem Formulation

To begin with, we present the formal problem setup. Specifically, we consider estimating a low-rank PSD matrix $\boldsymbol{M}^\natural$ from a few *rank-one measurements*. Suppose that the matrix of interest can be factorized as

$$\boldsymbol{M}^\natural = \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top} \in \mathbb{R}^{n\times n}, \tag{2.1}$$

where $\boldsymbol{X}^\natural \in \mathbb{R}^{n\times r}$ denotes the low-rank factor with $r \ll n$. We collect m measurements $\{y_i\}_{i=1}^m$ about $\boldsymbol{M}^\natural$ taking the form

$$y_i = \boldsymbol{a}_i^\top \boldsymbol{M}^\natural \boldsymbol{a}_i = \left\|\boldsymbol{a}_i^\top \boldsymbol{X}^\natural\right\|_2^2, \quad i = 1, \cdots, m, \tag{2.2}$$

where $\{\boldsymbol{a}_i\}_{i=1}^m$ represent the measurement vectors known *a priori*, of which $\boldsymbol{a}_i \in \mathbb{R}^n$ is the ith sensing vector composed of i.i.d. standard Gaussian entries, i.e. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, for $i = 1, \cdots, m$. One can think of $\{\boldsymbol{a}_i\boldsymbol{a}_i^\top\}_{i=1}^m$ as a set of linear sensing matrices (so

that $y_i = \langle \boldsymbol{a}_i \boldsymbol{a}_i^\top, \boldsymbol{M}^\natural \rangle$), which are all rank-one[1]. The underlying ground truth $\boldsymbol{X}^\natural$ is assumed to have full column rank but not necessarily having orthogonal columns. Define the condition number of $\boldsymbol{M}^\natural = \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}$ as

$$\kappa = \frac{\sigma_1^2\left(\boldsymbol{X}^\natural\right)}{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}. \tag{2.3}$$

Throughout this chapter, we assume the condition number is bounded by some constant independent of n and r, i.e. $\kappa = O(1)$. Our goal is to recover $\boldsymbol{X}^\natural$, up to (unrecoverable) orthonormal transformation, from the measurements $\boldsymbol{y} = \{y_i\}_{i=1}^m$ in a statistically and computationally efficient manner.

This problem spans a variety of important practical applications ranging from the covariance sketching scheme considered in [10] to array signal processing [35] and network traffic monitoring [36], from quantum state tomography [37] to compressive power spectrum estimation [38], and from non-coherent direction-of-arrival estimation based on magnitude measurements [39] to synthetic aperture radar imaging [40], with a few examples listed below.

- **Covariance sketching:** Consider a zero-mean data stream $\{\boldsymbol{x}_t\}_{t\in\mathcal{T}}$, whose covariance matrix $\boldsymbol{M}^\natural := \mathbb{E}[\boldsymbol{x}_t \boldsymbol{x}_t^\top]$ is (approximately) low-rank. To estimate the covariance matrix, one can collect m aggregated quadratic sketches of the form

$$y_i = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} (\boldsymbol{a}_i^\top \boldsymbol{x}_t)^2, \tag{2.4}$$

  which converges to $\mathbb{E}[(\boldsymbol{a}_i^\top \boldsymbol{x}_t)^2] = \boldsymbol{a}_i^\top \boldsymbol{M}^\natural \boldsymbol{a}_i$ as the number of data instances grows. This quadratic covariance sketching scheme can be performed under

[1]Given that y_i is a quadratic function with respect to both $\boldsymbol{X}^\natural$ and $\boldsymbol{a}_i$, the measurement scheme is also referred to as *quadratic sampling*.

minimal storage requirement and low sketching cost. See [10] for detailed descriptions.

- **Phase retrieval and mixed linear regression:** This problem subsumes as a special case the phase retrieval problem [25, 41, 42], which aims to estimate an unknown signal $\boldsymbol{x}^\natural \in \mathbb{R}^n$ from intensity measurements (which can often be modeled or approximated by quadratic measurements of the form $y_i = (\boldsymbol{a}_i^\top \boldsymbol{x}^\natural)^2$). This problem has found numerous applications in X-ray crystallography, optical imaging, astronomy, etc. Another related problem in machine learning is mixed linear regression with two components, where the data one collects are generated from one of two unknown regressors; see [43] for precise formulation.

- **Quantum state tomography:** Estimating the density operator of a quantum system can be formulated as a low-rank PSD matrix recovery problem using rank-one measurements, when the density operator is *almost pure* [17]. A problem of similar mathematical formulation occurs in phase space tomography [44], where the goal is to reconstruct the correlation function of a wave field.

- **Learning shallow polynomial neural networks:** Taking $\{\boldsymbol{a}_i, y_i\}_{i=1}^m$ as training data, our problem is equivalent to learning a one-hidden-layer fully-connected neural network with a quadratic activation function [45–47], where the output of the network is expressed as $y = \sum_{i=1}^r \sigma(\boldsymbol{a}^\top \boldsymbol{x}_i^\natural)$ with $\boldsymbol{X}^\natural = [\boldsymbol{x}_1^\natural, \boldsymbol{x}_2^\natural, \cdots, \boldsymbol{x}_r^\natural] \in \mathbb{R}^{n\times r}$ and the activation function $\sigma(z) = z^2$.

2.2 Vanilla Gradient Descent

Due to the quadratic nature of the measurements, the natural least-squares empirical risk formulation is highly nonconvex and in general challenging to solve. To be more specific, consider minimizing the squared loss:

$$f(\boldsymbol{X}) := \frac{1}{4m}\sum_{i=1}^{m}\left(y_i - \left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2\right)^2, \tag{2.5}$$

which aims to optimize a degree-4 polynomial in $\boldsymbol{X}$ and is NP hard in general. The problem, however, may become tractable under certain random designs, and may even be solvable using simple methods like gradient descent.

The algorithm studied herein is a combination of *vanilla gradient descent* and a judiciously designed spectral initialization. We attempt to optimize this function iteratively via gradient descent

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \mu_t \nabla f(\boldsymbol{X}_t), \qquad t = 0, 1, \cdots, \tag{2.6}$$

where $\boldsymbol{X}_t$ denotes the estimate in the tth iteration, μ_t is the step size/learning rate, and the gradient $\nabla f(\boldsymbol{X})$ is given by

$$\nabla f(\boldsymbol{X}) = \frac{1}{m}\sum_{i=1}^{m}\left(\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2 - y_i\right)\boldsymbol{a}_i\boldsymbol{a}_i^\top \boldsymbol{X}. \tag{2.7}$$

For initialization, similar to [48],[2] we apply the spectral method, which sets the columns of $\boldsymbol{X}_0$ as the top-r eigenvectors — properly scaled — of a matrix $\boldsymbol{Y}$ as defined in (2.8). The rationale is this: the mean of $\boldsymbol{Y}$ is given by

$$\mathbb{E}[\boldsymbol{Y}] = \frac{1}{2}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 \boldsymbol{I}_n + \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}, \tag{2.11}$$

[2]Compared with [48], when setting the eigenvalues in (2.9), we use the sample mean λ rather than $\lambda_{r+1}(\boldsymbol{Y})$ to estimate $\frac{1}{2}\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2$.

Algorithm 1: Gradient Descent with Spectral Initialization

Input: Measurements $\boldsymbol{y} = \{y_i\}_{i=1}^m$, and sensing vectors $\{\boldsymbol{a}_i\}_{i=1}^m$.

Parameters: Step size μ_t, rank r, and number of iterations T.

Initialization: Set $\boldsymbol{X}_0 = \boldsymbol{Z}_0 \boldsymbol{\Lambda}_0^{1/2}$, where the columns of $\boldsymbol{Z}_0 \in \mathbb{R}^{n\times r}$ contain the normalized eigenvectors corresponding to the r largest eigenvalues of the matrix

$$\boldsymbol{Y} = \frac{1}{2m}\sum_{i=1}^m y_i \boldsymbol{a}_i \boldsymbol{a}_i^\top, \tag{2.8}$$

and $\boldsymbol{\Lambda}_0$ is an $r \times r$ diagonal matrix, with the entries on the diagonal given as

$$[\boldsymbol{\Lambda}_0]_i = \lambda_i\big(\boldsymbol{Y}\big) - \lambda, \quad i = 1, \cdots, r, \tag{2.9}$$

where $\lambda = \frac{1}{2m}\sum_{i=1}^m y_i$ and $\lambda_i\left(\boldsymbol{Y}\right)$ is the ith largest eigenvalue of $\boldsymbol{Y}$.

Gradient loop: For $t = 0 : 1 : T-1$, do

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \mu_t \cdot \frac{1}{m}\sum_{i=1}^m \Big(\big\|\boldsymbol{a}_i^\top \boldsymbol{X}_t\big\|_2^2 - y_i\Big)\boldsymbol{a}_i \boldsymbol{a}_i^\top \boldsymbol{X}_t. \tag{2.10}$$

Output: $\boldsymbol{X}_T$.

and hence the principal components of $\boldsymbol{Y}$ form a reasonable estimate of $\boldsymbol{X}^\natural$, provided that there are sufficiently many samples. The full algorithm is described in Algorithm 1.

Before continuing, we demonstrate the effective and efficient performance of proposed algorithm with a numerical example. For each n, we generate an $n \times n$ PSD matrix $\boldsymbol{M}^\natural$ with rank $r = 5$ and all nonzero eigenvalues are equal to one, and set $m = 6nr$. Using a *constant* step size $\mu_t = 0.1$, the normalized recovery errors $\big\|\boldsymbol{X}_t\boldsymbol{X}_t^\top - \boldsymbol{M}^\natural\big\|_{\mathsf{F}} / \big\|\boldsymbol{M}^\natural\big\|_{\mathsf{F}}$ are shown in Figure 2.1 with respect to the iteration count, for $n = 100, 200, 500$ and 1000, respectively. These numerical results indicate that vanilla gradient descent (starting from an initial guess obtained by spectral method) exhibits remarkable linear convergence with a constant step size $\mu_t = 0.1$, although

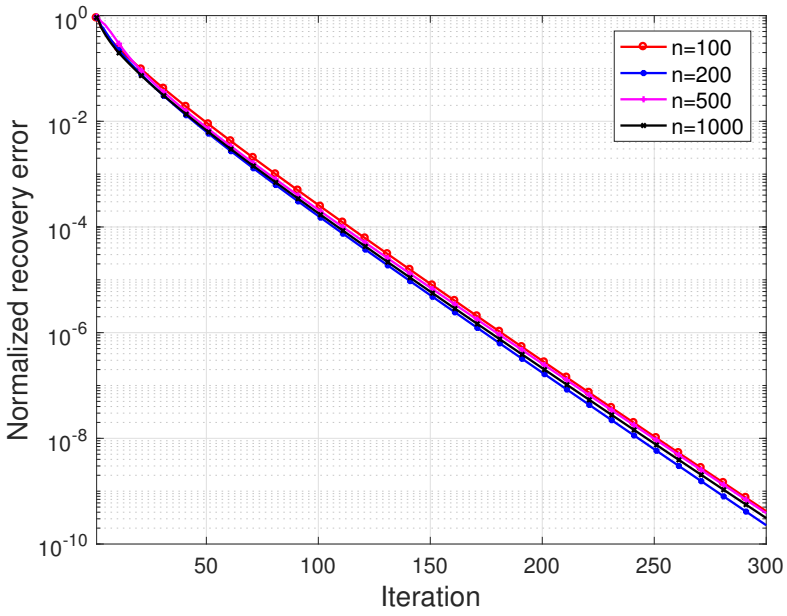


Figure 2.1: Normalized recovery error for low-rank PSD matrix recovery from rank-one measurements with respect to the iteration count in different problem sizes, when $r = 5$, $m = 6nr$ and $\mu_t = 0.1$.

the dimension of ground truth is varied from $n = 100$ to $n = 1000$. For all the cases, the normalized recovery error can be reduced to 10^{-8} within around 250 iterations. The convergence rate experiences only little changes even though the problem size varies. In comparisons, a conservative step size setting inversely proportional to n^4 is suggested in [48], which results in an overly pessimistic convergence rate, especially for large matrix recovery problems.

2.3 Performance Guarantees

Before proceeding to our main results, we pause here to introduce the metric used to assess the estimation error of the running iterates. Since $\left(\boldsymbol{X}^{\natural}\boldsymbol{P}\right)\left(\boldsymbol{X}^{\natural}\boldsymbol{P}\right)^{\top} =$

$\boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}$ for any orthonormal matrix $\boldsymbol{P} \in \mathbb{R}^{r\times r}$, $\boldsymbol{X}^\natural$ is recoverable up to orthonormal transforms. Hence, we define the error of the tth iterate $\boldsymbol{X}_t$ as

$$\operatorname{dist}\big(\boldsymbol{X}_t, \boldsymbol{X}^\natural\big) = \big\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\big\|_{\mathsf{F}}, \tag{2.12}$$

where $\boldsymbol{Q}_t$ is given by

$$\boldsymbol{Q}_t := \operatorname{argmin}_{\boldsymbol{P}\in\mathcal{O}^{r\times r}} \big\|\boldsymbol{X}_t\boldsymbol{P} - \boldsymbol{X}^\natural\big\|_{\mathsf{F}} \tag{2.13}$$

with $\mathcal{O}^{r\times r}$ denoting the set of all $r \times r$ orthonormal matrices. Accordingly, we have the following theoretical performance guarantees of Algorithm 1.

Theorem 1. *Suppose that $m \geq cnr^4\kappa^3 \log n$ with some large enough constant $c > 0$, and that the step size obeys $0 < \mu_t := \mu = \frac{c_4}{(r\kappa+\log n)^2\sigma_r^2(\boldsymbol{X}^\natural)}$. Then with probability at least $1 - O(mn^{-7})$, the iterates satisfy*

$$\operatorname{dist}\big(\boldsymbol{X}_t, \boldsymbol{X}^\natural\big) \leq c_1 \left(1 - 0.5\mu\sigma_r^2(\boldsymbol{X}^\natural)\right)^t \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}}, \tag{2.14}$$

for all $t \geq 0$. In addition,

$$\max_{1\leq l\leq m} \big\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right)\big\|_2 \leq c_2 \left(1 - 0.5\mu\sigma_r^2(\boldsymbol{X}^\natural)\right)^t \sqrt{\log n} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}}, \tag{2.15}$$

holds for all $0 \leq t \leq c_3 n^5$. Here, $c_1, \cdots, c_4$ are some universal positive constants.

Remark 1. The precise expression of required sample complexity in Theorem 1 can be written as $m \geq c\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^6}{\sigma_r^6(\boldsymbol{X}^\natural)} nr\log(\kappa n)$ with some large enough constant $c > 0$. By adjusting constants, with probability at least $1 - O(mn^{-7})$, (2.15) holds for $0 \leq t \leq O(n^{c_5})$ in any power $c_5 \geq 1$.

Theorem 1 has the following implications.

- **Near-optimal sample complexity when r is fixed:** Theorem 1 suggests that spectrally-initialized vanilla gradient descent succeeds as soon as $m = O(nr^4 \log n)$. When $r = O(1)$, this leads to near-optimal sample complexity up to logarithmic factor. In fact, once the spectral initialization is finished, a sample complexity at $m = O(nr^3 \log n)$ can guarantee the linear convergence to the global optima. To the best of our knowledge, this outperforms all performance guarantees in the literature obtained for any nonconvex method without requiring *resampling*.

- **Near-optimal computational complexity:** In order to achieve ϵ-accuracy, i.e. $\mathrm{dist}\left(\boldsymbol{X}_t, \boldsymbol{X}^\natural\right) \le \epsilon \|\boldsymbol{X}\|_{\mathsf{F}}$, it suffices to run gradient descent for

$$T = O\left(r^2 \mathrm{poly}\log(n) \log(1/\epsilon)\right) \tag{2.16}$$

  iterations. This results in a total computational complexity of

$$C = O(mnr \cdot T) = O\left(mnr^3 \mathrm{poly}\log(n) \log(1/\epsilon)\right). \tag{2.17}$$

  When r is fixed independent of m and n, the computational complexity scales linearly with mn (up to logarithmic factors), which is proportional to the time taken to read all data.

- **Implicit regularization:** Theorem 1 demonstrates that both the spectral initialization and the gradient descent updates provably control the entry-wise error $\max_{1\le l\le m} \left\|\boldsymbol{a}_l^\top \left(\boldsymbol{X}_t \boldsymbol{Q}_t - \boldsymbol{X}^\natural\right)\right\|_2$, and the iterates remain incoherent with respect to all the sensing vectors. In fact, the entry-wise error decreases linearly as well, which is not characterized in [49].

These findings significantly improve upon existing results that require either resampling (which is not sample-efficient and is not the algorithm one actually runs in practice [47, 50, 51]), or high iteration complexity (which results in high computation cost [48]), of which the statistical and computational guarantees to reach ϵ-accuracy are summarized in Table 2.1. We note our guarantee is the only one that achieves simultaneous near-optimal sample complexity and computational complexity.

Algorithms with resampling	Sample complexity	Computational complexity
AltMin-LRROM [50]	$O(nr^4 \log^2 n \log(\frac{1}{\epsilon}))$	$O(mnr \log(\frac{1}{\epsilon}))$
gFM [51]	$O(nr^3 \log(\frac{1}{\epsilon}))$	$O(mnr \log(\frac{1}{\epsilon}))$
EP-ROM [47]	$O(nr^2 \log^4 n \log(\frac{1}{\epsilon}))$	$O(mn^2 \log(\frac{1}{\epsilon}))$
AP-ROM [47]	$O(nr^3 \log^4 n \log(\frac{1}{\epsilon}))$	$O(mnr \log n \log(\frac{1}{\epsilon}))$
Algorithms without resampling	**Sample complexity**	**Computational complexity**
Convex [10]	$O(nr)$	$O(mn^2 \frac{1}{\sqrt{\epsilon}})$
GD [48]	$O(nr^6 \log^2 n)$	$O(mn^5 r^3 \log^4 n \log(\frac{1}{\epsilon}))$
GD (Algorithm 1, Ours)	$O(nr^4 \log n)$	$O(mnr \max\{\log^2 n, r^2\} \log(\frac{1}{\epsilon}))$

Table 2.1: Comparisons with existing results in terms of sample complexity and computational complexity to reach ϵ-accuracy. The top half of the table is concerned with algorithms that require resampling, while the bottom half of the table covers algorithms without resampling.

Theorem 1 is established using a fixed step size. According to our theoretical analysis, the incoherence condition (2.15) has a significant impact on the convergence rate. After a few iterations, the incoherence condition can be bounded independent of $\log n$, which suggests a larger step size and, therefore, faster convergence. Specifically, we have the following corollary.

Corollary 1. *Under the same setting of Theorem 1, after* $T_a = c_6 \max\{\kappa^2 r^2 \log n, \log^3 n\}$ *iterations, the step size can be relaxed as* $0 < \mu_t := \mu = \frac{c_7}{r^2\kappa^2\sigma_r^2(\boldsymbol{X}^\natural)}$*, with some universal constant* $c_6, c_7 > 0$*, then the iterates satisfy*

$$\operatorname{dist}\big(\boldsymbol{X}_t, \boldsymbol{X}^\natural\big) \le c_1 \left(1 - 0.5\mu\sigma_r^2(\boldsymbol{X}^\natural)\right)^t \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}, \tag{2.18}$$

for all $t \ge T_a$*, with probability at least* $1 - O(mn^{-7})$*.*

Remark 2. when $r = 1$, Corollary 1 will degenerate to the phase retrieval case which indicates that with sample complexity $m = O(n \log n)$, after certain iterations, a substantially more aggressive step size is feasible that is independent on the signal dimension n.

2.4 Surprising Effectiveness of Gradient Descent

Recently, gradient descent has been widely employed to address various nonconvex optimization problems due to its appealing efficiency from both statistical and computational perspectives. Despite the nonconvexity of natural least-squares empirical risk minimization

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) := \frac{1}{4m}\sum_{i=1}^{m}\left(y_i - \left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2\right)^2, \tag{2.19}$$

[48] showed that within a local neighborhood of $\boldsymbol{X}^\natural$, where $\boldsymbol{X}$ satisfies

$$\left\|\boldsymbol{X} - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \le \frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}, \tag{2.20}$$

$f(\boldsymbol{X})$ behaves like a strongly convex function, at least along certain descending directions. However, this region itself is not enough to guarantee computational efficiency, and consequently, the smoothness parameter derived in [48] is as large as n^2 (even

ignoring additional polynomial factors in r), leading to a step size as small as $O(1/n^4)$ and an iteration complexity of $O(n^4 \log(1/\epsilon))$. These are fairly pessimistic.

In order to improve computational guarantees, it might be tempting to employ appropriately designed regularization operations — such as truncation [52] and projection [53]. These explicit regularization operations are capable of stabilizing the search direction, and make sure the whole trajectory is in a basin of attraction with benign curvatures surrounding the ground truth. However, such explicit regularizations complicate algorithm implementations, as they typically introduce more tuning parameters.

Our work is inspired by [49], which uncovers the "implicit regularization" phenomenon of vanilla gradient descent for nonconvex estimation problems such as phase retrieval and low-rank matrix completion. In words, even without extra regularization operations, vanilla gradient descent always follows a path within some region around the global optimum with nice geometric structure, at least along certain directions. The current chapter demonstrates that a similar phenomenon persists in low-rank matrix recovery from rank-one measurements.

To describe this phenomenon in a precise manner, we need to specify which region enjoys the desired geometric properties. To this end, consider a local region around $\boldsymbol{X}^\natural$ where $\boldsymbol{X}$ is "incoherent"[3] with all sensing vectors in the following sense:

$$\max_{1\le l\le m} \left\|\boldsymbol{a}_l^\top \left(\boldsymbol{X}-\boldsymbol{X}^\natural\right)\right\|_2 \le \frac{1}{24}\sqrt{\log n}\cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}. \tag{2.21}$$

We term the intersection of (2.20) and (2.21) the *region of incoherence and contraction* (RIC). The nice feature of the RIC is this: within this region, the loss function $f(\boldsymbol{X})$

[3]This is called incoherent because if $\boldsymbol{X}$ is aligned (and hence coherent) with the sensing vectors, $\left\|\boldsymbol{a}_l^\top \left(\boldsymbol{X}-\boldsymbol{X}^\natural\right)\right\|_2$ can be $O(\sqrt{n})$ times larger than the right-hand side of (2.21).

enjoys a smoothness parameter that scales as $O(\max\{r, \log n\})$ (namely, $\|\nabla^2 f(\boldsymbol{x})\| \lesssim \max\{r, \log n\}$, which is much smaller than $O(n^2)$ provided in [48]). As is well known, a region enjoying a smaller smoothness parameter enables more aggressive progression of gradient descent.

A key question remains as to how to prove that the trajectory of gradient descent never leaves the RIC. This is, unfortunately, not guaranteed by standard optimization theory, which only ensures contraction of the Euclidean error. Rather, we need to exploit the statistical model of data generation, taking into consideration of the "homogeneity" of the samples together with the finite-sum form of the loss function. To address this issue, we resort to the leave-one-out trick [49, 54, 55] that produces auxiliary trajectories of gradient descent that use all but one sample. This allows us to establish the incoherence condition by leveraging the statistical independence of the leave-one-out trajectory with respect to the corresponding sensing vector that has been left out. Our theory refines the leave-one-out argument and further establishes linear contraction in terms of the entry-wise prediction error. In sum, our work highlights the substantial gain of jointly considering optimization and statistics in understanding learning algorithms.

2.5 Related Work

Instead of directly estimating $\boldsymbol{X}^\natural$, the problem of interest can be also solved by estimating $\boldsymbol{M}^\natural = \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}$ in higher dimension via nuclear norm minimization, which requires $O(nr)$ measurements for exact recovery [10, 17, 18]. See also [13, 14, 56–58] for the phase retrieval problem. However, nuclear norm minimization, often cast as

the semidefinite programming, is in general computationally expensive to deal with large-scale data.

On the other hand, nonconvex approaches have drawn intense attention in the past decade due to their ability to achieve computational and statistical efficiency all at once. Specifically, for the phase retrieval problem, Wirtinger flow (WF) and its variants [25, 49, 52, 59–62] have been proposed. As a *two-stage* algorithm, it consists of spectral initialization and iterative gradient updates. This strategy has found enormous success in solving other problems such as low-rank matrix recovery and completion [53, 63], blind deconvolution [64], and spectral compressed sensing [65]. We follow a similar route but analyze a more general problem that includes phase retrieval as a special case.

The work [48] is most close to ours, which studied the local convexity of the same loss function and developed performance guarantees for gradient descent using a similar, but different spectral initialization scheme. As discussed earlier, due to the pessimistic estimate of the smoothness parameter, they only allow a diminishing learning rate (or step size) of $O(1/n^4)$, leading to a high iteration complexity. We not only provide stronger computational guarantees, but also improve the sample complexity, compared with [48].

Several other existing works have suggested different approaches for low-rank PSD matrix recovery from rank-one measurements, including AltMin-LRROM [50], gFM [51], and AP-ROM and EP-ROM [47]. Comparing with these existing results as shown in Table 2.1, to the best of our knowledge, our work is the first nonconvex algorithm (without resampling) that achieves both near-optimal statistical and computational guarantees with respect to n. Iterative algorithms based on alternating minimization

or noisy power iterations [47, 50, 51] require a *fresh* set of samples at every iteration, which is never executed in practice, and the sample complexity grows unbounded for *exact* recovery.

Many nonconvex methods have been proposed and analyzed recently to solve the phase retrieval problem, including alternating minimization [66], the Kaczmarz method [67–69] and approximate message passing [70]. In [71], the Kaczmarz method is generalized to solve the problem studied in our work, but no theoretical performance guarantees are provided.

The local geometry studied in our work is in contrast to [72], which studied the global landscape of phase retrieval, and showed that there are no spurious local minima as soon as the sample complexity is above $O(n \log^3 n)$. It will be interesting to study the landscape property of the generalized model in our work.

Our model is also related to learning shallow neural networks. [73] studied the performance of gradient descent with resampling and an initialization provided by the tensor method for various activation functions, however their analysis did not cover quadratic activations. For quadratic activations, [45] adopts a greedy learning strategy, and can only guarantee sublinear convergence rate. Moreover, [46] studied the optimization landscape for an over-parameterized shallow neural network with quadratic activation, where r is larger than n.

2.6 Outline of Theoretical Analysis

This section provides the proof sketch of the main results, with the details deferred to the appendix. Our theoretical analysis is inspired by the work of [49] for phase retrieval and follows the general recipe outlined in [49], while significant changes and

elaborate derivations are needed. We refine the analysis to show that both the signal reconstruction error and the entry-wise error contract linearly, where the latter is not revealed by [49]. In below, we first characterize a RIC that enjoys both strong convexity and smoothness along certain directions. We then demonstrate — via an induction argument — that the iterates always stay within this nice region. Finally, the proof is complete by validating the desired properties of spectral initialization.

2.6.1 Local Geometry and Error Contraction

We start with characterizing a local region around $\boldsymbol{X}^\natural$, within which the loss function enjoys desired restricted strong convexity and smoothness properties. This requires exploring the property of the Hessian of $f(\boldsymbol{X})$, which is given by

$$\nabla^2 f(\boldsymbol{X}) = \frac{1}{m}\sum_{i=1}^{m}\left[\left(\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2 - y_i\right)\boldsymbol{I}_r + 2\boldsymbol{X}^\top \boldsymbol{a}_i\boldsymbol{a}_i^\top \boldsymbol{X}\right]\otimes\left(\boldsymbol{a}_i\boldsymbol{a}_i^\top\right). \tag{2.22}$$

Here, we use $\otimes$ to denote the Kronecker product and hence $\nabla^2 f(\boldsymbol{X}) \in \mathbb{R}^{nr\times nr}$. Now we are ready to state the following lemma regarding this local region, which will be referred to as the RIC throughout this chapter. The proof is given in Appendix B.1.

Lemma 1. *Suppose the sample size obeys* $m \geq c\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^4}{\sigma_r^4(\boldsymbol{X}^\natural)}nr\log(n\kappa)$ *for some sufficiently large constant* $c > 0$. *Then with probability at least* $1 - c_1 n^{-12} - me^{-1.5n} - mn^{-12}$, *we have*

$$\operatorname{vec}(\boldsymbol{V})^\top \nabla^2 f(\boldsymbol{X})\operatorname{vec}(\boldsymbol{V}) \geq 1.026\sigma_r^2(\boldsymbol{X}^\natural)\|\boldsymbol{V}\|_{\mathsf{F}}^2, \tag{2.23}$$

and

$$\left\|\nabla^2 f(\boldsymbol{X})\right\| \leq 1.5\sigma_r^2(\boldsymbol{X}^\natural)\log n + 6\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 \tag{2.24}$$

hold simultaneously for all matrices $\boldsymbol{X}$ *and* $\boldsymbol{V}$ *satisfying the following constraints:*

$$\left\|\boldsymbol{X} - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \leq \frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}; \tag{2.25a}$$

$$\max_{1\le l\le m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}-\boldsymbol{X}^\natural\right)\right\|_2\le\frac{1}{24}\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}, \tag{2.25b}$$

and $\boldsymbol{V}=\boldsymbol{T}_1\boldsymbol{Q}_{\boldsymbol{T}}-\boldsymbol{T}_2$ *satisfying*

$$\left\|\boldsymbol{T}_2-\boldsymbol{X}^\natural\right\|\le\frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|}, \tag{2.26}$$

where $\boldsymbol{Q}_{\boldsymbol{T}}:=\operatorname{argmin}_{\boldsymbol{P}\in\mathcal{O}^{r\times r}}\|\boldsymbol{T}_1\boldsymbol{P}-\boldsymbol{T}_2\|_{\mathsf{F}}$. *Here,* c_1 *is some absolute positive constant.*

The condition (2.25) on $\boldsymbol{X}$ formally characterizes the RIC, which enjoys the claimed restricted strong convexity (see (2.23)) and smoothness (see (2.24)). With Lemma 1 in mind, it is easy to see that if $\boldsymbol{X}_t$ lies within the RIC, the estimation error shrinks in the presence of a properly chosen step size. This is given in the lemma below whose proof can be found in Appendix B.2.

Lemma 2. *Suppose the sample size obeys* $m\ge c\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^4}{\sigma_r^4\left(\boldsymbol{X}^\natural\right)}nr\log\left(n\kappa\right)$ *for some sufficiently large constant* $c>0$. *Then with probability at least* $1-c_1n^{-12}-me^{-1.5n}-mn^{-12}$, *if* $\boldsymbol{X}_t$ *falls within the RIC as described in* (2.25), *we have*

$$\operatorname{dist}\left(\boldsymbol{X}_{t+1},\boldsymbol{X}^\natural\right)\le\left(1-0.513\mu\sigma_r^2(\boldsymbol{X}^\natural)\right)\operatorname{dist}\left(\boldsymbol{X}_t,\boldsymbol{X}^\natural\right),$$

provided that the step size obeys $0<\mu_t:=\mu\le\frac{1.026\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left(1.5\sigma_r^2(\boldsymbol{X}^\natural)\log n+6\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2\right)^2}$. *Here,* $c_1>0$ *is some universal constant.*

Assuming that the iterates $\{\boldsymbol{X}_t\}$ stay within the RIC (see (2.25)) for the first T_c iterations, according to Lemma 2, we have, by induction, that

$$\operatorname{dist}\left(\boldsymbol{X}_{T_c+1},\boldsymbol{X}^\natural\right)\le\left(1-0.513\mu\sigma_r^2(\boldsymbol{X}^\natural)\right)^{T_c+1}\operatorname{dist}\left(\boldsymbol{X}_0,\boldsymbol{X}^\natural\right)\le\frac{1}{24\sqrt{6}}\cdot\frac{\sqrt{\log n}}{\sqrt{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}$$

as soon as

$$T_c\ge c\max\left\{\log^2 n,\frac{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^4}{\sigma_r^4\left(\boldsymbol{X}^\natural\right)}\right\}\log n, \tag{2.27}$$

for some large enough constant c. The iterates when $t \geq T_c$ are easier to deal with; in fact, it is easily seen that $\boldsymbol{X}_{t+1}$ stays in the RIC since

$$\begin{aligned}\max_{1\leq l\leq m} \left\|\boldsymbol{a}_l^\top \left(\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^\natural\right)\right\|_2 &\leq \max_{1\leq l\leq m} \left\|\boldsymbol{a}_l\right\|_2 \left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^\natural\right\| \\ &\leq \sqrt{6n}\cdot\frac{1}{24\sqrt{6}}\cdot\frac{\sqrt{\log n}}{\sqrt{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} \qquad (2.28) \\ &= \frac{1}{24}\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}},\end{aligned}$$

where (2.28) follows from Lemma 16 for all $t \geq T_c$. Consequently, contraction of the estimation error $\mathrm{dist}\big(\boldsymbol{X}_t, \boldsymbol{X}^\natural\big)$ can be guaranteed by Lemma 1 for all $t \geq T_c$ with probability at least $1 - c_1 n^{-12} - me^{-1.5n} - mn^{-12}$.

2.6.2 Introducing Leave-One-Out Sequences

It has now become clear that the key remaining step is to verify that the iterates $\{\boldsymbol{X}_t\}$ satisfy (2.25) for the first T_c iterations, where T_c is on the order of (2.27). Verifying (2.25b) is conceptually hard since the iterates $\{\boldsymbol{X}_t\}$ are statistically dependent with all the sensing vectors $\{\boldsymbol{a}_i\}_{i=1}^m$. One may turn to some generic bounding techniques such as Cauchy-Schwarz inequality, which, however, usually would not yield a tight enough bound. To tackle this problem, for each $1 \leq l \leq m$, we introduce an auxiliary leave-one-out sequence $\{\boldsymbol{X}_t^{(l)}\}$, which discards a single measurement from consideration. Specifically, the sequence $\{\boldsymbol{X}_t^{(l)}\}$ is the gradient iterates operating on the following leave-one-out function

$$f^{(l)}\left(\boldsymbol{X}\right) := \frac{1}{4m}\sum_{i:i\neq l}\left(y_i - \left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2\right)^2. \qquad (2.29)$$

See Algorithm 2 for a formal definition of the leave-one-out sequences. Again, we want to emphasize that Algorithm 2 is just an auxiliary procedure useful for the theoretical analysis, and it does not need to be implemented in practice.

Algorithm 2: Leave-One-Out Versions

Input: Measurements $\{y_i\}_{i:i\neq l}$, and sensing vectors $\{\boldsymbol{a}_i\}_{i:i\neq l}$.

Parameters: Step size μ_t, rank r, and number of iterations T.

Initialization: $\boldsymbol{X}_0^{(l)} = \boldsymbol{Z}_0^{(l)} \boldsymbol{\Lambda}_0^{(l)1/2}$, where the columns of $\boldsymbol{Z}_0^{(l)} \in \mathbb{R}^{n\times r}$ contain the normalized eigenvectors corresponding to the r largest eigenvalues of the matrix

$$\boldsymbol{Y}^{(l)} = \frac{1}{2m}\sum_{i:i\neq l} y_i \boldsymbol{a}_i \boldsymbol{a}_i^\top, \tag{2.30}$$

and $\boldsymbol{\Lambda}_0^{(l)}$ is an $r\times r$ diagonal matrix, with the entries on the diagonal given as

$$\left[\boldsymbol{\Lambda}_0^{(l)}\right]_i = \lambda_i\big(\boldsymbol{Y}^{(l)}\big) - \lambda^{(l)}, \quad i = 1,\cdots,r, \tag{2.31}$$

where $\lambda^{(l)} = \frac{1}{2m}\sum_{i:i\neq l} y_i$ and $\lambda_i\big(\boldsymbol{Y}^{(l)}\big)$ is the ith largest eigenvalue of $\boldsymbol{Y}^{(l)}$.

Gradient loop: For $t = 0:1:T-1$, do

$$\boldsymbol{X}_{t+1}^{(l)} = \boldsymbol{X}_t^{(l)} - \mu_t \cdot \frac{1}{m}\sum_{i:i\neq l}\left(\left\|\boldsymbol{a}_i^\top \boldsymbol{X}_t^{(l)}\right\|_2^2 - y_i\right)\boldsymbol{a}_i\boldsymbol{a}_i^\top \boldsymbol{X}_t^{(l)}. \tag{2.32}$$

Output: $\boldsymbol{X}_T^{(l)}$.

2.6.3 Establishing Incoherence via Induction

Our proof is inductive in nature with the following induction hypotheses:

$$\left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \le C_1\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}; \tag{2.33a}$$

$$\max_{1\le l\le m}\left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}\right\|_{\mathsf{F}} \le C_3\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t \sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}; \tag{2.33b}$$

$$\max_{1\le l\le m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right)\right\|_2 \le C_2\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t \sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}, \tag{2.33c}$$

where $\boldsymbol{R}_t^{(l)} = \mathrm{argmin}_{\boldsymbol{P}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}_t^{(l)}\boldsymbol{P}\right\|_{\mathsf{F}}$, and the positive constants C_1, C_2 and C_3 satisfy

$$C_1 + C_3 \le \frac{1}{24}, \quad C_2 + \sqrt{6}C_3 \le \frac{1}{24}, \quad 5.86C_1 + 5.86C_3 + \sqrt{6}C_3 \le C_2. \tag{2.34}$$

Furthermore, the step size μ is chosen as

$$\mu = \frac{c_0 \sigma_r^2 \left(\boldsymbol{X}^\natural\right)}{\left(\sigma_r^2(\boldsymbol{X}^\natural)\log n + \|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2\right)^2} \tag{2.35}$$

with appropriate universal constant $c_0 > 0$.

Our goal is to show that if the tth iteration $\boldsymbol{X}_t$ satisfies the induction hypotheses (2.33), the $(t+1)$th iteration $\boldsymbol{X}_{t+1}$ also satisfies (2.33). It is straightforward to see that the hypothesis (2.33a) has already been established by Lemma 2, and we are left with (2.33b) and (2.33c). We first establish (2.33b) in the following lemma, which measures the proximity between $\boldsymbol{X}_t$ and the leave-one-out versions $\boldsymbol{X}_t^{(l)}$, whose proof is provided in Appendix B.3.

Lemma 3. *Suppose the sample size obeys* $m \geq c\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^4}{\sigma_r^4(\boldsymbol{X}^\natural)} nr\log(n\kappa)$ *for some sufficiently large constant* $c > 0$. *If the induction hypotheses* (2.33) *hold for the* t*th iteration, with probability at least* $1 - c_1 n^{-12} - me^{-1.5n} - mn^{-12}$, *we have*

$$\max_{1\leq l\leq m} \left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right\|_{\mathsf{F}} \leq C_3\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^{t+1}\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}},$$

as long as the step size satisfies (2.35). *Here,* $c_1 > 0$ *is some absolute constant.*

In addition, the incoherence property of $\boldsymbol{X}_{t+1}^{(l)}$ with respect to the lth sensing vector $\boldsymbol{a}_l$ is relatively easier to establish, due to their statistical independence. Combined with the proximity bound from Lemma 3, this allows us to justify the incoherence property of the original iterates $\boldsymbol{X}_{t+1}$, as summarized in the lemma below, whose proof is given in Appendix B.4.

Lemma 4. *Suppose the sample size obeys* $m \geq c\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^4}{\sigma_r^4(\boldsymbol{X}^\natural)} nr\log(n\kappa)$ *for some sufficiently large constant* $c > 0$. *If the induction hypotheses* (2.33) *hold for the* t*th*

iteration, with probability exceeding $1 - c_1 n^{-12} - me^{-1.5n} - 2mn^{-12}$,

$$\max_{1\le l\le m} \left\|\boldsymbol{a}_l^\top \left(\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^\natural\right)\right\|_2 \le C_2 \left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^{t+1} \sqrt{\log n}\cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}$$

holds as long as the step size satisfies (2.35). *Here,* $c_1 > 0$ *is some universal constant.*

2.6.4 Spectral Initialization

Finally, it remains to verify that the induction hypotheses hold for the initialization, i.e. the base case when $t = 0$. This is supplied by the following lemma, whose proof is given in Appendix B.5.

Lemma 5. *Suppose that the sample size exceeds* $m \ge c\frac{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^6}{\sigma_r^6\left(\boldsymbol{X}^\natural\right)} nr\log n$ *for some sufficiently large constant* $c > 0$. *Then* $\boldsymbol{X}_0$ *satisfies* (2.33) *with probability at least* $1 - c_1 n^{-12} - me^{-1.5n} - 3mn^{-12}$, *where* c_1 *is some absolute positive constant.*

2.7 Conclusion

In this chapter, we have shown that low-rank PSD matrices can be recovered from a near-minimal number of random rank-one measurements, via the vanilla gradient descent algorithm following spectral initialization. Our results significantly improve upon existing results in several ways, both computationally and statistically. In particular, our algorithm does not require resampling at every iteration (and hence requires fewer samples). The gradient iteration can provably employ a much more aggressive step size than what was suggested in prior literature (e.g. [48]), thus resulting in much smaller iteration complexity and hence lower computational cost. All of this is enabled by establishing the implicit regularization feature of gradient descent for nonconvex statistical estimation, where the iterates remain incoherent with the sensing vectors throughout the execution of the whole algorithm.

Chapter 3: Robust Matrix Recovery from Corrupted Rank-One Measurements

As discussed in Chapter 2, in many emerging applications of science and engineering, we are interested in estimating a low-rank PSD matrix $\boldsymbol{X}_0 \in \mathbb{R}^{n\times n}$ from a set of nonnegative magnitude measurements:

$$y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}_0 \rangle = \langle \boldsymbol{a}_i \boldsymbol{a}_i^\top, \boldsymbol{X}_0 \rangle = \boldsymbol{a}_i^\top \boldsymbol{X}_0 \boldsymbol{a}_i, \tag{3.1}$$

for $i = 1, \cdots, m$. The measurement y_i is quadratic in the sensing vector $\boldsymbol{a}_i \in \mathbb{R}^n$, but linear in $\boldsymbol{X}_0$, where the sensing matrix $\boldsymbol{A}_i = \boldsymbol{a}_i \boldsymbol{a}_i^\top$ is *rank-one*.

In this chapter, we focus on robust recovery of the low-rank PSD matrix when the measurements in (3.1) are further corrupted by outliers, possibly adversarial with arbitrary amplitudes. In practical applications, outliers are somewhat inevitable, which may be caused by sensor failures, malicious attacks, or reading errors [31–33]. In the application of covariance sketching, described in (2.4), a sufficient aggregation length $|\mathcal{T}|$ is necessary in order for each measurement to be well approximated by (3.1). Measurements which are not aggregated from a large enough $|\mathcal{T}|$ may be regarded as outliers. Therefore, it becomes critical to address robust recovery of $\boldsymbol{X}_0$ in the presence of outliers. Fortunately, it is reasonable to assume that the number of outliers is usually much smaller than the number of total measurements, making it

possible to leverage the sparsity of the outliers to faithfully recover the low-rank PSD matrix of interest. The results of this chapter are summarized in the papers [74, 75].

3.1 Problem Formulation

Let $\boldsymbol{X}_0 \in \mathbb{R}^{n\times n}$ be a rank-r PSD matrix, then the set of m measurements, which may be corrupted by either arbitrary outliers or bounded noise, can be represented as

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_0) + \boldsymbol{\eta} + \boldsymbol{w} \in \mathbb{R}^m, \tag{3.2}$$

The linear mapping $\mathcal{A}$: $\mathbb{R}^{n\times n} \to \mathbb{R}^m$ is defined as $\mathcal{A}(\boldsymbol{X}_0) := \{\mathcal{A}_i(\boldsymbol{X}_0)\}_{i=1}^m$ with $\mathcal{A}_i(\boldsymbol{X}_0) := \boldsymbol{a}_i^\top \boldsymbol{X}_0 \boldsymbol{a}_i$, where $\boldsymbol{a}_i \in \mathbb{R}^n$ is the ith sensing vector composed of i.i.d. standard Gaussian entries, for $i = 1, \cdots, m$. The vector $\boldsymbol{\eta} \in \mathbb{R}^m$ denotes the outlier vector, which is assumed to be sparse whose entries can be arbitrarily large. The fraction of nonzero entries is defined as $s := \|\boldsymbol{\eta}\|_0 / m$. Moreover, the vector $\boldsymbol{w} \in \mathbb{R}^m$ denotes the additive noise, which is assumed bounded as $\|\boldsymbol{w}\|_1 \leq \epsilon$. Our goal is to robustly recover $\boldsymbol{X}_0$ from the measurements $\boldsymbol{y}$.

3.2 Robust Recovery via Convex Relaxation

3.2.1 Robust-PhaseLift

To motivate our algorithm, consider the case when only the outlier vector $\boldsymbol{\eta}$ is present in (3.2) and the rank of $\boldsymbol{X}_0$ is known. One may seek a rank-r PSD matrix that minimizes the cardinality of the measurement residual to motivate outlier sparsity, given as

$$\hat{\boldsymbol{X}} = \text{argmin}_{\boldsymbol{X}\succeq 0} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_0, \quad \text{s.t.} \quad \text{rank}(\boldsymbol{X}) = r. \tag{3.3}$$

However, both the cardinality minimization and the rank constraint are NP-hard in general, making this method computationally infeasible. A common approach is to resort to convex relaxation, where we relax the cardinality minimization by its convex relaxation, i.e. the ℓ_1-norm, and meanwhile, drop the rank constraint, yielding:

$$\text{(Robust-PhaseLift:)} \quad \hat{\boldsymbol{X}} = \text{argmin}_{\boldsymbol{X}\succeq 0}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_1. \tag{3.4}$$

We denote the above convex program as the Robust-PhaseLift algorithm, since it coincides with the PhaseLift algorithm studied in [14, 56, 76] for phase retrieval[4]. The advantage of Robust-PhaseLift in (3.4) is that it does not require any prior knowledge of the noise bound, the rank of $\boldsymbol{X}_0$, nor the sparsity level of the outliers, and is free of any regularization parameter. It is worth emphasizing that in (3.4) only the PSD constraint of the solution is honored without explicitly motivating the low-rank structure, via for example, trace minimization[5].

3.2.2 Performance Guarantees

Encouragingly, we demonstrate that the algorithm (3.4) admits robust recovery of a rank-r PSD matrix as soon as the number of measurements is large enough, even with a fraction of arbitrary outliers in Theorem 2. To the best of our knowledge, this is the first theoretical performance guarantee of the robustness of (3.4) with respect to arbitrary outliers in the low-rank setting. Our main theorem is given as below.

Theorem 2. *Suppose that* $\|\boldsymbol{w}\|_1 \leq \epsilon$ *and* $s = \|\boldsymbol{\eta}\|_0/m$. *Assume the support of* $\boldsymbol{\eta}$ *is selected uniformly at random with the signs of its nonzero entries generated from*

[4]Note that there are a few different versions of PhaseLift in the literature which are not outlier-robust, therefore we rename (3.4) to Robust-PhaseLift for emphasis.

[5]The interested readers are invited to look up Figure 1 in [13] for an intuitive geometric interpretation in the noise-free and outlier-free case.

the Rademacher distribution as $\mathbb{P}\{\mathrm{sgn}(\eta_i) = -1\} = \mathbb{P}\{\mathrm{sgn}(\eta_i) = 1\} = 1/2$ *for each* $i \in \mathrm{supp}(\boldsymbol{\eta})$. *Then for a fixed rank-r PSD matrix* $\boldsymbol{X}_0 \in \mathbb{R}^{n\times n}$, *there exist some absolute constants* $c_1 > 0$ *and* $0 < s_0 < 1$ *such that as long as*

$$m \geq c_1 n r^2, \quad s \leq \frac{s_0}{r},$$

the solution to (3.4) *satisfies*

$$\left\|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\right\|_{\mathsf{F}} \leq c_2 \frac{r\epsilon}{m},$$

with probability exceeding $1 - \exp(-\gamma m/r^2)$ *for some constants* c_2 *and* γ.

Theorem 2 has the following consequences.

