Nonconvex Matrix Factorization from Rank-One Measurements

Yuanxin Li* Cong Ma† Yuxin Chen‡ Yuejie Chi§

February 2018; Revised: March 2020

Abstract

We consider the problem of recovering low-rank 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 bounded by a constant, this algorithm is guaranteed to converge to the ground truth (up to global ambiguity) with near-optimal sample complexity and computational complexity. To the best of our knowledge, this is the first guarantee that achieves near-optimality in both metrics. In particular, the key enabler of near-optimal computational guarantees is an implicit regularization phenomenon: without explicit regularization, both spectral initialization and the gradient descent iterates automatically stay within a region incoherent with the measurement vectors. This feature allows one to employ much more aggressive step sizes compared with the ones suggested in prior literature, without the need of sample splitting.

Keywords: matrix factorization, rank-one measurements, gradient descent, nonconvex optimization

1 Introduction

This paper is concerned with estimating a low-rank positive semidefinite matrix $\boldsymbol{M}^\natural \in \mathbb{R}^{n\times n}$ from a few *rank-one measurements*. Specifically, suppose that the matrix of interest can be factorized as

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

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

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

where $\{\boldsymbol{a}_i \in \mathbb{R}^n\}_{i=1}^m$ represent the measurement vectors known *a priori*. For instance, we will work with the Gaussian design model (namely, $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$) unless otherwise noted. 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 goal is to recover $\boldsymbol{M}^\natural$, or equivalently, the low-rank factor $\boldsymbol{X}^\natural$, from a limited number of rank-one measurements. This problem spans a variety of important practical applications, 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

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

*Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA; Email: yuanxinl@andrew.cmu.edu

†Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; Email: congm@princeton.edu

‡Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA; Email: yuxin.chen@princeton.edu

§Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA; Email: yuejiechi@cmu.edu

[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*.

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 minimal storage requirement and low sketching cost. See [1] for detailed descriptions.

- **Phase retrieval and mixed linear regression.** This problem subsumes as a special case the phase retrieval problem [2], 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 [3] for precise formulation.

- **Quantum state tomography.** Estimating the density operator of a quantum system can be formulated as a low-rank positive semidefinite matrix recovery problem using rank-one measurements, when the density operator is *almost pure* [4]. A problem of similar mathematical formulation occurs in phase space tomography [5], 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 [6, 7, 8], 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 $\sigma(z) = z^2$.

1.1 Main Contributions

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 the following optimization problem:

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

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. Our main finding is the following: under i.i.d. Gaussian design (i.e. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$), *vanilla gradient descent* combined with spectral initialization achieves appealing performance guarantees both statistically and computationally.

- Statistically, we show that gradient descent converges exactly to the true factor $\boldsymbol{X}^\natural$ (modulo unrecoverable global ambiguity), as soon as the number of measurements exceeds the order of $O(nr^4\kappa^{7/2}\log n)$. When r is fixed independent of n, this sample complexity is near-optimal up to some logarithmic factor with respect to n and r.

- Computationally, to achieve ϵ-accuracy, gradient descent requires an iteration complexity of $O(\kappa^2 r^2 \log(1/\epsilon))$ (up to logarithmic factors), with a per-iteration cost of $O(mnr)$. When r is fixed independent of m and n, the computational complexity scales linearly with mn, which is proportional to the time taken to read all data.

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 [9, 10, 8]), or high iteration complexity (which results in high computation cost [11]). In particular, our work is most related to [11] that also studied the effectiveness of gradient descent. The results in [11] require a sample complexity on the order of $nr^6\log^2 n$, as well as an iteration complexity of $O(n^4 r^2 \log(1/\epsilon))$ (up to logarithmic factors) to attain ϵ-accuracy. In comparison, our theory improves the sample complexity to $O(nr^4\log n)$ and, perhaps more importantly, establishes a much lower iteration complexity of $O(r^2\log(1/\epsilon))$ (up to logarithmic factor). To the best of our knowledge, this work is the first nonconvex method (without resampling) that achieves both near-optimal statistical and computational guarantees with respect to n.

1.2 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 (1), [11] showed that within a local neighborhood of $\boldsymbol{X}^\natural$, where $\boldsymbol{X}$ satisfies

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

$f(\boldsymbol{X})$ 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 [11] 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 [12] and projection [13]. 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 introduce more tuning parameters.

Our work is inspired by [14], 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 paper demonstrates that a similar phenomenon persists in low-rank matrix factorization 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"[2] with all sensing vectors in the following sense:

$$\max_{1\leq l\leq m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}-\boldsymbol{X}^\natural\right)\right\|_2 \leq \frac{1}{24}\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}. \tag{3}$$

We term the intersection of (2) and (3) the *Region of Incoherence and Contraction* (RIC). The nice feature of the RIC is this: within this region, the loss function $f(\boldsymbol{X})$ 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 [11]). 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. To address this issue, we resort to the leave-one-out trick [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26] 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 w.r.t. 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.

1.3 Notations

We use boldface lowercase (resp. uppercase) letters to represent vectors (resp. matrices). We denote by $\|\boldsymbol{x}\|_2$ the ℓ_2 norm of a vector $\boldsymbol{x}$, and $\boldsymbol{X}^\top$, $\|\boldsymbol{X}\|$ and $\|\boldsymbol{X}\|_{\mathsf{F}}$ the transpose, the spectral norm and the Frobenius norm of a matrix $\boldsymbol{X}$, respectively. The kth largest singular value of a matrix $\boldsymbol{X}$ is denoted by $\sigma_k\left(\boldsymbol{X}\right)$. 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 $\mathrm{vec}(\boldsymbol{V})$ to denote vectorization of a matrix $\boldsymbol{V}$. The notation $f(n)\lesssim g(n)$ or $f(n)=O(g(n))$ means that there exists a universal constant $c>0$ such that $|f(n)|\leq c|g(n)|$. In addition, we use c and C with different subscripts to represent positive numerical constants, whose values may change from line to line.

[2]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 (3).

2 Algorithms and Main Results

To begin with, we present the formal problem setup. Suppose we are given a set of m rank-one measurements as follows

$$y_i = \left\| \boldsymbol{a}_i^\top \boldsymbol{X}^\natural \right\|_2^2, \qquad i = 1, \cdots, m, \tag{4}$$

where $\boldsymbol{a}_i \in \mathbb{R}^n$ 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 $i = 1, \cdots, m$. The underlying ground truth $\boldsymbol{X}^\natural \in \mathbb{R}^{n \times r}$ 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

$$\kappa = \frac{\sigma_1^2\left(\boldsymbol{X}^\natural\right)}{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}. \tag{5}$$

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.

2.1 Vanilla Gradient Descent

The algorithm studied herein is a combination of vanilla gradient descent and a judiciously designed spectral initialization. Specifically, consider minimizing the squared loss:

$$f(\boldsymbol{X}) := \frac{1}{4m} \sum_{i=1}^m \left(y_i - \left\| \boldsymbol{a}_i^\top \boldsymbol{X} \right\|_2^2 \right)^2, \tag{6}$$

which is a nonconvex function. We attempt to optimize this function iteratively via gradient descent

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \mu_t \nabla f(\boldsymbol{X}_t), \qquad t = 0, 1, \cdots, \tag{7}$$

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

$$\nabla f(\boldsymbol{X}) = \frac{1}{m} \sum_{i=1}^m \left(\left\| \boldsymbol{a}_i^\top \boldsymbol{X} \right\|_2^2 - y_i \right) \boldsymbol{a}_i \boldsymbol{a}_i^\top \boldsymbol{X}. \tag{8}$$

For initialization, similar to [11],[3] 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 (9). The rationale is this: the mean of $\boldsymbol{Y}$ is given by

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

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.

2.2 Performance Guarantees

Before proceeding to our main results, we specify 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

$$\operatorname{dist}\left(\boldsymbol{X}_t, \boldsymbol{X}^\natural\right) = \left\| \boldsymbol{X}_t \boldsymbol{Q}_t - \boldsymbol{X}^\natural \right\|_\mathsf{F}, \tag{12}$$

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

$$\boldsymbol{Q}_t := \operatorname{argmin}_{\boldsymbol{P} \in \mathcal{O}^{r \times r}} \left\| \boldsymbol{X}_t \boldsymbol{P} - \boldsymbol{X}^\natural \right\|_\mathsf{F} \tag{13}$$

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.

[3] Compared with [11], when setting the eigenvalues in (10), 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

$$\boldsymbol{Y} = \frac{1}{2m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i\boldsymbol{a}_i^\top, \tag{9}$$

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

$$[\boldsymbol{\Lambda}_0]_i = \lambda_i\big(\boldsymbol{Y}\big) - \lambda, \quad i = 1, \cdots, r, \tag{10}$$

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

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

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \mu_t \cdot \frac{1}{m}\sum_{i=1}^{m}\left(\left\|\boldsymbol{a}_i^\top \boldsymbol{X}_t\right\|_2^2 - y_i\right)\boldsymbol{a}_i\boldsymbol{a}_i^\top \boldsymbol{X}_t. \tag{11}$$

Output: $\boldsymbol{X}_T$.

Theorem 1. *Fix $\boldsymbol{X}^\natural \in \mathbb{R}^{n\times r}$. Suppose that we have $y_i = \|\boldsymbol{a}_i^\top \boldsymbol{X}^\natural\|_2^2$ for $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_n)$ and $1 \le i \le m$. Suppose that $m \ge cnr^3(r+\sqrt{\kappa})\kappa^3\log n$ 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*

$$\mathrm{dist}\big(\boldsymbol{X}_t, \boldsymbol{X}^\natural\big) \le c_1\left(1 - 0.5\mu\sigma_r^2(\boldsymbol{X}^\natural)\right)^t \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}, \tag{14}$$

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

$$\max_{1\le l\le m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right)\right\|_2 \le c_2\left(1 - 0.5\mu\sigma_r^2(\boldsymbol{X}^\natural)\right)^t\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}, \tag{15}$$

for all $0 \le t \le c_3 n^5$. 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 $m \ge c\max\left\{\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}{\sigma_r(\boldsymbol{X}^\natural)}\sqrt{r}, \kappa\right\}\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^5}{\sigma_r^5(\boldsymbol{X}^\natural)}n\sqrt{r}\log(\kappa n)$ with some large enough constant $c > 0$. By adjusting constants, with probability at least $1 - O(mn^{-7})$, (15) holds for $0 \le t \le O(n^{c_5})$ in any power $c_5 \ge 1$.

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. dist $\left(\boldsymbol{X}_t, \boldsymbol{X}^\natural\right) \le \epsilon\|\boldsymbol{X}^\natural\|_{\mathsf{F}}$, it suffices to run gradient descent for $T = O\left(r^2\text{poly}\log(n)\log(1/\epsilon)\right)$ iterations. This results in a total computational complexity of $O\left(mnr^3\text{poly}\log(n)\log(1/\epsilon)\right)$.
- *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 [14].

Theorem 1 is established using a fixed step size. According to our theoretical analysis, the incoherence condition (15) has a significant impact on the convergence rate. After a few iterations, the incoherence condition can be bounded independent of $\log n$, which leads to a larger step size and 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*

$$\operatorname{dist}\big(\boldsymbol{X}_t, \boldsymbol{X}^\natural\big) \le c_1 \left(1 - 0.5\mu\sigma_r^2(\boldsymbol{X}^\natural)\right)^t \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}, \tag{16}$$

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

3 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 [1, 27, 4, 28]. See also [29, 30, 31, 32] 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 [33]. Specifically, for the phase retrieval problem, Wirtinger Flow (WF) and its variants [2, 12, 34, 14, 35, 36, 37] 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 [13, 38], blind deconvolution [39], and spectral compressed sensing [40]. We follow a similar route but analyze a more general problem that includes phase retrieval as a special case.

