Nonconvex Matrix Completion: From Geometric Analysis to Algorithmic Analysis

By

JI CHEN
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

APPLIED MATHEMATICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

____
Professor Xiaodong Li, Chair

____
Professor Naoki Saito

____
Professor Thomas Strohmer

Committee in Charge

2020

To my family

Contents

Nonconvex Matrix Completion: From Geometric Analysis to Algorithmic Analysis

Abstract

Techniques of matrix completion aim to impute a large portion of missing entries in a data matrix through a small portion of observed ones, with broad machine learning applications including collaborative filtering, system identification, global positioning, etc. This dissertation aims to analyze the nonconvex matrix problem from geometric and algorithmic perspectives.

The first part of the dissertation, i.e., Chapter 2 and 3, focuses on analyzing the nonconvex matrix completion problem from the geometric perspective. Geometric analysis has been conducted on various low-rank recovery problems including phase retrieval, matrix factorization and matrix completion in recent few years. Taking matrix completion as an example, with assumptions on the underlying matrix and the sampling rate, all the local minima of the nonconvex objective function were shown to be global minima, i.e., nonconvex optimization can recover the underlying matrix exactly. In Chapter 2, we propose a model-free framework for nonconvex matrix completion: We characterize how well local-minimum based low-rank factorization approximates the underlying matrix without any assumption on it. As an implication, a corollary of our main theorem improves the state-of-the-art sampling rate required for nonconvex matrix completion to rule out spurious local minima.

In practice, additional structures are usually employed in order to improve the accuracy of matrix completion. Examples include subspace constraints formed by side information in collaborative filtering, and skew symmetry in pairwise ranking. Chapter 3 performs a unified geometric analysis of nonconvex matrix completion with linearly parameterized factorization, which covers the aforementioned examples as special cases. Uniform upper bounds for estimation errors are established for all local minima, provided assumptions on the sampling rate and the underlying matrix are satisfied.

The second part of the dissertation (Chapter 4) focuses on algorithmic analysis of nonconvex matrix completion. Row-wise projection/regularization has become a widely adapted assumption due to its convenience for analysis, though it was observed to be unnecessary in numerical simulations. Recently the gap between theory and practice has been overcome for positive semidefinite

matrix completion via so called leave-one-out analysis. In Chapter 4, we extend the leave-one-out analysis to the rectangular case, and more significantly, improve the required sampling rate for convergence guarantee.

Acknowledgments

My advisor, Professor Xiaodong Li, has mentored me for five years during my graduate studies. His knowledge, intuition, and patience are the biggest supporters of my Ph.D. degree. I would like to thank Professor Naoki Saito and Professor Thomas Strohmer for being my dissertation committee, Professor Javier Arsuaga and Professor Yueyue Fan for attending my qualifying exam.

For my first project, I would like to acknowledge Taisong Jing for pointing us to the reference [**Bha13**]. I would like to acknowledge the anonymous reviewers of our paper [**CL19**] for their insightful comments, which urges us to improve our main results. All the authors of [**CL19**] contributed to the paper equally. For my second project, I would like to acknowledge my co-author Dekai Liu, it was really a nice experience to work with Dekai. All the authors of [**CLL19**] contributed to the paper equally. For my third project, I would like to acknowledge Professor Zongming Ma. All the authors of [**CLM20**] contributed to the paper equally. I would like to acknowledge the support from NSF CAREER Award DMS-184857 from my advisor Professor Xiaodong Li and NSF HDR TRIPODS grant CCF-1934568.

Thank Bohan Zhou, Yunshen Zhou, Yiqun Shao as well as Jiaxiang Li for being my officemate for years. Finally, thank my grandmother Yicheng Chen, my grandfather Guowei Ma, my mother Jiejie Ma, and my father Handong Chen for supporting me during my life.

CHAPTER 1

Introduction

Matrix completion techniques aim to predict missing entries in a data matrix from observed ones. Applications include *collaborative filtering* [**RS05**, **CR09**], in which unobserved user-item ratings are predicted with the available ones; *Global positioning in sensor networks* [**SY07**, **Sin08**, **OMK10**, **JM13**], in which some of the distances between sensors are unknown due to limitations such as power restrictions of the sensors; And *system identification* [**LV10**, **LHV13**], etc.

In most high-dimensional problems, low-complexity structures have to be imposed in order to perform non-trivial learning. In matrix completion algorithms, the low-complexity structure is the low-rankness of the ground truth. By imposing nuclear norm regularization in order to recover low-rank structures [**RFP10**], convex optimization methods have been widely used in the literature of matrix completion. By solving a nuclear norm minimization problem, [**CR09**] showed that exact recovery is possible for matrix completion. [**CT10**] gave a lower bound of number of entries required for exactly recovering the underlying matrix by any method. By improving the required sampling complexity from quadratic dependence on the rank of underlying matrix to linear dependence, [**CT10**] matched the aforementioned information theoretic lower bound on the dependence of rank. By adapting technologies including a recursive process designed in [**GLF$^+$10**] (referred as "golfing scheme" in [**Gro11**]) to the matrix completion problem, [**Rec11**] presented a much simplified proof comparing to prior works. Almost at the same time, by using golfing scheme, [**Gro11**] showed that exact recovery of Hermitian matrix can be achieved for any given matrix basis. Instead of incoherence conditions introduced in [**CR09**], [**NW12**] considered the matrix completion problem based on the measure of spikiness and low-rankness of matrices. [**MHT10**] considered a reformulated nuclear norm minimization problem in Lagrange form. Nuclear norm penalized estimator was also studied in [**KLT11**], and elastic penalty was considered in [**SZ12**]. Nuclear norm regularization has also been studied for robust principle component analysis, e.g., [**CLMW11**, **HKZ11**], etc.

Though convex optimization methods could have near-optimal theoretical guarantees for matrix completion with assumptions on incoherence conditions, they are in general unscalable to large data matrices whose dimensions are as high as hundreds of thousands. In contrast, nonconvex optimization methods have been proposed and analyzed in the literature due to computational convenience. Nonconvex optimization methods [**RS05**] based on low-rank factorization can reduce memory and computation costs and avoid iterative singular value decompositions, thereby much more scalable to large datasets than convex optimization. In [**KMO10a**, **KMO10b**], a nonconvex optimization has been proposed, in which the constraint is the Cartesian product of two Grassmann manifolds. With assumptions on the sampling complexity in comparison with the rank, incoherence and condition number of the matrix to complete, a method of alternating gradient descent with initialization is proven to converge to the global minimum and recover the low rank matrix accurately. Singular value projections (SVP) was employed in [**JMD10**] to recover the underlying low-rank matrix. Alternating minimization via the low-rank factorization $\boldsymbol{M} \approx \boldsymbol{X}\boldsymbol{Y}^\top$ was analyzed in [**JNS13**] provided independent samples are used to update $\boldsymbol{X}$ and $\boldsymbol{Y}$ in each step of the iteration. Their theoretical results were later improved and extended in [**Har14**, **HW14**, **ZWL15**].

Matrix completion algorithms with brand new samples in each iteration may be impractical given the observed entries are usually highly limited. Instead, gradient descent for a row-wise regularized nonconvex optimization was shown in [**SL16**] to converge to the global minimum and thereby recover the low-rank matrix, provided there hold assumptions on the sampling complexity and the low-rank matrix. In [**CW15**, **ZL16**, **YPCC16**], instead of introducing row-wise penalty within the nonconvex objective function, row-wise projection was employed for each iteration of gradient descent. With spectral initialization, projected gradient descent was guaranteed to converge to the global minimum geometrically. The row-wise regularization or projection has become a standard assumption for nonconvex matrix completion ever since, given they can explicitly control the $\ell_{2,\infty}$ norms of $\boldsymbol{X}$ and $\boldsymbol{Y}$, i.e., $\max_i \|\boldsymbol{X}_{i,\cdot}\|_2$ and $\max_i \|\boldsymbol{Y}_{i,\cdot}\|_2$, which is crucial in the theoretical analysis. Here $\boldsymbol{X}_{i,\cdot}$ denotes the i-th row of $\boldsymbol{X}$. However, it has been observed that row-wise regularization is numerically inactive in general, see, for example, [**Sun15**]. The gap between theory and practice was first overcame in [MWCC18], in which the matrix to complete is assumed to be positive semidefinite[1].

In summary, aforementioned theoretical analysis [SL16, CW15, ZL16, YPCC16, MWCC18] follow a two-step argument: First, with spectral initialization, the initial value can be shown to be located within a region close to the global minimum; Second, by analyzing the local geometry near global minimum, iterative methods such as gradient descent can be shown to converge geometrically.

Besides algorithmic analysis for nonconvex matrix completion, [GLM16, GJZ17] have been dedicated to the geometric analysis: Instead of consider local geometry near the global minimum, they analyzed the global geometry of the nonconvex objective function. With assumptions on the underlying low-rank matrix and sampling rate, [GLM16, GJZ17] showed that the regularized nonconvex objective function has no spurious local minima. That is, any local minimum is the global minimum, and thereby nonconvex methods recover the underlying low-rank matrix. Given the fact that under mild assumption, gradient descent can avoid strict saddle points almost surely [LSJR16, LPP+17, PP17, JGN+17], the no-spurious-local-minima result ensures that gradient descent with random initialization can converge to the global minimum.

It is also noteworthy that besides matrix completion, algorithmic and geometric nonconvex analyses have also been conducted for other low-rank recovery problems, such as phase retrieval [CLS15, SQW18, CLM16, CCFM19], matrix sensing [ZL15, TBS+15, LMZ18], blind deconvolution [LLSW19], etc.

1.1. Global geometry of nonconvex matrix completion, a model-free framework

To put it in the mathematical terms, the (positive semidefinite) matrix completion problem can be stated as follows: Let $\boldsymbol{M}$ be a $n \times n$ positive semidefinite matrix, and we would like to estimate the whole matrix from a small proportion of observed entries. To be specific, let $\Omega \subset [n] \times [n]$ be the index set that supports all observed entries, where $[n] \coloneqq \{1, 2, \ldots, n\}$. The observation is represented by $\mathcal{P}_\Omega(\boldsymbol{M})$, where the operator $\mathcal{P}_\Omega(\cdot)$ preserves the entries on Ω while changes the entries on Ω^c into zeros.

[1] Throughout this dissertation, in order to avoid confusion, positive semidefinite matrix is always assumed to be symmetric (or Hermitian in complex setup).

Note that any rank-r positive semidefinite matrix can be parameterized through the factorization $\boldsymbol{X}\boldsymbol{X}^\top$, where $\boldsymbol{X}$ has r columns. With this parameterization, the regularized least squares fitting proposed and further analyzed in [GLM16, GJZ17] is

$$f_{\text{psd}}(\boldsymbol{X}) \coloneqq \frac{1}{2p}\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M})\|_F^2 + \lambda G_\alpha(\boldsymbol{X}), \tag{1.1}$$

where

$$G_\alpha(\boldsymbol{X}) \coloneqq \sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^4. \tag{1.2}$$

Here $\boldsymbol{X}$ is an n-by-r matrix, $\boldsymbol{X}_{i,\cdot}$ denotes i-th row of $\boldsymbol{X}$, and λ and α are two tuning parameters.. The sampling rate p is usually unknown but is almost identical to its empirical version $|\Omega|/n^2$. Given (1.2) has been used in [GLM16, GJZ17], the reason why we introduced a fourth order penalty (1.2) here mainly consists of two parts: First, the fourth order penalty is twice continuously differentiable, which makes it possible for us to analyze the second order optimality condition of the objective function. Second, comparing to prior work [GLM16, GJZ17], technically speaking, the fourth term plays a crucial rule in our analysis, which we will see later in Chapter 2.

This optimization is obviously nonconvex, so standard optimization methods, such as gradient descents, may be attracted to local minima. A series of works in the literature, such as [GLM16, GJZ17], aimed to understand the nonconvex geometry of (1.1). In particular, people are interested in figuring out the conditions on the ground-truth low rank matrix $\boldsymbol{M}$ as well as the sampling rate of Ω, under which any local minimum $\widehat{\boldsymbol{X}}$ of (1.1) leads to an accurate estimate of $\boldsymbol{M}$ through $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top$. For example, [GLM16, GJZ17] showed that any local minimum $\widehat{\boldsymbol{X}}$ of (1.1) yields $\boldsymbol{M} = \widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top$, as long as $\boldsymbol{M}$ is exactly rank-r, the condition number of $\boldsymbol{M}$ is well-bounded, the incoherence parameter of the eigenspace of $\boldsymbol{M}$ is well-bounded, and the sampling rate is greater than a function of aforementioned quantities. However, in real applications, the aforementioned assumptions may not be satisfied. For example, it is not realistic to estimate the exact rank of underlying $\boldsymbol{M}$; the condition number and incoherence parameter can be extremely large due to small perturbations to $\boldsymbol{M}$ and rank mismatching caused by inaccurate estimation of rank, etc.

In order to address the aforementioned problems, our paper [CL19], which reduced to first three sections of Chapter 2, studied the theoretical properties of $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top$ with no assumptions on $\boldsymbol{M}$. Due to the fact that we do not assume $\boldsymbol{M}$ is exactly rank-r, there are actually two questions of interest: how close $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top$ is from $\boldsymbol{M}$, and how close $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top$ is from $\boldsymbol{M}_r$, the best rank-r approximation of $\boldsymbol{M}$ by spectral truncation. In comparison to [GLM16, GJZ17], our main contributions to be introduced in the next chapter include the following:

- Without assumptions imposed on $\boldsymbol{M}$ regarding its rank, eigenvalues and eigenvectors, our main result Theorem 2.1.2 are able to characterize how well any local-minimum based rank-r factorization $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top$ approximates $\boldsymbol{M}$ or $\boldsymbol{M}_r$. The sampling rate is only required to satisfy $p \geqslant C \log n/n$ for some absolute constant C. Therefore, for matrix completion applications, our framework provides more suitable guidelines than [GLM16, GJZ17]. In fact, the condition number and incoherence parameter of the matrix to complete may not satisfy the strong assumptions in [GLM16, GJZ17].
- When $\boldsymbol{M}$ is assumed to be exactly low-rank as in [GLM16, GJZ17], Corollary 2.1.3 improves the state-of-the-art no-spurious-local-minima results in [GLM16, GJZ17] for exact nonconvex matrix completion in terms of sampling rates. To be specific, assuming both condition numbers and incoherence parameters are both on the order of $O(1)$, our result improves the result in [GJZ17] from $\widetilde{O}(r^4/n)$ to $\widetilde{O}(r^2/n)$. Here $\widetilde{O}(\cdot)$ indicates that we ignore the logarithms.
- Theorem 2.1.2 also implies the conditions under which the nonconvex optimization (1.1) yields good low-rank approximation of $\boldsymbol{M}$ in the cases of large condition numbers, high incoherence parameters, or rank-mismatching.

On the other hand, [CL19] benefits from [GLM16, GJZ17] in various aspects. In order to characterize the properties of any local minimum $\widehat{\boldsymbol{X}}$, we follow the idea in [GJZ17] to combine the first and second order conditions of local minima linearly to construct an auxiliary function, denoted as $K(\boldsymbol{X})$ in this dissertation. If $\boldsymbol{M}$ is exactly rank-r and its eigenvalues and eigenvectors are well-bounded, [GJZ17] showed that $K(\boldsymbol{X}) \leqslant 0$ for all $\boldsymbol{X}$ as long as the sampling rate is large enough. This argument can be employed to prove that there are no spurious local minima.

However, $K(\boldsymbol{X}) \leqslant 0$ can be shown to hold for all $\boldsymbol{X}$ only if strong assumptions are imposed on $\boldsymbol{M}$. Therefore, we instead focus on analyzing the inequality $K(\widehat{\boldsymbol{X}}) \geqslant 0$ directly in the model-free manner, noting that $\widehat{\boldsymbol{X}}$ denotes a local minimum. Among a few novel technical ideas, the success of such model-free analysis relies crucially on the deterministic inequality (Lemma 2.3.5) that controls the difference between the function $K(\boldsymbol{X})$ and its population version $\mathbb{E}[K(\boldsymbol{X})]$ for any fixed $\boldsymbol{X}$.

Though most of the nonconvex matrix completion literature focus on the uniform sampling model, the model-free framework introduced in [CL19] enables us to derive an uniform approximation result: For any fixed sampling pattern Ω, Theorem 2.4.1 characterizes how well any local-minimum based rank-r factorization approximates the ground-truth. As an interesting byproduct, in Section 2.4, we find that minimizing (1.1) can still recover the ground-truth $\boldsymbol{M}$ very well when the uniformness of Ω is slightly violated. Furthermore, as a natural extention of model-free framework in [CL19], in Section 2.5, a model-free local minima analysis is conducted on nonconvex rectangular matrix completion with the following objective function,

$$f_{\text{rect}}(\boldsymbol{X}, \boldsymbol{Y}) := \frac{1}{2p}\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M})\|_F^2 + \frac{1}{8}\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}\|_F^2 + \lambda(G_\alpha(\boldsymbol{X}) + G_\alpha(\boldsymbol{Y})), \tag{1.3}$$

where $\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}$.

In summary, the theoretical analysis within [CL19] follows the framework of local minimum analysis for nonconvex optimization in the literature. For example, [BH89] has described the nonconvex landscape of the quadratic loss for PCA. [LW15] studied the local minima of regularized M-estimators. [SQW18] studied the global geometry of the phase retrieval problem. The conditions for no spurious local minima have been investigated in [BNS16] and [GLM16] for nonconvex matrix sensing and completion, respectively. The global geometry of nonconvex objective functions with underlying symmetric structures, including low-rank symmetric matrix factorization and sensing, has been studied in [LWL+16]. Global geometry of rectangular matrix factorization and sensing has been studied in [ZLTW17], where the issues of under-parameterization and over-parameterization have been investigated. Similar analysis has been extended to general low-rank optimization problems in [LZT17]. Matrix factorization has been further studied in [JGN+17] with a novel geometric characterization of saddle points, and this idea was later extended in [GJZ17], where a unified geometric analysis framework is proposed to study the landscapes of nonconvex matrix sensing, matrix completion and robust PCA. More recently, [**LLZ$^+$20**] analyzed the global geometry of low-rank (rectangular) matrix recovery without regularization.

1.1.1. Applications in memory-efficient kernel PCA. Kernel PCA [**SSM98**] is a widely used nonlinear dimension reduction technique in machine learning for the purpose of redundancy removal and preprocessing before prediction, classification or clustering. The method is implemented by finding a low-rank approximation of the kernel-based Gram matrix determined by the data sample. To be concrete, let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ be a data sample of size n and dimension d, and let $\boldsymbol{M}$ be the $n \times n$ positive semidefinite kernel matrix determined by a predetermined kernel function $K(\boldsymbol{x}, \boldsymbol{y})$ in that $M_{i,j} = K(\boldsymbol{z}_i, \boldsymbol{z}_j)$. Non-centered kernel PCA with r principal components amounts to finding the best rank-r approximation of $\boldsymbol{M}$.

However, when the sample size is large, the storage of the kernel matrix itself becomes challenging. Consider the example when the dimension d is in thousands while the sample size n is in millions. The memory cost for the data matrix is $d \times n$ and thus in billions, while the memory cost for the kernel matrix $\boldsymbol{M}$ is in trillions! On the other hand, if not storing $\boldsymbol{M}$, the implementation of standard iterative algorithms of SVD will involve one pass of computing all entries of $\boldsymbol{M}$ in each iteration, usually with formidable computational cost $O(n^2 d)$. Therefore, a natural question arises: *How to find low-rank approximations of $\boldsymbol{M}$ memory-efficiently?*

The following two are among the most well-known memory-efficient kernel PCA methods in the literature. One is Nyström method [**WS01**], which amounts to generating random partial columns of the kernel matrix, then finding a low-rank approximation based on generated columns. In order to generate random partial columns, uniform sampling without replacement was employed in [**WS01**], and different sampling strategies were proposed later, e.g., [**DM05**]. The method is convenient in implementation and efficient in both memory and computation, but relatively unstable in terms of approximation errors as will be shown in Section 2.2.

Another popular approach is stochastic approximation, e.g., Kernel Hebbian Algorithm (KHA) [**KFS05**], which is memory-efficient and approaches the exact principal component solution as the number of iterations goes to infinity with appropriately chosen learning rate [**KFS05**]. However, based on our experience, the method usually requires careful tuning of learning rates even for very slow convergence.

It is also worth mentioning that the randomized one-pass algorithm discussed in, e.g., [HMT11], where the theoretical properties of a random-projection based low-rank approximation method were fully analyzed. However, although the one-pass algorithm does not require the storage of the whole matrix $\boldsymbol{M}$, in kernel PCA one still needs to compute every entry of $\boldsymbol{M}$, which typically requires $O(n^2 d)$ computational complexity for kernel matrix.

As a result, we aim at finding a memory-efficient method as an alternative to the aforementioned approaches. In particular, we are interested in a method with desirable empirical properties: memory-efficient, no requirement on one or multiple passes to compute the complete kernel matrix, no requirement to tune the parameters carefully, and yielding stable results. To this end, we propose the following method based on entries sampling and nonconvex optimization: In the first step, Ω is generated to follow an Erdős-Rényi random graph with parameter p later specified in Model 2.1.1, and then a partial kernel matrix $\mathcal{P}_\Omega(\boldsymbol{M})$ is generated in that $M_{i,j} = K(\boldsymbol{z}_i, \boldsymbol{z}_j)$ for $(i,j) \in \Omega$. In the second step, the nonconvex optimization is implemented through gradient descent. Any local minimum of (1.1), $\widehat{X}$, is a solution of approximate kernel PCA in that $\boldsymbol{M} \approx \widehat{X}\widehat{X}^\top$.

To store the index set Ω and the sampled entries of $\boldsymbol{M}$ on Ω, the memory cost in the first step is $O(|\Omega|)$, which is comparable to the memory cost $O(nr + |\Omega|)$ in the second step. As to the computational complexity, besides the generation of Ω, the computational cost in the first step is typically $O(|\Omega| d)$, e.g., when the radial kernels or polynomial kernels are employed. This could be dominating the per-iteration computational complexity $O(|\Omega| r)$ in the second step when the target rank r is much smaller than the original dimension d.

Partial entries sampling plus nonconvex optimization has been proposed in the literature for scalable robust PCA and matrix completion [YPCC16]. However, to the best of our knowledge, [CL19] is the first to apply such an idea to memory-efficient kernel PCA. Moreover, the underlying signal matrix is assumed to be exactly low-rank in [YPCC16] while we make no assumptions on the positive semidefinite kernel matrix $\boldsymbol{M}$. Entry-sampling has been proposed in [AMS02,AM07] for scalable low-rank approximation. In particular, it is used to speed up kernel PCA in [AMS02], but spectral methods are subsequently employed after entries sampling as opposed to nonconvex optimization. It is also noteworthy that matrix completion techniques have been applied to certain kernel matrices when it is costly to generate each single entry [Gra02,PC10], wherein the proposed methods are not memory-efficient. In contrast, our method is memory-efficient in order to serve a different purpose.

1.2. Global geometry of nonconvex parameterized linear models

In practice, additional structures beyond low-rankness have been employed to improve efficiency and to reduce sample complexity for matrix completion. In collaborative filtering, for instance, side information about items and individuals has been used in the literature as subspace constraints for the matrix to complete [XJZ13, YZJ+13, Che15, EYW18, JD13, SCH+16]. Another example is pairwise ranking, where skew-symmetric structures are imposed in the implementation of matrix completion [JLYY11], see, also, [GL11, Cha15].

Interestingly, in [CLM20], which builds the main body of Chapter 3 in this dissertation, we observe that both examples, i.e., low-rank matrices with subspace constraints and skew-symmetric low-rank matrices, can be represented in the form $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{\xi})\boldsymbol{Y}(\boldsymbol{\xi})^\top$, where both factors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are linear and homogeneous in parameters $\boldsymbol{\xi} \in \mathbb{R}^d$. The details underlying the foregoing observations are as follows.

- Suppose $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ is known to be constrained in some pre-specified column and row spaces, with dimensions s_1 and s_2, respectively. Let $\widetilde{\boldsymbol{U}}$(and $\widetilde{\boldsymbol{V}}$) be a $n_1 \times s_1$(and $n_2 \times s_2$) matrix whose columns form an orthogonal basis for the given column(or row) space constraint for $\boldsymbol{M}$. Given the rank of $\boldsymbol{M}$, we know there must exist some (not necessarily unique) $\boldsymbol{\Xi}_A \in \mathbb{R}^{s_1 \times r}$ and $\boldsymbol{\Xi}_B \in \mathbb{R}^{s_2 \times r}$, such that

$$\boldsymbol{M} = \left(\widetilde{\boldsymbol{U}}\boldsymbol{\Xi}_A\right)\left(\widetilde{\boldsymbol{V}}\boldsymbol{\Xi}_B\right)^\top.$$

  Denote by $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B)$ a $(s_1 + s_2)r$-dimensional vector that contains all entries in $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$ (e.g., in the lexicographic order), and define the two linear mappings:

$$\boldsymbol{X}(\boldsymbol{\theta}) = \widetilde{\boldsymbol{U}}\boldsymbol{\Theta}_A \in \mathbb{R}^{n_1 \times r} \quad \text{and} \quad \boldsymbol{Y}(\boldsymbol{\theta}) = \widetilde{\boldsymbol{V}}\boldsymbol{\Theta}_B \in \mathbb{R}^{n_2 \times r}. \tag{1.4}$$

  Then the above parameterized factorization becomes $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{\xi})\boldsymbol{Y}(\boldsymbol{\xi})^\top$ with

$$\boldsymbol{\xi} = \mathrm{vec}(\boldsymbol{\Xi}_A, \boldsymbol{\Xi}_B).$$

- If $\boldsymbol{M}$ is a $n \times n$ rank-r skew-symmetric matrix (which implies that r is even), by the Youla decomposition [You61], it can be represented (not necessarily uniquely) as

  $$\boldsymbol{M} = \boldsymbol{\Xi}_A \boldsymbol{\Xi}_B^\top - \boldsymbol{\Xi}_B \boldsymbol{\Xi}_A^\top,$$

  where $\boldsymbol{\Xi}_A, \boldsymbol{\Xi}_B \in \mathbb{R}^{n \times \frac{r}{2}}$. Again, denote by $\boldsymbol{\theta} = \operatorname{vec}(\boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B)$ a (nr)-dimensional vector that contains all entries in $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$, and define the linear and homogeneous mappings

  $$\boldsymbol{X}(\boldsymbol{\theta}) = [\boldsymbol{\Theta}_A, -\boldsymbol{\Theta}_B] \in \mathbb{R}^{n \times r} \quad \text{and} \quad \boldsymbol{Y}(\boldsymbol{\theta}) = [\boldsymbol{\Theta}_B, \boldsymbol{\Theta}_A] \in \mathbb{R}^{n \times r}. \tag{1.5}$$

  We also have the factorization $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{\xi})\boldsymbol{Y}(\boldsymbol{\xi})^\top$ with $\boldsymbol{\xi} = \operatorname{vec}(\boldsymbol{\Xi}_A, \boldsymbol{\Xi}_B)$.

We are interested in recovering $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{\xi})\boldsymbol{Y}(\boldsymbol{\xi})^\top$ through the noisy observation $\mathcal{P}_\Omega(\boldsymbol{M} + \boldsymbol{N})$ via a nonconvex optimization similar to (1.3):

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\mathcal{P}_\Omega(\boldsymbol{X}(\boldsymbol{\theta})\boldsymbol{Y}(\boldsymbol{\theta})^\top - \boldsymbol{M} - \boldsymbol{N})\|_F^2 + pen(\boldsymbol{\theta}), \tag{1.6}$$

where $pen(\boldsymbol{\theta})$ is a penalty function that will be specified in (3.2). This optimization problem is nonconvex in $\boldsymbol{\theta}$. So it is natural to ask whether we can study the nonconvex geometry for (1.6) as [GLM16, GJZ17, CL19] did for the vanilla matrix completion problem (1.3).

As an initial step for this general question, in [CLM20], two key assumptions have been made on the parameterization $(\boldsymbol{X}(\boldsymbol{\theta}), \boldsymbol{Y}(\boldsymbol{\theta}))$ and the ground truth $\boldsymbol{M}$. The first assumption is that $\boldsymbol{X}(\boldsymbol{\theta})$ and $\boldsymbol{Y}(\boldsymbol{\theta})$ are linear and homogeneous in $\boldsymbol{\theta}$ as we required previously. The second assumption, referred to as *correlated parametric factorization*, is not easy to explain in non-mathematical terms, and its formal definition will be introduced in Section 3.1.2. This rather sophisticated assumption holds for various examples of parameterized low-rank factorization including low-rank matrices with subspace constraints (1.4) and low-rank skew-symmetric matrices (1.5). The verifications of the correlated parametric factorization assumption in these two examples will be given in Sections 3.3.1 and 3.4.1, respectively.

Under these assumptions, we will show in Section 3.1.3 that we can indeed analyze the nonconvex geometry for (1.6) in a comparable way to [GLM16, GJZ17, CL19] did for (1.3). To be specific, uniformly for all low-rank recovery $\widehat{\boldsymbol{M}} := \boldsymbol{X}(\hat{\boldsymbol{\xi}})\boldsymbol{Y}(\hat{\boldsymbol{\xi}})^\top$ with any local minimum $\hat{\boldsymbol{\xi}}$ of the nonconvex optimization (1.6), unified upper bounds are established for the estimation error

$\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_F^2$, as long as the sampling rate satisfies conditions that depends on rank, condition number, and eigenspace incoherence parameter of $\boldsymbol{M}$. Moreover, as corollaries, our main result implies local-minimum based estimation error bounds for the problems of subspace-constrained and skew-symmetric matrix completion.

1.3. Nonconvex Rectangular Matrix Completion via Gradient Descent without $\ell_{2,\infty}$ Regularization

As aforementioned, the $\ell_{2,\infty}$-norm regularization or projection has become a standard assumption for nonconvex matrix completion. Consider [ZL16] as an example. By assuming that $\operatorname{rank}(\boldsymbol{M}) = r$ is known and that Ω satisfies i.i.d. Bernoulli model with parameter p, i.e., Model 2.5.1, the nonconvex optimization

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n_1\times r},\boldsymbol{Y}\in\mathbb{R}^{n_2\times r}} f(\boldsymbol{X},\boldsymbol{Y}) := \frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right\|_F^2 + \frac{1}{8}\left\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}\right\|_F^2 \tag{1.7}$$

was proposed there to recover $\boldsymbol{M}$ through $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{Y}}^\top$. In order to show that (1.7) is able to recover $\boldsymbol{M}$ exactly, a projected gradient descent algorithm was proposed in [ZL16] where the projection depending on unknown parameters is intended to control the $\ell_{2,\infty}$ norms of the updates of $\boldsymbol{X}$ and $\boldsymbol{Y}$. It was shown that with spectral initialization, projected gradient decent is guaranteed to converge to the global minimum and recover $\boldsymbol{M}$ exactly, provided the sampling rate satisfies $p \geqslant C\mu r^2\kappa^2 \max(\mu, \log(n_1 \vee n_2))/(n_1 \wedge n_2)$ [ZL16]. Here μ is the incoherence parameter introduced in [CR09], κ is the condition number of the rank-r matrix $\boldsymbol{M}$, i.e., the ratio between the largest and smallest nonzero singular values of $\boldsymbol{M}$, and C is an absolute constant. On the other hand, it has also been pointed out in [ZL16] that the vanilla gradient descent without $\ell_{2,\infty}$-norm projection is observed to recover $\boldsymbol{M}$ exactly in simulations.

Similar $\ell_{2,\infty}$-norm regularizations have also been used in other related works, see, e.g., [CW15, YPCC16, WZG17], and a crucial question is how to control the $\ell_{2,\infty}$-norms of the updates of $\boldsymbol{X}$ and $\boldsymbol{Y}$ without explicit regularization that involves extra tuning parameters. This issue has been initiatively addressed in [MWCC18], in which the matrix to complete is assumed to be positive semidefinite, and the nonconvex optimization (1.7) is thereby reduced to

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}\right)\right\|_F^2. \tag{1.8}$$

[**MWCC18**] is focused on analyzing the convergence of vanilla gradient descent for (1.8). In particular, the leave-one-out technique well known in the regression analysis [**EKBB$^+$13**] is employed in order to control the $\ell_{2,\infty}$-norms of the updates of $\boldsymbol{X}$ in each step of iteration without explicit regularization or projection. [**MWCC18**] shows that vanilla gradient descent is guaranteed to recover $\boldsymbol{M}$, provided the sampling rate satisfies $p \geqslant C\,\mathrm{poly}(\kappa)\mu^3 r^3 \log^3 n/n$, which is somehow inferior to that in [**ZL16**]. This naturally raises several questions: Can we improve the required sampling rate from $O(\mathrm{poly}(\mu,\kappa,\log n)r^3/n)$ to $O(\mathrm{poly}(\mu,\kappa,\log n)r^2/n)$ for vanilla gradient descent without $\ell_{2,\infty}$-norm regularization? Or is explicit $\ell_{2,\infty}$-norm regularization/projection avoidable for achieving the $O(\mathrm{poly}(\mu,\kappa,\log n)r^2/n)$ sampling rate? Also, can we extend the nonconvex analysis in [**MWCC18**] to the rectangular case discussed in [**ZL16**]? Our work [**CLL19**] was intended to answer these questions. The materials included in [**CLL19**] reduce to Chapter 4 in this dissertation.

As mentioned before, [**CLL19**] aimed to establish the assumptions on the sampling complexity and the low-rank matrix $\boldsymbol{M}$, under which $\boldsymbol{M}$ can be recovered by the nonconvex optimization (1.7) via vanilla gradient descent. Roughly speaking, our main result states that as long as

$$p \geqslant C_S \mu^2 r^2 \kappa^{14} \log(n_1 \vee n_2)/(n_1 \wedge n_2)$$

with absolute constant C_S, vanilla gradient descent for (1.7) with spectral initialization is guaranteed to recover $\boldsymbol{M}$ accurately. Compared to [**MWCC18**], we have made several technical contributions including the following:

- By assuming the incoherence parameter $\mu = O(1)$ and the condition number $\kappa = O(1)$, regardless of the logarithms, the sampling rate $\widetilde{O}(r^3/n)$ in [**MWCC18**] is improved to $\widetilde{O}(r^2/(n_1 \wedge n_2))$, which is consistent with the result in [**ZL16**] where $\ell_{2,\infty}$-norm projected gradient descent is employed;
- The leave-one-out analysis for positive semidefinite matrix completion in [**MWCC18**] is extended to the rectangular case in [**CLL19**];

- In the case $\mu = O(1)$, $\kappa = O(1)$ and $r = O(1)$, the sampling rate $O(\log^3 n/n)$ in [MWCC18] is improved to $O(\log(n_1 \vee n_2)/(n_1 \wedge n_2))$ in our work, which is consistent with the result in [ZL16] where $\ell_{2,\infty}$-norm projected gradient descent is used.