- **Exact Recovery with Outliers:** When $\epsilon = 0$, Theorem 2 suggests the recovery is exact using Robust-PhaseLift (3.4), i.e. $\hat{\boldsymbol{X}} = \boldsymbol{X}_0$ even when a fraction of measurements are arbitrarily corrupted, as long as the number of measurements m is on the order of nr^2. Given there are at least nr unknowns, our measurement complexity is near-optimal up to a factor of r.

- **Stable Recovery with Bounded Noise:** In the presence of bounded noise, Theorem 2 suggests that the recovery performance decreases gracefully with the increase of ϵ, where the Frobenius norm of the reconstruction error is proportional to the per-entry noise level of the measurements.

- **Phase Retrieval:** When $r = 1$, the problem degenerates to the case of phase retrieval, and Theorem 2 recovers existing results in [76] for outlier-robust phase retrieval, where the measurement complexity is on the order of n, which is optimal up to a scaling factor.

Let us denote $\hat{\boldsymbol{X}}_r = \text{argmin}_{\text{rank}(\boldsymbol{Z})=r,\ \boldsymbol{Z}\succeq 0}\|\hat{\boldsymbol{X}} - \boldsymbol{Z}\|_{\mathsf{F}}$ as the best rank-r PSD matrix approximation of $\hat{\boldsymbol{X}}$, the solution to (3.4). Then Theorem 2 suggests that the estimate $\hat{\boldsymbol{X}}$ can be well approximated by a rank-r PSD matrix since

$$\|\hat{\boldsymbol{X}} - \hat{\boldsymbol{X}}_r\|_{\mathsf{F}} \leq \|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\mathsf{F}} \leq c_2\frac{r\epsilon}{m},$$

as long as the number of measurements is sufficiently large. Furthermore, we have

$$\|\hat{\boldsymbol{X}}_r - \boldsymbol{X}_0\|_{\mathsf{F}} \leq \|\hat{\boldsymbol{X}}_r - \hat{\boldsymbol{X}}\|_{\mathsf{F}} + \|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\mathsf{F}} \leq 2\|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\mathsf{F}} \leq 2c_2\frac{r\epsilon}{m},$$

indicating that $\hat{\boldsymbol{X}}_r$ provides an accurate estimate of $\boldsymbol{X}_0$ that is both exactly rank-r and PSD.

3.3 Related Work

In the absence of outliers, the PhaseLift algorithm in the following form

$$\min_{\boldsymbol{X}\succeq 0} \text{Tr}(\boldsymbol{X}) \quad \text{s.t.} \quad \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_1 \leq \epsilon, \tag{3.5}$$

has been proposed to solve the phase retrieval problem [13, 14, 77]. Later the same algorithm has been employed to recover low-rank PSD matrices in [10], where an order of nr measurements obtained from i.i.d. sub-Gaussian sensing vectors are shown to guarantee exact recovery in the noise-free case and stable recovery with bounded noise. One problem with the algorithm (3.5) is that the noise bound ϵ is assumed known. Furthermore, it is not amenable to handle outliers, since $\|\boldsymbol{y}-\mathcal{A}(\boldsymbol{X}_0)\|_1$ can be arbitrarily large with outliers and consequently the ground truth $\boldsymbol{X}_0$ quickly becomes infeasible for (3.5).

The proposed algorithm (3.4) is studied in [14,56,76] as a variant of PhaseLift for phase retrieval, corresponding to the case where $\boldsymbol{X}_0 = \boldsymbol{x}_0\boldsymbol{x}_0^\top$ is rank-one. It is shown

in [14, 56] that with $O(n)$ i.i.d. Gaussian sensing vectors, the algorithm succeeds with high probability. Compared with (3.5), the algorithm (3.4) eliminates trace minimization and leads to easier algorithm implementations. We note that [78] also considers a regularization-free algorithm for PSD matrix estimation that minimizes the ℓ_2-norm of the residual, which unfortunately, cannot handle outliers as Robust-PhaseLift (3.4). Our work is related to the work in [76] which first considered the robustness of the Robust-PhaseLift algorithm (3.4) in the presence of outliers for phase retrieval, establishing that the same guarantee holds even with a constant fraction of outliers. Our work extends the performance guarantee in [76] to the general low-rank PSD matrix case. Moreover, we show the proposed approach can be easily extended to recover low-rank Toeplitz PSD matrices via numerical experiments.

Broadly speaking, our problem is related to low-rank matrix recovery from an under-determined linear system [5, 79, 80], where the linear measurements are drawn from inner products with rank-one sensing matrices. It is due to this special structure of the sensing matrices that we can eliminate the trace minimization, and only consider the feasibility constraint for PSD matrices. Standard approaches for separating low-rank and sparse components [81–85] via convex optimization are given as

$$\min_{\boldsymbol{X}\succeq 0,\ \boldsymbol{\eta}} \mathrm{Tr}(\boldsymbol{X}) + \lambda\|\boldsymbol{\eta}\|_1, \quad \text{s.t.} \quad \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X}) - \boldsymbol{\eta}\|_1 \leq \epsilon,$$

where λ is a regularization parameter that requires to be tuned properly. In contrast, our algorithm is parameter-free.

3.4 Theoretical Analysis of Robust-PhaseLift

In this section we prove Theorem 2, and the roadmap of our proof is below. In Section 3.4.1, we first provide the sufficient conditions for an approximate dual certificate

that certifies the optimality of the proposed algorithm (3.4) in Lemma 6. Section 3.4.2 records a few lemmas that show $\mathcal{A}$ satisfies the required restricted isometry properties. Then, a dual certificate is constructed and validated for a fixed low-rank PSD matrix $\boldsymbol{X}_0$ in Section 3.4.3. Finally, the proof is concluded in Section 3.4.4.

First we introduce some additional notations. Let $\mathcal{S}$ be a subset of $\{1, 2, \cdots, m\}$, then $\mathcal{S}^{\perp}$ is the complement of $\mathcal{S}$ with respect to $\{1, 2, \cdots, m\}$. $\mathcal{A}_{\mathcal{S}}$ is the mapping operator $\mathcal{A}$ constrained on $\mathcal{S}$, which transforms a matrix $\boldsymbol{X}$ into a vector $\mathcal{A}_{\mathcal{S}}(\boldsymbol{X})$ whose entries equal $\boldsymbol{a}_i^{\top}\boldsymbol{X}\boldsymbol{a}_i$ for $i \in \mathcal{S}$ and are zero otherwise. Denote the adjoint operator of $\mathcal{A}$ by $\mathcal{A}^*(\boldsymbol{\mu}) = \sum_{i=1}^{m} \mu_i \boldsymbol{a}_i \boldsymbol{a}_i^{\top}$, where μ_i is the ith entry of $\boldsymbol{\mu}$, $1 \leq i \leq m$. Let the singular value decomposition of the fixed rank-r PSD matrix $\boldsymbol{X}_0$ be $\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\top}$, then the symmetric tangent space T at $\boldsymbol{X}_0$ is denoted by

$$T := \left\{ \boldsymbol{U}\boldsymbol{Z}^{\top} + \boldsymbol{Z}\boldsymbol{U}^{\top} \mid \boldsymbol{Z} \in \mathbb{R}^{n \times r} \right\}. \tag{3.6}$$

We denote by $\mathcal{P}_T$ and $\mathcal{P}_{T^{\perp}}$ the orthogonal projection onto T and its orthogonal complement, respectively. And for notational simplicity, we denote $\boldsymbol{H}_T := \mathcal{P}_T(\boldsymbol{H})$ and $\boldsymbol{H}_{T^{\perp}} := \boldsymbol{H} - \mathcal{P}_T(\boldsymbol{H})$ for any symmetric matrix $\boldsymbol{H} \in \mathbb{R}^{n \times n}$. Moreover, γ represents an absolute constant, whose value may change according to context.

3.4.1 Approximate Dual Certificate

The following lemma suggests that under certain appropriate restricted isometry preserving properties of $\mathcal{A}$, a properly constructed dual certificate can guarantee faithful recovery of the proposed algorithm (3.4), of which the proof is deferred to Appendix C.1.

Lemma 6 (Approximate Dual Certificate for (3.4))**.** *Denote a subset* $\mathcal{S}$ *with* $\frac{|\mathcal{S}|}{m} := \lceil \frac{s_0}{13\sqrt{2r}} \rceil$*, where* $0 < s_0 < 1$ *is some constant, and the support of* $\boldsymbol{\eta}$ *satisfies* $\operatorname{supp}(\boldsymbol{\eta}) \subseteq$

$\mathcal{S}$. *Suppose that the mapping* $\mathcal{A}$ *obeys that for all symmetric matrices* $\boldsymbol{X}$,

$$\frac{1}{m}\left\|\mathcal{A}\left(\boldsymbol{X}\right)\right\|_1 \leq \left(1+\frac{1}{10}\right)\left\|\boldsymbol{X}\right\|_1, \tag{3.7}$$

and

$$\frac{1}{|\mathcal{S}|}\left\|\mathcal{A}_{\mathcal{S}}\left(\boldsymbol{X}\right)\right\|_1 \leq \left(1+\frac{1}{10}\right)\left\|\boldsymbol{X}\right\|_1, \tag{3.8}$$

and for all matrices $\boldsymbol{X} \in T$,

$$\frac{1}{|\mathcal{S}^{\perp}|}\left\|\mathcal{A}_{\mathcal{S}^{\perp}}\left(\boldsymbol{X}\right)\right\|_1 > \frac{1}{5}\left(1-\frac{1}{12}\right)\left\|\boldsymbol{X}\right\|_{\mathsf{F}}. \tag{3.9}$$

Then if there exists a matrix $\boldsymbol{Y} = \mathcal{A}^*(\boldsymbol{\mu})$ *that satisfies*

$$\boldsymbol{Y}_{T^{\perp}} \preceq -\frac{1}{r}\boldsymbol{I}_{T^{\perp}}, \quad \left\|\boldsymbol{Y}_T\right\|_{\mathsf{F}} \leq \frac{1}{13r}, \tag{3.10}$$

and

$$\begin{cases} \mu_i = \frac{9}{m}\mathrm{sgn}(\eta_i), & i \in \mathrm{supp}(\boldsymbol{\eta}) \\ |\mu_i| \leq \frac{9}{m}, & i \notin \mathrm{supp}(\boldsymbol{\eta}) \end{cases}, \tag{3.11}$$

the solution to (3.4) *satisfies*

$$\left\|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\right\|_{\mathsf{F}} \leq c\frac{r\epsilon}{m}, \tag{3.12}$$

where c *is a constant.*

3.4.2 Restricted Isometry of $\mathcal{A}$

The first two conditions (3.7) and (3.8) in Lemma 6 are supplied straightforwardly in the following lemma as long as $m \geq cnr$ and $|\mathcal{S}| = c_1 m/r \geq c_2 n$ for some constants c, c_1 and c_2.

Lemma 7. *[13] Fix any* $\delta \in (0, \frac{1}{2})$ *and assume* $m \geq 20\delta^{-2}n$. *Then for all PSD matrices* $\boldsymbol{X}$, *one has*

$$(1-\delta)\left\|\boldsymbol{X}\right\|_1 \leq \frac{1}{m}\left\|\mathcal{A}\left(\boldsymbol{X}\right)\right\|_1 \leq (1+\delta)\left\|\boldsymbol{X}\right\|_1$$

with probability exceeding $1-2e^{-m\epsilon^2/2}$*, where* $\epsilon^2+\epsilon=\frac{\delta}{4}$*. The right hand side holds for all symmetric matrices.*

The third condition (3.9) in Lemma 6 can be obtained using the mixed-norm RIP-ℓ_2/ℓ_1 provided in [10] as long as $m \geq cnr$ and $|\mathcal{S}| \leq c_1 m$ for some constants c and c_1.

Lemma 8. *[10] Suppose the sensing vectors* $\boldsymbol{a}_i$*'s are composed of i.i.d. sub-Gaussian entries, then there exist positive universal constants* c_1*,* c_2 *and* c_3 *such that, provided that* $m > c_3 nr$*, for all matrices* $\boldsymbol{X}$ *of rank at most* r*, one has*

$$\left(1-\delta_r^{\mathrm{lb}}\right)\|\boldsymbol{X}\|_{\mathsf{F}} \leq \frac{2}{m}\|\mathcal{B}(\boldsymbol{X})\|_1 \leq \left(1+\delta_r^{\mathrm{ub}}\right)\|\boldsymbol{X}\|_{\mathsf{F}},$$

with probability exceeding $1-c_1e^{-c_2 m}$*, where* δ_r^{lb} *and* δ_r^{ub} *are defined as the RIP-*ℓ_2/ℓ_1 *constants. And the operator* $\mathcal{B}$ *represents the linear transformation that maps* $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ *to* $\{\mathcal{B}_i(\boldsymbol{X})\}_{i=1}^{m/2} \in \mathbb{R}^{m/2}$*, where* $\mathcal{B}_i(\boldsymbol{X}) := \langle \boldsymbol{a}_{2i-1}\boldsymbol{a}_{2i-1}^\top - \boldsymbol{a}_{2i}\boldsymbol{a}_{2i}^\top, \boldsymbol{X}\rangle$.

The third condition (3.9) can be easily validated from the lower bound by setting δ_r^{lb} appropriately, since $\|\mathcal{B}(\boldsymbol{X})\|_1 \leq \sum_{i=1}^{m/2}\left(\left|\langle \boldsymbol{a}_{2i-1}\boldsymbol{a}_{2i-1}^\top, \boldsymbol{X}\rangle\right| + \left|\langle \boldsymbol{a}_{2i}\boldsymbol{a}_{2i}^\top, \boldsymbol{X}\rangle\right|\right) = \|\mathcal{A}(\boldsymbol{X})\|_1$.

3.4.3 Construction of Dual Certificate

For notational simplicity, let $\alpha_0 := \mathbb{E}\left[Z^2\mathbb{I}_{\{|Z|\leq 3\}}\right] \approx 0.9707$, $\beta_0 := \mathbb{E}\left[Z^4\mathbb{I}_{\{|Z|\leq 3\}}\right] \approx 2.6728$ and $\theta_0 := \mathbb{E}\left[Z^6\mathbb{I}_{\{|Z|\leq 3\}}\right] \approx 11.2102$, where Z is a standard Gaussian random variable.

Consider that the singular value decomposition of a PSD matrix $\boldsymbol{X}_0$ of rank at most r can be represented as $\boldsymbol{X}_0 = \sum_{i=1}^{r}\lambda_i \boldsymbol{u}_i\boldsymbol{u}_i^\top$, then inspired by [14, 76], we

construct $\boldsymbol{Y}$ as

$$\begin{aligned}\boldsymbol{Y} &:= \frac{1}{m}\sum_{j\in\mathcal{S}^\perp}\Big[\frac{1}{r}\sum_{i=1}^{r}\left|\boldsymbol{a}_j^\top\boldsymbol{u}_i\right|^2\mathbb{I}_{\left\{\left|\boldsymbol{a}_j^\top\boldsymbol{u}_i\right|\le 3\right\}}-\big(\alpha_0+\frac{\beta_0-\alpha_0}{r}\big)\Big]\cdot\boldsymbol{a}_j\boldsymbol{a}_j^\top+\frac{9}{m}\sum_{j\in\mathcal{S}}\chi_j\boldsymbol{a}_j\boldsymbol{a}_j^\top\\ &:= \boldsymbol{Y}^{(0)}-\boldsymbol{Y}^{(1)}+\boldsymbol{Y}^{(2)},\end{aligned}\tag{3.13}$$

where

$$\boldsymbol{Y}^{(0)}=\frac{1}{m}\sum_{j\in\mathcal{S}^\perp}\left[\frac{1}{r}\sum_{i=1}^{r}\left|\boldsymbol{a}_j^\top\boldsymbol{u}_i\right|^2\mathbb{I}_{\left\{\left|\boldsymbol{a}_j^\top\boldsymbol{u}_i\right|\le 3\right\}}\right]\boldsymbol{a}_j\boldsymbol{a}_j^\top;\tag{3.14}$$

$$\boldsymbol{Y}^{(1)}=\frac{1}{m}\left(\alpha_0+\frac{\beta_0-\alpha_0}{r}\right)\sum_{j\in\mathcal{S}^\perp}\boldsymbol{a}_j\boldsymbol{a}_j^\top;\tag{3.15}$$

$$\boldsymbol{Y}^{(2)}=\frac{9}{m}\sum_{j\in\mathcal{S}}\chi_j\boldsymbol{a}_j\boldsymbol{a}_j^\top.\tag{3.16}$$

We set $\chi_j=\operatorname{sgn}(\eta_j)$ if $j\in\operatorname{supp}(\boldsymbol{\eta})$, otherwise χ_j's are i.i.d. Rademacher random variables with $\mathbb{P}\{\chi_j=1\}=\mathbb{P}\{\chi_j=-1\}=1/2$.

The construction immediately indicates that $\boldsymbol{Y}$ satisfies (3.11). Then, we can also show that $\boldsymbol{Y}$ satisfies (3.10) with high probability by separating the constructed $\boldsymbol{Y}$ into two parts and considering the bounds on $\boldsymbol{Y}^{(0)}-\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$, respectively. The theoretical results are summarized in following Lemma 9, which proves $\boldsymbol{Y}_{T^\perp}+\frac{1}{r}\boldsymbol{I}_{T^\perp}\preceq 0$ and Lemma 10, which proves $\|\boldsymbol{Y}_T\|_{\mathsf{F}}\le\frac{1}{13r}$, of which the proofs are given in Appendix C.2 and Appendix C.3, respectively.

Lemma 9. *Provided $m\ge cnr^2$ and $|\mathcal{S}|=c_1m/r\ge c_2nr$ for some constants c, c_1 and c_2, with probability at least $1-e^{-\gamma m/r^2}$, we have*

$$\left\|\boldsymbol{Y}_{T^\perp}+\frac{1.7}{r}\boldsymbol{I}_{T^\perp}\right\|\le\frac{0.25}{r}.$$

Lemma 10. *Provided $m\ge cnr^2$ and $|\mathcal{S}|=c_1m/r$, for some constants c and c_1, with probability at least $1-e^{-\gamma m/r^2}$, we have*

$$\|\boldsymbol{Y}_T\|_{\mathsf{F}}\le\frac{1}{15r}.$$

3.4.4 Proving Performance Guarantees of Robust-PhaseLift

The required restricted isometry properties of the linear mapping $\mathcal{A}$ are supplied in Section 3.4.2 and a valid appropriate dual certificate is constructed in Section 3.4.3, therefore, Theorem 2 can be straightforwardly obtained from the Lemma 6 in Section 3.4.1.

3.5 A Nonconvex Subgradient Descent Algorithm

In this section, we propose another algorithm for robust low-rank PSD matrix recovery from corrupted rank-one measurements assuming the rank (or its upper bound) of the PSD matrix $\boldsymbol{X}_0$ is known a priori as r. In this case, as in Chapter 2, we can decompose $\boldsymbol{X}_0$ as $\boldsymbol{X}_0 = \boldsymbol{U}_0\boldsymbol{U}_0^\top$ where $\boldsymbol{U}_0 \in \mathbb{R}^{n\times r}$ is the low-rank factor. Instead of directly recovering $\boldsymbol{X}_0$, we may aim at recovering $\boldsymbol{U}_0$ up to orthogonal transforms, since $(\boldsymbol{U}_0\boldsymbol{Q})(\boldsymbol{U}_0\boldsymbol{Q})^\top = \boldsymbol{U}_0\boldsymbol{U}_0$ for any orthonormal matrix $\boldsymbol{Q} \in \mathbb{R}^{r\times r}$. Consider relaxing of the loss function in (3.3) but keeping the rank constraint, then we obtain the following problem:

$$\hat{\boldsymbol{X}} = \text{argmin}_{\boldsymbol{X}\succeq 0}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_1, \quad \text{s.t.} \quad \text{rank}(\boldsymbol{X}) = r. \tag{3.17}$$

Since any rank-r PSD matrix $\boldsymbol{X}$ can be written as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^\top$ for some $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, (3.17) can be equivalently reformulated as

$$\hat{\boldsymbol{U}} = \text{argmin}_{\boldsymbol{U}\in\mathbb{R}^{n\times r}} f(\boldsymbol{U}), \tag{3.18}$$

with

$$f(\boldsymbol{U}) := \frac{1}{2m}\left\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top)\right\|_1 = \frac{1}{2m}\sum_{i=1}^{m}\left|y_i - \left\|\boldsymbol{U}^\top\boldsymbol{a}_i\right\|_2^2\right|.$$

Clearly, (3.18) is no longer convex. To illustrate, the first row of Figure 3.1 plots the value of the objective function in the negative logarithmic scale, i.e. $-\log f(\boldsymbol{U})$, under

different corruption scenarios when $\boldsymbol{U} \in \mathbb{R}^{2\times 1}$. For comparison, the second row of Figure 3.1 shows the loss function evaluated in ℓ_2-norm: $g(\boldsymbol{U}) = \frac{1}{4m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top)\|_2^2$, which is not robust to outliers.

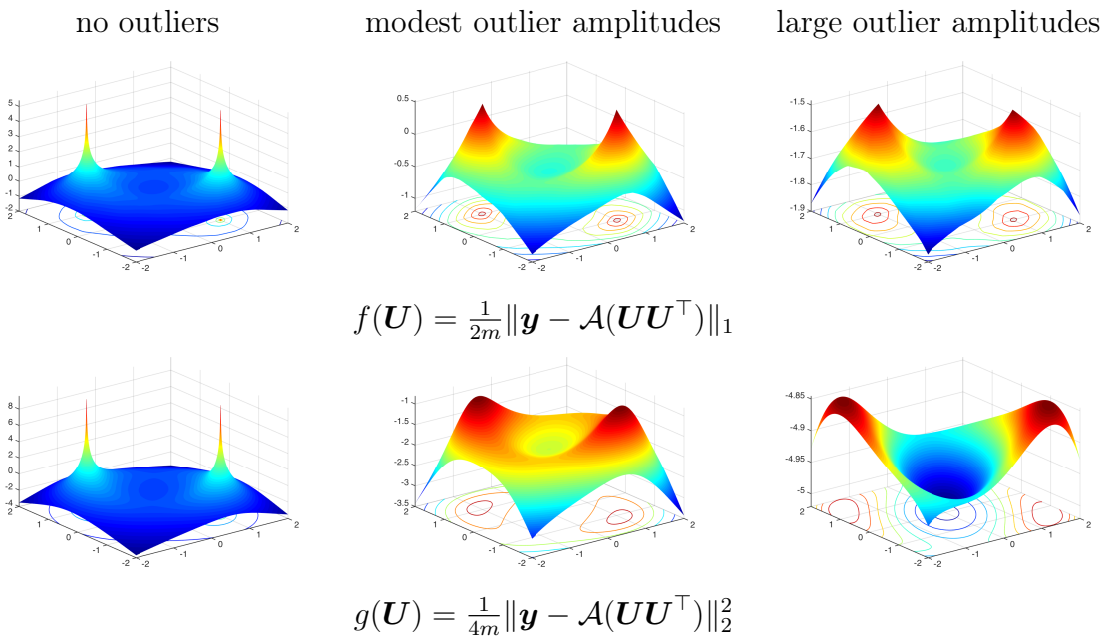


Figure 3.1: Illustrations of the objective function $-\log f(\boldsymbol{U})$ and its ℓ_2-norm counterpart $-\log g(\boldsymbol{U})$ (in negative logarithmic scales) under different corruption scenarios when $\boldsymbol{U} \in \mathbb{R}^{2\times 1}$. The number of measurements is $m = 100$ with i.i.d. Gaussian sensing vectors, and the fraction of outliers is $s = 0.2$ with uniformly selected support and amplitudes drawn from Unif$[0, 10]$ or Unif$[0, 100]$. It is interesting to observe that while large outliers completely distort $g(\boldsymbol{U})$, the proposed objective is quite robust with the ground truth being the only global optima of $f(\boldsymbol{U})$.

Motivated by the recent nonconvex approaches [25,48,52] of solving quadratic systems, we propose a *subgradient descent* algorithm to solve (3.18) effectively, working with a non-smooth function $f(\boldsymbol{U})$. Note that a subgradient of $f(\boldsymbol{U})$ with respect to

$\boldsymbol{U}$ can be given as

$$\partial f(\boldsymbol{U}) = -\frac{1}{m}\sum_{i=1}^{m} \operatorname{sgn}\left(y_i - \left\|\boldsymbol{U}^\top \boldsymbol{a}_i\right\|_2^2\right) \boldsymbol{a}_i \boldsymbol{a}_i^\top \boldsymbol{U}, \tag{3.19}$$

where the sign function $\operatorname{sgn}(\cdot)$ is defined as

$$\operatorname{sgn}(x) = \begin{cases} +1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}. \tag{3.20}$$

Our subgradient descent algorithm proceeds as below. Denote the estimate in the tth iteration by $\boldsymbol{U}^{(t)} \in \mathbb{R}^{n\times r}$. First, $\boldsymbol{U}^{(0)}$ is initialized as the best rank-r approximation of the following matrix with respect to Frobenius norm as

$$\boldsymbol{U}^{(0)}\left(\boldsymbol{U}^{(0)}\right)^\top = \operatorname{argmin}_{\operatorname{rank}(\boldsymbol{X})=r} \left\| \boldsymbol{X} - \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \right\|_{\mathsf{F}}^2. \tag{3.21}$$

Secondly, at the $(t+1)$th iteration, $t \geq 0$, we apply subgradient descent to refine the estimate as

$$\boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t)} - \mu_t \cdot \partial f(\boldsymbol{U}^{(t)}), \tag{3.22}$$

where the step size μ_t is adaptively set as

$$\mu_t = 0.05 \times \max\left\{2^{-t/1000}, 10^{-6}\right\},$$

which provides more accurate estimates using fewer iterations in the numerical simulations. The procedure is summarized in Algorithm 3, where the stopping rule in Algorithm 3 is simply put as a maximum number of iterations.

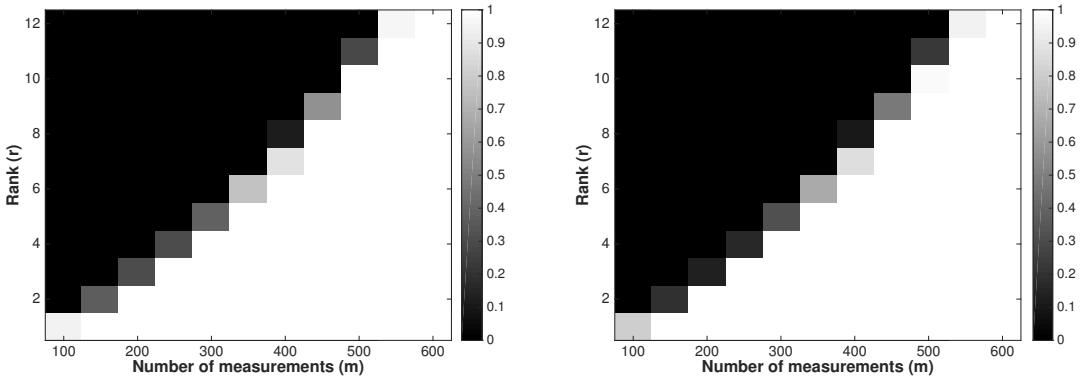
The main advantage of Algorithm 3 is its low memory and computational complexity. Given that it does not construct the full PSD matrix, the memory complexity is simply the size of $\boldsymbol{U}^{(t)}$, which is on the order of nr. The computational complexity per iteration is also low, which is on the order of mnr, that is linear in all the parameters. We demonstrate the excellent empirical performance of Algorithm 3 in Section 3.6.3.

Algorithm 3: Subgradient descent for solving (3.18)

Parameters: Rank r, number of iterations $T_{\max}$, and step size μ_t.
Input: Measurements $\boldsymbol{y}$, and sensing vectors $\{\boldsymbol{a}_i\}_{i=1}^m$.
Initialization: Initialize $\boldsymbol{U}^{(0)} \in \mathbb{R}^{n\times r}$ via (3.21).
For $t = 0 : 1 : T_{\max} - 1$ **do**
    update $\boldsymbol{U}^{(t+1)}$ via (3.22),
end for
Output: $\hat{\boldsymbol{U}} = \boldsymbol{U}^{(T_{\max})}$.

3.6 Numerical Examples

3.6.1 Performance of Convex Relaxation



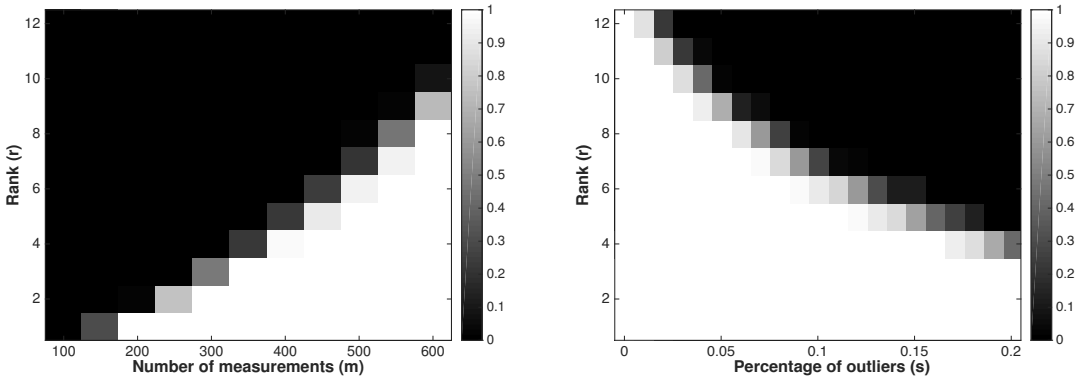
(a) (b)

Figure 3.2: Phase transitions of low-rank PSD matrix recovery with respect to the number of measurements and the rank, (a) with trace minimization; and (b) without trace minimization of noise-free measurements, when $n = 40$.

We first examine the performance of Robust-PhaseLift in (3.4). Let $n = 40$. We randomly generate a low-rank PSD matrix of rank-r as $\boldsymbol{X}_0 = \boldsymbol{U}_0\boldsymbol{U}_0^\top$, where $\boldsymbol{U}_0 \in \mathbb{R}^{n\times r}$ is composed of i.i.d. standard Gaussian variables. The sensing vectors are

also composed of i.i.d. standard Gaussian variables. Each Monte Carlo simulation is called successful if the normalized estimate error satisfies $\|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\mathsf{F}} / \|\boldsymbol{X}_0\|_{\mathsf{F}} \leq 10^{-6}$, where $\hat{\boldsymbol{X}}$ denotes the solution to (3.4). For each cell, the success rate is calculated by averaging over 100 Monte Carlo simulations.

Figure 3.2 shows the success rates of algorithms with respect to the number of measurements and the rank, with the trace minimization as in (3.5) in (a); and without the trace minimization as proposed in Robust-PhaseLift (3.4) in (b) for noise-free measurements. It can be seen that the performance of these two algorithms are almost equivalent, confirming a similar numerical observation for the phase retrieval problem [57] also holds in the low-rank setting, where trace minimization may be eliminated for low-rank PSD matrix recovery using rank-one measurements.



(a) (b)

Figure 3.3: Phase transitions of low-rank PSD matrix recovery with respect to (a) the number of measurements and the rank, with 5% of measurements corrupted by standard Gaussian variables; (b) the percentage of outliers and the rank, when the number of measurements is $m = 600$, when $n = 40$.

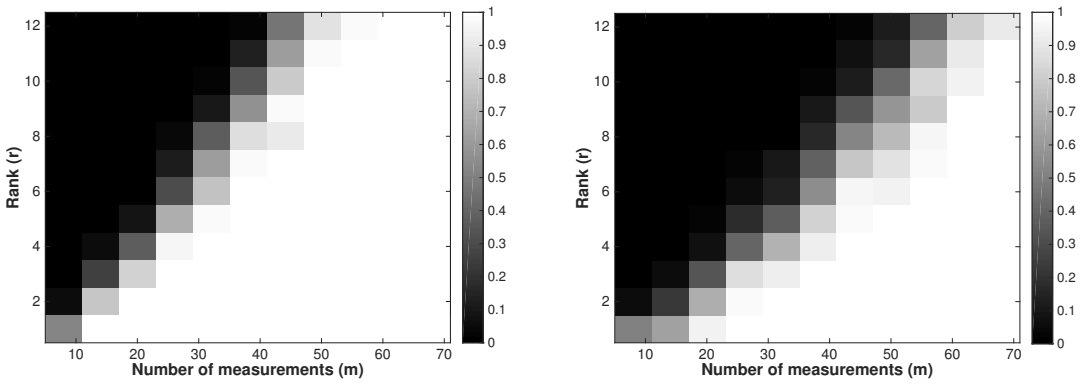
Figure 3.3 further shows the success rates of the Robust-PhaseLift algorithm (a) with respect to the number of measurements and the rank, when 5% of measurements are selected uniformly at random and corrupted by standard Gaussian variables; and (b) with respect to the percentage of outliers and the rank, for a fixed number of measurements $m = 600$. This also suggests possible room for improvements of our theoretical guarantee, as the numerical results indicate that the required measurement complexity for successful recovery has a seemingly linear relationship with r.

3.6.2 Convex Relaxation with Additional Toeplitz Structure

We next consider robust recovery of low-rank Toeplitz PSD matrices, where we allow complex-valued sensing vectors $\mathcal{A}(\boldsymbol{X}) = \{\boldsymbol{a}_i^H \boldsymbol{X} \boldsymbol{a}_i\}_{i=1}^m$ and complex-valued Toeplitz PSD matrices $\boldsymbol{X}$ with $(\cdot)^H$ denoting the Hermitian transpose. Estimating low-rank Toeplitz PSD matrices is of great interests for array signal processing [86]. We modify (3.4) by incorporating the Toeplitz constraint as:

$$\hat{\boldsymbol{X}} = \text{argmin}_{\boldsymbol{X} \succeq 0} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_1, \quad \text{s.t.} \quad \boldsymbol{X} \text{ is Toeplitz.} \tag{3.23}$$

Let $n = 64$, the Toeplitz PSD matrix $\boldsymbol{X}_0$ is generated as $\boldsymbol{X}_0 = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^H$, where $\boldsymbol{V} = [\boldsymbol{v}(f_1), \boldsymbol{v}(f_2), \cdots, \boldsymbol{v}(f_r)] \in \mathbb{C}^{n \times r}$ is a Vandermonde matrix with $\boldsymbol{v}(f_i) = [1, e^{j2\pi f_i}, \cdots, e^{j2\pi(n-1)f_i}]^\top$, $f_i \sim \text{Unif}[0,1]$, and $\boldsymbol{\Sigma} = \text{diag}\{\sigma_1^2, \cdots, \sigma_r^2\}$, with $\sigma_i^2 \sim \text{Unif}[0,1]$. Figure 3.4 shows the phase transitions of Toeplitz PSD matrix recovery with respect to the number of measurements and the rank without outliers in (a), and when 5% of measurements are selected uniformly at random and corrupted by standard Gaussian variables in (b). It can be seen that the low-rank Toeplitz PSD matrix can be robustly recovered from a sublinear number of measurements due to the additional Toeplitz structure. We note that a different covariance sketching scheme is considered



(a) (b)

Figure 3.4: Phase transitions of low-rank Toeplitz PSD matrix recovery with respect to the number of measurements and the rank, (a) without outliers, and (b) with 5% of measurements corrupted by standard Gaussian variables, when $n = 64$.

in [87–89] for estimating low-rank Toeplitz covariance matrices. Though not directly comparable to our measurement scheme, it may benefit from a similar parameter-free convex optimization to handle outliers.

3.6.3 Performance of Nonconvex Subgradient Descent

We next examine the performance of the nonconvex subgradient descent algorithm in Algorithm 3, where the number of iterations is set as $T_{\max} = 3 \times 10^4$, which is a large value to guarantee convergence when terminated. Denote the solution to Algorithm 3 by $\hat{\boldsymbol{U}}$, and each Monte Carlo simulation is deemed successful if the normalized estimate error satisfies $\|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\mathsf{F}}/\|\boldsymbol{X}_0\|_{\mathsf{F}} \leq 10^{-6}$, where $\hat{\boldsymbol{X}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^\top$ is the estimated low-rank PSD matrix. For each cell, the success rate is calculated by averaging over 100 Monte Carlo simulations.

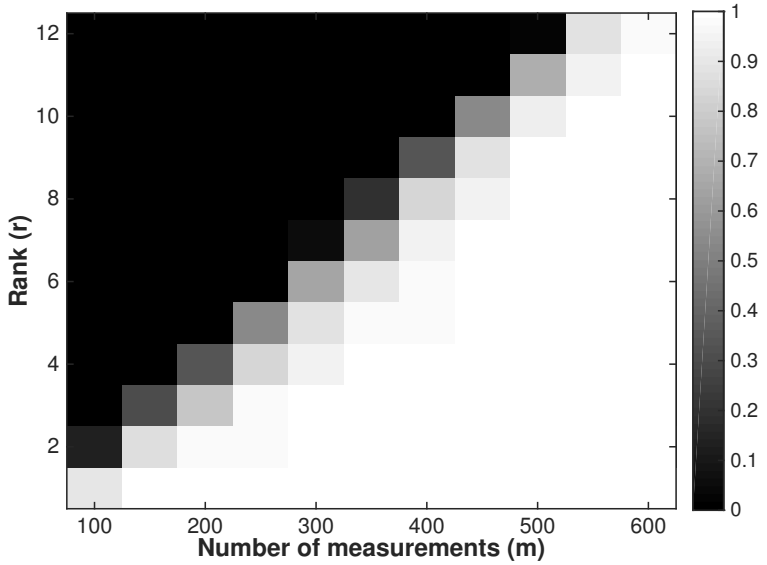
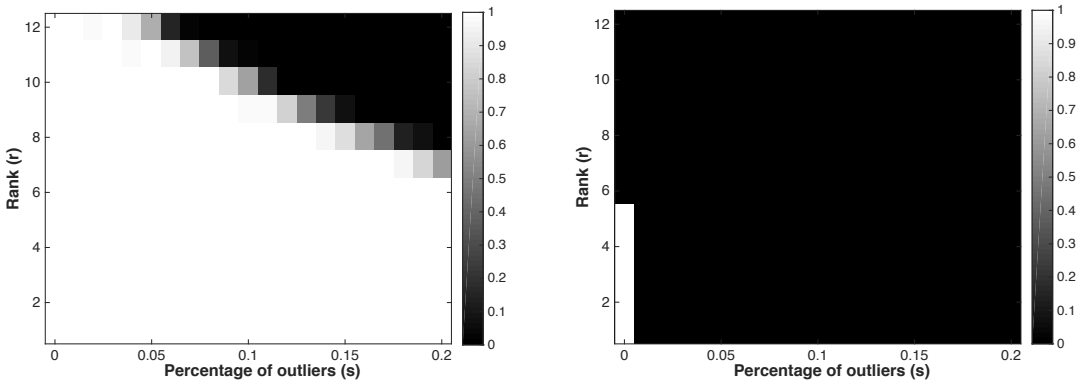


Figure 3.5: Phase transitions of low-rank PSD matrix recovery with respect to the number of measurements and the rank for the proposed Algorithm 3 using noise-free measurements, when $n = 40$.

Figure 3.5 shows the success rate of Algorithm 3 with respect to the number of measurements and the rank under the same setup of Figure 3.2 for noise-free measurements, when $n = 40$. Indeed, empirically Algorithm 3 performs similarly as the convex algorithms but with a much lower computational cost. Moreover, the proposed Algorithm 3 allows perfect recovery even in the presence of outliers. For comparison, we implement the extension of the WF algorithm in [25, 26, 48] in the low-rank case, that minimizes the squared ℓ_2-norm of the residual, where the update rule per iteration becomes

$$\boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t)} + \mu_t^{\mathrm{WF}} \cdot \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \|(\boldsymbol{U}^{(t)})^\top \boldsymbol{a}_i\|_2^2 \right) \boldsymbol{a}_i \boldsymbol{a}_i^\top \boldsymbol{U}^{(t)},$$

using the same initialization (3.21). The step size is set as $\mu_t^{\mathrm{WF}} = 0.1/\left\|\boldsymbol{U}_0\right\|_{\mathsf{F}}^2$. Figure



(a) (b)

Figure 3.6: Phase transitions of low-rank PSD matrix recovery with respect to the percentage of outliers and the rank using (a) the proposed Algorithm 3, and (b) the WF algorithm, when $n = 40$ and $m = 600$.

3.6 (a) shows the success rates of Algorithm 3 with respect to the percentage of outliers and the rank, under the same setup of Figure 3.3 (b), where the performance is even better than the convex counterpart in (3.4). In contrast, the WF algorithm performs poorly even with very few outliers, as shown in its success rate plot in Figure 3.6 (b), as the loss function used for WF is not robust to outliers.

3.6.4 Comparisons with Additional Bounded Noise

Finally, we compare the two proposed algorithms (Robust-PhaseLift in (3.4) and Algorithm 3), the WF algorithm and the PhaseLift algorithm in (3.5) when the measurements are corrupted by both outliers and bounded noise. Fix $n = 40$ and $r = 3$. The rank-r PSD matrix $\boldsymbol{X}_0$ and the sensing vectors are generated similarly as earlier. 5% of measurements are selected uniformly at random and corrupted by outliers obeying the Gaussian distribution $\mathcal{N}(0, 5^2)$. Moreover, each entry in the bounded

noise $\boldsymbol{w}$ is i.i.d. drawn from Unif$[-4/m, 4/m]$, thus $\|\boldsymbol{w}\|_1 \leq 4$. Figure 3.7 depicts the mean squared error $\|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\mathrm{F}}^2$ for different algorithms with respect to the number of measurements, where $\hat{\boldsymbol{X}}$ is the estimated PSD matrix. For the subgradient descent algorithm in Algorithm 3, various ranks are used as prior information, corresponding to the correct rank r, its underestimate $r-1$, and its overestimate $r+1$. It can

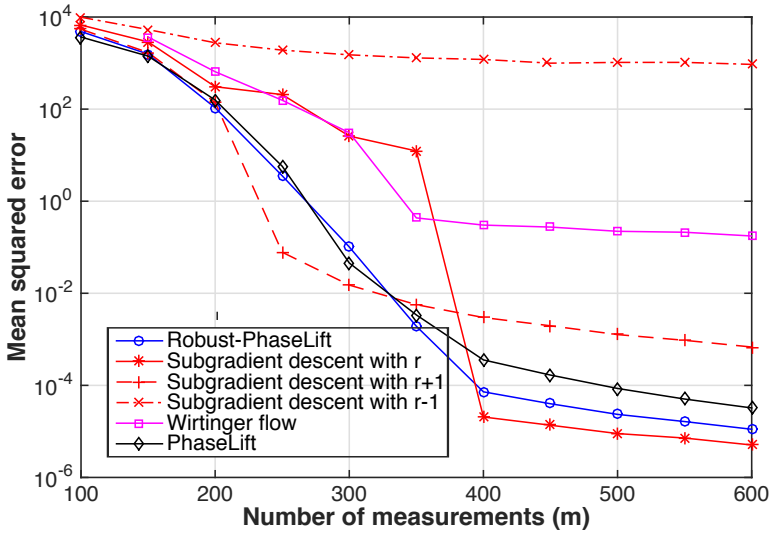


Figure 3.7: Comparisons of mean squared errors using different algorithms with respect to the number of measurements with 5% of outliers and bounded noise, when $n = 40$ and $r = 3$.

be seen that Algorithm 3 works well as long as the given rank provides an upper bound of the true rank, and it performs much better than the WF algorithm which is not outlier-robust. On the other hand, the PhaseLift algorithm (3.5) does not admit favorable performance if we set the constraint parameter as, for example, $c \cdot \epsilon$, for a small constant $c = 1, 2$ or 4, since the outliers do not fall into the prescribed

noise bound. In fact, it fails to return any feasible solution when the number and amplitudes of outliers is too large in our simulation (not shown). In order to obtain a favorable solution from PhaseLift, we use the oracle information to set the constraint parameter in (3.5) as $\epsilon = \|\boldsymbol{\eta} + \boldsymbol{w}\|_1$. In contrast, Robust-PhaseLift is parameter-free. It can be seen that Robust-PhaseLift, as well as Algorithm 3 with the correct model order, still achieve better performance than PhaseLift despite being aided by oracle information.

3.7 Conclusion

In this chapter, we address the problem of estimating a low-rank PSD matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ from rank-one measurements that are possibly corrupted by arbitrary outliers and bounded noise. This problem has many applications in covariance sketching, phase space tomography, and noncoherent detection in communications. It is shown that with an order of nr^2 random Gaussian sensing vectors, a PSD matrix of rank-r can be robustly recovered by minimizing the ℓ_1-norm of the observation residual within the semidefinite cone with high probability, even when a fraction of the measurements are adversarially corrupted. This convex formulation eliminates the need for trace minimization and tuning of parameters without prior knowledge of the outliers. Moreover, a nonconvex subgradient descent algorithm is proposed with excellent empirical performance, when additional information of the rank of the PSD matrix is available.

Chapter 4: Robust Matrix Recovery from Corrupted Linear Measurements

In this chapter, we focus on low-rank matrix recovery from random linear measurements in the presence of arbitrary outliers. Specifically, the sensing matrices are generated with i.i.d. standard Gaussian entries. Moreover, we assume that a small number of measurements are corrupted by outliers, possibly in an adversarial fashion with arbitrary amplitudes. The goal is to develop an efficient and robust algorithm that is able to handle a large number of adversarial outliers. The results of this chapter are summarized in the paper [90] and the paper submission [91].

4.1 Problem Formulation

Let $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ be a rank-r matrix that can be written as

$$\boldsymbol{M} = \boldsymbol{X}\boldsymbol{Y}^\top, \tag{4.1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}$ are the low-rank factors of $\boldsymbol{M}$. Define the condition number and the *average* condition number of $\boldsymbol{M}$ as

$$\kappa = \frac{\sigma_1(\boldsymbol{M})}{\sigma_r(\boldsymbol{M})}, \quad \text{and} \quad \bar{\kappa} = \frac{\|\boldsymbol{M}\|_{\mathsf{F}}}{\sqrt{r}\sigma_r(\boldsymbol{M})}, \tag{4.2}$$

respectively. Clearly, $\bar{\kappa} \leq \kappa$.

Let m be the number of measurements, and the set of sensing matrices are given as $\{\boldsymbol{A}_i\}_{i=1}^m$, where $\boldsymbol{A}_i \in \mathbb{R}^{n_1 \times n_2}$ is the ith sensing matrix. In particular, each entry of $\boldsymbol{A}_i$ is generated with i.i.d. standard Gaussian entries, i.e. $(\boldsymbol{A}_i)_{k,t} \sim \mathcal{N}(0,1)$. Denote the index set of corrupted measurements by $\mathcal{S}$, and correspondingly, the index set of clean measurements is given as the complementary set $\mathcal{S}^c$. Mathematically, the measurements $\boldsymbol{y} = \{y_i\}_{i=1}^m$ are given as

$$y_i = \begin{cases} \langle \boldsymbol{A}_i, \boldsymbol{M} \rangle, & \text{if} \quad i \in \mathcal{S}^c \\ \eta_i, & \text{if} \quad i \in \mathcal{S} \end{cases}, \tag{4.3}$$

where $\boldsymbol{\eta} = \{\eta_i\}_{i \in \mathcal{S}}$ is the set of outliers that can take arbitrary values. Denote the cardinality of $\mathcal{S}$ by $|\mathcal{S}| = s \cdot m$, where $0 \leq s < 1$ is the fraction of outliers. To simplify the notations, we define the linear maps $\mathcal{A}_i(\boldsymbol{M}) = \{\mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R} : \langle \boldsymbol{A}_i, \boldsymbol{M} \rangle\}$, and $\mathcal{A}(\boldsymbol{M}) = \{\mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}^m : \{\mathcal{A}_i(\boldsymbol{M})\}_{i=1}^m\}$.