The paper [11] is most close to our work, 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 [11].

Algorithms with resampling	Sample complexity	Computational complexity
AltMin-LRROM [9]	$O(nr^4 \log^2 n \log(\frac{1}{\epsilon}))$	$O(mnr \log(\frac{1}{\epsilon}))$
gFM [10]	$O(nr^3 \log(\frac{1}{\epsilon}))$	$O(mnr \log(\frac{1}{\epsilon}))$
EP-ROM [8]	$O(nr^2 \log^4 n \log(\frac{1}{\epsilon}))$	$O(mn^2 \log(\frac{1}{\epsilon}))$
AP-ROM [8]	$O(nr^3 \log^4 n \log(\frac{1}{\epsilon}))$	$O(mnr \log n \log(\frac{1}{\epsilon}))$
Algorithms without resampling	**Sample complexity**	**Computational complexity**
Convex [1]	$O(nr)$	$O(mn^2 \frac{1}{\sqrt{\epsilon}})$
GD [11]	$O(nr^6 \log^2 n)$	$O(mn^5 r^3 \log^4 n \log(\frac{1}{\epsilon}))$
GD (Algorithm 1, Ours)	$O(nr^4 \log n)$	$O(mnr \max\{\log^2 n, r^2\} \log(\frac{1}{\epsilon}))$

Table 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.

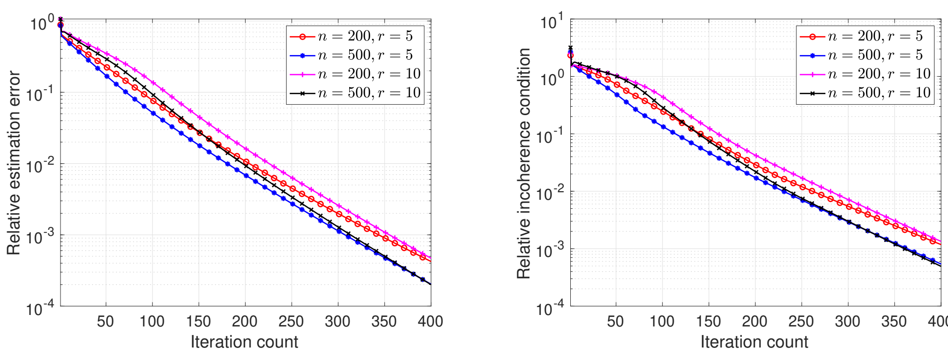
In spirit, our paper is also similar to [14], which studies the phase retrieval problem — a spectral case of the model we consider herein. Compared with the phase retrieval case, the extension from rank-one to rank-r case is highly nontrivial. In fact, none of the theorems or technical lemmas herein can be straightforwardly obtained without a significant amount of technical efforts. For instance, in the rank-one case, local strong convexity (cf. Lemma 1) holds uniformly within a local region surrounding the global optimum. However, this does not hold for the rank-r case, unless we restrict attention to highly restricted directions. See the restrictions on $\boldsymbol{V}$ in Lemma 1. This calls for more refined analysis in order to establish the uniform lower bound. In addition, with regards to spectral initialization, due to the non-uniform singular values, the perturbation bounds for the eigenvectors are more delicate to deal with. Last but not least, we also proved the linear convergence of the sample-wise incoherence measure, which has not been established even in the rank-one case in [14]; see Equation (15).

Several other existing works have suggested different approaches for low-rank matrix factorization from rank-one measurements, of which the statistical and computational guarantees to reach ϵ-accuracy are summarized in Table 1. We note our guarantee is the only one that achieves simultaneous near-optimal sample complexity and computational complexity. Iterative algorithms based on alternating minimization or noisy power iterations [9, 10, 8] 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 the Kaczmarz method [41, 42, 43] and approximate message passing [44]. In [45], the Kaczmarz method is generalized to solve the problem studied in this paper, but no theoretical performance guarantees are provided.

The local geometry studied in our paper is in contrast to [46], 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 paper.

Our model is also related to learning shallow neural networks. [47] 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, [6] adopts a greedy learning strategy, and can only guarantee sublinear convergence rate. Moreover, [7] studied the optimization landscape for an over-parameterized shallow neural network with quadratic activation, where r is larger than n.



(a) Relative estimation error (b) Relative incoherence condition

Figure 1: Performance of the proposed algorithm in regard to (a) relative estimation error, and (b) relative incoherence condition with respect to the iteration count using different problem sizes, when $m = 5nr$.

4 Numerical Experiments

In this section, we provide several numerical experiments to validate the effective and efficient performance of the proposed algorithm. During each experiment, given a pair of (n, r), the ground truth $\boldsymbol{X}^\natural \in \mathbb{R}^{n\times r}$ is generated with i.i.d. $\mathcal{N}(0, \frac{1}{n})$ entries. We first examine the relative estimation error $\mathrm{dist}\big(\boldsymbol{X}_t, \boldsymbol{X}^\natural\big)/\|\boldsymbol{X}^\natural\|_{\mathsf{F}}$ and the relative incoherence condition $\max_{1\le l\le m} \|\boldsymbol{a}_l^\top \big(\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\big)\|_2/\|\boldsymbol{X}^\natural\|_{\mathsf{F}}$ with respect to the iteration count using a constant step size $\mu_t = 0.03$, where the number of measurements set as $m = 5nr$. The convergence rates in Figure 1 are approximately linear, validating our theory.

We then examine the phase transitions of the proposed algorithm with respect to the number of measurements. Multiple Monte Carlo trials are conducted, and each trial is deemed a success if the relative estimation

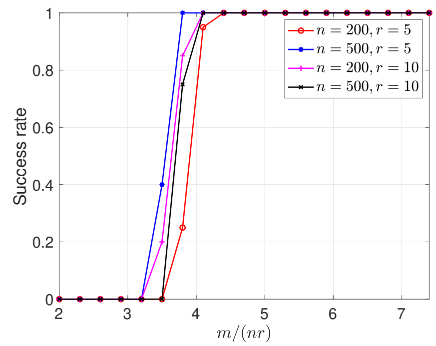


Figure 2: The success rate of the proposed algorithm with respect to the number of measurements $m/(nr)$ using different problem sizes.

error is less than 10^{-6} within $T = 1000$ iterations. Figure 2 depicts the success rate over 20 trials, where the proposed algorithm successfully recovers the ground truth as soon as the number of measurements is about 4 times above the degrees of freedom nr. These results suggest that the required sample complexity scales linearly with the degrees of freedom, and our theoretical guarantees are optimal up to logarithmic factors.

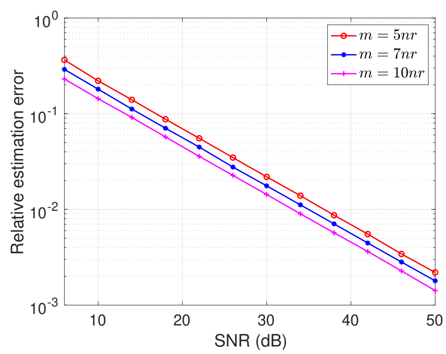


Figure 3: Relative estimation error with respect to SNR in different numbers of measurements, when $n = 200$ and $r = 5$.

Next, we numerically verify the stability of the proposed algorithm against additive noise, where each measurement is given as $y_i = \left\|\boldsymbol{a}_i^\top \boldsymbol{X}^\natural\right\|_2^2 + \epsilon_i$, where the noise ϵ_i is generated i.i.d. from $\mathcal{N}(0, \sigma^2)$. Figure 3 shows the estimation error with respect to SNR in different numbers of measurements when $n = 200$ and $r = 5$. As the noise variance σ^2 increases, the performance of the proposed algorithm degenerates smoothly. Increasing the number of measurements helps to improve the estimation accuracy.

Finally, we test the performance of the proposed algorithm when the measurement vectors $\boldsymbol{a}_i$'s are i.i.d. generated from a sub-Gaussian distribution under random initialization. Specifically, we consider a case where each entry in $\boldsymbol{a}_i$ is drawn i.i.d. from a uniform distribution $\mathcal{U}[-1, 1]$. We then implement gradient descent with a constant step size $\mu_t = 0.5$ starting from a random initialization, whose entries are generated i.i.d. following $\mathcal{N}(0, \frac{1}{n})$. Figure 4 shows the appealing convergence performance of the proposed algorithm.

5 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 [14] for phase retrieval and follows the general recipe outlined in [14], 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 [14]. In below, we first characterize a region of incoherence and contraction that enjoys both strong convexity

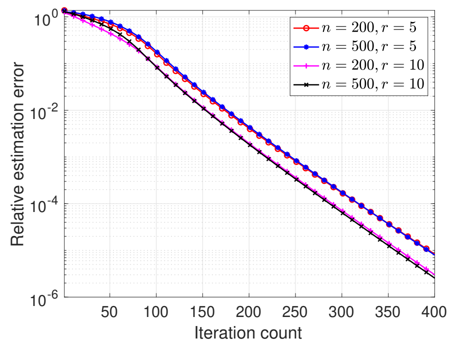


Figure 4: Relative estimation error with respect to the iteration count using different problem sizes when the sensing vectors are generated from sub-Gaussian distributions and a random initialization is employed, when $m = 5nr$.

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.

5.1 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

$$\nabla^2 f(\boldsymbol{X}) = \frac{1}{m}\sum_{i=1}^{m}\left[\left(\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2 - y_i\right)\boldsymbol{I}_r + 2\boldsymbol{X}^\top \boldsymbol{a}_i\boldsymbol{a}_i^\top \boldsymbol{X}\right] \otimes \left(\boldsymbol{a}_i\boldsymbol{a}_i^\top\right). \tag{17}$$

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 region of incoherence and contraction (RIC) throughout this paper. The proof is given in Appendix B.

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*

$$\operatorname{vec}(\boldsymbol{V})^\top \nabla^2 f(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{V}) \geq 1.026\sigma_r^2(\boldsymbol{X}^\natural)\left\|\boldsymbol{V}\right\|_\mathsf{F}^2, \tag{18}$$

and

$$\left\|\nabla^2 f(\boldsymbol{X})\right\| \leq 1.5\sigma_r^2(\boldsymbol{X}^\natural)\log n + 6\left\|\boldsymbol{X}^\natural\right\|_\mathsf{F}^2 \tag{19}$$

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

$$\left\|\boldsymbol{X} - \boldsymbol{X}^\natural\right\|_\mathsf{F} \leq \frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}, \tag{20a}$$

$$\max_{1\leq l\leq m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X} - \boldsymbol{X}^\natural\right)\right\|_2 \leq \frac{1}{24}\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}, \tag{20b}$$

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

$$\left\|\boldsymbol{T}_2 - \boldsymbol{X}^\natural\right\| \leq \frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|}, \tag{21}$$

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 (20) on $\boldsymbol{X}$ formally characterizes the RIC, which enjoys the claimed restricted strong convexity (see (18)) and smoothness (see (19)). 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 D.