To achieve these theoretical improvements and extensions, we need to make a series of modifications for the proof framework in [MWCC18]. The following technical novelties are worth highlighting, and the details are deferred to Chapter 4:

- In order to reduce the sampling rate $\widetilde{O}(r^3/n)$ in [MWCC18] to $\widetilde{O}(r^2/n)$ (assuming $\mu = O(1)$, $\kappa = O(1)$), a series of technical novelties are required. First, in the analysis of the spectral initialization for the gradient descent sequences and those for the leave-one-out sequences, $\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}) - \boldsymbol{M}\|$ is bounded in [MWCC18] basically based on Lemma 39 therein. Instead, we give tighter bounds by applying [Che15, Lemma 2] (Lemma 4.2.3 in this dissertation), and the difference is a factor of $\sqrt{r}$. Second, two pillar lemmas, Lemma 37 in [MWCC18] (restated as Lemma 4.3.1 in this dissertation) and a result in [Mat93] (restated as Lemma 4.3.3 in this dissertation), are repeatedly used in the leave-one-out analysis of [MWCC18]. We find that applying a concentration result introduced in [BJ14] and [LLR16] (restated as Lemma 2.3.6 in this dissertation) to verify the conditions in these lemmas could lead to sharper error bounds for the leave-one-out sequences. Third, also in the leave-one-out analysis, we need to modify the application of matrix Bernstein inequality in [MWCC18] in order to achieve sharper error bounds.
- In order to improve the orders of logarithms, we must improve the Hessian analysis in [MWCC18], i.e., Lemma 7 therein, and it turns out that Lemma 4.4 from [CL19] (Lemma 2.3.5 in this dissertation) and Lemma 9 from [ZL16] (Lemma C.1.1 in this dissertation) are effective to achieve this goal. These two lemmas are also effective in simplifying the proof in the Hessian analysis.

Leave-one-out analysis has been employed in [EKBB$^+$13] to establish the asymptotic sampling distribution for robust estimators in high/moderate dimensional regression. This technique has also been utilized in [AFWZ17] to control ℓ_∞ estimation errors for eigenvectors in stochastic spectral problems, with applications in exact spectral clustering in community detection without cleaning or regularization. As aforementioned, in [MWCC18], the authors have employed the leave-one-out technique to control $\ell_{2,\infty}$ estimation errors for the updates of low-rank factors in each step of gradient descent that solves (1.8). Besides matrix completion, they also show that similar techniques can be utilized to show the convergence of vanilla gradient descent in other low-rank recovery problems such as phase retrieval and blind deconvolution. Leave-one-out analysis has also been successfully employed in the study of Singular Value Projection (SVP) for matrix completion [**DC20**] and gradient descent with random initialization for phase retrieval [**CCFM19**].

Implicit regularization for gradient descent has also been studied in matrix sensing with over-parameterization. When the sampling matrices commute, it has been shown in [**GWB⁺17**] that gradient descent algorithm with near-origin starting point is guaranteed to recover the underlying low-rank matrix even under over-parameterized factorization. The result was later extended to the case in which the sensing operators satisfy certain RIP properties [**LMZ18**]. More recently, the balancing regularizer $\|\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y}\|_F^2$ has been shown to be unnecessary for rectangular matrix sensing [**MLC19**].

1.4. Notations

Throughout this dissertation, bold uppercase/lowercase characters denote matrices/vectors, respectively. For a given matrix $\boldsymbol{A}$, its (i,j)-th entry, i-th row, and j-th column are denoted as $A_{i,j}$, $\boldsymbol{A}_{i,\cdot}$, and $\boldsymbol{A}_{\cdot,j}$, respectively. Its spectral, Frobenius, and $\ell_{2,\infty}$ norms are denoted as $\|\boldsymbol{A}\|$, $\|\boldsymbol{A}\|_F$ and $\|\boldsymbol{A}\|_{2,\infty} := \max_i \|\boldsymbol{A}_{i,\cdot}\|_2^2$, respectively. Denote by $\text{colspan}(\boldsymbol{A})/\text{colspan}(\boldsymbol{A})$ the column/row space of $\boldsymbol{A}$. Deonte by $\boldsymbol{P}_{\boldsymbol{A}}$ the Euclidean projector onto $\text{colspan}(\boldsymbol{A})$. Denote $\boldsymbol{A} \succeq \boldsymbol{0}$ if $\boldsymbol{A}$ is a symmetric or Hermitian positive semidefinite matrix. For any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same dimensions, their matrix inner product is denoted as $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \text{trace}(\boldsymbol{A}^\top \boldsymbol{B}) = \sum_i \sum_j A_{i,j} B_{i,j}$, and their Hadamard/entrywise product is denoted as $\boldsymbol{A} \circ \boldsymbol{B}$ with entries $[\boldsymbol{A} \circ \boldsymbol{B}]_{i,j} = A_{i,j} B_{i,j}$. For any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\text{vec}(\boldsymbol{A}, \boldsymbol{B})$ denotes a vector consisting of all entries in $\boldsymbol{A}$ and $\boldsymbol{B}$ in some fixed order. Denote by $\boldsymbol{J}_{n_1 \times n_2}$ (or $\boldsymbol{J}$ when the dimensions are clear in the context) the $n_1 \times n_2$ matrix with all entries equal to one. Denote by $\mathsf{O}(r)$ the set of $r \times r$ orthogonal matrices. Let $n_{\min} := \min\{n_1, n_2\}$ and $n_{\max} := \max\{n_1, n_2\}$. Finally, denote by $C_1, C_2, \ldots$ and $C_v, C_c, \ldots$ fixed positive absolute constants. Furthermore, for notation convenience, in discussions we also use C to denote positive absolute constants which may vary line by line.

CHAPTER 2

Global Geometry of Nonconvex Matrix Completion, a Model-Free Framework

2.1. Model-free local minima analysis of nonconvex PSD matrix completion

Let $\boldsymbol{M}$ be an $n \times n$ positive semidefinite (PSD) matrix and let $r \ll n$ be a fixed integer. It is well known that a rank-r approximation of $\boldsymbol{M}$ can be obtained by truncating the spectral decomposition of $\boldsymbol{M}$. To be specific, let $\boldsymbol{M} = \sum_{i=1}^{n} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ be the spectral decomposition with $\sigma_1 \geqslant \ldots \geqslant \sigma_n \geqslant 0$. Then, the best rank-$r$ approximation of $\boldsymbol{M}$ is $\boldsymbol{M}_r = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$. If we denote $\boldsymbol{U}_r = [\sqrt{\sigma_1}\boldsymbol{u}_1, \ \ldots, \ \sqrt{\sigma_r}\boldsymbol{u}_r]$, then the best rank-$r$ approximation of $\boldsymbol{M}$ can be written as $\boldsymbol{M} \approx \boldsymbol{U}_r \boldsymbol{U}_r^\top$. By the well-known Eckart-Young-Mirsky Theorem [GVL12], $\boldsymbol{U}_r$ is actually the global minimum (up to rotation) to the following nonconvex optimization:

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} \|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}\|_F^2.$$

This factorization for low-rank approximation has been well-known in the literature, see, e.g., [BM03].

In this chapter, we are interested in the problem that how to find a rank-r approximation of $\boldsymbol{M}$ in the case that only partial entries are observed. Let $\Omega \subset [n] \times [n]$ be a symmetric index set, and we assume that $\boldsymbol{M}$ is only observed on the entries in Ω. For convenience of discussion, this subsampling is represented as $\mathcal{P}_\Omega(\boldsymbol{M})$ in that $\mathcal{P}_\Omega(\boldsymbol{M})_{i,j} = M_{i,j}$ if $(i,j) \in \Omega$ and $\mathcal{P}_\Omega(\boldsymbol{M})_{i,j} = 0$ if $(i,j) \notin \Omega$. We are interested in the following question,

How to find a rank-r approximation of $\boldsymbol{M}$ in a scalable manner only through $\mathcal{P}_\Omega(\boldsymbol{M})$?

We propose to find such a low-rank approximation through the following nonconvex optimization, which has been exactly proposed in [GLM16, GJZ17] for matrix completion. Denote $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times r}$. A rank-$r$ approximation of $\boldsymbol{M}$ can be found through minimizing (1.1). Following the framework of nonconvex optimization without initialization in [GLM16, GJZ17], our local-minimum based approximation for $\boldsymbol{M}$ is $\boldsymbol{M} \approx \widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top$ where $\widehat{\boldsymbol{X}}$ is any local minimum of (1.1).

Let's briefly discuss the memory and computational complexity to solve (1.1) via gradient descent. If Ω is symmetric and does not contain the diagonal entries as later specified in Model 2.1.1, the updating rule of gradient decent

$$\boldsymbol{X}^{(t+1)} = \boldsymbol{X}^{(t)} - \eta^{(t)}\nabla f(\boldsymbol{X}^{(t)}) \tag{2.1}$$

is equivalent to

$$\boldsymbol{x}_i^{(t+1)} := \boldsymbol{x}_i^{(t)} - \eta^{(t)} \left[\frac{2}{p} \sum_{j:(i,j)\in\Omega} \left(\langle \boldsymbol{x}_i^{(t)}, \boldsymbol{x}_j^{(t)} \rangle - M_{i,j} \right) \boldsymbol{x}_j^{(t)} + \frac{4\lambda}{\|\boldsymbol{x}_i^{(t)}\|_2} \left(\|\boldsymbol{x}_i^{(t)}\|_2 - \alpha \right)^3 1_{\{\|\boldsymbol{x}_i^{(t)}\|_2 \geqslant \alpha\}} \boldsymbol{x}_i^{(t)} \right],$$

where the memory cost is dominated by storing $\boldsymbol{X}^{(t)}$, $\boldsymbol{X}^{(t+1)}$, and $\boldsymbol{M}$ on Ω, which is generally $O(nr + |\Omega|)$. It is also obvious that the computational cost in each iteration is $O(|\Omega|r)$.

In this section, the following sampling scheme is employed:

Model 2.1.1 (Off-diagonal symmetric independent Ber(p) model). *Assume the index set Ω consists only of off-diagonal entries that are sampled symmetrically and independently with probability p, i.e.,*

(1) $(i,i) \notin \Omega$ for all $i = 1, \ldots, n$;

(2) For all $i < j$, sample $(i,j) \in \Omega$ independently with probability p;

(3) For all $i > j$, $(i,j) \in \Omega$ if and only if $(j,i) \in \Omega$.

Here we assume all diagonal entries are not in Ω for the generality of the formulation, although they are likely to be obtained in practice. For instance, all diagonal entries of the radial kernel matrix are ones. For any index set $\Omega \subset [n] \times [n]$, define the associated 0-1 matrix $\boldsymbol{\Omega} \in \{0,1\}^{n\times n}$ such that $\Omega_{i,j} = 1$ if and only if $(i,j) \in \Omega$. Then we can write $\mathcal{P}_\Omega(\boldsymbol{X}) = \boldsymbol{X} \circ \boldsymbol{\Omega}$ where $\circ$ denotes the Hadamard product.

Assume that the underlying positive semidefinite matrix $\boldsymbol{M}$ has the spectral decomposition

$$\boldsymbol{M} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top + \sum_{i=r+1}^{n} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top := \boldsymbol{M}_r + \boldsymbol{M}_{r+}, \tag{2.2}$$

where $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_n \geqslant 0$ are the spectrum, $\boldsymbol{u}_i \in \mathbb{R}^n$ are unit and mutually perpendicular eigenvectors. The matrix $\boldsymbol{M}_r \coloneqq \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ is the best rank-r approximation of $\boldsymbol{M}$ and $\boldsymbol{M}_{r+} \coloneqq \sum_{i=r+1}^n \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ denotes the residual part. In the case of multiple eigenvalues, the order in the eigenvalue decomposition (2.2) may not be unique. In this case, we consider the problem for any fixed order in (2.2) with the fixed $\boldsymbol{M}_r$.

THEOREM 2.1.2. *Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a positive semidefinite matrix with the spectral decomposition (2.2). Let Ω be sampled according to the off-diagonal symmetric Ber(p) model with $p \geqslant C_v \frac{\log n}{n}$ for some absolute constant C_v. Then in an event E_1 with probability $\mathbb{P}[E_1] \geqslant 1 - 2n^{-3}$, as long as the tuning parameters α and λ satisfy $100\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}} \leqslant \alpha \leqslant 200\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}}$ and $100\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p} \leqslant \lambda \leqslant 200\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}$, any local minimum $\widehat{\boldsymbol{X}} \in \mathbb{R}^{n\times r}$ of (1.1) satisfies*

$$
\begin{aligned}
\left\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}_r\right\|_F^2 \leqslant & C_1 \sum_{i=1}^r \left\{\left[C_2\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\|\boldsymbol{M}_r\|_{\ell_\infty} + C_2\sigma_{2r+1-i} - \sigma_i\right]_+\right\}^2 \\
& + C_1\frac{nr\|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2}{p}
\end{aligned}
\tag{2.3}
$$

and

$$
\begin{aligned}
\left\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\right\|_F^2 \leqslant & C_1 \sum_{i=1}^r \left\{\left[C_2\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\|\boldsymbol{M}_r\|_{\ell_\infty} + C_2\sigma_{2r+1-i} - \sigma_i\right]_+\right\}^2 \\
& + C_1\frac{nr\|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2}{p} + \|\boldsymbol{M}_{r+}\|_F^2
\end{aligned}
\tag{2.4}
$$

with C_1, C_2 absolute constants defined in the proof.

Model-free low-rank approximation from partial entries has been studied for spectral estimators in the literature. For example, under the settings of Theorem 2.1.2, the spectral low-rank approximation (denoted as $\boldsymbol{M}_{\text{approx}}$) discussed in [**KMO10b**, Theorem 1.1] is guaranteed to satisfy

$$
\|\boldsymbol{M}_{\text{approx}} - \boldsymbol{M}_r\|_F^2 \leqslant C\left\{\frac{nr\|\boldsymbol{M}_r\|_{\ell_\infty}^2}{p} + \frac{r\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+})\|^2}{p^2}\right\},
$$

with high probability. However, this cannot imply exact recovery even when $\boldsymbol{M}$ is of low rank and the sampling rate p satisfies the conditions specified in [**GJZ17**]. Similarly, the SVD-based USVT estimator introduced in [**Cha15**] does not imply exact recovery. In contrast, as will be discussed in the next subsection, Theorem 2.1.2 implies that any local minimum of (1.1) yields exact recovery of $\boldsymbol{M}$ with high probability under milder conditions than those in [GJZ17].

2.1.1. Implications in exact matrix completion. Assume in this subsection that the positive semidefinite matrix $\boldsymbol{M}$ is exactly rank-r, i.e.,

$$\boldsymbol{M} = \boldsymbol{M}_r = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top = \boldsymbol{U}_r \boldsymbol{U}_r^\top \tag{2.5}$$

where $\boldsymbol{U}_r = [\sqrt{\sigma_1}\boldsymbol{u}_1, \ \ldots, \ \sqrt{\sigma_r}\boldsymbol{u}_r]$. Furthermore, we assume its condition number $\kappa_r = \frac{\sigma_1}{\sigma_r}$ and eigen-space incoherence parameter [CR09] $\mu_r = \frac{n}{r}\max_i \sum_{j=1}^{r} u_{i,j}^2$ are well-bounded. This is a standard setup in the literature of nonconvex matrix completion, e.g., [KMO10a, SL16, CW15, ZL16, GLM16, YPCC16, GJZ17].

Notice that [GLM16] introduces a slightly different version of incoherence

$$\widetilde{\mu}_r := \frac{\sqrt{n}\|\boldsymbol{U}_r\|_{2,\infty}}{\|\boldsymbol{U}_r\|_F} = \sqrt{\frac{n\|\boldsymbol{M}_r\|_{\ell_\infty}}{\mathrm{trace}(\boldsymbol{M}_r)}} \tag{2.6}$$

as a measure of spikiness. Note that this is different from the spikiness defined in [NW12]. By the fact that $\|\boldsymbol{M}_r\|_{\ell_\infty} = \|\boldsymbol{U}_r\|_{2,\infty}^2 = \max_i \sum_{j=1}^{r} \sigma_j u_{i,j}^2$, the following relationship between μ and $\widetilde{\mu}$ is straightforward

$$\frac{\widetilde{\mu}_r^2}{\kappa_r} \leqslant \frac{\widetilde{\mu}_r^2\,\mathrm{trace}(\boldsymbol{M}_r)}{r\sigma_1} = \frac{n\|\boldsymbol{M}_r\|_{\ell_\infty}}{r\sigma_1} \leqslant \mu_r \leqslant \frac{n\|\boldsymbol{M}_r\|_{\ell_\infty}}{r\sigma_r} = \frac{\widetilde{\mu}_r^2\,\mathrm{trace}(\boldsymbol{M}_r)}{r\sigma_r} \leqslant \kappa_r \widetilde{\mu}_r^2. \tag{2.7}$$

Using $\|\boldsymbol{M}\|_{\ell_\infty} \leqslant \frac{r}{n}\sigma_1\mu_r$, Theorem 2.1.2 implies the following exact low-rank recovery results:

Corollary 2.1.3. *Under the assumptions of Theorem 2.1.2, if we further assume* $\mathrm{rank}(\boldsymbol{M}) = r$ *(i.e.,* $\boldsymbol{M} = \boldsymbol{M}_r$*) and*

$$p \geqslant 4C_2 \max\left\{\frac{\mu_r r \kappa_r \log n}{n}, \frac{\mu_r^2 r^2 \kappa_r^2}{n}\right\}$$

or

$$p \geqslant 4C_2 \max\left\{\frac{\widetilde{\mu}_r^2 r \kappa_r \log n}{n}, \frac{\widetilde{\mu}_r^4 r^2 \kappa_r^2}{n}\right\},$$

then in the event E_1 *with probability* $\mathbb{P}[E_1] \geqslant 1 - 2n^{-3}$*, any local minimum* $\widehat{\boldsymbol{X}} \in \mathbb{R}^{n\times r}$ *of objective function* $f(\boldsymbol{X})$ *defined in* (1.1) *satisfies* $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top = \boldsymbol{M}$.

Notice that our results are better than the state-of-the-art results for no spurious local minimum in [GJZ17], where the required sampling rate is $p \geqslant \frac{C}{n}\mu_r^3 r^4 \kappa_r^4 \log n$ (which also implies $p \geqslant \frac{C}{n}\widetilde{\mu}_r^6 r^4 \kappa_r^7 \log n$ by (2.7)).

2.1.2. Examples.

2.1.2. Examples. Besides improving the state-of-the-art no-spurious-local-minima results in nonconvex matrix completion, Theorem 2.1.2 is also capable of explaining some nontrivial phenomena in low-rank matrix completion in the presence of large condition numbers, high incoherence parameter, or mismatching between the selected and true ranks.

2.1.2.1. *Nonconvex matrix completion with large condition numbers and high eigen-space incoherence parameters.* Assume here $\boldsymbol{M}$ is exactly rank-r and its spectral decomposition is denoted as in (2.5). However, we assume that μ_r and κ_r can be extremely large, while the condition number and incoherence parameter for $\boldsymbol{M}_{r-1} = \sum_{i=1}^{r-1} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$, i.e., $\kappa_{r-1} = \frac{\sigma_1}{\sigma_{r-1}}$ and $\mu_{r-1} = \frac{n}{r-1} \max_i \sum_{j=1}^{r-1} u_{i,j}^2$, are well-bounded. We are interested in figuring out when the local minimum based rank-r factorization $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top$ approximates the original $\boldsymbol{M}$ well.

By $\|\boldsymbol{M}_r\|_{\ell_\infty} = \max_i \sum_{j=1}^r \sigma_j u_{i,j}^2$, we have

$$\|\boldsymbol{M}_r\|_{\ell_\infty} \leqslant \frac{r-1}{n}\sigma_1 \mu_{r-1} + \sigma_r \|\boldsymbol{u}_r\|_\infty^2.$$

Then by Theorem 2.1.2, if

$$p \geqslant C \max\left\{ \frac{\left[\mu_{r-1}\kappa_{r-1}(r-1) + n\frac{\sigma_r}{\sigma_{r-1}}\|\boldsymbol{u}_r\|_\infty^2\right]\log n}{n}, \frac{\left[\mu_{r-1}\kappa_{r-1}(r-1) + n\frac{\sigma_r}{\sigma_{r-1}}\|\boldsymbol{u}_r\|_\infty^2\right]^2}{n} \right\}$$

with some absolute constant C, in an event E with probability $\mathbb{P}[E] \geqslant 1 - 2n^{-3}$, for any local minimum $\widehat{\boldsymbol{X}} \in \mathbb{R}^{n\times r}$ of (1.1), $\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\|_F^2 \leqslant \frac{1}{100}\sigma_{r-1}^2$ holds. In other words, the relative approximation error satisfies $\frac{\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F} \leqslant \frac{1}{10\sqrt{r-1}}$.

Notice that $\|\boldsymbol{u}_r\|_\infty^2 \leqslant \frac{r}{n}\mu_r$ and $\frac{\sigma_r}{\sigma_{r-1}} = \frac{\kappa_{r-1}}{\kappa_r}$, so the above sampling rate requirement is satisfied as long as $\frac{\mu_r}{\kappa_r} \leqslant C\mu_{r-1}$ and

$$p \geqslant C \max\left\{ \frac{\mu_{r-1}\kappa_{r-1} r \log n}{n}, \frac{\mu_{r-1}^2 \kappa_{r-1}^2 r^2}{n} \right\}.$$

2.1.2.2. *Rank mismatching.* In this subsection, $\boldsymbol{M}$ is assumed to be exactly rank-R, i.e.,

$$\boldsymbol{M} = \boldsymbol{M}_R = \sum_{i=1}^{R} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top = \boldsymbol{U}_R \boldsymbol{U}_R^\top$$

where $\boldsymbol{U}_R = [\sqrt{\sigma_1}\boldsymbol{u}_1 \ \ \ldots \ \ \sqrt{\sigma_R}\boldsymbol{u}_R]$. However, we consider the case that the selected rank r is not the same as the true rank R, i.e., rank mismatching. As with Section 2.1.1, we assume the condition number $\kappa_R = \frac{\sigma_1}{\sigma_R}$ and eigen-space incoherence parameter $\mu_R = \frac{n}{R}\max_i \sum_{j=1}^{R} \sigma_j u_{i,j}^2$ are well-bounded. As with (2.7), there holds $\|\boldsymbol{M}\|_{\ell_\infty} \leqslant \frac{R}{n}\sigma_1 \mu_R$.

Case 1: $R < r$. Theorem 2.1.2 implies that if

$$p \geqslant C \max\left\{\frac{\mu_R \kappa_R R \log n}{n}, \frac{\mu_R^2 \kappa_R^2 R^2}{n}\right\}$$

for some absolute constant C, then in an event E with probability $\mathbb{P}[E] \geqslant 1 - 2n^{-3}$, any local minimum $\widehat{\boldsymbol{X}} \in \mathbb{R}^{n\times r}$ of (1.1) yields $\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\|_F^2 \leqslant \frac{1}{100}(r-R)\sigma_R^2$. This further yields the relative approximation error bound $\frac{\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F} \leqslant \frac{1}{10}\sqrt{\frac{r-R}{R}}$.

Case 2: $R > r$. Recall that $\|\boldsymbol{M}_r\|_{\ell_\infty} \leqslant \frac{r}{n}\sigma_1\mu_r$. Moreover,

$$\|\boldsymbol{M}_{r+}\|_{\ell_\infty} = \max_i \sum_{j=r+1}^{R} \sigma_j u_{i,j}^2 \leqslant \sigma_{r+1}\left(\max_i \sum_{j=1}^{R} u_{i,j}^2\right) = \frac{\mu_R R}{n}\sigma_{r+1}.$$

Theorem 2.1.2 implies that if

$$p \geqslant C \max\left\{\frac{\mu_r r \kappa_r \log n}{n}, \frac{\mu_r^2 r^2 \kappa_r^2}{n}, \frac{\mu_R^2 R^3}{n}\right\}$$

for some absolute constant C, then with high probability, any local minimum $\widehat{\boldsymbol{X}} \in \mathbb{R}^{n\times r}$ of (1.1) yields

$$\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}_r\|_F^2 \leqslant C(\sigma_{r+1}^2 + \ldots + \sigma_{2r}^2),$$

which implies that the relative error is well-controlled as long as $\sigma_{r+1}^2 + \ldots + \sigma_R^2$ accounts for a small proportion in $\sigma_1^2 + \ldots + \sigma_R^2$.

If we assume that $2C_2\sigma_{r+1} < \sigma_r$ where C_2 is specified in Theorem 2.1.2, under the same sampling rate requirement as above, Theorem 2.1.2 implies a much sharper result:

$$\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}_r\|_F^2 \leqslant \frac{1}{100}\sigma_{r+1}^2,$$

which yields the following (perhaps surprising) relative approximation error bound

$$\frac{\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}_r\|_F}{\|\boldsymbol{M}_r\|_F} \leqslant \frac{1}{10}\sqrt{\frac{\sigma_{r+1}^2}{\sigma_1^2 + \ldots + \sigma_r^2}} \leqslant \frac{1}{10\sqrt{r}}.$$

2.2. Simulations and applications in memory-efficient kernel PCA

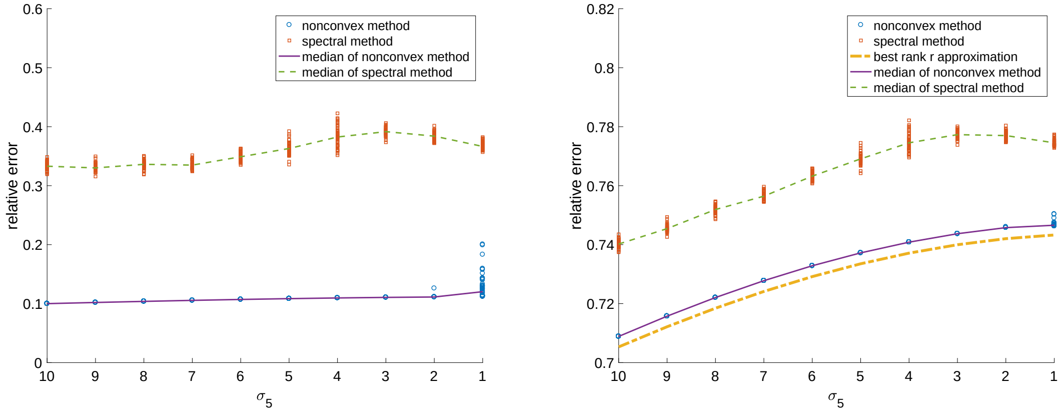
In the following simulations, we solve the following nonconvex optimization which is slightly different from (1.1):

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n\times r}} f(\boldsymbol{X}) \coloneqq \frac{1}{2}\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M})\|_F^2 + \lambda G_\alpha(\boldsymbol{X}).$$

The initialization $\boldsymbol{X}^{(0)}$ is constructed randomly with i.i.d. normal entries with mean 0 and variance 1. The step size $\eta^{(t)}$ for the gradient descent (2.1) is determined by Armijo's rule [**Arm66**]. The gradient descent algorithm is implemented with sparse matrix storage in Section 2.2.2 for the purpose of memory-efficient KPCA, while with full matrix storage in Section 2.2.1 to test the performance of general low-rank approximations from missing data. In each experiment, the iterations will be terminated when $\|\nabla f(\boldsymbol{X}^{(t)})\|_F \leqslant 10^{-3}$ or $\|\eta^{(t)}\nabla f(\boldsymbol{X}^{(t)})\|_F \leqslant 10^{-10}$ or the number of iterations surpasses 10^3. All methods are implemented in MATLAB. The experiments are running on a virtual computer with Linux KVM, with 12 cores of 2.00GHz Intel Xeon E5 processor and 16 GB memory.

2.2.1. Numerical simulations. In this section, we conduct numerical tests on a nonconvex optimization under different settings of spectrum for the 500×500 positive semidefinite matrix $\boldsymbol{M}$, whose eigenvectors are the same as the left singular vectors of a random 500×500 matrix with i.i.d. standard normal entries. The generation of eigenvalues for $\boldsymbol{M}$ will be further specified in each test. For each generated $\boldsymbol{M}$, the nonconvex optimization is implemented for 50 times with independent Ω's generated under the off-diagonal symmetric independent $\mathrm{Ber}(p)$ model. To implement the gradient descent algorithm (2.1), set $\alpha = 100\|\boldsymbol{M}\|_{\ell_\infty}$ and $\lambda = 100\|\boldsymbol{\Omega} - p\boldsymbol{J}\|$ (the performances of our method are empirically not sensitive to the choices of the tuning parameters). In each single numerical experiment, we also conduct spectral method proposed in [**AMS02**] to obtain an approximate low-rank approximation of $\boldsymbol{M}$ for the purpose of comparison.

2.2.1.1. *Full rank case.* Here $\boldsymbol{M}$ is assumed to have full rank, i.e., $\operatorname{rank}(\boldsymbol{M}) = 500$. To be specific, let $\sigma_1 = \cdots = \sigma_4 = 10$, $\sigma_6 = \cdots = \sigma_{500} = 1$, and $\sigma_5 = 10, 9, 8, \ldots, 2, 1$. The selected rank used in the nonconvex optimization is set as $r = 5$, and the sampling rate is set as $p = 0.2$. With different values of σ_5, the results of our implementations of the gradient descent are plotted in Figure 2.1. One can observe that the relative errors for our nonconvex method are well-bounded for different σ_5's, and much smaller than those for spectral low-rank approximation. The results indicate that our approach is able to approximate the "true" best rank-r approximation $\boldsymbol{M}_r$ accurately in the presence of heavy spectral tail and possibly large condition number σ_1/σ_5, even with only 20% observed entries.



(a) Relative error $\frac{\|\boldsymbol{M}_{\text{approx}} - \boldsymbol{M}_r\|_F}{\|\boldsymbol{M}_r\|_F}$.

(b) Relative error $\frac{\|\boldsymbol{M}_{\text{approx}} - \boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F}$.

FIGURE 2.1. Relative errors for full rank case.

2.2.1.2. *Low-rank matrix with large condition numbers.* Here $\boldsymbol{M}$ is assumed to be of exactly low rank with different condition numbers. Let $\sigma_1 = \cdots = \sigma_4 = 10$, $\sigma_5 = \frac{10}{\kappa}$, and $\sigma_6 = \cdots = \sigma_{500} = 0$. Here the condition number takes on values $\kappa = 10, 20, 30, 40, 50, 100, 200, \infty$, which implies $\operatorname{rank}(\boldsymbol{M}) = 5$ if $\kappa < \infty$ while $\operatorname{rank}(\boldsymbol{M}) = 4$ if $\kappa = \infty$. The selected rank is always assumed to be $r = 5$, while the sampling rate is always $p = 0.2$.

The performance of our nonconvex approach with various choices of κ is demonstrated in Figure 2.2. One can observe that our nononvex optimization approach yields exact recovery of $\boldsymbol{M}$ when $\kappa = 10$, while yields accurate low-rank approximation for $\boldsymbol{M}$ with relative errors almost always smaller than 0.3 when $\kappa \geqslant 20$. This fact is consistent with the example we discussed in Section

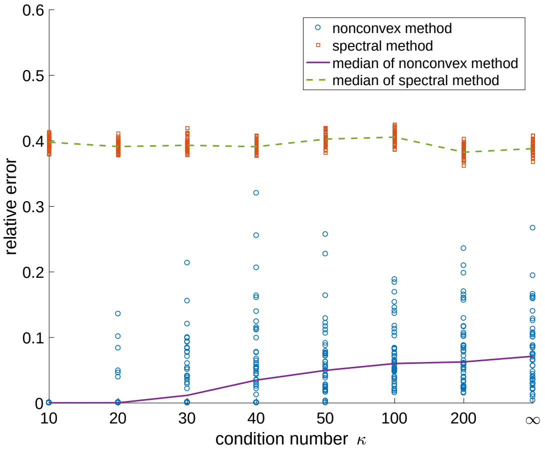
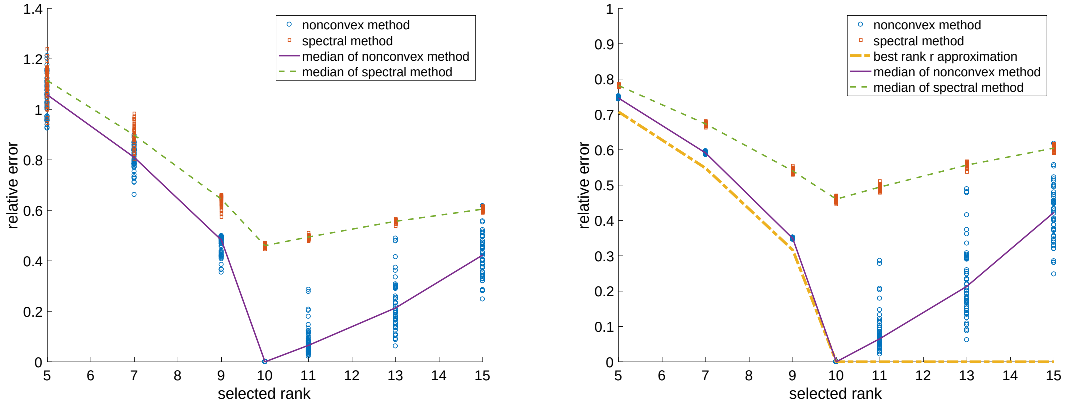


FIGURE 2.2. Relative error $\frac{\|\boldsymbol{M}_{\text{approx}}-\boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F}$ for low-rank matrix with extreme condition numbers.