Instead of recovering $\boldsymbol{M}$, we aim to directly recover its low-rank factors $(\boldsymbol{X}, \boldsymbol{Y})$ from the corrupted measurements $\boldsymbol{y}$, without a priori knowledge of statistical distribution or fractions of the outliers, in a computationally efficient and provably accurate manner. It is straightforward to see that for any orthonormal matrix $\boldsymbol{P} \in \mathbb{R}^{r \times r}$ and scalar $\gamma \in \mathbb{R}$ such that $\gamma \neq 0$, we have $(\gamma \boldsymbol{X}\boldsymbol{P})(\gamma^{-1}\boldsymbol{Y}\boldsymbol{P})^\top = \boldsymbol{X}\boldsymbol{Y}^\top$. To address the scaling ambiguity, we assume $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{Y}^\top \boldsymbol{Y}$, and consequently, $(\boldsymbol{X}, \boldsymbol{Y})$ can be recovered only up to orthonormal transformations. Hence, we measure the estimation accuracy by taking this into consideration. Let the estimates of low-rank factors be $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n_2 \times r}$, and define the augmented variables

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times r}, \quad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times r}. \tag{4.4}$$

Then the distance between $\boldsymbol{W}$ and $\boldsymbol{Z}$ is measured as

$$\operatorname{dist}(\boldsymbol{W}, \boldsymbol{Z}) = \min_{\boldsymbol{P} \in \mathcal{O}^{r \times r}} \|\boldsymbol{W} - \boldsymbol{Z}\boldsymbol{P}\|_{\mathsf{F}}. \tag{4.5}$$

Define

$$\boldsymbol{Q}_{(\boldsymbol{W},\boldsymbol{Z})} = \operatorname{argmin}_{\boldsymbol{P}\in\mathcal{O}^{r\times r}} \|\boldsymbol{W}-\boldsymbol{Z}\boldsymbol{P}\|_{\mathsf{F}}\,, \tag{4.6}$$

and then $\operatorname{dist}(\boldsymbol{W},\boldsymbol{Z}) = \|\boldsymbol{W}-\boldsymbol{Z}\boldsymbol{Q}\|_{\mathsf{F}}$, where the subscript of $\boldsymbol{Q}$ is dropped for notational simplicity.

This setting generalizes the outlier-free models studied in [5, 16, 26, 63], where convex and nonconvex approaches have been developed to accurately recover the low-rank matrix. Unfortunately, the vanilla gradient descent algorithm, of which the effectiveness has been demonstrated in [26, 63] for directly recovering the factors of low-rank matrices from random linear measurements, is very sensitive in the presence of even a single outlier, as the outliers can perturb the search directions arbitrarily. To handle outliers, existing convex optimization approaches (including Robust-PhaseLift in Chapter 3) based on sparse and low-rank decompositions can be applied using semidefinite programming [83, 84, 92]. However, their computational cost is very expensive. Therefore, our goal in this chapter is to develop a fast and robust nonconvex alternative that is globally convergent in a provable manner that can handle a large number of adversarial outliers.

4.2 Median-Truncated Gradient Descent

Define a quadratic loss function with respect to the ith measurement as

$$f_i(\boldsymbol{U},\boldsymbol{V}) = \frac{1}{4m}\left(y_i - \mathcal{A}_i(\boldsymbol{U}\boldsymbol{V}^\top)\right)^2, \tag{4.7}$$

where $\boldsymbol{U}\in\mathbb{R}^{n_1\times r}$ and $\boldsymbol{V}\in\mathbb{R}^{n_2\times r}$. In order to get rid of the impact of outliers, an ideal approach is to minimize an *oracle* loss function, expressed as

$$h_{\text{oracle}}(\boldsymbol{U},\boldsymbol{V}) = f_{\text{oracle}}(\boldsymbol{U},\boldsymbol{V}) + g(\boldsymbol{U},\boldsymbol{V})$$

$$= \sum_{i \in \mathcal{S}^c} f_i(\boldsymbol{U}, \boldsymbol{V}) + \frac{\lambda}{4} \left\| \boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{V}^\top \boldsymbol{V} \right\|_{\mathsf{F}}^2, \tag{4.8}$$

which aims to minimize the quadratic loss over only the *clean* measurements, in addition to a regularization term

$$g(\boldsymbol{U}, \boldsymbol{V}) = \frac{\lambda}{4} \left\| \boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{V}^\top \boldsymbol{V} \right\|_{\mathsf{F}}^2, \tag{4.9}$$

that aims at balancing the norm of the two factors. Nevertheless, it is impossible to minimize $h_{\text{oracle}}(\boldsymbol{U}, \boldsymbol{V})$ directly, since the oracle information regarding the support of outliers is absent. Moreover, the loss function is nonconvex, adding difficulty to its global optimization.

We propose a median-truncation strategy to robustify the gradient descent approach in [26, 63], which includes careful modifications on both initialization and local search. As it is widely known, the sample median is a more robust quantity to outliers, compared with the sample mean, which cannot be perturbed arbitrarily unless over half of the samples are outliers [93]. Therefore, it becomes an ideal metric to illuminate samples that are likely to be outliers and therefore should be eliminated during the gradient descent updates.

Specifically, we consider a gradient descent strategy where in each iteration, only a subset of all samples contribute to the search direction:

$$\begin{aligned} \boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \frac{\mu_t}{\|\boldsymbol{U}_0\|^2} \cdot \nabla_{\boldsymbol{U}} h_t(\boldsymbol{U}_t, \boldsymbol{V}_t); \\ \boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \frac{\mu_t}{\|\boldsymbol{V}_0\|^2} \cdot \nabla_{\boldsymbol{V}} h_t(\boldsymbol{U}_t, \boldsymbol{V}_t), \end{aligned} \tag{4.10}$$

where μ_t denotes the step size, and $\boldsymbol{W}_0 = [\boldsymbol{U}_0^\top, \boldsymbol{V}_0^\top]^\top$ is the initialization that will be specified later. Also, denote $\boldsymbol{W}_t = [\boldsymbol{U}_t^\top, \boldsymbol{V}_t^\top]^\top$. In particular, the iteration-varying loss function is given as

$$h_t(\boldsymbol{U}, \boldsymbol{V}) = \sum_{i \in \mathcal{E}^t} f_i(\boldsymbol{U}, \boldsymbol{V}) + g(\boldsymbol{U}, \boldsymbol{V}) := f_{\text{tr}}(\boldsymbol{U}, \boldsymbol{V}) + g(\boldsymbol{U}, \boldsymbol{V}), \tag{4.11}$$

where the set $\mathcal{E}^t$ varies at each iteration and includes only samples that are likely to be inliers. Denote the residual of the ith measurement at the tth iteration by

$$r_i^t = y_i - \mathcal{A}_i(\boldsymbol{U}_t \boldsymbol{V}_t^\top), \quad i = 1, \cdots, m, \tag{4.12}$$

and $\boldsymbol{r}^t = [r_1^t, r_2^t, \cdots, r_m^t]^\top = \boldsymbol{y} - \mathcal{A}(\boldsymbol{U}_t \boldsymbol{V}_t^\top)$. Then the set $\mathcal{E}^t$ is defined as

$$\mathcal{E}^t = \left\{ i \middle| |r_i^t| \leq \alpha_h \cdot \mathrm{med}\{|\boldsymbol{r}^t|\} \right\}, \tag{4.13}$$

where α_h is some small constant. In other words, only samples whose current absolute residuals are not too deviated from the sample median of the absolute residuals are included in the gradient update. As the estimate $(\boldsymbol{U}_t, \boldsymbol{V}_t)$ gets more accurate, we expect that the set $\mathcal{E}^t$ gets closer to the oracle set $\mathcal{S}^c$, and hence the gradient search is more accurate. Note that the set $\mathcal{E}^t$ varies per iteration, and therefore, can adaptively prune the outliers. The gradients of $h_t(\boldsymbol{U}, \boldsymbol{V})$ with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$ are given as

$$\begin{aligned} \nabla_{\boldsymbol{U}} h_t(\boldsymbol{U}, \boldsymbol{V}) &= \frac{1}{2m} \sum_{i \in \mathcal{E}^t} \left[\mathcal{A}_i \left(\boldsymbol{U}\boldsymbol{V}^\top \right) - y_i \right] \boldsymbol{A}_i \boldsymbol{V} + \lambda \boldsymbol{U} \left(\boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{V}^\top \boldsymbol{V} \right); \\ \nabla_{\boldsymbol{V}} h_t(\boldsymbol{U}, \boldsymbol{V}) &= \frac{1}{2m} \sum_{i \in \mathcal{E}^t} \left[\mathcal{A}_i \left(\boldsymbol{U}\boldsymbol{V}^\top \right) - y_i \right] \boldsymbol{A}_i^\top \boldsymbol{U} + \lambda \boldsymbol{V} \left(\boldsymbol{V}^\top \boldsymbol{V} - \boldsymbol{U}^\top \boldsymbol{U} \right). \end{aligned} \tag{4.14}$$

For initialization, we adopt a truncated spectral method, which uses the top singular vectors of a sample-weighted surrogate matrix, where again only the samples whose absolute values do not significantly digress from the sample median are included. To avoid statistical dependence in the theoretical analysis, we split the samples by using the sample median of m_2 samples to estimate $\|\boldsymbol{M}\|_{\mathsf{F}}$, and then using the rest of the samples to construct the truncated surrogate matrix to perform a spectral initialization. In practice, we find that this sample split is unnecessary, as demonstrated in the numerical simulations.

Algorithm 4: Median-Truncated Gradient Descent (median-TGD)

Parameters: Thresholds α_y and α_h, step size μ_t, average condition number bound $\bar{\kappa}_0$, rank r, and regularization parameter λ.

Input: Measurements $\boldsymbol{y} = \{y_i\}_{i=1}^m$, and sensing matrices $\{\boldsymbol{A}_i\}_{i=1}^m$.

Initialization:

1) Set $\boldsymbol{y}_1 = \{y_i\}_{i=1}^{m_1}$ and $\boldsymbol{y}_2 = \{y_i\}_{i=m_1+1}^{m}$, where $m_1 = \lceil m/2 \rceil$ and $m_2 = m - m_1$.

2) Take the rank-r SVD of the matrix

$$\boldsymbol{K} = \frac{1}{m_1} \sum_{i=1}^{m_1} y_i \boldsymbol{A}_i \mathbb{I}_{\{|y_i| \le \alpha_y \cdot \mathrm{med}(|\boldsymbol{y}_2|)\}}, \tag{4.15}$$

which is denoted by $\boldsymbol{C}_L \boldsymbol{\Sigma} \boldsymbol{C}_R^\top :=$ rank-r SVD of $\boldsymbol{K}$, where $\boldsymbol{C}_L \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{C}_R \in \mathbb{R}^{n_2 \times r}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$.

3) Initialize $\boldsymbol{U}_0 = \boldsymbol{C}_L \boldsymbol{\Sigma}^{1/2}$, and $\boldsymbol{V}_0 = \boldsymbol{C}_R \boldsymbol{\Sigma}^{1/2}$.

Gradient Loop: For $t = 0 : 1 : T-1$ do

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \frac{\mu_t}{\|\boldsymbol{U}_0\|^2} \cdot \left[\frac{1}{2m} \sum_{i=1}^{m} \left(\mathcal{A}_i \left(\boldsymbol{U}_t \boldsymbol{V}_t^\top \right) - y_i \right) \boldsymbol{A}_i \boldsymbol{V}_t \mathbb{I}_{\mathcal{E}_i^t} + \lambda \boldsymbol{U}_t \left(\boldsymbol{U}_t^\top \boldsymbol{U}_t - \boldsymbol{V}_t^\top \boldsymbol{V}_t \right) \right];$$

$$\boldsymbol{V}_{t+1} = \boldsymbol{V}_t - \frac{\mu_t}{\|\boldsymbol{V}_0\|^2} \cdot \left[\frac{1}{2m} \sum_{i=1}^{m} \left(\mathcal{A}_i \left(\boldsymbol{U}_t \boldsymbol{V}_t^\top \right) - y_i \right) \boldsymbol{A}_i^\top \boldsymbol{U}_t \mathbb{I}_{\mathcal{E}_i^t} + \lambda \boldsymbol{V}_t \left(\boldsymbol{V}_t^\top \boldsymbol{V}_t - \boldsymbol{U}_t^\top \boldsymbol{U}_t \right) \right],$$

where

$$\mathcal{E}_i^t = \left\{ \left| y_i - \mathcal{A}_i(\boldsymbol{U}_t \boldsymbol{V}_t^\top) \right| \le \alpha_h \cdot \mathrm{med} \left(\left| \boldsymbol{y} - \mathcal{A} \left(\boldsymbol{U}_t \boldsymbol{V}_t^\top \right) \right| \right) \right\}.$$

Output: $\hat{\boldsymbol{X}} = \boldsymbol{U}_T$, and $\hat{\boldsymbol{Y}} = \boldsymbol{V}_T$.

The details of the proposed algorithm, denoted as median-truncated gradient descent (median-TGD), are provided in Algorithm 4, where the stopping criterion is simply set as reaching a preset maximum number of iterations. In practice, it is also possible to set the stopping criteria by examining the progress between iterations. In sharp contrast to the standard gradient descent approach that exploits all samples in every iteration [26], both the initialization and the search directions are controlled more carefully in order to adaptively eliminate outliers, while maintaining a similar low computational cost.

Computationally, because the sample median can be computed in a linear time [94], our median-truncated gradient descent algorithm shares a similar attractive computational cost as [26, 63]. Specifically, the per-iteration computational complexity of the proposed algorithm is on the order of $O\left(mn^2 + 2n^2r + 4nr^2\right)$, where $n = (n_1 + n_2)/2$, which is linear with respect to m, while is quadratic with respect to n and r[6]. The proposed algorithm enjoys a lower computational complexity, compared with SVD-based methods [15] and alternating minimization [27], which usually require more than $O\left(mn^2 + n^3\right)$ or $O\left(mn^2 + m^2\right)$ operations during each iteration.

4.3 Performance Guarantees

Theorem 3 summarizes the performance guarantee of median-TGD in Algorithm 4 for low-rank matrix recovery using Gaussian measurements in the presence of sparse arbitrary outliers, when initialized within a proper neighborhood around the ground truth. As used before, we let $n = (n_1 + n_2)/2$ for convenience.

Theorem 3 (Exact recovery with sparse arbitrary outliers)**.** *Assume the measurement model* (4.3)*, where each $\boldsymbol{A}_i$ is generated with i.i.d. standard Gaussian entries. Suppose that the initialization $\boldsymbol{W}_0$ satisfies*

$$\operatorname{dist}\left(\boldsymbol{W}_0, \boldsymbol{Z}\right) \leq \frac{1}{24}\sigma_r\left(\boldsymbol{Z}\right).$$

Recall that $\kappa = \frac{\sigma_1(\boldsymbol{M})}{\sigma_r(\boldsymbol{M})}$. Set $\alpha_h = 6$ and $\lambda = \mathbb{E}\left[\xi^2 \mathbb{I}_{\{|\xi| \leq 0.65\alpha_h\}}\right]/4$ with $\xi \sim \mathcal{N}(0,1)$. There exist some constants $s_0 > 0$, $c_0 > 1, c_1 > 1$ such that with probability at least $1 - e^{-c_1 m}$, if $s \leq s_0$, and $m \geq c_1 nr \log n$, there exists a constant $\mu \leq \frac{1}{740}$, such that

[6]In practice, our algorithm can be applied to other measurement ensembles with more structures, such as sparsity, and the computational complexity can be further reduced.

with $\mu_t = \mu$, the estimates of median-TGD satisfy

$$\operatorname{dist}\left(\boldsymbol{W}_t, \boldsymbol{Z}\right) \leq \left(1 - \frac{\mu}{10\kappa}\right)^{t/2} \operatorname{dist}\left(\boldsymbol{W}_0, \boldsymbol{Z}\right).$$

Theorem 3 suggests that if the initialization $\boldsymbol{W}_0$ lies in the basin of attraction, median-TGD converges to the ground truth at a linear rate as long as the number m of measurements is on the order of $nr \log n$, even when a constant fraction of measurements are corrupted arbitrarily. In comparisons, the gradient descent algorithm by Tu et.al. [63] achieves the same convergence rate in a similar basin of attraction, with an order of nr measurements using outlier-free measurements. Therefore, our algorithm achieves robustness up to a constant fraction of outliers with a slight price of an additional logarithmic factor in the sample complexity.

Theorem 4 guarantees that the proposed truncated spectral method provides an initialization in the basin of attraction with high probability.

Theorem 4. *Assume the measurement model* (4.3)*, and $\bar{\kappa} \leq \bar{\kappa}_0$. Set $\alpha_y = 2\log(r^{1/4}\bar{\kappa}_0^{1/2} + 20)$. There exist some constants $s_1 > 0$ and $c_2, c_3, c_4 > 1$ such that with probability at least $1 - n^{-c_2} - \exp(-c_3 m)$, if $s \leq s_1/(\sqrt{r}\bar{\kappa})$, and $m \geq c_4 \alpha_y^2 \bar{\kappa}^2 nr^2 \log n$, we have*

$$\operatorname{dist}\left(\boldsymbol{W}_0, \boldsymbol{Z}\right) \leq \frac{1}{24}\sigma_r\left(\boldsymbol{Z}\right).$$

Theorem 4 suggests that the proposed initialization scheme is guaranteed to obtain a valid initialization in the basin of attraction with an order of $nr^2 \log n \log^2 r$ measurements when a fraction of $1/\sqrt{r}$ measurements are arbitrarily corrupted, assuming the average condition number $\bar{\kappa}$ is a small constant. In comparisons, in the outlier-free setting, Tu et.al. [63] requires an order of $nr^2\kappa^2$ measurements for a one-step spectral initialization, which is closest to our scheme. Therefore, our initialization achieves

robustness to a $1/\sqrt{r}$ fraction of outliers at a slight price of additional logarithmic factors in the sample complexity. It is worthwhile to note that in the absence of outliers, Tu et.al. [63] was able to further reduce the sample complexity of initialization to an order of nr by running multiple iterations of projected gradient descent. However, it is not clear whether such an iterative scheme can be generalized to the setting with outliers in our work.

In the case when the rank is a small constant, our results indicate that the proposed algorithm can tolerate a constant fraction of outliers with an order of $n \log n$ measurements, which is much smaller than the size of the matrix. Finally, we note that the parameter bounds in all theorems, including α_h, α_y and μ, are not optimized for performance, but mainly selected to establish the theoretical guarantees.

4.4 Related Work

Our work is amid the recent surge of nonconvex approaches for high-dimensional signal estimation, e.g. an incomplete and still growing list [25, 26, 53, 60, 63, 64, 95–98]. A series of recent work has demonstrated that, starting from a careful initialization, simple algorithms such as gradient descent [26, 53, 63, 95, 98, 99] and alternating minimization [27, 100] enjoy global convergence guarantees under near-optimal sample complexity. Some of these algorithms also converge at a linear rate, making them extremely appealing computationally. On the other hand, the global geometry of nonconvex low-rank matrix estimation has been investigated in [101–104], and it is proven that no spurious local optima, except strict saddle points, exist under suitable

coherence conditions and sufficiently large sample size. This implies global convergence from random initialization, provided the algorithm of choice can escape saddle points [105–107].

The most closely-related work is on low-rank matrix recovery using random linear measurements [26,63] in the absence of outliers, in the context of which our algorithm can be thought as a robust counterpart. Our particular approach is inspired by the work [97] on robust phase retrieval, which can be thought as robust recovery of a rank-one PSD matrix using rank-one measurement operators [10]. Our model in the current work differs as we tackle low-rank matrix recovery using random full-rank measurement operators, and thus non-trivial technical developments are necessary.

The concept of median has been adopted in various sub-domains of machine learning, for instance, K-median clustering [108] and resilient data aggregation for sensor networks [109]. The median-TGD algorithm presented here further extends the applications of median to robust high-dimensional estimation problems with theoretical guarantees. Another popular approach in robust estimation is to use the trimmed mean [93], which has found success in robustifying sparse regression [110] and subspace clustering [111]. However, using the trimmed mean needs an upper bound on the number of outliers, whereas median does not require such information. Recently, geometric median is adopted for robust empirical risk minimization as well [112–114].

It is worth mentioning that, besides convex method [84,92], nonconvex approaches for robust low-rank matrix completion have also been presented in [115–117], where the goal is to separate a low-rank matrix and sparse outliers from a small number of direct or linear measurements of their sum. The approaches typically use thresholding-based truncation for outlier removal and projected gradient descent for

low-rank matrix recovery, which are somewhat similar to our approach in terms of different ways to remove outliers. However, this line of work typically requires stronger assumptions on the outliers such as spread-ness conditions, while we allow arbitrary outliers.

4.5 Numerical Experiments

In this section, we evaluate the performance of the proposed median-TGD algorithm via conducting several numerical experiments. As mentioned earlier, for the initialization step, in practice we find it is not necessary to split the samples into two parts. Therefore, the matrix in (4.15) is changed instead to

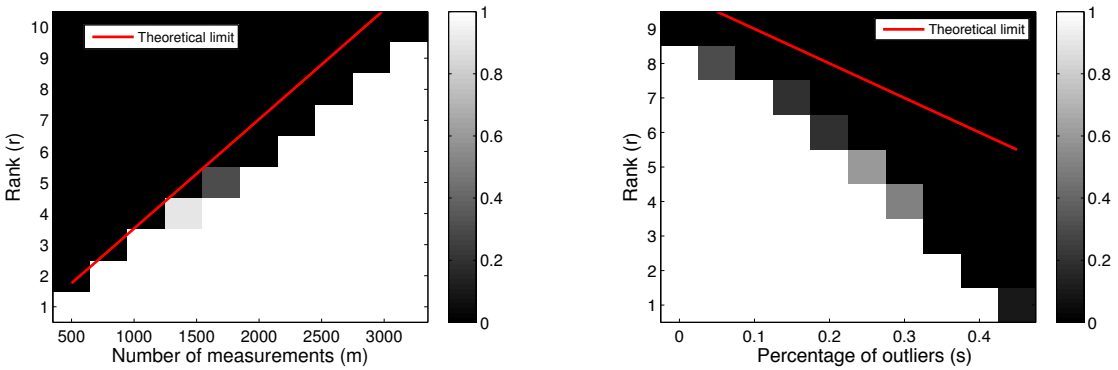
$$\boldsymbol{K} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{A}_i \mathbb{I}_{\{|y_i| \le \alpha_y \cdot \mathrm{med}(|\boldsymbol{y}|)\}}. \tag{4.16}$$

In particular, we check the trade-offs between the number of measurements, the rank and the fraction of outliers for accurate low-rank matrix recovery, and compare against the algorithm in [63], referred to as the vanilla gradient descent algorithm (vanilla-GD), to demonstrate the performance improvements in the presence of outliers due to median truncations.

Let $n_1 = 150$, $n_2 = 120$. We randomly generate a rank-r matrix as $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{Y}^\top$, where both $\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}$ are composed of i.i.d. standard Gaussian variables. The outliers are i.i.d. randomly generated following $\mathcal{N}(0, 10^4 \left\|\boldsymbol{M}\right\|_{\mathsf{F}}^2)$. We set $\alpha_y = 12$ and $\alpha_h = 6$, and pick a constant step size $\mu_t = 0.4$. In all experiments, the maximum number of iterations for median-TGD algorithm is set as $T = 10^4$ to guarantee convergence. Moreover, let $(\hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}})$ be the solution to the algorithm under examination, and the recovered low-rank matrix is given as $\hat{\boldsymbol{M}} = \hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}^\top$. Then, the normalized estimate error is defined as $\left\|\hat{\boldsymbol{M}} - \boldsymbol{M}\right\|_{\mathsf{F}} / \left\|\boldsymbol{M}\right\|_{\mathsf{F}}$.

4.5.1 Phase Transitions

We first examine the phase transitions of median-TGD algorithm with respect to the number of measurements, the rank and the percentage of outliers. Fix the percentage of outliers as $s = 5\%$. Given a pair of m and r, a ground truth $(\boldsymbol{X}, \boldsymbol{Y})$ is generated composed of i.i.d. standard Gaussian variables. Multiple Monte Carlo trials are carried out, and each trial is deemed a success if the normalized estimate error is less than 10^{-6}. Figure 4.1 (a) shows the success rates of median-TGD, averaged over



(a) (b)

Figure 4.1: Phase transitions of low-rank matrix recovery when $n_1 = 150$ and $n_2 = 120$. (a) Success rate with respect to the number of measurements and the rank, when 5% of measurements are corrupted by outliers. (b) Success rate with respect to the percentage of outliers and the rank, when $m = 2700$.

10 trials, with respect to the number of measurements and the rank, where the red line shows the theoretical limit defined as $r = (1 - s)m/(n_1 + n_2)$ by a heuristic count of the degrees of freedom. It can be seen that the required number of measurements for a successful matrix recovery scales linearly with the rank r, and the transition is sharp. We next examine the success rates of median-TGD with respect to the percentage

of outliers and the rank. Fix $m = 2700$. Under the same setup as Figure 4.1 (a), Figure 4.1 (b) shows the success rate of median-TGD, averaged over 10 trials, with respect to the rank and the percentage of outliers. The performance of median-TGD degenerates smoothly with the increase of the percentage of outliers. Similarly, the red line shows the theoretical limit as a comparison.

4.5.2 Stability to Additional Bounded Noise

We next examine the performance of median-TGD when the measurements are contaminated by both sparse outliers and dense noise. Here, the measurements are rewritten as

$$y_i = \begin{cases} \langle \boldsymbol{A}_i, \boldsymbol{M} \rangle + w_i, & \text{if} \quad i \in \mathcal{S}^c \\ \eta_i + w_i, & \text{if} \quad i \in \mathcal{S} \end{cases}, \tag{4.17}$$

where w_i, for $i = 1, \cdots, m$, denote the additional bounded noise. Fix $r = 5$ and $s = 5\%$. The dense noise is generated with i.i.d. random entries following $0.05\sigma_5(\boldsymbol{M}) \cdot \mathcal{U}[-1, 1]$. Figure 4.2 depicts the average normalized reconstruction errors with respect to the number of measurements using both median-TGD and vanilla-GD [63], where vanilla-GD is always given the true rank information, i.e. $r = 5$. The performance of median-TGD is comparable to that of vanilla-GD using outlier-free measurements, which cannot produce reliable estimates when the measurements are corrupted by outliers. Therefore, median-TGD can handle outliers in a much more robust manner. Moreover, the performance of median-GD is stable as long as an upper bound of the true rank is used.

We then compare the convergence rates of median-TGD and vanilla-GD under various outlier settings, by fixing $m = 2400$ while keeping the other settings the same as Figure 4.2. Figure 4.3 shows the normalized estimate error with respect to the

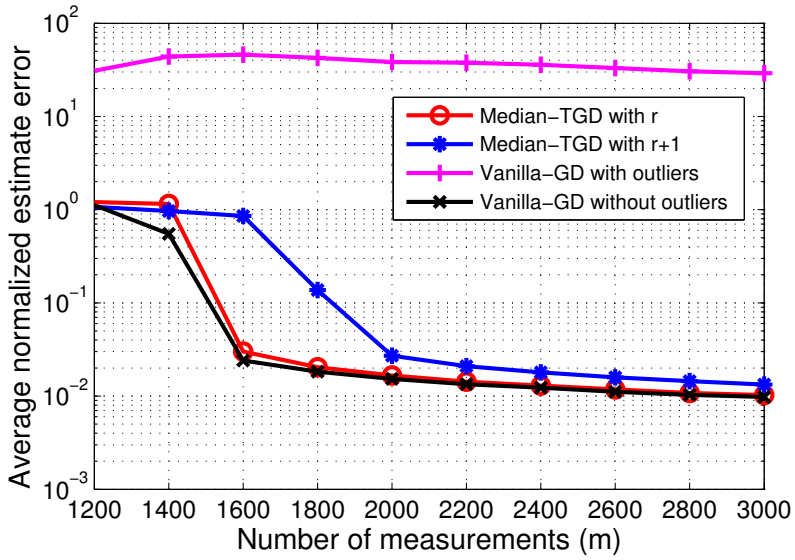
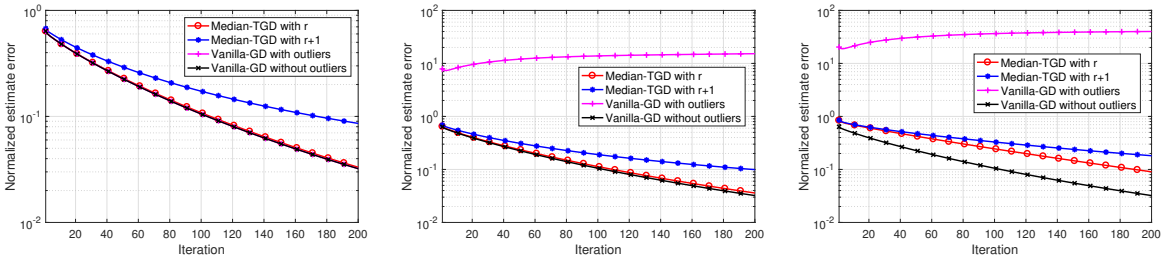


Figure 4.2: Comparisons of average normalized estimate errors between median-TGD and vanilla-GD in [63] with respect to the number of measurements, with 5% of measurements corrupted by outliers and additional bounded noise, when $n_1 = 150$, $n_2 = 120$, and $r = 5$.



(a) $s = 0$ (b) $s = 1\%$ (c) $s = 10\%$

Figure 4.3: Comparisons of convergence rates between median-TGD and vanilla-GD in different outlier-corruption scenarios, when $m = 2400$, $n_1 = 150$, $n_2 = 120$ and $r = 5$.

number of iterations of median-TGD and vanilla-GD with no outliers, 1% of outliers, and 10% of outliers, respectively. In the outlier-free case, both algorithms have comparable convergence rates. However, even with a few outliers, vanilla-GD suffers from a dramatical performance degradation, while median-TGD is robust against outliers and can still converge to an accurate estimate. Numerical experiments demonstrate the excellent empirical performance of the proposed algorithm for low-rank matrix recovery from outlier-corrupted measurements, which significantly outperforms the existing algorithms that are not resilient to outliers [26, 63].

4.6 Proof of Linear Convergence

In this section, we present the proof of Theorem 3. To obtain the performance guarantees, we establish that the proposed median-truncated gradient satisfies a so-called regularity condition (RC) [25], which is a sufficient condition for establishing the linear convergence to the ground truth. Since its debut in [25], the RC has been employed successfully in the analysis of phase retrieval [25, 60, 96, 97], blind deconvolution [64] and low-rank matrix recovery [26, 63, 98] in the recent literature, to name a few. However, our analysis is significantly more involved due to the fact that the truncation procedure involving low-rank matrices has not been tackled in the previous literature. In particular, we establish a new restricted isometry property (RIP) of the sample median for the class of low-rank matrices, which can be thought as an extension of the RIP for the sample mean in compressed sensing literature [5,16]. We remark that such a result can be of independent interest, and its establishment is non-trivial due to the nonlinear character of the median operation. Specifically, the roadmap of proof is below. Section 4.6.1 first establishes an RIP-like property for the

median of random linear measurements of low-rank matrices. Section 4.6.2 describes the RC, which is used to certify the linear convergence of the proposed algorithm. Section 4.6.3 proves several properties of the truncated gradient which are then used in Section 4.6.4 to prove the RC and finish the proof.

4.6.1 Concentration Property of Sample Median

To begin, we define below the quantile function of a population distribution and its corresponding sample version.

Definition 1 (Generalized quantile function)**.** *Let* $0 < \tau < 1$*. For a cumulative distribution function (CDF)* $F(x)$*, the generalized quantile function is defined as*

$$F^{-1}(\tau) = \inf\{x \in \mathbb{R} : F(x) \geq \tau\}.$$

For simplicity, denote $\theta_\tau(F) = F^{-1}(\tau)$ *as the* τ*-quantile of* F*. Moreover, for a sample collection* $\boldsymbol{y} = \{y_i\}_{i=1}^m$*, the sample* τ*-quantile* $\theta_\tau(\boldsymbol{y})$ *means* $\theta_\tau(\hat{F})$*, where* $\hat{F}$ *is the empirical distribution of the samples* $\boldsymbol{y}$*. Specifically,* $\operatorname{med}(\boldsymbol{y}) = \theta_{1/2}(\boldsymbol{y})$*.*

We establish a RIP-style concentration property for the sample median used in the truncation indicator of gradient descent, which provides theoretical footings on the success of the proposed algorithm. The concentration property of the sample p-quantile function $\theta_p(|\mathcal{A}(\boldsymbol{G})|)$ of all rank-$2r$ matrices $\boldsymbol{G}$ is formulated in the following proposition, of which the proof is shown in Appendix D.1.

Proposition 1. *Fix* $\epsilon \in (0,1)$*. If* $m \geq c_0(\epsilon^{-2}\log\epsilon^{-1})\, nr\log n$ *for some large enough constant* c_0*, then with probability at least* $1 - c_1\exp(-c_2 m\epsilon^2)$*, where* c_1 *and* c_2 *are some constants, we have for all rank-*$2r$ *matrices* $\boldsymbol{G} \in \mathbb{R}^{n_1\times n_2}$*,*

$$\theta_{\frac{1}{2}}(|\mathcal{A}(\boldsymbol{G})|) \in [0.6745 - \epsilon, 0.6745 + \epsilon]\,\|\boldsymbol{G}\|_{\mathsf{F}};$$

$$\theta_{0.49}\left(\left|\mathcal{A}\left(\boldsymbol{G}\right)\right|\right) \in \left[0.6588-\epsilon, 0.6588+\epsilon\right]\left\|\boldsymbol{G}\right\|_{\mathsf{F}};$$

$$\theta_{0.51}\left(\left|\mathcal{A}\left(\boldsymbol{G}\right)\right|\right) \in \left[0.6903-\epsilon, 0.6903+\epsilon\right]\left\|\boldsymbol{G}\right\|_{\mathsf{F}}.$$

Proposition 1 suggests that as long as m is on the order of $nr\log n$, the sample median $\theta_{\frac{1}{2}}\left(\left|\mathcal{A}\left(\boldsymbol{G}\right)\right|\right)$ concentrates around a scaled $\left\|\boldsymbol{G}\right\|_{\mathsf{F}}$ for all rank-$2r$ matrices $\boldsymbol{G}$, which resembles the matrix RIP in [5]. Based on Proposition 1, provided that $m \geq c_0 nr\log n$ for some large enough constant c_0, setting $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top$, we have

$$\theta_{0.49}, \theta_{\frac{1}{2}}, \theta_{0.51}\left(\left|\mathcal{A}\left(\boldsymbol{X}\boldsymbol{Y}^\top\right) - \mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^\top\right)\right|\right) \in \left[0.65, 0.70\right]\left\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top\right\|_{\mathsf{F}} \tag{4.18}$$

holds with probability at least $1 - c_1\exp\left(-c_2 m\right)$ for all $\boldsymbol{U} \in \mathbb{R}^{n_1\times r}$, $\boldsymbol{V} \in \mathbb{R}^{n_2\times r}$, $\boldsymbol{X} \in \mathbb{R}^{n_1\times r}$, and $\boldsymbol{Y} \in \mathbb{R}^{n_2\times r}$. On the other end, due to Lemma 22, we have

$$\begin{aligned}
&\operatorname{med}\left(\left|\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^\top\right)\right|\right) \geq \theta_{\frac{1}{2}-s}\left(\left|\mathcal{A}\left(\boldsymbol{X}\boldsymbol{Y}^\top\right) - \mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^\top\right)\right|\right);\\
&\operatorname{med}\left(\left|\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^\top\right)\right|\right) \leq \theta_{\frac{1}{2}+s}\left(\left|\mathcal{A}\left(\boldsymbol{X}\boldsymbol{Y}^\top\right) - \mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^\top\right)\right|\right).
\end{aligned}$$

As a result, when the fraction of corruption satisfies $s \leq 0.01$, the above equation together with (4.18) yields

$$0.65\left\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top\right\|_{\mathsf{F}} \leq \operatorname{med}\left(\left|\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^\top\right)\right|\right) \leq 0.70\left\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top\right\|_{\mathsf{F}}. \tag{4.19}$$

Therefore, an important consequence is that the truncation event $\mathcal{E}_i$ satisfies

$$\begin{aligned}
\mathbb{I}_{\mathcal{E}_i} &\geq \mathbb{I}_{\left\{\left|\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top\rangle - y_i\right| \leq 0.65\alpha_h\left\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}}\right\}};\\
\mathbb{I}_{\mathcal{E}_i} &\leq \mathbb{I}_{\left\{\left|\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top\rangle - y_i\right| \leq 0.70\alpha_h\left\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}}\right\}}.
\end{aligned} \tag{4.20}$$

4.6.2 Regularity Condition

We first introduce the so-called RC [25, 26, 63] that characterizes the benign curvature of the loss function around the ground truth, and guarantees the linear convergence of gradient descent to the ground truth.

We first rewrite the loss function in terms of the augmented variables in (4.4). Denote the matrix $\boldsymbol{B}_i = \begin{bmatrix} \mathbf{0} & \frac{1}{2}\boldsymbol{A}_i \\ \frac{1}{2}\boldsymbol{A}_i^\top & \mathbf{0} \end{bmatrix}$, and define $\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right) := \langle \boldsymbol{B}_i, \boldsymbol{W}\boldsymbol{W}^\top \rangle$ and $\mathcal{B}\left(\boldsymbol{W}\boldsymbol{W}^\top\right) := \left\{\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right)\right\}_{i=1}^m$, then we can have the equivalent representation

$$\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right) = \langle \boldsymbol{B}_i, \boldsymbol{W}\boldsymbol{W}^\top \rangle = \langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top \rangle = \mathcal{A}_i\left(\boldsymbol{U}\boldsymbol{V}^\top\right). \tag{4.21}$$

The regularizer can be rewritten as

$$g\left(\boldsymbol{W}\right) = \frac{\lambda}{4}\left\|\boldsymbol{W}^\top \boldsymbol{D}\boldsymbol{W}\right\|_{\mathsf{F}}^2, \tag{4.22}$$

where $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{I}_{n_1} & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{I}_{n_2} \end{bmatrix}$, and its gradient can be rewritten as

$$\nabla g(\boldsymbol{W}) = \lambda \boldsymbol{D}\boldsymbol{W}\left(\boldsymbol{W}^\top \boldsymbol{D}\boldsymbol{W}\right). \tag{4.23}$$

Then the truncated gradient can be rewritten as a function of $\boldsymbol{W}$,

$$\begin{aligned} \nabla h(\boldsymbol{W}) &= \frac{1}{m}\sum_{i=1}^m \left(\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right) - y_i\right)\boldsymbol{B}_i\boldsymbol{W}\mathbb{I}_{\mathcal{E}_i} + \lambda \boldsymbol{D}\boldsymbol{W}\left(\boldsymbol{W}^\top \boldsymbol{D}\boldsymbol{W}\right) \\ &:= \nabla f_{tr}(\boldsymbol{W}) + \nabla g(\boldsymbol{W}), \end{aligned} \tag{4.24}$$

where

$$\mathcal{E}_i = \left\{\left|y_i - \mathcal{B}_i(\boldsymbol{W}\boldsymbol{W}^\top)\right| \le \alpha_h \cdot \operatorname{med}\left(\left|\boldsymbol{y} - \mathcal{B}\left(\boldsymbol{W}\boldsymbol{W}^\top\right)\right|\right)\right\}. \tag{4.25}$$

Then the RC is defined in the following definition.

Definition 2 (Regularity Condition)**.** *Suppose $\boldsymbol{Z} \in \mathbb{R}^{(n_1+n_2)\times r}$ is the ground truth. The set of matrices that are in an ϵ-neighborhood of $\boldsymbol{Z}$ is defined as*

$$\mathcal{C}\left(\epsilon\right) = \left\{\boldsymbol{W} \in \mathbb{R}^{(n_1+n_2)\times r} : \operatorname{dist}\left(\boldsymbol{W}, \boldsymbol{Z}\right) \le \epsilon\right\}.$$

Then the function $h(\boldsymbol{W})$ is said to satisfy the RC, denoted by $\mathrm{RC}\left(\alpha, \beta, \epsilon\right)$*, if for all matrices $\boldsymbol{W} \in \mathcal{C}\left(\epsilon\right)$, the following inequality holds:*

$$\left\langle \nabla h\left(\boldsymbol{W}\right), \boldsymbol{W} - \boldsymbol{Z}\boldsymbol{Q}\right\rangle \ge \frac{\sigma_r^2\left(\boldsymbol{Z}\right)}{\alpha}\left\|\boldsymbol{W} - \boldsymbol{Z}\boldsymbol{Q}\right\|_{\mathsf{F}}^2 + \frac{1}{\beta\left\|\boldsymbol{Z}\right\|^2}\left\|\nabla h\left(\boldsymbol{W}\right)\right\|_{\mathsf{F}}^2, \tag{4.26}$$

where $\boldsymbol{Q}$ is an orthonormal matrix given in (4.6)*.*

The neighborhood $\mathcal{C}(\epsilon)$ is known as the basin of attraction. Interestingly, if $h(\boldsymbol{W})$ satisfies the RC, then initializing a simple gradient descent algorithm in the basin of attraction guarantees that the iterates converge at a linear rate to the ground truth, as summarized in the following lemma.

Lemma 11. *[25, 26, 63] Suppose that $h(\boldsymbol{W})$ satisfies $\mathrm{RC}(\alpha, \beta, \epsilon)$ and $\boldsymbol{W}_0 \in \mathcal{C}(\epsilon)$. Consider the gradient descent update*

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \frac{\mu}{\|\boldsymbol{Z}\|^2} \nabla h(\boldsymbol{W}_t) \tag{4.27}$$

with the step size $0 < \mu < \min\{\alpha/2, 2/\beta\}$. Then for all $t \geq 0$, we have $\boldsymbol{W}_t \in \mathcal{C}(\epsilon)$ and

$$\mathrm{dist}(\boldsymbol{W}_t, \boldsymbol{Z}) \leq \left(1 - \frac{2\mu}{\alpha\kappa}\right)^{t/2} \mathrm{dist}(\boldsymbol{W}_0, \boldsymbol{Z}).$$

Note that since the initialization satisfies $\mathrm{dist}(\boldsymbol{W}_0, \boldsymbol{Z}) \leq \frac{1}{24}\sigma_r(\boldsymbol{Z})$, by the triangle inequality we can guarantee that

$$\frac{23}{24}\|\boldsymbol{Z}\| \leq \|\boldsymbol{W}_0\| \leq \frac{25}{24}\|\boldsymbol{Z}\|,$$

which implies

$$\frac{23}{24\sqrt{2}}\|\boldsymbol{Z}\| \leq \|\boldsymbol{U}_0\| \leq \frac{25}{24\sqrt{2}}\|\boldsymbol{Z}\|;$$
$$\frac{23}{24\sqrt{2}}\|\boldsymbol{Z}\| \leq \|\boldsymbol{V}_0\| \leq \frac{25}{24\sqrt{2}}\|\boldsymbol{Z}\|,$$

where we use the fact $\|\boldsymbol{U}_0\| = \|\boldsymbol{V}_0\| = \|\boldsymbol{W}_0\|/\sqrt{2}$. Therefore, instead of proving the linear convergence of the actual update size $\frac{\mu}{\|\boldsymbol{U}_0\|^2}$ and $\frac{\mu}{\|\boldsymbol{V}_0\|^2}$, we prove it for the step size $\frac{\mu}{\|\boldsymbol{Z}\|^2}$ in (4.27), since they only differ by a constant scaling of μ. Hence, the rest of the proof is to verify that RC holds for the truncated gradient.

4.6.3 Properties of Truncated Gradient

We start by proving a few key properties of the truncated gradient $\nabla h(\boldsymbol{W}) = \nabla f_{tr}(\boldsymbol{W}) + \nabla g(\boldsymbol{W})$. Consider the measurement model with sparse outliers in (4.3). Define the truncation event

$$\tilde{\mathcal{E}}_i = \left\{ \left| \mathcal{B}_i(\boldsymbol{Z}\boldsymbol{Z}^\top) - \mathcal{B}_i(\boldsymbol{W}\boldsymbol{W}^\top) \right| \le \alpha_h \text{med}\left(\left| \boldsymbol{y} - \mathcal{B}\left(\boldsymbol{W}\boldsymbol{W}^\top\right) \right| \right) \right\},$$

which is the same as $\mathcal{E}_i$ except that the measurements used to calculate the residual are replaced by clean measurements. In particular, it is straight to see that (4.20) also holds for $\tilde{\mathcal{E}}_i$. Then we can write $\nabla f_{tr}\left(\boldsymbol{W}\right)$ as

$$\begin{aligned}
\nabla f_{tr}\left(\boldsymbol{W}\right) &= \frac{1}{m}\sum_{i=1}^{m}\left(\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right) - y_i\right)\boldsymbol{B}_i\boldsymbol{W}\mathbb{I}_{\mathcal{E}_i} \\
&= \frac{1}{m}\sum_{i\notin\mathcal{S}}\left(\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right) - \mathcal{B}_i\left(\boldsymbol{Z}\boldsymbol{Z}^\top\right)\right)\boldsymbol{B}_i\boldsymbol{W}\mathbb{I}_{\tilde{\mathcal{E}}_i} \\
&\quad + \frac{1}{m}\sum_{i\in\mathcal{S}}\left(\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right) - y_i\right)\boldsymbol{B}_i\boldsymbol{W}\mathbb{I}_{\mathcal{E}_i} \\
&= \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right) - \mathcal{B}_i\left(\boldsymbol{Z}\boldsymbol{Z}^\top\right)\right)\boldsymbol{B}_i\boldsymbol{W}\mathbb{I}_{\tilde{\mathcal{E}}_i}}_{:=\nabla^c f_{tr}(\boldsymbol{W})} \\
&\quad + \underbrace{\frac{1}{m}\sum_{i\in\mathcal{S}}\left[\left(\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right) - y_i\right)\mathbb{I}_{\mathcal{E}_i} - \left(\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right) - \mathcal{B}_i\left(\boldsymbol{Z}\boldsymbol{Z}^\top\right)\right)\mathbb{I}_{\tilde{\mathcal{E}}_i}\right]\boldsymbol{B}_i\boldsymbol{W}}_{:=\nabla^o f_{tr}(\boldsymbol{W})},
\end{aligned}$$

where $\nabla^c f_{tr}\left(\boldsymbol{W}\right)$ corresponds to the truncated gradient *as if* all measurements are clean, and $\nabla^o f_{tr}\left(\boldsymbol{W}\right)$ corresponds to the contribution of the outliers.

For notational simplicity, define

$$\boldsymbol{H} = \begin{bmatrix}\boldsymbol{H}_1\\ \boldsymbol{H}_2\end{bmatrix} = \boldsymbol{W} - \boldsymbol{Z}\boldsymbol{Q} = \begin{bmatrix}\boldsymbol{U} - \boldsymbol{X}\boldsymbol{Q}\\ \boldsymbol{V} - \boldsymbol{Y}\boldsymbol{Q}\end{bmatrix}, \tag{4.28}$$

where $\boldsymbol{Q}$ is given in (4.6). We have

$$\left\langle \nabla^c f_{tr}\left(\boldsymbol{W}\right), \boldsymbol{H}\right\rangle = \frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{B}_i, \boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{Z}\boldsymbol{Z}^\top\rangle \cdot \langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top\rangle \cdot \mathbb{I}_{\tilde{\mathcal{E}}_i}. \tag{4.29}$$

Define the set $\mathcal{D}$ as

$$\mathcal{D} = \left\{ i | \langle \boldsymbol{B}_i, \boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{Z}\boldsymbol{Z}^\top \rangle \cdot \langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top \rangle < 0 \right\}. \tag{4.30}$$

We can then split (4.29) and bound it as

$$\begin{aligned}
&\langle \nabla^c f_{tr}(\boldsymbol{W}), \boldsymbol{H} \rangle \\
&\geq \frac{1}{m} \sum_{i \notin \mathcal{D}} \langle \boldsymbol{B}_i, \boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{Z}\boldsymbol{Z}^\top \rangle \cdot \langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top \rangle \cdot \mathbb{I}_{\left\{ |\langle \boldsymbol{B}_i, \boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{Z}\boldsymbol{Z}^\top \rangle| \leq 0.65\alpha_h \left\| \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top \right\|_{\mathsf{F}} \right\}} \\
&\quad + \frac{1}{m} \sum_{i \in \mathcal{D}} \langle \boldsymbol{B}_i, \boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{Z}\boldsymbol{Z}^\top \rangle \cdot \langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top \rangle \cdot \mathbb{I}_{\left\{ |\langle \boldsymbol{B}_i, \boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{Z}\boldsymbol{Z}^\top \rangle| \leq 0.70\alpha_h \left\| \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top \right\|_{\mathsf{F}} \right\}} \\
&:= B_1 + B_2,
\end{aligned} \tag{4.31}$$

where

$$\begin{aligned}
B_1 &:= \frac{1}{2m} \sum_{i \notin \mathcal{D}} \langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top \rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{H}_1\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{H}_2^\top \rangle \\
&\qquad \cdot \mathbb{I}_{\left\{ |\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top \rangle| \leq 0.65\alpha_h \left\| \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top \right\|_{\mathsf{F}} \right\}}; \\
B_2 &:= \frac{1}{2m} \sum_{i \in \mathcal{D}} \langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top \rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{H}_1\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{H}_2^\top \rangle \\
&\qquad \cdot \mathbb{I}_{\left\{ |\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top \rangle| \leq 0.70\alpha_h \left\| \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top \right\|_{\mathsf{F}} \right\}}.
\end{aligned}$$

The first term in (4.31) can be lower bounded by Proposition 2, whose proof is given in Appendix D.2.