Lemma 2. *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* $1 - c_1 n^{-12} - me^{-1.5n} - mn^{-12}$*, if* $\boldsymbol{X}_t$ *falls within the RIC as described in* (20)*, we have*

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

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

Assuming that the iterates $\{\boldsymbol{X}_t\}$, stay within the RIC (see (20)) for the first T_c iterations, according to Lemma 2, we have, by induction, that

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

as soon as

$$T_c \geq c \max\left\{\log^2 n, \frac{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^4}{\sigma_r^4\left(\boldsymbol{X}^\natural\right)}\right\} \log n, \tag{22}$$

for some large enough constant c. Notice that due to the high probability nature of each induction step, the union bound can only tolerate a polynomial number of induction steps, say $T_c = c_3 n^5$. After $t \geq T_c$, showing that $\boldsymbol{X}_{t+1}$ stays in the RIC is more immediate because the distance between $\boldsymbol{X}$ and $\boldsymbol{X}^\natural$ has decreased by enough so that the simple Cauchy-Schwarz inequality suffices to prove. In particular, we have

$$\begin{aligned}
\max_{1 \leq l \leq m} \left\|\boldsymbol{a}_l^\top \left(\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^\natural\right)\right\|_2 &\leq \max_{1 \leq l \leq m} \left\|\boldsymbol{a}_l\right\|_2 \left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^\natural\right\| \\
&\leq \sqrt{6n} \cdot \frac{1}{24\sqrt{6}} \cdot \frac{\sqrt{\log n}}{\sqrt{n}} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}} \qquad (23) \\
&= \frac{1}{24}\sqrt{\log n} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}},
\end{aligned}$$

where (23) follows from Lemma 10 for all $t \geq T_c$. Consequently, contraction of the estimation error $\operatorname{dist}\left(\boldsymbol{X}_t, \boldsymbol{X}^\natural\right)$ can be guaranteed by Lemma 1 for all $t \geq T_c$ with probability at least $1 - c_1 n^{-12} - me^{-1.5n} - mn^{-12}$.

5.2 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 (20) for the first T_c iterations, where T_c is on the order of (22). Verifying (20b) is conceptually hard since the iterates $\{\boldsymbol{X}_t\}$ are statistically dependent with all the sensing vectors $\{\boldsymbol{a}_i\}_{i=1}^m$. 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

$$f^{(l)}\left(\boldsymbol{X}\right) := \frac{1}{4m} \sum_{i: i \neq l} \left(y_i - \left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2\right)^2. \tag{24}$$

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

$$\boldsymbol{Y}^{(l)} = \frac{1}{2m}\sum_{i:i\neq l} y_i \boldsymbol{a}_i\boldsymbol{a}_i^\top, \tag{25}$$

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

$$\left[\boldsymbol{\Lambda}_0^{(l)}\right]_i = \lambda_i\big(\boldsymbol{Y}^{(l)}\big) - \lambda^{(l)}, \quad i = 1,\cdots,r, \tag{26}$$

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

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

$$\boldsymbol{X}_{t+1}^{(l)} = \boldsymbol{X}_t^{(l)} - \mu_t\cdot\frac{1}{m}\sum_{i:i\neq l}\left(\left\|\boldsymbol{a}_i^\top\boldsymbol{X}_t^{(l)}\right\|_2^2 - y_i\right)\boldsymbol{a}_i\boldsymbol{a}_i^\top\boldsymbol{X}_t^{(l)}. \tag{27}$$

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

5.3 Establishing Incoherence via Induction

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

$$\left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right\|_\mathsf{F} \le C_1\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}, \tag{28a}$$

$$\max_{1\le l\le m}\left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}\right\|_\mathsf{F} \le C_3\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\kappa\|\boldsymbol{X}^\natural\|_\mathsf{F}}, \tag{28b}$$

$$\max_{1\le l\le m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right)\right\|_2 \le C_2\left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}, \tag{28c}$$

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

$$C_1 + C_3 \le \frac{1}{24},\quad C_2 + \sqrt{6}C_3 \le \frac{1}{24},\quad 5.86C_1 + 29.3C_3 + 5\sqrt{6}C_3 \le C_2. \tag{29}$$

Furthermore, the step size μ is chosen as

$$\mu = \frac{c_0\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left(\sigma_r^2(\boldsymbol{X}^\natural)\log n + \|\boldsymbol{X}^\natural\|_\mathsf{F}^2\right)^2} \tag{30}$$

with appropriate universal constant $c_0 > 0$.

Our goal is to show that if the tth iteration $\boldsymbol{X}_t$ satisfies the induction hypotheses (28), then the $(t+1)$th iteration $\boldsymbol{X}_{t+1}$ also satisfies (28). It is straightforward to see that the hypothesis (28a) has already been established by Lemma 2, and we are left with (28b) and (28c). We first establish (28b) 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 E.

Lemma 3. *Suppose the sample size obeys $m \ge c\frac{\|\boldsymbol{X}^\natural\|_\mathsf{F}^4}{\sigma_r^4(\boldsymbol{X}^\natural)}nr\log(n\kappa)$ for some sufficiently large constant $c>0$. If the induction hypotheses* (28) *hold for the tth iteration, with probability at least $1-c_1n^{-12}-me^{-1.5n}-mn^{-12}$, we have*

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

as long as the step size obeys (30)*. 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 F.

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* (28) *hold for the* t*th iteration, with probability exceeding* $1 - c_1 n^{-12} - me^{-1.5n} - 2mn^{-12}$,

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

holds as long as the step size satisfies (30)*. Here,* $c_1 > 0$ *is some universal constant.*

5.4 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 G.

Lemma 5. *Suppose that the sample size exceeds* $m \geq c \max\left\{\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}{\sigma_r(\boldsymbol{X}^\natural)}\sqrt{r}, \kappa\right\} \frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^5}{\sigma_r^5(\boldsymbol{X}^\natural)} n\sqrt{r}\log n$ *for some sufficiently large constant* $c > 0$*. Then* $\boldsymbol{X}_0$ *satisfies* (28) *with probability at least* $1 - c_1 n^{-12} - me^{-1.5n} - 3mn^{-12}$*, where* c_1 *is some absolute positive constant.*

5.5 Putting things together

With Lemmata 1-5 in place, we are ready to put things together and prove the desired result Theorem 1.

Proof of Theorem 1. Lemma 5 justifies (14) and (15) in Theorem 1 for $t = 0$, i.e. spectral initialization. Given this base case, Lemma 2, together with Lemmata 3-4 establishes (14) and (15) in Theorem 1 for $1 \leq t \leq T_c$ in an inductive manner. Further built upon these, the proof is complete by repeatedly applying Lemma 2; see the paragraph after Lemma 2 for a complete argument of this part. □

6 Conclusions

In this paper, we have shown that low-rank positive semidefinite 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. [11]), 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.

There are several problems that are worth exploring in future investigation. For example, our theory reveals the typical size of the fitting error of $\boldsymbol{X}_t$ (i.e. $y_i - \|\boldsymbol{a}_i^\top \boldsymbol{X}_t\|_2$) in the presence of noiseless data, which would serve as a helpful benchmark when separating sparse outliers in the more realistic scenario. Another direction is to explore whether implicit regularization remains valid for learning shallow neural networks [47]. Since the current work can be viewed as learning a one-hidden-layer fully-connected network with a quadratic activation function $\sigma(z) = z^2$, it would be of great interest to study if the techniques utilized herein can be used to develop strong guarantees when the activation function takes other forms. Below we list a few further considerations that are worth discussion.

- *Random initialization.* The current paper focuses on a judiciously designed initialization scheme, namely spectral initialization. It turns out that for fast convergence of gradient descent, spectral initialization is not necessary and random initialization suffices; see Figure 4 for the numerical evidence. However, establishing the global convergence for gradient descent with random initialization is challenging. For example, [48] proves this for the case with $r = 1$. How to extend that to the general rank case remains an interesting and challenging problem. One roadblock is to construct appropriate sign-flipping sequences as in [48] to decouple the dependency of the gradient iterates on the data.

- *Sample complexity w.r.t. r and κ.* Last but not least, our sample complexity (i.e. $O(nr^4)$) is sub-optimal when the rank r is allowed to grow with the problem dimension n. Some of the difficulties stem from establishing the local strong convexity of the nonconvex loss function (cf. Lemma 1). Specifically, it is challenging to prove the local strong convexity with near optimal sample complexity $O(nr)$. Improving the theoretical support under optimal sample complexity $O(nr)$ remains a challenging problem.

- *Universal recovery guarantees?* Throughout the paper, we have assumed that the signal $\boldsymbol{X}^\natural$ is fixed and independent of the measurement vectors. One might naturally wonder whether the recovery guarantees continue to hold when we allow $\boldsymbol{X}^\natural$ to be arbitrary. This, however, might be highly nontrivial. In particular, establishing the local strong convexity and smoothness as shown in Lemma 1 is difficulty, if not impossible, when $\boldsymbol{X}^\natural$ is allowed to be arbitrary. Here the Hessian matrix $\nabla^2 f(\boldsymbol{X})$ does not necessarily concentrate around its mean with very few samples, which might preclude us from obtaining the desired strong convexity and smoothness conditions for the loss function.

Acknowledgements

The work of Y. Li and Y. Chi is supported in part by AFOSR under the grant FA9550-15-1-0205, by ONR under the grant N00014-18-1-2142, by ARO under the grant W911NF-18-1-0303, and by NSF under the grants CAREER ECCS-1818571, CCF-1704245 and CCF-1901199. The work of Y. Chen is supported in part by the grants AFOSR YIP award FA9550-19-1-0030, ONR N00014-19-1-2120, ARO W911NF-18-1-0303, NSF CCF-1907661 and IIS-1900140, and the Yang Family and David T. Wilkinson Innovation Funds.

Appendices

A Technical Lemmas

In this section, we document a few useful lemmas that are used throughout the proof.

Lemma 6. *[38, Lemma 5.4] For any matrices $\boldsymbol{X}$, $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, we have*

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

Lemma 7 (Covering number for low-rank matrices)**.** *[49, 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 obeying $\left|\bar{\mathcal{S}}_r\right| \leq (9/\epsilon)^{(n_1+n_2+1)r}$.*

Lemma 8. *[50, 2] Suppose $x_1, \cdots, x_m$ are i.i.d. real-valued random variables obeying $x_i \leq b$ for some deterministic number $b > 0$, $\mathbb{E}\left[x_i\right] = 0$, and $\mathbb{E}\left[x_i^2\right] = d^2$. 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 9. *[51, 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 10. *[2] 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 11. *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 [52, 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 12. *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*

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

Moreover, for any order $k \geq 1$, we have $\mathbb{E}\left[\|\boldsymbol{a}^\top \boldsymbol{H}\|_2^{2k} \right] \leq c_k \|\boldsymbol{H}\|_{\mathsf{F}}^{2k}$, 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

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

we can derive

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

and

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

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 $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_r] \in \mathbb{R}^{n\times r}$, $\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_r\}$, 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 $b_i \sim \mathcal{N}\left(0, \sigma_i^2\right)$ due to the fact $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_r$. Since $\mathbb{E}\left[b_i^{2t}\right] = \sigma_i^{2t}(2t-1)!! \leq c_k\sigma_i^{2t}$ 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

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

as claimed.