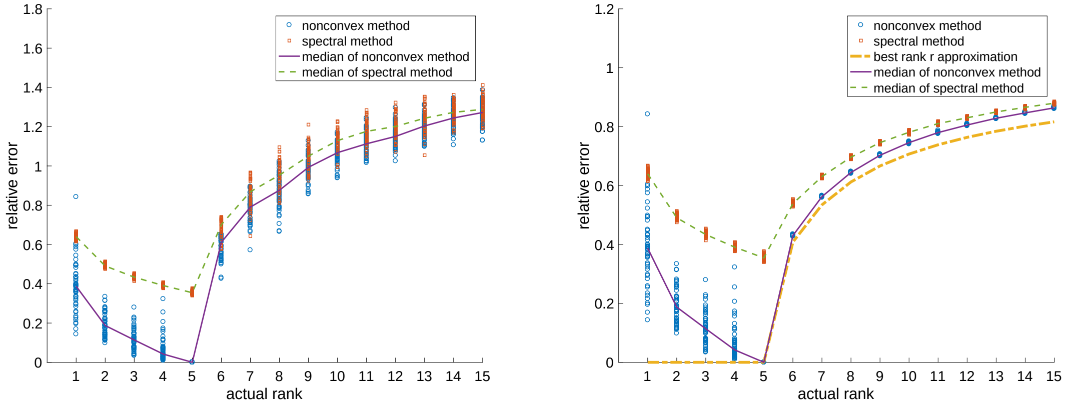
2.1.2.1, where we have shown that under certain incoherence conditions, the relative approximation error can be well-bounded even when $\kappa_r = \infty$.

2.2.1.3. *Rank mismatching.* In this section, we consider rank mismatching, i.e., the rank of $\boldsymbol{M}$ is low but different from the selected rank r. In particular, we consider two settings for simulation: First, we fix $\boldsymbol{M}$ with $\text{rank}(\boldsymbol{M}) = 10$, while the nonconvex optimization is implemented with selected rank $r = 5, 7, 9, 10, 11, 13, 15$; Second, the matrix $\boldsymbol{M}$ is randomly generated with rank from 1 to 15, while the selected rank is always $r = 5$. The sampling rate is fixed as $p = 0.2$. We perform the simulation on two sets of spectrums: For the first one, all the nonzero eigenvalues are 10; And the second one has decreasing eigenvalues: $\sigma_1 = 20, \sigma_2 = 18, \cdots, \sigma_{10} = 2$ for the case of fixed $\text{rank}(\boldsymbol{M})$, $\sigma_1 = 30, \cdots, \sigma_{\text{rank}(\boldsymbol{M})} = 32 - 2 \times \text{rank}(\boldsymbol{M})$ for the case of fixed selected rank r. Numerical results for the case of fixed $\text{rank}(\boldsymbol{M})$ are demonstrated in Figure 2.3 (constant nonzero eigenvalues) and Figure 2.5 (decreasing nonzero eigenvalues), while the case of fixed selected rank in Figure 2.4 (constant nonzero eigenvalues) and Figure 2.6 (decreasing nonzero eigenvalues). One can observe from these figures that if the selected rank r is less than the actual rank $\text{rank}(\boldsymbol{M})$, for the approximation of $\boldsymbol{M}$, our nonconvex approach performs almost as well as the complete-data based best low-rank approximation $\boldsymbol{M}_r$. Another interesting phenomenon is that our nonconvex method outperforms simple spectral methods in the approximation of either $\boldsymbol{M}$ or $\boldsymbol{M}_r$ significantly if the selected rank is greater than or equal to the true rank.



(a) Relative error $\frac{\|\boldsymbol{M}_{\text{approx}}-\boldsymbol{M}_r\|_F}{\|\boldsymbol{M}_r\|_F}$. (b) Relative error $\frac{\|\boldsymbol{M}_{\text{approx}}-\boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F}$.

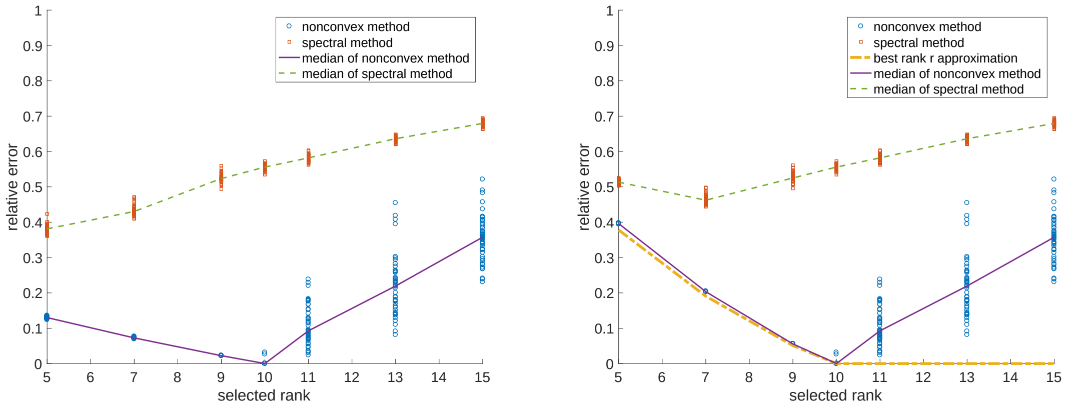
FIGURE 2.3. Relative errors for rank mismatching for a fixed $\boldsymbol{M}$ with $\operatorname{rank}(\boldsymbol{M}) = 10$.



(a) Relative error $\frac{\|\boldsymbol{M}_{\text{approx}}-\boldsymbol{M}_r\|_F}{\|\boldsymbol{M}_r\|_F}$. (b) Relative error $\frac{\|\boldsymbol{M}_{\text{approx}}-\boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F}$.

FIGURE 2.4. Relative errors for rank mismatching, fixed selected rank.

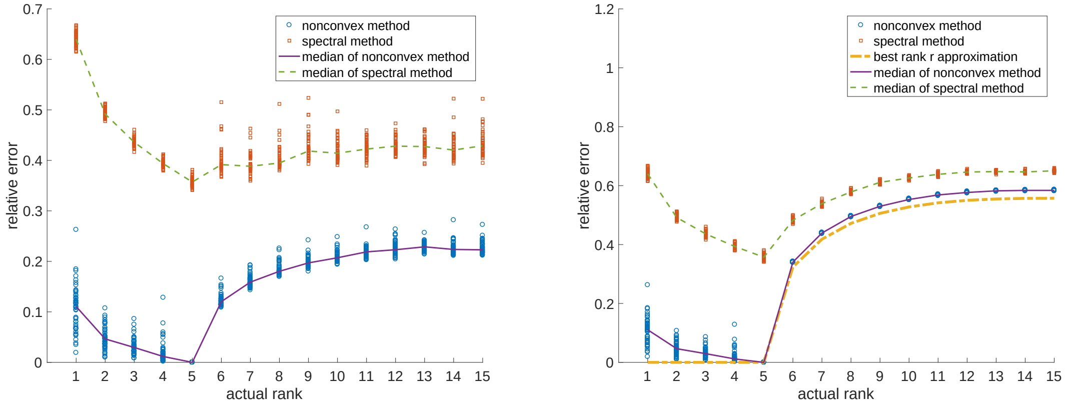
2.2.2. Empirical performance of memory-efficient kernel PCA. In order to study the empirical performance of our memory-efficient kernel PCA approach, we apply it to the synthetic data set in [**Wan12**]. The data set is an i.i.d. sample with sample size $n = 10,000$ and dimension $d = 3$, and the data points are partitioned into two classes independently with equal probabilities. Points in the first class are first generated uniformly at random on the three-dimensional sphere $\{\boldsymbol{x} : \|\boldsymbol{x}\|_2 = 0.3\}$, while points in the second class are first generated uniformly at random on the three-dimensional sphere $\{\boldsymbol{x} : \|\boldsymbol{x}\|_2 = 1\}$. Every point is then perturbed independently by



(a) Relative error $\frac{\|\boldsymbol{M}_{\text{approx}}-\boldsymbol{M}_r\|_F}{\|\boldsymbol{M}_r\|_F}$.

(b) Relative error $\frac{\|\boldsymbol{M}_{\text{approx}}-\boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F}$.

Figure 2.5. Relative errors for rank mismatching for a fixed $\boldsymbol{M}$ with $\operatorname{rank}(\boldsymbol{M}) = 10$.



(a) Relative error $\frac{\|\boldsymbol{M}_{\text{approx}}-\boldsymbol{M}_r\|_F}{\|\boldsymbol{M}_r\|_F}$.

(b) Relative error $\frac{\|\boldsymbol{M}_{\text{approx}}-\boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F}$.

Figure 2.6. Relative errors for rank mismatching, fixed selected rank.

$\mathcal{N}(\boldsymbol{0}, \frac{1}{100}\boldsymbol{I}_3)$ noise. We aim to implement memory-efficient uncentered kernel PCA with $r = 2$ on this dataset with the radial kernel $\exp(-\|\boldsymbol{x} - \boldsymbol{y}\|_2^2)$ in order to cluster the data points.

To implement the Nyström method [WS01], 50 columns (and corresponding rows) are selected uniformly at random without replacement, then a rank-2 approximation of the kernel matrix $\boldsymbol{M}$ can be efficiently constructed with a smaller scale factorization. The effective sampling rate for Nyström method is $p_{\text{Nys}} = \frac{2\times 50n - 50^2}{n^2} \approx 0.01$. In contrast, in addition to recording the selected entry values, our nonconvex optimization method also requires to record the row and column indices for each selected entry. By using sparse matrix storage schemes like compressed sparse row

(CSR) format [Saa03], it needs $2n^2 p_{NCVX} + n + 1$ entries to store the sparse matrix. Therefore, if $p_{NCVX} \geqslant \frac{3}{n}$, the nonconvex approach requires at most 2.5 times as much memory as Nyström method for the same sampling complexity. Therefore, we choose the sampling rate $p_{\text{NCVX}} = \frac{p_{\text{Nys}}}{2.5}$ in the implementation of the nonconvex optimization such that the memory consumption is less costly than the Nyström method.

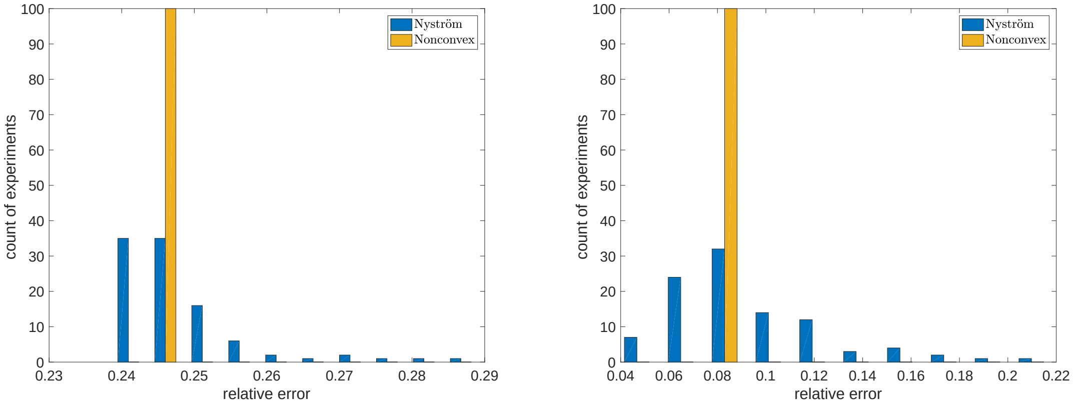
Fixing such a synthetic data set, we apply both the Nyström method and our approach (with $\alpha = 100\|\boldsymbol{M}\|_{\ell_\infty} = 100$ and $\lambda = 500\sqrt{np_{\text{NCVX}}}$) for 100 times. Denote by $\boldsymbol{M}$ the ground truth of the kernel matrix, by $\boldsymbol{M}_2$ the ground truth of the best rank-2 approximation of $\boldsymbol{M}$, and by $\boldsymbol{M}_{\text{approx}}$ the memory efficient rank-2 approximation obtained by Nyström method or our nonconvex optimization. The left and right panels of Figure 2.7 compare the two methods in approximating $\boldsymbol{M}_2$ and $\boldsymbol{M}$ respectively based on the distributions of relative errors throughout the 100 Monte Carlo simulations. One can see that our approach is comparable with the Nyström method in terms of median performance, but much more stable.

Both Nyström method and our nonconvex optimization give approximation in the form of $\boldsymbol{M} \approx \widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top$, so clustering analysis can be directly implemented based on $\widehat{\boldsymbol{X}}$. We implement k-means on the rows of $\widehat{\boldsymbol{X}}$ with 20 repetitions, and Figure 2.8 compares the two methods in the distribution of clustering accuracies. It clearly shows that our nonconvex optimization yields accurate clustering throughout the 100 tests while the Nyström method results in poor clustering occasionally.

Moreover, during the iterations of the nonconvex method, the regularization term never activate throughout the 100 simulations. Therefore, empirically speaking, the performances of our numerical tests will remain the same if we simply set $\lambda = 0$.

2.3. Proof of Theorem 2.1.2

In this section, we give a proof for Theorem 2.1.2. In Section 2.3.1, we will present some useful supporting lemmas; in Section 2.3.2, we present a proof for our main result Theorem 2.1.2; we leave proof of lemmas used in former subsections to the appendix. Our proof ideas benefit from those in [GJZ17] as well as [ZLTW17], [JGN+17].



(a) Relative error $\frac{\|\boldsymbol{M}_{\text{approx}}-\boldsymbol{M}_2\|_F}{\|\boldsymbol{M}_2\|_F}$.

(b) Relative error $\frac{\|\boldsymbol{M}_{\text{approx}}-\boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F}$.

FIGURE 2.7. Relative errors for Nyström method with sampling rate $p_{\text{Nys}} \approx 0.01$ and nonconvex method with sampling rate $p_{\text{NCVX}} = \frac{p_{\text{Nys}}}{2.5}$.

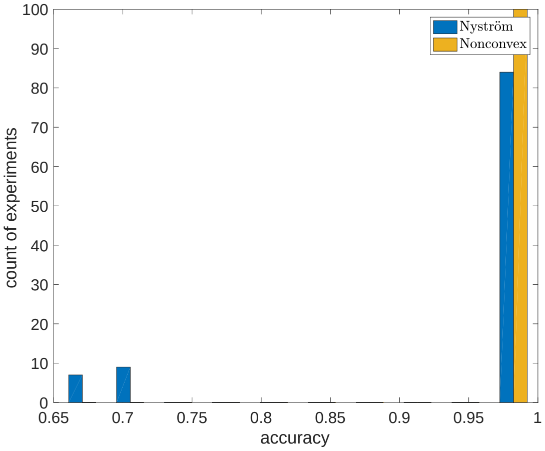


FIGURE 2.8. Clustering accuracy for Nyström method with sampling rate $p_{\text{Nys}} \approx 0.01$ and nonconvex method with sampling rate $p_{\text{NCVX}} = \frac{p_{\text{Nys}}}{2.5}$.

2.3.1. Supporting lemmas. In this section, we give some useful supporting lemmas. The following lemma is well known in the literature, see, e.g., [**Vu18**] and [**BVH16**].

LEMMA 2.3.1. *There is a constant $C_v > 0$ such that the following holds. If Ω is sampled according to the off-diagonal symmetric Ber(p) model with $p \geqslant C_v \frac{\log n}{n}$, then in an event E_v with probability $\mathbb{P}[E_v] \geqslant 1 - n^{-3}$,*

$$\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \leqslant C_v\sqrt{np}.$$

The following eigen-space incoherence parameter has been proposed in [CR09].

Definition 2.3.2 ([CR09]). *For any subspace $\mathcal{U}$ of $\mathbb{R}^n$ of dimension r, denote $\boldsymbol{P}_{\mathcal{U}} : \mathbb{R}^n \to \mathbb{R}^n$ as the orthogonal projection onto $\mathcal{U}$. Define*

$$\mu(\mathcal{U}) \coloneqq \frac{n}{r} \max_{1 \leqslant i \leqslant n} \|\boldsymbol{P}_{\mathcal{U}} \boldsymbol{e}_i\|_2^2, \tag{2.8}$$

where $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ represents the standard orthogonal basis of $\mathbb{R}^n$.

As with Theorem 4.1 in [CR09], for the off-diagonal symmetric Ber(p) model, we also have:

Lemma 2.3.3. *Let Ω be sampled according to the off-diagonal symmetric Ber(p) model. Define*

$$\mathcal{T} \coloneqq \{\boldsymbol{M} \in \mathbb{R}^{n \times n} \mid (\boldsymbol{I} - \boldsymbol{P}_{\mathcal{U}})\boldsymbol{M}(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{U}}) = \boldsymbol{0},\ \boldsymbol{M} \text{ symmetric}\},$$

where $\mathcal{U}$ is a fixed subspace of $\mathbb{R}^n$. Let $\mathcal{P}_{\mathcal{T}}$ be the Euclidean projection on to $\mathcal{T}$: For any symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$,

$$\mathcal{P}_{\mathcal{T}}(\boldsymbol{M}) = \boldsymbol{P}_{\mathcal{U}}\boldsymbol{M} + \boldsymbol{M}\boldsymbol{P}_{\mathcal{U}} - \boldsymbol{P}_{\mathcal{U}}\boldsymbol{M}\boldsymbol{P}_{\mathcal{U}}.$$

Then there is an absolute constant C_{Ca}, if $p \geqslant C_{Ca} \frac{\mu(\mathcal{U}) \dim(\mathcal{U}) \log n}{n}$ with $\mu(\mathcal{U})$ defined in (2.8), in an event E_{Ca} with probability $\mathbb{P}[E_{Ca}] \geqslant 1 - n^{-3}$, we have

$$p^{-1}\|\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}} - p\mathcal{P}_{\mathcal{T}}\| \leqslant 10^{-5}.$$

In [Gro11] and [GN10], similar results are given for symmetric uniform sampling with/without replacement. The proof of Lemma 2.3.3 is very similar to that in [Rec11].

The first and second order optimally conditions of $f(\boldsymbol{X})$ satisfy the following properties:

Lemma 2.3.4 ([GLM16, Proposition 4.1]). *The first order optimality condition of objective function (1.1) is*

$$\nabla f(\boldsymbol{X}) = 2\mathcal{P}_{\Omega}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M})\boldsymbol{X} + \lambda \nabla G_\alpha(\boldsymbol{X}) = \boldsymbol{0},$$

and the second order optimality condition requires that for any $\boldsymbol{H} \in \mathbb{R}^{n\times r}$, we have

$$\begin{aligned}
&\operatorname{vec}(\boldsymbol{H})^\top \nabla^2 f(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{H}) \\
=&\|\mathcal{P}_\Omega(\boldsymbol{H}\boldsymbol{X}^\top + \boldsymbol{X}\boldsymbol{H}^\top)\|_F^2 + 2\langle \mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}), \mathcal{P}_\Omega(\boldsymbol{H}\boldsymbol{H}^\top)\rangle + \lambda \operatorname{vec}(\boldsymbol{H})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{H}) \\
\geqslant& 0.
\end{aligned}$$

In the sequel, we are going to present our key lemma which will be used multiple times throughout this section. For any matrix $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{n_1\times n_2}$, any set $\Omega_0 \in [n_1]\times[n_2]$ and any real number $t \in \mathbb{R}$, we introduce following notation for simplicity of notations:

$$D_{\Omega_0,t}(\boldsymbol{M}_1, \boldsymbol{M}_2) := \langle \mathcal{P}_{\Omega_0}(\boldsymbol{M}_1), \mathcal{P}_{\Omega_0}(\boldsymbol{M}_2)\rangle - t\langle \boldsymbol{M}_1, \boldsymbol{M}_2\rangle. \tag{2.9}$$

More specifically, sometimes we use $D(\boldsymbol{M}_1, \boldsymbol{M}_2)$ as a shortcut of $D_{\Omega,p}(\boldsymbol{M}_1, \boldsymbol{M}_2)$. Our key lemma is given as follows:

Lemma 2.3.5. *Let Ω_0 be any index set in $[n_1]\times[n_2]$, and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{n_1\times n_2}$ be defined correspondingly as in Section 2.1. For any $\boldsymbol{A} \in \mathbb{R}^{n_1\times r_1}, \boldsymbol{B} \in \mathbb{R}^{n_1\times r_2}, \boldsymbol{C} \in \mathbb{R}^{n_2\times r_1}, \boldsymbol{D} \in \mathbb{R}^{n_2\times r_2}$, and any $t \in \mathbb{R}$, there holds*

$$|D_{\Omega_0,t}(\boldsymbol{A}\boldsymbol{C}^\top, \boldsymbol{B}\boldsymbol{D}^\top)| \leqslant \|\boldsymbol{\Omega}_0 - t\boldsymbol{J}\| \sqrt{\sum_{k=1}^{n_1} \|\boldsymbol{A}_{k,\cdot}\|_2^2 \|\boldsymbol{B}_{k,\cdot}\|_2^2} \sqrt{\sum_{k=1}^{n_2} \|\boldsymbol{C}_{k,\cdot}\|_2^2 \|\boldsymbol{D}_{k,\cdot}\|_2^2} \tag{2.10}$$

We will use this result for $\Omega_0 = \Omega, t = p$ for multiple times later. Note that here we do not make any assumptions on Ω_0 and this is a deterministic result. The proof of this lemma is deferred to Section A.1.0.1. This result extends the following lemma given in [BJ14] and [LLR16]:

Lemma 2.3.6 ([BJ14, LLR16]). *Suppose matrix $\boldsymbol{M} \in \mathbb{R}^{n_1\times n_2}$ can be decomposed as $\boldsymbol{M} = \boldsymbol{B}\boldsymbol{D}^\top$, let $\Omega_0 \subset [n_1]\times[n_2]$ be any index set. Then for any $t \in \mathbb{R}$, we have*

$$\|\mathcal{P}_{\Omega_0}(\boldsymbol{M}) - t\boldsymbol{M}\| \leqslant \|\boldsymbol{\Omega}_0 - t\boldsymbol{J}\| \|\boldsymbol{B}\|_{2,\infty} \|\boldsymbol{D}\|_{2,\infty}.$$

Lemma 2.3.5 is applied in our proof of Lemma 2.3.9 in replace of Theorem D.1 in [GLM16] to derive tighter control of perturbation terms, i.e., $K_2(\boldsymbol{X}), K_3(\boldsymbol{X})$ and $K_4(\boldsymbol{X})$ defined in (2.14). Their result is given here for the purpose of comparison.

Lemma 2.3.7 ([GLM16, Theorem D.1]). *With high probability over the choice of Ω, for any two rank-r matrices $\boldsymbol{W}, \boldsymbol{Z} \in \mathbb{R}^{n\times n}$, we have*

$$\begin{aligned} &|\langle \mathcal{P}_\Omega(\boldsymbol{W}), \mathcal{P}_\Omega(\boldsymbol{Z})\rangle - p\langle \boldsymbol{W}, \boldsymbol{Z}\rangle| \\ =&O\left(\|\boldsymbol{W}\|_{\ell_\infty}\|\boldsymbol{Z}\|_{\ell_\infty} nr\log n + \sqrt{pnr\|\boldsymbol{W}\|_{\ell_\infty}\|\boldsymbol{Z}\|_{\ell_\infty}\|\boldsymbol{W}\|_F\|\boldsymbol{Z}\|_F \log n}\right). \end{aligned} \tag{2.11}$$

In [SL16], [CW15] and [ZL16], upper bounds are given to $\|\mathcal{P}_\Omega(\boldsymbol{H}\boldsymbol{H}^\top)\|_F^2$ for any $\boldsymbol{H}$. To be more precise, they assume Ω is sampled according to the i.i.d. Bernoulli model with probability p. If $p \geqslant C\frac{\log n}{n}$ for some sufficient large absolute constant C, there holds

$$\|\mathcal{P}_\Omega(\boldsymbol{H}\boldsymbol{H}^\top)\|_F^2 - p\|\boldsymbol{H}\|_F^4 \leqslant C\sqrt{np}\sum_{i=1}^n \|\boldsymbol{H}_{i,\cdot}\|_2^4 \tag{2.12}$$

with high probability. In contrast, by combining Lemma 2.3.1 and Lemma 2.3.5, there holds

$$|\|\mathcal{P}_\Omega(\boldsymbol{H}\boldsymbol{H}^\top)\|_F^2 - p\|\boldsymbol{H}\boldsymbol{H}^\top\|_F^2| \leqslant C\sqrt{np}\sum_{i=1}^n \|\boldsymbol{H}_{i,\cdot}\|_2^4 \tag{2.13}$$

with high probability. This is tighter than (2.12) in that $\|\boldsymbol{H}\boldsymbol{H}^\top\|_F \leqslant \|\boldsymbol{H}\|_F^2$. Moreover, comparing to (2.12), our result (2.13) directly measures the difference between $\|\mathcal{P}_\Omega(\boldsymbol{H}\boldsymbol{H}^\top)\|_F^2$ and its expectation $p\|\boldsymbol{H}\boldsymbol{H}^\top\|_F^2$, which makes the model-free analysis possible.

2.3.2. A proof of Theorem 2.1.2.

This section aims to prove Theorem 2.1.2. The proof is basically divided into two parts: In Section 2.3.2.1, we discuss the landscape of objective function $f(\boldsymbol{X})$ and then define the auxiliary function $K(\boldsymbol{X})$. We show that the span of local minima of $f(\boldsymbol{X})$ can be controlled by the superlevel set of $K(\boldsymbol{X})$: $\{\boldsymbol{X} \in \mathbb{R}^{n\times r} \mid K(\boldsymbol{X}) \geqslant 0\}$. In Section 2.3.2.2, we give a uniform upper bound of $K(\boldsymbol{X})$ in order to control the above superlevel set.

2.3.2.1. *Landscape of objective function f and auxiliary function K.* Denote $\boldsymbol{U}_r$ as $\boldsymbol{U}_r := [\sqrt{\sigma_1}\boldsymbol{u}_1 \ \ldots \ \sqrt{\sigma_r}\boldsymbol{u}_r]$. For a given $\boldsymbol{X} \in \mathbb{R}^{n\times r}$, suppose that $\boldsymbol{X}^\top\boldsymbol{U}_r$ has SVD $\boldsymbol{X}^\top\boldsymbol{U}_r = \boldsymbol{A}\boldsymbol{D}\boldsymbol{B}^\top$, and let $\boldsymbol{R}_{\boldsymbol{X},\boldsymbol{U}_r} := \boldsymbol{B}\boldsymbol{A}^\top \in \mathsf{O}(r)$ and $\boldsymbol{U} := \boldsymbol{U}_r\boldsymbol{R}_{\boldsymbol{X},\boldsymbol{U}_r}$, where $\mathsf{O}(r)$ denotes the set of $r\times r$ orthogonal matrices $\{\boldsymbol{R} \in \mathbb{R}^{r\times r} \mid \boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{R}\boldsymbol{R}^\top = \boldsymbol{I}\}$. Then $\boldsymbol{X}^\top\boldsymbol{U} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top$ is a positive semidefinite matrix. Then also holds $\boldsymbol{U}_r\boldsymbol{U}_r^\top = \boldsymbol{U}\boldsymbol{U}^\top$.

Denote $\boldsymbol{\Delta} := \boldsymbol{X} - \boldsymbol{U}$, and define the following auxiliary function introduced in [**JGN+17**] and [**GJZ17**]:

$$K(\boldsymbol{X}) := \operatorname{vec}(\boldsymbol{\Delta})^\top \nabla^2 f(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{\Delta}) - 4\langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta}\rangle.$$

The first and second order optimality conditions for any local minimum $\widehat{\boldsymbol{X}}$ imply that $K(\widehat{\boldsymbol{X}}) \geqslant 0$. In other words, we have

$$\{\text{All local minima of } f(\boldsymbol{X})\} \subset \{\boldsymbol{X} \in \mathbb{R}^{n\times r} \mid K(\boldsymbol{X}) \geqslant 0\}.$$

Figure 2.9 illustrates the relationship between local minima of f and the superlevel set of $K(\boldsymbol{X})$.

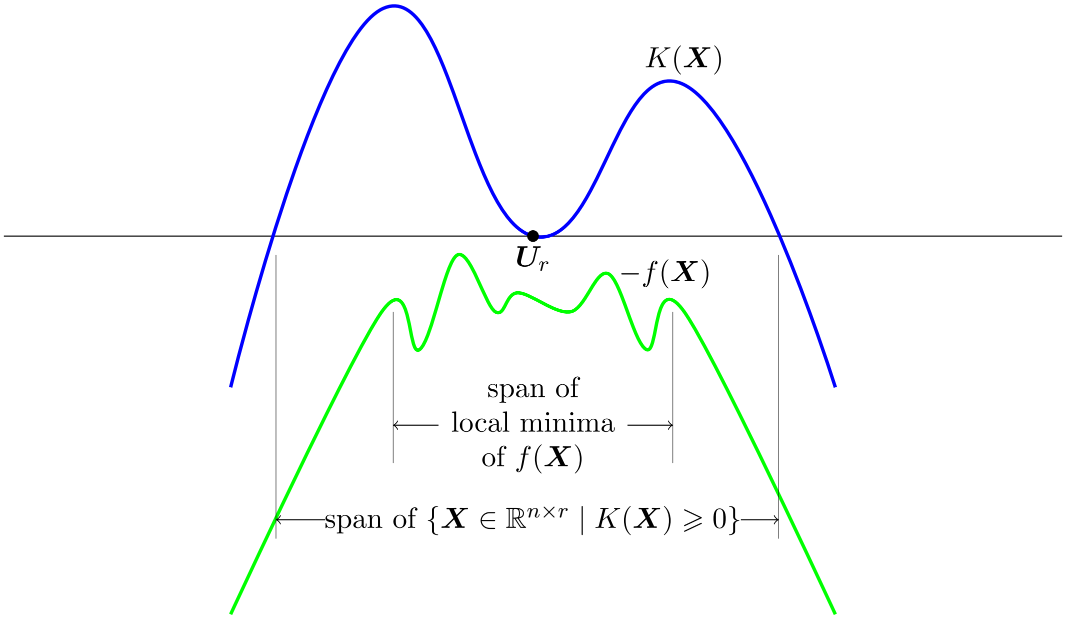


FIGURE 2.9. Landscape of $-f(\boldsymbol{X}), K(\boldsymbol{X})$ and $\boldsymbol{U}_r$.

To study the properties of the local minima of $f(\boldsymbol{X})$, we can consider the superlevel set of $K(\boldsymbol{X})$: $\{\boldsymbol{X} \in \mathbb{R}^{n\times r} \mid K(\boldsymbol{X}) \geqslant 0\}$ instead. In order to get a clear representation of $K(\boldsymbol{X})$, one can plug in the formulas of gradient and Hessian in Lemma 2.3.4. By repacking terms in [**GJZ17**, Lemma 7], and given $\langle \boldsymbol{U}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle = 0$, due to the definition of $\boldsymbol{U}$ and $\boldsymbol{M}_{r+}$, $K(\boldsymbol{X})$ can be decomposed as follows:

LEMMA 2.3.8 ([**GJZ17**, Lemma 7]). *Uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n\times r}$, as well as corresponding $\boldsymbol{U}$ and $\boldsymbol{\Delta}$ defined above, we have*

$$
\begin{aligned}
K(\boldsymbol{X}) = &\underbrace{\left(\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 - 3\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2\right)}_{K_1(\boldsymbol{X})} \\
&+ \underbrace{\frac{1}{p}D_{\Omega,p}(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top) - \frac{3}{p}D_{\Omega,p}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top)}_{K_2(\boldsymbol{X})} \\
&+ \underbrace{\lambda\left(\operatorname{vec}(\boldsymbol{\Delta})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{\Delta}) - 4\langle \nabla G_\alpha(\boldsymbol{X}), \boldsymbol{\Delta}\rangle\right)}_{K_3(\boldsymbol{X})} \\
&+ \underbrace{\frac{6}{p}D_{\Omega,p}(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}) + \frac{8}{p}D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}) + 6\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle}_{K_4(\boldsymbol{X})},
\end{aligned} \tag{2.14}
$$

where $D_{\Omega,p}(\cdot,\cdot)$ is defined in (2.9).

Notice that in Theorem 2.1.2, we are only concerned about the difference between $\boldsymbol{X}\boldsymbol{X}^\top$ and $\boldsymbol{M}_r$ (or $\boldsymbol{M}$), which remains the same by replacing $\boldsymbol{X}$ with $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{R}$, for any $\boldsymbol{R} \in \mathsf{O}(r)$. On the other hand, by the definition of $\boldsymbol{R}_{\boldsymbol{X},\boldsymbol{U}_r}$, we have $\boldsymbol{R}_{\boldsymbol{X}\boldsymbol{R},\boldsymbol{U}_r} = \boldsymbol{R}_{\boldsymbol{X},\boldsymbol{U}_r}\boldsymbol{R}$ for any $\boldsymbol{R} \in \mathsf{O}(r)$, which implies $\widetilde{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{R}$ and $\widetilde{\boldsymbol{\Delta}} = \boldsymbol{\Delta}\boldsymbol{R}$. Now we have

$$
\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top = \boldsymbol{X}\boldsymbol{X}^\top, \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^\top = \boldsymbol{U}\boldsymbol{U}^\top, \widetilde{\boldsymbol{\Delta}}\widetilde{\boldsymbol{\Delta}}^\top = \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Delta}}^\top = \boldsymbol{U}\boldsymbol{\Delta}^\top,
$$

which means $K_i(\widetilde{\boldsymbol{X}}) = K_i(\boldsymbol{X})$ for $i = 1, 2, 4$. As for K_3, by [GJZ17, Lemma 18], we have

$$
\begin{aligned}
&\operatorname{vec}(\boldsymbol{\Delta})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{\Delta}) - 4\langle \nabla G_\alpha(\boldsymbol{X}), \boldsymbol{\Delta}\rangle \\
=&4\sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^3 \frac{\|\boldsymbol{X}_{i,\cdot}\|_2^2 \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2 - \langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot}\rangle^2}{\|\boldsymbol{X}_{i,\cdot}\|_2^3} + 12\sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^2 \frac{\langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot}\rangle^2}{\|\boldsymbol{X}_{i,\cdot}\|_2^2} \\
&- 16\sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^3 \frac{\langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot}\rangle}{\|\boldsymbol{X}_{i,\cdot}\|_2}.
\end{aligned}
$$

Since $\boldsymbol{R} \in \mathsf{O}(r)$, we have $\|\widetilde{\boldsymbol{X}}_{i,\cdot}\|_2 = \|\boldsymbol{X}_{i,\cdot}\|_2$, $\|\widetilde{\boldsymbol{\Delta}}_{i,\cdot}\|_2 = \|\boldsymbol{\Delta}_{i,\cdot}\|_2$ and $\langle \widetilde{\boldsymbol{X}}_{i,\cdot}, \widetilde{\boldsymbol{\Delta}}_{i,\cdot}\rangle = \langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot}\rangle$, so we have $K_3(\widetilde{\boldsymbol{X}}) = K_3(\boldsymbol{X})$. Putting things together, we have $K(\widetilde{\boldsymbol{X}}) = K(\boldsymbol{X})$.