Proposition 2. *Provided $m \geq c_1 nr$, we have*

$$\begin{aligned}
B_1 \geq\ & \frac{\gamma_1}{2} \langle \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top, \boldsymbol{H}_1\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{H}_2^\top \rangle \\
& - 0.0006\alpha_h \left\| \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top \right\|_{\mathsf{F}} \left\| \boldsymbol{H}_1\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{H}_2^\top \right\|_{\mathsf{F}}
\end{aligned} \tag{4.32}$$

holds for all $\boldsymbol{Z}, \boldsymbol{W} \in \mathbb{R}^{(n_1+n_2)\times r}$ with probability at least $1 - \exp(-c_2 m)$, where $\gamma_1 = \mathbb{E}\left[\xi^2 \mathbb{I}_{\{|\xi| \leq 0.65\alpha_h\}}\right]$ with $\xi \sim \mathcal{N}(0,1)$, and $c_1, c_2 > 0$ are numerical constants.

The second term in (4.31) can be lower bounded by Proposition 3, whose proof is given in Appendix D.3.

Proposition 3. *Provided $m \geq c_1 nr$, we have*

$$B_2 \geq -0.36\alpha_h \left\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}} \left\|\boldsymbol{H}_1\boldsymbol{H}_2^\top\right\|_{\mathsf{F}} \tag{4.33}$$

holds for all $\boldsymbol{Z}, \boldsymbol{W} \in \mathbb{R}^{(n_1+n_2)\times r}$ with probability at least $1-\exp\left(-c_2 m\right)$, where $c_1, c_2 > 0$ are numerical constants.

The contribution of outliers $\nabla^o f_{tr}\left(\boldsymbol{W}\right)$ can be bounded by the following proposition, whose proof is given in Appendix D.4.

Proposition 4. *Provided $m \geq c_1 nr \log n$, we have*

$$\left|\langle \nabla^o f_{tr}\left(\boldsymbol{W}\right), \boldsymbol{H}\rangle\right| \leq 0.71\alpha_h\sqrt{s}\left\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top\right\|_{\mathsf{F}} \left\|\boldsymbol{H}_1\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{H}_2^\top\right\|_{\mathsf{F}} \tag{4.34}$$

holds for all $\boldsymbol{Z}, \boldsymbol{W} \in \mathbb{R}^{(n_1+n_2)\times r}$ with probability at least $1-\exp\left(-c_2 m\right)$, where $c_1, c_2 > 0$ are numerical constants.

On the other end, Proposition 5 establishes an upper bound for $\left\|\nabla f_{tr}\left(\boldsymbol{W}\right)\right\|_{\mathsf{F}}^2$, whose proof is given in Appendix D.5.

Proposition 5. *Provided $m \geq c_1 nr \log n$, we have*

$$\left\|\nabla f_{tr}\left(\boldsymbol{W}\right)\right\|_{\mathsf{F}}^2 \leq 0.25\alpha_h^2 \left\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}}^2 \left\|\boldsymbol{W}\right\|^2 \tag{4.35}$$

holds for all $\boldsymbol{Z}, \boldsymbol{W} \in \mathbb{R}^{(n_1+n_2)\times r}$ with probability at least $1-\exp\left(-c_2 m\right)$, where $c_1, c_2 > 0$ are numerical constants.

Moreover, for the regularizer, we have

$$\langle \nabla g(\boldsymbol{W}), \boldsymbol{H}\rangle = \lambda\langle \boldsymbol{D}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{D}\boldsymbol{W}\right), \boldsymbol{H}\rangle$$

$$
\begin{aligned}
&= \lambda\langle \boldsymbol{U}\left(\boldsymbol{U}^\top\boldsymbol{U}-\boldsymbol{V}^\top\boldsymbol{V}\right),\boldsymbol{H}_1\rangle + \lambda\langle \boldsymbol{V}\left(\boldsymbol{V}^\top\boldsymbol{V}-\boldsymbol{U}^\top\boldsymbol{U}\right),\boldsymbol{H}_2\rangle \\
&= \lambda\langle \boldsymbol{U}\boldsymbol{U}^\top,\boldsymbol{H}_1\boldsymbol{U}^\top\rangle + \lambda\langle \boldsymbol{V}\boldsymbol{V}^\top,\boldsymbol{H}_2\boldsymbol{V}^\top\rangle \\
&\quad - \lambda\langle \boldsymbol{U}\boldsymbol{V}^\top,\boldsymbol{H}_1\boldsymbol{V}^\top+\boldsymbol{U}\boldsymbol{H}_2^\top\rangle, \qquad (4.36)
\end{aligned}
$$

and

$$
\begin{aligned}
\|\nabla g(\boldsymbol{W})\|_{\mathsf{F}}^2 &= \lambda^2\left\|\boldsymbol{D}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{D}\boldsymbol{W}\right)\right\|_{\mathsf{F}}^2 \\
&= \lambda^2\left\|\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{D}\boldsymbol{W}\right)\right\|_{\mathsf{F}}^2 \\
&\le \lambda^2\|\boldsymbol{W}\|^2\left\|\boldsymbol{W}^\top\boldsymbol{D}\boldsymbol{W}\right\|_{\mathsf{F}}^2 \\
&= \lambda^2\|\boldsymbol{W}\|^2\left\|(\boldsymbol{H}+\boldsymbol{Z}\boldsymbol{Q})^\top\boldsymbol{D}\left(\boldsymbol{H}+\boldsymbol{Z}\boldsymbol{Q}\right)\right\|_{\mathsf{F}}^2 \\
&= \lambda^2\|\boldsymbol{W}\|^2\left\|\boldsymbol{H}^\top\boldsymbol{D}\boldsymbol{H}+\boldsymbol{H}^\top\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Q}+(\boldsymbol{Z}\boldsymbol{Q})^\top\boldsymbol{D}\boldsymbol{H}\right\|_{\mathsf{F}}^2 \qquad (4.37) \\
&\le \lambda^2\|\boldsymbol{W}\|^2\left(\left\|\boldsymbol{H}^\top\boldsymbol{D}\boldsymbol{H}\right\|_{\mathsf{F}}+2\left\|\boldsymbol{H}^\top\boldsymbol{D}\boldsymbol{Z}\right\|_{\mathsf{F}}\right)^2, \qquad (4.38)
\end{aligned}
$$

where (4.37) follows from $\boldsymbol{X}^\top\boldsymbol{X}=\boldsymbol{Y}^\top\boldsymbol{Y}$.

4.6.4 Certifying Regularity Condition with Sparse Outliers

We are now ready to establish the RC in the neighborhood where $\|\boldsymbol{H}\|_{\mathsf{F}} \le \frac{1}{24}\sigma_r(\boldsymbol{Z})$. Recall that based on Propositions 2, 3 and 4, and (4.36), we have

$$
\begin{aligned}
&\langle\nabla h(\boldsymbol{W}),\boldsymbol{W}-\boldsymbol{Z}\boldsymbol{Q}\rangle \\
&\ge \langle\nabla f_{\mathrm{tr}}(\boldsymbol{W}),\boldsymbol{H}\rangle + \langle\nabla g(\boldsymbol{W}),\boldsymbol{H}\rangle \\
&\ge \langle\nabla f_{\mathrm{tr}}^c(\boldsymbol{W}),\boldsymbol{H}\rangle - |\langle\nabla f_{\mathrm{tr}}^o(\boldsymbol{W}),\boldsymbol{H}\rangle| + \langle\nabla g(\boldsymbol{W}),\boldsymbol{H}\rangle \\
&\ge \frac{\gamma_1}{2}\langle\boldsymbol{U}\boldsymbol{V}^\top-\boldsymbol{X}\boldsymbol{Y}^\top,\boldsymbol{H}_1\boldsymbol{V}^\top+\boldsymbol{U}\boldsymbol{H}_2^\top\rangle - 0.0006\alpha_h\left\|\boldsymbol{U}\boldsymbol{V}^\top-\boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}}\left\|\boldsymbol{H}_1\boldsymbol{V}^\top+\boldsymbol{U}\boldsymbol{H}_2^\top\right\|_{\mathsf{F}} \\
&\quad - 0.36\alpha_h\left\|\boldsymbol{U}\boldsymbol{V}^\top-\boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}}\left\|\boldsymbol{H}_1\boldsymbol{H}_2^\top\right\|_{\mathsf{F}} - 0.71\alpha_h\sqrt{s}\left\|\boldsymbol{U}\boldsymbol{V}^\top-\boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}}\|\boldsymbol{H}_1\boldsymbol{V}^\top+\boldsymbol{U}\boldsymbol{H}_2^\top\|_{\mathsf{F}} \\
&\quad + \langle\nabla g(\boldsymbol{W}),\boldsymbol{H}\rangle. \qquad (4.39)
\end{aligned}
$$

Set $\alpha_h = 6$, and we have $\gamma_1 \approx 0.998348$. Set $\lambda = \gamma_1/4$, then we can write

$$\begin{aligned}
&\frac{\gamma_1}{2}\langle \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top, \boldsymbol{H}_1\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{H}_2^\top\rangle + \langle \nabla g(\boldsymbol{W}), \boldsymbol{H}\rangle \\
&= 2\lambda\langle \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top, \boldsymbol{H}_1\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{H}_2^\top\rangle + \lambda\langle \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{H}_1\boldsymbol{U}^\top\rangle + \lambda\langle \boldsymbol{V}\boldsymbol{V}^\top, \boldsymbol{H}_2\boldsymbol{V}^\top\rangle \\
&\quad - \lambda\langle \boldsymbol{U}\boldsymbol{V}^\top, \boldsymbol{H}_1\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{H}_2^\top\rangle \\
&= \lambda\langle \boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{Z}\boldsymbol{Z}^\top, \boldsymbol{H}\boldsymbol{W}^\top\rangle - \lambda\langle \boldsymbol{X}\boldsymbol{Y}^\top, \boldsymbol{H}_1\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{H}_2^\top\rangle \\
&\quad + \lambda\langle \boldsymbol{X}\boldsymbol{X}^\top, \boldsymbol{H}_1\boldsymbol{U}^\top\rangle + \lambda\langle \boldsymbol{Y}\boldsymbol{Y}^\top, \boldsymbol{H}_2\boldsymbol{V}^\top\rangle, \qquad (4.40)
\end{aligned}$$

where the last three terms can be re-arranged as

$$\begin{aligned}
&\langle \boldsymbol{X}\boldsymbol{X}^\top, \boldsymbol{H}_1\boldsymbol{U}^\top\rangle + \langle \boldsymbol{Y}\boldsymbol{Y}^\top, \boldsymbol{H}_2\boldsymbol{V}^\top\rangle - \langle \boldsymbol{X}\boldsymbol{Y}^\top, \boldsymbol{H}_1\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{H}_2^\top\rangle \\
&= \langle (\boldsymbol{X}\boldsymbol{Q})^\top\boldsymbol{U}, (\boldsymbol{X}\boldsymbol{Q})^\top\boldsymbol{H}_1\rangle + \langle (\boldsymbol{Y}\boldsymbol{Q})^\top\boldsymbol{V}, (\boldsymbol{Y}\boldsymbol{Q})^\top\boldsymbol{H}_2\rangle \\
&\quad - \langle (\boldsymbol{Y}\boldsymbol{Q})^\top\boldsymbol{V}, (\boldsymbol{X}\boldsymbol{Q})^\top\boldsymbol{H}_1\rangle - \langle (\boldsymbol{Y}\boldsymbol{Q})^\top\boldsymbol{H}_2, (\boldsymbol{X}\boldsymbol{Q})^\top\boldsymbol{U}\rangle \\
&= \langle (\boldsymbol{Y}\boldsymbol{Q})^\top\boldsymbol{V}, (\boldsymbol{Y}\boldsymbol{Q})^\top\boldsymbol{H}_2 - (\boldsymbol{X}\boldsymbol{Q})^\top\boldsymbol{H}_1\rangle + \langle (\boldsymbol{X}\boldsymbol{Q})^\top\boldsymbol{H}_1 - (\boldsymbol{Y}\boldsymbol{Q})^\top\boldsymbol{H}_2, (\boldsymbol{X}\boldsymbol{Q})^\top\boldsymbol{U}\rangle \\
&= \langle (\boldsymbol{Y}\boldsymbol{Q})^\top\boldsymbol{V} - (\boldsymbol{X}\boldsymbol{Q})^\top\boldsymbol{U}, (\boldsymbol{Y}\boldsymbol{Q})^\top(\boldsymbol{V} - \boldsymbol{Y}\boldsymbol{Q}) - (\boldsymbol{X}\boldsymbol{Q})^\top(\boldsymbol{U} - \boldsymbol{X}\boldsymbol{Q})\rangle \\
&= \|(\boldsymbol{Y}\boldsymbol{Q})^\top\boldsymbol{V} - (\boldsymbol{X}\boldsymbol{Q})^\top\boldsymbol{U}\|_{\mathsf{F}}^2 = \|\boldsymbol{H}^\top\boldsymbol{D}\boldsymbol{Z}\|_{\mathsf{F}}^2, \qquad (4.41)
\end{aligned}$$

where (4.41) follows from $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{Y}^\top\boldsymbol{Y}$. Moreover, using the facts that $\boldsymbol{H}^\top\boldsymbol{Z}\boldsymbol{Q}$ and $\boldsymbol{H}^\top\boldsymbol{W}$ are symmetric matrices and $\boldsymbol{W}^\top\boldsymbol{Z}\boldsymbol{Q} \succeq 0$ [48], we have the first term in (4.40) bounded as

$$\begin{aligned}
&\langle \boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{Z}\boldsymbol{Z}^\top, \boldsymbol{H}\boldsymbol{W}^\top\rangle \\
&= \|\boldsymbol{H}\boldsymbol{Q}^\top\boldsymbol{Z}^\top\|_{\mathsf{F}}^2 + \|\boldsymbol{H}^\top\boldsymbol{Z}\|_{\mathsf{F}}^2 + \|\boldsymbol{H}\boldsymbol{H}^\top\|_{\mathsf{F}}^2 + 3\langle \boldsymbol{H}\boldsymbol{H}^\top, \boldsymbol{H}\boldsymbol{Q}^\top\boldsymbol{Z}^\top\rangle \\
&\geq \|\boldsymbol{H}\boldsymbol{Q}^\top\boldsymbol{Z}^\top\|_{\mathsf{F}}^2 + \|\boldsymbol{H}^\top\boldsymbol{Z}\|_{\mathsf{F}}^2 + \|\boldsymbol{H}\boldsymbol{H}^\top\|_{\mathsf{F}}^2 - 3\|\boldsymbol{H}\boldsymbol{H}^\top\|_{\mathsf{F}}\|\boldsymbol{H}\boldsymbol{Q}^\top\boldsymbol{Z}^\top\|_{\mathsf{F}} \qquad (4.42) \\
&\geq \|\boldsymbol{H}\boldsymbol{Q}^\top\boldsymbol{Z}^\top\|_{\mathsf{F}}^2 + \|\boldsymbol{H}^\top\boldsymbol{Z}\|_{\mathsf{F}}^2 + \|\boldsymbol{H}\boldsymbol{H}^\top\|_{\mathsf{F}}^2 - \frac{1}{8}\|\boldsymbol{H}\boldsymbol{Q}^\top\boldsymbol{Z}^\top\|_{\mathsf{F}}^2 \qquad (4.43) \\
&\geq \frac{7}{8}\|\boldsymbol{H}\boldsymbol{Q}^\top\boldsymbol{Z}^\top\|_{\mathsf{F}}^2 + \|\boldsymbol{H}^\top\boldsymbol{Z}\|_{\mathsf{F}}^2 + \|\boldsymbol{H}\boldsymbol{H}^\top\|_{\mathsf{F}}^2, \qquad (4.44)
\end{aligned}$$

where (4.42) follows from Cauchy-Schwarz inequality, (4.43) follows from $\left\|\boldsymbol{H}\boldsymbol{H}^{\top}\right\|_{\mathsf{F}} \leq \|\boldsymbol{H}\|_{\mathsf{F}}^2 \leq \frac{1}{24}\sigma_r\left(\boldsymbol{Z}\boldsymbol{Q}\right)\|\boldsymbol{H}\|_{\mathsf{F}} \leq \frac{1}{24}\left\|\boldsymbol{H}\boldsymbol{Q}^{\top}\boldsymbol{Z}^{\top}\right\|_{\mathsf{F}}$, where we used $\|\boldsymbol{H}\|_{\mathsf{F}} \leq \frac{1}{24}\sigma_r\left(\boldsymbol{Z}\right) = \frac{1}{24}\sigma_r\left(\boldsymbol{Z}\boldsymbol{Q}\right)$. In addition, we have

$$\begin{aligned}\|\boldsymbol{H}_1\boldsymbol{V}^{\top} + \boldsymbol{U}\boldsymbol{H}_2^{\top}\|_{\mathsf{F}} &\leq \sqrt{2}\|\boldsymbol{H}\boldsymbol{W}^{\top}\|_{\mathsf{F}} \\ &\leq \sqrt{2}\|\boldsymbol{H}\boldsymbol{H}\|_{\mathsf{F}} + \sqrt{2}\|\boldsymbol{H}\boldsymbol{Q}^{\top}\boldsymbol{Z}^{\top}\|_{\mathsf{F}} \\ &\leq \frac{25}{24}\sqrt{2}\|\boldsymbol{H}\boldsymbol{Q}^{\top}\boldsymbol{Z}^{\top}\|_{\mathsf{F}}, \end{aligned} \tag{4.45}$$

and

$$\begin{aligned}\left\|\boldsymbol{U}\boldsymbol{V}^{\top} - \boldsymbol{X}\boldsymbol{Y}^{\top}\right\|_{\mathsf{F}} &\leq \frac{1}{\sqrt{2}}\left\|\boldsymbol{W}\boldsymbol{W}^{\top} - \boldsymbol{Z}\boldsymbol{Z}^{\top}\right\|_{\mathsf{F}} \\ &= \frac{1}{\sqrt{2}}\left\|\boldsymbol{H}\boldsymbol{H}^{\top} + \boldsymbol{Z}\boldsymbol{Q}\boldsymbol{H}^{\top} + \boldsymbol{H}\left(\boldsymbol{Z}\boldsymbol{Q}\right)^{\top}\right\|_{\mathsf{F}} \\ &\leq \frac{1}{\sqrt{2}}\left\|\boldsymbol{H}\boldsymbol{H}^{\top}\right\|_{\mathsf{F}} + \sqrt{2}\left\|\boldsymbol{H}\boldsymbol{Q}^{\top}\boldsymbol{Z}^{\top}\right\|_{\mathsf{F}} \\ &\leq \frac{49}{48}\sqrt{2}\left\|\boldsymbol{H}\boldsymbol{Q}^{\top}\boldsymbol{Z}^{\top}\right\|_{\mathsf{F}}, \end{aligned} \tag{4.46}$$

and

$$\left\|\boldsymbol{H}_1\boldsymbol{H}_2^{\top}\right\|_{\mathsf{F}} \leq \frac{1}{\sqrt{2}}\left\|\boldsymbol{H}\boldsymbol{H}^{\top}\right\|_{\mathsf{F}}. \tag{4.47}$$

Plugging (4.44), (4.45), (4.46) and (4.47) into (4.39), we have

$$\begin{aligned}&\left\langle \nabla h\left(\boldsymbol{W}\right), \boldsymbol{W} - \boldsymbol{Z}\boldsymbol{Q}\right\rangle \\ &\geq \left[\frac{7}{8}\lambda - (0.0006 + 0.71\sqrt{s})\alpha_h\frac{25\cdot 49}{24^2} - 0.36\alpha_h\frac{49}{2\cdot 24^2}\right]\|\boldsymbol{H}\boldsymbol{Q}^{\top}\boldsymbol{Z}^{\top}\|_{\mathsf{F}}^2 \\ &\quad + \lambda\|\boldsymbol{H}^{\top}\boldsymbol{Z}\|_{\mathsf{F}}^2 + \lambda\left\|\boldsymbol{H}^{\top}\boldsymbol{D}\boldsymbol{Z}\right\|_{\mathsf{F}}^2 + \lambda\|\boldsymbol{H}\boldsymbol{H}^{\top}\|_{\mathsf{F}}^2 \\ &\geq \left(0.1188 - 9.06\sqrt{s}\right)\|\boldsymbol{H}\boldsymbol{Q}^{\top}\boldsymbol{Z}^{\top}\|_{\mathsf{F}}^2 + \lambda\|\boldsymbol{H}^{\top}\boldsymbol{Z}\|_{\mathsf{F}}^2 + \lambda\left\|\boldsymbol{H}^{\top}\boldsymbol{D}\boldsymbol{Z}\right\|_{\mathsf{F}}^2 \\ &\quad + \lambda\|\boldsymbol{H}\boldsymbol{H}^{\top}\|_{\mathsf{F}}^2, \end{aligned} \tag{4.48}$$

where (4.48) follows from the setting $\alpha_h = 6$ and $\lambda = \gamma_1/4$.

On the other end, since

$$\left\|\boldsymbol{H}^{\top}\boldsymbol{D}\boldsymbol{H}\right\|_{\mathsf{F}}^{2}=\left\|\boldsymbol{H}_{1}^{\top}\boldsymbol{H}_{1}-\boldsymbol{H}_{2}^{\top}\boldsymbol{H}_{2}\right\|_{\mathsf{F}}^{2}\leq 2\left(\left\|\boldsymbol{H}_{1}\boldsymbol{H}_{1}^{\top}\right\|_{\mathsf{F}}^{2}+\left\|\boldsymbol{H}_{2}\boldsymbol{H}_{2}^{\top}\right\|_{\mathsf{F}}^{2}\right)\leq 2\left\|\boldsymbol{H}\boldsymbol{H}^{\top}\right\|_{\mathsf{F}}^{2},$$

from Proposition 5 and (4.38) we have

$$\begin{aligned}
&\left\|\nabla h\left(\boldsymbol{W}\right)\right\|_{\mathsf{F}}^{2}\\
&\leq 2\left\|\nabla f_{tr}\left(\boldsymbol{W}\right)\right\|_{\mathsf{F}}^{2}+2\left\|\nabla g\left(\boldsymbol{W}\right)\right\|_{\mathsf{F}}^{2}\\
&\leq 0.5\alpha_{h}^{2}\left\|\boldsymbol{U}\boldsymbol{V}^{\top}-\boldsymbol{X}\boldsymbol{Y}^{\top}\right\|_{\mathsf{F}}^{2}\left\|\boldsymbol{W}\right\|^{2}+2\lambda^{2}\left\|\boldsymbol{W}\right\|^{2}\left(\left\|\boldsymbol{H}^{\top}\boldsymbol{D}\boldsymbol{H}\right\|_{\mathsf{F}}+2\left\|\boldsymbol{H}^{\top}\boldsymbol{D}\boldsymbol{Z}\right\|_{\mathsf{F}}\right)^{2}\\
&\leq\left(0.5\alpha_{h}^{2}\left\|\boldsymbol{U}\boldsymbol{V}^{\top}-\boldsymbol{X}\boldsymbol{Y}^{\top}\right\|_{\mathsf{F}}^{2}+4\lambda^{2}\left\|\boldsymbol{H}^{\top}\boldsymbol{D}\boldsymbol{H}\right\|_{\mathsf{F}}^{2}+16\lambda^{2}\left\|\boldsymbol{H}^{\top}\boldsymbol{D}\boldsymbol{Z}\right\|_{\mathsf{F}}^{2}\right)\left\|\boldsymbol{W}\right\|^{2}\\
&\leq\left(0.5\alpha_{h}^{2}2\left(\frac{49}{48}\right)^{2}\left\|\boldsymbol{H}\boldsymbol{Q}^{\top}\boldsymbol{Z}^{\top}\right\|_{\mathsf{F}}^{2}+8\lambda^{2}\left\|\boldsymbol{H}\boldsymbol{H}^{\top}\right\|_{\mathsf{F}}^{2}+16\lambda^{2}\left\|\boldsymbol{H}^{\top}\boldsymbol{D}\boldsymbol{Z}\right\|_{\mathsf{F}}^{2}\right)\left(\frac{25}{24}\right)^{2}\left\|\boldsymbol{Z}\right\|^{2}\\
&\leq\left(40.8\left\|\boldsymbol{H}\boldsymbol{Q}^{\top}\boldsymbol{Z}^{\top}\right\|_{\mathsf{F}}^{2}+1.1\left\|\boldsymbol{H}^{\top}\boldsymbol{D}\boldsymbol{Z}\right\|_{\mathsf{F}}^{2}\right)\left\|\boldsymbol{Z}\right\|^{2}.
\end{aligned}$$

Therefore, if we let $\alpha=20$ and $\beta=1000$, we have the right hand side of RC as

$$\begin{aligned}
&\frac{\sigma_{r}^{2}\left(\boldsymbol{Z}\right)}{\alpha}\left\|\boldsymbol{H}\right\|_{\mathsf{F}}^{2}+\frac{1}{\beta\left\|\boldsymbol{Z}\right\|^{2}}\left\|\nabla h\left(\boldsymbol{W}\right)\right\|_{\mathsf{F}}^{2}\\
&\leq\frac{\sigma_{r}^{2}\left(\boldsymbol{Z}\right)}{20}\left\|\boldsymbol{H}\right\|_{\mathsf{F}}^{2}+0.0408\left\|\boldsymbol{H}\boldsymbol{Q}^{\top}\boldsymbol{Z}^{\top}\right\|_{\mathsf{F}}^{2}+0.0011\left\|\boldsymbol{H}^{\top}\boldsymbol{D}\boldsymbol{Z}\right\|_{\mathsf{F}}^{2}\\
&\leq 0.0908\left\|\boldsymbol{H}\boldsymbol{Q}^{\top}\boldsymbol{Z}^{\top}\right\|_{\mathsf{F}}^{2}+0.0011\left\|\boldsymbol{H}^{\top}\boldsymbol{D}\boldsymbol{Z}\right\|_{\mathsf{F}}^{2}.
\end{aligned}$$

Consequently, matching it with the (4.48), we conclude that when s is a sufficiently small constant, RC holds with parameters $(20,100,\sigma_{r}(\boldsymbol{Z})/24)$. Note that the parameters α,β,s have not been optimized in the proof.

4.7 Proof of Robust Initialization

As in the description of Algorithm 4, we split the samples into two portions $\{\boldsymbol{y}_{1},\boldsymbol{y}_{2}\}$ in the initialization stage for the convenience of theoretical analysis. We use the measurements $\boldsymbol{y}_{2}=\{y_{i}\}_{i=m_{1}+1}^{m}$ to estimate $\left\|\boldsymbol{M}\right\|_{\mathsf{F}}$ via the sample median of

$\boldsymbol{y}_2$. Then, we employ the rest of measurements $\boldsymbol{y}_1 = \{y_i\}_{i=1}^{m_1}$ to generate initialization via the truncated spectral method. Besides, denote the outlier fraction of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ by $s_1 = |\mathcal{S}_1|/m_1$ and $s_2 = |\mathcal{S}_2|/m_2$, respectively, where $\mathcal{S}_1$ and $\mathcal{S}_2$ are the corresponding outlier supports of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. Hence, $\max\{s_1, s_2\} \leq 2s$.

Due to Lemma 22, provided s_2 is small, we have

$$\theta_{\frac{1}{2}-s_2}\left(\{|\mathcal{A}_i(\boldsymbol{M})|\}_{i=m_1+1}^{m}\right) \leq \mathrm{med}\left(|\boldsymbol{y}_2|\right) \leq \theta_{\frac{1}{2}+s_2}\left(\{|\mathcal{A}_i(\boldsymbol{M})|\}_{i=m_1+1}^{m}\right). \tag{4.49}$$

Following Proposition 1, if $s_2 \leq 2s < 0.01$, we have that provided $m \geq c_1 nr \log n$ for some large constant c_1,

$$0.65\,\|\boldsymbol{M}\|_{\mathsf{F}} \leq \mathrm{med}\left(|\boldsymbol{y}_2|\right) \leq 0.70\,\|\boldsymbol{M}\|_{\mathsf{F}} \tag{4.50}$$

holds with probability at least $1 - \exp\left(-c_2 m\right)$ for some constant c_2.

Therefore, (4.50) guarantees that the threshold used in the truncation is on the order of $\|\boldsymbol{M}\|_{\mathsf{F}}$. To emphasize the independence between the measurements used for norm estimation via the sample median and the rest of the measurements used in the truncated spectral method, we define $C_M := \mathrm{med}\left(|\boldsymbol{y}_2|\right)$, which satisfies (4.50). Rewrite (4.15) as

$$\boldsymbol{K} = (1 - s_1)\boldsymbol{K}_1 + s_1 \boldsymbol{K}_2$$

where

$$\boldsymbol{K}_1 = \frac{1}{|\mathcal{S}_1^c|}\sum_{i\in\mathcal{S}_1^c} \mathcal{A}_i(\boldsymbol{M})\boldsymbol{A}_i \mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})|\leq \alpha_y C_M\}}, \quad \boldsymbol{K}_2 = \frac{1}{|\mathcal{S}_1|}\sum_{i\in\mathcal{S}_1} y_i \boldsymbol{A}_i \mathbb{I}_{\{|y_i|\leq \alpha_y C_M\}}, \tag{4.51}$$

where $\mathcal{S}_1^c$ is the the complementary set of $\mathcal{S}_1$. Note that

$$\mathbb{E}[\boldsymbol{K}_1] = \frac{1}{|\mathcal{S}_1^c|}\sum_{i\in\mathcal{S}_1^c} \mathbb{E}\left[\left(\mathcal{A}_i(\boldsymbol{M})\right)\boldsymbol{A}_i \mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})|\leq \alpha_y C_M\}}\right] = \gamma_2 \boldsymbol{M},$$

where $\gamma_2 := \mathbb{E}\left[\xi^2 \mathbb{I}_{\left\{|\xi| \le \alpha_y C_M / \|\boldsymbol{M}\|_{\mathsf{F}}\right\}}\right] \le 1$ with $\xi \sim \mathcal{N}(0,1)$, and

$$\mathbb{E}[\boldsymbol{K}_2] = \frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} y_i \mathbb{E}[\boldsymbol{A}_i] \mathbb{I}_{\{|y_i| \le \alpha_y C_M\}} = \boldsymbol{0}.$$

We have the following proposition on the concentration of $\boldsymbol{K}$, of which the proof is given in Appendix D.6.

Proposition 6. *With probability at least $1 - n^{-c_1}$, we have*

$$\|\boldsymbol{K} - (1 - s_1)\gamma_2 \boldsymbol{M}\| \le C\alpha_y \sqrt{\frac{n \log n}{m}} \|\boldsymbol{M}\|_{\mathsf{F}}, \tag{4.52}$$

provided that $m \ge c_2 \log n$, where $c_1, c_2, C > 1$ are numerical constants.

Let $\epsilon := C\alpha_y \sqrt{\frac{n \log n}{m}}$ for short-hand notations. Denote $\tilde{n} = \min\{n_1, n_2\}$. Let $\sigma_1(\boldsymbol{K}) \ge \sigma_2(\boldsymbol{K}) \ge \cdots \sigma_{\tilde{n}}(\boldsymbol{K})$ be the singular values of $\boldsymbol{K}$ in a nonincreasing order, and $\sigma_1(\boldsymbol{M}) \ge \sigma_2(\boldsymbol{M}) \ge \cdots \sigma_{\tilde{n}}(\boldsymbol{M})$ be the singular values of $\boldsymbol{M}$ in a nonincreasing order. Since $\boldsymbol{M}$ has rank r, we know $\sigma_{r+1}(\boldsymbol{M}) = \cdots = \sigma_{\tilde{n}}(\boldsymbol{M}) = 0$. By the Weyl's inequality and (4.52), we have

$$|\sigma_i(\boldsymbol{K}) - (1 - s_1)\gamma_2 \sigma_i(\boldsymbol{M})| \le \epsilon \|\boldsymbol{M}\|_{\mathsf{F}}, \quad i = 1, \cdots, \tilde{n}, \tag{4.53}$$

which implies

$$\sigma_i(\boldsymbol{K}) \le \epsilon \|\boldsymbol{M}\|_{\mathsf{F}}, \quad i \ge (r+1). \tag{4.54}$$

By definition, $\boldsymbol{U}_0 = \boldsymbol{C}_L \boldsymbol{\Sigma}^{1/2}$, $\boldsymbol{V}_0 = \boldsymbol{C}_R \boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{W}_0 = \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$, where $\boldsymbol{C}_L \boldsymbol{\Sigma} \boldsymbol{C}_R^\top :=$ rank-r SVD of $\boldsymbol{K}$, with $\boldsymbol{C}_L \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{C}_R \in \mathbb{R}^{n_2 \times r}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$. Recall $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix}$, then according to Lemma 23, we have

$$\begin{aligned} \left\|\boldsymbol{W}_0 \boldsymbol{W}_0^\top - \boldsymbol{Z}\boldsymbol{Z}^\top\right\|_{\mathsf{F}} &\le 2 \left\|\boldsymbol{U}_0 \boldsymbol{V}_0^\top - \boldsymbol{M}\right\|_{\mathsf{F}} \\ &= 2 \left\|\boldsymbol{C}_L \boldsymbol{\Sigma} \boldsymbol{C}_R^\top - (1 - s_1)\gamma_2 \boldsymbol{M}\right\|_{\mathsf{F}} + 2 \left\|\left((1 - s_1)\gamma_2 - 1\right) \boldsymbol{M}\right\|_{\mathsf{F}} \end{aligned}$$

$$
\begin{aligned}
&\leq 2\sqrt{2r}\left(\left\|\boldsymbol{C}_L\boldsymbol{\Sigma}\boldsymbol{C}_R^\top-\boldsymbol{K}\right\|+\|\boldsymbol{K}-(1-s_1)\gamma_2\boldsymbol{M}\|\right)\\
&\quad+2\left|(1-s_1)\gamma_2-1\right|\cdot\|\boldsymbol{M}\|_{\mathsf{F}}\\
&\leq 2\sqrt{2r}\left(\sigma_{r+1}(\boldsymbol{K})+\epsilon\|\boldsymbol{M}\|_{\mathsf{F}}\right)+2\left|(1-s_1)\gamma_2-1\right|\cdot\|\boldsymbol{M}\|_{\mathsf{F}}\\
&\leq \left(4\sqrt{2r}\epsilon+2s_1\gamma_2+2(1-\gamma_2)\right)\|\boldsymbol{M}\|_{\mathsf{F}}.
\end{aligned}
$$

By Lemma 12, we have

$$
\begin{aligned}
\operatorname{dist}\left(\boldsymbol{W}_0,\boldsymbol{Z}\right)&\leq\frac{\left\|\boldsymbol{W}_0\boldsymbol{W}_0^\top-\boldsymbol{Z}\boldsymbol{Z}^\top\right\|_{\mathsf{F}}}{\sqrt{2\left(\sqrt{2}-1\right)}\sigma_r\left(\boldsymbol{Z}\right)}\\
&\leq\frac{\left(4\sqrt{2r}\epsilon+2s_1\gamma_2+2(1-\gamma_2)\right)\|\boldsymbol{M}\|_{\mathsf{F}}}{\sqrt{2\left(\sqrt{2}-1\right)}\sigma_r\left(\boldsymbol{Z}\right)}\\
&=\frac{\left(2\sqrt{2r}\epsilon+s_1\gamma_2+(1-\gamma_2)\right)\|\boldsymbol{M}\|_{\mathsf{F}}}{\sqrt{\sqrt{2}-1}\sqrt{\sigma_r\left(\boldsymbol{M}\right)}},
\end{aligned}
$$

where we use the fact that for all i, $\sigma_i(\boldsymbol{X})=\sigma_i(\boldsymbol{Y})=\sigma_i(\boldsymbol{Z})/\sqrt{2}=\sqrt{\sigma_i\left(\boldsymbol{M}\right)}$.

Therefore, we have $\operatorname{dist}\left(\boldsymbol{W}_0,\boldsymbol{Z}\right)\leq\frac{1}{24}\sigma_r\left(\boldsymbol{Z}\right)$ if

$$
\max\{\sqrt{r}\epsilon,s_1,1-\gamma_2\}\leq c\frac{\sigma_r(\boldsymbol{M})}{\|\boldsymbol{M}\|_{\mathsf{F}}}=\frac{c}{\sqrt{r}\bar{\kappa}}.
$$

To be more specific, we need $s_1<2s\leq c_1/(\sqrt{r}\bar{\kappa})$, $m>c_2\alpha_y^2nr^2\bar{\kappa}^2\log n$, and

$$
1-\gamma_2=\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\xi^2\mathbb{I}_{\left\{|\xi|>\alpha_yC_M/\|\boldsymbol{M}\|_{\mathsf{F}}\right\}}\right]\leq\frac{1}{35\sqrt{r}\bar{\kappa}}.
$$

The last condition can be satisfied by setting $\alpha_y=2\log\left(r^{1/4}\bar{\kappa}_0^{1/2}+20\right)$, as long as $\bar{\kappa}_0$ is an upper bound of $\bar{\kappa}$ such that $\bar{\kappa}\leq\bar{\kappa}_0$.

4.8 Conclusion

In this chapter, we present a median-truncated gradient descent algorithm to improve the robustness of low-rank matrix recovery from random linear measurements in

the presence of outliers. The effectiveness of the proposed algorithm is provably guaranteed by theoretical analysis, and validated through various numerical experiments as well.

Chapter 5: Future Work

The current work in this dissertation yields several intriguing open problems and potential directions for future research work.

First, in the work of low-rank PSD matrix recovery from clean rank-one measurements, as described in Chapter 2, when taking low-rank factorization in (2.1), we assume knowing the true rank r, and consequently, the obtained low-rank factor has full column rank. However, in practice, it is more common to know only an upper bound of the rank of the underlying ground truth, and in such a case, the current theoretical analysis may no longer be valid. Instead, based on the notations in Chapter 2, we consider employing gradient descent to solve an nonconvex optimization problem over a regularized loss function, formulated as

$$\min_{\boldsymbol{X}} \frac{1}{4m} \sum_{i=1}^{m} \left(y_i - \left\| \boldsymbol{a}_i^\top \boldsymbol{X} \right\|_2^2 \right)^2 + \tau \left\| \boldsymbol{X} \right\|_{\mathsf{F}}^2 . \tag{5.1}$$

It is interesting to justify the performance of the designed algorithm when the rank is over-estimated by taking advantage of the powerful leave-one-out strategy.

Next, even though we have proposed a nonconvex algorithm based on subgradient descent for robust low-rank PSD matrix recovery when the rank-one measurements are corrupted by arbitrary outliers, as described in Chapter 3, the theoretical analysis still lacks and we hope to close this gap in the future. Moreover, owing to the efficiency

of median truncation to mitigate the impact of outliers, which has be demonstrated in the full-rank linear sensing model in Chapter 4, we anticipate that such a modified gradient descent can work well with the rank-one sensing model as well to improve the recovery robustness.

Finally, another potential research direction is to solve the low-rank matrix recovery problems via stochastic gradient descent (SGD). SGD only needs to calculate the gradient of a single sample or a batch of few samples during each iteration, hence, it requires much less computational cost and storage space, compared with standard gradient descent using all of the samples, and has the potential to be adopted on online streaming data [118]. So it is of practical importance to consider SGD for low-rank matrix recovery and characterize the corresponding performance guarantees.

Appendix A: Supportive Lemmas

In this section, we document several useful technical lemmas that are used throughout the proofs.

Lemma 12. *[63, Lemma 5.4] For any matrices $\boldsymbol{X}, \boldsymbol{U} \in \mathbb{R}^{n\times r}$, we have*

$$\left\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top\right\|_{\mathsf{F}} \geq \sqrt{2(\sqrt{2}-1)}\sigma_r\left(\boldsymbol{X}\right)\mathrm{dist}(\boldsymbol{X},\boldsymbol{U}),$$

where $\mathrm{dist}(\boldsymbol{X},\boldsymbol{U}) := \min_{\boldsymbol{P}\in\mathcal{O}^{r\times r}} \|\boldsymbol{X}\boldsymbol{P}-\boldsymbol{U}\|_{\mathsf{F}}$.

Lemma 13 (Covering number for low-rank matrices)**.** *[16, Lemma 3.1] Let $\mathcal{S}_r = \{\boldsymbol{X} \in \mathbb{R}^{n_1\times n_2}, \mathrm{rank}(\boldsymbol{X}) \leq r, \|\boldsymbol{X}\|_{\mathsf{F}} = 1\}$. Then there exists an ϵ-net $\bar{\mathcal{S}}_r \subset \mathcal{S}_r$ with respect to the Frobenius norm whose cardinality obeys*

$$\left|\bar{\mathcal{S}}_r\right| \leq (9/\epsilon)^{(n_1+n_2+1)r}.$$

Lemma 14. *[25, 119] Suppose $x_1, \cdots, x_m$ are i.i.d. real-valued random variables obeying $x_i \leq b$ for some deterministic number $b > 0$, $\mathbb{E}[x_i] = 0$, and $\mathbb{E}[x_i^2] = d^2$. Setting $\sigma^2 = m\cdot\max\{b^2, d^2\}$, we have*

$$\mathbb{P}\left(\sum_{i=1}^m x_i \geq t\right) \leq \min\left\{\exp\left(-\frac{t^2}{2\sigma^2}\right), 25\left(1-\Phi\left(\frac{t}{\sigma}\right)\right)\right\},$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard Gaussian variable.

Lemma 15. *[120, Theorem 5.39] Suppose the $\boldsymbol{a}_i$'s are i.i.d. random vectors following $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $i = 1, \cdots, m$. Then for every $t \geq 0$ and $0 < \delta \leq 1$,*

$$\left\| \boldsymbol{I}_n - \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{a}_i \boldsymbol{a}_i^\top \right\| \leq \delta$$

holds with probability at least $1 - 2e^{-ct^2}$, where $\delta = C\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}$. On this event, for all $\boldsymbol{W} \in \mathbb{R}^{n \times r}$, there exists

$$\left| \frac{1}{m} \sum_{i=1}^{m} \left\| \boldsymbol{a}_i^\top \boldsymbol{W} \right\|_2^2 - \|\boldsymbol{W}\|_{\mathsf{F}}^2 \right| \leq \delta \|\boldsymbol{W}\|_{\mathsf{F}}^2 .$$

Lemma 16. *[25] Suppose the $\boldsymbol{a}_i$'s are i.i.d. random vectors following $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $i = 1, \cdots, m$. Then with probability at least $1 - me^{-1.5n}$, we have*

$$\max_{1 \leq i \leq m} \|\boldsymbol{a}_i\|_2 \leq \sqrt{6n}.$$

Lemma 17. *Fix $\boldsymbol{W} \in \mathbb{R}^{n \times r}$. Suppose the $\boldsymbol{a}_i$'s are i.i.d. random vectors following $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $i = 1, \cdots, m$. Then with probability at least $1 - mrn^{-13}$, we have*

$$\max_{1 \leq i \leq m} \left\| \boldsymbol{a}_i^\top \boldsymbol{W} \right\|_2 \leq 5.86 \sqrt{\log n} \|\boldsymbol{W}\|_{\mathsf{F}} .$$

Proof. Define $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_r]$, then we can write $\left\| \boldsymbol{a}_i^\top \boldsymbol{W} \right\|_2^2 = \sum_{k=1}^{r} \left(\boldsymbol{a}_i^\top \boldsymbol{w}_k \right)^2$. Recognize that $\left(\boldsymbol{a}_i^\top \frac{\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|_2} \right)^2$ follows the χ^2 distribution with 1 degree of freedom. It then follows from [121, Lemma 1] that

$$\mathbb{P}\left(\left(\boldsymbol{a}_i^\top \frac{\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|_2} \right)^2 \geq 1 + 2\sqrt{t} + 2t \right) \leq \exp(-t),$$

for any $t > 0$. Taking $t = 13 \log n$ yields

$$\mathbb{P}\left(\left(\boldsymbol{a}_i^\top \boldsymbol{w}_k \right)^2 \leq 34.3 \|\boldsymbol{w}_k\|_2^2 \log n \right) \geq 1 - n^{-13}.$$

Finally, taking the union bound, we obtain

$$\max_{1 \leq i \leq m} \left\| \boldsymbol{a}_i^\top \boldsymbol{W} \right\|_2^2 \leq \sum_{k=1}^{r} 34.3 \|\boldsymbol{w}_k\|_2^2 \log n = 34.3 \|\boldsymbol{W}\|_{\mathsf{F}}^2 \log n$$

with probability at least $1 - mrn^{-13}$. □

Lemma 18. *Suppose $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Then for any fixed matrices $\boldsymbol{X}$, $\boldsymbol{H} \in \mathbb{R}^{n\times r}$, we have*

$$\mathbb{E}\left[\left\|\boldsymbol{a}^\top \boldsymbol{H}\right\|_2^2 \left\|\boldsymbol{a}^\top \boldsymbol{X}\right\|_2^2\right] = \left\|\boldsymbol{H}\right\|_{\mathsf{F}}^2 \left\|\boldsymbol{X}\right\|_{\mathsf{F}}^2 + 2\left\|\boldsymbol{H}^\top \boldsymbol{X}\right\|_{\mathsf{F}}^2;$$
$$\mathbb{E}\left[\left(\boldsymbol{a}^\top \boldsymbol{H}\boldsymbol{X}^\top \boldsymbol{a}\right)^2\right] = \left(\mathrm{Tr}\big(\boldsymbol{H}^\top \boldsymbol{X}\big)\right)^2 + \mathrm{Tr}\big(\boldsymbol{H}^\top \boldsymbol{X}\boldsymbol{H}^\top \boldsymbol{X}\big) + \left\|\boldsymbol{H}\boldsymbol{X}^\top\right\|_{\mathsf{F}}^2.$$

Moreover, for any order $k \geq 1$, we have $\mathbb{E}\big[\left\|\boldsymbol{a}^\top \boldsymbol{H}\right\|_2^{2k}\big] \leq c_k \left\|\boldsymbol{H}\right\|_{\mathsf{F}}^{2k}$, where $c_k > 0$ is a numerical constant that depends only on k.