As a simple remark, the bound on $\mathbb{E}[\|\boldsymbol{a}^\top \boldsymbol{H}\|_2^{2k}]$ can also be obtained via a corollary of Hanson-Wright inequality. In particular, [53, Theorem 6.3.2] tells us that $\|\|\boldsymbol{a}^\top \boldsymbol{H}\|_2 - \|\boldsymbol{H}\|_{\mathsf{F}}\|_{\psi_2} \lesssim \|\boldsymbol{H}\|$. Here $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm of a random variable; see [53]. As a result, $\|\|\boldsymbol{a}^\top \boldsymbol{H}\|_2\|_{\psi_2} \lesssim \|\boldsymbol{H}\|_{\mathsf{F}}$, which immediately implies $\mathbb{E}\left[\|\boldsymbol{a}^\top \boldsymbol{H}\|_2^{2k}\right] \leq c_k\|\boldsymbol{H}\|_{\mathsf{F}}^{2k}$ for some $c_k > 0$ that depends only on k. □

Lemma 13. *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 $0 < \delta \leq 1$, suppose $m \geq c\delta^{-2}n\log n$ for some sufficiently large constant $c > 0$. Then we have*

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

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

Proof. This proof adapts the results of [2, 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 [54, Theorem 1.9], we have

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

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i(1)\right)^4-3\right|\geq\delta\right)\leq e^2\cdot e^{-\left(c_2\delta^2 m\right)^{1/4}};$$
$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i(1)\right)^6-15\right|\geq\delta\right)\leq e^2\cdot e^{-\left(c_3\delta^2 m\right)^{1/6}}.$$

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

$$\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i(1)\right)^2-1\right|\leq\delta,\ \left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i(1)\right)^4-3\right|\leq\delta,\ \text{and}\ \left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i(1)\right)^6-15\right|\leq\delta, \tag{31}$$

with probability at least $1-c_4n^{-13}$ for some constant $c_4>0$. Moreover, following [52, 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

$$\max_{1\leq i\leq m}\left|\boldsymbol{a}_i(1)\right|\leq\sqrt{36.5\log m}, \tag{32}$$

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

With (31) and (32), the results in [2, 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 $\boldsymbol{x}_k^\natural$, $k=1,\ldots,r$ and taking the union bound, we have

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

□

B Proof of Lemma 1

The crucial ingredient for proving the lower bound (18) is the following lemma, whose proof is provided in Appendix C.

Lemma 14. *Suppose $m\geq c\frac{\left\|\boldsymbol{X}^\natural\right\|_\mathsf{F}^4}{\sigma_r^4(\boldsymbol{X}^\natural)}nr\log\left(n\kappa\right)$ with some large enough positive constant c, then with probability at least $1-c_1n^{-12}-me^{-1.5n}$, we have*

$$\text{vec}\left(\boldsymbol{V}\right)^\top\nabla^2 f(\boldsymbol{X})\text{vec}\left(\boldsymbol{V}\right)\geq 2\text{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right)+1.204\sigma_r^2(\boldsymbol{X}^\natural)\left\|\boldsymbol{V}\right\|_\mathsf{F}^2, \tag{33}$$

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

With Lemma 14 in place, we are ready to prove (18). Let $\boldsymbol{V} = \boldsymbol{T}_1\boldsymbol{Q}_{\boldsymbol{T}} - \boldsymbol{T}_2$ satisfy the assumptions in Lemma 1, then we can demonstrate that

$$
\begin{aligned}
&\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) \\
&= \mathrm{Tr}\left(\left(\boldsymbol{X}^{\natural} - \boldsymbol{T}_2 + \boldsymbol{T}_2\right)^{\top}\boldsymbol{V}\left(\boldsymbol{X}^{\natural} - \boldsymbol{T}_2 + \boldsymbol{T}_2\right)^{\top}\boldsymbol{V}\right) \\
&= \mathrm{Tr}\left(\left(\boldsymbol{X}^{\natural} - \boldsymbol{T}_2\right)^{\top}\boldsymbol{V}\left(\boldsymbol{X}^{\natural} - \boldsymbol{T}_2\right)^{\top}\boldsymbol{V}\right) + 2\mathrm{Tr}\left(\left(\boldsymbol{X}^{\natural} - \boldsymbol{T}_2\right)^{\top}\boldsymbol{V}\boldsymbol{T}_2^{\top}\boldsymbol{V}\right) + \mathrm{Tr}\left(\boldsymbol{T}_2^{\top}\boldsymbol{V}\boldsymbol{T}_2^{\top}\boldsymbol{V}\right) \\
&\geq \mathrm{Tr}\left(\boldsymbol{T}_2^{\top}\boldsymbol{V}\boldsymbol{T}_2^{\top}\boldsymbol{V}\right) - \left\|\boldsymbol{X}^{\natural} - \boldsymbol{T}_2\right\|^2\|\boldsymbol{V}\|_{\mathsf{F}}^2 - 2\left\|\boldsymbol{X}^{\natural} - \boldsymbol{T}_2\right\|\|\boldsymbol{T}_2\|\|\boldsymbol{V}\|_{\mathsf{F}}^2 \\
&= \left\|\boldsymbol{T}_2^{\top}\boldsymbol{V}\right\|_{\mathsf{F}}^2 - \left\|\boldsymbol{X}^{\natural} - \boldsymbol{T}_2\right\|^2\|\boldsymbol{V}\|_{\mathsf{F}}^2 - 2\left\|\boldsymbol{X}^{\natural} - \boldsymbol{T}_2\right\|\|\boldsymbol{T}_2\|\|\boldsymbol{V}\|_{\mathsf{F}}^2 &\quad (34)\\
&\geq -\left[\left(\frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\|\boldsymbol{X}^{\natural}\|}\right)^2 + 2\cdot\frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\|\boldsymbol{X}^{\natural}\|}\cdot\left(\frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\left\|\boldsymbol{X}^{\natural}\right\|} + \left\|\boldsymbol{X}^{\natural}\right\|\right)\right]\|\boldsymbol{V}\|_{\mathsf{F}}^2 &\quad (35)\\
&\geq -0.0886\sigma_r^2(\boldsymbol{X}^{\natural})\,\|\boldsymbol{V}\|_{\mathsf{F}}^2, &\quad (36)
\end{aligned}
$$

where (34) follows from the fact that $\boldsymbol{T}_2^{\top}\boldsymbol{V} \in \mathbb{R}^{r\times r}$ is a symmetric matrix [55, Theorem 2], (35) 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 (36) is based on the fact $\left\|\boldsymbol{X}^{\natural}\right\| \geq \sigma_r(\boldsymbol{X}^{\natural})$. Combining (36) with Lemma 14, we establish the lower bound (18).

To prove the upper bound (19) asserted in the lemma, we make the observation that the Hessian in (17) 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\| &\quad (37)\\
&= \Bigg\|\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(\left\|\boldsymbol{a}_i^{\top}\boldsymbol{X}\right\|_2^2 - \left\|\boldsymbol{a}_i^{\top}\boldsymbol{X}^{\natural}\right\|_2^2\right)\right]\left(\boldsymbol{a}_i\boldsymbol{a}_i^{\top}\right) \\
&\qquad + \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) - 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 + \underbrace{2\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} + \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}, &\quad (38)
\end{aligned}
$$

where (37) follows from the fact $\|\boldsymbol{I}\otimes\boldsymbol{A}\| = \|\boldsymbol{A}\|$. It is seen from Lemma 13 that

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

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

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

With regards to the first term B_1, note that by Lemma 11 and (20b), 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,

$$B_1 \le 1.471\sigma_r^2\left(\boldsymbol{X}^\natural\right)\log n\left\|\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\boldsymbol{a}_i^\top\right\| \le 1.48\sigma_r^2\left(\boldsymbol{X}^\natural\right)\log n, \tag{39}$$

where the last line follows from Lemma 9. The proof is then finished by combining (38) with the preceding bounds on B_1, B_2 and B_3.

C Proof of Lemma 14

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

$$\begin{aligned}
&\mathrm{vec}\left(\boldsymbol{V}\right)^\top \nabla^2 f(\boldsymbol{X})\mathrm{vec}\left(\boldsymbol{V}\right) \\
&= \frac{1}{m}\sum_{i=1}^m \mathrm{vec}\left(\boldsymbol{V}\right)^\top\left[\left[\left(\left\|\boldsymbol{a}_i^\top\boldsymbol{X}\right\|_2^2 - y_i\right)\boldsymbol{I}_r + 2\boldsymbol{X}^\top\boldsymbol{a}_i\boldsymbol{a}_i^\top\boldsymbol{X}\right]\otimes\left(\boldsymbol{a}_i\boldsymbol{a}_i^\top\right)\right]\mathrm{vec}\left(\boldsymbol{V}\right) \\
&= \frac{1}{m}\sum_{i=1}^m\left(\left\|\boldsymbol{a}_i^\top\boldsymbol{X}\right\|_2^2 - y_i\right)\mathrm{vec}\left(\boldsymbol{V}\right)^\top\mathrm{vec}\left(\boldsymbol{a}_i\boldsymbol{a}_i^\top\boldsymbol{V}\right) + \frac{1}{m}\sum_{i=1}^m\mathrm{vec}\left(\boldsymbol{V}\right)^\top\mathrm{vec}\left(2\boldsymbol{a}_i\boldsymbol{a}_i^\top\boldsymbol{V}\boldsymbol{X}^\top\boldsymbol{a}_i\boldsymbol{a}_i^\top\boldsymbol{X}\right) \\
&= \frac{1}{m}\sum_{i=1}^m\left[\left(\left\|\boldsymbol{a}_i^\top\boldsymbol{X}\right\|_2^2 - \left\|\boldsymbol{a}_i^\top\boldsymbol{X}^\natural\right\|_2^2\right)\left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2 + 2\left(\boldsymbol{a}_i^\top\boldsymbol{X}\boldsymbol{V}^\top\boldsymbol{a}_i\right)^2\right].
\end{aligned} \tag{40}$$

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 (40) as

$$\begin{aligned}
&p\left(\boldsymbol{V},\boldsymbol{H},t\right) \\
&:= \underbrace{\frac{1}{m}\sum_{i=1}^m\left[\left\|\boldsymbol{a}_i^\top\boldsymbol{X}\right\|_2^2\left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2 + 2\left(\boldsymbol{a}_i^\top\boldsymbol{X}\boldsymbol{V}^\top\boldsymbol{a}_i\right)^2\right]}_{:=q(\boldsymbol{V},\boldsymbol{H},t)} - \frac{1}{m}\sum_{i=1}^m\left\|\boldsymbol{a}_i^\top\boldsymbol{X}^\natural\right\|_2^2\left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2.
\end{aligned} \tag{41}$$

The aim is thus to control $p\left(\boldsymbol{V},\boldsymbol{H},t\right)$ 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 (41). Let $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r]$, then by Lemma 13,

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

$$\leq \delta \left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2} \sum_{k=1}^{r} \|\boldsymbol{v}_k\|_2^2 = \delta \left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2} \|\boldsymbol{V}\|_{\mathsf{F}}^{2}.$$

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 $1 - c_1 r n^{-13}$,

$$\frac{1}{m} \sum_{i=1}^{m} \left\|\boldsymbol{a}_i^\top \boldsymbol{X}^{\natural}\right\|_2^2 \left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^2 \leq \left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^2 \left\|\boldsymbol{V}\right\|_{\mathsf{F}}^2 + 2\left\|\boldsymbol{X}^{\natural\top} \boldsymbol{V}\right\|_{\mathsf{F}}^2 + \frac{1}{24} \sigma_r^2(\boldsymbol{X}^{\natural}) \|\boldsymbol{V}\|_{\mathsf{F}}^2, \tag{42}$$

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 (41), 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.