Therefore, if we want to show that any $\boldsymbol{X}$ with $K(\boldsymbol{X}) \geqslant 0$ satisfies (2.3) and (2.4) with high probability, without loss of generality, we can assume that $\boldsymbol{X}$ satisfies the property that $\boldsymbol{X}^\top \boldsymbol{U}_r$ is a positive semidefinite matrix, i.e., $\boldsymbol{U} = \boldsymbol{U}_r$.

2.3.2.2. *Proof of Theorem 2.1.2.* In order to prove our main result, we first give a uniform upper bound of $K(\boldsymbol{X})$. Then for any local minimum $\widehat{\boldsymbol{X}}$, $K(\widehat{\boldsymbol{X}}) \geqslant 0$, the property enables us to solve for the range of possible $\widehat{\boldsymbol{X}}$.

Lemma 2.3.9. *Assume that tuning parameters α, λ satisfy $100\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}} \leqslant \alpha \leqslant 200\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}}$, $100\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p} \leqslant \lambda \leqslant 200\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}$, and $p \geqslant C_v\frac{\log n}{n}$ with absolute constant C_v defined in Lemma 2.3.1. Then, in an event E_1 with probability $\mathbb{P}[E_1] \geqslant 1 - 2n^{-3}$, uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n\times r}$ and corresponding $\boldsymbol{\Delta}$ defined as before, we have*

$$\sum_{i=2}^{4} K_i(\boldsymbol{X}) \leqslant 10^{-3}\left[\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right] + \psi, \tag{2.15}$$

where

$$\psi := C_3\sum_{i=1}^{r}\left\{\left[C_2\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\|\boldsymbol{M}_r\|_{\ell_\infty} + C_2\sigma_{2r+1-i} - \sigma_i\right]_+\right\}^2 + C_3\frac{nr\|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2}{p} \tag{2.16}$$

and C_2, C_3 are absolute constants defined within the proof.

Note in our proof of Theorem 2.1.2, we only use probabilistic tools in the above lemma to control perturbation terms, i.e., $K_2(\boldsymbol{X}), K_3(\boldsymbol{X}), K_4(\boldsymbol{X})$. The rest part of the proof is purely deterministic.

Recall by the way we define $\boldsymbol{\Delta}$,

$$\begin{aligned}\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2 =& \|\boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 \\ &= \|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 + 2\|\boldsymbol{\Delta}\boldsymbol{U}^\top\|_F^2 + 2\langle\boldsymbol{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top\rangle + 4\langle\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top\rangle.\end{aligned} \tag{2.17}$$

Now denote $a := \|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F = \|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F, b := \|\boldsymbol{\Delta}^\top\boldsymbol{U}\|_F$. Putting Lemma 2.3.8 and Lemma 2.3.9 together, and using (2.17), we have

$$\begin{aligned}K(\boldsymbol{X}) \leqslant& 1.001\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 - 3\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2 + 10^{-3}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + \psi \\ =& 1.001a^2 - 3\left[\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 + 2\|\boldsymbol{\Delta}\boldsymbol{U}^\top\|_F^2 + 2\langle\boldsymbol{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top\rangle + 4\langle\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top\rangle\right] \\ &+ 10^{-3}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + \psi,\end{aligned} \tag{2.18}$$

By the definition of matrix inner product, we have

$$\begin{aligned}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 =&\langle \boldsymbol{U}\boldsymbol{\Delta}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top\rangle = \operatorname{trace}(\boldsymbol{\Delta}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Delta}^\top) = \operatorname{trace}(\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Delta}^\top\boldsymbol{\Delta}) \\ =&\langle \boldsymbol{U}^\top\boldsymbol{U}, \boldsymbol{\Delta}^\top\boldsymbol{\Delta}\rangle,\end{aligned} \tag{2.19}$$

and

$$\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top\rangle = \operatorname{trace}(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\boldsymbol{U}\boldsymbol{\Delta}^\top) = \operatorname{trace}(\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\boldsymbol{U}) = \langle \boldsymbol{\Delta}^\top\boldsymbol{\Delta}, \boldsymbol{\Delta}^\top\boldsymbol{U}\rangle. \tag{2.20}$$

Here we use the fact that trace$(\boldsymbol{AB})$ = trace$(\boldsymbol{BA})$ for any matrix $\boldsymbol{A}$ and $\boldsymbol{B}$ with suitable size. Moreover, since we choose $\boldsymbol{U}$ such that $\boldsymbol{U}^\top\boldsymbol{X}$ is positive semidefinite, $\boldsymbol{U}^\top\boldsymbol{\Delta} = \boldsymbol{\Delta}^\top\boldsymbol{U}$ and $\boldsymbol{U}^\top(\boldsymbol{\Delta}+\boldsymbol{U}) \succeq \mathbf{0}$. Therefore, we also have

$$\begin{aligned}\langle \boldsymbol{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top\rangle =&\operatorname{trace}(\boldsymbol{U}\boldsymbol{\Delta}^\top\boldsymbol{U}\boldsymbol{\Delta}^\top) = \operatorname{trace}(\boldsymbol{\Delta}\boldsymbol{U}^\top\boldsymbol{\Delta}\boldsymbol{U}^\top) = \operatorname{trace}(\boldsymbol{U}^\top\boldsymbol{\Delta}\boldsymbol{U}^\top\boldsymbol{\Delta}) \\ =&\langle \boldsymbol{\Delta}^\top\boldsymbol{U}, \boldsymbol{U}^\top\boldsymbol{\Delta}\rangle = \langle \boldsymbol{\Delta}^\top\boldsymbol{U}, \boldsymbol{\Delta}^\top\boldsymbol{U}\rangle = \|\boldsymbol{\Delta}^\top\boldsymbol{U}\|_F^2\end{aligned} \tag{2.21}$$

and

$$\langle \boldsymbol{\Delta}^\top\boldsymbol{\Delta}, \boldsymbol{U}^\top\boldsymbol{U} + \boldsymbol{\Delta}^\top\boldsymbol{U}\rangle = \langle \boldsymbol{\Delta}^\top\boldsymbol{\Delta}, (\boldsymbol{U}+\boldsymbol{\Delta})^\top\boldsymbol{U}\rangle \geqslant 0. \tag{2.22}$$

Here (2.22) also uses the fact that inner product of two positive semidefinite matrices is non-negative.

By putting (2.18), (2.19), (2.20), (2.21) together,

$$\begin{aligned}K(\boldsymbol{X}) \leqslant&1.001a^2 - 3\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 - 6\|\boldsymbol{\Delta}\boldsymbol{U}^\top\|_F^2 - 6\langle \boldsymbol{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top\rangle - 12\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top\rangle \\ &+ 10^{-3}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + \psi \\ =&-1.999a^2 - 6\langle \boldsymbol{U}^\top\boldsymbol{U}, \boldsymbol{\Delta}^\top\boldsymbol{\Delta}\rangle - 6\|\boldsymbol{\Delta}^\top\boldsymbol{U}\|_F^2 - 12\langle \boldsymbol{\Delta}^\top\boldsymbol{\Delta}, \boldsymbol{\Delta}^\top\boldsymbol{U}\rangle \\ &+ 10^{-3}\langle \boldsymbol{U}^\top\boldsymbol{U}, \boldsymbol{\Delta}^\top\boldsymbol{\Delta}\rangle + \psi \\ =&-1.999a^2 - \langle \boldsymbol{\Delta}^\top\boldsymbol{\Delta}, 5.999\boldsymbol{U}^\top\boldsymbol{U} + 12\boldsymbol{\Delta}^\top\boldsymbol{U}\rangle - 6b^2 + \psi.\end{aligned} \tag{2.23}$$

Therefore, combining with (2.22),

$$\begin{aligned}K(\boldsymbol{X}) \leqslant&-1.999a^2 - 6.001\langle \boldsymbol{\Delta}^\top\boldsymbol{\Delta}, \boldsymbol{\Delta}^\top\boldsymbol{U}\rangle - 6b^2 + \psi \\ \leqslant&-1.999a^2 + 6.001ab - 6b^2 + \psi\end{aligned} \tag{2.24}$$

holds for all $\boldsymbol{X} \in \mathbb{R}^{n\times r}$. For the last line, we apply Cauchy-Schwarz inequality for matrices, i.e.,

$$|\langle \boldsymbol{\Delta}^\top \boldsymbol{\Delta}, \boldsymbol{\Delta}^\top \boldsymbol{U}\rangle| \leqslant \|\boldsymbol{\Delta}^\top \boldsymbol{\Delta}\|_F \|\boldsymbol{\Delta}^\top \boldsymbol{U}\|_F = ab.$$

Note that for any local minimum $\widehat{\boldsymbol{X}}$, we have $K(\widehat{\boldsymbol{X}}) \geqslant 0$. Replacing $\boldsymbol{X}$ with $\widehat{\boldsymbol{X}}$ in (2.24), there holds

$$-1.999a^2 + 6.001ab - 6b^2 + \psi \geqslant 0,$$

which further implies

$$0 \leqslant a \leqslant 2\sqrt{\psi}, \qquad 0 \leqslant b \leqslant \sqrt{\psi}. \tag{2.25}$$

From (2.23), we have

$$K(\widehat{\boldsymbol{X}}) \leqslant -1.999a^2 - \langle \boldsymbol{\Delta}^\top \boldsymbol{\Delta}, 5.999\boldsymbol{U}^\top \boldsymbol{U} + 12\boldsymbol{\Delta}^\top \boldsymbol{U}\rangle - 6b^2 + \psi.$$

Recall from (2.19), $\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 = \langle \boldsymbol{U}^\top \boldsymbol{U}, \boldsymbol{\Delta}^\top \boldsymbol{\Delta}\rangle$, and $K(\widehat{\boldsymbol{X}}) \geqslant 0$. Therefore, combining with (2.25),

$$\begin{aligned}
5.999\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 \leqslant & -1.999a^2 - \langle \boldsymbol{\Delta}^\top \boldsymbol{\Delta}, 12\boldsymbol{\Delta}^\top \boldsymbol{U}\rangle - 6b^2 + \psi \\
\leqslant & -1.999a^2 + 12\|\boldsymbol{\Delta}^\top \boldsymbol{\Delta}\|_F \|\boldsymbol{\Delta}^\top \boldsymbol{U}\|_F - 6b^2 + \psi \\
\leqslant & -1.999a^2 + 12ab - 6b^2 + \psi \\
\leqslant & 25\psi.
\end{aligned} \tag{2.26}$$

From (2.18),

$$K(\widehat{\boldsymbol{X}}) \leqslant 1.001\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 - 3\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2 + 10^{-3}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + \psi.$$

Using the fact that $K(\widehat{\boldsymbol{X}}) \geqslant 0$ again, we have

$$3\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2 \leqslant 1.001\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 + 10^{-3}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + \psi$$

Combining with (2.25), (2.26), we futher have

$$3\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2 \leqslant 1.001a^2 + 0.005\psi + \psi \leqslant 6\psi. \tag{2.27}$$

Therefore, (2.3) is directly implied by (2.27) via choosing $C_1 = 2C_3$. Notice that

$$\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\|_F^2 = \|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2 - 2\langle \widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top, \boldsymbol{M}_{r+}\rangle + \|\boldsymbol{M}_{r+}\|_F^2 \leqslant \|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2 + \|\boldsymbol{M}_{r+}\|_F^2$$

where the inequality holds since $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top \succeq \mathbf{0}$ and $\boldsymbol{M}_{r+} \succeq \mathbf{0}$. Therefore, (2.4) is implied by (2.3).

2.4. Other sampling models, a uniform approximation theorem

Most of the matrix completion literature concentrates on the uniform sampling model as defined in Model 2.1.1, and limited literature talks about other models. [**BJ14**] considered sampling from a d-regular graph satisfies certain assumption, and [**LLR16**] considered weighted low-rank matrix completion, just to name a few.

In the existing nonconvex matrix completion literature, theoretical analysis of uniform sampling model relies heavily on the uniform randomness, and many useful tools will fail if the uniform randomness is violated. For example, [**CG18**] considered a semi-random sampling model: After the uniform sampling, additional entries are allowed to reveal. Intuitively speaking, since we are given more information than before, we would expect a better approximation. However, it is still open what is the best way to use those extra information. In [**CG18**], exact recovery is not guaranteed even if we are given more information than what is needed under uniform model.

Given the fact that deterministic inequality Lemma 2.3.5 does not make any assumption on the sampling model, it becomes possible to extend the model-free framework to other sampling models introduced in [**BJ14**, **LLR16**, **CG18**]. Actually, given Lemma 2.3.5, we are able to derive a uniform approximation control of any fixed sampling pattern:

Theorem 2.4.1 (Uniform approximation). *Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a positive semidefinite matrix with the spectral decomposition* (2.2)*. Let Ω be any fixed sampling pattern. For any fixed $t \in (0,\infty)$, let*

$$\phi(t) := \frac{\|\boldsymbol{\Omega} - t\boldsymbol{J}\|}{t}.$$

Then as long as the tuning parameters α and λ satisfy $100\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}} \leqslant \alpha \leqslant 200\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}}$ and $100\phi(t) \leqslant \lambda \leqslant 200\phi(t)$, any local minimum $\widehat{\boldsymbol{X}} \in \mathbb{R}^{n\times r}$ of

$$\frac{1}{2t}\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M})\|_F^2 + \lambda G_\alpha(\boldsymbol{X})$$

satisfies

$$\begin{aligned}\left\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}_r\right\|_F^2 \leqslant & C_1 \sum_{i=1}^r \left\{[C_2\phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty} + C_2\sigma_{2r+1-i} - \sigma_i]_+\right\}^2 \\ & + C_1\phi^2(t)r\|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2 + 2\phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=1}^r \sigma_i.\end{aligned}$$

and

$$\begin{aligned}\left\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\right\|_F^2 \leqslant & C_1 \sum_{i=1}^r \left\{[C_2\phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty} + C_2\sigma_{2r+1-i} - \sigma_i]_+\right\}^2 \\ & + C_1\phi^2(t)r\|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2 + 2\phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=1}^r \sigma_i + \|\boldsymbol{M}_{r+}\|_F^2.\end{aligned}$$

Where constants C_1, C_2 are defined in Theorem 2.1.2.

When Ω follows the uniform sampling as in Model 2.1.1, by taking $t = p$ and using Lemma 2.3.1 on $\phi(p)$, with high probability, the approximation upper bounds in Theorem 2.4.1 match the bounds in Theorem 2.1.2, except an extra $2\phi(p)\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=1}^r \sigma_i$ term.

Moreover, when $\operatorname{rank}(\boldsymbol{M}) = r$, for any fixed Ω, when $\Omega \neq [n] \times [n]$, i.e., there are missing entries, the extra term $2\phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=1}^r \sigma_i$ prohibits the nonconvex approach to achieve exact recovery. This is actually not a big surprise due to the fact that universal exact recovery does not exist even for rank-two (incoherent) matrices given $n^2/4$ observed entries [**BJ14**, Claim 5.2].

From Theorem 2.4.1, one can observe that $\phi(t)$ plays a special role in the approximation error upper bound. This matches the simulations in universal recovery literature, for example, [**BJ14**].

As an interesting byproduct of Theorem 2.4.1, we are able to analyze the following semi-random model:

Model 2.4.2. *Indices of revealed entries are first sampled uniformly with probability p, as described in Model 2.1.1. The revealed index set is denoted as Ω_{unif}. After that, an adversary is allowed to see the ground truth matrix and Ω_{unif}. The adversary is allowed to reveal extra indices or even cover/drop revealed entries in Ω_{unif}. The set of affected indices is denoted as Ω_{adv}. And the final set of revealed indices is denoted as Ω. For notation simplicity, we define three matrices*

$\mathbf{\Omega}_{unif}$, $\mathbf{\Omega}_{adv}$, $\mathbf{\Omega}$ *as following. For any* i, j,

$$[\mathbf{\Omega}_{unif}]_{i,j} = \begin{cases} 1 & \textit{if } (i,j) \in \Omega_{unif} \\ 0 & \textit{otherwise}, \end{cases}$$

$$[\mathbf{\Omega}_{adv}]_{i,j} = \begin{cases} 1 & \textit{if } (i,j) \notin \Omega_{unif} \textit{ and } (i,j) \in \Omega_{adv} \\ -1 & \textit{if } (i,j) \in \Omega_{unif} \textit{ and } (i,j) \in \Omega_{adv} \\ 0 & \textit{otherwise}, \end{cases}$$

$$[\mathbf{\Omega}]_{i,j} = \begin{cases} 1 & \textit{if } (i,j) \in \Omega \\ 0 & \textit{otherwise}. \end{cases}$$

Therefore, we have $\mathbf{\Omega} = \mathbf{\Omega}_{unif} + \mathbf{\Omega}_{adv}$.

More discussion about the semi-random model can be found in [**CG18**] and references therein. Given the above defined semi-random model, we are able to conclude the following result.

Corollary 2.4.3. *Let* $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ *be a rank-r positive semidefinite matrix. Let* Ω *be sampled according to the semi-random model defined in Model 2.4.2. For any* $\varepsilon \in (0,1)$, *if*

$$p \geqslant \max\left\{4C_2^2C_v^2\frac{\mu^2r^2\kappa^2}{n}, \frac{16C_v^2}{\varepsilon^2}\frac{\mu^2r^2\kappa^2}{n}, C_v\frac{\log n}{n}\right\}$$

and $|\Omega_{adv}| \leqslant np$. *Then as long as the tuning parameters* α *and* λ *satisfy* $100\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}} \leqslant \alpha \leqslant 200\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}}$ *and* $100\frac{\|\mathbf{\Omega}-p\boldsymbol{J}\|}{p} \leqslant \lambda \leqslant 200\frac{\|\mathbf{\Omega}-p\boldsymbol{J}\|}{p}$, *in an event* E_{semi} *with probability* $\mathbb{P}[E_{semi}] \geqslant 1 - n^{-3}$, *any local minimum* $\widehat{\boldsymbol{X}} \in \mathbb{R}^{n\times r}$ *of* (1.1) *satisfies*

$$\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\|_F^2 \leqslant \varepsilon\|\boldsymbol{M}\|_F^2.$$

The above result still holds if $|\Omega_{\rm adv}| = O(np)$. Therefore, the above corollary reveals the fact that in the semi-random model, if the adversary does not change the uniform revealed index set too much, i.e., $|\Omega_{\rm adv}| = O(np)$, we can still recover $\boldsymbol{M}$ quite well using the original objective function (1.1). Comparing to [**CG18**, Theorem 1.2], we have a lower sampling complexity requirement (in terms of μ, r and κ) and do not need an extra pre-processing. On the other hand side, our corollary can only handle the situation when $|\Omega_{\rm adv}|$ is small, while [**CG18**, Theorem 1.2] can cover a wider range of semi-random models.

Proof of Corollary 2.4.3. The above corollary is a simple implication of our uniform approximation theorem Theorem 2.4.1. By taking $t = p$, we can see that

$$\phi(p) = \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p} = \frac{\|\boldsymbol{\Omega}_{\text{unif}} + \boldsymbol{\Omega}_{\text{adv}} - p\boldsymbol{J}\|}{p} \leqslant \frac{\|\boldsymbol{\Omega}_{\text{unif}} - p\boldsymbol{J}\|}{p} + \frac{\|\boldsymbol{\Omega}_{\text{adv}}\|}{p}$$

From Lemma 2.3.1 we see $\frac{\|\boldsymbol{\Omega}_{\text{unif}} - p\boldsymbol{J}\|}{p} \leqslant C_v\sqrt{\frac{n}{p}}$, and by the fact that $\frac{\|\boldsymbol{\Omega}_{\text{adv}}\|}{p} \leqslant \frac{\|\boldsymbol{\Omega}_{\text{adv}}\|_F}{p} = \frac{\sqrt{|\Omega_{\text{adv}}|}}{p} \leqslant \sqrt{\frac{n}{p}}$, we can see $\phi(p) \leqslant 2C_v\sqrt{\frac{n}{p}}$. Plugging it back to Theorem 2.4.1 finishes the proof. □

2.5. Rectangular case

In Section 2.1, a model-free framework of local minima analysis is proposed for PSD matrix completion with following objective function:

$$f_{\text{psd}}(\boldsymbol{X}) := \frac{1}{2p}\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M})\|_F^2 + \lambda G_\alpha(\boldsymbol{X}).$$

On the other hand side, the global geometry of nonconvex rectangular matrix completion is analysed in [GJZ17] with following objective function:

$$f_{\text{rect}}(\boldsymbol{X}, \boldsymbol{Y}) := \frac{1}{2p}\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M})\|_F^2 + \frac{1}{8}\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}\|_F^2 + \lambda(G_\alpha(\boldsymbol{X}) + G_\alpha(\boldsymbol{Y})).$$

Therefore, one natural question arise: Can we extend the model-free framework to the rectangular case? The answer is yes.

First of all, suppose the index set Ω satisfies following model:

Model 2.5.1. *For rectangular matrix completion, the index set Ω is assumed to follow the independent Ber(p) model, i.e., each entry is sampled independently with probability p.*

Moreover, suppose that matrix $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ has following singular value decomposition:

$$\boldsymbol{M} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top + \sum_{i=r+1}^{n_1 \wedge n_2} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top := \boldsymbol{M}_r + \boldsymbol{M}_{r+}, \tag{2.28}$$

where $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_n \geqslant 0$ are the spectrum. $\boldsymbol{u}_i \in \mathbb{R}^{n_1}$ are unit and mutually perpendicular singular vectors, and so are $\boldsymbol{v}_i \in \mathbb{R}^{n_2}$. Similar to what we have in Section 2.1, the matrix $\boldsymbol{M}_r := \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ is the best rank-r approximation of $\boldsymbol{M}$ and $\boldsymbol{M}_{r+} := \sum_{i=r+1}^{n} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ denotes the residual part. Similar to what we did in the PSD case, let

$$\boldsymbol{U}_r \coloneqq [\sqrt{\sigma_1}\boldsymbol{u}_1, \ldots, \sqrt{\sigma_r}\boldsymbol{u}_r], \quad \boldsymbol{V}_r \coloneqq [\sqrt{\sigma_1}\boldsymbol{v}_1, \ldots, \sqrt{\sigma_r}\boldsymbol{v}_r]$$

and

$$\boldsymbol{W}_r \coloneqq \begin{bmatrix} \boldsymbol{U}_r \\ \boldsymbol{V}_r \end{bmatrix}.$$

Then we have the following result:

THEOREM 2.5.2. *Let $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ be an n_1-by-n_2 matrix with the spectral decomposition (2.28). Let Ω be sampled according to the Ber(p) model with $p \geqslant C_v \frac{\log(n_1 \vee n_2)}{n_1 \wedge n_2}$ for some absolute constant C_v. Then in an event E_2 with probability $\mathbb{P}[E_2] \geqslant 1 - 2(n_1+n_2)^{-3}$, as long as the tuning parameters α and λ satisfy $100\|\boldsymbol{W}_r\|_{2,\infty} \leqslant \alpha \leqslant 200\|\boldsymbol{W}_r\|_{2,\infty}$ and $100\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p} \leqslant \lambda \leqslant 200\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}$, any local minimum $(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}) \in \mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r}$ of (1.3) satisfies*

$$\begin{aligned}\left\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{Y}}^\top - \boldsymbol{M}_r\right\|_F^2 \leqslant & C_1 \sum_{i=1}^r \left\{\left[C_2\left(\sqrt{\frac{n_1 \vee n_2}{p}} + \frac{\log(n_1 \vee n_2)}{p}\right)\|\boldsymbol{W}_r\|_{2,\infty}^2 + C_2\sigma_{2r+1-i} - \sigma_i\right]_+\right\}^2 \\ & + C_1 \frac{(n_1 \vee n_2) r \|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2}{p}\end{aligned}$$

and

$$\begin{aligned}\left\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{Y}}^\top - \boldsymbol{M}\right\|_F^2 \leqslant & C_1 \sum_{i=1}^r \left\{\left[C_2\left(\sqrt{\frac{n_1 \vee n_2}{p}} + \frac{\log(n_1 \vee n_2)}{p}\right)\|\boldsymbol{W}_r\|_{2,\infty}^2 + C_2\sigma_{2r+1-i} - \sigma_i\right]_+\right\}^2 \\ & + C_1 \frac{(n_1 \vee n_2) r \|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2}{p} + \|\boldsymbol{M}_{r+}\|_F^2\end{aligned}$$

with C_1, C_2 absolute constants defined in Theorem 2.1.2.

The proof of Theorem 2.5.2 is basically a mimic of proof of Theorem 2.1.2. Therefore, we leave the sketch of the proof to the appendix.

By following the proof techniques introduced in Section 2.4, we are also able to achieve a uniform approximation result under rectangular setup analogs to Theorem 2.4.1.

By letting eigen-space incoherence parameter for rectangular matrix as

$$\mu_r = \max\left\{\frac{n_1}{r}\max_i \sum_{j=1}^r u_{i,j}^2, \frac{n_2}{r}\max_i \sum_{j=1}^r v_{i,j}^2\right\},$$

we can also derive an exact recovery result from Theorem 2.5.2.

Corollary 2.5.3. *Under the assumptions of Theorem 2.5.2, if we further assume* $\operatorname{rank}(\boldsymbol{M}) = r$ *(i.e., $\boldsymbol{M} = \boldsymbol{M}_r$) and*

$$p \geqslant 4C_2 \max\left\{\frac{\mu_r r \kappa_r \log(n_1 \vee n_2)}{n_1 \wedge n_2}, \frac{\mu_r^2 r^2 \kappa_r^2}{n_1 \wedge n_2}\right\}$$

then in the event E_2 with probability $\mathbb{P}[E_2] \geqslant 1 - 2(n_1 + n_2)^{-3}$, any local minimum $(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}) \in \mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r}$ of objective function $f_{rect}(\boldsymbol{X}, \boldsymbol{Y})$ defined in (1.3) satisfies $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{Y}}^\top = \boldsymbol{M}$.

This result improves the sampling complexity required for no spurious local minimum analysis [**GJZ17**], which is $p \geqslant C\mu_r^4 r^6 \kappa_r^6 \log(n_1 \vee n_2)/(n_1 \wedge n_2)$.

CHAPTER 3

Global Geometry of Nonconvex Parameterized Linear Models

Similar to the discussions in Chapter 1 and Chapter 2, the vanilla noisy matrix completion problem can be stated as follows: Let $\boldsymbol{M}$ be an $n_1 \times n_2$ matrix, and we would like to estimate the whole matrix from a small proportion of noisy observed entries. To be specific, let $\Omega \subset [n_1] \times [n_2]$ be the index set that supports all observed entries. The observation is represented by $\mathcal{P}_\Omega(\boldsymbol{M}+\boldsymbol{N})$, where $\boldsymbol{N}$ is a matrix that represents noise or perturbation. Given selected rank r, the following regularized least squares fitting is proposed and further analyzed in [GJZ17]

$$f_{\text{rect}}(\boldsymbol{X},\boldsymbol{Y}) := \frac{1}{2p}\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M} - \boldsymbol{N})\|_F^2 + \frac{1}{8}\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}\|_F^2 + \lambda(G_\alpha(\boldsymbol{X}) + G_\alpha(\boldsymbol{Y})), \tag{3.1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n_1\times r}, \boldsymbol{Y} \in \mathbb{R}^{n_2\times r}$.

In Section 2.5, the global geometry is analyzed in the noiseless case where $\boldsymbol{N} = \boldsymbol{0}$. If we further assume $\boldsymbol{M}$ is a rank-r well-conditioned matrix, its eigenspace incoherence parameter [CR09] is well-bounded, and the sampling rate satisfies $p \gtrsim (r^2 \log(n_1 \vee n_2))/(n_1 \wedge n_2)$, any local minimum of (3.1) gives $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{Y}}^\top = \boldsymbol{M}$, i.e., there is no spurious local minimum.

In this Chapter, we are devoted to study the nonconvex geometry for (1.6) as [GLM16, GJZ17, CL19] did for the vanilla matrix completion problem (3.1).

3.1. Main Results

3.1.1. Method. We first give the specific form of (1.6). Plugging the parametric form $\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{\theta})$ and $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{\theta})$ into the nonconvex optimization (3.1), we have the following optimization:

$$\begin{aligned}\tilde{f}(\boldsymbol{\theta}) :=& f_{\text{rect}}(\boldsymbol{X}(\boldsymbol{\theta}),\boldsymbol{Y}(\boldsymbol{\theta}))\\ =& \frac{1}{2p}\|\mathcal{P}_\Omega(\boldsymbol{X}(\boldsymbol{\theta})\boldsymbol{Y}(\boldsymbol{\theta})^\top - \boldsymbol{M} - \boldsymbol{N})\|_F^2 + \frac{1}{8}\|\boldsymbol{X}(\boldsymbol{\theta})^\top\boldsymbol{X}(\boldsymbol{\theta}) - \boldsymbol{Y}(\boldsymbol{\theta})^\top\boldsymbol{Y}(\boldsymbol{\theta})\|_F^2\\ &+ \lambda(G_\alpha(\boldsymbol{X}(\boldsymbol{\theta})) + G_\alpha(\boldsymbol{Y}(\boldsymbol{\theta}))).\end{aligned} \tag{3.2}$$

Prior to investigating the theoretical properties of (3.2), let us first specialize it to the completion of subspace-constrained and skew-symmetric low-rank matrices, where the parameterization takes the forms (1.4) and (1.5), respectively.

- In the case of matrix completion with subspace constraints, denote $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B)$, and the linear mappings $\boldsymbol{X}(\boldsymbol{\theta})$ and $\boldsymbol{Y}(\boldsymbol{\theta})$ are defined as in (1.4). Without loss of generality, assume that both $\widetilde{\boldsymbol{U}}$ and $\widetilde{\boldsymbol{V}}$ consist of orthonormal basis, i.e., $\widetilde{\boldsymbol{U}}^\top \widetilde{\boldsymbol{U}} = \boldsymbol{I}_{s_1}$ and $\widetilde{\boldsymbol{V}}^\top \widetilde{\boldsymbol{V}} = \boldsymbol{I}_{s_2}$. Then the parameterization (1.4) implies the following

$$\begin{cases} \boldsymbol{X}(\boldsymbol{\theta})\boldsymbol{Y}(\boldsymbol{\theta})^\top = \widetilde{\boldsymbol{U}}\boldsymbol{\Theta}_A\boldsymbol{\Theta}_B^\top\widetilde{\boldsymbol{V}}^\top, \\ \boldsymbol{X}(\boldsymbol{\theta})^\top\boldsymbol{X}(\boldsymbol{\theta}) = \boldsymbol{\Theta}_A^\top\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}}\boldsymbol{\Theta}_A = \boldsymbol{\Theta}_A^\top\boldsymbol{\Theta}_A, \\ \boldsymbol{Y}(\boldsymbol{\theta})^\top\boldsymbol{Y}(\boldsymbol{\theta}) = \boldsymbol{\Theta}_B^\top\widetilde{\boldsymbol{V}}^\top\widetilde{\boldsymbol{V}}\boldsymbol{\Theta}_B = \boldsymbol{\Theta}_B^\top\boldsymbol{\Theta}_B. \end{cases}$$

  Substituting them into (3.2), we have the objective function:

$$\begin{aligned} f_{\text{subspace}}(\boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B) \coloneqq & \frac{1}{2p}\|\mathcal{P}_\Omega(\widetilde{\boldsymbol{U}}\boldsymbol{\Theta}_A\boldsymbol{\Theta}_B^\top\widetilde{\boldsymbol{V}}^\top - \boldsymbol{M} - \boldsymbol{N})\|_F^2 + \frac{1}{8}\|\boldsymbol{\Theta}_A^\top\boldsymbol{\Theta}_A - \boldsymbol{\Theta}_B^\top\boldsymbol{\Theta}_B\|_F^2 \\ & + \lambda(G_\alpha(\widetilde{\boldsymbol{U}}\boldsymbol{\Theta}_A) + G_\alpha(\widetilde{\boldsymbol{V}}\boldsymbol{\Theta}_B)). \end{aligned} \tag{3.3}$$

- In the case of skew-symmetric matrix completion, again, denote $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B)$, and the linear mappings $\boldsymbol{X}(\boldsymbol{\theta})$ and $\boldsymbol{Y}(\boldsymbol{\theta})$ are defined as in (1.5). Straightforward calculation gives

$$\begin{cases} \boldsymbol{X}(\boldsymbol{\theta})\boldsymbol{Y}(\boldsymbol{\theta})^\top = \boldsymbol{\Theta}_A\boldsymbol{\Theta}_B^\top - \boldsymbol{\Theta}_B\boldsymbol{\Theta}_A^\top, \\ \boldsymbol{X}(\boldsymbol{\theta})^\top\boldsymbol{X}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\Theta}_A^\top\boldsymbol{\Theta}_A & -\boldsymbol{\Theta}_A^\top\boldsymbol{\Theta}_B \\ -\boldsymbol{\Theta}_B^\top\boldsymbol{\Theta}_A & \boldsymbol{\Theta}_B^\top\boldsymbol{\Theta}_B \end{bmatrix}, \\ \boldsymbol{Y}(\boldsymbol{\theta})^\top\boldsymbol{Y}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\Theta}_B^\top\boldsymbol{\Theta}_B & \boldsymbol{\Theta}_B^\top\boldsymbol{\Theta}_A \\ \boldsymbol{\Theta}_A^\top\boldsymbol{\Theta}_B & \boldsymbol{\Theta}_A^\top\boldsymbol{\Theta}_A \end{bmatrix}. \end{cases}$$

Substituting them into (3.2), we have the objective function

$$
\begin{aligned}
f_{\text{skew}}(\boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B) :=& \frac{1}{2p}\|\mathcal{P}_\Omega(\boldsymbol{\Theta}_A\boldsymbol{\Theta}_B^\top - \boldsymbol{\Theta}_B\boldsymbol{\Theta}_A^\top - \boldsymbol{M} - \boldsymbol{N})\|_F^2 + \frac{1}{4}\|\boldsymbol{\Theta}_A^\top\boldsymbol{\Theta}_A - \boldsymbol{\Theta}_B^\top\boldsymbol{\Theta}_B\|_F^2 \\
&+ \frac{1}{4}\|\boldsymbol{\Theta}_A^\top\boldsymbol{\Theta}_B + \boldsymbol{\Theta}_B^\top\boldsymbol{\Theta}_A\|_F^2 + 2\lambda G_\alpha([\boldsymbol{\Theta}_B, \boldsymbol{\Theta}_A]).
\end{aligned}
\tag{3.4}
$$

Here we use the fact $G_\alpha([\boldsymbol{\Theta}_B, \boldsymbol{\Theta}_A]) = G_\alpha([\boldsymbol{\Theta}_A, -\boldsymbol{\Theta}_B])$.