Proof. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_r]$ and $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_r]$. Based on the simple facts

$$\mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{a})^2 \boldsymbol{a}\boldsymbol{a}^\top\right] = \left\|\boldsymbol{x}\right\|_2^2 \boldsymbol{I}_n + 2\boldsymbol{x}\boldsymbol{x}^\top,$$
$$\mathbb{E}\left[(\boldsymbol{a}^\top \boldsymbol{x}_i)(\boldsymbol{a}^\top \boldsymbol{x}_j)\boldsymbol{a}\boldsymbol{a}^\top\right] = \boldsymbol{x}_i \boldsymbol{x}_j^\top + \boldsymbol{x}_j \boldsymbol{x}_i^\top + \boldsymbol{x}_i^\top \boldsymbol{x}_j \boldsymbol{I}_n,$$

for any fixed vectors $\boldsymbol{x}$, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j \in \mathbb{R}^n$, we can derive

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{a}^\top \boldsymbol{H}\right\|_2^2 \left\|\boldsymbol{a}^\top \boldsymbol{X}\right\|_2^2\right] &= \sum_{i=1}^r \sum_{j=1}^r \mathbb{E}\left[\left(\boldsymbol{a}^\top \boldsymbol{h}_i\right)^2 \left(\boldsymbol{a}^\top \boldsymbol{x}_j\right)^2\right] \\
&= \sum_{i=1}^r \sum_{j=1}^r \left[\left\|\boldsymbol{h}_i\right\|_2^2 \left\|\boldsymbol{x}_j\right\|_2^2 + 2\left(\boldsymbol{h}_i^\top \boldsymbol{x}_j\right)^2\right] \\
&= \left\|\boldsymbol{H}\right\|_{\mathsf{F}}^2 \left\|\boldsymbol{X}\right\|_{\mathsf{F}}^2 + 2\left\|\boldsymbol{H}^\top \boldsymbol{X}\right\|_{\mathsf{F}}^2,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[\left(\boldsymbol{a}^\top \boldsymbol{H}\boldsymbol{X}^\top \boldsymbol{a}\right)^2\right] &= \mathbb{E}\left[\sum_{i=1}^r \left(\boldsymbol{a}^\top \boldsymbol{h}_i\right)^2 \left(\boldsymbol{a}^\top \boldsymbol{x}_i\right)^2 + \sum_{i\neq j} \left(\boldsymbol{a}^\top \boldsymbol{h}_i\right)\left(\boldsymbol{a}^\top \boldsymbol{x}_i\right)\left(\boldsymbol{a}^\top \boldsymbol{h}_j\right)\left(\boldsymbol{a}^\top \boldsymbol{x}_j\right)\right] \\
&= \sum_{i=1}^r \left[\left\|\boldsymbol{h}_i\right\|_2^2 \left\|\boldsymbol{x}_i\right\|_2^2 + 2\left(\boldsymbol{h}_i^\top \boldsymbol{x}_i\right)^2\right] \\
&\quad + \sum_{i\neq j} \left[\left(\boldsymbol{h}_i^\top \boldsymbol{x}_i\right)\left(\boldsymbol{h}_j^\top \boldsymbol{x}_j\right) + \left(\boldsymbol{h}_i^\top \boldsymbol{h}_j\right)\left(\boldsymbol{x}_i^\top \boldsymbol{x}_j\right) + \left(\boldsymbol{h}_i^\top \boldsymbol{x}_j\right)\left(\boldsymbol{x}_i^\top \boldsymbol{h}_j\right)\right] \\
&= \left(\mathrm{Tr}\big(\boldsymbol{H}^\top \boldsymbol{X}\big)\right)^2 + \left\|\boldsymbol{H}\boldsymbol{X}^\top\right\|_{\mathsf{F}}^2 + \mathrm{Tr}\big(\boldsymbol{H}^\top \boldsymbol{X}\boldsymbol{H}^\top \boldsymbol{X}\big).
\end{aligned}$$

Finally, to bound $\mathbb{E}\left[\left\|\boldsymbol{a}^\top \boldsymbol{H}\right\|_2^{2k}\right]$ for an arbitrary $\boldsymbol{H} \in \mathbb{R}^{n\times r}$, we write the singular value decomposition of $\boldsymbol{H}$ as $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_r] \in \mathbb{R}^{n\times r}$, $\boldsymbol{\Sigma} = \operatorname{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_r\}$, and $\boldsymbol{V} \in \mathbb{R}^{r\times r}$. This gives

$$\left\|\boldsymbol{a}^\top \boldsymbol{H}\right\|_2^2 = \sum_{i=1}^r \sigma_i^2 (\boldsymbol{a}^\top \boldsymbol{u}_i)^2.$$

Let $b_i = \sigma_i \boldsymbol{a}^\top \boldsymbol{u}_i$ for $i = 1, \cdots, r$, which are independent random variables obeying $b_i \sim \mathcal{N}(0, \sigma_i^2)$ due to the fact $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_r$. Since $\mathbb{E}[b_i^{2t}] = \sigma_i^{2t}(2t-1)!! \le c_k \sigma_i^{2t}$ for any $i = 1, \cdots, r$ and $t = 1, \cdots, k$, where c_k is some large enough constant depending only on k, we arrive at

$$\mathbb{E}\left[\left(\sum_{i=1}^r b_i^2\right)^k\right] \le c_k \left(\sum_{i=1}^r \sigma_i^2\right)^k = c_k \left\|\boldsymbol{H}\right\|_{\mathsf{F}}^{2k},$$

as claimed. □

Lemma 19. *Fix $\boldsymbol{X}^\natural \in \mathbb{R}^{n\times r}$. Suppose the $\boldsymbol{a}_i$'s are i.i.d. random vectors following $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $i = 1, \cdots, m$. For any $0 < \delta \le 1$, suppose $m \ge c\delta^{-2} n \log n$ for some sufficiently large constant $c > 0$. Then we have*

$$\left\|\frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{a}_i^\top \boldsymbol{X}^\natural\right\|_2^2 \boldsymbol{a}_i \boldsymbol{a}_i^\top - \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 \boldsymbol{I}_n - 2\boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}\right\| \le \delta \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2,$$

with probability at least $1 - c_1 r n^{-13}$, where $c_1 > 0$ is some absolute constant.

Proof. This proof adapts the results of [25, Lemma 7.4] with refining the probabilities. Let $\boldsymbol{a}(1)$ be the first element of a vector $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Based on [122, Theorem 1.10], we have

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m (\boldsymbol{a}_i(1))^2 - 1\right| \ge \delta\right) \le e^2 \cdot e^{-\left(c_1\delta^2 m\right)^{1/2}};$$
$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m (\boldsymbol{a}_i(1))^4 - 3\right| \ge \delta\right) \le e^2 \cdot e^{-\left(c_2\delta^2 m\right)^{1/4}};$$

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i(1)\right)^6-15\right|\geq\delta\right)\leq e^2\cdot e^{-\left(c_3\delta^2 m\right)^{1/6}}.$$

So, by setting $m\gg\delta^{-2}n$, we have

$$\begin{aligned}\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i(1)\right)^2-1\right|&\leq\delta;\\ \left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i(1)\right)^4-3\right|&\leq\delta;\\ \left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i(1)\right)^6-15\right|&\leq\delta,\end{aligned}\tag{A.1}$$

with probability at least $1-c_4n^{-13}$ for some constant $c_4>0$. Moreover, following [121, Lemma 1], we know

$$\mathbb{P}\left(\left(\boldsymbol{a}_i(1)\right)^2\geq 1+2\sqrt{t}+2t\right)\leq\exp\left(-t\right),$$

which gives

$$\mathbb{P}\left(\left(\boldsymbol{a}_i(1)\right)^2\geq 36.5\log m\right)\leq\exp\left(-14\log m\right)=m^{-14},$$

if setting $t=14\log m$. Therefore, as long as $m\geq cn$, we have

$$\max_{1\leq i\leq m}\left|\boldsymbol{a}_i(1)\right|\leq\sqrt{36.5\log m},\tag{A.2}$$

with probability at least $1-c_5n^{-13}$ for some constant $c_5>0$.

With (A.1) and (A.2), the results in [25, Lemma 7.4] imply that for any $0<\delta\leq 1$, as soon as $m\geq c\delta^{-2}n\log n$ for some sufficiently large constant c, with probability at least $1-c_1n^{-13}$,

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)^2\boldsymbol{a}_i\boldsymbol{a}_i^\top-\left\|\boldsymbol{x}\right\|_2^2\boldsymbol{I}-2\boldsymbol{x}\boldsymbol{x}^\top\right\|\leq\delta\left\|\boldsymbol{x}\right\|_2^2$$

holds for any fixed vector $\boldsymbol{x}\in\mathbb{R}^n$. Let $\boldsymbol{X}^\natural=[\boldsymbol{x}_1^\natural,\boldsymbol{x}_2^\natural,\cdots,\boldsymbol{x}_r^\natural]$. Instantiating the above bound for the set of vectors $\boldsymbol{x}_k^\natural$, $k=1,\cdots,r$ and taking the union bound, we have

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\left\|\boldsymbol{a}_i^\top\boldsymbol{X}^\natural\right\|_2^2\boldsymbol{a}_i\boldsymbol{a}_i^\top-\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2\boldsymbol{I}-2\boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}\right\|$$

$$\leq \sum_{k=1}^{r} \left\| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{x}_k^\natural \right)^2 \boldsymbol{a}_i \boldsymbol{a}_i^\top - \left\| \boldsymbol{x}_k^\natural \right\|_2^2 \boldsymbol{I} - 2 \boldsymbol{x}_k^\natural \boldsymbol{x}_k^{\natural\top} \right\| \leq \delta \sum_{k=1}^{r} \left\| \boldsymbol{x}_k^\natural \right\|_2^2 = \delta \left\| \boldsymbol{X}^\natural \right\|_{\mathsf{F}}^2.$$

□

Lemma 20. *[97, Lemma 1] Suppose $F(\cdot)$ is cumulative distribution function (i.e. non-decreasing and right-continuous) with continuous density function $f(\cdot)$. Assume the samples $\{x_i\}_{i=1}^m$ are i.i.d. drawn from f. Let $0 < p < 1$. If $l < f(\theta) < L$ for all θ in $\{\theta : |\theta - \theta_p| \leq \epsilon\}$, then*

$$|\theta_p(\{x_i\}_{i=1}^m) - \theta_p(F)| < \epsilon$$

holds with probability at least $1 - 2\exp\left(-2m\epsilon^2 l^2\right)$.

Lemma 21. *[97, Lemma 2] Given a vector $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]$, where we order the entries in a non-decreasing manner $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$. Given another vector $\boldsymbol{y} = [y_1, y_2, \cdots, y_n]$, then*

$$|x_{(k)} - y_{(k)}| \leq \|\boldsymbol{x} - \boldsymbol{y}\|_\infty,$$

holds for all $k = 1, \cdots, n$.

Lemma 22. *[97, Lemma 3] Consider clean samples $\{\tilde{y}_i\}_{i=1}^m$. If a fraction s of them are corrupted by outliers, one obtains contaminated samples $\{y_i\}_{i=1}^m$, which contain sm corrupted samples and $(1-s)m$ clean samples. Then for a quantile p such that $s < p < 1 - s$, we have*

$$\theta_{p-s}\left(\{\tilde{y}_i\}_{i=1}^m\right) \leq \theta_p\left(\{y_i\}_{i=1}^m\right) \leq \theta_{p+s}\left(\{\tilde{y}_i\}_{i=1}^m\right).$$

Lemma 23. *[123, Lemma 4] For any matrix $\boldsymbol{Z}_i$ of the form $\boldsymbol{Z}_i = \begin{bmatrix} \boldsymbol{U}_i \boldsymbol{\Sigma}_i^{\frac{1}{2}} \boldsymbol{Q}_i \\ \boldsymbol{V}_i \boldsymbol{\Sigma}_i^{\frac{1}{2}} \boldsymbol{Q}_i \end{bmatrix}$, where $\boldsymbol{U}_i$, $\boldsymbol{V}_i$ and $\boldsymbol{Q}_i$ are unitary matrices and $\boldsymbol{\Sigma}_i \succeq 0$ is a diagonal matrix, for $i = 1, 2$, we have*

$$\left\| \boldsymbol{Z}_1 \boldsymbol{Z}_1^\top - \boldsymbol{Z}_2 \boldsymbol{Z}_2^\top \right\|_{\mathsf{F}} \leq 2 \left\| \boldsymbol{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^\top - \boldsymbol{U}_2 \boldsymbol{\Sigma}_2 \boldsymbol{V}_2^\top \right\|_{\mathsf{F}}.$$

Lemma 24 (Orlicz-norm version Bernstein's inequality)**.** *[124, Proposition 2] Let $\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_m$ be a finite sequence of independent zero-mean random matrices with dimensions $d_1 \times d_2$. Suppose $\|\boldsymbol{S}_i\|_{\psi_2} \leq B$, and define*

$$\sigma_{\boldsymbol{S}}^2 = \max\left\{\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\boldsymbol{S}_i\boldsymbol{S}_i^\top\right]\right\|, \left\|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\boldsymbol{S}_i^\top\boldsymbol{S}_i\right]\right\|\right\}.$$

Then there exists a constant $C > 0$ such that, for all $t > 0$, with probability at least $1 - e^{-t}$

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{S}_i\right\| \leq C\max\left\{\sigma_{\boldsymbol{S}}\sqrt{\frac{t+\log(d_1+d_2)}{m}}, B\sqrt{\log\left(\frac{B}{\sigma_{\boldsymbol{S}}}\right)}\frac{t+\log(d_1+d_2)}{m}\right\}.$$

Lemma 25. *Suppose $\boldsymbol{A}_i \in \mathbb{R}^{n_1\times n_2}$'s are sensing matrices, each generated with i.i.d. Gaussian entries, for $i = 1, \cdots, m$. Let $n = (n_1+n_2)/2$, and $m \geq n$. Then*

$$\max_{i=1,2,\cdots,m}\|\boldsymbol{A}_i\|_{\mathsf{F}} \leq 2\sqrt{n(n+m)} \tag{A.3}$$

holds with probability exceeding $1 - m\cdot\exp(-n(n+m))$.

Proof. Let $\boldsymbol{A}$ be a sensing matrix, generated with i.i.d. standard Gaussian entries, and $\boldsymbol{A}_{k,t}$ be the entry of $\boldsymbol{A}$ with index (k,t), then we know $\boldsymbol{A}_{k,t} \sim \mathcal{N}(0,1)$. Since $\|\boldsymbol{A}\|_{\mathsf{F}}^2 = \sum_{k,t}\boldsymbol{A}_{k,t}^2$, $\|\boldsymbol{A}\|_{\mathsf{F}}^2$ is a Chi-squared random variable with degree of freedom as n_1n_2. According to [121, Lemma 1], we have

$$\mathbb{P}\left\{\|\boldsymbol{A}\|_{\mathsf{F}}^2 \geq \left(1+2\sqrt{\lambda}+2\lambda\right)n_1n_2\right\} \leq \exp(-\lambda n_1n_2),$$

for any $\lambda > 0$. Let $\lambda = (n+m)/n$. It is clear that $\lambda \geq 2$ for $m \geq n$. Moreover, $2\lambda \geq 2\sqrt{\lambda}+1$ for $\lambda \geq 2$. Thus, we obtain

$$\mathbb{P}\left\{\|\boldsymbol{A}\|_{\mathsf{F}}^2 \geq 4n(n+m)\right\} \leq \exp(-n(n+m)).$$

Therefore, the proof is completed by applying the union bound. □

Lemma 26 (Restricted Isometry Property)**.** *[16] Fix* $0 < \delta < 1$*. For every* $1 \leq r \leq \min\{n_1, n_2\}$*, there exist positive constants* c_0 *and* c_1 *depending only on* δ *such that provided* $m \geq c_0(n_1 + n_2)r$*,*

$$(1-\delta)\left\|\boldsymbol{M}\right\|_{\mathsf{F}} \leq \frac{1}{\sqrt{m}}\left\|\mathcal{A}\left(\boldsymbol{M}\right)\right\|_2 \leq (1+\delta)\left\|\boldsymbol{M}\right\|_{\mathsf{F}}$$

holds for all matrices $\boldsymbol{M}$ *of rank at most* r *with probability at least* $1 - \exp\left(-c_1 m\right)$*.*

Appendix B: Technical Proofs in Chapter 2

B.1 Proof of Lemma 1

The crucial ingredient for proving the lower bound (2.23) is the following lemma, whose proof is provided in Appendix B.6.

Lemma 27. *Suppose* $m \geq c\frac{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{4}}{\sigma_r^4\left(\boldsymbol{X}^{\natural}\right)} nr\log\left(n\kappa\right)$ *with some large enough positive constant* c*, then with probability at least* $1-c_1 n^{-12}-me^{-1.5n}$*, we have*

$$\operatorname{vec}\left(\boldsymbol{V}\right)^{\top}\nabla^2 f(\boldsymbol{X})\operatorname{vec}\left(\boldsymbol{V}\right) \geq 2\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) + 1.204\sigma_r^2(\boldsymbol{X}^{\natural})\left\|\boldsymbol{V}\right\|_{\mathsf{F}}^2, \tag{B.1}$$

for all matrices $\boldsymbol{X}$ *and* $\boldsymbol{V}$ *where* $\boldsymbol{X}$ *satisfies* $\left\|\boldsymbol{X}-\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}} \leq \frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}}$*. Here,* $c_1>0$ *is some universal constant.*

With Lemma 27 in place, we are ready to prove (2.23). Let $\boldsymbol{V}=\boldsymbol{T}_1\boldsymbol{Q}_{\boldsymbol{T}}-\boldsymbol{T}_2$ satisfy the assumptions in Lemma 1, then we can demonstrate that

$$\begin{aligned}
&\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right)\\
&=\mathrm{Tr}\left(\left(\boldsymbol{X}^{\natural}-\boldsymbol{T}_2+\boldsymbol{T}_2\right)^{\top}\boldsymbol{V}\left(\boldsymbol{X}^{\natural}-\boldsymbol{T}_2+\boldsymbol{T}_2\right)^{\top}\boldsymbol{V}\right)\\
&=\mathrm{Tr}\left(\left(\boldsymbol{X}^{\natural}-\boldsymbol{T}_2\right)^{\top}\boldsymbol{V}\left(\boldsymbol{X}^{\natural}-\boldsymbol{T}_2\right)^{\top}\boldsymbol{V}\right)+2\mathrm{Tr}\left(\left(\boldsymbol{X}^{\natural}-\boldsymbol{T}_2\right)^{\top}\boldsymbol{V}\boldsymbol{T}_2^{\top}\boldsymbol{V}\right)+\mathrm{Tr}\left(\boldsymbol{T}_2^{\top}\boldsymbol{V}\boldsymbol{T}_2^{\top}\boldsymbol{V}\right)\\
&\geq \mathrm{Tr}\left(\boldsymbol{T}_2^{\top}\boldsymbol{V}\boldsymbol{T}_2^{\top}\boldsymbol{V}\right)-\left\|\boldsymbol{X}^{\natural}-\boldsymbol{T}_2\right\|^2\left\|\boldsymbol{V}\right\|_{\mathsf{F}}^2-2\left\|\boldsymbol{X}^{\natural}-\boldsymbol{T}_2\right\|\left\|\boldsymbol{T}_2\right\|\left\|\boldsymbol{V}\right\|_{\mathsf{F}}^2\\
&=\left\|\boldsymbol{T}_2^{\top}\boldsymbol{V}\right\|_{\mathsf{F}}^2-\left\|\boldsymbol{X}^{\natural}-\boldsymbol{T}_2\right\|^2\left\|\boldsymbol{V}\right\|_{\mathsf{F}}^2-2\left\|\boldsymbol{X}^{\natural}-\boldsymbol{T}_2\right\|\left\|\boldsymbol{T}_2\right\|\left\|\boldsymbol{V}\right\|_{\mathsf{F}}^2
\end{aligned} \tag{B.2}$$

$$\geq -\left[\left(\frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|}\right)^2 + 2\cdot\frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|}\cdot\left(\frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|} + \left\|\boldsymbol{X}^\natural\right\|\right)\right]\|\boldsymbol{V}\|_{\mathsf{F}}^2 \tag{B.3}$$

$$\geq -0.0886\sigma_r^2(\boldsymbol{X}^\natural)\,\|\boldsymbol{V}\|_{\mathsf{F}}^2\,, \tag{B.4}$$

where (B.2) follows from the fact that $\boldsymbol{T}_2^\top\boldsymbol{V} \in \mathbb{R}^{r\times r}$ is a symmetric matrix [125, Theorem 2], (B.3) arises from the fact $\left\|\boldsymbol{T}_2^\top\boldsymbol{V}\right\|_{\mathsf{F}}^2 \geq 0$ as well as the assumptions of Lemma 1, and (B.4) is based on the fact $\left\|\boldsymbol{X}^\natural\right\| \geq \sigma_r(\boldsymbol{X}^\natural)$. Combining (B.4) with Lemma 27, we establish the lower bound (2.23).

To prove the upper bound (2.24) asserted in the lemma, we make the observation that the Hessian in (2.22) satisfies

$$\begin{aligned}
&\left\|\nabla^2 f(\boldsymbol{X})\right\| \\
&= \left\|\frac{1}{m}\sum_{i=1}^m\left[\left(\|\boldsymbol{a}_i^\top\boldsymbol{X}\|_2^2 - \|\boldsymbol{a}_i^\top\boldsymbol{X}^\natural\|_2^2\right)\boldsymbol{I}_r + 2\boldsymbol{X}^\top\boldsymbol{a}_i\boldsymbol{a}_i^\top\boldsymbol{X}\right]\otimes\left(\boldsymbol{a}_i\boldsymbol{a}_i^\top\right)\right\| \\
&\leq \left\|\frac{1}{m}\sum_{i=1}^m\left[\left|\boldsymbol{a}_i^\top\left(\boldsymbol{X}+\boldsymbol{X}^\natural\right)\left(\boldsymbol{X}-\boldsymbol{X}^\natural\right)^\top\boldsymbol{a}_i\right|\boldsymbol{I}_r + 2\left\|\boldsymbol{a}_i^\top\boldsymbol{X}\right\|_2^2\boldsymbol{I}_r\right]\otimes\left(\boldsymbol{a}_i\boldsymbol{a}_i^\top\right)\right\| \\
&\leq \left\|\frac{1}{m}\sum_{i=1}^m\left[\left(\|\boldsymbol{a}_i^\top\boldsymbol{X}\|_2 + \|\boldsymbol{a}_i^\top\boldsymbol{X}^\natural\|_2\right)\cdot\left\|\boldsymbol{a}_i^\top\left(\boldsymbol{X}-\boldsymbol{X}^\natural\right)\right\|_2 + 2\left\|\boldsymbol{a}_i^\top\boldsymbol{X}\right\|_2^2\right]\boldsymbol{a}_i\boldsymbol{a}_i^\top\right\| \qquad \text{(B.5)} \\
&= \Bigg\|\frac{1}{m}\sum_{i=1}^m\left(\|\boldsymbol{a}_i^\top\boldsymbol{X}\|_2 + \|\boldsymbol{a}_i^\top\boldsymbol{X}^\natural\|_2\right)\cdot\left\|\boldsymbol{a}_i^\top\left(\boldsymbol{X}-\boldsymbol{X}^\natural\right)\right\|_2\cdot\left(\boldsymbol{a}_i\boldsymbol{a}_i^\top\right) \\
&\qquad + \frac{1}{m}\sum_{i=1}^m 2\left(\left\|\boldsymbol{a}_i^\top\boldsymbol{X}\right\|_2^2 - \left\|\boldsymbol{a}_i^\top\boldsymbol{X}^\natural\right\|_2^2\right)\cdot\left(\boldsymbol{a}_i\boldsymbol{a}_i^\top\right) + \frac{1}{m}\sum_{i=1}^m 2\left\|\boldsymbol{a}_i^\top\boldsymbol{X}^\natural\right\|_2^2\left(\boldsymbol{a}_i\boldsymbol{a}_i^\top\right) \\
&\qquad - 2\left(\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2\boldsymbol{I}_n + 2\boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}\right) + 2\left(\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2\boldsymbol{I}_n + 2\boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}\right)\Bigg\| \\
&\leq \underbrace{\left\|\frac{3}{m}\sum_{i=1}^m\left(\|\boldsymbol{a}_i^\top\boldsymbol{X}\|_2 + \|\boldsymbol{a}_i^\top\boldsymbol{X}^\natural\|_2\right)\cdot\left\|\boldsymbol{a}_i^\top\left(\boldsymbol{X}-\boldsymbol{X}^\natural\right)\right\|_2\left(\boldsymbol{a}_i\boldsymbol{a}_i^\top\right)\right\|}_{:=B_1} \\
&\quad + 2\underbrace{\left\|\frac{1}{m}\sum_{i=1}^m\left\|\boldsymbol{a}_i^\top\boldsymbol{X}^\natural\right\|_2^2\left(\boldsymbol{a}_i\boldsymbol{a}_i^\top\right) - \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2\boldsymbol{I}_n - 2\boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}\right\|}_{:=B_2}
\end{aligned}$$

$$+ \underbrace{2\left\|\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2}\boldsymbol{I}_n + 2\boldsymbol{X}^{\natural}\boldsymbol{X}^{\natural\top}\right\|}_{:=B_3}, \tag{B.6}$$

where (B.5) follows from the fact $\|\boldsymbol{I}\otimes\boldsymbol{A}\| = \|\boldsymbol{A}\|$. It is seen from Lemma 19 that

$$B_2 \le \delta\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2} \le 0.02\sigma_r^2\big(\boldsymbol{X}^{\natural}\big),$$

when setting $\delta \le 0.02\frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2}}$. Moreover, it is straightforward to check that

$$B_3 \le 6\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2}.$$

With regards to the first term B_1, note that by Lemma 17 and (2.25b), we can bound

$$\left\|\boldsymbol{a}_i^{\top}\boldsymbol{X}\right\|_2 \le \left\|\boldsymbol{a}_i^{\top}\boldsymbol{X}^{\natural}\right\|_2 + \left\|\boldsymbol{a}_i^{\top}(\boldsymbol{X}-\boldsymbol{X}^{\natural})\right\|_2 \le 5.86\sqrt{\log n}\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}} + \frac{1}{24}\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}}$$

for $1\le i\le m$, and therefore,

$$B_1 \le 1.471\sigma_r^2\big(\boldsymbol{X}^{\natural}\big)\log n\left\|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{a}_i\boldsymbol{a}_i^{\top}\right\| \le 1.48\sigma_r^2\big(\boldsymbol{X}^{\natural}\big)\log n, \tag{B.7}$$

where the last inequality follows from Lemma 15. The proof is then finished by combining (B.6) with the preceding bounds on B_1, B_2 and B_3.

B.2 Proof of Lemma 2

We first note that

$$\begin{aligned}
\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2} &\le \left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_t - \boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2} && \text{(B.8)}\\
&= \left\|\left(\boldsymbol{X}_t - \mu\nabla f\left(\boldsymbol{X}_t\right)\right)\boldsymbol{Q}_t - \boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2} \\
&= \left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \mu\nabla f\left(\boldsymbol{X}_t\boldsymbol{Q}_t\right) - \boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2} && \text{(B.9)}\\
&= \left\|\boldsymbol{x}_t - \boldsymbol{x}^{\natural} - \mu\cdot\mathrm{vec}\left(\nabla f\left(\boldsymbol{X}_t\boldsymbol{Q}_t\right) - \nabla f\left(\boldsymbol{X}^{\natural}\right)\right)\right\|_2^2, && \text{(B.10)}
\end{aligned}$$

where we write

$$\boldsymbol{x}_t := \mathrm{vec}\left(\boldsymbol{X}_t\boldsymbol{Q}_t\right) \quad \text{and} \quad \boldsymbol{x}^{\natural} := \mathrm{vec}\big(\boldsymbol{X}^{\natural}\big).$$

Here, (B.8) follows from the definition of $\boldsymbol{Q}_{t+1}$ (see (2.13)), (B.9) holds owing to the identity $\nabla f\left(\boldsymbol{X}_t\right)\boldsymbol{Q}_t = \nabla f\left(\boldsymbol{X}_t\boldsymbol{Q}_t\right)$ for $\boldsymbol{Q}_t \in \mathcal{O}^{r\times r}$, and (B.10) arises from the fact that $\nabla f\left(\boldsymbol{X}^\natural\right) = \boldsymbol{0}$. Let

$$\boldsymbol{X}_t(\tau) = \boldsymbol{X}^\natural + \tau\big(\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\big),$$

where $\tau \in [0,1]$. Then, by the fundamental theorem of calculus for vector-valued functions [126],

$$\begin{aligned}
\text{RHS of (B.10)} &= \left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural - \mu\cdot\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right)\left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)d\tau\right\|_2^2 \qquad \text{(B.11)}\\
&= \left\|\left(\boldsymbol{I} - \mu\cdot\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right)\mathrm{d}\tau\right)\left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)\right\|_2^2\\
&= \left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)^\top\left(\boldsymbol{I} - \mu\cdot\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right)\mathrm{d}\tau\right)^2\left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)\\
&= \left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural\right\|_2^2 - 2\mu\cdot\left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)^\top\left(\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right)\mathrm{d}\tau\right)\left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)\\
&\qquad + \mu^2\cdot\left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)^\top\left(\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right)\mathrm{d}\tau\right)^2\left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)\\
&\le \left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural\right\|_2^2 - 2\mu\cdot\left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)^\top\left(\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right)\mathrm{d}\tau\right)\left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)\\
&\qquad + \mu^2\cdot\left\|\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right)\mathrm{d}\tau\right\|^2\left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural\right\|_2^2. \qquad \text{(B.12)}
\end{aligned}$$

It is easy to verify that $\boldsymbol{X}_t(\tau)$ satisfies (2.25) for any $\tau \in [0,1]$, since

$$\left\|\boldsymbol{X}_t(\tau) - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} = \tau\left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \le \frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}},$$

and

$$\max_{1\le l\le m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t(\tau) - \boldsymbol{X}^\natural\right)\right\|_2 = \tau\cdot\max_{1\le l\le m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right)\right\|_2 \le \frac{1}{24}\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}.$$

Lemma 1 then implies that

$$\left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)^\top\left(\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right)\mathrm{d}\tau\right)\left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right) \ge 1.026\sigma_r^2\left(\boldsymbol{X}^\natural\right)\left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural\right\|_2^2,$$

and

$$\left\|\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right) \mathrm{d}\tau\right\| \leq 1.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\log n + 6\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2.$$

Substituting the above two inequalities into (B.10) and (B.12) gives

$$\begin{aligned}
&\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 \\
&\leq \left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural\right\|_2^2 - 2\mu\cdot 1.026\sigma_r^2\left(\boldsymbol{X}^\natural\right)\left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural\right\|_2^2 \\
&\quad + \mu^2\cdot\left(1.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\log n + 6\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2\right)^2\left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural\right\|_2^2 \\
&= \left[1 - 2.052\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu + \left(1.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\log n + 6\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2\right)^2\mu^2\right]\left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 \\
&\leq \left(1 - 1.026\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)\left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2,
\end{aligned}$$

with the proviso that $\mu \leq \frac{1.026\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left(1.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\log n + 6\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2\right)^2}$. This allows us to conclude that

$$\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \leq \left(1 - 0.513\sigma_r^2(\boldsymbol{X}^\natural)\mu\right)\left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right\|_{\mathsf{F}}.$$

B.3 Proof of Lemma 3

Recognizing that

$$\begin{aligned}
\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right\|_{\mathsf{F}} &\leq \left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_t^{(l)}\boldsymbol{Q}_t^\top\boldsymbol{Q}_{t+1}\right\|_{\mathsf{F}} \\
&= \left\|\boldsymbol{X}_{t+1} - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_t^{(l)}\boldsymbol{Q}_t^\top\right\|_{\mathsf{F}} = \left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_t - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_t^{(l)}\right\|_{\mathsf{F}},
\end{aligned}$$

we will focus on bounding $\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_t - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_t^{(l)}\right\|_{\mathsf{F}}$. Since

$$\begin{aligned}
&\boldsymbol{X}_{t+1}\boldsymbol{Q}_t - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_t^{(l)} \\
&= \left(\boldsymbol{X}_t - \mu\nabla f\left(\boldsymbol{X}_t\right)\right)\boldsymbol{Q}_t - \left(\boldsymbol{X}_t^{(l)} - \mu\nabla f^{(l)}\left(\boldsymbol{X}_t^{(l)}\right)\right)\boldsymbol{R}_t^{(l)} \\
&= \boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \mu\nabla f\left(\boldsymbol{X}_t\right)\boldsymbol{Q}_t + \mu\nabla f^{(l)}\left(\boldsymbol{X}_t^{(l)}\right)\boldsymbol{R}_t^{(l)} \\
&= \boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \mu\frac{1}{m}\sum_{i=1}^m\left(\left\|\boldsymbol{a}_i^\top\boldsymbol{X}_t\right\|_2^2 - y_i\right)\boldsymbol{a}_i\boldsymbol{a}_i^\top\boldsymbol{X}_t\boldsymbol{Q}_t
\end{aligned}$$

$$
\begin{aligned}
&+\mu\frac{1}{m}\sum_{i=1}^{m}\left(\left\|\boldsymbol{a}_i^\top\boldsymbol{X}_t^{(l)}\right\|_2^2-y_i\right)\boldsymbol{a}_i\boldsymbol{a}_i^\top\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}-\mu\frac{1}{m}\left(\left\|\boldsymbol{a}_l^\top\boldsymbol{X}_t^{(l)}\right\|_2^2-y_l\right)\boldsymbol{a}_l\boldsymbol{a}_l^\top\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}\\
&=\underbrace{\boldsymbol{X}_t\boldsymbol{Q}_t-\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}-\mu\nabla f\left(\boldsymbol{X}_t\boldsymbol{Q}_t\right)+\mu\nabla f\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}\right)}_{:=\boldsymbol{S}_{t,1}^{(l)}}\\
&\quad\underbrace{-\mu\frac{1}{m}\left(\left\|\boldsymbol{a}_l^\top\boldsymbol{X}_t^{(l)}\right\|_2^2-y_l\right)\boldsymbol{a}_l\boldsymbol{a}_l^\top\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}}_{:=\boldsymbol{S}_{t,2}^{(l)}},
\end{aligned}
$$

we aim to control $\left\|\boldsymbol{S}_{t,1}^{(l)}\right\|_{\mathsf{F}}$ and $\left\|\boldsymbol{S}_{t,2}^{(l)}\right\|_{\mathsf{F}}$ separately.

We first bound the term $\left\|\boldsymbol{S}_{t,2}^{(l)}\right\|_{\mathsf{F}}$, which is easier to handle. Observe that by Cauchy-Schwarz inequality,

$$
\begin{aligned}
\left|\left\|\boldsymbol{a}_l^\top\boldsymbol{X}_t^{(l)}\right\|_2^2-y_l\right| &= \left|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}-\boldsymbol{X}^\natural\right)\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}+\boldsymbol{X}^\natural\right)^\top\boldsymbol{a}_l\right|\\
&\le \left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}-\boldsymbol{X}^\natural\right)\right\|_2\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}+\boldsymbol{X}^\natural\right)\right\|_2. \qquad \text{(B.13)}
\end{aligned}
$$

The first term in (B.13) can be bounded by

$$
\begin{aligned}
&\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}-\boldsymbol{X}^\natural\right)\right\|_2\\
&\le \left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}-\boldsymbol{X}_t\boldsymbol{Q}_t\right)\right\|_2+\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t\boldsymbol{Q}_t-\boldsymbol{X}^\natural\right)\right\|_2\\
&\le \sqrt{6n}\left\|\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}-\boldsymbol{X}_t\boldsymbol{Q}_t\right\|+C_2\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\\
&\le \sqrt{6n}C_3\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\\
&\quad+C_2\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\\
&= \left(\sqrt{6}C_3+C_2\right)\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}, \qquad \text{(B.14)}
\end{aligned}
$$

where we have used the triangle inequality, Lemma 16, as well as the induction hypotheses (2.33c) and (2.33b). Similarly, the second term in (B.13) can be bounded as

$$
\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}+\boldsymbol{X}^\natural\right)\right\|_2\le\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}-\boldsymbol{X}^\natural\right)\right\|_2+2\left\|\boldsymbol{a}_l^\top\boldsymbol{X}^\natural\right\|_2
$$

$$\leq \left(\sqrt{6}C_3 + C_2\right)\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} + 11.72\sqrt{\log n}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}$$
$$\leq \left(\sqrt{6}C_3 + C_2 + 11.72\right)\sqrt{\log n}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}, \tag{B.15}$$

where we have used (B.14), Lemma 17, and $\sigma_r^2\left(\boldsymbol{X}^\natural\right) \leq \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2$. Similarly, we can also obtain

$$\left\|\boldsymbol{a}_l^\top \boldsymbol{X}_t^{(l)}\right\|_2 \leq \left(\sqrt{6}C_3 + C_2 + 5.86\right)\sqrt{\log n}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}.$$

Substituting (B.14) and (B.15) into (B.13), and using the above inequality, we get

$$\begin{aligned}
\left\|\boldsymbol{S}_{t,2}^{(l)}\right\|_{\mathsf{F}} &= \mu\frac{1}{m}\cdot\left|\left\|\boldsymbol{a}_l^\top \boldsymbol{X}_t^{(l)}\right\|_2^2 - y_l\right|\cdot\left\|\boldsymbol{a}_l\boldsymbol{a}_l^\top \boldsymbol{X}_t^{(l)}\right\|_{\mathsf{F}} \\
&\leq C_4^2\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\cdot\mu\frac{1}{m}\cdot\sigma_r^2\left(\boldsymbol{X}^\natural\right)\log n\cdot\left\|\boldsymbol{a}_l\right\|_2\left\|\boldsymbol{a}_l^\top \boldsymbol{X}_t^{(l)}\right\|_2 \\
&\leq \sqrt{6}C_4^3\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\cdot\mu\frac{1}{m}\cdot\sigma_r^2\left(\boldsymbol{X}^\natural\right)\log n\cdot\sqrt{n}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}\sqrt{\log n} \\
&= \sqrt{6}C_4^3\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\cdot\mu\frac{\sqrt{n}\cdot(\log n)^{3/2}}{m}\sigma_r^2\left(\boldsymbol{X}^\natural\right)\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}},
\end{aligned} \tag{B.16}$$

where $C_4 := \sqrt{6}C_3 + C_2 + 11.72$.

Next, we turn to $\left\|\boldsymbol{S}_{t,1}^{(l)}\right\|_{\mathsf{F}}$. By defining

$$\boldsymbol{s}_{t,1}^{(l)} = \mathrm{vec}\big(\boldsymbol{S}_{t,1}^{(l)}\big), \quad \boldsymbol{x}_t = \mathrm{vec}\left(\boldsymbol{X}_t\boldsymbol{Q}_t\right), \quad \text{and} \quad \boldsymbol{x}_t^{(l)} = \mathrm{vec}\big(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}\big),$$

we can write

$$\begin{aligned}
\boldsymbol{s}_{t,1}^{(l)} &= \boldsymbol{x}_t - \boldsymbol{x}_t^{(l)} - \mu\cdot\mathrm{vec}\left(\nabla f\left(\boldsymbol{X}_t\boldsymbol{Q}_t\right) - \nabla f(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)})\right) \\
&= \boldsymbol{x}_t - \boldsymbol{x}_t^{(l)} - \mu\cdot\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t^{(l)}(\tau)\right)\left(\boldsymbol{x}_t - \boldsymbol{x}_t^{(l)}\right)\mathrm{d}\tau \\
&= \left(\boldsymbol{I} - \mu\cdot\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t^{(l)}(\tau)\right)\mathrm{d}\tau\right)\left(\boldsymbol{x}_t - \boldsymbol{x}_t^{(l)}\right).
\end{aligned}$$

Here, the second line follows from the fundamental theorem of calculus for vector-valued functions [126], where

$$\boldsymbol{X}_t^{(l)}(\tau) = \boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} + \tau\left(\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}\right), \tag{B.17}$$

for $\tau \in [0,1]$. Using very similar algebra as in Appendix B.2, we obtain

$$\begin{aligned}\left\|\boldsymbol{S}_{t,1}^{(l)}\right\|_{\mathsf{F}}^2 \le \left\|\boldsymbol{x}_t - \boldsymbol{x}_t^{(l)}\right\|_2^2 + \mu^2 \left\|\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t^{(l)}(\tau)\right) \mathrm{d}\tau\right\|^2 \left\|\boldsymbol{x}_t - \boldsymbol{x}_t^{(l)}\right\|_2^2 \\ - 2\mu \cdot \left(\boldsymbol{x}_t - \boldsymbol{x}_t^{(l)}\right)^\top \left(\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t^{(l)}(\tau)\right) \mathrm{d}\tau\right) \left(\boldsymbol{x}_t - \boldsymbol{x}_t^{(l)}\right). \qquad \text{(B.18)}\end{aligned}$$

It is easy to verify that for all $\tau \in [0,1]$,

$$\begin{aligned}\left\|\boldsymbol{X}_t^{(l)}(\tau) - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} &= \left\|(1-\tau)\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \boldsymbol{X}_t\boldsymbol{Q}_t\right) + \boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \\ &\le (1-\tau)\left\|\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \boldsymbol{X}_t\boldsymbol{Q}_t\right\|_{\mathsf{F}} + \left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \\ &\le C_3\sqrt{\frac{\log n}{n}} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} + C_1\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} \qquad \text{(B.19)} \\ &= \left(C_3\sqrt{\frac{\log n}{n}} + C_1\right)\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} \le \frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}, \qquad \text{(B.20)}\end{aligned}$$

where (B.19) follows from the induction hypotheses (2.33a) and (2.33b), and (B.20) follows as long as $C_1 + C_3 \le \frac{1}{24}$. Further, for all $1 \le l \le m$, by the induction hypothesis (2.33b) and (2.33c),

$$\begin{aligned}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}(\tau) - \boldsymbol{X}^\natural\right)\right\|_2 &\le (1-\tau)\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \boldsymbol{X}_t\boldsymbol{Q}_t\right)\right\|_2 + \left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right)\right\|_2 \\ &\le \|\boldsymbol{a}_l\|_2\left\|\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \boldsymbol{X}_t\boldsymbol{Q}_t\right\| + C_2\sqrt{\log n} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} \\ &\le \sqrt{6n}C_3\sqrt{\frac{\log n}{n}} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} + C_2\sqrt{\log n} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} \\ &\le \left(\sqrt{6}C_3 + C_2\right)\sqrt{\log n} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} \le \frac{1}{24}\sqrt{\log n} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}},\end{aligned}$$

as long as $\sqrt{6}C_3 + C_2 \le \frac{1}{24}$. Therefore, Lemma 1 holds for $\boldsymbol{X}_t^{(l)}(\tau)$, and similar to Appendix B.2, (B.18) can be further bounded by

$$\left\|\boldsymbol{S}_{t,1}^{(l)}\right\|_{\mathsf{F}} \le \left(1 - 0.513\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)\left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}\right\|_{\mathsf{F}} \qquad \text{(B.21)}$$

as long as $\mu \leq \frac{1.026\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left(1.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\log n+6\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2\right)^2}$. Consequently, combining (B.16) and (B.21), we can get

$$\begin{aligned}\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right\|_{\mathsf{F}} &\leq \left\|\boldsymbol{S}_{t,1}^{(l)}\right\|_{\mathsf{F}}+\left\|\boldsymbol{S}_{t,2}^{(l)}\right\|_{\mathsf{F}} \\ &\leq \left(1-0.513\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)\left\|\boldsymbol{X}_t\boldsymbol{Q}_t-\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}\right\|_{\mathsf{F}} \\ &\quad+\sqrt{6}C_4^3\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\cdot\mu\frac{\sqrt{n}\cdot(\log n)^{3/2}}{m}\sigma_r^2\left(\boldsymbol{X}^\natural\right)\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}} \\ &\leq C_3\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^{t+1}\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}, \qquad \text{(B.22)}\end{aligned}$$

where (B.22) follows from the induction hypothesis (2.33b), as long as $m \geq c\frac{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2}{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}n\log n$ for some large enough constant $c>0$.

B.4 Proof of Lemma 4

For any $1 \leq l \leq m$, by the statistical independence of $\boldsymbol{a}_l$ and $\boldsymbol{X}_{t+1}^{(l)}$ and by Lemma 17, we have

$$\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}-\boldsymbol{X}^\natural\right)\right\|_2 \leq 5.86\sqrt{\log n}\left\|\boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}-\boldsymbol{X}^\natural\right\|_{\mathsf{F}}.$$

Further, by the triangle inequality, Lemma 16, Lemma 3 and Lemma 2, we can deduce that

$$\begin{aligned}&\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{X}^\natural\right)\right\|_2 \\ &\leq \left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right)\right\|_2+\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}-\boldsymbol{X}^\natural\right)\right\|_2 \\ &\leq \|\boldsymbol{a}_l\|_2\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right\|+5.86\sqrt{\log n}\left\|\boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}-\boldsymbol{X}^\natural\right\|_{\mathsf{F}} \\ &\leq \sqrt{6n}\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right\|+5.86\sqrt{\log n}\cdot\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right\|_{\mathsf{F}} \\ &\quad+5.86\sqrt{\log n}\cdot\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{X}^\natural\right\|_{\mathsf{F}} \\ &\leq \left(\sqrt{6n}+5.86\sqrt{\log n}\right)C_3\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^{t+1}\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\end{aligned}$$

$$
\begin{aligned}
&+ 5.86\sqrt{\log n}C_1\left(1-0.513\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^{t+1}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\\
&\le\left(\sqrt{6}C_3+5.86C_1+5.86C_3\sqrt{\frac{\log n}{n}}\right)\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^{t+1}\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\\
&\le C_2\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^{t+1}\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}},
\end{aligned}
$$

where the last line follows as long as $\sqrt{6}C_3+5.86C_1+5.86C_3\le C_2$. The proof is then finished by applying the union bound for all $1\le l\le m$.

B.5 Proof of Lemma 5

Define

$$
\begin{aligned}
\boldsymbol{\Sigma}_0&=\operatorname{diag}\left\{\lambda_1\left(\boldsymbol{Y}\right),\lambda_2\left(\boldsymbol{Y}\right),\cdots,\lambda_r\left(\boldsymbol{Y}\right)\right\}=\boldsymbol{\Lambda}_0+\lambda\boldsymbol{I};\\
\boldsymbol{\Sigma}_0^{(l)}&=\operatorname{diag}\left\{\lambda_1\left(\boldsymbol{Y}^{(l)}\right),\lambda_2\left(\boldsymbol{Y}^{(l)}\right),\cdots,\lambda_r\left(\boldsymbol{Y}^{(l)}\right)\right\}=\boldsymbol{\Lambda}_0^{(l)}+\lambda^{(l)}\boldsymbol{I},\quad 1\le l\le m,
\end{aligned}
$$

then by definition we have $\boldsymbol{Y}\boldsymbol{Z}_0=\boldsymbol{Z}_0\boldsymbol{\Sigma}_0$, $\boldsymbol{Y}^{(l)}\boldsymbol{Z}_0^{(l)}=\boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}$, and

$$
\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_0^{(l)}-\boldsymbol{Z}_0^\top\boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}=\frac{1}{2m}y_l\boldsymbol{Z}_0^\top\boldsymbol{a}_l\boldsymbol{a}_l^\top\boldsymbol{Z}_0^{(l)}. \tag{B.23}
$$

Moreover, let $\boldsymbol{Z}_{0,c}$ and $\boldsymbol{Z}_{0,c}^{(l)}$ be the complement matrices of $\boldsymbol{Z}_0$ and $\boldsymbol{Z}_0^{(l)}$, respectively, such that both $[\boldsymbol{Z}_0,\boldsymbol{Z}_{0,c}]$ and $\left[\boldsymbol{Z}_0^{(l)},\boldsymbol{Z}_{0,c}^{(l)}\right]$ are orthonormal matrices. Below we will prove the induction hypotheses (2.33) in the base case when $t=0$ one by one.