C.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 12,

$$\begin{aligned}
\mathbb{E}[G_i] &= \mathbb{E}\left[\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2 \left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^2\right] + 2\mathbb{E}\left[\left(\boldsymbol{a}_i^\top \boldsymbol{X} \boldsymbol{V}^\top \boldsymbol{a}_i\right)^2\right] \\
&= \|\boldsymbol{X}\|_{\mathsf{F}}^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\left\|\boldsymbol{X}^\top \boldsymbol{V}\right\|_{\mathsf{F}}^2 + 2\left(\mathrm{Tr}\left(\boldsymbol{X}^\top \boldsymbol{V}\right)\right)^2 + 2\left\|\boldsymbol{X}\boldsymbol{V}^\top\right\|_{\mathsf{F}}^2 + 2\mathrm{Tr}\left(\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X}^\top \boldsymbol{V}\right) \\
&\leq \|\boldsymbol{X}\|_{\mathsf{F}}^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\|\boldsymbol{X}\|^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\|\boldsymbol{X}\|_{\mathsf{F}}^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\|\boldsymbol{X}\|^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\|\boldsymbol{X}\|^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 \\
&\leq 9\|\boldsymbol{X}\|_{\mathsf{F}}^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 = 9\left\|\boldsymbol{X}^{\natural} + t\frac{\sigma_r^2(\boldsymbol{X}^{\natural})}{\|\boldsymbol{X}^{\natural}\|_{\mathsf{F}}}\boldsymbol{H}\right\|_{\mathsf{F}}^2 &\quad (43) \\
&\leq 18\left(\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^2 + t^2 \frac{\sigma_r^4(\boldsymbol{X}^{\natural})}{\|\boldsymbol{X}^{\natural}\|_{\mathsf{F}}^2} \|\boldsymbol{H}\|_{\mathsf{F}}^2\right) \leq 18.002\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^2, &\quad (44)
\end{aligned}$$

where (43) follows $\|\boldsymbol{V}\|_{\mathsf{F}} = 1$ and $\boldsymbol{X} = \boldsymbol{X}^{\natural} + t\frac{\sigma_r^2(\boldsymbol{X}^{\natural})}{\|\boldsymbol{X}^{\natural}\|_{\mathsf{F}}}\boldsymbol{H}$, and (44) arises from the calculations with $\|\boldsymbol{H}\|_{\mathsf{F}} = 1$ and $t \leq 1/24$. Therefore, if we define $T_i = \mathbb{E}[G_i] - G_i$, we have $\mathbb{E}[T_i] = 0$ and

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

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

$$\begin{aligned}
&\mathbb{E}\left[T_i^2\right] = \mathbb{E}\left[G_i^2\right] - \left(\mathbb{E}[G_i]\right)^2 \leq \mathbb{E}\left[G_i^2\right] \\
&= \mathbb{E}\left[\left(\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2 \left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^2 + 2\left(\boldsymbol{a}_i^\top \boldsymbol{X}\boldsymbol{V}^\top \boldsymbol{a}_i\right)^2\right)^2\right] \\
&= \mathbb{E}\left[\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^4 \left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^4\right] + 4\mathbb{E}\left[\left(\boldsymbol{a}_i^\top \boldsymbol{X}\boldsymbol{V}^\top \boldsymbol{a}_i\right)^4\right] + 4\mathbb{E}\left[\left(\boldsymbol{a}_i^\top \boldsymbol{X}\boldsymbol{V}^\top \boldsymbol{a}_i\right)^2 \left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2 \left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^2\right] \\
&\leq 9\mathbb{E}\left[\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^4 \left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^4\right] &\quad (45) \\
&\leq 9\sqrt{\mathbb{E}\left[\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^8\right] \mathbb{E}\left[\left\|\boldsymbol{a}_i^\top \boldsymbol{V}\right\|_2^8\right]} &\quad (46)
\end{aligned}$$

$$\leq 9c_4 \|\boldsymbol{X}\|_{\mathsf{F}}^4 \|\boldsymbol{V}\|_{\mathsf{F}}^4 = 9c_4 \big\|\boldsymbol{X}\big\|_{\mathsf{F}}^4 \tag{47}$$

$$= 9c_4 \left\|\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}\boldsymbol{H}\right\|_{\mathsf{F}}^4 \lesssim \|\boldsymbol{X}^\natural\|_{\mathsf{F}}^4,$$

where (45) follows from the Cauchy-Schwarz inequality, (46) comes from the Hölder's inequality, and (47) is a consequence of Lemma 12. Apply Lemma 8 to arrive at

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m T_i \geq \frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^\natural\right)\right) \leq \exp\left(-c\frac{m\sigma_r^4\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^4}\right), \tag{48}$$

which further leads to

$$\begin{aligned}
&q\left(\boldsymbol{V}, \boldsymbol{H}, t\right) = \frac{1}{m}\sum_{i=1}^m G_i = \mathbb{E}\left[G_i\right] - \frac{1}{m}\sum_{i=1}^m T_i \\
&\geq \mathbb{E}\left[G_i\right] - \frac{1}{24}\sigma_r^2\big(\boldsymbol{X}^\natural\big) \\
&= \|\boldsymbol{X}\|_{\mathsf{F}}^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\big\|\boldsymbol{X}^\top\boldsymbol{V}\big\|_{\mathsf{F}}^2 + 2\left(\mathrm{Tr}\big(\boldsymbol{X}^\top\boldsymbol{V}\big)\right)^2 + 2\big\|\boldsymbol{X}\boldsymbol{V}^\top\big\|_{\mathsf{F}}^2 + 2\mathrm{Tr}\left(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X}^\top\boldsymbol{V}\right) - \frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^\natural\right) \\
&\geq \|\boldsymbol{X}\|_{\mathsf{F}}^2 \|\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\big\|\boldsymbol{X}^\top\boldsymbol{V}\big\|_{\mathsf{F}}^2 + 2\big\|\boldsymbol{X}\boldsymbol{V}^\top\big\|_{\mathsf{F}}^2 + 2\mathrm{Tr}\left(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X}^\top\boldsymbol{V}\right) - \frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^\natural\right).
\end{aligned} \tag{49}$$

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)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2}\|\boldsymbol{H}\|_{\mathsf{F}}^2 + 2t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\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)}{\|\boldsymbol{X}^\natural\|_{\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)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2}\big\|\boldsymbol{H}^\top\boldsymbol{V}\big\|_{\mathsf{F}}^2 + 2t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\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)}{\|\boldsymbol{X}^\natural\|_{\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)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2}\big\|\boldsymbol{H}\boldsymbol{V}^\top\big\|_{\mathsf{F}}^2 + 2t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\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)}{\|\boldsymbol{X}^\natural\|_{\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); \\
\mathrm{Tr}\left(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X}^\top\boldsymbol{V}\right) &= \mathrm{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)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}\mathrm{Tr}\left(\boldsymbol{H}^\top\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) + t^2\frac{\sigma_r^4\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2}\mathrm{Tr}\left(\boldsymbol{H}^\top\boldsymbol{V}\boldsymbol{H}^\top\boldsymbol{V}\right) \\
&\geq \mathrm{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)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}\big\|\boldsymbol{X}^\natural\big\|\|\boldsymbol{H}\|\|\boldsymbol{V}\|_{\mathsf{F}}^2 - t^2\frac{\sigma_r^4\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2}\|\boldsymbol{H}\|^2\|\boldsymbol{V}\|_{\mathsf{F}}^2 \\
&\geq \mathrm{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 (49), yields

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

$$\geq \left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2}+2\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right\|_{\mathsf{F}}^{2}+2\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right)+2\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)-\left(\frac{15}{24}+\frac{1}{12\cdot 24}\right)\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)$$
$$\geq \left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2}+2\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right\|_{\mathsf{F}}^{2}+2\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right)+1.371\sigma_r^2\left(\boldsymbol{X}^{\natural}\right).$$

C.2 Covering Arguments

Since we have obtained a lower bound on $q(\boldsymbol{V},\boldsymbol{H},t)$ 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 10. For notational convenience, we define

$$g(\boldsymbol{V},\boldsymbol{H},t)=q(\boldsymbol{V},\boldsymbol{H},t)-\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}^{2}-2\left\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right\|_{\mathsf{F}}^{2}-2\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right)-1.371\sigma_r^2\left(\boldsymbol{X}^{\natural}\right).$$

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

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

we have

$$\begin{aligned}&|g(\boldsymbol{V}_1,\boldsymbol{H},t)-g(\boldsymbol{V}_2,\boldsymbol{H},t)|\\&\leq|q(\boldsymbol{V}_1,\boldsymbol{H},t)-q(\boldsymbol{V}_2,\boldsymbol{H},t)|+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]-\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|\\&\quad+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|+\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\\&=216\epsilon n^2\left\|\boldsymbol{X}^{\natural}+t\frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\|\boldsymbol{X}^{\natural}\|_{\mathsf{F}}}\boldsymbol{H}\right\|^2+8\left\|\boldsymbol{X}^{\natural}\right\|^{2}\epsilon\\&\leq 432\epsilon n^2\left(\left\|\boldsymbol{X}^{\natural}\right\|^2+t^2\frac{\sigma_r^4\left(\boldsymbol{X}^{\natural}\right)}{\|\boldsymbol{X}^{\natural}\|_{\mathsf{F}}^2}\|\boldsymbol{H}\|^2\right)+8\left\|\boldsymbol{X}^{\natural}\right\|^{2}\epsilon\\&\leq\left(432.75n^2+8\right)\epsilon\left\|\boldsymbol{X}^{\natural}\right\|^2\leq\frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^{\natural}\right),\end{aligned}$$

as long as $\epsilon = \frac{\sigma_r^2(\boldsymbol{X}^\natural)}{10584 n^2 \|\boldsymbol{X}^\natural\|^2}$. Based on Lemma 7, the cardinality of this ϵ-net will be

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

Secondly, consider the ϵ-net covering argument for $\boldsymbol{H}$. Suppose $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ obey $\|\boldsymbol{H}_1\|_\mathsf{F} = 1$, $\|\boldsymbol{H}_2\|_\mathsf{F} = 1$, and $\|\boldsymbol{H}_1 - \boldsymbol{H}_2\|_\mathsf{F} \leq \epsilon$. 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[\left\|\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\boldsymbol{H}_1\right)\right\|_2^2\left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2 + 2\left(\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\boldsymbol{H}_1\right)\boldsymbol{V}^\top\boldsymbol{a}_i\right)^2\right] \\
&\quad - \frac{1}{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)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\boldsymbol{H}_2\right)\right\|_2^2\left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2 + 2\left(\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\boldsymbol{H}_2\right)\boldsymbol{V}^\top\boldsymbol{a}_i\right)^2\right]\Bigg| \\
&\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\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\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)}{\|\boldsymbol{X}^\natural\|_\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)}{\|\boldsymbol{X}^\natural\|_\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)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\boldsymbol{H}_2\right)\boldsymbol{V}^\top\boldsymbol{a}_i\right)^2\right| \\
&\leq 6n \cdot \sqrt{6n} \cdot t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\epsilon \cdot 2\sqrt{6n} \cdot \frac{25}{24}\|\boldsymbol{X}^\natural\| + 2 \cdot 6n \cdot t\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\epsilon \cdot 12n \cdot \frac{25}{24}\|\boldsymbol{X}^\natural\| \\
&\leq \frac{75}{8}\epsilon n^2\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\|\boldsymbol{X}^\natural\| \leq \frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^\natural\right),
\end{aligned}$$

as long as $\epsilon = \frac{1}{225n^2} \cdot \frac{\|\boldsymbol{X}^\natural\|_\mathsf{F}}{\|\boldsymbol{X}^\natural\|}$. Based on Lemma 7, the cardinality of this ϵ-net will be