For any local minimum of (3.2), namely, $\hat{\boldsymbol{\xi}}$, we are interested in analyzing the estimation error $\|\boldsymbol{M} - \boldsymbol{X}(\hat{\boldsymbol{\xi}})\boldsymbol{Y}(\hat{\boldsymbol{\xi}})^\top\|_F^2$. To this end, given $\tilde{f}(\boldsymbol{\theta})$ is smooth, it is natural to study the stationarity and optimality conditions, i.e., $\nabla^2\tilde{f}(\boldsymbol{\theta}) \succeq \mathbf{0}_{d\times d}$ and $\nabla\tilde{f}(\boldsymbol{\theta}) = \mathbf{0}$. How to employ these two conditions in order to control any local minimum is the key to deriving our main result presented later in this section.

3.1.2. Assumptions. In order to study the estimation error $\|\boldsymbol{M} - \boldsymbol{X}(\hat{\boldsymbol{\xi}})\boldsymbol{Y}(\hat{\boldsymbol{\xi}})^\top\|_F^2$ where $\hat{\boldsymbol{\xi}}$ is any local minimum of the nonconvex program (3.2), we start with some assumptions on the matrix $\boldsymbol{M}$, the parametrization $(\boldsymbol{X}(\boldsymbol{\theta}), \boldsymbol{Y}(\boldsymbol{\theta}))$, and the support of the observed entries Ω.

In this chapter, we assume $\boldsymbol{M}$ is of rank r, and it has reduced singular value decomposition as

$$
\boldsymbol{M} = \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top
$$

where $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0$. The condition number is denoted as $\kappa = \kappa_r := \sigma_1/\sigma_r$. Moreover, the incoherence parameter [**CR09**] $\mu = \mu_r$ for $\boldsymbol{M}$ is denoted as

$$
\mu = \max\left\{\mu(\operatorname{colspan}([\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r])), \mu(\operatorname{colspan}([\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r]))\right\},
$$

where $\mu(\cdot)$ is defined in (2.8).

Next, we assume both $\boldsymbol{X}(\boldsymbol{\theta})$ and $\boldsymbol{Y}(\boldsymbol{\theta})$ to be linear mappings.

Assumption 3.1.1 (Homogeneity and linearity). *Both $\boldsymbol{X}(\boldsymbol{\theta}) \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{Y}(\boldsymbol{\theta}) \in \mathbb{R}^{n_2\times r}$ are homogeneous linear functions in $\boldsymbol{\theta}$, i.e., $\boldsymbol{X}(t\boldsymbol{\theta}_1) = t\boldsymbol{X}(\boldsymbol{\theta}_1)$, $\boldsymbol{Y}(t\boldsymbol{\theta}_1) = t\boldsymbol{Y}(\boldsymbol{\theta}_1)$, $\boldsymbol{X}(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) = \boldsymbol{X}(\boldsymbol{\theta}_1) + \boldsymbol{X}(\boldsymbol{\theta}_2)$ and $\boldsymbol{Y}(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) = \boldsymbol{Y}(\boldsymbol{\theta}_1) + \boldsymbol{Y}(\boldsymbol{\theta}_2)$ for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$ and $t \in \mathbb{R}$.*

As mentioned in the previous section, the next assumption, referred to as the *Correlated Parametric Factorization*, is the key assumption in analyzing the theoretical properties of the local minima of (3.2). It will be verified for low-rank factorization with subspace constraints (1.4) in Section 3.3.1, and for skew-symmetric low-rank factorization (1.5) in Section 3.4.1, respectively.

Assumption 3.1.2 (Correlated Parametric Factorization of $\boldsymbol{M}$). *The rank-r matrix $\boldsymbol{M}$ and the parameterization $(\boldsymbol{X}(\boldsymbol{\theta}), \boldsymbol{Y}(\boldsymbol{\theta}))$ are said to satisfy the correlated parameterized factorization, if for any $\boldsymbol{\theta} \in \mathbb{R}^d$, there exits $\boldsymbol{\xi} \in \mathbb{R}^d$ (not necessarily unique), such that*

$$\begin{cases} \boldsymbol{M} = \boldsymbol{X}(\boldsymbol{\xi})\boldsymbol{Y}(\boldsymbol{\xi})^\top, \\ \boldsymbol{X}(\boldsymbol{\xi})^\top \boldsymbol{X}(\boldsymbol{\xi}) = \boldsymbol{Y}(\boldsymbol{\xi})^\top \boldsymbol{Y}(\boldsymbol{\xi}), \\ \boldsymbol{X}(\boldsymbol{\theta})^\top \boldsymbol{X}(\boldsymbol{\xi}) + \boldsymbol{Y}(\boldsymbol{\theta})^\top \boldsymbol{Y}(\boldsymbol{\xi}) \succeq \mathbf{0}. \end{cases} \tag{3.5}$$

Recall that the support of the observed entries is $\Omega \subset [n_1] \times [n_2]$. For generality, we consider here two scenarios where the entries are observed independently with certain probability p. For rectangular matrix completion, the index set Ω is assumed to follow the independent $\mathrm{Ber}(p)$ model (Model 2.5.1). For square matrix completion ($n_1 = n_2 \coloneqq n$), the index set Ω is assumed to follow the off-diagonal symmetric independent $\mathrm{Ber}(p)$ model (Model 2.1.1)

3.1.3. Theoretical results.

Our main theorem is the following.

Theorem 3.1.3. *Let $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ be a rank-r matrix. The parameters μ and κ are defined in Section 3.1.2. Suppose that $\boldsymbol{M}$, $\boldsymbol{X}(\boldsymbol{\theta})$ and $\boldsymbol{Y}(\boldsymbol{\theta})$ satisfy Assumptions 3.1.1 and 3.1.2, and that Ω, the support of observed entries, satisfies either Model 2.1.1 or 2.5.1. Moreover, let the sampling rate p and the tuning parameters α and λ in (3.1) satisfy the following inequalities:*

$$\begin{cases} p \geqslant C_4 \max\left\{\frac{1}{n_1 \wedge n_2}\mu r \log(n_1 \vee n_2), \frac{n_1 \vee n_2}{(n_1 \wedge n_2)^2}\mu^2 r^2 \kappa^2\right\}, \\ C_5\sqrt{\frac{n_1 \vee n_2}{p}} \leqslant \lambda \leqslant 10C_5\sqrt{\frac{n_1 \vee n_2}{p}}, \\ C_5\sqrt{\frac{\mu r \sigma_1}{n_1 \wedge n_2}} \leqslant \alpha \leqslant 10C_5\sqrt{\frac{\mu r \sigma_1}{n_1 \wedge n_2}}. \end{cases} \tag{3.6}$$

Then, in an event E_3 with probability $\mathbb{P}[E_3] \geqslant 1 - 2(n_1 + n_2)^{-3}$, any local minimum $\hat{\boldsymbol{\xi}}$ of (3.2) satisfies

$$\|\boldsymbol{M} - \widehat{\boldsymbol{M}}\|_F^2 \leqslant \frac{C_6 r}{p^2}\varphi^2,$$

where

$$\varphi := \max_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d} \|\boldsymbol{P}_{\boldsymbol{X}(\boldsymbol{\theta}_1)} \mathcal{P}_\Omega(\boldsymbol{N}) \boldsymbol{P}_{\boldsymbol{Y}(\boldsymbol{\theta}_2)}\|. \tag{3.7}$$

Here C_4, C_5, C_6 are fixed absolute constants.

In particular, if there is no noise, i.e., $\boldsymbol{N} = \mathbf{0}$, then with high probability any local minimum $\hat{\boldsymbol{\xi}}$ leads to $\widehat{\boldsymbol{M}} = \boldsymbol{M}$. In other words, there is no spurious local minimum.

As a simple application, existing results of the landscape analysis for nonconvex positive semidefinite (PSD) matrix completion can be viewed as corollaries of Theorem 3.1.3. In fact, consider the example of low-rank PSD matrix completion, we have $n_1 = n_2 = n$, and the ground truth can be decomposed as $\boldsymbol{M} = \boldsymbol{\Xi}_0 \boldsymbol{\Xi}_0^\top$ for some $\boldsymbol{\Xi}_0 \in \mathbb{R}^{n \times r}$. For any $\boldsymbol{\Theta} \in \mathbb{R}^{n \times r}$, denote $\boldsymbol{\theta} := \text{vec}(\boldsymbol{\Theta})$, and define the linear mappings $\boldsymbol{X}(\boldsymbol{\theta}) = \boldsymbol{Y}(\boldsymbol{\theta}) = \boldsymbol{\Theta}$. This parameterization implies $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{\xi}_0)\boldsymbol{Y}(\boldsymbol{\xi}_0)^\top$ for $\boldsymbol{\xi}_0 := \text{vec}(\boldsymbol{\Xi}_0)$. Assumption 3.1.1 is obviously satisfied. Assumption 3.1.2 can be straightly verified by taking the SVD of $\boldsymbol{\Theta}^\top \boldsymbol{\Xi}_0 = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ and letting $\boldsymbol{\Xi} = \boldsymbol{\Xi}_0 \boldsymbol{V}\boldsymbol{U}^\top$; see, e.g., [**CW15**, Lemma 1].

The factorization now becomes $\boldsymbol{X}(\boldsymbol{\theta})\boldsymbol{Y}(\boldsymbol{\theta})^\top = \boldsymbol{\Theta}\boldsymbol{\Theta}^\top$, so the nonconvex parametric matrix completion (3.2) thereby takes the following form:

$$f_{\text{psd}}(\boldsymbol{\Theta}) := \frac{1}{2p}\left\|\mathcal{P}_\Omega(\boldsymbol{\Theta}\boldsymbol{\Theta}^\top - \boldsymbol{M} - \boldsymbol{N})\right\|_F^2 + 2\lambda G_\alpha(\boldsymbol{\Theta}),$$

which is the nonconvex program that has been used in [**GLM16**, **GJZ17**, **CL19**] for PSD matrix completion. Since Assumptions 3.1.1 and 3.1.2 are verified, as a corollary of Theorem 3.1.3, there are no spurious local minima for the above objective function as long as the tuning parameters are suitably chosen, and the sampling rate satisfies $p \geqslant \frac{C_4}{n} \max\left\{\mu r \log n, \mu^2 r^2 \kappa^2\right\}$, which is exactly the same as the no-spurious result introduced in Chapter 2. Furthermore, consider the special noisy case in which the entries of noise matrix $\boldsymbol{N}$ are i.i.d. Gaussian random variables with mean 0 and variance σ^2. Then $\|\mathcal{P}_\Omega(\boldsymbol{N})\|^2 = O((np + \log^2 n)\sigma^2)$ (see, e.g., [**CW15**, Lemma 11]). Theorem 3.1.3 implies that estimation error bound $\|\boldsymbol{M}-\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Theta}}^\top\|_F^2 = O((\frac{nr}{p}+\frac{r\log^2 n}{p^2})\sigma^2)$, which matches the state-of-the-art results in the literature of noisy matrix completion; see, e.g., [**KMO10b**, **CW15**, **MWCC18**].

In the next two subsections, we explain how to apply Theorem 3.1.3 to studying the theoretical properties of nonconvex optimizations for subspace-constrained and skew-symmetric matrix completions; that is, (3.3) and (3.4).

3.1.4. Nonconvex subspace constrained matrix completion. In the case of matrix completion with subspace constraints, clearly, the linear mappings $\boldsymbol{X}(\boldsymbol{\theta})$ and $\boldsymbol{Y}(\boldsymbol{\theta})$ defined in (1.4) satisfy Assumption 3.1.1. The verification of Assumption 3.1.2 is summarized as the following lemma, the proof of which is given later in Section 3.3.1.

Lemma 3.1.4. *Let $\boldsymbol{M}\in\mathbb{R}^{n_1\times n_2}$ be a rank-r matrix whose column space and row space are constrained in* $\operatorname{colspan}(\widetilde{\boldsymbol{U}})$ *and* $\operatorname{colspan}(\widetilde{\boldsymbol{V}})$. *Then the parameterization $\boldsymbol{X}(\boldsymbol{\theta})$ and $\boldsymbol{Y}(\boldsymbol{\theta})$ defined in* (1.4) *as well as $\boldsymbol{M}\in\mathbb{R}^{n_1\times n_2}$ satisfy Assumption 3.1.2.*

With the assumptions verified, Theorem 3.1.3 implies the following corollary for nonconvex matrix completion with subspace constraints, i.e., (3.3).

Corollary 3.1.5. *Let $\boldsymbol{M}\in\mathbb{R}^{n_1\times n_2}$ be a rank-r matrix. The parameters μ and κ are defined in Section 3.1.2. Assume that the columns of $\widetilde{\boldsymbol{U}}\in\mathbb{R}^{n_1\times s_1}$ constitute an orthonormal basis of the column space constraint for $\boldsymbol{M}$, while the columns of $\widetilde{\boldsymbol{V}}\in\mathbb{R}^{n_2\times s_2}$ constitute an orthonormal basis of the row space constraint. The support of observation, Ω, is assumed to follow from Model 2.5.1, and that the entries of the noise matrix $\boldsymbol{N}$ are i.i.d. centered sub-exponential random variables satisfying the Bernstein condition [**Wai19**, (2.15)] with parameter b and variance ν^2.*

If the sampling rate p and the tuning parameters α,λ satisfy (3.6). *Then, uniformly in an event $E_{subspace}$ with probability $\mathbb{P}[E_{subspace}]\geqslant 1-3(n_1+n_2)^{-3}$, any local minimum $(\widehat{\boldsymbol{\Xi}}_A,\widehat{\boldsymbol{\Xi}}_B)$ of $f_{subspace}(\boldsymbol{\Theta}_A,\boldsymbol{\Theta}_B)$ in* (3.3) *satisfies:*

$$\begin{aligned}
&\|\widetilde{\boldsymbol{U}}\widehat{\boldsymbol{\Xi}}_A\widehat{\boldsymbol{\Xi}}_B^\top\widetilde{\boldsymbol{V}}^\top-\boldsymbol{M}\|_F^2\\
&\leqslant\frac{C_7 r}{p^2}\left(\nu^2 p(s_1+s_2)\log(n_1+n_2)+b^2\frac{\mu_{\widetilde{U}}\mu_{\widetilde{V}}s_1s_2}{n_1n_2}\log^2(n_1+n_2)\right).
\end{aligned}\tag{3.8}$$

Here $\mu_{\widetilde{U}}=\frac{n_1}{s_1}\|\widetilde{\boldsymbol{U}}\|_{2,\infty}^2$, $\mu_{\widetilde{V}}=\frac{n_2}{s_2}\|\widetilde{\boldsymbol{V}}\|_{2,\infty}^2$, and C_7 is some fixed positive absolute constant.

To the best of our knowledge, existing theoretical works on matrix completion with subspace constraints are majorly focused on the noiseless case [YZJ+13, XJZ13, Che15, JD13, EYW18], while statistical convergence rates under the noisy case have not been studied in detail in the literature. Consider again the case that $\boldsymbol{N}$ consists of i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. This gives $b = \sigma$ and variance $\nu^2 = \sigma^2$. For simplicity of discussion, also assume $s_1 = s_2 = s$, $n_1 = n_2 = n$, $\mu = O(1)$, $\mu_{\widetilde{U}} = O(1)$, $\mu_{\widetilde{V}} = O(1)$ and $\kappa = O(1)$. Then Corollary 3.1.5 implies that as long as $p \gtrsim \frac{1}{n} \max\left\{r \log n, r^2\right\}$, there holds $\|\widetilde{\boldsymbol{U}}\widehat{\boldsymbol{\Xi}}_A \widehat{\boldsymbol{\Xi}}_B^\top \widetilde{\boldsymbol{V}}^\top - \boldsymbol{M}\|_F^2 \lesssim \sigma^2 sr(\log n)/p$. We have explained in the previous subsection that the error rates for matrix completion without subspace constraints are $O(\sigma^2 nr/p)$. Therefore, Corollary 3.1.5 indicates that the estimation error can be significantly reduced if the dimensions of the subspace constraints are much lower than the ambient dimensions.

In the noiseless case, we should admit that the sampling rates requirement

$$p \gtrsim \frac{1}{n} \max\left\{r \log n, r^2\right\}$$

is possibly suboptimal, since it is worse than the state-of-the-art sampling rates requirement if convex optimization is employed; see, e.g., [Che15]. This gap is not technically easy to fill, and narrowing it seems beyond the scope of the current paper since Corollary 3.1.5 serves as an example to showcase the usefulness of our main result Theorem 3.1.3. We are interested in narrowing this gap in some future work.

3.1.5. Nonconvex Skew-symmetric Matrix Completion. In the case of rank-r skew-symmetric matrix completion, linear mappings $\boldsymbol{X}(\boldsymbol{\theta})$ and $\boldsymbol{Y}(\boldsymbol{\theta})$ defined in (1.5) evidently satisfies the linearity and homogeneity. Assumption 3.1.2 is verified through the following result with the proof deferred to Section 3.4.1.

Lemma 3.1.6. *Let $\boldsymbol{M}$ be a rank-r skew-symmetric matrix. Then, the parameterization $\boldsymbol{X}(\boldsymbol{\theta})$, $\boldsymbol{Y}(\boldsymbol{\theta})$ defined in* (1.5) *as well as $\boldsymbol{M}$ satisfy Assumption 3.1.2.*

Given Assumption 3.1.2 is justified for the parametric form (1.5), our main result Theorem 3.1.3 implies the following estimation upper bound result for nonconvex skew-symmetric matrix completion (3.4).

THEOREM 3.1.7. *Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a rank-r skew-symmetric matrix. The parameters μ and κ are defined in Section 3.1.2. The support of the observed entries Ω is assumed to follow Model 2.1.1. Assume that the noise matrix $\boldsymbol{N}$ is a skew-symmetric matrix, whose upper triangular part of $\boldsymbol{N}$ consists of i.i.d. centered sub-exponential random variables satisfying the Bernstein condition with parameter b and variance ν^2. Suppose that the sampling rate p and the tuning parameters α and λ satisfy (3.6). Then, uniformly in an event E_{skew} with probability $\mathbb{P}[E_{skew}] \geqslant 1 - 3n^{-3}$, any local minimum $(\widehat{\boldsymbol{\Xi}}_A, \widehat{\boldsymbol{\Xi}}_B)$ of $f_{skew}(\boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B)$ defined in (3.4) satisfies*

$$\|\widehat{\boldsymbol{\Xi}}_A\widehat{\boldsymbol{\Xi}}_B^\top - \widehat{\boldsymbol{\Xi}}_B\widehat{\boldsymbol{\Xi}}_A^\top - \boldsymbol{M}\|_F^2 \leqslant \frac{C_8 r}{p^2}\left(\nu^2 pn\log n + b^2\log^2 n\right).$$

Where C_8 is a fixed positive absolute constant.

As with the discussion in Section 3.1.4, if the upper triangular part of noise matrix $\boldsymbol{N}$ consists of i.i.d. Gaussian random variables with mean 0 and variance σ^2, and the sampling rate satisfis $p \gtrsim \frac{1}{n}\max\left\{r\log n, r^2\right\}$, then the estimation error satisfies $\|\widehat{\boldsymbol{\Xi}}_A\widehat{\boldsymbol{\Xi}}_B^\top - \widehat{\boldsymbol{\Xi}}_B\widehat{\boldsymbol{\Xi}}_A^\top - \boldsymbol{M}\|_F^2 = O(\sigma^2 nr(\log n)/p)$, which is comparable to the aforementioned state-of-the-art result $O(\sigma^2 nr/p)$ up to a logarithmic factor.

3.2. Experiments

In this section, we conduct numerical experiments to test the performance of the proposed nonconvex optimization for parametric matrix completion (3.2). As leading examples, we consider both (3.3) for matrix completion with subspace constraints and (3.4) for skew-symmetric matrix completion.

Notice that in order to implement (3.2), we need the knowledge of p, and to choose α and λ properly. In all simulations, we replaced p in (3.2) with the estimated value $\hat{p} := \frac{|\Omega|}{n_1 n_2}$, and set the parameters as $\lambda = 100\sqrt{(n_1+n_2)\hat{p}}$ and $\alpha = 100$. We solved the nonconvex optimization by gradient descent, and initialized at $\boldsymbol{\theta}_0$ with i.i.d. standard normal entries. At each step of the gradient descent, the step size was selected through line search. To be specific, at each update of $\vec{\theta}$, the step size was set to be $\max\{2^{-k}, 10^{-10}\}$ for $k := \min\{t \mid t = 0,1,2,3,\cdots, \tilde{f}(\boldsymbol{\theta} - 2^{-t}\nabla\tilde{f}(\boldsymbol{\theta})) \leqslant \tilde{f}(\boldsymbol{\theta})\}$. The gradient descent iteration was terminated either after 500 iterations or as soon as the update on $\boldsymbol{\theta}$ satisfied $\|\nabla\tilde{f}(\boldsymbol{\theta}))\|_2^2 \leqslant 10^{-10}$.

In the following subsections, more implementation details of (3.3) and (3.4) are explained, respectively.

3.2.1. Nonconvex matrix completion with subspace constraints. In all implementations of (3.3), we set $n_1 = n_2 = 500$ and $r = 2$. We generated $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{40} \in \mathbb{R}^{500}$ as 40 left singular vectors of a 500×500 random matrix with i.i.d. standard normal entries, and generated $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{40} \in \mathbb{R}^{500}$ similarly from another random matrix with the same distribution. The ground truth was fixed at $\boldsymbol{M} = \boldsymbol{u}_1 \boldsymbol{v}_1^\top + \boldsymbol{u}_2 \boldsymbol{v}_2^\top$ (so $\|\boldsymbol{M}\|_F^2 = 2$). The dimensions of subspace constraints were fixed at $s_1 = s_2 = s = 10, 20, 30, 40$, and we let $\widetilde{\boldsymbol{U}} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s]$ and $\widetilde{\boldsymbol{V}} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_s]$.

In the noisy case, $\boldsymbol{N}$ consisted of i.i.d. Gaussian entries with mean 0 and variance $\sigma^2 = \frac{1}{500^2}$, and so $\mathbb{E}\|\boldsymbol{N}\|_F^2 = 1$. The sampling rate was chosen as $p = 1 \times 0.005, 2 \times 0.005, \ldots, 20 \times 0.005$. For each fixed pair of (p, s), gradient descent was implemented to solve (3.3) with the input $\mathcal{P}_\Omega(\boldsymbol{M} + \boldsymbol{N})$, and the reported relative error was averaged over 10 independent generations of the support of observations Ω and the noise $\boldsymbol{N}$. Figure 3.1 indicates a positive dependency between the the dimension of constraints s and the average estimation error as expected in light of the theoretical result (3.8).

In the noiseless case, an experiment is considered a success if and only if $\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_F / \|\boldsymbol{M}\|_F \leqslant 10^{-3}$. The sampling rate is chosen as $p = 1 \times 10^{-4}, 2 \times 10^{-4}, \ldots, 20 \times 10^{-4}$. Figure 3.2 illustrates a positive dependency between the dimension s and required sample size for consistent successes. As noted before, this dependency has not been explained in our theoretical results, although this phenomenon should be expected in light of the theoretical results for convex approaches [**YZJ**$^+$**13**, **XJZ13**, **Che15**] as well as alternating minimization [**JD13**]. We plan to study this dependency for nonconvex landscape analysis in the future.

3.2.2. Nonconvex skew-symmetric matrix completion. If the ground-truth low rank matrix $\boldsymbol{M}$ is known to be skew-symmetric, we have two nonconvex optimization programs to use in order to recover $\boldsymbol{M}$ from $\mathcal{P}_\Omega(\boldsymbol{M})$: nonconvex skew-symmetric matrix completion (3.4) and the original rectangular matrix completion (3.1). In fact, it has been shown in [**GL11**, Theorem 3] that if the initial input is skew-symmetric, some rectangular matrix completion algorithms, such as singular value projection (SVP) [**JMD10**], will also lead to a skew-symmetric result. This thus

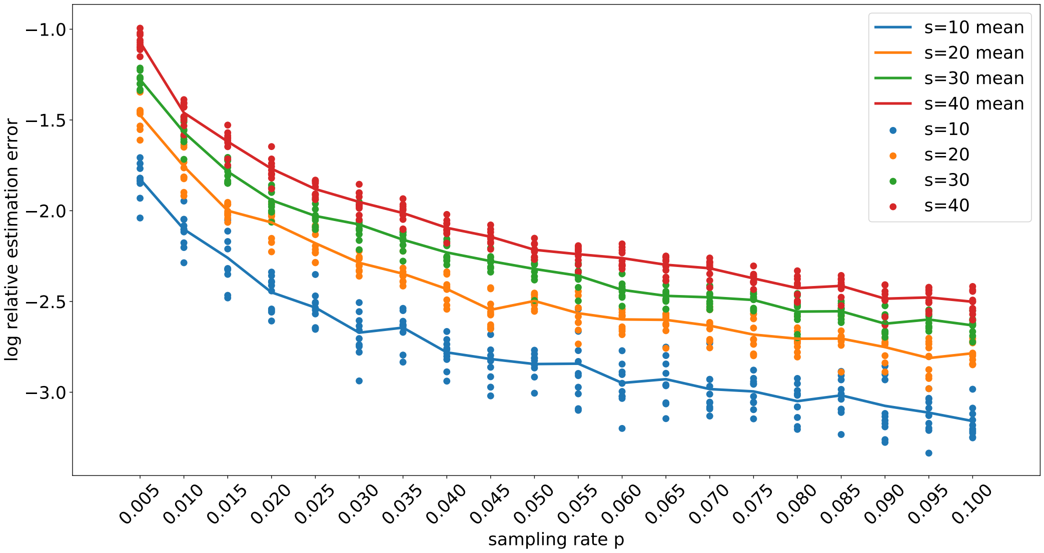


FIGURE 3.1. Logarithm of relative estimation error $\log_{10}(\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_F^2/\|\boldsymbol{M}\|_F^2)$ of nonconvex subspace constrained matrix completion. Here we set the dimension of ground truth $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ as $n_1 = n_2 = 500$, rank of $\boldsymbol{M}$ as $r = 2$, dimension of the column/row subspace constraint as $s_1 = s_2 = s$ and noise level as $\sigma^2 = \frac{1}{n_1 n_2} = \frac{1}{500^2}$. Each dot in the plot represents one trail of the numerical experiment, and the curves represent the mean of 10 independent trials for given s.

raises a natural question: Is there any advantage to use (3.4) over the vanilla approach (3.1)? We make this comparison here empirically by simulations. For the ease of comparison, we focus on the noiseless case.

For all simulations, the matrix size was fixed at $n = 500$ while the rank was fixed at $r = 4, 10, 20$. For each r, we generated r orthonormal vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{r/2}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{r/2} \in \mathbb{R}^{500}$ from left singular vectors of a 500×500 random matrix with i.i.d. standard normal entries. The ground truth was then constructed as

$$\boldsymbol{M} = \boldsymbol{u}_1\boldsymbol{v}_1^\top - \boldsymbol{v}_1\boldsymbol{u}_1^\top + \ldots + \boldsymbol{u}_{r/2}\boldsymbol{v}_{r/2}^\top - \boldsymbol{v}_{r/2}\boldsymbol{u}_{r/2}^\top.$$

The sampling rate was fixed at $p = 1 \times 10^{-2}, 2 \times 10^{-2}, \ldots, 20 \times 10^{-2}$. For each fixed pair (r, p), we generated 10 independent copies of $\Omega \in [500] \times [500]$ from Model 2.1.1. For each simulated data set, we implement both (3.4) and (3.1) with gradient descents. Figure 3.3 plots the relative estimation errors as well as the corresponding medians in logarithmic scale by implementing (3.4) and (3.1),

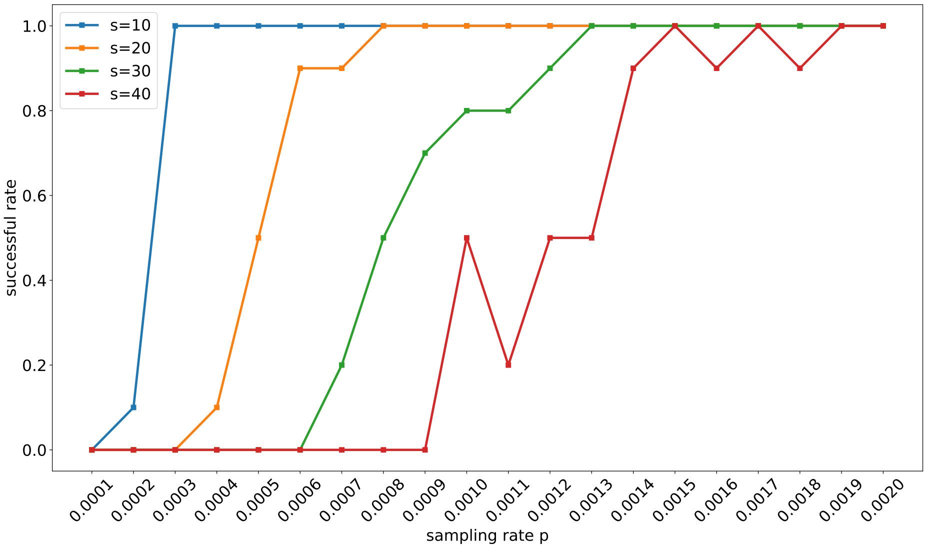


FIGURE 3.2. Rates of success for noiseless nonconvex subspace constrained matrix completion. Here we set the dimension of ground truth $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ as $n_1 = n_2 = 500$, rank of $\boldsymbol{M}$ as $r = 2$, dimension of the column/row subspace constraint as $s_1 = s_2 = s$.

respectively. The comparison indicates that (3.4) and (3.1) are essentially equally successful when the rank of the skew-symmetric matrix is 4 or 10, but (3.4) seems slightly more successful than (3.1) in terms of the empirically required sample sizes in the settings we considered.

3.3. Analysis of nonconvex subspace constrained matrix completion

This section mainly consists of two parts: First we give a proof of Lemma 3.1.4. Then we give a proof of Corollary 3.1.5.

3.3.1. Proof of Lemma 3.1.4.

PROOF. The homogeneous linearity (i.e., Assumption 3.1.1) of $(\boldsymbol{X}(\boldsymbol{\theta}), \boldsymbol{Y}(\boldsymbol{\theta}))$ is directly from the definition (1.4).

In order to show the parameterization satisfies Assumption 3.1.2, we want to show that for any $\boldsymbol{\theta} = \text{vec}([\boldsymbol{\Theta}_A^\top, \boldsymbol{\Theta}_B^\top]^\top) \in \mathbb{R}^{r(s_1+s_2)}$, there exits a $\boldsymbol{\xi} \in \mathbb{R}^{r(s_1+s_2)}$ that satisfies (3.5).

In order to do so, let $\boldsymbol{S} := \widetilde{\boldsymbol{U}}^\top \boldsymbol{M} \widetilde{\boldsymbol{V}}$. Then $\boldsymbol{S} \in \mathbb{R}^{s_1 \times s_2}$. Recall that $\widetilde{\boldsymbol{U}}$ consists of an orthonormal basis of the column space constraint, and $\widetilde{\boldsymbol{V}}$ consists of an orthonormal basis of the column row

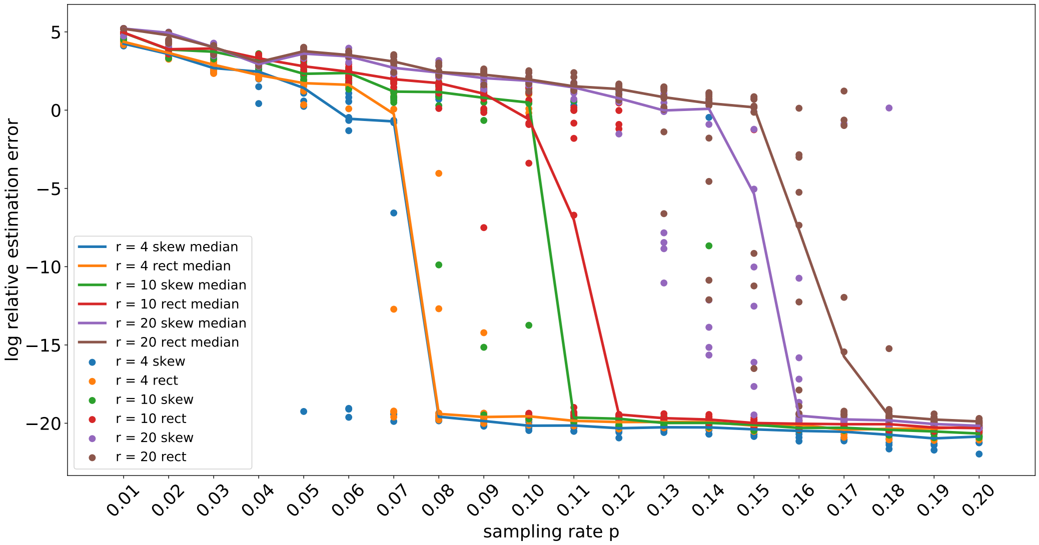


FIGURE 3.3. Logarithm of relative estimation error $\log_{10}(\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_F^2/\|\boldsymbol{M}\|_F^2)$. In this plot, skew-symmetric matrix completion and regular rectangular matrix completion are used to complete noiseless skew-symmetric matrices with rank $r = 4, 10, 20$. Here we set the dimension of ground truth $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ as $n = 500$. Each dot in the plot represents one trail of the numerical experiment, and the curves represent the median of 10 independent trials.

constraint of $\boldsymbol{M}$. Therefore, $\boldsymbol{P}_{\widetilde{\boldsymbol{U}}} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^\top$, $\boldsymbol{P}_{\widetilde{\boldsymbol{V}}} = \widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^\top$ and $\boldsymbol{M}$ can be represented as $\boldsymbol{M} = \widetilde{\boldsymbol{U}}\boldsymbol{S}\widetilde{\boldsymbol{V}}^\top$. Since $\boldsymbol{M}$ is of rank r, by the orthogonality of $\widetilde{\boldsymbol{U}}$ and $\widetilde{\boldsymbol{V}}$, $\operatorname{rank}(\boldsymbol{S}) = r$. Let the reduced SVD of $\boldsymbol{S}$ be

$$\boldsymbol{S} = \boldsymbol{S}_L \boldsymbol{\Lambda} \boldsymbol{S}_R^\top, \tag{3.9}$$

where $\boldsymbol{S}_L \in \mathbb{R}^{s_1\times r}$, $\boldsymbol{S}_R \in \mathbb{R}^{s_2\times r}$, $\boldsymbol{S}_L^\top\boldsymbol{S}_L = \boldsymbol{I}_{s_1}, \boldsymbol{S}_R^\top\boldsymbol{S}_R = \boldsymbol{I}_{s_2}$ and $\boldsymbol{\Lambda} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ is a $r \times r$ diagonal matrix with $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_r$. Moreover, by letting $\boldsymbol{U}^\star := \widetilde{\boldsymbol{U}}\boldsymbol{S}_L \in \mathbb{R}^{n_1\times r}$, $\boldsymbol{V}^\star := \widetilde{\boldsymbol{V}}\boldsymbol{S}_R \in \mathbb{R}^{n_2\times r}$, we can verify that $\boldsymbol{M} = \boldsymbol{U}^\star\boldsymbol{\Lambda}\boldsymbol{V}^{\star\top}$ is a reduced SVD of $\boldsymbol{M}$.