B.5.1 Proof of (2.33a)

From Lemma 12, we have

$$
\begin{aligned}
\left\|\boldsymbol{X}_0\boldsymbol{Q}_0-\boldsymbol{X}^\natural\right\|_{\mathsf{F}}&\le\frac{1}{\sqrt{2\left(\sqrt{2}-1\right)}\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\boldsymbol{X}_0\boldsymbol{X}_0^\top-\boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}\right\|_{\mathsf{F}}\\
&=\frac{1}{\sqrt{2\left(\sqrt{2}-1\right)}\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\boldsymbol{Z}_0\boldsymbol{\Lambda}_0\boldsymbol{Z}_0^\top-\boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}\right\|_{\mathsf{F}}
\end{aligned}
$$

$$\leq \frac{\sqrt{r}}{\sqrt{2\left(\sqrt{2}-1\right)}\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top-\boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}-\lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^\top\right\|. \quad \text{(B.24)}$$

The last term in (B.24) can be further bounded as

$$\begin{aligned}
&\left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top-\boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}-\lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^\top\right\| \\
&\leq \left\|\boldsymbol{Y}-\frac{1}{2}\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2\boldsymbol{I}-\boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}\right\|+\left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top-\boldsymbol{Y}+\frac{1}{2}\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2\boldsymbol{Z}_{0,c}\boldsymbol{Z}_{0,c}^\top\right\| \\
&\quad+\left\|\frac{1}{2}\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2\boldsymbol{Z}_0\boldsymbol{Z}_0^\top-\lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^\top\right\| \\
&\leq \delta\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2+\delta\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2+\delta\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2=3\delta\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2, \qquad \text{(B.25)}
\end{aligned}$$

where (B.25) follows from

$$\|\boldsymbol{Y}-\mathbb{E}[\boldsymbol{Y}]\|=\left\|\boldsymbol{Y}-\frac{1}{2}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2\boldsymbol{I}-\boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}\right\|\leq\delta\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2$$

via Lemma 19, the Weyl's inequality, and

$$|\lambda-\mathbb{E}[\lambda]|=\left|\lambda-\frac{1}{2}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2\right|\leq\delta\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2$$

via Lemma 15. Plugging (B.25) into (B.24), we have

$$\left\|\boldsymbol{X}_0\boldsymbol{Q}_0-\boldsymbol{X}^\natural\right\|_{\mathsf{F}}\leq\frac{3}{\sqrt{2\left(\sqrt{2}-1\right)}}\cdot\frac{\delta\sqrt{r}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2}{\sigma_r\left(\boldsymbol{X}^\natural\right)}.$$

Setting $\delta=c\frac{\sigma_r^3\left(\boldsymbol{X}^\natural\right)}{\sqrt{r}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^3}$ for a sufficiently small constant c, i.e. $m\gtrsim\frac{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^6}{\sigma_r^6\left(\boldsymbol{X}^\natural\right)}nr\log n$, we get $\left\|\boldsymbol{X}_0\boldsymbol{Q}_0-\boldsymbol{X}^\natural\right\|_{\mathsf{F}}\leq C_1\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}$. Following similar procedures, we can also show $\left\|\boldsymbol{X}_0^{(l)}\boldsymbol{Q}_0^{(l)}-\boldsymbol{X}^\natural\right\|_{\mathsf{F}}\leq C_1\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}$.

B.5.2 Proof of (2.33b)

Following Weyl's inequality, by (2.33a), we have

$$\left|\sigma_i\left(\boldsymbol{X}_0\right)-\sigma_i\left(\boldsymbol{X}^\natural\right)\right|\leq C_1\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}},$$

and similarly,

$$\left|\sigma_i\big(\boldsymbol{X}_0^{(l)}\big) - \sigma_i\left(\boldsymbol{X}^\natural\right)\right| \le C_1 \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}},$$

for $i = 1, \cdots, r$. Combined with Lemma 12, there exists some constant c such that

$$\begin{aligned}
&\left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}\right\|_{\mathsf{F}} \\
&\le \frac{1}{\sqrt{2\left(\sqrt{2}-1\right)}\sigma_r\left(\boldsymbol{X}_0\right)}\left\|\boldsymbol{X}_0\boldsymbol{X}_0^\top - \boldsymbol{X}_0^{(l)}\boldsymbol{X}_0^{(l)\top}\right\|_{\mathsf{F}} \\
&\le \frac{c}{\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\boldsymbol{X}_0\boldsymbol{X}_0^\top - \boldsymbol{X}_0^{(l)}\boldsymbol{X}_0^{(l)\top}\right\|_{\mathsf{F}} \\
&= \frac{c}{\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\boldsymbol{Z}_0\boldsymbol{\Lambda}_0\boldsymbol{Z}_0^\top - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Lambda}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}} \\
&= \frac{c}{\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}\boldsymbol{Z}_0^{(l)\top} - \lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^\top + \lambda^{(l)}\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}} \\
&\le \frac{c}{\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}} \\
&\quad + \frac{c}{\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^\top - \lambda^{(l)}\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}}. \qquad \text{(B.26)}
\end{aligned}$$

We will bound each term in (B.26), respectively. For the first term, we have

$$\begin{aligned}
&\left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}} \\
&= \left\|\left[\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_0^{(l)} - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}, \boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_{0,c}^{(l)}\right]\right\|_{\mathsf{F}} \\
&\le \left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_0^{(l)} - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}\right\|_{\mathsf{F}} + \left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_{0,c}^{(l)}\right\|_{\mathsf{F}} \\
&\le \left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_0^{(l)} - \boldsymbol{Z}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}\right\|_{\mathsf{F}} + \left\|\boldsymbol{Z}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)} - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}\right\|_{\mathsf{F}} + \|\boldsymbol{Y}\|\left\|\boldsymbol{Z}_0^\top\boldsymbol{Z}_{0,c}^{(l)}\right\|_{\mathsf{F}} \\
&\le \left\|\boldsymbol{Z}_0 \cdot \frac{1}{2m}y_l\boldsymbol{Z}_0^\top\boldsymbol{a}_l\boldsymbol{a}_l^\top\boldsymbol{Z}_0^{(l)}\right\|_{\mathsf{F}} + \left\|\boldsymbol{Z}_0\boldsymbol{Z}_0^\top - \boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}}\|\boldsymbol{Y}^{(l)}\| \\
&\quad + \|\boldsymbol{Y}\|\left\|\boldsymbol{Z}_0^\top\boldsymbol{Z}_{0,c}^{(l)}\right\|_{\mathsf{F}}, \qquad \text{(B.27)}
\end{aligned}$$

where the last inequality follows from (B.23). Note that the first term in (B.27) can be bounded as

$$\left\|\boldsymbol{Z}_0 \cdot \frac{1}{2m}y_l\boldsymbol{Z}_0^\top\boldsymbol{a}_l\boldsymbol{a}_l^\top\boldsymbol{Z}_0^{(l)}\right\|_{\mathsf{F}} \le \frac{1}{2m}\left\|\boldsymbol{a}_l^\top\boldsymbol{X}^\natural\right\|_2^2\left\|\boldsymbol{a}_l^\top\boldsymbol{Z}_0^{(l)}\right\|_2\left\|\boldsymbol{a}_l^\top\boldsymbol{Z}_0\right\|_2$$

$$\lesssim \frac{\sqrt{n}\cdot(\log n)^{3/2}\cdot\sqrt{r}}{m}\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2}, \tag{B.28}$$

which follows from Lemma 16 and Lemma 17. The second term in (B.27) can be bounded as

$$\begin{aligned}\left\|\boldsymbol{Z}_0\boldsymbol{Z}_0^{\top}-\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}} &= \left\|\boldsymbol{Z}_0\left(\boldsymbol{Z}_0-\boldsymbol{Z}_0^{(l)}\boldsymbol{T}_0^{(l)}\right)^{\top}+\left(\boldsymbol{Z}_0-\boldsymbol{Z}_0^{(l)}\boldsymbol{T}_0^{(l)}\right)\left(\boldsymbol{Z}_0^{(l)}\boldsymbol{T}_0^{(l)}\right)^{\top}\right\|_{\mathsf{F}}\\ &\le 2\left\|\boldsymbol{Z}_0-\boldsymbol{Z}_0^{(l)}\boldsymbol{T}_0^{(l)}\right\|_{\mathsf{F}} \le 2\sqrt{2}\left\|\boldsymbol{Z}_0^{\top}\boldsymbol{Z}_{0,c}^{(l)}\right\|_{\mathsf{F}},\end{aligned}$$

where $\boldsymbol{T}_t^{(l)}=\operatorname{argmin}_{\boldsymbol{P}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{Z}_t-\boldsymbol{Z}_t^{(l)}\boldsymbol{P}\right\|_{\mathsf{F}}$, and the last line follows from the fact $\left\|\boldsymbol{Z}_0-\boldsymbol{Z}_0^{(l)}\boldsymbol{T}_0^{(l)}\right\|_{\mathsf{F}}\le\sqrt{2}\left\|\boldsymbol{Z}_0^{\top}\boldsymbol{Z}_{0,c}^{(l)}\right\|_{\mathsf{F}}$ [127]. Putting this together with the third term in (B.27), we have

$$\begin{aligned}&\left\|\boldsymbol{Z}_0\boldsymbol{Z}_0^{\top}-\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}}\left\|\boldsymbol{Y}^{(l)}\right\|+\|\boldsymbol{Y}\|\left\|\boldsymbol{Z}_0^{\top}\boldsymbol{Z}_{0,c}^{(l)}\right\|_{\mathsf{F}}\\ &\le\left(2\sqrt{2}\left\|\boldsymbol{Y}^{(l)}\right\|+\|\boldsymbol{Y}\|\right)\left\|\boldsymbol{Z}_0^{\top}\boldsymbol{Z}_{0,c}^{(l)}\right\|_{\mathsf{F}}\\ &\lesssim\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2}\frac{\left\|\left(\frac{1}{m}y_l\boldsymbol{a}_l\boldsymbol{a}_l^{\top}\right)\boldsymbol{Z}_0^{(l)}\right\|_{\mathsf{F}}}{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)} \qquad\text{(B.29)}\\ &\lesssim\frac{\left\|\boldsymbol{a}_l^{\top}\boldsymbol{X}^{\natural}\right\|_2^2\left\|\boldsymbol{a}_l^{\top}\boldsymbol{Z}_0^{(l)}\right\|_2\|\boldsymbol{a}_l\|_2}{m}\frac{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^2}{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}\\ &\lesssim\frac{\sqrt{n}\cdot(\log n)^{3/2}\cdot\sqrt{r}}{m}\frac{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^4}{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}, \qquad\text{(B.30)}\end{aligned}$$

where (B.29) follows from Lemma 19 and the Davis-Kahan $\sin\Theta$ theorem [128], and (B.30) follows from Lemma 16 and Lemma 17.

For the second term in (B.26), we have

$$\begin{aligned}&\left\|\lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^{\top}-\lambda^{(l)}\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}}\\ &=\left\|\lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^{\top}-\lambda\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}+\lambda\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}-\lambda^{(l)}\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}}\\ &\le\lambda\cdot\left\|\boldsymbol{Z}_0\boldsymbol{Z}_0^{\top}-\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}}+\left|\lambda-\lambda^{(l)}\right|\cdot\left\|\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}}\end{aligned}$$

$$\lesssim \frac{\sqrt{n}\cdot(\log n)^{3/2}\cdot\sqrt{r}}{m}\frac{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^4}{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}+\frac{y_l}{2m}\sqrt{r} \tag{B.31}$$

$$\lesssim \frac{\sqrt{n}\cdot(\log n)^{3/2}\cdot\sqrt{r}}{m}\frac{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^4}{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}+\frac{\sqrt{r}\cdot\log n}{m}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2, \tag{B.32}$$

where the first term of (B.31) is bounded similarly as (B.30), and (B.32) follows from Lemma 17. Combining (B.28), (B.30), and (B.32), we obtain

$$\left\|\boldsymbol{X}_0\boldsymbol{Q}_0-\boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}\right\|_{\mathsf{F}}\lesssim\frac{\sqrt{n}\cdot(\log n)^{3/2}\cdot\sqrt{r}}{m}\frac{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^4}{\sigma_r^3\left(\boldsymbol{X}^\natural\right)}\lesssim\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}},$$

where the last inequality holds as long as $m\gtrsim\frac{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^5}{\sigma_r^5\left(\boldsymbol{X}^\natural\right)}n\sqrt{r}\log n=nr^3\log n$.

B.5.3 Proof of (2.33c)

For every $1\le l\le m$, from (2.33a) and (2.33b), we have

$$\left\|\boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}-\boldsymbol{X}^\natural\right\|_{\mathsf{F}}\le\left\|\boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}-\boldsymbol{X}_0\boldsymbol{Q}_0\right\|_{\mathsf{F}}+\left\|\boldsymbol{X}_0\boldsymbol{Q}_0-\boldsymbol{X}^\natural\right\|_{\mathsf{F}}\lesssim\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}.$$

This further gives

$$\begin{aligned}
&\max_{1\le l\le m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_0\boldsymbol{Q}_0-\boldsymbol{X}^\natural\right)\right\|_2\\
&\le\max_{1\le l\le m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_0\boldsymbol{Q}_0-\boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}\right)\right\|_2+\max_{1\le l\le m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}-\boldsymbol{X}^\natural\right)\right\|_2\\
&\le\max_{1\le l\le m}\left\|\boldsymbol{a}_l\right\|_2\left\|\boldsymbol{X}_0\boldsymbol{Q}_0-\boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}\right\|+\max_{1\le l\le m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}-\boldsymbol{X}^\natural\right)\right\|_2\\
&\lesssim\sqrt{n}\cdot\max_{1\le l\le m}\left\|\boldsymbol{X}_0\boldsymbol{Q}_0-\boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}\right\|+\sqrt{\log n}\cdot\max_{1\le l\le m}\left\|\boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}-\boldsymbol{X}^\natural\right\|_2 \qquad \text{(B.33)}\\
&\lesssim\sqrt{n}\cdot\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}+\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} \qquad \text{(B.34)}\\
&\lesssim\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}},
\end{aligned}$$

where (B.33) follows from Lemma 16 and Lemma 17, and (B.34) follows from (2.33b).

B.5.4 Finishing the Proof

The proof of Lemma 5 is now complete by appropriately adjusting the constants.

B.6 Proof of Lemma 27

Without loss of generality, we assume $\|\boldsymbol{V}\|_{\mathsf{F}} = 1$. Write

$$\begin{aligned}
&\operatorname{vec}(\boldsymbol{V})^\top \nabla^2 f(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{V}) \\
&= \frac{1}{m}\sum_{i=1}^m \operatorname{vec}(\boldsymbol{V})^\top \left[\left[\left(\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2 - y_i\right)\boldsymbol{I}_r + 2\boldsymbol{X}^\top \boldsymbol{a}_i \boldsymbol{a}_i^\top \boldsymbol{X}\right] \otimes \left(\boldsymbol{a}_i \boldsymbol{a}_i^\top\right)\right] \operatorname{vec}(\boldsymbol{V}) \\
&= \frac{1}{m}\sum_{i=1}^m \left(\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2 - y_i\right) \operatorname{vec}(\boldsymbol{V})^\top \operatorname{vec}\left(\boldsymbol{a}_i \boldsymbol{a}_i^\top \boldsymbol{V}\right) \\
&\quad + \frac{1}{m}\sum_{i=1}^m \operatorname{vec}(\boldsymbol{V})^\top \operatorname{vec}\left(2\boldsymbol{a}_i \boldsymbol{a}_i^\top \boldsymbol{V}\boldsymbol{X}^\top \boldsymbol{a}_i \boldsymbol{a}_i^\top \boldsymbol{X}\right) \\
&= \frac{1}{m}\sum_{i=1}^m \left[\left(\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2 - \left\|\boldsymbol{a}_i^\top \boldsymbol{X}^\natural\right\|_2^2\right)\left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^2 + 2\left(\boldsymbol{a}_i^\top \boldsymbol{X}\boldsymbol{V}^\top \boldsymbol{a}_i\right)^2\right]. \qquad \text{(B.35)}
\end{aligned}$$

In what follows, we let $\boldsymbol{X} = \boldsymbol{X}^\natural + t\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}\boldsymbol{H}$ with $t \le 1/24$ and $\|\boldsymbol{H}\|_{\mathsf{F}} = 1$ which immediately obeys $\left\|\boldsymbol{X} - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \le \frac{1}{24}\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}$, and express the right-hand side of (B.35) as

$$\begin{aligned}
p(\boldsymbol{V}, \boldsymbol{H}, t) := &\underbrace{\frac{1}{m}\sum_{i=1}^m \left[\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2 \left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^2 + 2\left(\boldsymbol{a}_i^\top \boldsymbol{X}\boldsymbol{V}^\top \boldsymbol{a}_i\right)^2\right]}_{:= q(\boldsymbol{V}, \boldsymbol{H}, t)} \\
&- \frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{a}_i^\top \boldsymbol{X}^\natural\right\|_2^2 \left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^2. \qquad \text{(B.36)}
\end{aligned}$$

The aim is thus to control $p(\boldsymbol{V}, \boldsymbol{H}, t)$ for all matrices satisfying $\|\boldsymbol{H}\|_{\mathsf{F}} = 1$ and $\|\boldsymbol{V}\|_{\mathsf{F}} = 1$, and for all t obeying $t \le 1/24$.

We first bound the second term in (B.36). Let $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r]$, then by Lemma 19,

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{a}_i^\top \boldsymbol{X}^\natural\right\|_2^2 \left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^2 - \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 \left\|\boldsymbol{V}\right\|_{\mathsf{F}}^2 - 2\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right\|_{\mathsf{F}}^2\right| \\
&= \left|\frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{a}_i^\top \boldsymbol{X}^\natural\right\|_2^2 \sum_{k=1}^r \left(\boldsymbol{a}_i^\top \boldsymbol{v}_k\right)^2 - \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 \sum_{k=1}^r \|\boldsymbol{v}_k\|_2^2 - 2\sum_{k=1}^r \left\|\boldsymbol{X}^{\natural\top}\boldsymbol{v}_k\right\|_2^2\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{r} \left| \frac{1}{m} \sum_{i=1}^{m} \left\| \boldsymbol{a}_i^\top \boldsymbol{X}^\natural \right\|_2^2 \left(\boldsymbol{a}_i^\top \boldsymbol{v}_k \right)^2 - \left\| \boldsymbol{X}^\natural \right\|_{\mathsf{F}}^2 \left\| \boldsymbol{v}_k \right\|_2^2 - 2 \left\| \boldsymbol{X}^{\natural\top} \boldsymbol{v}_k \right\|_2^2 \right| \\
&= \sum_{k=1}^{r} \left| \boldsymbol{v}_k^\top \left(\frac{1}{m} \sum_{i=1}^{m} \left\| \boldsymbol{a}_i^\top \boldsymbol{X}^\natural \right\|_2^2 \boldsymbol{a}_i \boldsymbol{a}_i^\top - \left\| \boldsymbol{X}^\natural \right\|_{\mathsf{F}}^2 - 2 \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top} \right) \boldsymbol{v}_k \right| \\
&\leq \sum_{k=1}^{r} \left\| \boldsymbol{v}_k \right\|_2^2 \left\| \frac{1}{m} \sum_{i=1}^{m} \left\| \boldsymbol{a}_i^\top \boldsymbol{X}^\natural \right\|_2^2 \boldsymbol{a}_i \boldsymbol{a}_i^\top - \left\| \boldsymbol{X}^\natural \right\|_{\mathsf{F}}^2 - 2 \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top} \right\| \\
&\leq \delta \left\| \boldsymbol{X}^\natural \right\|_{\mathsf{F}}^2 \sum_{k=1}^{r} \left\| \boldsymbol{v}_k \right\|_2^2 = \delta \left\| \boldsymbol{X}^\natural \right\|_{\mathsf{F}}^2 \left\| \boldsymbol{V} \right\|_{\mathsf{F}}^2 .
\end{aligned}
$$

By setting $\delta \leq \frac{1}{24} \frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2}$, we see that with probability at least $1 - c_1 r n^{-13}$,

$$
\frac{1}{m} \sum_{i=1}^{m} \left\| \boldsymbol{a}_i^\top \boldsymbol{X}^\natural \right\|_2^2 \left\| \boldsymbol{a}_i^\top \boldsymbol{V} \right\|_2^2 \leq \left\| \boldsymbol{X}^\natural \right\|_{\mathsf{F}}^2 \left\| \boldsymbol{V} \right\|_{\mathsf{F}}^2 + 2 \left\| \boldsymbol{X}^{\natural\top} \boldsymbol{V} \right\|_{\mathsf{F}}^2 + \frac{1}{24} \sigma_r^2(\boldsymbol{X}^\natural) \left\| \boldsymbol{V} \right\|_{\mathsf{F}}^2 , \quad \text{(B.37)}
$$

holds simultaneously for all matrices $\boldsymbol{V}$, as long as $m \gtrsim \frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^4}{\sigma_r^4(\boldsymbol{X}^\natural)} n \log n$.

Next, we turn to the first term $q(\boldsymbol{V}, \boldsymbol{H}, t)$ in (B.36), and we need to accommodate all matrices satisfying $\|\boldsymbol{H}\|_{\mathsf{F}} = 1$ and $\|\boldsymbol{V}\|_{\mathsf{F}} = 1$, and all scalars obeying $t \leq 1/24$. The strategy is that we first establish the bound of $q(\boldsymbol{V}, \boldsymbol{H}, t)$ for any fixed $\boldsymbol{H}$, $\boldsymbol{V}$ and t, and then extend the result to a uniform bound for all $\boldsymbol{H}$, $\boldsymbol{V}$ and t by covering arguments.

B.6.1 Bound with Fixed Matrices and Scalar

Recall that

$$
q(\boldsymbol{V}, \boldsymbol{H}, t) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{\left[\left\| \boldsymbol{a}_i^\top \boldsymbol{X} \right\|_2^2 \left\| \boldsymbol{a}_i^\top \boldsymbol{V} \right\|_2^2 + 2 \left(\boldsymbol{a}_i^\top \boldsymbol{X} \boldsymbol{V}^\top \boldsymbol{a}_i \right)^2 \right]}_{:= G_i} .
$$

We will start by assuming that $\boldsymbol{X}$ and $\boldsymbol{V}$ are both fixed and statistically independent of $\{\boldsymbol{a}_i\}_{i=1}^m$. In view of Lemma 18,

$$
\mathbb{E}[G_i] = \mathbb{E} \left[\left\| \boldsymbol{a}_i^\top \boldsymbol{X} \right\|_2^2 \left\| \boldsymbol{a}_i^\top \boldsymbol{V} \right\|_2^2 \right] + 2 \mathbb{E} \left[\left(\boldsymbol{a}_i^\top \boldsymbol{X} \boldsymbol{V}^\top \boldsymbol{a}_i \right)^2 \right]
$$

$$
\begin{aligned}
&= \|\boldsymbol{X}\|_{\mathsf{F}}^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\left\|\boldsymbol{X}^\top \boldsymbol{V}\right\|_{\mathsf{F}}^2 + 2\left(\mathrm{Tr}\left(\boldsymbol{X}^\top \boldsymbol{V}\right)\right)^2 + 2\left\|\boldsymbol{X}\boldsymbol{V}^\top\right\|_{\mathsf{F}}^2 + 2\mathrm{Tr}\left(\boldsymbol{X}^\top \boldsymbol{V}\boldsymbol{X}^\top \boldsymbol{V}\right) \\
&\leq \|\boldsymbol{X}\|_{\mathsf{F}}^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\|\boldsymbol{X}\|^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\|\boldsymbol{X}\|_{\mathsf{F}}^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\|\boldsymbol{X}\|^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\|\boldsymbol{X}\|^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 \\
&\leq 9\|\boldsymbol{X}\|_{\mathsf{F}}^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 = 9\left\|\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}\right\|_{\mathsf{F}}^2 \qquad \text{(B.38)} \\
&\leq 18\left(\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 + t^2\frac{\sigma_r^4\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2}\|\boldsymbol{H}\|_{\mathsf{F}}^2\right) \leq 18.002\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2, \qquad \text{(B.39)}
\end{aligned}
$$

where (B.38) follows $\|\boldsymbol{V}\|_{\mathsf{F}} = 1$ and $\boldsymbol{X} = \boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}$, and (B.39) arises from the calculations with $\|\boldsymbol{H}\|_{\mathsf{F}} = 1$ and $t \leq 1/24$. Therefore, if we define $T_i = \mathbb{E}\left[G_i\right] - G_i$, we have $\mathbb{E}\left[T_i\right] = 0$ and

$$
T_i \leq \mathbb{E}\left[G_i\right] \leq 18.002\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2,
$$

due to $G_i \geq 0$. In addition,

$$
\begin{aligned}
&\mathbb{E}\left[T_i^2\right] = \mathbb{E}\left[G_i^2\right] - \left(\mathbb{E}\left[G_i\right]\right)^2 \leq \mathbb{E}\left[G_i^2\right] \\
&= \mathbb{E}\left[\left(\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2\left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^2 + 2\left(\boldsymbol{a}_i^\top \boldsymbol{X}\boldsymbol{V}^\top \boldsymbol{a}_i\right)^2\right)^2\right] \\
&= \mathbb{E}\left[\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^4\left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^4\right] + 4\mathbb{E}\left[\left(\boldsymbol{a}_i^\top \boldsymbol{X}\boldsymbol{V}^\top \boldsymbol{a}_i\right)^4\right] + 4\mathbb{E}\left[\left(\boldsymbol{a}_i^\top \boldsymbol{X}\boldsymbol{V}^\top \boldsymbol{a}_i\right)^2\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2\left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^2\right] \\
&\leq 9\mathbb{E}\left[\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^4\left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^4\right] \qquad \text{(B.40)} \\
&\leq 9\sqrt{\mathbb{E}\left[\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^8\right]\mathbb{E}\left[\left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^8\right]} \qquad \text{(B.41)} \\
&\leq 9c_4\|\boldsymbol{X}\|_{\mathsf{F}}^4\|\boldsymbol{V}\|_{\mathsf{F}}^4 = 9c_4\left\|\boldsymbol{X}\right\|_{\mathsf{F}}^4 \qquad \text{(B.42)} \\
&= 9c_4\left\|\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}\right\|_{\mathsf{F}}^4 \lesssim \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^4,
\end{aligned}
$$

where (B.40) follows from the Cauchy-Schwarz inequality, (B.41) comes from the Hölder's inequality, and (B.42) is a consequence of Lemma 18. Apply Lemma 14 to arrive at

$$
\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m T_i \geq \frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^\natural\right)\right) \leq \exp\left(-c\frac{m\sigma_r^4\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^4}\right), \qquad \text{(B.43)}
$$

which further leads to

$$
\begin{aligned}
& q\left(\boldsymbol{V}, \boldsymbol{H}, t\right) \\
& = \frac{1}{m}\sum_{i=1}^{m} G_i = \mathbb{E}\left[G_i\right] - \frac{1}{m}\sum_{i=1}^{m} T_i \\
& \geq \mathbb{E}\left[G_i\right] - \frac{1}{24}\sigma_r^2\big(\boldsymbol{X}^\natural\big) \\
& = \|\boldsymbol{X}\|_{\mathsf{F}}^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\big\|\boldsymbol{X}^\top\boldsymbol{V}\big\|_{\mathsf{F}}^2 + 2\left(\mathrm{Tr}\big(\boldsymbol{X}^\top\boldsymbol{V}\big)\right)^2 + 2\big\|\boldsymbol{X}\boldsymbol{V}^\top\big\|_{\mathsf{F}}^2 + 2\mathrm{Tr}\left(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X}^\top\boldsymbol{V}\right) \\
& \quad - \frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^\natural\right) \\
& \geq \|\boldsymbol{X}\|_{\mathsf{F}}^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\big\|\boldsymbol{X}^\top\boldsymbol{V}\big\|_{\mathsf{F}}^2 + 2\big\|\boldsymbol{X}\boldsymbol{V}^\top\big\|_{\mathsf{F}}^2 + 2\mathrm{Tr}\left(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X}^\top\boldsymbol{V}\right) \\
& \quad - \frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^\natural\right). \qquad \text{(B.44)}
\end{aligned}
$$

Substituting $\boldsymbol{X} = \boldsymbol{X}^\natural + t\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}\boldsymbol{H}$ for $\boldsymbol{X}$, and using the facts $\|\boldsymbol{H}\|_{\mathsf{F}} = 1$, $\|\boldsymbol{V}\|_{\mathsf{F}} = 1$ and $t \leq 1/24$, we can calculate the following bounds:

$$
\begin{aligned}
\|\boldsymbol{X}\|_{\mathsf{F}}^2 &= \big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}^2 + t^2\frac{\sigma_r^4\left(\boldsymbol{X}^\natural\right)}{\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}^2}\|\boldsymbol{H}\|_{\mathsf{F}}^2 + 2t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}}\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{H}\right) \\
&\geq \big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}^2 - 2t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}}\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}\|\boldsymbol{H}\|_{\mathsf{F}} \\
&\geq \big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}^2 - \frac{1}{12}\sigma_r^2\left(\boldsymbol{X}^\natural\right); \\
\big\|\boldsymbol{X}^\top\boldsymbol{V}\big\|_{\mathsf{F}}^2 &= \big\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\big\|_{\mathsf{F}}^2 + t^2\frac{\sigma_r^4\left(\boldsymbol{X}^\natural\right)}{\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}^2}\big\|\boldsymbol{H}^\top\boldsymbol{V}\big\|_{\mathsf{F}}^2 + 2t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}}\mathrm{Tr}\left(\boldsymbol{V}^\top\boldsymbol{H}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) \\
&\geq \big\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\big\|_{\mathsf{F}}^2 - 2t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}}\big\|\boldsymbol{X}^\natural\big\|\,\|\boldsymbol{H}\|\,\|\boldsymbol{V}\|_{\mathsf{F}}^2 \\
&\geq \big\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\big\|_{\mathsf{F}}^2 - \frac{1}{12}\sigma_r^2\left(\boldsymbol{X}^\natural\right); \\
\big\|\boldsymbol{X}\boldsymbol{V}^\top\big\|_{\mathsf{F}}^2 &= \big\|\boldsymbol{X}^\natural\boldsymbol{V}^\top\big\|_{\mathsf{F}}^2 + t^2\frac{\sigma_r^4\left(\boldsymbol{X}^\natural\right)}{\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}^2}\big\|\boldsymbol{H}\boldsymbol{V}^\top\big\|_{\mathsf{F}}^2 + 2t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}}\mathrm{Tr}\left(\boldsymbol{V}\boldsymbol{H}^\top\boldsymbol{X}^\natural\boldsymbol{V}^\top\right) \\
&\geq \big\|\boldsymbol{X}^\natural\boldsymbol{V}^\top\big\|_{\mathsf{F}}^2 - 2t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}}\big\|\boldsymbol{X}^\natural\big\|\,\|\boldsymbol{H}\|\,\|\boldsymbol{V}\|_{\mathsf{F}}^2 \\
&\geq \big\|\boldsymbol{X}^\natural\boldsymbol{V}^\top\big\|_{\mathsf{F}}^2 - \frac{1}{12}\sigma_r^2\left(\boldsymbol{X}^\natural\right);
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Tr}\left(\boldsymbol{X}^{\top}\boldsymbol{V}\boldsymbol{X}^{\top}\boldsymbol{V}\right) &= \operatorname{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) + 2t\frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}}\operatorname{Tr}\left(\boldsymbol{H}^{\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) \\
&\quad + t^2\frac{\sigma_r^4\left(\boldsymbol{X}^{\natural}\right)}{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^2}\operatorname{Tr}\left(\boldsymbol{H}^{\top}\boldsymbol{V}\boldsymbol{H}^{\top}\boldsymbol{V}\right) \\
&\geq \operatorname{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) - 2t\frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}}\left\|\boldsymbol{X}^{\natural}\right\|\|\boldsymbol{H}\|\|\boldsymbol{V}\|_{\mathsf{F}}^2 \\
&\quad - t^2\frac{\sigma_r^4\left(\boldsymbol{X}^{\natural}\right)}{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^2}\|\boldsymbol{H}\|^2\|\boldsymbol{V}\|_{\mathsf{F}}^2 \\
&\geq \operatorname{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) - \left(\frac{1}{12}+\frac{1}{24^2}\right)\sigma_r^2\left(\boldsymbol{X}^{\natural}\right),
\end{aligned}
$$

which, combining with (B.44), yields

$$
\begin{aligned}
&q\left(\boldsymbol{V},\boldsymbol{H},t\right) \\
&\geq \left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^2 + 2\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right\|_{\mathsf{F}}^2 + 2\left\|\boldsymbol{X}^{\natural}\boldsymbol{V}^{\top}\right\|_{\mathsf{F}}^2 + 2\operatorname{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) \\
&\quad - \left(\frac{15}{24}+\frac{1}{12\cdot 24}\right)\sigma_r^2\left(\boldsymbol{X}^{\natural}\right) \\
&\geq \left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^2 + 2\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right\|_{\mathsf{F}}^2 + 2\operatorname{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) + 2\sigma_r^2\left(\boldsymbol{X}^{\natural}\right) - \left(\frac{15}{24}+\frac{1}{12\cdot 24}\right)\sigma_r^2\left(\boldsymbol{X}^{\natural}\right) \\
&\geq \left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^2 + 2\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right\|_{\mathsf{F}}^2 + 2\operatorname{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) + 1.371\sigma_r^2\left(\boldsymbol{X}^{\natural}\right).
\end{aligned}
$$

B.6.2 Covering Arguments

Since we have obtained a lower bound on $q\left(\boldsymbol{V},\boldsymbol{H},t\right)$ for fixed $\boldsymbol{V}$, $\boldsymbol{H}$ and t, we now move on to extending it to a uniform bound that covers all $\boldsymbol{V}$, $\boldsymbol{H}$ and t simultaneously. Towards this, we will invoke the ϵ-net covering arguments for all $\boldsymbol{V}$, $\boldsymbol{H}$ and t, respectively, and will rely on the fact $\max_{1\leq i\leq m}\|\boldsymbol{a}_i\|_2 \leq \sqrt{6n}$ asserted in Lemma 16. For notational convenience, we define

$$
\begin{aligned}
g\left(\boldsymbol{V},\boldsymbol{H},t\right) &:= q\left(\boldsymbol{V},\boldsymbol{H},t\right) \\
&\quad - \left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^2 - 2\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right\|_{\mathsf{F}}^2 - 2\operatorname{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) - 1.371\sigma_r^2\left(\boldsymbol{X}^{\natural}\right).
\end{aligned}
$$

First, consider the ϵ-net covering argument for $\boldsymbol{V}$. Suppose $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are such that $\|\boldsymbol{V}_1\|_{\mathsf{F}} = 1$, $\|\boldsymbol{V}_2\|_{\mathsf{F}} = 1$, and $\|\boldsymbol{V}_1 - \boldsymbol{V}_2\|_{\mathsf{F}} \leq \epsilon$. Then, since

$$\left| \left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}_1\right\|_{\mathsf{F}}^2 - \left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}_2\right\|_{\mathsf{F}}^2 \right| \leq \left(\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}_1\right\|_{\mathsf{F}} + \left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}_2\right\|_{\mathsf{F}}\right) \left\|\boldsymbol{X}^{\natural\top}\left(\boldsymbol{V}_1 - \boldsymbol{V}_2\right)\right\|_{\mathsf{F}} \leq 2\left\|\boldsymbol{X}^{\natural}\right\|^2 \epsilon,$$

and

$$\begin{aligned}
&\left|\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}_1\boldsymbol{X}^{\natural\top}\boldsymbol{V}_1\right) - \mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}_2\boldsymbol{X}^{\natural\top}\boldsymbol{V}_2\right)\right| \\
&\leq \left|\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}_1\boldsymbol{X}^{\natural\top}\boldsymbol{V}_1\right) - \mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}_1\boldsymbol{X}^{\natural\top}\boldsymbol{V}_2\right)\right| \\
&\quad + \left|\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}_1\boldsymbol{X}^{\natural\top}\boldsymbol{V}_2\right) - \mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}_2\boldsymbol{X}^{\natural\top}\boldsymbol{V}_2\right)\right| \\
&\leq \left\|\boldsymbol{X}^{\natural}\right\|^2 \|\boldsymbol{V}_1\|_{\mathsf{F}} \|\boldsymbol{V}_1 - \boldsymbol{V}_2\|_{\mathsf{F}} + \left\|\boldsymbol{X}^{\natural}\right\|^2 \|\boldsymbol{V}_2\|_{\mathsf{F}} \|\boldsymbol{V}_1 - \boldsymbol{V}_2\|_{\mathsf{F}} \leq 2\left\|\boldsymbol{X}^{\natural}\right\|^2 \epsilon,
\end{aligned}$$

we have

$$\begin{aligned}
&\left|g\left(\boldsymbol{V}_1, \boldsymbol{H}, t\right) - g\left(\boldsymbol{V}_2, \boldsymbol{H}, t\right)\right| \\
&\leq \left|q\left(\boldsymbol{V}_1, \boldsymbol{H}, t\right) - q\left(\boldsymbol{V}_2, \boldsymbol{H}, t\right)\right| + 2\left| \left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}_1\right\|_{\mathsf{F}}^2 - \left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}_2\right\|_{\mathsf{F}}^2 \right| \\
&\quad + 2\left|\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}_1\boldsymbol{X}^{\natural\top}\boldsymbol{V}_1\right) - \mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}_2\boldsymbol{X}^{\natural\top}\boldsymbol{V}_2\right)\right| \\
&\leq \left| \frac{1}{m}\sum_{i=1}^m \left[\left\|\boldsymbol{a}_i^\top\boldsymbol{X}\right\|_2^2 \left\|\boldsymbol{a}_i^\top\boldsymbol{V}_1\right\|_2^2 + 2\left(\boldsymbol{a}_i^\top\boldsymbol{X}\boldsymbol{V}_1^\top\boldsymbol{a}_i\right)^2\right] \right. \\
&\qquad \left. - \frac{1}{m}\sum_{i=1}^m \left[\left\|\boldsymbol{a}_i^\top\boldsymbol{X}\right\|_2^2 \left\|\boldsymbol{a}_i^\top\boldsymbol{V}_2\right\|_2^2 + 2\left(\boldsymbol{a}_i^\top\boldsymbol{X}\boldsymbol{V}_2^\top\boldsymbol{a}_i\right)^2\right] \right| + 8\left\|\boldsymbol{X}^{\natural}\right\|^2\epsilon \\
&\leq \frac{1}{m}\sum_{i=1}^m \left| \left\|\boldsymbol{a}_i^\top\boldsymbol{X}\right\|_2^2 \left\|\boldsymbol{a}_i^\top\boldsymbol{V}_1\right\|_2^2 - \left\|\boldsymbol{a}_i^\top\boldsymbol{X}\right\|_2^2 \left\|\boldsymbol{a}_i^\top\boldsymbol{V}_2\right\|_2^2 \right| \\
&\quad + \frac{2}{m}\sum_{i=1}^m \left| \left(\boldsymbol{a}_i^\top\boldsymbol{X}\boldsymbol{V}_1^\top\boldsymbol{a}_i\right)^2 - \left(\boldsymbol{a}_i^\top\boldsymbol{X}\boldsymbol{V}_2^\top\boldsymbol{a}_i\right)^2 \right| + 8\left\|\boldsymbol{X}^{\natural}\right\|^2\epsilon \\
&\leq \frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{a}_i^\top\boldsymbol{X}\right\|_2^2 \cdot \left(\left\|\boldsymbol{a}_i^\top\boldsymbol{V}_1\right\|_2 + \left\|\boldsymbol{a}_i^\top\boldsymbol{V}_2\right\|_2\right) \cdot \left\|\boldsymbol{a}_i^\top\left(\boldsymbol{V}_1 - \boldsymbol{V}_2\right)\right\|_2 \\
&\quad + \frac{2}{m}\sum_{i=1}^m \left|\boldsymbol{a}_i^\top\boldsymbol{X}\left(\boldsymbol{V}_1 + \boldsymbol{V}_2\right)^\top\boldsymbol{a}_i\right| \cdot \left|\boldsymbol{a}_i^\top\boldsymbol{X}\left(\boldsymbol{V}_1 - \boldsymbol{V}_2\right)^\top\boldsymbol{a}_i\right| + 8\left\|\boldsymbol{X}^{\natural}\right\|^2\epsilon \\
&\leq 6n \cdot \|\boldsymbol{X}\|^2 \cdot 2\sqrt{6n} \cdot \sqrt{6n} \cdot \epsilon + 2 \cdot 12n \cdot \|\boldsymbol{X}\| \cdot 6n \cdot \|\boldsymbol{X}\|\,\epsilon + 8\left\|\boldsymbol{X}^{\natural}\right\|^2\epsilon
\end{aligned}$$

$$
\begin{aligned}
&= 216\epsilon n^2 \left\| \boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H} \right\|^2 + 8\left\|\boldsymbol{X}^\natural\right\|^2\epsilon \\
&\le 432\epsilon n^2 \left(\left\|\boldsymbol{X}^\natural\right\|^2 + t^2\frac{\sigma_r^4\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2}\left\|\boldsymbol{H}\right\|^2 \right) + 8\left\|\boldsymbol{X}^\natural\right\|^2\epsilon \\
&\le \left(432.75n^2+8\right)\epsilon\left\|\boldsymbol{X}^\natural\right\|^2 \le \frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^\natural\right),
\end{aligned}
$$

as long as $\epsilon = \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{10584n^2\left\|\boldsymbol{X}^\natural\right\|^2}$. Based on Lemma 13, the cardinality of this ϵ-net will be

$$
\left(\frac{9}{\epsilon}\right)^{(n+r+1)r} = \left(\frac{9\cdot 10584n^2\left\|\boldsymbol{X}^\natural\right\|^2}{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}\right)^{(n+r+1)r} \le \exp\left(cnr\log\left(n\kappa\right)\right).
$$

Secondly, consider the ϵ-net covering argument for $\boldsymbol{H}$. Suppose $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ obey $\left\|\boldsymbol{H}_1\right\|_{\mathsf{F}} = 1$, $\left\|\boldsymbol{H}_2\right\|_{\mathsf{F}} = 1$, and $\left\|\boldsymbol{H}_1 - \boldsymbol{H}_2\right\|_{\mathsf{F}} \le \epsilon$. Then one has

$$
\begin{aligned}
&\left|g\left(\boldsymbol{V},\boldsymbol{H}_1,t\right) - g\left(\boldsymbol{V},\boldsymbol{H}_2,t\right)\right| \\
&= \left|q\left(\boldsymbol{V},\boldsymbol{H}_1,t\right) - q\left(\boldsymbol{V},\boldsymbol{H}_2,t\right)\right| \\
&= \Bigg| \frac{1}{m}\sum_{i=1}^m \left\| \boldsymbol{a}_i^\top \left(\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}_1 \right) \right\|_2^2 \left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2 \\
&\quad + \frac{1}{m}\sum_{i=1}^m 2\left(\boldsymbol{a}_i^\top \left(\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}_1 \right)\boldsymbol{V}^\top\boldsymbol{a}_i \right)^2 \\
&\quad - \frac{1}{m}\sum_{i=1}^m \left\| \boldsymbol{a}_i^\top \left(\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}_2 \right) \right\|_2^2 \left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2 \\
&\quad - \frac{1}{m}\sum_{i=1}^m 2\left(\boldsymbol{a}_i^\top \left(\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}_2 \right)\boldsymbol{V}^\top\boldsymbol{a}_i \right)^2 \Bigg| \\
&\le \frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2 \cdot \left| \left\| \boldsymbol{a}_i^\top \left(\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}_1 \right) \right\|_2^2 - \left\| \boldsymbol{a}_i^\top \left(\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}_2 \right) \right\|_2^2 \right| \\
&\quad + \frac{2}{m}\sum_{i=1}^m \left| \left(\boldsymbol{a}_i^\top \left(\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}_1 \right)\boldsymbol{V}^\top\boldsymbol{a}_i \right)^2 - \left(\boldsymbol{a}_i^\top \left(\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}_2 \right)\boldsymbol{V}^\top\boldsymbol{a}_i \right)^2 \right| \\
&\le 6n\cdot\sqrt{6n}\cdot t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\epsilon\cdot 2\sqrt{6n}\cdot\frac{25}{24}\left\|\boldsymbol{X}^\natural\right\| + 2\cdot 6n\cdot t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\epsilon\cdot 12n\cdot\frac{25}{24}\left\|\boldsymbol{X}^\natural\right\|
\end{aligned}
$$

$$\leq \frac{75}{8}\epsilon n^2 \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\left\|\boldsymbol{X}^\natural\right\| \leq \frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^\natural\right),$$

as long as $\epsilon = \frac{1}{225n^2}\cdot\frac{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}{\left\|\boldsymbol{X}^\natural\right\|}$. Based on Lemma 13, the cardinality of this ϵ-net will be

$$\left(\frac{9}{\epsilon}\right)^{(n+r+1)r} = \left(9\cdot 225n^2\cdot\frac{\left\|\boldsymbol{X}^\natural\right\|}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\right)^{(n+r+1)r} \leq \exp\left(cnr\log n\right).$$

Finally, consider the ϵ-net covering argument for all t, such that $t \leq 1/24$. Suppose t_1 and t_2 satisfy $t_1 \leq 1/24$, $t_2 \leq 1/24$ and $|t_1 - t_2| \leq \epsilon$. Then we get

$$\begin{aligned}
&\left|g\left(\boldsymbol{V},\boldsymbol{H},t_1\right) - g\left(\boldsymbol{V},\boldsymbol{H},t_2\right)\right| \\
&= \left|q\left(\boldsymbol{V},\boldsymbol{H},t_1\right) - q\left(\boldsymbol{V},\boldsymbol{H},t_2\right)\right| \\
&= \left|\frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t_1\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}\right)\right\|_2^2 \left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2\right. \\
&\quad + \frac{1}{m}\sum_{i=1}^m 2\left(\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t_1\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}\right)\boldsymbol{V}^\top\boldsymbol{a}_i\right)^2 \\
&\quad - \frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t_2\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}\right)\right\|_2^2 \left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2 \\
&\quad \left. - \frac{1}{m}\sum_{i=1}^m 2\left(\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t_2\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}\right)\boldsymbol{V}^\top\boldsymbol{a}_i\right)^2\right| \\
&\leq \frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2 \cdot \left|\left\|\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t_1\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}\right)\right\|_2^2 - \left\|\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t_2\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}\right)\right\|_2^2\right| \\
&\quad + \frac{2}{m}\sum_{i=1}^m \left|\left(\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t_1\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}\right)\boldsymbol{V}^\top\boldsymbol{a}_i\right)^2 - \left(\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t_2\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\boldsymbol{H}\right)\boldsymbol{V}^\top\boldsymbol{a}_i\right)^2\right| \\
&\leq 6n\cdot\sqrt{6n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\epsilon\cdot 2\sqrt{6n}\cdot\frac{25}{24}\left\|\boldsymbol{X}^\natural\right\| + 2\cdot 6n\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\epsilon\cdot 12n\cdot\frac{25}{24}\left\|\boldsymbol{X}^\natural\right\| \\
&\leq 225\epsilon n^2\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\left\|\boldsymbol{X}^\natural\right\| \leq \frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^\natural\right),
\end{aligned}$$

as long as $\epsilon = \frac{1}{5400n^2}\cdot\frac{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}{\left\|\boldsymbol{X}^\natural\right\|}$. The cardinality of this ϵ-net will be $\frac{1/24}{\epsilon} \leq cn^2\cdot\frac{\left\|\boldsymbol{X}^\natural\right\|}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}$.

Therefore, when $m \geq c\frac{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^4}{\sigma_r^4\left(\boldsymbol{X}^\natural\right)} nr\log\left(n\kappa\right)$ with some large enough constant c, for all matrices $\boldsymbol{V}$ and $\boldsymbol{X}$ such that $\left\|\boldsymbol{X}-\boldsymbol{X}^\natural\right\|_{\mathsf{F}} \leq \frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}$, we have

$$q\left(\boldsymbol{V},\boldsymbol{H},t\right) \geq \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 + 2\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right\|_{\mathsf{F}}^2 + 2\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) + 1.246\sigma_r^2\left(\boldsymbol{X}^\natural\right), \quad \text{(B.45)}$$

with probability at least $1-e^{-c_1 nr\log\left(n\kappa\right)} - me^{-1.5n}$.