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

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| \\
&= \Bigg|\frac{1}{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)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\boldsymbol{H}\right)\right\|_2^2\left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2 + 2\left(\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t_1\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\boldsymbol{H}\right)\boldsymbol{V}^\top\boldsymbol{a}_i\right)^2\right] \\
&\quad - \frac{1}{m}\sum_{i=1}^m\left[\left\|\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t_2\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\boldsymbol{H}\right)\right\|_2^2\left\|\boldsymbol{a}_i^\top\boldsymbol{V}\right\|_2^2 + 2\left(\boldsymbol{a}_i^\top\left(\boldsymbol{X}^\natural + t_2\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\boldsymbol{H}\right)\boldsymbol{V}^\top\boldsymbol{a}_i\right)^2\right]\Bigg| \\
&\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)}{\|\boldsymbol{X}^\natural\|_\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)}{\|\boldsymbol{X}^\natural\|_\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)}{\|\boldsymbol{X}^\natural\|_\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)}{\|\boldsymbol{X}^\natural\|_\mathsf{F}}\boldsymbol{H}\right)\boldsymbol{V}^\top\boldsymbol{a}_i\right)^2\right|
\end{aligned}$$

$$\le 6n \cdot \sqrt{6n} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}\epsilon \cdot 2\sqrt{6n} \cdot \frac{25}{24}\|\boldsymbol{X}^\natural\| + 2 \cdot 6n \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}\epsilon \cdot 12n \cdot \frac{25}{24}\|\boldsymbol{X}^\natural\|$$
$$\le 225\epsilon n^2 \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}\|\boldsymbol{X}^\natural\| \le \frac{1}{24}\sigma_r^2\left(\boldsymbol{X}^\natural\right),$$

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

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

$$q(\boldsymbol{V}, \boldsymbol{H}, t) \ge \|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2 + 2\|\boldsymbol{X}^{\natural\top}\boldsymbol{V}\|_{\mathsf{F}}^2 + 2\mathrm{Tr}\left(\boldsymbol{X}^{\natural\top}\boldsymbol{V}\boldsymbol{X}^{\natural\top}\boldsymbol{V}\right) + 1.246\sigma_r^2\left(\boldsymbol{X}^\natural\right), \tag{50}$$

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

C.3 Finishing the Proof

Combining (42) and (50), we can prove

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

as claimed.

D Proof of Lemma 2

We first note that

$$\begin{aligned}
\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^\natural\|_{\mathsf{F}}^2 &\le \|\boldsymbol{X}_{t+1}\boldsymbol{Q}_t - \boldsymbol{X}^\natural\|_{\mathsf{F}}^2 &(51)\\
&= \|(\boldsymbol{X}_t - \mu\nabla f(\boldsymbol{X}_t))\boldsymbol{Q}_t - \boldsymbol{X}^\natural\|_{\mathsf{F}}^2\\
&= \|\boldsymbol{X}_t\boldsymbol{Q}_t - \mu\nabla f(\boldsymbol{X}_t\boldsymbol{Q}_t) - \boldsymbol{X}^\natural\|_{\mathsf{F}}^2 &(52)\\
&= \left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural - \mu\cdot\mathrm{vec}\left(\nabla f(\boldsymbol{X}_t\boldsymbol{Q}_t) - \nabla f\left(\boldsymbol{X}^\natural\right)\right)\right\|_2^2, &(53)
\end{aligned}$$

where we write

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

Here, (51) follows from the definition of $\boldsymbol{Q}_{t+1}$ (see (13)), (52) holds owing to the identity $\nabla f(\boldsymbol{X}_t)\boldsymbol{Q}_t = \nabla f(\boldsymbol{X}_t\boldsymbol{Q}_t)$ for $\boldsymbol{Q}_t \in \mathcal{O}^{r\times r}$, and (53) arises from the fact that $\nabla f\left(\boldsymbol{X}^\natural\right) = \boldsymbol{0}$. Let

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

where $\tau \in [0, 1]$. Then, by the fundamental theorem of calculus for vector-valued functions [56],

$$\text{RHS of (53)} = \left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural - \mu\cdot\int_0^1 \nabla^2 f(\boldsymbol{X}_t(\tau))\left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)d\tau\right\|_2^2 \tag{54}$$

$$
\begin{aligned}
&= \left\| \left(\boldsymbol{I} - \mu \cdot \int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right) \mathrm{d}\tau \right) \left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right) \right\|_2^2 \\
&= \left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)^\top \left(\boldsymbol{I} - \mu \cdot \int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right) \mathrm{d}\tau \right)^2 \left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right) \\
&= \left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural\right\|_2^2 - 2\mu \cdot \left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)^\top \left(\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right) \mathrm{d}\tau \right) \left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right) \\
&\qquad + \mu^2 \cdot \left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)^\top \left(\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right) \mathrm{d}\tau \right)^2 \left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right) \\
&\leq \left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural\right\|_2^2 - 2\mu \cdot \left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right)^\top \left(\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right) \mathrm{d}\tau \right) \left(\boldsymbol{x}_t - \boldsymbol{x}^\natural\right) \\
&\qquad + \mu^2 \cdot \left\| \int_0^1 \nabla^2 f\left(\boldsymbol{X}_t(\tau)\right) \mathrm{d}\tau \right\|^2 \left\|\boldsymbol{x}_t - \boldsymbol{x}^\natural\right\|_2^2 . \qquad (55)
\end{aligned}
$$

It is easy to verify that $\boldsymbol{X}_t(\tau)$ satisfies (20) for any $\tau \in [0,1]$, since

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

and

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

Lemma 1 then implies that

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

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 (53) and (55) gives

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

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

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

E 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

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

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,

$$\begin{aligned}
\left|\left\|\boldsymbol{a}_l^\top\boldsymbol{X}_t^{(l)}\right\|_2^2 - y_l\right| &= \left|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \boldsymbol{X}^\natural\right)\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} + \boldsymbol{X}^\natural\right)^\top\boldsymbol{a}_l\right| \\
&\le \left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \boldsymbol{X}^\natural\right)\right\|_2\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} + \boldsymbol{X}^\natural\right)\right\|_2. \qquad (56)
\end{aligned}$$

The first term in (56) can be bounded by

$$\begin{aligned}
&\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \boldsymbol{X}^\natural\right)\right\|_2 \\
&\le \left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \boldsymbol{X}_t\boldsymbol{Q}_t\right)\right\|_2 + \left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right)\right\|_2 \\
&\le \sqrt{6n}\left\|\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \boldsymbol{X}_t\boldsymbol{Q}_t\right\| + C_2\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} \\
&\le \sqrt{6n}C_3\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\kappa\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} + C_2\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} \\
&\le \left(\sqrt{6}C_3 + C_2\right)\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t\sqrt{\log n}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}, \qquad (57)
\end{aligned}$$

where we have used the triangle inequality, Lemma 10, as well as the induction hypotheses (28c) and (28b). Similarly, the second term in (56) can be bounded as

$$\begin{aligned}
\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} + \boldsymbol{X}^\natural\right)\right\|_2 &\le \left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \boldsymbol{X}^\natural\right)\right\|_2 + 2\left\|\boldsymbol{a}_l^\top\boldsymbol{X}^\natural\right\|_2 \\
&\le \left(\sqrt{6}C_3 + C_2\right)\sqrt{\log n}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} + 11.72\sqrt{\log n}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}} \\
&\le \left(\sqrt{6}C_3 + C_2 + 11.72\right)\sqrt{\log n}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}, \qquad (58)
\end{aligned}$$

where we have used (57), Lemma 11, and $\sigma_r^2\left(\boldsymbol{X}^\natural\right) \le \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2$. Similarly, we can also obtain

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

Substituting (57) and (58) into (56), and using the above inequality, we get

$$\left\|\boldsymbol{S}_{t,2}^{(l)}\right\|_{\mathsf{F}} = \mu\frac{1}{m}\cdot\left|\left\|\boldsymbol{a}_l^\top\boldsymbol{X}_t^{(l)}\right\|_2^2 - y_l\right|\cdot\left\|\boldsymbol{a}_l\boldsymbol{a}_l^\top\boldsymbol{X}_t^{(l)}\right\|_{\mathsf{F}}$$

$$
\begin{aligned}
&\leq C_4^2 \left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t \cdot \mu \frac{1}{m}\cdot \sigma_r^2\left(\boldsymbol{X}^\natural\right)\log n \cdot \|\boldsymbol{a}_l\|_2 \left\|\boldsymbol{a}_l^\top \boldsymbol{X}_t^{(l)}\right\|_2 \\
&\leq \sqrt{6} C_4^3 \left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t \cdot \mu \frac{1}{m}\cdot \sigma_r^2\left(\boldsymbol{X}^\natural\right)\log n \cdot \sqrt{n}\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}\sqrt{\log n} \\
&= \sqrt{6} C_4^3 \left(1-0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t \cdot \mu \frac{\sqrt{n}\cdot(\log n)^{3/2}}{m}\sigma_r^2\left(\boldsymbol{X}^\natural\right)\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}, \qquad (59)
\end{aligned}
$$

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 [56], where

$$
\boldsymbol{X}_t^{(l)}(\tau) = \boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} + \tau\left(\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}\right), \qquad (60)
$$

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

$$
\begin{aligned}
\left\|\boldsymbol{S}_{t,1}^{(l)}\right\|_{\mathsf{F}}^2 \leq{}& \left\|\boldsymbol{x}_t - \boldsymbol{x}_t^{(l)}\right\|_2^2 + \mu^2\left\|\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t^{(l)}(\tau)\right)\mathrm{d}\tau\right\|^2\left\|\boldsymbol{x}_t - \boldsymbol{x}_t^{(l)}\right\|_2^2 \\
&- 2\mu\cdot\left(\boldsymbol{x}_t - \boldsymbol{x}_t^{(l)}\right)^\top\left(\int_0^1 \nabla^2 f\left(\boldsymbol{X}_t^{(l)}(\tau)\right)\mathrm{d}\tau\right)\left(\boldsymbol{x}_t - \boldsymbol{x}_t^{(l)}\right). \qquad (61)
\end{aligned}
$$

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

$$
\begin{aligned}
\left\|\boldsymbol{X}_t^{(l)}(\tau) - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} &= \left\|(1-\tau)\left(\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \boldsymbol{X}_t\boldsymbol{Q}_t\right) + \boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \\
&\leq (1-\tau)\left\|\boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)} - \boldsymbol{X}_t\boldsymbol{Q}_t\right\|_{\mathsf{F}} + \left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \\
&\leq C_3\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\kappa\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} + C_1\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} \qquad (62) \\
&\leq \left(C_3\sqrt{\frac{\log n}{n}} + C_1\right)\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} \leq \frac{1}{24}\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}, \qquad (63)
\end{aligned}
$$

where (62) follows from the induction hypotheses (28a) and (28b), and (63) follows as long as $C_1 + C_3 \leq \frac{1}{24}$. Further, for all $1 \leq l \leq m$, by the induction hypothesis (28b) and (28c),

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

as long as $\sqrt{6}C_3 + C_2 \leq \frac{1}{24}$. Therefore, Lemma 1 holds for $\boldsymbol{X}_t^{(l)}(\tau)$, and similar to Appendix D, (61) can be further bounded by

$$\left\|\boldsymbol{S}_{t,1}^{(l)}\right\|_{\mathsf{F}} \leq \left(1 - 0.513\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)\left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}\right\|_{\mathsf{F}} \tag{64}$$

as long as $\mu \leq \frac{1.026\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left(1.5\sigma_r^2(\boldsymbol{X}^\natural)\log n + 6\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2\right)^2}$. Consequently, combining (59) and (64), we can get

$$\begin{aligned}
\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right\|_{\mathsf{F}} &\leq \left\|\boldsymbol{S}_{t,1}^{(l)}\right\|_{\mathsf{F}} + \left\|\boldsymbol{S}_{t,2}^{(l)}\right\|_{\mathsf{F}} \\
&\leq \left(1 - 0.513\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)\left\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}\right\|_{\mathsf{F}} \\
&\quad + \sqrt{6}C_4^3\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^t \cdot \mu\frac{\sqrt{n}\cdot(\log n)^{3/2}}{m}\sigma_r^2\left(\boldsymbol{X}^\natural\right)\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}} \\
&\leq C_3\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^{t+1}\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\kappa\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}},
\end{aligned} \tag{65}$$

where (65) follows from the induction hypothesis (28b), as long as $m \geq c\kappa\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2}{\sigma_r^2(\boldsymbol{X}^\natural)}n\log n$ for some large enough constant $c > 0$.