Define

$$\boldsymbol{\Xi}_A^\star := \boldsymbol{S}_L\boldsymbol{\Lambda}^{1/2} \in \mathbb{R}^{s_1\times r}, \quad \boldsymbol{\Xi}_B^\star := \boldsymbol{S}_R\boldsymbol{\Lambda}^{1/2} \in \mathbb{R}^{s_2\times r}.$$

For any $\mathbf{\Theta}_A \in \mathbb{R}^{s_1 \times r}$ and , $\mathbf{\Theta}_B \in \mathbb{R}^{s_2 \times r}$, by considering the SVD of $(\mathbf{\Theta}_A^\top \mathbf{\Xi}_A^\star + \mathbf{\Theta}_B^\top \mathbf{\Xi}_B^\star)$, we know there exits an $r \times r$ orthogonal matrix $\boldsymbol{T} \in O(r)$ [CW15, Lemma 1], such that

$$(\mathbf{\Theta}_A^\top \mathbf{\Xi}_A^\star + \mathbf{\Theta}_B^\top \mathbf{\Xi}_B^\star)\boldsymbol{T} \succeq \mathbf{0}.$$

Let $\boldsymbol{\xi} = \mathrm{vec}([(\mathbf{\Xi}_A^\star \boldsymbol{T})^\top, (\mathbf{\Xi}_B^\star \boldsymbol{T})^\top]^\top)$, then

$$\boldsymbol{X}(\boldsymbol{\xi}) = \widetilde{\boldsymbol{U}}\mathbf{\Xi}_A^\star \boldsymbol{T} \in \mathbb{R}^{n_1 \times r} \quad \text{and} \quad \boldsymbol{Y}(\boldsymbol{\xi}) = \widetilde{\boldsymbol{V}}\mathbf{\Xi}_B^\star \boldsymbol{T} \in \mathbb{R}^{n_2 \times r}.$$

Keeping in mind that both $\widetilde{\boldsymbol{U}}$ and $\widetilde{\boldsymbol{V}}$ are orthonormal basis matrices, the conditions in (3.5) can be verified one by one:

$$\boldsymbol{X}(\boldsymbol{\xi})\boldsymbol{Y}(\boldsymbol{\xi})^\top = \widetilde{\boldsymbol{U}}\mathbf{\Xi}_A^\star \boldsymbol{T}(\widetilde{\boldsymbol{V}}\mathbf{\Xi}_B^\star \boldsymbol{T})^\top = \widetilde{\boldsymbol{U}}\mathbf{\Xi}_A^\star {\mathbf{\Xi}_B^\star}^\top \widetilde{\boldsymbol{V}}^\top = \widetilde{\boldsymbol{U}}\boldsymbol{S}_L \mathbf{\Lambda} \boldsymbol{S}_R^\top \widetilde{\boldsymbol{V}}^\top = \boldsymbol{M}.$$

The last equality is by (3.9).

$$\begin{aligned}\boldsymbol{X}(\boldsymbol{\xi})^\top \boldsymbol{X}(\boldsymbol{\xi}) &= (\widetilde{\boldsymbol{U}}\mathbf{\Xi}_A^\star \boldsymbol{T})^\top \widetilde{\boldsymbol{U}}\mathbf{\Xi}_A^\star \boldsymbol{T} = \boldsymbol{T}^\top {\mathbf{\Xi}_A^\star}^\top \mathbf{\Xi}_A^\star \boldsymbol{T} = \boldsymbol{T}^\top \mathbf{\Lambda}^{1/2} \boldsymbol{S}_L^\top \boldsymbol{S}_L \mathbf{\Lambda}^{1/2} \boldsymbol{T} = \boldsymbol{T}^\top \mathbf{\Lambda} \boldsymbol{T} \\ &= \boldsymbol{T}^\top {\mathbf{\Xi}_B^\star}^\top \mathbf{\Xi}_B^\star \boldsymbol{T} = (\widetilde{\boldsymbol{V}}\mathbf{\Xi}_B^\star \boldsymbol{T})^\top \widetilde{\boldsymbol{V}}\mathbf{\Xi}_B^\star \boldsymbol{T} = \boldsymbol{Y}(\boldsymbol{\xi})^\top \boldsymbol{Y}(\boldsymbol{\xi}).\end{aligned}$$

Here we use the fact $\boldsymbol{S}_L^\top \boldsymbol{S}_L = \boldsymbol{S}_R^\top \boldsymbol{S}_R = \boldsymbol{I}_r$. Moreover,

$$\begin{aligned}&\boldsymbol{X}(\boldsymbol{\theta})^\top \boldsymbol{X}(\boldsymbol{\xi}) + \boldsymbol{Y}(\boldsymbol{\theta})^\top \boldsymbol{Y}(\boldsymbol{\xi}) \\ =&(\widetilde{\boldsymbol{U}}\mathbf{\Theta}_A)^\top \widetilde{\boldsymbol{U}}\mathbf{\Xi}_A^\star \boldsymbol{T} + (\widetilde{\boldsymbol{V}}\mathbf{\Theta}_B)^\top \widetilde{\boldsymbol{V}}\mathbf{\Xi}_B^\star \boldsymbol{T} \\ =&(\mathbf{\Theta}_A^\top \mathbf{\Xi}_A^\star + \mathbf{\Theta}_B^\top \mathbf{\Xi}_B^\star)\boldsymbol{T} \\ \succeq&\mathbf{0}.\end{aligned}$$

Therefore, the parameterization $(\boldsymbol{X}(\boldsymbol{\theta}), \boldsymbol{Y}(\boldsymbol{\theta}))$ satisfies Assumption 3.1.2. □

3.3.2. Proof of Corollary 3.1.5. Since the assumptions of Theorem 3.1.3 are satisfied, therefore, in the event E_3 defined in Theorem 3.1.3,

$$\|\boldsymbol{M} - \widehat{\boldsymbol{M}}\|_F^2 \leqslant \frac{C_6 r}{p^2}\varphi^2.$$

Therefore, it suffices to show that

$$\varphi^2 \leqslant \frac{C_7}{C_6}\left(p(s_1+s_2)\log(n_1+n_2)\nu^2 + b^2\frac{\mu_{\widetilde{U}}\mu_{\widetilde{V}}s_1s_2}{n_1n_2}\log^2(n_1+n_2)\right). \tag{3.10}$$

By (B.6) and (1.4), we have for any $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\operatorname{colspan}(\boldsymbol{X}(\boldsymbol{\theta})) \subset \operatorname{colspan}(\widetilde{\boldsymbol{U}}), \quad \operatorname{colspan}(\boldsymbol{Y}(\boldsymbol{\theta})) \subset \operatorname{colspan}(\widetilde{\boldsymbol{V}}).$$

Therefore, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$,

$$\|\boldsymbol{P}_{\boldsymbol{X}(\boldsymbol{\theta}_1)}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P}_{\boldsymbol{Y}(\boldsymbol{\theta}_2)}\| = \|\boldsymbol{P}_{\boldsymbol{X}(\boldsymbol{\theta}_1)}\boldsymbol{P}_{\widetilde{\boldsymbol{U}}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P}_{\widetilde{\boldsymbol{V}}}\boldsymbol{P}_{\boldsymbol{Y}(\boldsymbol{\theta}_2)}\| \leqslant \|\boldsymbol{P}_{\widetilde{\boldsymbol{U}}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P}_{\widetilde{\boldsymbol{V}}}\|.$$

So we have

$$\varphi \leqslant \|\boldsymbol{P}_{\widetilde{\boldsymbol{U}}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P}_{\widetilde{\boldsymbol{V}}}\|.$$

Therefore, (3.10) can be proved by the following lemma.

LEMMA 3.3.1. *Assume that the support of observation Ω follows from Model 2.5.1. We assume that the entries of the noise matrix $\boldsymbol{N}$ are i.i.d. centered sub-exponential random variables satisfying the Bernstein condition with parameter b and variance ν^2. $\widetilde{\boldsymbol{U}}$ and $\widetilde{\boldsymbol{V}}$ are defined in Chapter 1. Then in an event $E_{subspace\_N}$ with probability $\mathbb{P}[E_{subspace\_N}] \geqslant 1-(n_1+n_2)^{-3}$, we have*

$$\|\boldsymbol{P}_{\widetilde{\boldsymbol{U}}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P}_{\widetilde{\boldsymbol{V}}}\| \leqslant C_w\left(\sqrt{p(s_1+s_2)\log(n_1+n_2)}\nu + b\sqrt{\frac{\mu_{\widetilde{U}}\mu_{\widetilde{V}}s_1s_2}{n_1n_2}}\log(n_1+n_2)\right).$$

for some absolute constant C_w defined in the proof.

The proof of Lemma 3.3.1 is mainly following the discussion in [Wai19, Example 6.18] as well as [Wai19, Example 6.14] and is deferred to appendix.

Letting $E_{\text{subspace}} = E_3 \cap E_{\text{subspace\_N}}$, and $C_7 = 2C_6C_w^2$ finishes the proof.

3.4. Analysis of nonconvex skew-symmetric matrix completion

In this section, we first give a proof of Lemma 3.1.6. Then we give a proof of Theorem 3.1.7.

3.4.1. Proof of Lemma 3.1.6.

Proof. The homogeneous linearity of $(\boldsymbol{X}(\boldsymbol{\theta}), \boldsymbol{Y}(\boldsymbol{\theta}))$ follows directly from the definition (1.5). Therefore, the only thing remains to be verified is that for any $\boldsymbol{\theta} = \text{vec}([\boldsymbol{\Theta}_A^\top, \boldsymbol{\Theta}_B^\top]^\top) \in \mathbb{R}^{nr}$, there exits an $\boldsymbol{\xi} \in \mathbb{R}^{nr}$ that satisfies (3.5).

Recall that $\boldsymbol{M}$ is a rank-r skew-symmetric matrix, where r is even. Then its Youla decomposition [You61] can be written as

$$\boldsymbol{M} \coloneqq \lambda_1 \boldsymbol{\phi}_1 \boldsymbol{\psi}_1^\top - \lambda_1 \boldsymbol{\psi}_1 \boldsymbol{\phi}_1^\top + \lambda_2 \boldsymbol{\phi}_2 \boldsymbol{\psi}_2^\top - \lambda_2 \boldsymbol{\psi}_2 \boldsymbol{\phi}_2^\top + \ldots + \lambda_{r/2} \boldsymbol{\phi}_{r/2} \boldsymbol{\psi}_{r/2}^\top - \lambda_{r/2} \boldsymbol{\psi}_{r/2} \boldsymbol{\phi}_{r/2}^\top,$$

where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_{r/2} > 0$ and $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{r/2}, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{r/2}$ are unit vectors in $\mathbb{R}^n$. Moreover, $\boldsymbol{\phi}_i$'s and $\boldsymbol{\psi}_i$'s are pairwise perpendicular to each other, i.e., for any $i, j \in [r/2]$, $\boldsymbol{\phi}_i^\top \boldsymbol{\psi}_j = 0$, $\boldsymbol{\phi}_i^\top \boldsymbol{\phi}_j = 0$ if $i \neq j$, and $\boldsymbol{\psi}_i^\top \boldsymbol{\psi}_j = 0$ if $i \neq j$.

Let

$$\boldsymbol{\Xi}_A^\star = [\sqrt{\lambda_1}\boldsymbol{\phi}_1, \ldots, \sqrt{\lambda_{r/2}}\boldsymbol{\phi}_{r/2}] \in \mathbb{R}^{n \times \frac{r}{2}} \quad \text{and} \quad \boldsymbol{\Xi}_B^\star = [\sqrt{\lambda_1}\boldsymbol{\psi}_1, \ldots, \sqrt{\lambda_{r/2}}\boldsymbol{\psi}_{r/2}] \in \mathbb{R}^{n \times \frac{r}{2}}.$$

It is straightforward to verify that

$$\boldsymbol{M} = \boldsymbol{\Xi}_A^\star \boldsymbol{\Xi}_B^{\star\top} - \boldsymbol{\Xi}_B^\star \boldsymbol{\Xi}_A^{\star\top}.$$

Recall the fact that for any $i, j \in [r/2]$, $\boldsymbol{\phi}_i^\top \boldsymbol{\psi}_j = 0$; $\boldsymbol{\phi}_i^\top \boldsymbol{\phi}_j = 0$ if $i \neq j$ and $\boldsymbol{\phi}_i^\top \boldsymbol{\phi}_j = 1$ if $i = j$; $\boldsymbol{\psi}_i^\top \boldsymbol{\psi}_j = 0$ if $i \neq j$ and $\boldsymbol{\psi}_i^\top \boldsymbol{\psi}_j = 1$ if $i = j$. Therefore,

$$\boldsymbol{\Xi}_A^{\star\top} \boldsymbol{\Xi}_B^\star = \boldsymbol{0} \text{ and } \boldsymbol{\Xi}_A^{\star\top} \boldsymbol{\Xi}_A^\star = \boldsymbol{\Xi}_B^{\star\top} \boldsymbol{\Xi}_B^\star = \text{diag}(\lambda_1, \ldots, \lambda_{r/2}). \tag{3.11}$$

For any $\boldsymbol{\theta} = \text{vec}([\boldsymbol{\Theta}_A^\top, \boldsymbol{\Theta}_B^\top]^\top)$ with $\boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B \in \mathbb{R}^{n \times \frac{r}{2}}$, consider the singular value decomposition of the complex matrix $(\boldsymbol{\Theta}_A + \sqrt{-1}\boldsymbol{\Theta}_B)^H (\boldsymbol{\Xi}_A^\star + \sqrt{-1}\boldsymbol{\Xi}_B^\star)$ ($\boldsymbol{A}^H$ is conjugate transpose of complex matrix $\boldsymbol{A}$), $(\boldsymbol{\Theta}_A + \sqrt{-1}\boldsymbol{\Theta}_B)^H (\boldsymbol{\Xi}_A^\star + \sqrt{-1}\boldsymbol{\Xi}_B^\star) = \boldsymbol{A}\boldsymbol{D}\boldsymbol{B}^H$, where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{\frac{r}{2} \times \frac{r}{2}}$ are complex unitary matrices and $\boldsymbol{D} \in \mathbb{R}^{\frac{r}{2} \times \frac{r}{2}}$ is a real diagonal matrix. Therefore, $\boldsymbol{B}\boldsymbol{A}^H$ is also a complex unitary matrix, decompose it as $\boldsymbol{B}\boldsymbol{A}^H = \boldsymbol{R}_1 + \sqrt{-1}\boldsymbol{R}_2$ with $\boldsymbol{R}_1, \boldsymbol{R}_2 \in \mathbb{R}^{\frac{r}{2} \times \frac{r}{2}}$. Therefore,

$$(\boldsymbol{\Theta}_A + \sqrt{-1}\boldsymbol{\Theta}_B)^H (\boldsymbol{\Xi}_A^\star + \sqrt{-1}\boldsymbol{\Xi}_B^\star)(\boldsymbol{R}_1 + \sqrt{-1}\boldsymbol{R}_2) = \boldsymbol{A}\boldsymbol{D}\boldsymbol{B}^H\boldsymbol{B}\boldsymbol{A}^H = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^H \succeq \boldsymbol{0},$$

that is, it is a Hermitian positive semidefinite matrix. Let

$$\boldsymbol{\Xi}_A = \boldsymbol{\Xi}_A^\star \boldsymbol{R}_1 - \boldsymbol{\Xi}_B^\star \boldsymbol{R}_2 \text{ and } \boldsymbol{\Xi}_B = \boldsymbol{\Xi}_A^\star \boldsymbol{R}_2 + \boldsymbol{\Xi}_B^\star \boldsymbol{R}_1.$$

Then there holds $(\boldsymbol{\Xi}_A^\star + \sqrt{-1}\boldsymbol{\Xi}_B^\star)(\boldsymbol{R}_1 + \sqrt{-1}\boldsymbol{R}_2) = \boldsymbol{\Xi}_A + \sqrt{-1}\boldsymbol{\Xi}_B$, and

$$(\boldsymbol{\Theta}_A + \sqrt{-1}\boldsymbol{\Theta}_B)^H(\boldsymbol{\Xi}_A + \sqrt{-1}\boldsymbol{\Xi}_B) \succeq \mathbf{0},$$

which is equivalent to the following r-by-r real matrix is positive semidefinite:

$$\begin{bmatrix} \boldsymbol{\Theta}_A^\top \boldsymbol{\Xi}_A + \boldsymbol{\Theta}_B^\top \boldsymbol{\Xi}_B & \boldsymbol{\Theta}_B^\top \boldsymbol{\Xi}_A - \boldsymbol{\Theta}_A^\top \boldsymbol{\Xi}_B \\ \boldsymbol{\Theta}_A^\top \boldsymbol{\Xi}_B - \boldsymbol{\Theta}_B^\top \boldsymbol{\Xi}_A & \boldsymbol{\Theta}_A^\top \boldsymbol{\Xi}_A + \boldsymbol{\Theta}_B^\top \boldsymbol{\Xi}_B \end{bmatrix} \succeq \mathbf{0}.$$

Also, since $\boldsymbol{R}_1 + \sqrt{-1}\boldsymbol{R}_2$ is unitary, we have

$$\boldsymbol{R} := \begin{bmatrix} \boldsymbol{R}_1 & -\boldsymbol{R}_2 \\ \boldsymbol{R}_2 & \boldsymbol{R}_1 \end{bmatrix} \in \mathsf{O}(r).$$

Let $\boldsymbol{\xi} = \operatorname{vec}([\boldsymbol{\Xi}_A^\top, \boldsymbol{\Xi}_B^\top]^\top)$. Then we have

$$\begin{aligned} \boldsymbol{X}(\boldsymbol{\xi}) &= [\boldsymbol{\Xi}_A, -\boldsymbol{\Xi}_B] \\ &= [\boldsymbol{\Xi}_A^\star \boldsymbol{R}_1 - \boldsymbol{\Xi}_B^\star \boldsymbol{R}_2, -\boldsymbol{\Xi}_A^\star \boldsymbol{R}_2 - \boldsymbol{\Xi}_B^\star \boldsymbol{R}_1] \\ &= [\boldsymbol{\Xi}_A^\star, -\boldsymbol{\Xi}_B^\star] \begin{bmatrix} \boldsymbol{R}_1 & -\boldsymbol{R}_2 \\ \boldsymbol{R}_2 & \boldsymbol{R}_1 \end{bmatrix} \\ &= [\boldsymbol{\Xi}_A^\star, -\boldsymbol{\Xi}_B^\star]\, \boldsymbol{R}, \end{aligned}$$

and similarly

$$\begin{aligned} \boldsymbol{Y}(\boldsymbol{\xi}) &= [\boldsymbol{\Xi}_B, \boldsymbol{\Xi}_A] \\ &= [\boldsymbol{\Xi}_A^\star \boldsymbol{R}_2 + \boldsymbol{\Xi}_B^\star \boldsymbol{R}_1, \boldsymbol{\Xi}_A^\star \boldsymbol{R}_1 - \boldsymbol{\Xi}_B^\star \boldsymbol{R}_2] \\ &= [\boldsymbol{\Xi}_B^\star, \boldsymbol{\Xi}_A^\star] \begin{bmatrix} \boldsymbol{R}_1 & -\boldsymbol{R}_2 \\ \boldsymbol{R}_2 & \boldsymbol{R}_1 \end{bmatrix} \\ &= [\boldsymbol{\Xi}_B^\star, \boldsymbol{\Xi}_A^\star]\, \boldsymbol{R}. \end{aligned}$$

It is then straightforward to verify that

$$\boldsymbol{X}(\boldsymbol{\xi})\boldsymbol{Y}(\boldsymbol{\xi})^\top = [\boldsymbol{\Xi}_A, -\boldsymbol{\Xi}_B]\,[\boldsymbol{\Xi}_B, \boldsymbol{\Xi}_A]^\top = [\boldsymbol{\Xi}_A^\star, -\boldsymbol{\Xi}_B^\star]\,[\boldsymbol{\Xi}_B^\star, \boldsymbol{\Xi}_A^\star]^\top = \boldsymbol{M}.$$

In order to further verify $\boldsymbol{X}(\boldsymbol{\xi})^\top \boldsymbol{X}(\boldsymbol{\xi}) = \boldsymbol{Y}(\boldsymbol{\xi})^\top \boldsymbol{Y}(\boldsymbol{\xi})$, it suffices to prove

$$[\boldsymbol{\Xi}_A^\star, -\boldsymbol{\Xi}_B^\star]^\top [\boldsymbol{\Xi}_A^\star, -\boldsymbol{\Xi}_B^\star] = [\boldsymbol{\Xi}_B^\star, \boldsymbol{\Xi}_A^\star]^\top [\boldsymbol{\Xi}_B^\star, \boldsymbol{\Xi}_A^\star],$$

which is guaranteed by ${\boldsymbol{\Xi}_A^\star}^\top \boldsymbol{\Xi}_B^\star = \boldsymbol{0}$ and ${\boldsymbol{\Xi}_A^\star}^\top \boldsymbol{\Xi}_A^\star = {\boldsymbol{\Xi}_B^\star}^\top \boldsymbol{\Xi}_B^\star$ as was shown in (3.11).

Finally, straightforward calculation gives

$$\begin{aligned}\boldsymbol{X}(\boldsymbol{\theta})^\top \boldsymbol{X}(\boldsymbol{\xi}) + \boldsymbol{Y}(\boldsymbol{\theta})^\top \boldsymbol{Y}(\boldsymbol{\xi}) &= [\boldsymbol{\Theta}_A, -\boldsymbol{\Theta}_B]^\top [\boldsymbol{\Xi}_A, -\boldsymbol{\Xi}_B] + [\boldsymbol{\Theta}_B, \boldsymbol{\Theta}_A]^\top [\boldsymbol{\Xi}_B, \boldsymbol{\Xi}_A] \\ &= \begin{bmatrix} \boldsymbol{\Theta}_A^\top \boldsymbol{\Xi}_A + \boldsymbol{\Theta}_B^\top \boldsymbol{\Xi}_B & \boldsymbol{\Theta}_B^\top \boldsymbol{\Xi}_A - \boldsymbol{\Theta}_A^\top \boldsymbol{\Xi}_B \\ \boldsymbol{\Theta}_A^\top \boldsymbol{\Xi}_B - \boldsymbol{\Theta}_B^\top \boldsymbol{\Xi}_A & \boldsymbol{\Theta}_A^\top \boldsymbol{\Xi}_A + \boldsymbol{\Theta}_B^\top \boldsymbol{\Xi}_B \end{bmatrix} \succeq \boldsymbol{0}.\end{aligned}$$

□

3.4.2. Proof of Theorem 3.1.7. Following the lines in Section 3.3.2, it suffices to show that

$$\varphi^2 \leqslant \frac{C_8}{C_6}\left(pn\log n\nu^2 + b^2\log^2 n\right).$$

Recall the fact that $\varphi \leqslant \|\mathcal{P}_\Omega(\boldsymbol{N})\|$, then the proof can be done by employing the following Lemma.

Lemma 3.4.1. *Let the support of the observed entries Ω satisfy Model 2.1.1. We assume that the noise matrix $\boldsymbol{N}$ is a skew-symmetric matrix, and upper triangular part of $\boldsymbol{N}$ consists of i.i.d. centered sub-exponential random variables satisfying the Bernstein condition with parameter b and variance ν^2. Then in an event E_{skew_N} with probability $\mathbb{P}[E_{skew\_N}] \geqslant 1 - n^{-3}$, we have*

$$\|\mathcal{P}_\Omega(\boldsymbol{N})\| \leqslant C_{w'}\left(\sqrt{pn\log n}\nu + b\log n\right).$$

for some absolute constant $C_{w'}$.

The proof is almost exactly the same with proof of Lemma 3.3.1. Therefore, we omit the proof here. We can finish the proof of Theorem 3.1.7 by letting $E_{\text{skew}} = E_3 \cap E_{\text{skew\_N}}$ and $C_8 = 2C_6C_{w'}^2$.

CHAPTER 4

Nonconvex Rectangular Matrix Completion via Gradient Descent without $\ell_{2,\infty}$ Regularization

4.1. Algorithm and Main Results

In this Chapter, the following nonconvex optimization is considered.

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n_1\times r},\boldsymbol{Y}\in\mathbb{R}^{n_2\times r}} f(\boldsymbol{X},\boldsymbol{Y}) := \frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right\|_F^2 + \frac{1}{8}\left\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}\right\|_F^2 \tag{4.1}$$

Our setup for the above nonconvex optimization (4.1) is the same as that in former chapters: the matrix $\boldsymbol{M}$ is of rank-r; the sampling scheme Ω satisfies the i.i.d. Bernoulli model with parameter p, i.e., Model 2.5.1. In Section 4.1.1, we give the formula of the gradient descent. And in Section 4.1.2, we present the main result.

4.1.1. Gradient descent and spectral initialization. We consider the initialization through a simple singular value decomposition: Let

$$\boldsymbol{M}^0 := \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}) \approx \widetilde{\boldsymbol{X}}^0\boldsymbol{\Sigma}^0(\widetilde{\boldsymbol{Y}}^0)^\top \tag{4.2}$$

be the top-r partial singular value decomposition of $\boldsymbol{M}^0$. In other words, the columns of $\widetilde{\boldsymbol{X}}^0 \in \mathbb{R}^{n_1\times r}$ consist of the leading r left singular vectors of $\boldsymbol{M}^0$; the diagonal entries of the diagonal matrix $\boldsymbol{\Sigma}^0 \in \mathbb{R}^{r\times r}$ consist of the corresponding leading r singular values; and the columns of $\widetilde{\boldsymbol{Y}}^0 \in \mathbb{R}^{n_2\times r}$ consist of the corresponding leading r right singular vectors. Let

$$\boldsymbol{X}^0 = \widetilde{\boldsymbol{X}}^0(\boldsymbol{\Sigma}^0)^{1/2}, \quad \boldsymbol{Y}^0 = \widetilde{\boldsymbol{Y}}^0(\boldsymbol{\Sigma}^0)^{1/2}. \tag{4.3}$$

We choose $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ as the initialization for the gradient descent.

The nonconvex optimization (4.1) yields the following formula for gradients:

$$\nabla_X f(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\boldsymbol{Y} + \frac{1}{2}\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}\right),$$

$$\nabla_Y f(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{p}\left[\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right]^\top\boldsymbol{X} + \frac{1}{2}\boldsymbol{Y}\left(\boldsymbol{Y}^\top\boldsymbol{Y} - \boldsymbol{X}^\top\boldsymbol{X}\right).$$

Then the gradient descent algorithm solving (4.1) with some fixed step size η can be explicitly stated as follows:

$$\begin{aligned}\boldsymbol{X}^{t+1} &= \boldsymbol{X}^t - \frac{\eta}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^t\left(\boldsymbol{Y}^t\right)^\top - \boldsymbol{M}\right)\boldsymbol{Y}^t - \frac{\eta}{2}\boldsymbol{X}^t\left(\left(\boldsymbol{X}^t\right)^\top\boldsymbol{X}^t - \left(\boldsymbol{Y}^t\right)^\top\boldsymbol{Y}^t\right),\\ \boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^t - \frac{\eta}{p}\left[\mathcal{P}_\Omega\left(\boldsymbol{X}^t\left(\boldsymbol{Y}^t\right)^\top - \boldsymbol{M}\right)\right]^\top\boldsymbol{X}^t - \frac{\eta}{2}\boldsymbol{Y}^t\left(\left(\boldsymbol{Y}^t\right)^\top\boldsymbol{Y}^t - \left(\boldsymbol{X}^t\right)^\top\boldsymbol{X}^t\right).\end{aligned} \tag{4.4}$$

For any m, we obtain an estimate of $\boldsymbol{M}$ after m iterations as $\widehat{\boldsymbol{M}}^m = \boldsymbol{X}^m(\boldsymbol{Y}^m)^\top$. We aim to study how close the estimate $\widehat{\boldsymbol{M}}^m$ is from the ground truth $\boldsymbol{M}$ under certain assumptions of the sampling complexity.

4.1.2. Main results. In this section, we specify the conditions for $\boldsymbol{M}$ and Ω to guarantee the convergence of the vanilla gradient descent (4.4) with the spectral initialization (4.3). To begin with, we list some necessary assumptions and notations as follows: First, $\operatorname{rank}(\boldsymbol{M}) = r$ is assumed to be known and thereby used in the nonconvex optimization (4.1). The singular value decomposition of $\boldsymbol{M}$ is $\boldsymbol{M} = \widetilde{\boldsymbol{U}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{V}}^\top = \boldsymbol{U}\boldsymbol{V}^\top$ where

$$\boldsymbol{U} = \widetilde{\boldsymbol{U}}(\boldsymbol{\Sigma})^{1/2} \in \mathbb{R}^{n_1\times r}, \quad \text{and} \quad \boldsymbol{V} = \widetilde{\boldsymbol{V}}(\boldsymbol{\Sigma})^{1/2} \in \mathbb{R}^{n_2\times r}.$$

Second, denote by μ the subspace incoherence parameter of the rank-r matrix $\boldsymbol{M}$ as in [CR09], i.e.,

$$\mu := \max(\mu(\operatorname{colspan}(\boldsymbol{U})), \mu(\operatorname{colspan}(\boldsymbol{V})))$$

just like in former chapters. Recall for any r-dimensional subspace $\mathcal{U}$ of $\mathbb{R}^n$, its incoherence parameter is defined as $\mu(\mathcal{U}) := \frac{n}{r}\max\limits_{1\leqslant i\leqslant n}\|\mathcal{P}_{\mathcal{U}}\boldsymbol{e}_i\|_2^2$ with $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n$ being the standard orthogonal basis. Third, denote the condition number of $\boldsymbol{M}$ as $\kappa = \sigma_1(\boldsymbol{M})/\sigma_r(\boldsymbol{M})$, where $\sigma_1(\boldsymbol{M})$ and $\sigma_r(\boldsymbol{M})$ are the first and the r-th singular value of $\boldsymbol{M}$. Finally, assume that there is some absolute constant $C_{10} \geqslant 1$ such that $1/C_{10} \leqslant n_1/n_2 \leqslant C_{10}$. With these assumptions and notations, our main result is stated as follows:

THEOREM 4.1.1. *Let Ω be sampled according to the i.i.d. Bernoulli model with the parameter p. If $p \geqslant C_S \frac{\mu^2 r^2 \kappa^{14} \log(n_1 \vee n_2)}{n_1 \wedge n_2}$ for some absolute constant C_S, then, as long as the gradient descent step size η in (4.4) satisfies $\eta \leqslant \frac{\sigma_r(\boldsymbol{M})}{200\sigma_1^2(\boldsymbol{M})}$, in an event E with probability $\mathbb{P}[E] \geqslant 1-(n_1+n_2)^{-3}$, the gradient descent iteration (4.4) starting from the spectral initialization (4.3) converges linearly for at least the first $(n_1+n_2)^3$ steps:*

$$\min_{\boldsymbol{R}\in\mathsf{O}(r)} \left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\|_F \leqslant \rho^t \sqrt{\sigma_r(\boldsymbol{M})},$$

$0 \leqslant t \leqslant (n_1+n_2)^3$. Here $\mathsf{O}(r)$ denotes the set of $r \times r$ orthogonal matrices, and $\rho \coloneqq 1-0.05\eta\sigma_r(\boldsymbol{M})$ satisfies $0 < \rho < 1$. If additionally assume $\eta \geqslant \frac{\sigma_r(\boldsymbol{M})}{1000\sigma_1^2(\boldsymbol{M})}$, the above inequality implies

$$\min_{\boldsymbol{R}\in\mathsf{O}(r)} \left\| \begin{bmatrix} \boldsymbol{X}^T \\ \boldsymbol{Y}^T \end{bmatrix} \boldsymbol{R} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\|_F \leqslant e^{-(n_1+n_2)^3/C_R} \sqrt{\sigma_r(\boldsymbol{M})}$$

for $T \coloneqq (n_1+n_2)^3$ and an absolute constant $C_R > 0$.

4.2. The Leave-one-out Sequences and the Roadmap of the Proof

The proof framework of Theorem 4.1.1 relies crucially on extending the leave-one-out sequences in [**MWCC18**] from positive definite matrix completion to rectangular matrix completion (4.1). Roughly speaking, the proof consists of three major parts: some local properties for the Hessian of the nonconvex objective function $f(\boldsymbol{X},\boldsymbol{Y})$ defined in (4.1), error bounds for the initialization $(\boldsymbol{X}^0,\boldsymbol{Y}^0)$ and those of the leave-one-out sequences $(\boldsymbol{X}^{0,(l)},\boldsymbol{Y}^{0,(l)})$, error bounds for the gradient sequence $(\boldsymbol{X}^t,\boldsymbol{Y}^t)$ and the leave-one-out sequences $(\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)})$. We first give the definition of the leave-one-out sequences rigorously.