B.6.3 Finishing the Proof

Combining (B.37) and (B.45), we can prove

$$\begin{aligned}
\mathrm{vec}\left(\boldsymbol{V}\right)^\top \nabla^2 f(\boldsymbol{X})\mathrm{vec}\left(\boldsymbol{V}\right) &\geq \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 + 2\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right\|_{\mathsf{F}}^2 + 2\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) + 1.246\sigma_r^2\left(\boldsymbol{X}^\natural\right) \\
&\quad - \frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{a}_i^\top\boldsymbol{X}^\natural\right\|_2^2\left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2 \\
&\geq \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 + 2\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right\|_{\mathsf{F}}^2 + 2\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) + 1.246\sigma_r^2\left(\boldsymbol{X}^\natural\right) \\
&\quad - \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 - 2\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right\|_{\mathsf{F}}^2 - \frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^\natural\right) \\
&\geq 2\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) + 1.204\sigma_r^2\left(\boldsymbol{X}^\natural\right),
\end{aligned}$$

as claimed.

Appendix C: Technical Proofs in Chapter 3

C.1 Proof of Lemma 6: Approximate Dual Certificate

Denote the solution to (3.4) by $\hat{\boldsymbol{X}} = \boldsymbol{X}_0 + \boldsymbol{H} \neq \boldsymbol{X}_0$, then we have $\hat{\boldsymbol{X}} \succeq 0$, $\boldsymbol{H}_{T^\perp} \succeq 0$, and furthermore,

$$\begin{aligned}\|\mathcal{A}(\boldsymbol{H}) - (\boldsymbol{\eta} + \boldsymbol{w})\|_1 &= \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X}_0 + \boldsymbol{H})\|_1 \\ &= \|\boldsymbol{y} - \mathcal{A}(\hat{\boldsymbol{X}})\|_1 \\ &\leq \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X}_0)\|_1 = \|\boldsymbol{\eta} + \boldsymbol{w}\|_1,\end{aligned}$$

where the inequality follows from the optimality of $\hat{\boldsymbol{X}}$ since both $\hat{\boldsymbol{X}}$ and $\boldsymbol{X}_0$ are feasible to (3.4). Since

$$\|\mathcal{A}(\boldsymbol{H}) - (\boldsymbol{\eta} + \boldsymbol{w})\|_1 = \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{H}) - \boldsymbol{\eta} - \boldsymbol{w}_{\mathcal{S}}\|_1 + \|\mathcal{A}_{\mathcal{S}^\perp}(\boldsymbol{H}) - \boldsymbol{w}_{\mathcal{S}^\perp}\|_1,$$

and

$$\|\boldsymbol{\eta} + \boldsymbol{w}\|_1 = \|\boldsymbol{\eta} + \boldsymbol{w}_{\mathcal{S}}\|_1 + \|\boldsymbol{w}_{\mathcal{S}^\perp}\|_1,$$

where $\boldsymbol{w}_{\mathcal{S}} \in \mathbb{R}^m$ is a vector whose entries are same with $\boldsymbol{w}$ on indices in $\mathcal{S}$, and otherwise are zeros, and $\boldsymbol{w} = \boldsymbol{w}_{\mathcal{S}} + \boldsymbol{w}_{\mathcal{S}^\perp}$, we have

$$\|\mathcal{A}_{\mathcal{S}^\perp}(\boldsymbol{H})\|_1 \leq \|\mathcal{A}_{\mathcal{S}^\perp}(\boldsymbol{H}) - \boldsymbol{w}_{\mathcal{S}^\perp}\|_1 + \|\boldsymbol{w}_{\mathcal{S}^\perp}\|_1$$

$$\leq \|\boldsymbol{\eta}+\boldsymbol{w}\|_1 - \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{H})-\boldsymbol{\eta}-\boldsymbol{w}_{\mathcal{S}}\|_1 + \|\boldsymbol{w}_{\mathcal{S}^\perp}\|_1$$
$$\leq \|\boldsymbol{\eta}+\boldsymbol{w}_{\mathcal{S}}\|_1 - \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{H})-\boldsymbol{\eta}-\boldsymbol{w}_{\mathcal{S}}\|_1 + 2\|\boldsymbol{w}_{\mathcal{S}^\perp}\|_1$$
$$\leq \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{H})\|_1 + 2\|\boldsymbol{w}_{\mathcal{S}^\perp}\|_1,$$

where the last inequality follows from the triangle inequality. We could further bound

$$\begin{aligned}\|\mathcal{A}_{\mathcal{S}^\perp}(\boldsymbol{H}_T)\|_1 &\leq \|\mathcal{A}_{\mathcal{S}^\perp}(\boldsymbol{H})\|_1 + \|\mathcal{A}_{\mathcal{S}^\perp}(\boldsymbol{H}_{T^\perp})\|_1 \\ &\leq \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{H})\|_1 + \|\mathcal{A}_{\mathcal{S}^\perp}(\boldsymbol{H}_{T^\perp})\|_1 + 2\|\boldsymbol{w}_{\mathcal{S}^\perp}\|_1 \\ &\leq \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{H}_T)\|_1 + \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{H}_{T^\perp})\|_1 + \|\mathcal{A}_{\mathcal{S}^\perp}(\boldsymbol{H}_{T^\perp})\|_1 + 2\|\boldsymbol{w}_{\mathcal{S}^\perp}\|_1 \\ &= \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{H}_T)\|_1 + \|\mathcal{A}(\boldsymbol{H}_{T^\perp})\|_1 + 2\|\boldsymbol{w}_{\mathcal{S}^\perp}\|_1. \end{aligned} \tag{C.1}$$

Our assumptions on $\mathcal{A}$ imply that

$$\begin{aligned}\left(1+\frac{1}{10}\right)\mathrm{Tr}\left(\boldsymbol{H}_{T^\perp}\right) &\geq \frac{1}{m}\left\|\mathcal{A}\left(\boldsymbol{H}_{T^\perp}\right)\right\|_1 \\ &\geq \frac{1}{m}\left(\|\mathcal{A}_{\mathcal{S}^\perp}(\boldsymbol{H}_T)\|_1 - \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{H}_T)\|_1 - 2\|\boldsymbol{w}_{\mathcal{S}^\perp}\|_1\right) \\ &\geq \frac{|\mathcal{S}^\perp|}{5m}\left(1-\frac{1}{12}\right)\|\boldsymbol{H}_T\|_{\mathsf{F}} - \frac{|\mathcal{S}|}{m}\left(1+\frac{1}{10}\right)\|\boldsymbol{H}_T\|_1 - \frac{2\epsilon}{m},\end{aligned}$$

where the first inequality follows from (3.7) due to $\left\|\boldsymbol{H}_{T^\perp}\right\|_1 = \mathrm{Tr}\left(\boldsymbol{H}_{T^\perp}\right)$, as $\boldsymbol{H}_{T^\perp} \succeq 0$, the second inequality follows from (C.1), and the last inequality follows from (3.8) and (3.9). This gives

$$\mathrm{Tr}\left(\boldsymbol{H}_{T^\perp}\right) \geq \left(\frac{|\mathcal{S}^\perp|}{6m} - \frac{|\mathcal{S}|}{m}\sqrt{2r}\right)\|\boldsymbol{H}_T\|_{\mathsf{F}} - \frac{2\epsilon}{m}, \tag{C.2}$$

where we use the inequality $\|\boldsymbol{H}_T\|_1 \leq \sqrt{2r}\|\boldsymbol{H}_T\|_{\mathsf{F}}$.

On the other hand, since $\boldsymbol{\mu}/(9/m)$ is a subgradient of the ℓ_1-norm at $\boldsymbol{\eta}$ from (3.11), we have

$$\|\boldsymbol{\eta}\|_1 + \left\langle \frac{m}{9}\boldsymbol{\mu}, \boldsymbol{w}-\mathcal{A}(\boldsymbol{H})\right\rangle \leq \|\boldsymbol{w}+\boldsymbol{\eta}-\mathcal{A}(\boldsymbol{H})\|_1 \leq \|\boldsymbol{\eta}+\boldsymbol{w}\|_1 \leq \|\boldsymbol{\eta}\|_1 + \|\boldsymbol{w}\|_1,$$

which, by a simple transformation, is

$$\langle \boldsymbol{\mu}, \mathcal{A}(\boldsymbol{H}) \rangle \geq \langle \boldsymbol{\mu}, \boldsymbol{w} \rangle - \frac{9}{m} \|\boldsymbol{w}\|_1 \geq -\left(\|\boldsymbol{\mu}\|_\infty + \frac{9}{m} \right) \|\boldsymbol{w}\|_1 \geq -\frac{18\epsilon}{m}.$$

Then with

$$\langle \boldsymbol{H}, \boldsymbol{Y} \rangle = \langle \mathcal{A}(\boldsymbol{H}), \boldsymbol{\mu} \rangle,$$

we can get

$$\begin{aligned} -\frac{18\epsilon}{m} &\leq \langle \mathcal{A}(\boldsymbol{H}), \boldsymbol{\mu} \rangle = \langle \boldsymbol{H}, \boldsymbol{Y} \rangle = \langle \boldsymbol{H}_T, \boldsymbol{Y}_T \rangle + \langle \boldsymbol{H}_{T^\perp}, \boldsymbol{Y}_{T^\perp} \rangle \\ &\leq \|\boldsymbol{Y}_T\|_{\mathsf{F}} \|\boldsymbol{H}_T\|_{\mathsf{F}} - \frac{1}{r} \langle \boldsymbol{H}_{T^\perp}, \boldsymbol{I}_{T^\perp} \rangle \\ &\leq \frac{1}{13r} \|\boldsymbol{H}_T\|_{\mathsf{F}} - \frac{1}{r} \mathrm{Tr}(\boldsymbol{H}_{T^\perp}), \end{aligned}$$

which gives

$$\mathrm{Tr}(\boldsymbol{H}_{T^\perp}) \leq \frac{1}{13} \|\boldsymbol{H}_T\|_{\mathsf{F}} + \frac{18r\epsilon}{m}. \tag{C.3}$$

Combining with (C.2), we know

$$\left(\frac{|\mathcal{S}^\perp|}{6m} - \frac{|\mathcal{S}|}{m} \sqrt{2r} \right) \|\boldsymbol{H}_T\|_{\mathsf{F}} - \frac{2\epsilon}{m} \leq \frac{1}{13} \|\boldsymbol{H}_T\|_{\mathsf{F}} + \frac{18r\epsilon}{m}.$$

Since $\frac{|\mathcal{S}^\perp|}{6m} - \frac{|\mathcal{S}|}{m}\sqrt{2r} - \frac{1}{13} > 0$ under the assumption on $\frac{|\mathcal{S}|}{m}$ in Lemma 6, we have

$$\|\boldsymbol{H}_T\|_{\mathsf{F}} \leq \frac{20r\epsilon}{m\left(\frac{|\mathcal{S}^\perp|}{6m} - \frac{|\mathcal{S}|}{m}\sqrt{2r} - \frac{1}{13} \right)} \leq c_1 \frac{r\epsilon}{m},$$

where c_1 is some fixed constant. Finally, we have

$$\begin{aligned} \|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\mathsf{F}} &\leq \|\boldsymbol{H}_T\|_{\mathsf{F}} + \|\boldsymbol{H}_{T^\perp}\|_{\mathsf{F}} \\ &\leq \|\boldsymbol{H}_T\|_{\mathsf{F}} + \mathrm{Tr}(\boldsymbol{H}_{T^\perp}) \\ &\leq \left(1 + \frac{1}{13} \right) \|\boldsymbol{H}_T\|_{\mathsf{F}} + \frac{18r\epsilon}{m} \leq c\frac{r\epsilon}{m}, \end{aligned}$$

for some constant c.

C.2 Proof of Lemma 9

First, by standard results in random matrix theory [120, Corollary 5.35], we have

$$\left\| \frac{m}{|\mathcal{S}^\perp|} \boldsymbol{Y}^{(1)} - \left(\alpha_0 + \frac{\beta_0 - \alpha_0}{r} \right) \boldsymbol{I} \right\| \le \frac{\beta_0}{40r},$$

with probability at least $1 - 2e^{-\gamma |\mathcal{S}^\perp|/r^2}$ for some constant γ provided $\left|\mathcal{S}^\perp\right| \ge cnr^2$ for some constant c. In particular, this gives

$$\left\| \frac{m}{|\mathcal{S}^\perp|} \boldsymbol{Y}^{(1)}_{T^\perp} - \left(\alpha_0 + \frac{\beta_0 - \alpha_0}{r} \right) \boldsymbol{I}_{T^\perp} \right\| \le \frac{\beta_0}{40r}. \tag{C.4}$$

Let $\boldsymbol{a}_j' = (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{a}_j$ be the projection of $\boldsymbol{a}_j$ onto the orthogonal complement of the column space of $\boldsymbol{U}$, then we have

$$\boldsymbol{Y}^{(0)}_{T^\perp} = \frac{1}{m} \sum_{j \in \mathcal{S}^\perp} \boldsymbol{\epsilon}_j \boldsymbol{\epsilon}_j^\top,$$

where $\boldsymbol{\epsilon}_j = \left(\frac{1}{r} \sum_{i=1}^r \left| \boldsymbol{a}_j^\top \boldsymbol{u}_i \right|^2 \mathbb{I}_{\left\{ \left| \boldsymbol{a}_j^\top \boldsymbol{u}_i \right| \le 3 \right\}} \right)^{1/2} \boldsymbol{a}_j'$, $j \in \mathcal{S}^\perp$, are i.i.d. copies of a zero-mean, isotropic and sub-Gaussian random vector $\boldsymbol{\epsilon}$, which satisfies $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \alpha_0 \boldsymbol{I}_{T^\perp}$. Following [120, Theorem 5.39], we have

$$\left\| \frac{m}{|\mathcal{S}^\perp|} \boldsymbol{Y}^{(0)}_{T^\perp} - \alpha_0 \boldsymbol{I}_{T^\perp} \right\| \le \frac{\alpha_0}{40r}, \tag{C.5}$$

with probability at least $1 - 2e^{-\gamma |\mathcal{S}^\perp|/r^2}$ for some constant γ provided $\left|\mathcal{S}^\perp\right| \ge cnr^2$ for some constant c. As a result, if $m \ge cnr^2$ for some large constant c and $|\mathcal{S}| \le c_1 m$ for some constant c_1 small enough, with probability at least $1 - e^{-\gamma m/r^2}$, there exists

$$\begin{aligned}
&\left\| \boldsymbol{Y}^{(0)}_{T^\perp} - \boldsymbol{Y}^{(1)}_{T^\perp} + \frac{\beta_0 - \alpha_0}{r} \boldsymbol{I}_{T^\perp} \right\| \\
&\le \left\| \boldsymbol{Y}^{(0)}_{T^\perp} - \boldsymbol{Y}^{(1)}_{T^\perp} + \frac{\beta_0 - \alpha_0}{r} \frac{\left|\mathcal{S}^\perp\right|}{m} \boldsymbol{I}_{T^\perp} \right\| + \left(1 - \frac{\left|\mathcal{S}^\perp\right|}{m} \right) \frac{\beta_0 - \alpha_0}{r} \\
&\le \left\| \frac{m}{|\mathcal{S}^\perp|} \boldsymbol{Y}^{(0)}_{T^\perp} - \frac{m}{|\mathcal{S}^\perp|} \boldsymbol{Y}^{(1)}_{T^\perp} + \frac{\beta_0 - \alpha_0}{r} \boldsymbol{I}_{T^\perp} \right\| + \frac{|\mathcal{S}|}{m} \frac{\beta_0 - \alpha_0}{r}
\end{aligned}$$

$$\leq \frac{\beta_0}{30r} + \frac{\alpha_0}{60r}. \tag{C.6}$$

Next, let's check $\left\|\boldsymbol{Y}^{(2)}_{T^\perp}\right\|$. Since $\boldsymbol{Y}^{(2)} = \frac{1}{m}\sum_{j\in\mathcal{S}} 9\chi_j \boldsymbol{a}_j\boldsymbol{a}_j^\top$, where $\mathbb{E}[9\chi_j\boldsymbol{a}_j\boldsymbol{a}_j^\top] = \boldsymbol{0}$, by [120, Theorem 5.39] we have

$$\left\|\boldsymbol{Y}^{(2)}\right\| = \frac{|\mathcal{S}|}{m}\left\|\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} 9\chi_j\boldsymbol{a}_j\boldsymbol{a}_j^\top\right\| \leq \frac{1}{10r},$$

with probability at least $1-2\exp(-\gamma m/r)$ as long as $m \geq cnr^2$ and $|\mathcal{S}| = c_1 m/r \geq c_2 nr$, for some constants c, c_1 and c_2. In particular, this gives

$$\left\|\boldsymbol{Y}^{(2)}_{T^\perp}\right\| \leq \frac{1}{10r}. \tag{C.7}$$

Putting this together with (C.6), we can obtain that if $m \geq cnr^2$ and $|\mathcal{S}| = c_1 m/r \geq c_2 nr$ for some constants c, c_1 and c_2, with probability at least $1-e^{-\gamma m/r^2}$,

$$\left\|\boldsymbol{Y}_{T^\perp} + \frac{1.7}{r}\boldsymbol{I}_{T^\perp}\right\| = \left\|\boldsymbol{Y}^{(0)}_{T^\perp} - \boldsymbol{Y}^{(1)}_{T^\perp} + \boldsymbol{Y}^{(2)}_{T^\perp} + \frac{1.7}{r}\boldsymbol{I}_{T^\perp}\right\| \left(\frac{\alpha_0}{60} + \frac{\beta_0}{30} + 0.11\right)\frac{1}{r} \leq \frac{0.25}{r}.$$

C.3 Proof of Lemma 10

Let $\tilde{\boldsymbol{Y}} = \left(\boldsymbol{Y}^{(0)} - \boldsymbol{Y}^{(1)}\right)\boldsymbol{U}$, and $\tilde{\boldsymbol{Y}}' = \left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top\right)\tilde{\boldsymbol{Y}}$ be the projection of $\tilde{\boldsymbol{Y}}$ onto the orthogonal complement of $\boldsymbol{U}$, then we have

$$\left\|\boldsymbol{Y}^{(0)}_T - \boldsymbol{Y}^{(1)}_T\right\|_{\mathsf{F}}^2 = \left\|\boldsymbol{U}^\top\tilde{\boldsymbol{Y}}\right\|_{\mathsf{F}}^2 + 2\left\|\tilde{\boldsymbol{Y}}'\right\|_{\mathsf{F}}^2. \tag{C.8}$$

First consider the term $\|\boldsymbol{U}^\top\tilde{\boldsymbol{Y}}\|_{\mathsf{F}}^2$ in (C.8), where the kth column of $\boldsymbol{U}^\top\tilde{\boldsymbol{Y}}$ can be expressed explicitly as

$$\begin{aligned}\left(\boldsymbol{U}^\top\tilde{\boldsymbol{Y}}\right)_k &= \frac{1}{m}\sum_{j\in\mathcal{S}^\perp}\left[\frac{1}{r}\sum_{i=1}^r \left|\boldsymbol{a}_j^\top\boldsymbol{u}_i\right|^2 \mathbb{I}_{\left\{|\boldsymbol{a}_j^\top\boldsymbol{u}_i|\leq 3\right\}} - \left(\alpha_0 + \frac{\beta_0-\alpha_0}{r}\right)\right]\cdot\left(\boldsymbol{a}_j^\top\boldsymbol{u}_k\right)\left(\boldsymbol{U}^\top\boldsymbol{a}_j\right) \\ &:= \frac{1}{m}\boldsymbol{\Phi}\boldsymbol{c}_k,\end{aligned}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{r\times|\mathcal{S}^{\perp}|}$ is constructed by $\boldsymbol{U}^{\top}\boldsymbol{a}_j$'s, and $\boldsymbol{c}_k \in \mathbb{R}^{|\mathcal{S}^{\perp}|}$ is composed of $c_{k,j}$'s, each one expressed as

$$c_{k,j} = \left[\frac{1}{r}\sum_{i=1}^{r}\left|\boldsymbol{a}_j^{\top}\boldsymbol{u}_i\right|^2 \mathbb{I}_{\left\{\left|\boldsymbol{a}_j^{\top}\boldsymbol{u}_i\right|\leq 3\right\}} - \left(\alpha_0 + \frac{\beta_0-\alpha_0}{r}\right)\right]\left(\boldsymbol{a}_j^{\top}\boldsymbol{u}_k\right),$$

with

$$\begin{aligned}\mathbb{E}[c_{k,j}^2] &= \frac{1}{r^2}\left(\theta_0 + (r-1)\,\beta_0 - \beta_0^2 - (r-1)\,\alpha_0^2\right)\\ &= \frac{1}{r}\left(\beta_0 - \alpha_0^2\right) + \frac{1}{r^2}\left(\theta_0 + \alpha_0^2 - \beta_0^2 - \beta_0\right) \leq \frac{4.07}{r}.\end{aligned}$$

Note that $c_{k,j}^2$'s are i.i.d. sub-exponential random variables with $\left\|c_{k,j}^2\right\|_{\psi_1} \leq K$, for some constant K, then according to [120, Corollary 5.17],

$$\mathbb{P}\left\{\left|\sum_{j\in\mathcal{S}^{\perp}}\left(c_{k,j}^2 - \mathbb{E}c_{k,j}^2\right)\right| \geq \frac{\epsilon}{r}\left|\mathcal{S}^{\perp}\right|\right\} \leq 2\exp\left(-c\frac{\epsilon^2\left|\mathcal{S}^{\perp}\right|}{K^2r^2}\right),$$

which shows that as long as $|\mathcal{S}| \leq c_1 m$, for some constants c and c_1,

$$\|\boldsymbol{c}_k\|_2^2 \leq \frac{4.07+c}{r}m \leq \frac{4.1m}{r}$$

holds with probability at least $1-e^{-\gamma m/r^2}$. Furthermore, for a fixed vector $\boldsymbol{x} \in \mathbb{R}^{|\mathcal{S}^{\perp}|}$ obeying $\|\boldsymbol{x}\|_2 = 1$, $\|\boldsymbol{\Phi}\boldsymbol{x}\|_2^2$ is distributed as a chi-square random variable with r degrees of freedom. From [121, Lemma 1], we have

$$\|\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 \leq \frac{m}{12000r^2},$$

with probability at least $1-e^{-\gamma m/r^2}$, provided $m \geq cnr^2$ for some sufficiently large constant c. Therefore, we can obtain

$$\left\|\left(\boldsymbol{U}^{\top}\tilde{\boldsymbol{Y}}\right)_k\right\|_2^2 = \frac{1}{m^2}\left\|\boldsymbol{\Phi}\frac{\boldsymbol{c}_k}{\|\boldsymbol{c}_k\|_2}\right\|_2^2\|\boldsymbol{c}_k\|_2^2 \leq \frac{1}{2700r^3},$$

which yields

$$\|\boldsymbol{U}^{\top}\tilde{\boldsymbol{Y}}\|_{\mathsf{F}}^2 = \sum_{k=1}^{r}\left\|\left(\boldsymbol{U}^{\top}\tilde{\boldsymbol{Y}}\right)_k\right\|_2^2 \leq \frac{1}{2700r^2}, \tag{C.9}$$

with probability at least $1 - e^{-\gamma m/r^2}$, when $m \geq cnr^2$ and $|\mathcal{S}| \leq c_1 m$.

To bound the second term in (C.8), we could adopt the same techniques as before. The kth column of $\tilde{\boldsymbol{Y}}'$ can be expressed explicitly as

$$\begin{aligned}\left(\tilde{\boldsymbol{Y}}'\right)_k &= \frac{1}{m}\sum_{j\in\mathcal{S}^\perp}\left[\frac{1}{r}\sum_{i=1}^{r}\left|\boldsymbol{a}_j^\top\boldsymbol{u}_i\right|^2\mathbb{I}_{\left\{|\boldsymbol{a}_j^\top\boldsymbol{u}_i|\leq 3\right\}} - \left(\alpha_0 + \frac{\beta_0-\alpha_0}{r}\right)\right]\cdot\left(\boldsymbol{a}_j^\top\boldsymbol{u}_k\right)\left(\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^\top\right)\boldsymbol{a}_j \\ &:= \frac{1}{m}\sum_{j\in\mathcal{S}^\perp} c_{k,j}\boldsymbol{a}_j' := \frac{1}{m}\boldsymbol{\Psi}\boldsymbol{c}_k,\end{aligned}$$

where $\boldsymbol{\Psi} \in \mathbb{R}^{n\times|\mathcal{S}^\perp|}$ is constructed by $\boldsymbol{a}_j'$'s, each of which, as a reminder, is the projection of $\boldsymbol{a}_j$ onto the orthogonal complement of the column space of $\boldsymbol{U}$ as $\boldsymbol{a}_j' = \left(\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^\top\right)\boldsymbol{a}_j$. Equivalently, $\boldsymbol{\Psi} = \left(\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^\top\right)\boldsymbol{A}$, where $\boldsymbol{A} \in \mathbb{R}^{n\times|\mathcal{S}^\perp|}$ is constructed by $\boldsymbol{a}_j$'s, $j \in \mathcal{S}^\perp$. For a fixed vector $\boldsymbol{x} \in \mathbb{R}^{|\mathcal{S}^\perp|}$ obeying $\|\boldsymbol{x}\|_2 = 1$, we have $\|\boldsymbol{\Psi}\boldsymbol{x}\|_2^2 = \left\|\left(\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^\top\right)\boldsymbol{A}\boldsymbol{x}\right\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{x}\|_2^2$, where $\|\boldsymbol{A}\boldsymbol{x}\|_2^2$ is distributed as a chi-square random variable with n degrees of freedom. Again [121, Lemma 1] tells us

$$\|\boldsymbol{\Psi}\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq \frac{m}{12000r^2},$$

with probability exceeding $1 - e^{-\gamma m/r^2}$, provided $m \geq cnr^2$ for a sufficiently large constant c. Hence,

$$\left\|\left(\tilde{\boldsymbol{Y}}'\right)_k\right\|_2^2 \leq \frac{1}{m^2}\left\|\boldsymbol{\Psi}\frac{\boldsymbol{c}_k}{\|\boldsymbol{c}_k\|_2}\right\|_2^2\|\boldsymbol{c}_k\|_2^2 \leq \frac{1}{2700r^3},$$

which leads to

$$\left\|\tilde{\boldsymbol{Y}}'\right\|_{\mathsf{F}}^2 = \sum_{k=1}^{r}\left\|\left(\tilde{\boldsymbol{Y}}'\right)_k\right\|_2^2 \leq \frac{1}{2700r^2}. \tag{C.10}$$

Then, combining (C.9) and (C.10), we know that

$$\left\|\boldsymbol{Y}_T^{(0)} - \boldsymbol{Y}_T^{(1)}\right\|_{\mathsf{F}} = \sqrt{\left\|\boldsymbol{U}^\top\tilde{\boldsymbol{Y}}\right\|_{\mathsf{F}}^2 + 2\left\|\tilde{\boldsymbol{Y}}'\right\|_{\mathsf{F}}^2} \leq \frac{1}{30r}. \tag{C.11}$$

Next, let's check $\|\boldsymbol{Y}_T^{(2)}\|_{\mathsf{F}}^2$, which can be written as

$$\|\boldsymbol{Y}_T^{(2)}\|_{\mathsf{F}}^2 = \|\boldsymbol{U}^\top\bar{\boldsymbol{Y}}\|_{\mathsf{F}}^2 + 2\|\bar{\boldsymbol{Y}}'\|_{\mathsf{F}}^2,$$

where $\bar{\boldsymbol{Y}} = \boldsymbol{Y}^{(2)}\boldsymbol{U}$ and $\bar{\boldsymbol{Y}}' = (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\bar{\boldsymbol{Y}}$. For the first term $\|\boldsymbol{U}^\top\bar{\boldsymbol{Y}}\|_{\mathsf{F}}^2$, the kth column of $\boldsymbol{U}^\top\bar{\boldsymbol{Y}}$ can be formulated explicitly as

$$\left(\boldsymbol{U}^\top\bar{\boldsymbol{Y}}\right)_k = \frac{1}{m}\sum_{j\in\mathcal{S}} 9\chi_j\left(\boldsymbol{a}_j^\top\boldsymbol{u}_k\right)\left(\boldsymbol{U}^\top\boldsymbol{a}_j\right) := \frac{1}{m}\bar{\boldsymbol{\Phi}}\boldsymbol{d}_k,$$

where $\bar{\boldsymbol{\Phi}} \in \mathbb{R}^{r\times|\mathcal{S}|}$ is constructed by $\boldsymbol{U}^\top\boldsymbol{a}_j$'s, and $\boldsymbol{d}_k \in \mathbb{R}^{|\mathcal{S}|}$ is composed of $d_{k,j}$'s, each one expressed as

$$d_{k,j} = 9\chi_j\left(\boldsymbol{a}_j^\top\boldsymbol{u}_k\right),$$

with $\mathbb{E}[d_{k,j}^2] = 81$. Note that $d_{k,j}^2$'s are i.i.d. sub-exponential random variables with $\left\|d_{k,j}^2\right\|_{\psi_1} \le K$, for some constant K, then based on [120, Corollary 5.17],

$$\mathbb{P}\left\{\left|\sum_{j\in\mathcal{S}}\left(d_{k,j}^2 - \mathbb{E}d_{k,j}^2\right)\right| \ge \epsilon\,|\mathcal{S}|\right\} \le 2\exp\left(-c_1\frac{\epsilon^2\,|\mathcal{S}|}{K^2}\right),$$

which indicates that if $|\mathcal{S}| = cm/r$, for some constant c,

$$\|\boldsymbol{d}_k\|_2^2 \le (81 + c_1)\,|\mathcal{S}| \le 82\,|\mathcal{S}| := \delta_0\,|\mathcal{S}|$$

holds with probability at least $1 - e^{-\gamma m/r}$. And for a fixed vector $\boldsymbol{x} \in \mathbb{R}^{|\mathcal{S}|}$ obeying $\|\boldsymbol{x}\|_2 = 1$, $\left\|\bar{\boldsymbol{\Phi}}\boldsymbol{x}\right\|_2^2$ is also a chi-square random variable with r degrees of freedom, so

$$\left\|\bar{\boldsymbol{\Phi}}\boldsymbol{x}\right\|_2^2 \le \frac{m}{2700\delta_0 c r^2},$$

with probability at least $1 - e^{-\gamma m/r^2}$, provided $m \ge c_1 n r^2$ for some sufficiently large constant c_1. Thus we have

$$\left\|\left(\boldsymbol{U}^\top\bar{\boldsymbol{Y}}\right)_k\right\|_2^2 = \frac{1}{m^2}\left\|\bar{\boldsymbol{\Phi}}\frac{\boldsymbol{d}_k}{\|\boldsymbol{d}_k\|_2}\right\|_2^2\|\boldsymbol{d}_k\|_2^2 \le \frac{1}{2700r^3},$$

which gives

$$\|\boldsymbol{U}^\top\bar{\boldsymbol{Y}}\|_{\mathsf{F}}^2 = \sum_{k=1}^{r}\left\|\left(\boldsymbol{U}^\top\bar{\boldsymbol{Y}}\right)_k\right\|_2^2 \le \frac{1}{2700r^2}, \tag{C.12}$$

with probability at least $1 - e^{-\gamma m/r^2}$, when $m \geq c_1 n r^2$ and $|\mathcal{S}| = cm/r$, for some appropriate constants c and c_1.

Now consider the second term $\|\bar{\boldsymbol{Y}}'\|_{\mathsf{F}}^2$ in $\|\boldsymbol{Y}_T^{(2)}\|_{\mathsf{F}}^2$, where the kth column of $\bar{\boldsymbol{Y}}'$ can be expressed explicitly as

$$\left(\bar{\boldsymbol{Y}}'\right)_k = \frac{1}{m}\sum_{j\in\mathcal{S}} 9\chi_j(\boldsymbol{a}_j^\top \boldsymbol{u}_k)\left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top\right)\boldsymbol{a}_j := \frac{1}{m}\sum_{j\in\mathcal{S}} d_{k,j}\boldsymbol{a}_j' := \frac{1}{m}\bar{\boldsymbol{\Psi}}\boldsymbol{d}_k,$$

where $\bar{\boldsymbol{\Psi}} \in \mathbb{R}^{n\times|\mathcal{S}|}$ is constructed by $\boldsymbol{a}_j'$'s. Also, we can decompose $\bar{\boldsymbol{\Psi}}$ as $\bar{\boldsymbol{\Psi}} = \left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top\right)\bar{\boldsymbol{A}}$, where $\bar{\boldsymbol{A}} \in \mathbb{R}^{n\times|\mathcal{S}|}$ is constructed by $\boldsymbol{a}_j$'s, $j \in \mathcal{S}$. For a fixed vector $\boldsymbol{x} \in \mathbb{R}^{|\mathcal{S}|}$ obeying $\|\boldsymbol{x}\|_2 = 1$, we have $\left\|\bar{\boldsymbol{\Psi}}\boldsymbol{x}\right\|_2^2 = \left\|\left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top\right)\bar{\boldsymbol{A}}\boldsymbol{x}\right\|_2^2 \leq \left\|\bar{\boldsymbol{A}}\boldsymbol{x}\right\|_2^2$, where $\left\|\bar{\boldsymbol{A}}\boldsymbol{x}\right\|_2^2$ is a chi-square random variable with n degrees of freedom as well. Since we already know that provided $m \geq c_1 n r^2$ for a sufficiently large constant c_1,

$$\left\|\bar{\boldsymbol{\Psi}}\boldsymbol{x}\right\|_2^2 \leq \left\|\bar{\boldsymbol{A}}\boldsymbol{x}\right\|_2^2 \leq \frac{m}{2700\delta_0 c r^2},$$

with probability exceeding $1 - e^{-\gamma m/r^2}$, we can have

$$\left\|\left(\bar{\boldsymbol{Y}}'\right)_k\right\|_2^2 \leq \frac{1}{m^2}\left\|\bar{\boldsymbol{\Psi}}\frac{\boldsymbol{d}_k}{\|\boldsymbol{d}_k\|_2}\right\|_2^2 \|\boldsymbol{d}_k\|_2^2 \leq \frac{1}{2700 r^3},$$

and a further result

$$\left\|\bar{\boldsymbol{Y}}'\right\|_{\mathsf{F}}^2 = \sum_{k=1}^{r}\left\|\left(\bar{\boldsymbol{Y}}'\right)_k\right\|_2^2 \leq \frac{1}{2700 r^2}, \tag{C.13}$$

which, combining with (C.12), leads to

$$\left\|\boldsymbol{Y}_T^{(2)}\right\|_{\mathsf{F}} = \sqrt{\left\|\boldsymbol{U}^\top\bar{\boldsymbol{Y}}\right\|_{\mathsf{F}}^2 + 2\left\|\bar{\boldsymbol{Y}}'\right\|_{\mathsf{F}}^2} \leq \frac{1}{30r}. \tag{C.14}$$

Finally, we can obtain that if $m \geq cnr^2$ and $|\mathcal{S}| = c_1 m/r$, for some constants c and c_1, with probability at least $1 - e^{-\gamma m/r^2}$,

$$\|\boldsymbol{Y}_T\|_{\mathsf{F}} = \left\|\boldsymbol{Y}_T^{(0)} - \boldsymbol{Y}_T^{(1)} + \boldsymbol{Y}_T^{(2)}\right\|_{\mathsf{F}} \leq \frac{1}{15r}. \tag{C.15}$$

Appendix D: Technical Proofs in Chapter 4

D.1 Proof of Proposition 1

Due to scaling invariance, without loss of generality, it is sufficient to consider all rank-$2r$ matrices with unit Frobenius norm. First, we fix the rank-$2r$ matrix $\boldsymbol{G}_0 \in \mathbb{R}^{n_1 \times n_2}$, and then generalize to all rank-$2r$ matrices by a covering argument. Note that $|\mathcal{A}_i(\boldsymbol{G}_0)|$, $i = 1, \cdots, m$, are i.i.d. copies of $|\langle \boldsymbol{A}, \boldsymbol{G}_0 \rangle|$, where $\boldsymbol{A}$ is generated with i.i.d. standard Gaussian entries. Since $\|\boldsymbol{G}_0\|_{\mathsf{F}} = 1$, $\langle \boldsymbol{A}, \boldsymbol{G}_0 \rangle$ follows the distribution $\mathcal{N}(0,1)$, and $|\langle \boldsymbol{A}, \boldsymbol{G}_0 \rangle|$ follows a folded normal distribution, whose probability density function and cumulative distribution function are denoted by f_1 and F_1, respectively. It is known from Lemma 20 that

$$0.6745 - \epsilon \leq \operatorname{med}(|\mathcal{A}(\boldsymbol{G}_0)|) \leq 0.6745 + \epsilon, \tag{D.1}$$

with probability at least $1 - 2\exp(-cm\epsilon^2)$ for a small ϵ, where c is a constant around 2×0.6356^2. Similar arguments extend to other quantiles. From Lemma 20, we have

$$0.6588 - \epsilon \leq \theta_{0.49}(|\mathcal{A}(\boldsymbol{G}_0)|) \leq 0.6588 + \epsilon; \tag{D.2}$$

$$0.6903 - \epsilon \leq \theta_{0.51}(|\mathcal{A}(\boldsymbol{G}_0)|) \leq 0.6903 + \epsilon, \tag{D.3}$$

with probability at least $1 - 2\exp(-cm\epsilon^2)$ for a small ϵ, where c is a constant around 2×0.6287^2.

Next, we extend the results to all rank-$2r$ matrices $\boldsymbol{G}$ with $\|\boldsymbol{G}\|_{\mathsf{F}} = 1$ via a covering argument. We argue for the median and similar arguments extend to other quantiles straightforwardly. Let $\mathcal{N}_\tau$ be a τ-net covering all rank-$2r$ matrices with respect to the Frobenius norm. Let $n = (n_1 + n_2)/2$, then from Lemma 13, $|\mathcal{N}_\tau| \leq (9/\tau)^{2r(2n+1)}$. Taking the union bound, we obtain

$$0.6745 - \epsilon \leq \operatorname{med}\left(|\mathcal{A}(\boldsymbol{G}_0)|\right) \leq 0.6745 + \epsilon, \quad \forall \boldsymbol{G}_0 \in \mathcal{N}_\tau, \tag{D.4}$$

with probability at least $1 - (9/\tau)^{2r(2n+1)} \exp\left(-cm\epsilon^2\right)$. Set $\tau = \epsilon/(2\sqrt{n(n+m)})$. Under this event and (A.3), which holds with probability at least $1 - m\exp\left(-n\left(n+m\right)\right)$ from Lemma 25, for any rank-$2r$ matrix $\boldsymbol{G}$ with $\|\boldsymbol{G}\|_{\mathsf{F}} = 1$, there exists $\boldsymbol{G}_0 \in \mathcal{N}_\tau$ such that $\|\boldsymbol{G} - \boldsymbol{G}_0\|_{\mathsf{F}} \leq \tau$, and

$$\begin{aligned}
\left|\operatorname{med}\left(|\mathcal{A}\left(\boldsymbol{G}_0\right)|\right) - \operatorname{med}\left(|\mathcal{A}\left(\boldsymbol{G}\right)|\right)\right| &\leq \max_{i=1,\cdots,m} \big|\, |\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle| - |\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle| \,\big| && \text{(D.5)}\\
&\leq \max_{i=1,\cdots,m} |\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle - \langle \boldsymbol{A}_i, \boldsymbol{G}\rangle| && \text{(D.6)}\\
&\leq \max_{i=1,\cdots,m} \|\boldsymbol{G}_0 - \boldsymbol{G}\|_{\mathsf{F}} \|\boldsymbol{A}_i\|_{\mathsf{F}}\\
&\leq \tau \max_{i=1,\cdots,m} \|\boldsymbol{A}_i\|_{\mathsf{F}} \leq \epsilon, && \text{(D.7)}
\end{aligned}$$

where (D.5) follows from Lemma 21, and (D.6) follows from the fact $||a| - |b|| \leq |a - b|$.

The rest of the proof is then to argue that (D.7) holds with probability at least $1 - c_1 \exp\left(-c_2 m\epsilon^2\right)$ for some constants c_1 and c_2, as long as $m \geq c_0\left(\epsilon^{-2}\log\epsilon^{-1}\right) nr\log\left(nr\right)$ for some sufficiently large constant c_0. Note that

$$\begin{aligned}
(9/\tau)^{2r(2n+1)} &= \exp\left(2r\left(2n+1\right)\left(\log 18 + \log(\epsilon^{-1}) + \frac{1}{2}\log n + \frac{1}{2}\log\left(n+m\right)\right)\right)\\
&\leq \exp\left(5nr\log m + c_3 nr\log\epsilon^{-1}\right).
\end{aligned}$$

It is straightforward to verify $c_3 nr \log \epsilon^{-1} \leq c_4 m\epsilon^2$, where $2c_4 < c - c_2$, based on the specific setting of m, as long as c_0 is large enough. Then, it suffices to show

$$5nr \log m < c_5 m\epsilon^2, \tag{D.8}$$

where $c_5 < c - c_4 - c_2$, when $m \geq c_0 \left(\epsilon^{-2} \log \epsilon^{-1}\right) nr \log (nr)$ for some large enough constant c_0.

First, for any fixed n, if (D.8) holds for some m and $m \geq (5/c_5)\, \epsilon^{-2} nr$, (D.8) holds for a larger m as well, since

$$5nr \log (m+1) = 5nr \log m + \frac{5nr}{m} \log \left(1 + \frac{1}{m}\right)^m \leq 5nr \log m + 5nr/m \leq c_5 (m+1)\, \epsilon^2.$$

Next, we show that for any fixed n, we can find a constant c_0 such that (D.8) holds as long as $m = c_0 \left(\epsilon^{-2} \log \epsilon^{-1}\right) nr \log (nr)$. Pick a small enough $\epsilon < 1/e$ that is fixed throughout the proof. Given c_5, we can always find a large enough c_0 such that $\frac{1}{3} \log c_0 < c_5 c_0/15 - 5/3$. Then as long as $nr \geq 3$, we can get $\frac{1}{3} \log c_0 < (c_5 c_0/15 - 5/3) \log \epsilon^{-1} \log nr$, which further yields $\frac{1}{3} \log c_0 + \log \epsilon^{-1} + \frac{2}{3} \log nr < (c_5 c_0/15)$ $\cdot \log \epsilon^{-1} \log nr$. As a result, we have

$$\begin{aligned}
(c_5 c_0/5) \log \epsilon^{-1} \log nr &> \log c_0 + 3 \log \epsilon^{-1} + 2 \log nr \\
&= \log \left(c_0 \epsilon^{-3} (nr)^2\right) \\
&> \log \left(c_0 \left(\epsilon^{-2} \log \epsilon^{-1}\right) nr \log (nr)\right),
\end{aligned}$$

which implies (D.8).

D.2 Proof of Proposition 2

We prove the following lemma which directly implies Proposition 2.

Lemma 28. *Under the conditions of Proposition 2, we have*

$$\frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle\cdot\langle \boldsymbol{A}_i, \boldsymbol{T}\rangle\cdot\mathbb{I}_{\left\{|\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle|\le 0.65\alpha_h\|\boldsymbol{G}\|_{\mathsf{F}}\right\}}\mathbb{I}_{\{\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A}_i,\boldsymbol{T}\rangle\ge 0\}} \ge \gamma_1\langle \boldsymbol{G},\boldsymbol{T}\rangle - 0.0011\alpha_h\|\boldsymbol{G}\|_{\mathsf{F}}\|\boldsymbol{T}\|_{\mathsf{F}}$$

holds with high probability for all rank-2r matrices $\boldsymbol{G},\boldsymbol{T}\in\mathbb{R}^{n_1\times n_2}$.

Specializing Lemma 28 to $\boldsymbol{G}=\boldsymbol{U}\boldsymbol{V}^\top-\boldsymbol{X}\boldsymbol{Y}^\top$ and $\boldsymbol{T}=\boldsymbol{H}_1\boldsymbol{V}^\top+\boldsymbol{U}\boldsymbol{H}_2^\top$ yields Proposition 2. The rest of the proof is dedicated to proving Lemma 28. Without loss of generality, we can assume $\|\boldsymbol{G}\|_{\mathsf{F}}=\|\boldsymbol{T}\|_{\mathsf{F}}=1$. Define an auxiliary function as

$$\chi(t)=\begin{cases}1, & |t|<0.65\alpha_h-\delta;\\ \frac{1}{\delta}\left(0.65\alpha_h-|t|\right), & 0.65\alpha_h-\delta\le|t|\le 0.65\alpha_h;\\ 0, & |t|>0.65\alpha_h,\end{cases}$$

where δ is a sufficiently small constant. The function $\chi(t)$ is a Lipschitz function with the Lipschitz constant $1/\delta$. We have

$$\begin{aligned}&\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A}_i,\boldsymbol{T}\rangle\cdot\mathbb{I}_{\{|\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle|\le 0.65\alpha_h-\delta\}}\cdot\mathbb{I}_{\{\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A}_i,\boldsymbol{T}\rangle\ge 0\}}\\ &\le\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A}_i,\boldsymbol{T}\rangle\cdot\chi(\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle)\cdot\mathbb{I}_{\{\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A}_i,\boldsymbol{T}\rangle\ge 0\}}\\ &\le\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A}_i,\boldsymbol{T}\rangle\cdot\mathbb{I}_{\{|\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle|\le 0.65\alpha_h\}}\cdot\mathbb{I}_{\{\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A}_i,\boldsymbol{T}\rangle\ge 0\}}.\end{aligned}\tag{D.9}$$

Let $\zeta_i=\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A}_i,\boldsymbol{T}\rangle\cdot\chi(\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle)\cdot\mathbb{I}_{\{\langle \boldsymbol{A}_i,\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A}_i,\boldsymbol{T}\rangle\ge 0\}}$, $i=1,\cdots,m$, of which each can be considered as an i.i.d. copy of ζ, defined as $\zeta=\langle \boldsymbol{A},\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A},\boldsymbol{T}\rangle\cdot\chi(\langle \boldsymbol{A},\boldsymbol{G}\rangle)\cdot\mathbb{I}_{\{\langle \boldsymbol{A},\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A},\boldsymbol{T}\rangle\ge 0\}}$. From (D.9), we have

$$\begin{aligned}\mathbb{E}\left[\zeta\right]&\ge\mathbb{E}\left[\langle \boldsymbol{A},\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A},\boldsymbol{T}\rangle\cdot\mathbb{I}_{\{|\langle \boldsymbol{A},\boldsymbol{G}\rangle|\le 0.65\alpha_h-\delta\}}\cdot\mathbb{I}_{\{\langle \boldsymbol{A},\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A},\boldsymbol{T}\rangle\ge 0\}}\right]\\ &\ge\mathbb{E}\left[\langle \boldsymbol{A},\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A},\boldsymbol{T}\rangle\cdot\mathbb{I}_{\{|\langle \boldsymbol{A},\boldsymbol{G}\rangle|\le 0.65\alpha_h-\delta\}}\right]\\ &=\left\langle\mathbb{E}\left[\langle \boldsymbol{A},\boldsymbol{G}\rangle\boldsymbol{A}\cdot\mathbb{I}_{\{|\langle \boldsymbol{A},\boldsymbol{G}\rangle|\le 0.65\alpha_h-\delta\}}\right],\boldsymbol{T}\right\rangle=\gamma_1\cdot\langle \boldsymbol{G},\boldsymbol{T}\rangle,\end{aligned}$$

where $\gamma_1=\mathbb{E}\left[\xi^2\mathbb{I}_{\{|\xi|\le 0.65\alpha_h-\delta\}}\right]$ with $\xi\sim\mathcal{N}(0,1)$. Moreover, for $p\ge 0$,

$$\left(\mathbb{E}\left[|\zeta|^p\right]\right)^{1/p}\le\left(\mathbb{E}\left[\left|\langle \boldsymbol{A},\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A},\boldsymbol{T}\rangle\cdot\mathbb{I}_{\{|\langle \boldsymbol{A},\boldsymbol{G}\rangle|\le 0.65\alpha_h\}}\cdot\mathbb{I}_{\{\langle \boldsymbol{A},\boldsymbol{G}\rangle\cdot\langle \boldsymbol{A},\boldsymbol{T}\rangle\ge 0\}}\right|^p\right]\right)^{1/p}$$

$$\leq \left(\mathbb{E}\left[\left|\langle \boldsymbol{A}, \boldsymbol{G}\rangle \cdot \langle \boldsymbol{A}, \boldsymbol{T}\rangle \cdot \mathbb{I}_{\{|\langle \boldsymbol{A}, \boldsymbol{G}\rangle| \leq 0.65\alpha_h\}}\right|^p\right]\right)^{1/p}$$
$$\leq 0.65\alpha_h \left(\mathbb{E}\left[|\langle \boldsymbol{A}, \boldsymbol{T}\rangle|^p\right]\right)^{1/p} \leq 0.65c\alpha_h\sqrt{p},$$

which indicates that ζ is a sub-Gaussian random variable with $\|\zeta\|_{\psi_2} \leq 0.65c\alpha_h$. Then applying the Hoeffding-type inequality [120, Proposition 5.10], we have for any $t \geq 0$,

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\zeta_i - \mathbb{E}\left[\zeta\right]\right| \geq t\right\} \leq \exp\left(-cmt^2/\alpha_h^2\right),$$

for some $c > 0$. Let $t = \varepsilon\alpha_h$, where ε is small enough. Then

$$\frac{1}{m}\sum_{i=1}^{m}\zeta_i \geq \mathbb{E}\left[\zeta\right] - \varepsilon\alpha_h \geq \gamma_1\langle \boldsymbol{G}, \boldsymbol{T}\rangle - \varepsilon\alpha_h \tag{D.10}$$

holds with probability at least $1 - \exp\left(-cm\varepsilon^2\right)$.