F 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 11, we have

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

Since following Lemma 2,

$$\begin{aligned}
\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^\natural\right\|\left\|\boldsymbol{X}^\natural\right\| &\leq \left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^\natural\right\|_{\mathsf{F}}\left\|\boldsymbol{X}^\natural\right\| \\
&\leq C_1\left(1 - 0.513\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^{t+1}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\cdot\left\|\boldsymbol{X}^\natural\right\| \\
&\leq \frac{1}{2}\sigma_r^2\left(\boldsymbol{X}^\natural\right),
\end{aligned}$$

as long as $C_1 \leq \frac{1}{2}$, and following Lemma 3,

$$\begin{aligned}
\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right\|\left\|\boldsymbol{X}^\natural\right\| &\leq \left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right\|_{\mathsf{F}}\left\|\boldsymbol{X}^\natural\right\| \\
&\leq C_3\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^\natural\right)\mu\right)^{t+1}\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\kappa\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}\cdot\left\|\boldsymbol{X}^\natural\right\| \\
&\leq \frac{1}{4}\sigma_r^2\left(\boldsymbol{X}^\natural\right),
\end{aligned}$$

as long as $C_3 \leq \frac{1}{4}$, we can invoke Lemma 37 in [14] and get

$$\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{Q}_{t+1}^{(l)}\right\|_{\mathsf{F}} \leq 5\kappa\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right\|_{\mathsf{F}}.$$

Further, by the triangle inequality, Lemma 10, Lemma 3 and Lemma 2, we can deduce that

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

$$
\begin{aligned}
&+ 5.86\sqrt{\log n}\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^{\natural}\right\|_{\mathsf{F}} \\
&\leq \left(\sqrt{6n} + 5.86\sqrt{\log n}\right)\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{Q}_{t+1}^{(l)}\right\|_{\mathsf{F}} + 5.86\sqrt{\log n}\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^{\natural}\right\|_{\mathsf{F}} \\
&\leq 5\left(\sqrt{6n} + 5.86\sqrt{\log n}\right)\kappa\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right\|_{\mathsf{F}} + 5.86\sqrt{\log n}\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^{\natural}\right\|_{\mathsf{F}} \\
&\leq 5\left(\sqrt{6n} + 5.86\sqrt{\log n}\right)\kappa \cdot C_3\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)\mu\right)^{t+1}\sqrt{\frac{\log n}{n}} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\kappa\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}} \\
&+ 5.86\sqrt{\log n} \cdot C_1\left(1 - 0.513\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)\mu\right)^{t+1} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}} \\
&\leq \left(5\sqrt{6}C_3 + 5.86C_1 + 29.3C_3\sqrt{\frac{\log n}{n}}\right)\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)\mu\right)^{t+1}\sqrt{\log n} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}} \\
&\leq C_2\left(1 - 0.5\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)\mu\right)^{t+1}\sqrt{\log n} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^{\natural}\right)}{\left\|\boldsymbol{X}^{\natural}\right\|_{\mathsf{F}}},
\end{aligned}
$$

where the last line follows as long as $5\sqrt{6}C_3 + 5.86C_1 + 29.3C_3 \leq C_2$. The proof is then finished by applying the union bound for all $1 \leq l \leq m$.

G Proof of Lemma 5

Define

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

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

$$
\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^{\top}\boldsymbol{Z}_0^{(l)} - \boldsymbol{Z}_0^{\top}\boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)} = \frac{1}{2m}y_l\boldsymbol{Z}_0^{\top}\boldsymbol{a}_l\boldsymbol{a}_l^{\top}\boldsymbol{Z}_0^{(l)}. \tag{66}
$$

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 (28) in the base case when $t = 0$ one by one.

G.1 Proof of (28a)

From Lemma 6, we have

$$
\begin{aligned}
\left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}^{\natural}\right\|_{\mathsf{F}} &\leq \frac{1}{\sqrt{2\left(\sqrt{2}-1\right)}\sigma_r\left(\boldsymbol{X}^{\natural}\right)}\left\|\boldsymbol{X}_0\boldsymbol{X}_0^{\top} - \boldsymbol{X}^{\natural}\boldsymbol{X}^{\natural\top}\right\|_{\mathsf{F}} \\
&= \frac{1}{\sqrt{2\left(\sqrt{2}-1\right)}\sigma_r\left(\boldsymbol{X}^{\natural}\right)}\left\|\boldsymbol{Z}_0\boldsymbol{\Lambda}_0\boldsymbol{Z}_0^{\top} - \boldsymbol{X}^{\natural}\boldsymbol{X}^{\natural\top}\right\|_{\mathsf{F}} \\
&\leq \frac{\sqrt{r}}{\sqrt{2\left(\sqrt{2}-1\right)}\sigma_r\left(\boldsymbol{X}^{\natural}\right)}\left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^{\top} - \boldsymbol{X}^{\natural}\boldsymbol{X}^{\natural\top} - \lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^{\top}\right\|.
\end{aligned} \tag{67}
$$

The last term in (67) can be further bounded as

$$
\begin{aligned}
&\left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^{\top} - \boldsymbol{X}^{\natural}\boldsymbol{X}^{\natural\top} - \lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^{\top}\right\| \\
&\leq \left\|\boldsymbol{Y} - \frac{1}{2}\|\boldsymbol{X}^{\natural}\|_{\mathsf{F}}^2\boldsymbol{I} - \boldsymbol{X}^{\natural}\boldsymbol{X}^{\natural\top}\right\| + \left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^{\top} - \boldsymbol{Y} + \frac{1}{2}\|\boldsymbol{X}^{\natural}\|_{\mathsf{F}}^2\boldsymbol{Z}_{0,c}\boldsymbol{Z}_{0,c}^{\top}\right\| + \left\|\frac{1}{2}\|\boldsymbol{X}^{\natural}\|_{\mathsf{F}}^2\boldsymbol{Z}_0\boldsymbol{Z}_0^{\top} - \lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^{\top}\right\|
\end{aligned}
$$

$$\leq \delta \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 + \delta \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 + \delta \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 = 3\delta \left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2, \tag{68}$$

where (68) 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 13, the Weyl's inequality, and

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

via Lemma 9. Plugging (68) into (67), 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 $\delta = c\frac{\sigma_r^3(\boldsymbol{X}^\natural)}{\sqrt{r}\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^3}$ for a sufficiently small constant c, i.e. $m \gtrsim \frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^6}{\sigma_r^6(\boldsymbol{X}^\natural)} nr\log n$, we get $\left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \leq C_1\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}$. Following similar procedures, we can also show $\left\|\boldsymbol{X}_0^{(l)}\boldsymbol{Q}_0^{(l)} - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \leq C_1\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}$.

G.2 Proof of (28b)

Following Weyl's inequality, by (28a), we have

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

and similarly, $\left|\sigma_i\left(\boldsymbol{X}_0^{(l)}\right) - \sigma_i\left(\boldsymbol{X}^\natural\right)\right| \leq C_1\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}$, for $i = 1, \cdots, r$. Combined with Lemma 6, there exists some constant c such that

$$\begin{aligned}
&\left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}\right\|_{\mathsf{F}} \leq \frac{1}{\sqrt{2\left(\sqrt{2}-1\right)}\sigma_r\left(\boldsymbol{X}_0\right)}\left\|\boldsymbol{X}_0\boldsymbol{X}_0^\top - \boldsymbol{X}_0^{(l)}\boldsymbol{X}_0^{(l)\top}\right\|_{\mathsf{F}} \\
&\leq \frac{c}{\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\boldsymbol{X}_0\boldsymbol{X}_0^\top - \boldsymbol{X}_0^{(l)}\boldsymbol{X}_0^{(l)\top}\right\|_{\mathsf{F}} \\
&= \frac{c}{\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\boldsymbol{Z}_0\boldsymbol{\Lambda}_0\boldsymbol{Z}_0^\top - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Lambda}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}} \\
&= \frac{c}{\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}\boldsymbol{Z}_0^{(l)\top} - \lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^\top + \lambda^{(l)}\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}} \\
&\leq \frac{c}{\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}} + \frac{c}{\sigma_r\left(\boldsymbol{X}^\natural\right)}\left\|\lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^\top - \lambda^{(l)}\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}}. &(69)
\end{aligned}$$

We will bound each term in (69), respectively. For the first term, we have

$$\begin{aligned}
&\left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}} = \left\|\left[\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_0^{(l)} - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}, \boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_{0,c}^{(l)}\right]\right\|_{\mathsf{F}} \\
&\leq \left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_0^{(l)} - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}\right\|_{\mathsf{F}} + \left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_{0,c}^{(l)}\right\|_{\mathsf{F}} \\
&\leq \left\|\boldsymbol{Z}_0\boldsymbol{\Sigma}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_0^{(l)} - \boldsymbol{Z}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}\right\|_{\mathsf{F}} + \left\|\boldsymbol{Z}_0\boldsymbol{Z}_0^\top\boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)} - \boldsymbol{Z}_0^{(l)}\boldsymbol{\Sigma}_0^{(l)}\right\|_{\mathsf{F}} + \|\boldsymbol{Y}\|\left\|\boldsymbol{Z}_0^\top\boldsymbol{Z}_{0,c}^{(l)}\right\|_{\mathsf{F}} \\
&\leq \left\|\boldsymbol{Z}_0 \cdot \frac{1}{2m}y_l\boldsymbol{Z}_0^\top\boldsymbol{a}_l\boldsymbol{a}_l^\top\boldsymbol{Z}_0^{(l)}\right\|_{\mathsf{F}} + \left\|\boldsymbol{Z}_0\boldsymbol{Z}_0^\top - \boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_{\mathsf{F}}\left\|\boldsymbol{Y}^{(l)}\right\| + \|\boldsymbol{Y}\|\left\|\boldsymbol{Z}_0^\top\boldsymbol{Z}_{0,c}^{(l)}\right\|_{\mathsf{F}}, &(70)
\end{aligned}$$

where the last line follows from (66). Note that the first term in (70) can be bounded as

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

$$\lesssim \frac{\sqrt{n}\cdot(\log n)^{3/2}\cdot\sqrt{r}}{m}\left\|\boldsymbol{X}^\natural\right\|_\mathsf{F}^2, \tag{71}$$

which follows Lemma 10 and Lemma 11. The second term in (70) 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 $\left\|\boldsymbol{Z}_0 - \boldsymbol{Z}_0^{(l)}\boldsymbol{T}_0^{(l)}\right\|_\mathsf{F} \le \sqrt{2}\left\|\boldsymbol{Z}_0^\top\boldsymbol{Z}_{0,c}^{(l)}\right\|_\mathsf{F}$ [57]. Putting this together with the third term in (70), we have

$$\begin{aligned}\left\|\boldsymbol{Z}_0\boldsymbol{Z}_0^\top - \boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_\mathsf{F}\left\|\boldsymbol{Y}^{(l)}\right\| + \|\boldsymbol{Y}\|\left\|\boldsymbol{Z}_0^\top\boldsymbol{Z}_{0,c}^{(l)}\right\|_\mathsf{F} &\le \left(2\sqrt{2}\left\|\boldsymbol{Y}^{(l)}\right\| + \|\boldsymbol{Y}\|\right)\left\|\boldsymbol{Z}_0^\top\boldsymbol{Z}_{0,c}^{(l)}\right\|_\mathsf{F} \\ &\lesssim \left\|\boldsymbol{X}^\natural\right\|_\mathsf{F}^2\frac{\left\|\left(\frac{1}{m}y_l\boldsymbol{a}_l\boldsymbol{a}_l^\top\right)\boldsymbol{Z}_0^{(l)}\right\|_\mathsf{F}}{\sigma_r^2\left(\boldsymbol{X}^\natural\right)} \qquad (72) \\ &\lesssim \frac{\left\|\boldsymbol{a}_l^\top\boldsymbol{X}^\natural\right\|_2^2\left\|\boldsymbol{a}_l^\top\boldsymbol{Z}_0^{(l)}\right\|_2\|\boldsymbol{a}_l\|_2}{m}\frac{\left\|\boldsymbol{X}^\natural\right\|_\mathsf{F}^2}{\sigma_r^2\left(\boldsymbol{X}^\natural\right)} \\ &\lesssim \frac{\sqrt{n}\cdot(\log n)^{3/2}\cdot\sqrt{r}}{m}\frac{\left\|\boldsymbol{X}^\natural\right\|_\mathsf{F}^4}{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}, \qquad (73)\end{aligned}$$

where (72) follows from Lemma 13 and the Davis-Kahan $\sin\Theta$ theorem [58], and (73) follows from Lemma 10 and Lemma 11.