4.2.1. Leave-one-out sequences.

Let's start with the following notations:

- Denote by $\Omega_{-i,\cdot} \coloneqq \{(k,l) \in \Omega : k \neq i\}$ the subset of Ω where entries in the i-th row are removed;

- Denote by $\Omega_{\cdot,-j} \coloneqq \{(k,l) \in \Omega : l \neq j\}$ the subset of Ω where entries in the j-th column are removed;
- Denote by $\Omega_{i,\cdot} \coloneqq \{(i,k) \in \Omega\}$ the subset of Ω where only entries in the i-th row are kept;
- Denote by $\Omega_{\cdot,j} \coloneqq \{(k,j) \in \Omega\}$ the subset of Ω where only entries in the j-th column are kept;
- The definitions of the projectors $\mathcal{P}_{\Omega_{-i,\cdot}}$, $\mathcal{P}_{\Omega_{\cdot,-j}}$, $\mathcal{P}_{\Omega_{i,\cdot}}$ and $\mathcal{P}_{\Omega_{\cdot,j}}$ are similar to that of $\mathcal{P}_\Omega$;
- Denote by $\mathcal{P}_{i,\cdot}(\cdot)/\mathcal{P}_{\cdot,j}(\cdot) : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ the orthogonal projector that transforms a matrix by keeping its i-th row/j-th column and setting all other entries into zeros:

$$(\mathcal{P}_{i,\cdot}(\boldsymbol{M}))_{k,l} = \begin{cases} M_{k,l} & \text{if } k = i \\ 0 & \text{otherwise,} \end{cases}, \quad (\mathcal{P}_{\cdot,j}(\boldsymbol{M}))_{k,l} = \begin{cases} M_{k,l} & \text{if } l = j \\ 0 & \text{otherwise.} \end{cases}$$

These notations facilitate the leave-one-out analysis in rectangular matrix completion, in which each row/column is associated with a separate "leave-one-out" sequence. The initialization for the "leave-one-out" sequences are defined similarly to the initialization $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ for the gradient descent flow. To be concrete, for the i-th row, define

$$\boldsymbol{M}^{0,(i)} \coloneqq \frac{1}{p}\mathcal{P}_{\Omega_{-i,\cdot}}(\boldsymbol{M}) + \mathcal{P}_{i,\cdot}(\boldsymbol{M}),$$

i.e., the i-th row of $\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M})$ is replaced with the complete i-th row of $\boldsymbol{M}$. Similarly, for the j-th column, define

$$\boldsymbol{M}^{0,(n_1+j)} \coloneqq \frac{1}{p}\mathcal{P}_{\Omega_{\cdot,-j}}(\boldsymbol{M}) + \mathcal{P}_{\cdot,j}(\boldsymbol{M}),$$

i.e., the j-th column of $\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M})$ is replaced with the complete j-th column of $\boldsymbol{M}$. In short, we write

$$\boldsymbol{M}^{0,(l)} \coloneqq \begin{cases} \left(\frac{1}{p}\mathcal{P}_{\Omega_{-l,\cdot}} + \mathcal{P}_{l,\cdot}\right)(\boldsymbol{M}) & 1 \leqslant l \leqslant n_1 \\ \left(\frac{1}{p}\mathcal{P}_{\Omega_{\cdot,-(l-n_1)}} + \mathcal{P}_{\cdot,l-n_1}\right)(\boldsymbol{M}) & n_1 + 1 \leqslant l \leqslant n_1 + n_2. \end{cases} \tag{4.5}$$

For $1 \leqslant l \leqslant n_1+n_2$, as with the spectral initialization for gradient descent, let $\widetilde{\boldsymbol{X}}^{0,(l)}\boldsymbol{\Sigma}^{0,(l)}\left(\widetilde{\boldsymbol{Y}}^{0,(l)}\right)^\top$ be top-r partial singular value decomposition of $\boldsymbol{M}^{0,(l)}$. Further, as with the definition of $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$

in (4.3), we define the initialization for the l-th leave-one-out sequence as

$$\boldsymbol{X}^{0,(l)} = \widetilde{\boldsymbol{X}}^{0,(l)} \left(\boldsymbol{\Sigma}^{0,(l)}\right)^{1/2}, \quad \boldsymbol{Y}^{0,(l)} = \widetilde{\boldsymbol{Y}}^{0,(l)} \left(\boldsymbol{\Sigma}^{0,(l)}\right)^{1/2}. \tag{4.6}$$

It is clear that if $1 \leqslant l \leqslant n_1$, $(\boldsymbol{X}^{0,(l)}, \boldsymbol{Y}^{0,(l)})$ is the initialization for the leave-one-out sequence associated with the l-th row, while if $n_1 + 1 \leqslant l \leqslant n_1 + n_2$, $(\boldsymbol{X}^{0,(l)}, \boldsymbol{Y}^{0,(l)})$ is associated with the $(l - n_1)$-th column.

Starting with $(\boldsymbol{X}^{0,(l)}, \boldsymbol{Y}^{0,(l)})$, we define the l-th leave-one-out sequence by considering the corresponding modification of the nonconvex optimization (4.1). For $1 \leqslant l \leqslant n_1$, the nonconvex optimization (4.1) is modified as

$$\min_{\substack{\boldsymbol{X}\in\mathbb{R}^{n_1\times r}\\ \boldsymbol{Y}\in\mathbb{R}^{n_2\times r}}} f(\boldsymbol{X}, \boldsymbol{Y}) := \frac{1}{2p} \left\| \left(\mathcal{P}_{\Omega_{-l,\cdot}} + p\mathcal{P}_{l,\cdot}\right) \left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right) \right\|_F^2 + \frac{1}{8} \left\| \boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y} \right\|_F^2 .$$

The leave-one-out sequence associated with the l-th row is defined as the corresponding gradient descent sequence with the same step size η:

$$\begin{aligned} \boldsymbol{X}^{t+1,(l)} =& \boldsymbol{X}^{t,(l)} - \frac{\eta}{p}\mathcal{P}_{\Omega_{-l,\cdot}} \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{M}\right) \boldsymbol{Y}^{t,(l)} - \eta\mathcal{P}_{l,\cdot} \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{M}\right) \boldsymbol{Y}^{t,(l)} \\ & - \frac{\eta}{2}\boldsymbol{X}^{t,(l)} \left((\boldsymbol{X}^{t,(l)})^\top\boldsymbol{X}^{t,(l)} - (\boldsymbol{Y}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)}\right) \end{aligned} \tag{4.7}$$

and

$$\begin{aligned} \boldsymbol{Y}^{t+1,(l)} =& \boldsymbol{Y}^{t,(l)} - \frac{\eta}{p} \left[\mathcal{P}_{\Omega_{-l,\cdot}} \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{M}\right)\right]^\top \boldsymbol{X}^{t,(l)} \\ & - \eta \left[\mathcal{P}_{l,\cdot} \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{M}\right)\right]^\top \boldsymbol{X}^{t,(l)} \\ & - \frac{\eta}{2}\boldsymbol{Y}^{t,(l)} \left((\boldsymbol{Y}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)} - (\boldsymbol{X}^{t,(l)})^\top\boldsymbol{X}^{t,(l)}\right) \end{aligned} \tag{4.8}$$

Similarly, for $n_1 + 1 \leqslant l \leqslant n_1 + n_2$, consider the nonconvex optimization

$$\min_{\substack{\boldsymbol{X}\in\mathbb{R}^{n_1\times r}\\ \boldsymbol{Y}\in\mathbb{R}^{n_2\times r}}} f(\boldsymbol{X}, \boldsymbol{Y}) := \frac{1}{2p} \left\| \left(\mathcal{P}_{\Omega_{\cdot,-(l-n_1)}} + p\mathcal{P}_{\cdot,l-n_1}\right) \left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right) \right\|_F^2 + \frac{1}{8} \left\| \boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y} \right\|_F^2 .$$

Subsequently, the leave-one-out sequence associated with the $(l - n_1)$-th column is defined as the sequence:

$$\begin{aligned} \boldsymbol{X}^{t+1,(l)} =& \boldsymbol{X}^{t,(l)} - \frac{\eta}{p}\mathcal{P}_{\Omega_{\cdot,-(l-n_1)}}\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{M}\right)\boldsymbol{Y}^{t,(l)} \\ & - \eta\mathcal{P}_{\cdot,l-n_1}\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{M}\right)\boldsymbol{Y}^{t,(l)} \\ & - \frac{\eta}{2}\boldsymbol{X}^{t,(l)}\left((\boldsymbol{X}^{t,(l)})^\top\boldsymbol{X}^{t,(l)} - (\boldsymbol{Y}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)}\right) \end{aligned} \tag{4.9}$$

and

$$\begin{aligned} \boldsymbol{Y}^{t+1,(l)} =& \boldsymbol{Y}^{t,(l)} - \frac{\eta}{p}\left[\mathcal{P}_{\Omega_{\cdot,-(l-n_1)}}\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{M}\right)\right]^\top\boldsymbol{X}^{t,(l)} \\ & - \eta\left[\mathcal{P}_{\cdot,l-n_1}\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{M}\right)\right]^\top\boldsymbol{X}^{t,(l)} \\ & - \frac{\eta}{2}\boldsymbol{Y}^{t,(l)}\left((\boldsymbol{Y}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)} - (\boldsymbol{X}^{t,(l)})^\top\boldsymbol{X}^{t,(l)}\right). \end{aligned} \tag{4.10}$$

These $n_1 + n_2$ leave-one-out sequences will be employed to prove the convergence of vanilla gradient descent (4.4) as with [**MWCC18**] as will be detailed in next few sections.

4.2.2. Local properties of the Hessian. As with [**MWCC18**, Lemma 7], we characterize some local properties of the Hessian of the objective function $f(\boldsymbol{X}, \boldsymbol{Y})$:

LEMMA 4.2.1. *If the sampling rate satisfies*

$$p \geqslant C_{S1}\frac{\mu r\kappa\log(n_1 \vee n_2)}{n_1 \wedge n_2}$$

for some absolute constant C_{S1}, then on an event E_H with probability $\mathbb{P}[E_H] \geqslant 1 - 3(n_1 + n_2)^{-11}$, we have

$$\operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right)^\top \nabla^2 f(\boldsymbol{X}, \boldsymbol{Y}) \operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right) \geqslant \frac{1}{5}\sigma_r(\boldsymbol{M})\left\|\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right\|_F^2 \tag{4.11}$$

and

$$\|\nabla^2 f(\boldsymbol{X}, \boldsymbol{Y})\| \leqslant 5\sigma_1(\boldsymbol{M}), \tag{4.12}$$

uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}$ *satisfying*

$$\left\| \begin{bmatrix} \boldsymbol{X} - \boldsymbol{U} \\ \boldsymbol{Y} - \boldsymbol{V} \end{bmatrix} \right\|_{2,\infty} \leqslant \frac{1}{500\kappa\sqrt{n_1+n_2}}\sqrt{\sigma_1(\boldsymbol{M})} \tag{4.13}$$

and all $\boldsymbol{D_X} \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{D_Y} \in \mathbb{R}^{n_2 \times r}$ *such that* $\begin{bmatrix} \boldsymbol{D_X} \\ \boldsymbol{D_Y} \end{bmatrix}$ *is in the set*

(4.14)
$$\left\{ \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{Y}_1 \end{bmatrix} \widehat{\boldsymbol{R}} - \begin{bmatrix} \boldsymbol{X}_2 \\ \boldsymbol{Y}_2 \end{bmatrix} : \left\| \begin{bmatrix} \boldsymbol{X}_2 - \boldsymbol{U} \\ \boldsymbol{Y}_2 - \boldsymbol{V} \end{bmatrix} \right\| \leqslant \frac{\sqrt{\sigma_1(\boldsymbol{M})}}{500\kappa}, \widehat{\boldsymbol{R}} := \operatorname*{argmin}_{\boldsymbol{R} \in \mathsf{O}(r)} \left\| \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{Y}_1 \end{bmatrix} \boldsymbol{R} - \begin{bmatrix} \boldsymbol{X}_2 \\ \boldsymbol{Y}_2 \end{bmatrix} \right\|_F \right\}.$$

The proof is similar to [**MWCC18**, Lemma 7], but as mentioned in Chapter 1, we apply Lemma 4.4 from [**CL19**] and Lemma 9 from [**ZL16**] to improve the order of logarithms. The details are relegated to Section C.1.

4.2.3. Analysis of the initializations for the Leave-one-out sequences. As with Lemma 13 in [**MWCC18**], we now specify how close the spectral initialization $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ in (4.3) and its leave-one-out counterparts $(\boldsymbol{X}^{0,(l)}, \boldsymbol{Y}^{0,(l)})$ in (4.6) are from the ground truth $(\boldsymbol{U}, \boldsymbol{V})$ (recall that $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{V}^\top$). To begin with, we list some convenient notations for several orthogonal matrices that relate $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$, $(\boldsymbol{X}^{0,(l)}, \boldsymbol{Y}^{0,(l)})$ and $(\boldsymbol{U}, \boldsymbol{V})$:

$$\boldsymbol{R}^0 := \operatorname*{argmin}_{\boldsymbol{R} \in \mathsf{O}(r)} \left\| \begin{bmatrix} \boldsymbol{X}^0 \\ \boldsymbol{Y}^0 \end{bmatrix} \boldsymbol{R} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\|_F,$$

$$\boldsymbol{R}^{0,(l)} := \operatorname*{argmin}_{\boldsymbol{R} \in \mathsf{O}(r)} \left\| \begin{bmatrix} \boldsymbol{X}^{0,(l)} \\ \boldsymbol{Y}^{0,(l)} \end{bmatrix} \boldsymbol{R} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\|_F$$

and

$$\boldsymbol{T}^{0,(l)} := \operatorname*{argmin}_{\boldsymbol{R} \in \mathsf{O}(r)} \left\| \begin{bmatrix} \boldsymbol{X}^0 \\ \boldsymbol{Y}^0 \end{bmatrix} \boldsymbol{R}^0 - \begin{bmatrix} \boldsymbol{X}^{0,(l)} \\ \boldsymbol{Y}^{0,(l)} \end{bmatrix} \boldsymbol{R} \right\|_F. \tag{4.15}$$

Lemma 4.2.2. *If*

$$p \geqslant C_{S2} \frac{\mu^2 r^2 \kappa^6 \log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

then on an event $E_{init} \subset E_H$ (defined in Lemma 4.2.1) with probability $\mathbb{P}[E_{init}] \geqslant 1-(n_1+n_2)^{-10}$, there hold the following inequalities

$$\left\| \begin{bmatrix} \boldsymbol{X}^0 \\ \boldsymbol{Y}^0 \end{bmatrix} \boldsymbol{R}^0 - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \leqslant C_I \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2)p}} \sqrt{\sigma_1(\boldsymbol{M})}, \tag{4.16}$$

$$\left\| \left(\begin{bmatrix} \boldsymbol{X}^{0,(l)} \\ \boldsymbol{Y}^{0,(l)} \end{bmatrix} \boldsymbol{R}^{0,(l)} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right)_{l,\cdot} \right\|_2 \leqslant 100 C_I \sqrt{\frac{\mu^2 r^2 \kappa^7 \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})}, \tag{4.17}$$

$$\left\| \begin{bmatrix} \boldsymbol{X}^0 \\ \boldsymbol{Y}^0 \end{bmatrix} \boldsymbol{R}^0 - \begin{bmatrix} \boldsymbol{X}^{0,(l)} \\ \boldsymbol{Y}^{0,(l)} \end{bmatrix} \boldsymbol{T}^{0,(l)} \right\|_F \leqslant C_I \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})}. \tag{4.18}$$

For $l = 1, 2, \cdots, n_1+n_2$. Here C_I and C_{S2} are two fixed absolute constants.

The detailed proof of Lemma 4.2.2 is deferred to Appendix C.2, while we here highlight some key ideas in the proof. First, in order to transform the problem of rectangular matrix completion into symmetric matrix completion, the trick of "symmetric dilation" introduced in [**Pau02**, **AFWZ17**] is employed. Moreover, a major technical novelty in our proof is to replace [**MWCC18**, Lemma 39] with [**Che15**, Lemma 2] to obtain sharper error bounds as shown in (4.16), (4.17) and (4.18). We restate that lemma here:

Lemma 4.2.3 (Modification of [**Che15**, Lemma 2]). *Let $\boldsymbol{A}$ be any fixed $n_1 \times n_2$ matrix, and let the index set $\Omega \in [n_1] \times [n_2]$ satisfy the i.i.d. Bernoulli model with parameter p. Denote*

$$\overline{\boldsymbol{A}} \coloneqq \begin{bmatrix} \boldsymbol{0} & \boldsymbol{A} \\ \boldsymbol{A}^\top & \boldsymbol{0} \end{bmatrix},$$

$$\overline{\Omega} \coloneqq \{(i,j) \,|1 \leqslant i,j \leqslant n_1+n_2, (i, j-n_1) \in \Omega \text{ or } (j, i-n_1) \in \Omega\}.$$

There is an absolute constant C_{14} and an event E_{Ch} with probability $\mathbb{P}[E_{Ch}] \geqslant 1-(n_1+n_2)^{-11}$, such that for all $1 \leqslant l \leqslant n_1+n_2$, there holds

$$\begin{aligned} \left\| \frac{1}{p}\mathcal{P}_{\overline{\Omega}_{-l}}(\overline{\boldsymbol{A}}) + \mathcal{P}_l(\overline{\boldsymbol{A}}) - \overline{\boldsymbol{A}} \right\| &\leqslant \left\| \frac{1}{p}\mathcal{P}_{\overline{\Omega}}(\overline{\boldsymbol{A}}) - \overline{\boldsymbol{A}} \right\| \\ &\leqslant C_{14} \left(\frac{\log(n_1 \vee n_2)}{p} \|\overline{\boldsymbol{A}}\|_{\ell_\infty} + \sqrt{\frac{\log(n_1 \vee n_2)}{p}} \|\overline{\boldsymbol{A}}\|_{2,\infty} \right). \end{aligned} \tag{4.19}$$

Here

$$\mathcal{P}_{\overline{\Omega}_{-l}}(\overline{\boldsymbol{A}}) := \sum_{(i,j)\in\overline{\Omega}, i\neq l, j\neq l} \overline{A}_{i,j}\boldsymbol{e}_i\boldsymbol{e}_j^\top,$$
$$\mathcal{P}_l(\overline{\boldsymbol{A}}) := \sum_{(i,j)\in[n_1+n_2]\times[n_1+n_2], i=l \text{ or } j=l} \overline{A}_{i,j}\boldsymbol{e}_i\boldsymbol{e}_j^\top,$$

and $\boldsymbol{e}_1, \dots \boldsymbol{e}_{n_1+n_2}$ *are the standard basis of* $\mathbb{R}^{n_1+n_2}$.

The second inequality in (4.19) is directly implied by [Che15]. In fact, [Che15] yields the bound for $\left\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{A}) - \boldsymbol{A}\right\|$. On the other hand, the equalities

$$\begin{aligned}\left\|\frac{1}{p}\mathcal{P}_{\overline{\Omega}}(\overline{\boldsymbol{A}}) - \overline{\boldsymbol{A}}\right\| &= \left\|\begin{bmatrix} \mathbf{0} & \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{A}) - \boldsymbol{A} \\ \left(\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{A}) - \boldsymbol{A}\right)^\top & \mathbf{0}\end{bmatrix}\right\| \\ &= \left\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{A}) - \boldsymbol{A}\right\|\end{aligned}$$

as well as $\|\overline{\boldsymbol{A}}\|_{\ell_\infty} = \|\boldsymbol{A}\|_{\ell_\infty}$ and $\|\overline{\boldsymbol{A}}\|_{2,\infty} = \max\{\|\boldsymbol{A}\|_{2,\infty}, \|\boldsymbol{A}^\top\|_{2,\infty}\}$ translate the bound in [Che15] to our result. As to the first inequality in (4.19), it holds due simply to the fact that $\frac{1}{p}\mathcal{P}_{\overline{\Omega}_{-l}}(\overline{\boldsymbol{A}}) + \mathcal{P}_l(\overline{\boldsymbol{A}}) - \overline{\boldsymbol{A}}$ is essentially a submatrix of $\frac{1}{p}\mathcal{P}_{\overline{\Omega}}(\overline{\boldsymbol{A}}) - \overline{\boldsymbol{A}}$ (the l-th column and l-th row are changed to zeros.)

4.2.4. Analysis for the leave-one-out sequences. In this section we are about to introduce the lemma that guarantees the convergence of the gradient descent for the nonconvex optimization (4.1) with the leave-one-out technique. To be concrete, we are going to control certain distances between the gradient descent sequence $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ in (4.4), the leave-one-out sequences $(\boldsymbol{X}^{t,(l)}, \boldsymbol{Y}^{t,(l)})$ in (4.7), (4.8), (4.9) and (4.10), and the low-rank factors $(\boldsymbol{U}, \boldsymbol{V})$. Again, we denote some orthogonal matrices that relate $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$, $(\boldsymbol{X}^{t,(l)}, \boldsymbol{Y}^{t,(l)})$ and $(\boldsymbol{U}, \boldsymbol{V})$ for $1 \leqslant l \leqslant n_1 + n_2$:

$$
\begin{aligned}
\boldsymbol{R}^t &:= \operatorname*{argmin}_{\boldsymbol{R} \in \mathsf{O}(r)} \left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\|_F, \\
\boldsymbol{R}^{t,(l)} &:= \operatorname*{argmin}_{\boldsymbol{R} \in \mathsf{O}(r)} \left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{R} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\|_F, \\
\boldsymbol{T}^{t,(l)} &:= \operatorname*{argmin}_{\boldsymbol{R} \in \mathsf{O}(r)} \left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{R} \right\|_F.
\end{aligned}
\tag{4.20}
$$

LEMMA 4.2.4. *Suppose that the the step size satisfies*

$$\eta \leqslant \frac{\sigma_r(\boldsymbol{M})}{200\sigma_1^2(\boldsymbol{M})},$$

and that the sampling rate satisfies

$$p \geqslant C_{S3} \frac{\mu^2 r^2 \kappa^{14} \log(n_1 \vee n_2)}{n_1 \wedge n_2}$$

for some absolute constant C_{S3}.

For any fixed $t \geqslant 0$, if on an event $E_{gd}^t \subset E_H$ (defined in Lemma 4.2.1) there hold

$$\left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \leqslant C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} \sqrt{\sigma_1(\boldsymbol{M})}, \tag{4.21}$$

$$\left\| \left(\begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{R}^{t,(l)} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right)_{l,\cdot} \right\|_2 \leqslant 100 C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})}, \tag{4.22}$$

$$\left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} \right\|_F \leqslant C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})}, \tag{4.23}$$

$$\left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\|_{2,\infty} \leqslant 110 C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{12} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})}, \tag{4.24}$$

for all $1 \leqslant l \leqslant n_1 + n_2$*, where* C_I *is the absolute constant defined in Lemma 4.2.2 and* $\rho := 1 - 0.05\eta\sigma_r(\boldsymbol{M})$*, then on an event* $E_{gd}^{t+1} \subset E_{gd}^{t}$ *satisfying* $\mathbb{P}[E_{gd}^{t} \backslash E_{gd}^{t+1}] \leqslant (n_1+n_2)^{-10}$*, the above inequalities (4.21), (4.22), (4.23) and (4.24) also hold for* $t+1$*.*

If we translate the inequalities (70) in [**MWCC18**] in terms of $\sqrt{\sigma_1(\boldsymbol{M})}$, a straightforward comparison shows that our bounds are $O(\sqrt{r})$ tighter. Our key technical novelty for this improvement has been summarized in Chapter 1 and is thereby omitted here. The detailed proof is deferred to Section 4.3.

4.2.5. Proof of the main theorem. We are now ready to give a proof for the main theorem based upon the above lemmas:

Proof of Theorem 4.1.1. We choose $C_S = C_{S2} + C_{S3} + 2C_I^2$ where C_{S2}, C_{S3} and C_I are defined in Lemma 4.2.2 and 4.2.4. Then the requirements on the sampling rate p in both Lemma 4.2.2 and 4.2.4 are satisfied. By Lemma 4.2.2, the inequalities (4.16), (4.17) and (4.18) hold on the event E_{init} defined there, which implies that the inequalities (4.21), (4.22) and (4.23) hold for $t=0$ on E_{init}. Moreover, (4.24) can be straightforwardly implied by (4.21), (4.22) and (4.23) (the proof is deferred to Section 4.3.5), and thereby also holds for $t=0$. Let $E_{gd}^0 = E_{init}$. By applying Lemma 4.2.4 iteratively for $t = 1, 2, \ldots, (n_1+n_2)^3$, we know on an event $E := E_{gd}^{(n_1+n_2)^3} \subset \cdots \subset E_{gd}^0 = E_{init}$ there holds

$$\left\|\begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}\right\| \leqslant C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2)p}} \sqrt{\sigma_1(\boldsymbol{M})}$$

for all t satisfying $0 \leqslant t \leqslant (n_1+n_2)^3$ and $\rho = 1 - 0.05\eta\sigma_r(\boldsymbol{M})$. This further implies that

$$\begin{aligned} \left\|\begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}\right\|_F &\leqslant \sqrt{2r} \left\|\begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}\right\| \\ &\leqslant \sqrt{2r} C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2)p}} \sqrt{\sigma_1(\boldsymbol{M})} \\ &\leqslant \rho^t \sqrt{\sigma_r(\boldsymbol{M})}, \end{aligned} \tag{4.25}$$

where the last inequality is due to our assumption

$$p \geqslant 2C_I^2 \frac{\mu r^2 \kappa^7 \log(n_1 \vee n_2)}{n_1 \wedge n_2}.$$

Lemma 4.2.4 also implies that

$$\mathbb{P}[E_{gd}^{(n_1+n_2)^3}] \geqslant 1 - \left(1 + (n_1+n_2)^3\right)(n_1+n_2)^{-10} \geqslant 1 - (n_1+n_2)^{-3},$$

which gives the proof of the first part of Theorem 4.1.1. If we assume additionally that $\eta \geqslant \frac{\sigma_r(\boldsymbol{M})}{1000\sigma_1^2(\boldsymbol{M})}$, which directly gives $0 < \rho \leqslant 1 - 5 \times 10^{-5}$. This implies that

$$\rho^{(n_1+n_2)^3} \leqslant \exp(\log(1 - 5 \times 10^{-5})(n_1+n_2)^3) \leqslant \exp(-(n_1+n_2)^3/C_R).$$

for some absolute constant C_R. □

4.3. Proof of Lemma 4.2.4

In this section, we give the proof of Lemma 4.2.4. Within the proof, we will mainly follow the proof structure introduced in [**MWCC18**], and useful lemmas from [**MWCC18**] such as Lemma 4.3.1 and Lemma 4.3.3 are intensively used. Moreover, we use Lemma 2.3.6 throughout this section to simplify the proof, and we also conduct a more meticulous application of the matrix Bernstein inequality. These efforts result an $O(\sqrt{r})$ tighter on our error bounds.

4.3.1. Key Lemmas. In this subsection, we list some useful lemmas which will be used to prove Lemma 4.2.4.

First, we need a lemma from [**MWCC18**]:

LEMMA 4.3.1 ([**MWCC18**, Lemma 37]). *Suppose $\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{n\times r}$ are matrices such that*

$$\|\boldsymbol{X}_1 - \boldsymbol{X}_0\|\|\boldsymbol{X}_0\| \leqslant \frac{\sigma_r^2(\boldsymbol{X}_0)}{2}, \quad \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|\|\boldsymbol{X}_0\| \leqslant \frac{\sigma_r^2(\boldsymbol{X}_0)}{4}. \tag{4.26}$$

Denote

$$\boldsymbol{R}_1 \coloneqq \underset{\boldsymbol{R}\in\mathsf{O}(r)}{\operatorname{argmin}} \|\boldsymbol{X}_1\boldsymbol{R} - \boldsymbol{X}_0\|_F,$$

$$\boldsymbol{R}_2 \coloneqq \underset{\boldsymbol{R}\in\mathsf{O}(r)}{\operatorname{argmin}} \|\boldsymbol{X}_2\boldsymbol{R} - \boldsymbol{X}_0\|_F.$$

Then the following two inequalities hold true:

$$\begin{aligned}
\|\boldsymbol{X}_1\boldsymbol{R}_1 - \boldsymbol{X}_2\boldsymbol{R}_2\| \leqslant& 5\frac{\sigma_1^2(\boldsymbol{X}_0)}{\sigma_r^2(\boldsymbol{X}_0)}\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|,\\
\|\boldsymbol{X}_1\boldsymbol{R}_1 - \boldsymbol{X}_2\boldsymbol{R}_2\|_F \leqslant& 5\frac{\sigma_1^2(\boldsymbol{X}_0)}{\sigma_r^2(\boldsymbol{X}_0)}\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_F.
\end{aligned}$$

In order to proceed, we also need a control of $\|\boldsymbol{\Omega} - p\boldsymbol{J}\|$, right here, in order to incorporate the assumption that $1/C_{10} \leqslant n_1/n_2 \leqslant C_{10}$, we have a slightly modified version of Lemma A.4.6:

Lemma 4.3.2. *There is a constant $C_{13} > 0$ such that if $p \geqslant C_{13}\frac{\log(n_1\vee n_2)}{n_1\wedge n_2}$, then on an event E_S with probability $\mathbb{P}[E_S] \geqslant 1-(n_1+n_2)^{-11}$, we have*

$$\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \leqslant C_{13}\sqrt{(n_1\wedge n_2)p}.$$

Here we use the assumption that $1/C_{10} < n_1/n_2 < C_{10}$ and C_{13} is dependent on C_{10}.