Next, a covering argument is needed to extend (D.10) to all rank-$2r$ matrices $(\boldsymbol{G}, \boldsymbol{T})$ with unit Frobenius norm. Let $\mathcal{N}_\tau$ be a τ-net covering all rank-$2r$ matrices with respect to the Frobenius norm, and define

$$\mathcal{M}_\tau = \{(\boldsymbol{G}_0, \boldsymbol{T}_0) : (\boldsymbol{G}_0, \boldsymbol{T}_0) \in \mathcal{N}_\tau \times \mathcal{N}_\tau\}$$

such that for any pair of rank-$2r$ matrices $(\boldsymbol{G}, \boldsymbol{T})$ with $\|\boldsymbol{G}\|_{\mathsf{F}} = \|\boldsymbol{T}\|_{\mathsf{F}} = 1$, there exists $(\boldsymbol{G}_0, \boldsymbol{T}_0) \in \mathcal{M}_\tau$ with $\|\boldsymbol{G}_0\|_{\mathsf{F}} = \|\boldsymbol{T}_0\|_{\mathsf{F}} = 1$ satisfying $\|\boldsymbol{G}_0 - \boldsymbol{G}\|_{\mathsf{F}} \leq \tau$ and $\|\boldsymbol{T}_0 - \boldsymbol{T}\|_{\mathsf{F}} \leq \tau$. Since both $\operatorname{rank}(\boldsymbol{G}) \leq 2r$ and $\operatorname{rank}(\boldsymbol{T}) \leq 2r$, then Lemma 13 guarantees $|\mathcal{M}_\tau| \leq (9/\tau)^{2r(2n+1)} \cdot (9/\tau)^{2r(2n+1)} \leq (9/\tau)^{4r(2n+1)}$. Taking the union bound gives for all $(\boldsymbol{G}_0, \boldsymbol{T}_0) \in \mathcal{M}_\tau$,

$$\frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}_0\rangle \cdot \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle\right) \cdot \mathbb{I}_{\{\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}_0\rangle \geq 0\}} \geq \gamma_1 \cdot \langle \boldsymbol{G}_0, \boldsymbol{T}_0\rangle - \varepsilon\alpha_h$$

with probability at least $1 - (9/\tau)^{4r(2n+1)} \exp\left(-c\varepsilon^2 m\right)$. Furthermore,

$$\Big|\frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}\rangle \cdot \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle\right) \cdot \mathbb{I}_{\{\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}\rangle \geq 0\}}$$

$$
\begin{aligned}
&\quad - \frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}_0\rangle \cdot \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle\right) \cdot \mathbb{I}_{\{\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}_0\rangle \geq 0\}}\Big| \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\Big|\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}\rangle \cdot \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle\right) \cdot \mathbb{I}_{\{\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}\rangle \geq 0\}} \\
&\qquad\qquad - \langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}_0\rangle \cdot \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle\right) \cdot \mathbb{I}_{\{\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}_0\rangle \geq 0\}}\Big| \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\left|\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}\rangle \cdot \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle\right) - \langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}_0\rangle \cdot \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle\right)\right| \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\left|\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle \cdot \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle\right) - \langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle \cdot \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle\right)\right| \cdot \left|\langle \boldsymbol{A}_i, \boldsymbol{T}\rangle\right| \\
&\quad + \frac{1}{m}\sum_{i=1}^{m}\left|\langle \boldsymbol{A}_i, \boldsymbol{T} - \boldsymbol{T}_0\rangle\right| \cdot \left|\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle \cdot \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle\right)\right| \\
&\leq \frac{0.65\alpha_h}{\delta}\Bigg(\frac{1}{m}\sum_{i=1}^{m}\left|\langle \boldsymbol{A}_i, \boldsymbol{G} - \boldsymbol{G}_0\rangle\right| \cdot \left|\langle \boldsymbol{A}_i, \boldsymbol{T}\rangle\right| \\
&\qquad\qquad + \frac{1}{m}\sum_{i=1}^{m}\left|\langle \boldsymbol{A}_i, \boldsymbol{T} - \boldsymbol{T}_0\rangle\right| \cdot \left|\langle \boldsymbol{A}_i, \boldsymbol{G}_0\rangle\right|\Bigg) \qquad \text{(D.11)} \\
&\leq \frac{0.65\alpha_h}{\delta}\Bigg(\frac{1}{\sqrt{m}}\|\mathcal{A}(\boldsymbol{G} - \boldsymbol{G}_0)\|_2 \cdot \frac{1}{\sqrt{m}}\|\mathcal{A}(\boldsymbol{T})\|_2 \\
&\qquad\qquad + \frac{1}{\sqrt{m}}\|\mathcal{A}(\boldsymbol{T} - \boldsymbol{T}_0)\|_2 \cdot \frac{1}{\sqrt{m}}\|\mathcal{A}(\boldsymbol{G}_0)\|_2\Bigg) \qquad \text{(D.12)} \\
&\leq \frac{c_2\alpha_h}{\delta}\left(\|\boldsymbol{G} - \boldsymbol{G}_0\|_{\mathsf{F}}\|\boldsymbol{T}\|_{\mathsf{F}} + \|\boldsymbol{T} - \boldsymbol{T}_0\|_{\mathsf{F}}\|\boldsymbol{G}_0\|_{\mathsf{F}}\right) \qquad \text{(D.13)} \\
&\leq \frac{c_2\alpha_h\tau}{\delta},
\end{aligned}
$$

where (D.11) follows from the Lipschitz property of $t\chi(t)$, (D.12) follows from the Cauchy-Schwarz inequality, and (D.13) follows from Lemma 26.

Let $\tau = c_1\delta\varepsilon$, then provided $m \geq c_2\varepsilon^{-2}\left(\log\frac{1}{\delta\varepsilon}\right)nr$,

$$
\frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}\rangle \cdot \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle\right) \cdot \mathbb{I}_{\{\langle \boldsymbol{A}_i, \boldsymbol{G}\rangle \cdot \langle \boldsymbol{A}_i, \boldsymbol{T}\rangle \geq 0\}} \geq \gamma_1 \cdot \langle \boldsymbol{G}, \boldsymbol{T}\rangle - 1.1\varepsilon\alpha_h
$$

holds for all rank-$2r$ matrices $\boldsymbol{G}$ and $\boldsymbol{T}$ with probability at least $1 - \exp\left(-c\varepsilon^2 m\right)$. The proof is finished by setting δ arbitrarily small and $\varepsilon = 0.001$.

D.3 Proof of Proposition 3

First, note that due to the definition of $\mathcal{D}$ in (4.30), $-B_2$ can be written as

$$\begin{aligned}
-B_2 &= \frac{1}{2m}\sum_{i\in\mathcal{D}} \left|\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\rangle\right| \cdot \left|\langle \boldsymbol{A}_i, \boldsymbol{H}_1\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{H}_2^\top\rangle\right| \\
&\qquad\qquad \cdot \mathbb{I}_{\left\{\left|\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\rangle\right| \le 0.70\alpha_h \left\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}}\right\}} \\
&= \frac{1}{m}\sum_{i\in\mathcal{D}} \left|\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\rangle\right| \cdot \left|\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top\rangle\right| \cdot \mathbb{I}_{\left\{\left|\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\rangle\right| \le 0.70\alpha_h \left\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}}\right\}} \\
&\le 0.70\alpha_h \left\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}} \cdot \frac{1}{m}\sum_{i\in\mathcal{D}} \left|\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top\rangle\right|. \qquad \text{(D.14)}
\end{aligned}$$

Note that when $i \in \mathcal{D}$, we have the following lemma, whose proof is given in Appendix D.7.

Lemma 29. *If $i \in \mathcal{D}$, one has $\left|\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top\rangle\right| < \frac{1}{2}\left|\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{H}^\top\rangle\right|$.*

Plugging Lemma 29 into (D.14), we obtain

$$\begin{aligned}
-B_2 &\le 0.35\alpha_h \left\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}} \cdot \frac{1}{m}\sum_{i\in\mathcal{D}} \left|\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{H}^\top\rangle\right| \\
&\le 0.35\alpha_h \left\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}} \frac{1}{m}\sqrt{m}\left(\sum_{i\in\mathcal{D}} \left|\langle \boldsymbol{A}_i, \boldsymbol{H}_1\boldsymbol{H}_2^\top\rangle\right|^2\right)^{1/2} \qquad \text{(D.15)} \\
&\le 0.35\alpha_h \left\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}} \frac{1}{\sqrt{m}}\|\mathcal{A}(\boldsymbol{H}_1\boldsymbol{H}_2^\top)\|_2 \\
&\le 0.35\left(1+\delta\right)\alpha_h \left\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_{\mathsf{F}} \left\|\boldsymbol{H}_1\boldsymbol{H}_2^\top\right\|_{\mathsf{F}}, \qquad \text{(D.16)}
\end{aligned}$$

where (D.15) follows from the Cauchy-Schwarz inequality and the last inequality follows from Lemma 26.

D.4 Proof of Proposition 4

First, note that by the definitions of $\mathcal{E}_i$ and $\tilde{\mathcal{E}}_i$, we have

$$\left|\left(\mathcal{A}_i\left(\boldsymbol{U}\boldsymbol{V}^\top\right) - y_i\right)\mathbb{I}_{\mathcal{E}_i}\right| \le \alpha_h \mathrm{med}\left(\left|\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^\top\right)\right|\right);$$

$$\left|\left(\mathcal{A}_i\left(\boldsymbol{U}\boldsymbol{V}^\top\right)-\mathcal{A}_i\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)\right)\mathbb{I}_{\tilde{\mathcal{E}}_i}\right|\leq\alpha_h\text{med}\left(\left|\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^\top\right)\right|\right).$$

Then we further obtain

$$\begin{aligned}
&\left|\langle\nabla^o f_{tr}\left(\boldsymbol{W}\right),\boldsymbol{H}\rangle\right| \\
&\leq\frac{1}{m}\sum_{i\in\mathcal{S}}\left|\left[\left(\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right)-y_i\right)\mathbb{I}_{\mathcal{E}_i}-\left(\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right)-\mathcal{B}_i\left(\boldsymbol{Z}\boldsymbol{Z}^\top\right)\right)\mathbb{I}_{\tilde{\mathcal{E}}_i}\right]\langle\boldsymbol{B}_i,\boldsymbol{H}\boldsymbol{W}^\top\rangle\right| \\
&=\frac{1}{m}\sum_{i\in\mathcal{S}}\left|\left[\left(\mathcal{A}_i\left(\boldsymbol{U}\boldsymbol{V}^\top\right)-y_i\right)\mathbb{I}_{\mathcal{E}_i}-\left(\mathcal{A}_i\left(\boldsymbol{U}\boldsymbol{V}^\top\right)-\mathcal{A}_i\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)\right)\mathbb{I}_{\tilde{\mathcal{E}}_i}\right]\langle\boldsymbol{B}_i,\boldsymbol{H}\boldsymbol{W}^\top\rangle\right| \\
&\leq\frac{2\alpha_h}{m}\text{med}\left(\left|\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^\top\right)\right|\right)\sum_{i\in\mathcal{S}}\left|\langle\boldsymbol{B}_i,\boldsymbol{H}\boldsymbol{W}^\top\rangle\right| \\
&\leq\frac{2\alpha_h}{m}\text{med}\left(\left|\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^\top\right)\right|\right)\sqrt{|\mathcal{S}|}\left(\sum_{i\in\mathcal{S}}\left|\langle\boldsymbol{B}_i,\boldsymbol{H}\boldsymbol{W}^\top\rangle\right|^2\right)^{1/2} \qquad \text{(D.17)} \\
&\leq\alpha_h\sqrt{\frac{|\mathcal{S}|}{m}}\text{med}\left(\left|\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^\top\right)\right|\right)\left(\frac{1}{m}\sum_{i=1}^m\left|\langle\boldsymbol{A}_i,\boldsymbol{H}_1\boldsymbol{V}^\top+\boldsymbol{U}\boldsymbol{H}_2^\top\rangle\right|^2\right)^{1/2} \\
&\leq 0.70\alpha_h\sqrt{s}\left\|\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{U}\boldsymbol{V}^\top\right\|_{\mathsf{F}}\cdot(1+\delta)\|\boldsymbol{H}_1\boldsymbol{V}^\top+\boldsymbol{U}\boldsymbol{H}_2^\top\|_{\mathsf{F}} \qquad \text{(D.18)} \\
&\leq 0.71\alpha_h\sqrt{s}\left\|\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{U}\boldsymbol{V}^\top\right\|_{\mathsf{F}}\|\boldsymbol{H}_1\boldsymbol{V}^\top+\boldsymbol{U}\boldsymbol{H}_2^\top\|_{\mathsf{F}},
\end{aligned}$$

where (D.17) follows from the Cauchy-Schwarz inequality, (D.18) follows from (4.19) and Lemma 26, and the last inequality follows by setting δ sufficiently small.

D.5 Proof of Proposition 5

Since $\|\nabla f_{tr}\left(\boldsymbol{W}\right)\|_{\mathsf{F}}^2=\max_{\|\boldsymbol{G}\|_{\mathsf{F}}=1}\left|\langle\nabla f_{tr}\left(\boldsymbol{W}\right),\boldsymbol{G}\rangle\right|^2$, it is sufficient to upper bound $\left|\langle\nabla f_{tr}\left(\boldsymbol{W}\right),\boldsymbol{G}\rangle\right|^2$ for any arbitrary $\boldsymbol{G}=\begin{bmatrix}\boldsymbol{G}_1^\top & \boldsymbol{G}_2^\top\end{bmatrix}^\top\in\mathbb{R}^{(n_1+n_2)\times r}$ with $\boldsymbol{G}_1\in\mathbb{R}^{n_1\times r}$ and $\boldsymbol{G}_2\in\mathbb{R}^{n_2\times r}$ satisfying $\|\boldsymbol{G}\|_{\mathsf{F}}=1$. We have

$$\begin{aligned}
\left|\langle\nabla f_{tr}\left(\boldsymbol{W}\right),\boldsymbol{G}\rangle\right|^2&=\left|\langle\frac{1}{m}\sum_{i=1}^m\left(\mathcal{B}_i\left(\boldsymbol{W}\boldsymbol{W}^\top\right)-y_i\right)\boldsymbol{B}_i\boldsymbol{W}\mathbb{I}_{\mathcal{E}_i},\boldsymbol{G}\rangle\right|^2 \\
&=\left|\left\langle\frac{1}{m}\sum_{i=1}^m\left(\mathcal{A}_i\left(\boldsymbol{U}\boldsymbol{V}^\top\right)-y_i\right)\boldsymbol{B}_i\boldsymbol{W}\mathbb{I}_{\mathcal{E}_i},\boldsymbol{G}\right\rangle\right|^2
\end{aligned}$$

$$
\begin{aligned}
&= \left| \frac{1}{m} \sum_{i=1}^{m} \left(\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top \rangle - y_i \right) \cdot \langle \boldsymbol{B}_i, \boldsymbol{G}\boldsymbol{W}^\top \rangle \cdot \mathbb{I}_{\mathcal{E}_i} \right|^2 \\
&\leq \left(\frac{1}{m} \sum_{i=1}^{m} \left(\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top \rangle - y_i \right)^2 \cdot \mathbb{I}_{\mathcal{E}_i} \right) \\
&\quad \cdot \left(\frac{1}{m} \sum_{i=1}^{m} \left| \langle \boldsymbol{A}_i, \frac{1}{2} \left(\boldsymbol{G}_1 \boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{G}_2^\top \right) \rangle \right|^2 \right), \qquad \text{(D.19)}
\end{aligned}
$$

where (D.19) follows from the Cauchy-Schwarz inequality. Due to (4.20), we have

$$
\begin{aligned}
&\frac{1}{m} \sum_{i=1}^{m} \left(\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top \rangle - y_i \right)^2 \cdot \mathbb{I}_{\mathcal{E}_i} \\
&\leq \frac{1}{m} \sum_{i=1}^{m} \left(\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top \rangle - y_i \right)^2 \cdot \mathbb{I}_{\left\{ |\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{V}^\top \rangle - y_i| \leq 0.70 \alpha_h \left\| \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top \right\|_{\mathsf{F}} \right\}} \\
&\leq 0.70^2 \alpha_h^2 \left\| \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top \right\|_{\mathsf{F}}^2. \qquad \text{(D.20)}
\end{aligned}
$$

From Lemma 26, we have

$$
\begin{aligned}
\frac{1}{m} \sum_{i=1}^{m} \left| \langle \boldsymbol{A}_i, \frac{1}{2} \left(\boldsymbol{G}_1 \boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{G}_2^\top \right) \rangle \right|^2 &\leq \frac{1}{4} (1+\delta)^2 \left\| \boldsymbol{G}_1 \boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{G}_2^\top \right\|_{\mathsf{F}}^2 \\
&\leq \frac{1}{2} (1+\delta)^2 \left(\left\| \boldsymbol{G}_1 \boldsymbol{V}^\top \right\|_{\mathsf{F}}^2 + \left\| \boldsymbol{U}\boldsymbol{G}_2^\top \right\|_{\mathsf{F}}^2 \right) \\
&\leq \frac{1}{2} (1+\delta)^2 \max \left\{ \|\boldsymbol{U}\|^2, \|\boldsymbol{V}\|^2 \right\} \\
&\leq \frac{1}{2} (1+\delta)^2 \|\boldsymbol{W}\|^2. \qquad \text{(D.21)}
\end{aligned}
$$

Plugging (D.20) and (D.21) into (D.19), we have

$$
\left| \langle \nabla f_{tr} (\boldsymbol{W}), \boldsymbol{G} \rangle \right|^2 \leq \frac{1}{2} \cdot 0.70^2 (1+\delta)^2 \alpha_h^2 \left\| \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{X}\boldsymbol{Y}^\top \right\|_{\mathsf{F}}^2 \|\boldsymbol{W}\|^2,
$$

and the proof is completed by setting δ small enough.

D.6 Proof of Proposition 6

First, consider the bound of $\|\boldsymbol{K}_1 - \mathbb{E}[\boldsymbol{K}_1]\|$. Define

$$
\boldsymbol{S}_i = \mathcal{A}_i (\boldsymbol{M}) \boldsymbol{A}_i \mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})| \leq \alpha_y C_M\}} - \gamma_2 \boldsymbol{M}, \quad i \in \mathcal{S}_1^c,
$$

which satisfies $\mathbb{E}[\boldsymbol{S}_i] = \mathbf{0}$, and $\boldsymbol{K}_1 - \mathbb{E}[\boldsymbol{K}_1] = \frac{1}{|\mathcal{S}_1^c|}\sum_{i\in\mathcal{S}_1^c}\boldsymbol{S}_i$.

Based on [120, Proposition 5.34], we know

$$\mathbb{P}\left\{\left|\|\boldsymbol{A}_i\| - \mathbb{E}\left[\|\boldsymbol{A}_i\|\right]\right| > t\right\} \leq 2e^{-t^2/2},$$

which shows $\|\boldsymbol{A}_i\| - \mathbb{E}[\|\boldsymbol{A}_i\|]$ is a sub-Gaussian random variable satisfying $\left\|\|\boldsymbol{A}_i\| - \mathbb{E}[\|\boldsymbol{A}_i\|]\right\|_{\psi_2} \leq c$. Then, we have $\|\boldsymbol{A}_i\|_{\psi_2} \leq \mathbb{E}[\|\boldsymbol{A}_i\|] + c \leq 2\sqrt{n} + c$, where the last inequality follows from the fact $\mathbb{E}[\|\boldsymbol{A}_i\|] \leq 2\sqrt{n}$. As a result, we can calculate

$$\begin{aligned}
\|\boldsymbol{S}_i\|_{\psi_2} &= \left\|\mathcal{A}_i(\boldsymbol{M})\,\boldsymbol{A}_i\mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})|\leq\alpha_y C_M\}} - \gamma_2\boldsymbol{M}\right\|_{\psi_2} \\
&\leq \left\|\mathcal{A}_i(\boldsymbol{M})\,\boldsymbol{A}_i\mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})|\leq\alpha_y C_M\}}\right\|_{\psi_2} + \gamma_2\|\boldsymbol{M}\| \\
&\leq \alpha_y C_M\|\boldsymbol{A}_i\|_{\psi_2} + \gamma_2\|\boldsymbol{M}\| \leq c_1\sqrt{n}\alpha_y\|\boldsymbol{M}\|_{\mathsf{F}},
\end{aligned}$$

where c_1 is some constant. Moreover, we have

$$\begin{aligned}
&\sigma^2_{\boldsymbol{S}_i} \\
&:= \max\left\{\left\|\frac{1}{|\mathcal{S}_1^c|}\sum_{i\in\mathcal{S}_1^c}\mathbb{E}\left[\boldsymbol{S}_i\boldsymbol{S}_i^\top\right]\right\|, \left\|\frac{1}{|\mathcal{S}_1^c|}\sum_{i\in\mathcal{S}_1^c}\mathbb{E}\left[\boldsymbol{S}_i^\top\boldsymbol{S}_i\right]\right\|\right\} \\
&= \max\left\{\left\|\mathbb{E}\left[\left(\mathcal{A}_i(\boldsymbol{M})\,\boldsymbol{A}_i\mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})|\leq\alpha_y C_M\}} - \gamma_2\boldsymbol{M}\right)\left(\mathcal{A}_i(\boldsymbol{M})\,\boldsymbol{A}_i\mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})|\leq\alpha_y C_M\}} - \gamma_2\boldsymbol{M}\right)^\top\right]\right\|,\right. \\
&\qquad \left.\left\|\mathbb{E}\left[\left(\mathcal{A}_i(\boldsymbol{M})\,\boldsymbol{A}_i\mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})|\leq\alpha_y C_M\}} - \gamma_2\boldsymbol{M}\right)^\top\left(\mathcal{A}_i(\boldsymbol{M})\,\boldsymbol{A}_i\mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})|\leq\alpha_y C_M\}} - \gamma_2\boldsymbol{M}\right)\right]\right\|\right\} \\
&= \max\left\{\left\|\mathbb{E}\left[\left(\mathcal{A}_i(\boldsymbol{M})\right)^2\boldsymbol{A}_i\boldsymbol{A}_i^\top\mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})|\leq\alpha_y C_M\}}\right] - \gamma_2^2\boldsymbol{M}\boldsymbol{M}^\top\right\|,\right. \\
&\qquad \left.\left\|\mathbb{E}\left[\left(\mathcal{A}_i(\boldsymbol{M})\right)^2\boldsymbol{A}_i^\top\boldsymbol{A}_i\mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})|\leq\alpha_y C_M\}}\right] - \gamma_2^2\boldsymbol{M}^\top\boldsymbol{M}\right\|\right\} \\
&\leq \max\left\{\left\|\mathbb{E}\left[\left(\mathcal{A}_i(\boldsymbol{M})\right)^2\boldsymbol{A}_i\boldsymbol{A}_i^\top\mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})|\leq\alpha_y C_M\}}\right]\right\| + \gamma_2^2\|\boldsymbol{M}\|^2,\right. \\
&\qquad \left.\left\|\mathbb{E}\left[\left(\mathcal{A}_i(\boldsymbol{M})\right)^2\boldsymbol{A}_i^\top\boldsymbol{A}_i\mathbb{I}_{\{|\mathcal{A}_i(\boldsymbol{M})|\leq\alpha_y C_M\}}\right]\right\| + \gamma_2^2\|\boldsymbol{M}\|^2\right\} \\
&\leq \alpha_y^2 C_M^2\max\left\{\left\|\mathbb{E}\left[\boldsymbol{A}_i\boldsymbol{A}_i^\top\right]\right\|, \left\|\mathbb{E}\left[\boldsymbol{A}_i^\top\boldsymbol{A}_i\right]\right\|\right\} + \gamma_2^2\|\boldsymbol{M}\|^2 \\
&\leq c_2 n\alpha_y^2\|\boldsymbol{M}\|_{\mathsf{F}}^2,
\end{aligned}$$

where c_2 is some constant. By Lemma 24, we have

$$\left\| \frac{1}{|\mathcal{S}_1^c|} \sum_{i \in \mathcal{S}_1^c} \boldsymbol{S}_i \right\| \leq C\sqrt{n}\alpha_y \left\|\boldsymbol{M}\right\|_{\mathsf{F}} \max\left\{ \sqrt{\frac{t+\log(2n)}{|\mathcal{S}_1^c|}}, \frac{t+\log(2n)}{|\mathcal{S}_1^c|} \right\},$$

with probability at least $1-e^{-t}$, where C is some constant. Set $t = c\log n$. As long as $|\mathcal{S}_1^c| = (1-s_1)m/2 \geq c'\log n$, we have

$$\|\boldsymbol{K}_1 - \mathbb{E}\left[\boldsymbol{K}_1\right]\| \leq C\alpha_y \|\boldsymbol{M}\|_{\mathsf{F}} \sqrt{\frac{n\log n}{m}} \tag{D.22}$$

holds with probability at least $1-n^{-c}$ for some $c > 1$.

Next, we employ the same technique to bound $\|\boldsymbol{K}_2 - \mathbb{E}\left[\boldsymbol{K}_2\right]\|$. Define $\boldsymbol{T}_i = y_i \boldsymbol{A}_i \mathbb{I}_{\{|y_i| \leq \alpha_y C_M\}}$, which satisfies $\mathbb{E}[\boldsymbol{T}_i] = \boldsymbol{0}$ and $\boldsymbol{K}_2 - \mathbb{E}\left[\boldsymbol{K}_2\right] = \frac{1}{|\mathcal{S}_1|}\sum_{i\in\mathcal{S}_1} \boldsymbol{T}_i$. We have

$$\left\|\boldsymbol{T}_i\right\|_{\psi_2} = \left\| y_i \boldsymbol{A}_i \mathbb{I}_{\{|y_i| \leq \alpha_y C_M\}} \right\|_{\psi_2} \leq \alpha_y C_M \left\|\boldsymbol{A}_i\right\|_{\psi_2} \leq c_1\sqrt{n}\alpha_y \left\|\boldsymbol{M}\right\|_{\mathsf{F}},$$

where c_1 is some constant, and

$$\begin{aligned}
\sigma_{\boldsymbol{T}_i}^2 &:= \max\left\{ \left\| \frac{1}{|\mathcal{S}_1|} \sum_{i\in\mathcal{S}_1} \mathbb{E}\left[\boldsymbol{T}_i \boldsymbol{T}_i^\top\right] \right\|, \left\| \frac{1}{|\mathcal{S}_1|} \sum_{i\in\mathcal{S}_1} \mathbb{E}\left[\boldsymbol{T}_i^\top \boldsymbol{T}_i\right] \right\| \right\} \\
&\leq \alpha_y^2 C_M^2 \max\left\{ \left\|\mathbb{E}\left[\boldsymbol{A}_i \boldsymbol{A}_i^\top\right]\right\|, \left\|\mathbb{E}\left[\boldsymbol{A}_i^\top \boldsymbol{A}_i\right]\right\| \right\} \leq c_2 n \alpha_y^2 \left\|\boldsymbol{M}\right\|_{\mathsf{F}}^2,
\end{aligned}$$

where c_2 is some constant. Again, by Lemma 24 we have

$$\left\| \frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} \boldsymbol{T}_i \right\| \leq C\sqrt{n}\alpha_y \left\|\boldsymbol{M}\right\|_{\mathsf{F}} \max\left\{ \sqrt{\frac{t+\log(2n)}{|\mathcal{S}_1|}}, \frac{t+\log(2n)}{|\mathcal{S}_1|} \right\}$$

with probability at least $1-e^{-t}$. Then by setting $t = c\log n$, and recalling $|\mathcal{S}_1| = s_1 m/2$, we have with probability at least $1-n^{-c}$,

$$\|\boldsymbol{K}_2\| \leq C\sqrt{n}\alpha_y \left\|\boldsymbol{M}\right\|_{\mathsf{F}} \max\left\{ \sqrt{\frac{\log n}{s_1 m}}, \frac{\log n}{s_1 m} \right\}. \tag{D.23}$$

Combing (D.22) and (D.23), we have with probability at least $1-n^{-c}$,

$$\|\boldsymbol{K} - (1-s_1)\gamma_2 \boldsymbol{M}\|$$

$$
\begin{aligned}
&\leq (1-s_1)\left\|\boldsymbol{K}_1-\gamma_2\boldsymbol{M}\right\|+s_1\left\|\boldsymbol{K}_2\right\| \\
&\leq C\alpha_y\|\boldsymbol{M}\|_{\mathsf{F}}\sqrt{\frac{n\log n}{m}}+C\sqrt{n}\alpha_y\left\|\boldsymbol{M}\right\|_{\mathsf{F}}\max\left\{\sqrt{\frac{s_1\log n}{m}},\frac{\log n}{m}\right\} \\
&\leq C\alpha_y\|\boldsymbol{M}\|_{\mathsf{F}}\sqrt{\frac{n\log n}{m}},
\end{aligned}
$$

provided that $m > c_2 \log n$ for large enough c_2.

D.7 Proof of Lemma 29

Since $\boldsymbol{H} = \boldsymbol{W} - \boldsymbol{Z}\boldsymbol{Q}$, we can write

$$
\begin{aligned}
\langle \boldsymbol{B}_i, \boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{Z}\boldsymbol{Z}^\top\rangle &= \langle \boldsymbol{B}_i, \boldsymbol{W}\boldsymbol{W}^\top - (\boldsymbol{Z}\boldsymbol{Q})(\boldsymbol{Z}\boldsymbol{Q})^\top\rangle \\
&= \langle \boldsymbol{B}_i, \boldsymbol{W}\boldsymbol{W}^\top - (\boldsymbol{W}-\boldsymbol{H})(\boldsymbol{W}-\boldsymbol{H})^\top\rangle \\
&= 2\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top\rangle - \langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{H}^\top\rangle.
\end{aligned}
$$

Therefore, $i \in \mathcal{D}$ if and only if

$$
\left(2\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top\rangle - \langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{H}^\top\rangle\right)\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top\rangle < 0. \tag{D.24}
$$

If $\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top\rangle > 0$, we know $\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top\rangle < \frac{1}{2}\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{H}^\top\rangle$; if $\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top\rangle < 0$, we know $\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top\rangle > \frac{1}{2}\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{H}^\top\rangle$. Therefore, we have $|\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{W}^\top\rangle| < \frac{1}{2}|\langle \boldsymbol{B}_i, \boldsymbol{H}\boldsymbol{H}^\top\rangle|$.

Bibliography

[1] J. Bennett and S. Lanning, "The Netflix Prize," in *Proceedings of KDD cup and workshop*, vol. 2007. New York, NY, USA, 2007, p. 35.

[2] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[3] J. D. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 713–719.

[4] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[5] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[6] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[7] S. Negahban and M. J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, pp. 1069–1097, 2011.

[8] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Communications of the ACM*, vol. 55, no. 6, pp. 111–119, 2012.

[9] Y. Chen and Y. Chi, "Robust spectral compressed sensing via structured matrix completion," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6576–6601, 2014.

[10] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4034–4059, 2015.

[11] M. A. Davenport and J. Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 608–622, 2016.

[12] Y. Chen and Y. Chi, "Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization," *IEEE Signal Processing Magazine*, vol. 35, no. 4, 2018.

[13] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.

[14] E. J. Candès and X. Li, "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," *Foundations of Computational Mathematics*, vol. 14, no. 5, pp. 1017–1026, 2014.

[15] P. Jain, R. Meka, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Advances in Neural Information Processing Systems*, 2010, pp. 937–945.

[16] E. J. Candès and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[17] R. Kueng, H. Rauhut, and U. Terstiege, "Low rank matrix recovery from rank one measurements," *Applied and Computational Harmonic Analysis*, vol. 42, no. 1, pp. 88–116, 2017.

[18] T. T. Cai and A. Zhang, "ROP: Matrix recovery via rank-one projections," *The Annals of Statistics*, vol. 43, no. 1, pp. 102–138, 2015.

[19] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[20] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[21] Y. Chen, "Incoherence-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2909–2923, 2015.

[22] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[23] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[24] S. Burer and R. D. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Mathematical Programming*, vol. 95, no. 2, pp. 329–357, 2003.

[25] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[26] Q. Zheng and J. Lafferty, "A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements," in *Advances in Neural Information Processing Systems*, 2015, pp. 109–117.

[27] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, 2013, pp. 665–674.

[28] J. Tanner and K. Wei, "Low rank matrix completion by alternating steepest descent methods," *Applied and Computational Harmonic Analysis*, vol. 40, no. 2, pp. 417–429, 2016.

[29] P. Jain and P. Netrapalli, "Fast exact matrix completion with finite samples," in *Conference on Learning Theory*, 2015, pp. 1007–1034.

[30] C. Jin, S. M. Kakade, and P. Netrapalli, "Provable efficient online matrix completion via non-convex stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2016, pp. 4520–4528.

[31] F. De la Torre and M. J. Black, "Robust principal component analysis for computer vision," in *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV)*, vol. 1. IEEE, 2001, pp. 362–369.

[32] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma, "Robust photometric stereo via low-rank matrix completion and recovery," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 703–717.

[33] D. S. Weller, A. Pnueli, G. Divon, O. Radzyner, Y. C. Eldar, and J. A. Fessler, "Undersampled phase retrieval with outliers," *IEEE transactions on computational imaging*, vol. 1, no. 4, pp. 247–258, 2015.

[34] Y. Li, C. Ma, Y. Chen, and Y. Chi, "Nonconvex matrix factorization from rank-one measurements," *arXiv preprint arXiv:1802.06286*, 2018.

[35] L. L. Scharf, "The SVD and reduced rank signal processing," *Signal processing*, vol. 25, no. 2, pp. 113–133, 1991.

[36] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," in *ACM SIGMETRICS Performance evaluation review*, vol. 32, no. 1. ACM, 2004, pp. 61–72.

[37] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Physical review letters*, vol. 105, no. 15, p. 150401, 2010.

[38] D. D. Arianandа and G. Leus, "Compressive wideband power spectrum estimation," *IEEE Transactions on signal processing*, vol. 60, no. 9, pp. 4775–4789, 2012.

[39] H. Kim, A. M. Haimovich, and Y. C. Eldar, "Non-coherent direction of arrival estimation from magnitude-only measurements," *Signal Processing Letters, IEEE*, vol. 22, no. 7, pp. 925–929, 2015.

[40] E. Mason, I.-Y. Son, and B. Yazıcı, "Passive synthetic aperture radar imaging using low-rank matrix recovery methods," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 8, pp. 1570–1582, 2015.

[41] N. E. Hurt, *Phase retrieval and zero crossings: mathematical methods in image reconstruction*. Springer Science & Business Media, 2001, vol. 52.

[42] J. Miao, T. Ishikawa, Q. Shen, and T. Earnest, "Extending X-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes," *Annu. Rev. Phys. Chem.*, vol. 59, pp. 387–410, 2008.

[43] Y. Chen, X. Yi, and C. Caramanis, "A convex formulation for mixed regression with two components: Minimax optimal rates," in *Conference on Learning Theory*, 2014, pp. 560–604.

[44] L. Tian, J. Lee, S. B. Oh, and G. Barbastathis, "Experimental compressive phase space tomography," *Optics express*, vol. 20, no. 8, pp. 8296–8308, 2012.

[45] R. Livni, S. Shalev-Shwartz, and O. Shamir, "On the computational efficiency of training neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 855–863.

[46] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *arXiv preprint arXiv:1707.04926*, 2017.

[47] M. Soltani and C. Hegde, "Towards provable learning of polynomial neural networks using low-rank matrix estimation," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 1417–1426.

[48] S. Sanghavi, R. Ward, and C. D. White, "The local convexity of solving systems of quadratic equations," *Results in Mathematics*, vol. 71, no. 3-4, pp. 569–608, 2017.

[49] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution," *arXiv preprint arXiv:1711.10467*, 2017.

[50] K. Zhong, P. Jain, and I. S. Dhillon, "Efficient matrix sensing using rank-1 Gaussian measurements," in *International Conference on Algorithmic Learning Theory*. Springer, 2015, pp. 3–18.

[51] M. Lin and J. Ye, "A non-convex one-pass framework for generalized factorization machine and rank-one matrix sensing," in *Advances in Neural Information Processing Systems*, 2016, pp. 1633–1641.

[52] Y. Chen and E. J. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Communications on Pure and Applied Mathematics*, vol. 70, no. 5, pp. 822–883, 2017.

[53] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," *arXiv preprint arXiv:1509.03025*, 2015.

[54] Y. Chen, J. Fan, C. Ma, and K. Wang, "Spectral method and regularized MLE are both optimal for top-K ranking," *arXiv preprint arXiv:1707.09971*, 2017.

[55] Y. Zhong and N. Boumal, "Near-optimal bounds for phase synchronization," *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 989–1016, 2018.

[56] L. Demanet and P. Hand, "Stable optimizationless recovery from phaseless linear measurements," *Journal of Fourier Analysis and Applications*, vol. 20, no. 1, pp. 199–221, 2014.

[57] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, MaxCut and complex semidefinite programming," *Mathematical Programming*, vol. 149, no. 1-2, pp. 47–81, 2015.

[58] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging: a contemporary overview," *Signal Processing Magazine, IEEE*, vol. 32, no. 3, pp. 87–109, 2015.

[59] T. T. Cai, X. Li, and Z. Ma, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow," *The Annals of Statistics*, vol. 44, no. 5, pp. 2221–2251, 2016.

[60] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, "A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms," *Journal of Machine Learning Research*, vol. 18, no. 141, pp. 1–35, 2017.

[61] M. Soltanolkotabi, "Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization," *arXiv preprint arXiv:1702.06175*, 2017.

[62] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 773–794, 2018.

[63] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via Procrustes flow," in *International Conference on Machine Learning*, 2016, pp. 964–973.

[64] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Applied and Computational Harmonic Analysis*, 2018. [Online]. Available: https://doi.org/10.1016/j.acha.2018.01.001

[65] J.-F. Cai, T. Wang, and K. Wei, "Spectral compressed sensing via projected gradient descent," *arXiv preprint arXiv:1707.09726*, 2017.

[66] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[67] K. Wei, "Solving systems of phaseless equations via Kaczmarz methods: A proof of concept study," *Inverse Problems*, vol. 31, no. 12, p. 125008, 2015.

[68] Y. S. Tan and R. Vershynin, "Phase retrieval via randomized Kaczmarz: theoretical guarantees," *Information and Inference: A Journal of the IMA*, p. iay005, 2018. [Online]. Available: http://dx.doi.org/10.1093/imaiai/iay005

[69] H. Jeong and C. S. Güntürk, "Convergence of the randomized Kaczmarz method for phase retrieval," *arXiv preprint arXiv:1706.10291*, 2017.

[70] J. Ma, J. Xu, and A. Maleki, "Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2-regularization," *arXiv preprint arXiv:1801.01170*, 2018.

[71] Y. Chi and Y. M. Lu, "Kaczmarz method for solving quadratic equations," *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1183–1187, 2016.

[72] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 2379–2383.

[73] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon, "Recovery guarantees for one-hidden-layer neural networks," in *International Conference on Machine Learning*, 2017, pp. 4140–4149.

[74] Y. Li, Y. Sun, and Y. Chi, "Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 397–408, 2017.

[75] Y. Sun, Y. Li, and Y. Chi, "Outlier-robust recovery of low-rank positive semidefinite matrices from magnitude measurements," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4069–4073.

[76] P. Hand, "PhaseLift is robust to a constant fraction of arbitrary errors," *Applied and Computational Harmonic Analysis*, vol. 42, no. 3, pp. 550–562, 2017.

[77] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM review*, vol. 57, no. 2, pp. 225–251, 2015.

[78] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege, "Stable low-rank matrix recovery via null space properties," *Information and Inference: A Journal of the IMA*, vol. 5, no. 4, pp. 405–441, 2016.

[79] W. Dai, O. Milenkovic, and E. Kerman, "Subspace evolution and transfer (SET) for low-rank matrix completion," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3120–3132, 2011.

[80] M. Wang, W. Xu, and A. Tang, "A unique "nonnegative" solution to an underdetermined system: From vectors to matrices," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 1007–1016, 2011.

[81] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[82] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[83] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive principal component pursuit," *Information and Inference: A Journal of the IMA*, vol. 2, no. 1, pp. 32–68, 2013.

[84] X. Li, "Compressed sensing and matrix completion with constant proportion of corruptions," *Constructive Approximation*, vol. 37, pp. 73–99, 2013.

[85] G. Mateos and G. B. Giannakis, "Robust PCA as bilinear decomposition with outlier-sparsity regularization," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5176–5190, 2012.

[86] Y. I. Abramovich, D. A. Gray, A. Y. Gorokhov, and N. K. Spencer, "Positive-definite Toeplitz completion in DOA estimation for nonuniform linear antenna arrays. I. Fully augmentable arrays," *IEEE Transactions on Signal Processing*, vol. 46, no. 9, pp. 2458–2471, 1998.

[87] H. Qiao and P. Pal, "Generalized nested sampling for compressing low rank toeplitz matrices," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1844–1848, 2015.

[88] D. Romero, D. D. Ariananda, Z. Tian, and G. Leus, "Compressive covariance sensing: Structure-based compressive sensing beyond sparsity," *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 78–93, 2016.

[89] D. Romero, R. López-Valcarce, and G. Leus, "Compression limits for random vectors with linearly parameterized second-order statistics," *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1410–1425, 2015.

[90] Y. Li, Y. Chi, H. Zhang, and Y. Liang, "Non-convex low-rank matrix recovery from corrupted random linear measurements," in *Sampling Theory and Applications (SampTA), 2017 International Conference on*. IEEE, 2017, pp. 134–137.

[91] ——, "Nonconvex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent," *arXiv preprint arXiv:1709.08114*, 2017.

[92] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis, "Low-rank matrix recovery from errors and erasures," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4324–4337, 2013.

[93] P. J. Huber, *Robust Statistics*. Springer, 2011.

[94] R. J. Tibshirani, "Fast computation of the median by successive binning," *arXiv preprint arXiv:0806.3301*, 2008.

[95] T. Zhao, Z. Wang, and H. Liu, "A nonconvex optimization framework for low rank matrix estimation," in *Advances in Neural Information Processing Systems*, 2015, pp. 559–567.

[96] Y. Chen and E. J. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in *Advances in Neural Information Processing Systems*, 2015, pp. 739–747.

[97] H. Zhang, Y. Chi, and Y. Liang, "Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow," *arXiv preprint arXiv:1603.03805*, 2016.

[98] D. Park, A. Kyrillidis, S. Bhojanapalli, C. Caramanis, and S. Sanghavi, "Provable Burer-Monteiro factorization for a class of norm-constrained matrix problems," *arXiv preprint arXiv:1606.01316*, 2016.

[99] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[100] M. Hardt, "Understanding alternating minimization for matrix completion," in *IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, 2014, pp. 651–660.

[101] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, 2016, pp. 3873–3881.

[102] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[103] Q. Li and G. Tang, "The nonconvex geometry of low-rank matrix optimizations with general objective functions," *arXiv preprint arXiv:1611.03060*, 2016.

[104] X. Li, Z. Wang, J. Lu, R. Arora, J. Haupt, H. Liu, and T. Zhao, "Symmetry, saddle points, and global geometry of nonconvex matrix factorization," *arXiv preprint arXiv:1612.09296*, 2016.

[105] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points–online stochastic gradient for tensor decomposition," in *Conference on Learning Theory*, 2015, pp. 797–842.

[106] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Conference on Learning Theory*, 2016, pp. 1246–1257.

[107] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *International Conference on Machine Learning*, 2017, pp. 1724–1732.

[108] K. Chen, "On k-median clustering in high dimensions," in *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*. Society for Industrial and Applied Mathematics, 2006, pp. 1177–1185.

[109] D. Wagner, "Resilient aggregation in sensor networks," in *Proceedings of the 2nd ACM workshop on Security of ad hoc and sensor networks*. ACM, 2004, pp. 78–87.

[110] Y. Chen, C. Caramanis, and S. Mannor, "Robust sparse regression under adversarial corruption," in *International Conference on Machine Learning*, 2013, pp. 774–782.

[111] C. Qu and H. Xu, "Subspace clustering with irrelevant features via robust dantzig selector," in *Advances in Neural Information Processing Systems*, 2015, pp. 757–765.

[112] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar, "Robust estimation via robust gradient estimation," *arXiv preprint arXiv:1802.06485*, 2018.

[113] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," *arXiv preprint arXiv:1803.01498*, 2018.

[114] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, p. 44, 2017.

[115] X. Yi, D. Park, Y. Chen, and C. Caramanis, "Fast algorithms for robust PCA via gradient descent," in *Advances in neural information processing systems*, 2016, pp. 4152–4160.

[116] Y. Cherapanamjeri, K. Gupta, and P. Jain, "Nearly optimal robust matrix completion," in *International Conference on Machine Learning*, 2017, pp. 797–805.

[117] X. Zhang, L. Wang, and Q. Gu, "A unified framework for low-rank plus sparse matrix recovery," *arXiv preprint arXiv:1702.06525*, 2017.

[118] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *International Conference on Machine Learning*, 2016, pp. 1225–1234.

[119] V. Bentkus, "An inequality for tail probabilities of martingales with differences bounded from one side," *Journal of Theoretical Probability*, vol. 16, no. 1, pp. 161–173, 2003.

[120] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *Compressed Sensing, Theory and Applications*, pp. 210 – 268, 2012.

[121] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, pp. 1302–1338, 2000.

[122] W. Schudy and M. Sviridenko, "Concentration and moment inequalities for polynomials of independent random variables," in *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2012, pp. 437–446.

[123] Q. Zheng and J. Lafferty, "Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent," *arXiv preprint arXiv:1605.07051*, 2016.

[124] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, pp. 2302–2329, 2011.

[125] J. M. Ten Berge, "Orthogonal procrustes rotation for two or more matrices," *Psychometrika*, vol. 42, no. 2, pp. 267–276, 1977.

[126] S. Lang, "Real and functional analysis, volume 142 of graduate texts in mathematics," *Springer-Verlag, New York,*, vol. 10, pp. 11–13, 1993.

[127] Y. Yu, T. Wang, and R. J. Samworth, "A useful variant of the davis–kahan theorem for statisticians," *Biometrika*, vol. 102, no. 2, pp. 315–323, 2014.

[128] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. iii," *SIAM Journal on Numerical Analysis*, vol. 7, no. 1, pp. 1–46, 1970.