For the second term in (69), we have

$$\begin{aligned}\left\|\lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^\top - \lambda^{(l)}\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_\mathsf{F} &= \left\|\lambda\boldsymbol{Z}_0\boldsymbol{Z}_0^\top - \lambda\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top} + \lambda\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top} - \lambda^{(l)}\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_\mathsf{F} \\ &\le \lambda\cdot\left\|\boldsymbol{Z}_0\boldsymbol{Z}_0^\top - \boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_\mathsf{F} + \left|\lambda - \lambda^{(l)}\right|\cdot\left\|\boldsymbol{Z}_0^{(l)}\boldsymbol{Z}_0^{(l)\top}\right\|_\mathsf{F} \\ &\lesssim \frac{\sqrt{n}\cdot(\log n)^{3/2}\cdot\sqrt{r}}{m}\frac{\left\|\boldsymbol{X}^\natural\right\|_\mathsf{F}^4}{\sigma_r^2\left(\boldsymbol{X}^\natural\right)} + \frac{y_l}{2m}\sqrt{r} \qquad (74) \\ &\lesssim \frac{\sqrt{n}\cdot(\log n)^{3/2}\cdot\sqrt{r}}{m}\frac{\left\|\boldsymbol{X}^\natural\right\|_\mathsf{F}^4}{\sigma_r^2\left(\boldsymbol{X}^\natural\right)} + \frac{\sqrt{r}\cdot\log n}{m}\left\|\boldsymbol{X}^\natural\right\|_\mathsf{F}^2, \qquad (75)\end{aligned}$$

where the first term of (74) is bounded similarly as (73), and (75) follows from Lemma 11. Combining (71), (73), and (75), we obtain

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

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

G.3 Proof of (28c)

Since from (28a) and (28b),

$$\left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}^\natural\right\|\left\|\boldsymbol{X}^\natural\right\| \le \left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}^\natural\right\|_\mathsf{F}\left\|\boldsymbol{X}^\natural\right\| \lesssim \sigma_r^2\left(\boldsymbol{X}^\natural\right),$$

and for every $1 \le l \le m$,

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

with proper constants, following Lemma 37 in [14], we have

$$\left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}_0^{(l)}\boldsymbol{Q}_0^{(l)}\right\|_{\mathsf{F}} \leq 5\kappa \left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}\right\|_{\mathsf{F}},$$

which implies that for every $1 \leq l \leq m$ we can get

$$\begin{aligned}
\left\|\boldsymbol{X}_0^{(l)}\boldsymbol{Q}_0^{(l)} - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} &\leq \left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}_0^{(l)}\boldsymbol{Q}_0^{(l)}\right\|_{\mathsf{F}} + \left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \\
&\lesssim \kappa \left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}\right\|_{\mathsf{F}} + \left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \\
&\lesssim \kappa\sqrt{\frac{\log n}{n}} \cdot \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\kappa\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} + \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} \\
&\lesssim \frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}}.
\end{aligned}$$

This further gives

$$\begin{aligned}
&\max_{1\leq l\leq m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}^\natural\right)\right\|_2 \\
&\leq \max_{1\leq l\leq m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}_0^{(l)}\boldsymbol{Q}_0^{(l)}\right)\right\|_2 + \max_{1\leq l\leq m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_0^{(l)}\boldsymbol{Q}_0^{(l)} - \boldsymbol{X}^\natural\right)\right\|_2 \\
&\leq \max_{1\leq l\leq m}\left\|\boldsymbol{a}_l\right\|_2\left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}_0^{(l)}\boldsymbol{Q}_0^{(l)}\right\| + \max_{1\leq l\leq m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_0^{(l)}\boldsymbol{Q}_0^{(l)} - \boldsymbol{X}^\natural\right)\right\|_2 \\
&\lesssim \sqrt{n}\cdot\max_{1\leq l\leq m}\left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}_0^{(l)}\boldsymbol{Q}_0^{(l)}\right\| + \sqrt{\log n}\cdot\max_{1\leq l\leq m}\left\|\boldsymbol{X}_0^{(l)}\boldsymbol{Q}_0^{(l)} - \boldsymbol{X}^\natural\right\|_2 &&(76)\\
&\lesssim \sqrt{n}\cdot\kappa\max_{1\leq l\leq m}\left\|\boldsymbol{X}_0\boldsymbol{Q}_0 - \boldsymbol{X}_0^{(l)}\boldsymbol{R}_0^{(l)}\right\| + \sqrt{\log n}\cdot\max_{1\leq l\leq m}\left\|\boldsymbol{X}_0^{(l)}\boldsymbol{Q}_0^{(l)} - \boldsymbol{X}^\natural\right\|_2 \\
&\lesssim \sqrt{n}\cdot\kappa\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\kappa\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} + \sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}} &&(77)\\
&\lesssim \sqrt{\log n}\cdot\frac{\sigma_r^2\left(\boldsymbol{X}^\natural\right)}{\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}},
\end{aligned}$$

where (76) follows from Lemma 10 and Lemma 11, and (77) follows from (28b).

G.4 Finishing the Proof

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

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 742–769, 2018.

[8] M. Soltani and C. Hegde, "Towards provable learning of polynomial neural networks using low-rank matrix estimation," in *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, 2018.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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," *accepted to Foundations of Computational Mathematics*, pp. 1–182, 2019.

[15] Y. Zhong and N. Boumal, "Near-optimal bounds for phase synchronization," *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 989–1016, 2018.

[16] Y. Chen, J. Fan, C. Ma, and K. Wang, "Spectral method and regularized MLE are both optimal for top-k ranking," *Annals of statistics*, vol. 47, no. 4, pp. 2204–2235, 2019.

[17] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu, "On robust regression with high-dimensional predictors," *Proceedings of the National Academy of Sciences*, vol. 110, no. 36, pp. 14 557–14 562, 2013.

[18] Y. Chen, J. Fan, C. Ma, and Y. Yan, "Inference and uncertainty quantification for noisy matrix completion," *Proceedings of the National Academy of Sciences*, vol. 116, no. 46, pp. 22 931–22 937, 2019.

[19] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan, "Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization," *arXiv preprint arXiv:1902.07698*, 2019.

[20] L. Ding and Y. Chen, "The leave-one-out approach for matrix completion: Primal and dual analysis," *arXiv preprint arXiv:1803.07554*, 2018.

[21] J. Chen, D. Liu, and X. Li, "Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization," *arXiv preprint arXiv:1901.06116*, 2019.

[22] N. El Karoui, "On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators," *Probability Theory and Related Fields*, vol. 170, no. 1-2, pp. 95–175, 2018.

[23] C. Cai, G. Li, H. V. Poor, and Y. Chen, "Nonconvex low-rank tensor completion from noisy data," in *Advances in Neural Information Processing Systems*, 2019, pp. 1861–1872.

[24] C. Cai, G. Li, Y. Chi, H. V. Poor, and Y. Chen, "Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees," *arXiv preprint arXiv:1910.04267*, 2019.

[25] J. Dong and Y. Shi, "Nonconvex demixing from bilinear measurements," *IEEE Transactions on Signal Processing*, vol. 66, no. 19, pp. 5152–5166, 2018.

[26] Y. Chen, J. Fan, C. Ma, and Y. Yan, "Bridging convex and nonconvex optimization in robust PCA: Noise, outliers, and missing data," *arXiv preprint arXiv:2001.05484*, 2020.

[27] T. Cai and A. Zhang, "ROP: Matrix recovery via rank-one projections," *The Annals of Statistics*, vol. 43, no. 1, pp. 102–138, 2015.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5239–5269, 2019.

[34] 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.

[35] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, "A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 5164–5198, 2017.

[36] M. Soltanolkotabi, "Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization," *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 2374–2400, 2019.

[37] 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, 2017.

[38] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR. org, 2016, pp. 964–973.

[39] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Applied and computational harmonic analysis*, vol. 47, no. 3, pp. 893–934, 2019.

[40] J.-F. Cai, T. Wang, and K. Wei, "Spectral compressed sensing via projected gradient descent," *SIAM Journal on Optimization*, vol. 28, no. 3, pp. 2625–2653, 2018.

[41] K. Wei, "Solving systems of phaseless equations via Kaczmarz methods: A proof of concept study," *Inverse Problems*, vol. 31, no. 12, p. 125008, 2015.

[42] Y. S. Tan and R. Vershynin, "Phase retrieval via randomized Kaczmarz: Theoretical guarantees," *Information and Inference: A Journal of the IMA*, vol. 8, no. 1, pp. 97–123, 2019.

[43] H. Jeong and C. S. Güntürk, "Convergence of the randomized kaczmarz method for phase retrieval," *arXiv preprint arXiv:1706.10291*, 2017.

[44] J. Ma, J. Xu, and A. Maleki, "Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2-regularization," *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3600–3629, 2019.

[45] Y. Chi and Y. M. Lu, "Kaczmarz method for solving quadratic equations," *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1183–1187, 2016.

[46] 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.

[47] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon, "Recovery guarantees for one-hidden-layer neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, pp. 4140–4149.

[48] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval," *Mathematical Programming*, vol. 176, no. 1-2, pp. 5–37, 2019.

[49] 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.

[50] 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.

[51] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *Compressed Sensing, Theory and Applications*, pp. 210 – 268, 2012.

[52] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, vol. 28, no. 5, pp. 1302–1338, 2000.

[53] R. Vershynin, "High dimensional probability," 2017.

[54] W. Schudy and M. Sviridenko, "Concentration and moment inequalities for polynomials of independent random variables," *arXiv preprint arXiv:1104.4997*, 2011.

[55] J. M. Ten Berge, "Orthogonal procrustes rotation for two or more matrices," *Psychometrika*, vol. 42, no. 2, pp. 267–276, 1977.

[56] S. Lang, *Real and functional analysis.* Springer Science & Business Media, 2012, vol. 142.

[57] 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.

[58] 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.