Finally, we need a lemma to control the norm of $\operatorname{sgn}(\boldsymbol{C}+\boldsymbol{E}) - \operatorname{sgn}(\boldsymbol{C})$ by the norm of $\boldsymbol{E}$:

Lemma 4.3.3 ([**Mat93**, **MWCC18**]). *Let $\boldsymbol{C} \in \mathbb{R}^{r\times r}$ be a nonsingular matrix. Then for any matrix $\boldsymbol{E} \in \mathbb{R}^{r\times r}$ with $\|\boldsymbol{E}\| \leqslant \sigma_r(\boldsymbol{C})$ and any unitarily invariant norm $|||\cdot|||$, one have*

$$|||\operatorname{sgn}(\boldsymbol{C}+\boldsymbol{E}) - \operatorname{sgn}(\boldsymbol{C})||| \leqslant \frac{2}{\sigma_{r-1}(\boldsymbol{C})+\sigma_r(\boldsymbol{C})}|||\boldsymbol{E}|||.$$

4.3.2. Proof of (4.21). For the spectral norm, first consider the auxiliary iterates defined as following:

$$\begin{aligned}
\widetilde{\boldsymbol{X}}^{t+1} :=& \boldsymbol{X}^t\boldsymbol{R}^t - \frac{\eta}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^t\left(\boldsymbol{Y}^t\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{V}\\
&- \frac{\eta}{2}\boldsymbol{U}(\boldsymbol{R}^t)^\top\left(\left(\boldsymbol{X}^t\right)^\top\boldsymbol{X}^t - \left(\boldsymbol{Y}^t\right)^\top\boldsymbol{Y}^t\right)\boldsymbol{R}^t,\\
\widetilde{\boldsymbol{Y}}^{t+1} :=& \boldsymbol{Y}^t\boldsymbol{R}^t - \frac{\eta}{p}\left[\mathcal{P}_\Omega\left(\boldsymbol{X}^t\left(\boldsymbol{Y}^t\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right]^\top\boldsymbol{U}\\
&- \frac{\eta}{2}\boldsymbol{V}(\boldsymbol{R}^t)^\top\left(\left(\boldsymbol{Y}^t\right)^\top\boldsymbol{Y}^t - \left(\boldsymbol{X}^t\right)^\top\boldsymbol{X}^t\right)\boldsymbol{R}^t.
\end{aligned}\tag{4.27}$$

Denote

$$\widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1} := \boldsymbol{X}^t\boldsymbol{R}^t - \eta\left(\boldsymbol{X}^t\left(\boldsymbol{Y}^t\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{V} - \frac{\eta}{2}\boldsymbol{U}(\boldsymbol{R}^t)^\top\left(\left(\boldsymbol{X}^t\right)^\top\boldsymbol{X}^t - \left(\boldsymbol{Y}^t\right)^\top\boldsymbol{Y}^t\right)\boldsymbol{R}^t$$

and

$$\widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1} := \boldsymbol{Y}^t\boldsymbol{R}^t - \eta\left(\boldsymbol{X}^t\left(\boldsymbol{Y}^t\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)^\top\boldsymbol{U} - \frac{\eta}{2}\boldsymbol{V}(\boldsymbol{R}^t)^\top\left(\left(\boldsymbol{Y}^t\right)^\top\boldsymbol{Y}^t - \left(\boldsymbol{X}^t\right)^\top\boldsymbol{X}^t\right)\boldsymbol{R}^t.$$

Then by triangle inequality, we have the following decomposition:

$$\begin{aligned}
\left\|\begin{bmatrix}\boldsymbol{X}^{t+1}\\ \boldsymbol{Y}^{t+1}\end{bmatrix}\boldsymbol{R}^{t+1} - \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\| &\leqslant \left\|\begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix} - \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\| + \left\|\begin{bmatrix}\boldsymbol{X}^{t+1}\\ \boldsymbol{Y}^{t+1}\end{bmatrix}\boldsymbol{R}^{t+1} - \begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix}\right\| \\
&\leqslant \underbrace{\left\|\begin{bmatrix}\widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix} - \begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix}\right\|}_{\alpha_1} + \underbrace{\left\|\begin{bmatrix}\widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix} - \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\|}_{\alpha_2} \\
&\quad + \underbrace{\left\|\begin{bmatrix}\boldsymbol{X}^{t+1}\\ \boldsymbol{Y}^{t+1}\end{bmatrix}\boldsymbol{R}^{t+1} - \begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix}\right\|}_{\alpha_3}.
\end{aligned} \tag{4.28}$$

4.3.2.1. *Analysis of* α_1. First for α_1, since

$$\begin{aligned}
&\begin{bmatrix}\widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix} - \begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix} \\
=&\eta\begin{bmatrix}\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{V}\\ \frac{1}{p}\left[\mathcal{P}_\Omega\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right]^\top\boldsymbol{U}\end{bmatrix} - \eta\begin{bmatrix}\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{V}\\ \left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)^\top\boldsymbol{U}\end{bmatrix},
\end{aligned}$$

and using the facts $\left\|\begin{bmatrix}\boldsymbol{A}\\ \boldsymbol{B}\end{bmatrix}\right\| \leqslant \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$ and $\|\boldsymbol{U}\| = \|\boldsymbol{V}\|$, we have

$$\begin{aligned}
\alpha_1 =&\eta\left\|\begin{bmatrix}\left(\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right) - \left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right)\boldsymbol{V}\\ \left(\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right) - \left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right)^\top\boldsymbol{U}\end{bmatrix}\right\| \\
\leqslant&2\eta\|\boldsymbol{U}\|\left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right) - \left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right\| \\
\leqslant&2\eta\|\boldsymbol{U}\|\left\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{\Delta}_{\boldsymbol{X}}^t\boldsymbol{V}^\top) - \boldsymbol{\Delta}_{\boldsymbol{X}}^t\boldsymbol{V}^\top\right\| + 2\eta\|\boldsymbol{U}\|\left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\right) - \boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\right\| \\
&+ 2\eta\|\boldsymbol{U}\|\left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}^t(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\right) - \boldsymbol{\Delta}_{\boldsymbol{X}}^t(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\right\|.
\end{aligned}$$

Here we denote $\boldsymbol{\Delta}_{\boldsymbol{X}}^t := \boldsymbol{X}^t\boldsymbol{R}^t - \boldsymbol{U}, \boldsymbol{\Delta}_{\boldsymbol{Y}}^t := \boldsymbol{Y}^t\boldsymbol{R}^t - \boldsymbol{V}$, and $\boldsymbol{\Delta}^t := \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}}^t \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^t \end{bmatrix}$. The last inequality uses the fact that

$$\begin{aligned} \boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top =& \boldsymbol{X}^t\boldsymbol{R}^t(\boldsymbol{R}^t)^\top(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top \\ =&(\boldsymbol{\Delta}_{\boldsymbol{X}}^t + \boldsymbol{U})(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t + \boldsymbol{V})^\top - \boldsymbol{U}\boldsymbol{V}^\top \\ =&\boldsymbol{\Delta}_{\boldsymbol{X}}^t\boldsymbol{V}^\top + \boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top + \boldsymbol{\Delta}_{\boldsymbol{X}}^t(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top. \end{aligned} \tag{4.29}$$

Using Lemma 2.3.6, we can show that

$$\alpha_1 \leqslant \frac{2\eta}{p}\|\boldsymbol{U}\|\|\boldsymbol{\Omega} - p\boldsymbol{J}\|(\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty} + \|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|_{2,\infty} + \|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|_{2,\infty}\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|_{2,\infty}).$$

From (4.24), if

$$p \geqslant 110^2 C_I^2 \frac{\mu r \kappa^{11}\log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

then

$$\|\boldsymbol{\Delta}^t\|_{2,\infty} \leqslant \sqrt{\frac{\mu r \kappa}{n_1 \wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}.$$

Here we also use the fact that $\rho < 1$. Recall that $\|\boldsymbol{U}\|_{2,\infty}, \|\boldsymbol{V}\|_{2,\infty} \leqslant \sqrt{\frac{\mu r \kappa}{n_1 \wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}$, and by (4.24),

$$\alpha_1 \leqslant \frac{2\eta}{p}\sqrt{\sigma_1(\boldsymbol{M})}\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \times 3\left(110 C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{12}\log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\sqrt{\sigma_1(\boldsymbol{M})} \times \sqrt{\frac{\mu r \kappa}{n_1 \wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}\right).$$

Moreover, using Lemma 4.3.2, if in addition

$$p \geqslant (C_{13} + 16 \times 660^2)\frac{\mu^2 r^2 \kappa^9 \log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

then on the event $E_{gd}^t \subset E_H \subset E_S$, we have

$$\begin{aligned}\alpha_1 \leqslant & \frac{2\eta}{p}\sqrt{\sigma_1(\boldsymbol{M})}\sqrt{(n_1 \wedge n_2)p} \\ & \times 3\left(110C_I\rho^t\sqrt{\frac{\mu^2 r^2 \kappa^{12}\log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\sqrt{\sigma_1(\boldsymbol{M})} \times \sqrt{\frac{\mu r \kappa}{n_1 \wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}\right) \\ = & 660\eta C_I\rho^t\sqrt{\frac{\mu^3 r^3 \kappa^{13}\log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p^2}}\sqrt{\sigma_1(\boldsymbol{M})}^3 \\ \leqslant & 0.25\eta\sigma_r(\boldsymbol{M})C_I\rho^t\sqrt{\frac{\mu r \kappa^6\log(n_1 \vee n_2)}{(n_1 \wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})}.\end{aligned} \tag{4.30}$$

4.3.2.2. *Analysis of α_2.* Since $\boldsymbol{\Delta}_{\boldsymbol{X}}^t = \boldsymbol{X}^t\boldsymbol{R}^t - \boldsymbol{U}$ and $\boldsymbol{\Delta}_{\boldsymbol{Y}}^t = \boldsymbol{Y}^t\boldsymbol{R}^t - \boldsymbol{V}$, we have

$$\begin{aligned}&(\boldsymbol{R}^t)^\top\left[\left(\boldsymbol{X}^t\right)^\top\boldsymbol{X}^t - \left(\boldsymbol{Y}^t\right)^\top\boldsymbol{Y}^t\right]\boldsymbol{R}^t \\ &= \left(\boldsymbol{\Delta}_{\boldsymbol{X}}^t + \boldsymbol{U}\right)^\top\left(\boldsymbol{\Delta}_{\boldsymbol{X}}^t + \boldsymbol{U}\right) - \left(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t + \boldsymbol{V}\right)^\top\left(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t + \boldsymbol{V}\right) \\ &= \left(\boldsymbol{\Delta}_{\boldsymbol{X}}^t\right)^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t + \left(\boldsymbol{\Delta}_{\boldsymbol{X}}^t\right)^\top\boldsymbol{U} + \boldsymbol{U}^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t + \boldsymbol{U}^\top\boldsymbol{U} \\ &\quad - \left[\left(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\right)^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t + \left(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\right)^\top\boldsymbol{V} + \boldsymbol{V}^\top\left(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\right) + \boldsymbol{V}^\top\boldsymbol{V}\right] \\ &= \left(\boldsymbol{\Delta}_{\boldsymbol{X}}^t\right)^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t + \left(\boldsymbol{\Delta}_{\boldsymbol{X}}^t\right)^\top\boldsymbol{U} + \boldsymbol{U}^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t - \left(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\right)^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t - \left(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\right)^\top\boldsymbol{V} - \boldsymbol{V}^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t.\end{aligned} \tag{4.31}$$

Therefore, for α_2,

$$\begin{aligned}
&\begin{bmatrix} \widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{X}^t\boldsymbol{R}^t - \eta\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{V} \\ \boldsymbol{Y}^t\boldsymbol{R}^t - \eta\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)^\top\boldsymbol{U} \end{bmatrix} \\
&\quad + \begin{bmatrix} -\frac{\eta}{2}\boldsymbol{U}(\boldsymbol{R}^t)^\top\left(\left(\boldsymbol{X}^t\right)^\top\boldsymbol{X}^t - \left(\boldsymbol{Y}^t\right)^\top\boldsymbol{Y}^t\right)\boldsymbol{R}^t - \boldsymbol{U} \\ -\frac{\eta}{2}\boldsymbol{V}(\boldsymbol{R}^t)^\top\left(\left(\boldsymbol{Y}^t\right)^\top\boldsymbol{Y}^t - \left(\boldsymbol{X}^t\right)^\top\boldsymbol{X}^t\right)\boldsymbol{R}^t - \boldsymbol{V} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}}^t - \eta\boldsymbol{\Delta}_{\boldsymbol{X}}^t\boldsymbol{V}^\top\boldsymbol{V} - \eta\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\boldsymbol{V} - \frac{\eta}{2}\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{X}}^t)^\top\boldsymbol{U} \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^t - \eta\boldsymbol{V}(\boldsymbol{\Delta}_{\boldsymbol{X}}^t)^\top\boldsymbol{U} - \eta\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\boldsymbol{U}^\top\boldsymbol{U} - \frac{\eta}{2}\boldsymbol{V}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\boldsymbol{V} \end{bmatrix} \\
&\quad + \begin{bmatrix} -\frac{\eta}{2}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t + \frac{\eta}{2}\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\boldsymbol{V} + \frac{\eta}{2}\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t + \eta\boldsymbol{\mathcal{E}}_1 \\ -\frac{\eta}{2}\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t + \frac{\eta}{2}\boldsymbol{V}(\boldsymbol{\Delta}_{\boldsymbol{X}}^t)^\top\boldsymbol{U} + \frac{\eta}{2}\boldsymbol{V}\boldsymbol{U}^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t + \eta\boldsymbol{\mathcal{E}}_2 \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}}^t - \eta\boldsymbol{\Delta}_{\boldsymbol{X}}^t\boldsymbol{V}^\top\boldsymbol{V} - \eta\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t + \frac{\eta}{2}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^t - \eta\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\boldsymbol{U}^\top\boldsymbol{U} - \eta\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t + \frac{\eta}{2}\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t \end{bmatrix} \\
&\quad + \begin{bmatrix} \frac{\eta}{2}\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t - \frac{\eta}{2}\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\boldsymbol{V} - \frac{\eta}{2}\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{X}}^t)^\top\boldsymbol{U} + \eta\boldsymbol{\mathcal{E}}_1 \\ \frac{\eta}{2}\boldsymbol{V}\boldsymbol{U}^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t - \frac{\eta}{2}\boldsymbol{V}(\boldsymbol{\Delta}_{\boldsymbol{X}}^t)^\top\boldsymbol{U} - \frac{\eta}{2}\boldsymbol{V}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\boldsymbol{V} + \eta\boldsymbol{\mathcal{E}}_2 \end{bmatrix}.
\end{aligned} \tag{4.32}$$

Here

$$\boldsymbol{\mathcal{E}}_1 :=-\boldsymbol{\Delta}_{\boldsymbol{X}}^t(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\boldsymbol{V} - \frac{1}{2}\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{X}}^t)^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t + \frac{1}{2}\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t, \tag{4.33}$$

$$\boldsymbol{\mathcal{E}}_2 :=-\boldsymbol{\Delta}_{\boldsymbol{Y}}^t(\boldsymbol{\Delta}_{\boldsymbol{X}}^t)^\top\boldsymbol{U} - \frac{1}{2}\boldsymbol{V}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t + \frac{1}{2}\boldsymbol{V}(\boldsymbol{\Delta}_{\boldsymbol{X}}^t)^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t \tag{4.34}$$

denote terms with at least two $\boldsymbol{\Delta}_{\boldsymbol{X}}^t$'s and $\boldsymbol{\Delta}_{\boldsymbol{Y}}^t$'s. By the way we define $\boldsymbol{R}^t$ in (4.20),

$$\begin{bmatrix} \boldsymbol{X}^t\boldsymbol{R}^t \\ \boldsymbol{Y}^t\boldsymbol{R}^t \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$$

is positive semidefinite. Therefore,

$$\begin{bmatrix} \boldsymbol{X}^t\boldsymbol{R}^t - \boldsymbol{U} \\ \boldsymbol{Y}^t\boldsymbol{R}^t - \boldsymbol{V} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} = (\boldsymbol{\Delta}_{\boldsymbol{X}}^t)^\top \boldsymbol{U} + (\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top \boldsymbol{V}$$

is symmetric. Plugging this fact back to (4.32) we have

$$\begin{aligned}
&\begin{bmatrix} \widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \\
=&\begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}}^t - \eta\boldsymbol{\Delta}_{\boldsymbol{X}}^t\boldsymbol{V}^\top\boldsymbol{V} - \eta\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t + \eta\boldsymbol{\mathcal{E}}_1 \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^t - \eta\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\boldsymbol{U}^\top\boldsymbol{U} - \eta\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t + \eta\boldsymbol{\mathcal{E}}_2 \end{bmatrix} \\
=&\frac{1}{2}\begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}}^t \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^t \end{bmatrix}(\boldsymbol{I} - 2\eta\boldsymbol{U}^\top\boldsymbol{U}) + \frac{1}{2}\left(\boldsymbol{I} - 2\eta\begin{bmatrix} \boldsymbol{U}\boldsymbol{U}^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}\boldsymbol{V}^\top \end{bmatrix}\right)\begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}}^t \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^t \end{bmatrix} + \eta\boldsymbol{\mathcal{E}},
\end{aligned}$$

where $\boldsymbol{\mathcal{E}} := \begin{bmatrix} \boldsymbol{\mathcal{E}}_1 \\ \boldsymbol{\mathcal{E}}_2 \end{bmatrix}$. Here the last equality uses the fact that $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{V}^\top\boldsymbol{V}$. Recall that we define $\boldsymbol{U}$ by $\widetilde{\boldsymbol{U}}\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{V}$ by $\widetilde{\boldsymbol{V}}\boldsymbol{\Sigma}^{1/2}$, $\boldsymbol{U}\boldsymbol{U}^\top$ and $\boldsymbol{V}\boldsymbol{V}^\top$ share the same eigenvalues. And $\|\boldsymbol{U}\boldsymbol{U}^\top\| = \|\boldsymbol{V}\boldsymbol{V}^\top\| = \sigma_1(\boldsymbol{M})$. Therefore, we have

$$\begin{aligned}
\alpha_2 =&\left\|\begin{bmatrix} \widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}\right\| \\
\leqslant&\frac{1}{2}\|\boldsymbol{I} - 2\eta\boldsymbol{U}^\top\boldsymbol{U}\|\|\boldsymbol{\Delta}^t\| + \frac{1}{2}\|\boldsymbol{\Delta}^t\|\left\|\boldsymbol{I} - 2\eta\begin{bmatrix} \boldsymbol{U}\boldsymbol{U}^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}\boldsymbol{V}^\top \end{bmatrix}\right\| + \eta\|\boldsymbol{\mathcal{E}}\| \\
\leqslant&(1 - \eta\sigma_r(\boldsymbol{M}))\|\boldsymbol{\Delta}^t\| + \eta\|\boldsymbol{\mathcal{E}}\|.
\end{aligned}$$

The last inequality uses the fact that $\eta \leqslant \frac{\sigma_r(\boldsymbol{M})}{200\sigma_1^2(\boldsymbol{M})}$. By the definition of $\boldsymbol{\mathcal{E}}$,

$$\|\boldsymbol{\mathcal{E}}\| \leqslant 4\|\boldsymbol{\Delta}^t\|^2\|\boldsymbol{U}\|$$

holds. From (4.21) and since

$$p \geqslant 1600C_I^2\frac{\mu r\kappa^8\log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

on the event E^t_{gd},

$$\|\boldsymbol{\mathcal{E}}\| \leqslant 4 \times C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2)p}} \sigma_1(\boldsymbol{M})\|\boldsymbol{\Delta}^t\| \leqslant 0.1\sigma_r(\boldsymbol{M})\|\boldsymbol{\Delta}^t\|$$

holds. Therefore, we have

$$\alpha_2 \leqslant (1 - 0.9\eta\sigma_r(\boldsymbol{M}))\|\boldsymbol{\Delta}^t\|. \tag{4.35}$$

4.3.2.3. *Analysis of* α_3*.* Now we can start to control α_3. Rewrite α_3 as

$$\begin{aligned} \alpha_3 &= \left\| \begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix} \boldsymbol{R}^{t+1} - \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} \right\| \\ &= \left\| \begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix} \boldsymbol{R}^{t}(\boldsymbol{R}^{t})^\top \boldsymbol{R}^{t+1} - \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} \right\|. \end{aligned}$$

We want to apply Lemma 4.3.1 with

$$\boldsymbol{X}_0 = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}, \ \boldsymbol{X}_1 = \begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix} \boldsymbol{R}^t, \text{ and } \boldsymbol{X}_2 = \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix}. \tag{4.36}$$

By the way we define $\boldsymbol{U}$ and $\boldsymbol{V}$, we have $\sigma_1(\boldsymbol{X}_0) = \sqrt{2\sigma_1(\boldsymbol{M})}$, $\sigma_2(\boldsymbol{X}_0) = \sqrt{2\sigma_2(\boldsymbol{M})}$, $\cdots$, $\sigma_r(\boldsymbol{X}_0) = \sqrt{2\sigma_r(\boldsymbol{M})}$, and $\sigma_1(\boldsymbol{X}_0)/\sigma_r(\boldsymbol{X}_0) = \sqrt{\kappa}$. In order to proceed, we first assume we can apply Lemma 4.3.1 here:

Claim 4.3.4. *Under the setup of Lemma 4.2.4, on the event* $E^t_{gd} \subset E_H \subset E_S$*, the requirement of Lemma 4.3.1 to apply here is satisfied with* $\boldsymbol{X}_0$*,* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *defined as in* (4.36)*. Moreover, by applying Lemma 4.3.1, we have*

$$\alpha_3 = \left\| \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix} \boldsymbol{R}^{t+1} \right\| \leqslant 0.5\eta\sigma_r(\boldsymbol{M})\|\boldsymbol{\Delta}^t\|. \tag{4.37}$$

Now by putting the estimations of $\alpha_1, \alpha_2, \alpha_3$, (4.30), (4.35), (4.37) together,

$$\begin{aligned}
&\left\| \begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix} \boldsymbol{R}^{t+1} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \\
\leqslant& \alpha_1 + \alpha_2 + \alpha_3 \\
\leqslant& (1 - 0.9\eta\sigma_r(\boldsymbol{M}))\|\boldsymbol{\Delta}^t\| + 0.25\eta\sigma_r(\boldsymbol{M}) C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} \sqrt{\sigma_1(\boldsymbol{M})} \\
&+ 0.5\eta\sigma_r(\boldsymbol{M})\|\boldsymbol{\Delta}^t\| \\
\leqslant& C_I \rho^{t+1} \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} \sqrt{\sigma_1(\boldsymbol{M})}
\end{aligned} \tag{4.38}$$

holds on the event $E_{gd}^t \subset E_H \subset E_S$, where the last inequality uses (4.21) and $\rho = 1 - 0.05\eta\sigma_r(\boldsymbol{M})$.

Proof of Claim 4.3.4. By the definition of $\boldsymbol{R}^{t+1}$ in (4.20), we can verify that

$$\boldsymbol{R}_1 = (\boldsymbol{R}^t)^\top \boldsymbol{R}^{t+1}.$$

Recall $\boldsymbol{R}_1$ is defined in Lemma 4.3.1. Now we want to show that $\boldsymbol{R}_2 = \boldsymbol{I}$. In other words, we want to show

$$\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}^\top \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} \succeq \boldsymbol{0}.$$

First, from (4.27),

$$\begin{aligned}
&\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}^\top \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} \\
=& \boldsymbol{U}^\top \boldsymbol{X}^t \boldsymbol{R}^t - \frac{\eta}{p} \boldsymbol{U}^\top \mathcal{P}_\Omega \left(\boldsymbol{X}^t (\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top \right) \boldsymbol{V} - \frac{\eta}{2} \boldsymbol{U}^\top \boldsymbol{U} (\boldsymbol{R}^t)^\top \left((\boldsymbol{X}^t)^\top \boldsymbol{X}^t - (\boldsymbol{Y}^t)^\top \boldsymbol{Y}^t \right) \boldsymbol{R}^t \\
&+ \boldsymbol{V}^\top \boldsymbol{Y}^t \boldsymbol{R}^t - \frac{\eta}{p} \boldsymbol{V}^\top \left[\mathcal{P}_\Omega \left(\boldsymbol{X}^t (\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top \right) \right]^\top \boldsymbol{U} - \frac{\eta}{2} \boldsymbol{V}^\top \boldsymbol{V} (\boldsymbol{R}^t)^\top \left((\boldsymbol{Y}^t)^\top \boldsymbol{Y}^t - (\boldsymbol{X}^t)^\top \boldsymbol{X}^t \right) \boldsymbol{R}^t \\
=& \boldsymbol{U}^\top \boldsymbol{X}^t \boldsymbol{R}^t + \boldsymbol{V}^\top \boldsymbol{Y}^t \boldsymbol{R}^t - \frac{\eta}{p} \boldsymbol{U}^\top \mathcal{P}_\Omega \left(\boldsymbol{X}^t (\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top \right) \boldsymbol{V} - \frac{\eta}{p} \boldsymbol{V}^\top \left[\mathcal{P}_\Omega \left(\boldsymbol{X}^t (\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top \right) \right]^\top \boldsymbol{U},
\end{aligned}$$

where the last equation holds since $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{V}^\top\boldsymbol{V}$. By the definition of $\boldsymbol{R}^t$, $\boldsymbol{U}^\top\boldsymbol{X}^t\boldsymbol{R}^t + \boldsymbol{V}^\top\boldsymbol{Y}^t\boldsymbol{R}^t$ is positive semidefinite, therefore symmetric. Therefore, $\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}^\top \begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix}$ is symmetric. Moreover, we have

$$\begin{aligned}\left\|\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}^\top \begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix} - \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\| \leqslant & \left\|\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\| \left\|\begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix} - \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\| \\ \leqslant & 2\sqrt{\sigma_1(\boldsymbol{M})}(\alpha_1+\alpha_2),\end{aligned}$$

where the last inequality holds by triangle inequality and the definition of α_1 and α_2 in (4.28). From (4.30) and (4.35),

$$\alpha_1+\alpha_2 \leqslant (1-0.9\eta\sigma_r(\boldsymbol{M}))\|\boldsymbol{\Delta}^t\| + 0.25\eta\sigma_r(\boldsymbol{M})C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})}$$

holds on the event E_{gd}^t. Therefore, from (4.21), and the fact that

$$p \geqslant 16C_I^2\frac{\mu r\kappa^8\log(n_1\vee n_2)}{n_1\wedge n_2}$$

and

$$\eta \leqslant \frac{\sigma_r(\boldsymbol{M})}{200\sigma_1^2(\boldsymbol{M})},$$

we have

$$\begin{aligned}&\left\|\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}^\top \begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix} - \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\| \\ \leqslant & \left\|\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\| \left\|\begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix} - \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\| \\ \leqslant & 2\sqrt{\sigma_1(\boldsymbol{M})}(1-0.65\eta\sigma_r(\boldsymbol{M}))C_I\rho^t \times \sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})} \\ \leqslant & 0.5\sigma_r(\boldsymbol{M}) \leqslant 0.5\sigma_r^2(\boldsymbol{X}_0)\end{aligned}$$

on the event E_{gd}^t. By the fact that $\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} = \boldsymbol{U}^\top \boldsymbol{U} + \boldsymbol{V}^\top \boldsymbol{V} = 2\boldsymbol{U}^\top \boldsymbol{U}$, we have

$$\lambda_r \left(\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right) = 2\sigma_r(\boldsymbol{M}).$$

By the construction of $\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}^\top \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix}$, it is an $r \times r$ symmetric matrix. By the Weyl's inequality, for all $i = 1, \cdots, r$, any two symmetric matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{r\times r}$ satisfies

$$|\lambda_i(\boldsymbol{A}) - \lambda_i(\boldsymbol{B})| \leqslant \|\boldsymbol{A} - \boldsymbol{B}\|.$$

Therefore, we have

$$\lambda_r \left(\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}^\top \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} \right) \geqslant 1.5\sigma_r(\boldsymbol{M}),$$

and $\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}^\top \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} \succeq \boldsymbol{0}$. Therefore, we have

$$\boldsymbol{I} = \boldsymbol{R}_2 = \operatorname*{argmin}_{\boldsymbol{R} \in \mathsf{O}(r)} \left\| \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} \boldsymbol{R} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\|_F.$$

Now we want to verify condition (4.26) of Lemma 4.3.1 is valid here. Since we have already shown

$$\left\| \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \left\| \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \leqslant 0.5\sigma_r^2(\boldsymbol{X}_0),$$

the first inequality is verified. Moreover, by the definition of $\boldsymbol{X}^{t+1}$ and $\boldsymbol{Y}^{t+1}$,

$$\begin{aligned} \boldsymbol{X}^{t+1}\boldsymbol{R}^t = & \boldsymbol{X}^t\boldsymbol{R}^t - \frac{\eta}{p}\mathcal{P}_\Omega(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top)\boldsymbol{Y}^t\boldsymbol{R}^t \\ & - \frac{\eta}{2}(\boldsymbol{X}^t\boldsymbol{R}^t)(\boldsymbol{R}^t)^\top((\boldsymbol{X}^t)^\top\boldsymbol{X}^t - (\boldsymbol{Y}^t)^\top\boldsymbol{Y}^t)\boldsymbol{R}^t, \end{aligned}$$

$$\begin{aligned}\boldsymbol{Y}^{t+1}\boldsymbol{R}^t =& \boldsymbol{Y}^t\boldsymbol{R}^t - \frac{\eta}{p}\left[\mathcal{P}_\Omega(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top)\right]^\top \boldsymbol{X}^t\boldsymbol{R}^t \\ &- \frac{\eta}{2}(\boldsymbol{Y}^t\boldsymbol{R}^t)(\boldsymbol{R}^t)^\top((\boldsymbol{Y}^t)^\top\boldsymbol{Y}^t - (\boldsymbol{X}^t)^\top\boldsymbol{X}^t)\boldsymbol{R}^t.\end{aligned}$$

Hence,

$$\begin{aligned}&\left\|\begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix} - \begin{bmatrix}\boldsymbol{X}^{t+1}\\ \boldsymbol{Y}^{t+1}\end{bmatrix}\boldsymbol{R}^t\right\| \\ =&\eta\left\|\begin{bmatrix}\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{\Delta}_{\boldsymbol{Y}}^t \\ \frac{1}{p}\left[\mathcal{P}_\Omega\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right]^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t\end{bmatrix} + \begin{bmatrix}\frac{1}{2}\boldsymbol{\Delta}_{\boldsymbol{X}}^t(\boldsymbol{R}^t)^\top\left((\boldsymbol{X}^t)^\top\boldsymbol{X}^t - (\boldsymbol{Y}^t)^\top\boldsymbol{Y}^t\right)\boldsymbol{R}^t \\ \frac{1}{2}\boldsymbol{\Delta}_{\boldsymbol{Y}}^t(\boldsymbol{R}^t)^\top\left((\boldsymbol{Y}^t)^\top\boldsymbol{Y}^t - (\boldsymbol{X}^t)^\top\boldsymbol{X}^t\right)\boldsymbol{R}^t\end{bmatrix}\right\| \\ \leqslant&\eta\left\|\begin{bmatrix}\boldsymbol{0} & \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right) \\ \frac{1}{p}\left[\mathcal{P}_\Omega\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right]^\top & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Delta}_{\boldsymbol{X}}^t \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^t\end{bmatrix}\right\| \\ &+\frac{\eta}{2}(\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\| + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|)\left\|(\boldsymbol{R}^t)^\top\left((\boldsymbol{X}^t)^\top\boldsymbol{X}^t - (\boldsymbol{Y}^t)^\top\boldsymbol{Y}^t\right)\boldsymbol{R}^t\right\| \\ \leqslant&\eta\left(\left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right\| + \left\|(\boldsymbol{R}^t)^\top\left((\boldsymbol{X}^t)^\top\boldsymbol{X}^t - (\boldsymbol{Y}^t)^\top\boldsymbol{Y}^t\right)\boldsymbol{R}^t\right\|\right)\|\boldsymbol{\Delta}^t\|.\end{aligned}\tag{4.39}$$

In order to bound $\left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right\|$. Recalling (4.29) and combining with Lemma 2.3.6 we have

$$\begin{aligned}&\left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^t(\boldsymbol{Y}^t)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right\| \\ \leqslant&\left\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{\Delta}_{\boldsymbol{X}}^t\boldsymbol{V}^\top) - \boldsymbol{\Delta}_{\boldsymbol{X}}^t\boldsymbol{V}^\top\right\| + \|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\boldsymbol{V}^\top\| + \left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\right) - \boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\right\| \\ &+\left\|\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\right\| + \left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}^t(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\right) - \boldsymbol{\Delta}_{\boldsymbol{X}}^t(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\right\| + \left\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t(\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\right\| \\ \leqslant&\frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p}(\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty} + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|_{2,\infty}\|\boldsymbol{U}\|_{2,\infty} + \|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|_{2,\infty}\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|_{2,\infty}) \\ &+\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|\|\boldsymbol{V}\| + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|\|\boldsymbol{U}\| + \|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|.\end{aligned}\tag{4.40}$$

And in addition , from (4.31),

$$\begin{aligned}&\left\|(\boldsymbol{R}^t)^\top\left((\boldsymbol{X}^t)^\top\boldsymbol{X}^t - (\boldsymbol{Y}^t)^\top\boldsymbol{Y}^t\right)\boldsymbol{R}^t\right\| \\ =&\left\|\boldsymbol{U}^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t + (\boldsymbol{\Delta}_{\boldsymbol{X}}^t)^\top\boldsymbol{U} + (\boldsymbol{\Delta}_{\boldsymbol{X}}^t)^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^t - \boldsymbol{V}^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t - (\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\boldsymbol{V} - (\boldsymbol{\Delta}_{\boldsymbol{Y}}^t)^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\right\| \\ \leqslant&2\|\boldsymbol{U}\|\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\| + 2\|\boldsymbol{V}\|\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\| + \|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|^2 + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|^2.\end{aligned}\tag{4.41}$$

Combining the estimations (4.40) and (4.41) together and plugging back into (4.39) we have

$$\begin{aligned}
&\left\|\begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix} \boldsymbol{R}^t \right\| \\
&\leqslant \eta \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p} (\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty} + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|_{2,\infty}\|\boldsymbol{U}\|_{2,\infty} + \|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|_{2,\infty}\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|_{2,\infty})\|\boldsymbol{\Delta}^t\| \\
&\quad + \eta \left(\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|\|\boldsymbol{V}\| + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|\|\boldsymbol{U}\| + \|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|\right) \|\boldsymbol{\Delta}^t\| \\
&\quad + \eta \left(2\|\boldsymbol{U}\|\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\| + 2\|\boldsymbol{V}\|\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\| + \|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|^2 + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|^2\right) \|\boldsymbol{\Delta}^t\|.
\end{aligned} \tag{4.42}$$

From (4.21), (4.24) and

$$p \geqslant 110^2 C_I^2 \frac{\mu r \kappa^{11} \log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

we have

$$\|\boldsymbol{\Delta}^t\| \leqslant C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2)p}} \sqrt{\sigma_1(\boldsymbol{M})} \leqslant \sqrt{\sigma_1(\boldsymbol{M})}$$

and

$$\|\boldsymbol{\Delta}^t\|_{2,\infty} \leqslant 110 C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{12} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})} \leqslant \sqrt{\frac{\mu r \kappa}{n_1 \wedge n_2}} \sqrt{\sigma_1(\boldsymbol{M})}.$$

Therefore, by applying Lemma 4.3.2 and given

$$p \geqslant (6600 C_I + 32400 C_I^2) \frac{\mu^{1.5} r^{1.5} \kappa^{10} \log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

we have

$$
\begin{aligned}
&\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}(\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}+\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|_{2,\infty}\|\boldsymbol{U}\|_{2,\infty}+\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|_{2,\infty}\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|_{2,\infty})\\
&+\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|\|\boldsymbol{V}\|+\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|\|\boldsymbol{U}\|+\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|+2\|\boldsymbol{U}\|\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|+2\|\boldsymbol{V}\|\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|+\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|^2+\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|^2\\
\leqslant&3\sqrt{\frac{n_1\wedge n_2}{p}}\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}110C_I\rho^t\times\sqrt{\frac{\mu^2r^2\kappa^{12}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sigma_1(\boldsymbol{M})\\
&+9C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sigma_1(\boldsymbol{M})\\
\leqslant&330C_I\rho^t\sqrt{\frac{\mu^3r^3\kappa^{13}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p^2}}\sigma_1(\boldsymbol{M})+9C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sigma_1(\boldsymbol{M})\\
\leqslant&\frac{1}{10\kappa}\sigma_r(\boldsymbol{M})\leqslant\frac{1}{2}\sigma_1(\boldsymbol{M}).
\end{aligned}
$$

Therefore, by plugging back to (4.42),

$$
\left\|\begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix}-\begin{bmatrix}\boldsymbol{X}^{t+1}\\ \boldsymbol{Y}^{t+1}\end{bmatrix}\boldsymbol{R}^t\right\|\leqslant\frac{1}{10\kappa}\eta\sigma_r(\boldsymbol{M})\|\boldsymbol{\Delta}^t\|, \tag{4.43}
$$

and

$$
\begin{aligned}
\left\|\begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1}\\ \widetilde{\boldsymbol{Y}}^{t+1}\end{bmatrix}-\begin{bmatrix}\boldsymbol{X}^{t+1}\\ \boldsymbol{Y}^{t+1}\end{bmatrix}\boldsymbol{R}^t\right\|\left\|\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\|\leqslant&\eta\frac{1}{2}\sigma_1(\boldsymbol{M})2\sqrt{\sigma_1(\boldsymbol{M})}\|\boldsymbol{\Delta}^t\|\\
\leqslant&\eta\sqrt{\sigma_1(\boldsymbol{M})}^3C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})}\\
\leqslant&\frac{1}{4}\sigma_r(\boldsymbol{M})\leqslant\frac{1}{4}\sigma_r^2(\boldsymbol{X}_0)
\end{aligned}
$$

holds on the event E_{gd}^t. Here the second inequality holds due to (4.21), and the third inequality follows $p\geqslant C_I^2\frac{\mu r\kappa^6\log(n_1\vee n_2)}{n_1\wedge n_2}$, and $\eta\leqslant\frac{\sigma_r(\boldsymbol{M})}{200\sigma_1^2(\boldsymbol{M})}$.

Therefore, all the requirements in (4.26) of Lemma 4.3.1 is valid, and Lemma 4.3.1 can be applied with $\boldsymbol{X}_0$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ defined as in (4.36). By applying Lemma 4.3.1,

$$\begin{aligned}\alpha_3 =& \left\| \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix} \boldsymbol{R}^{t+1} \right\| \\ \leqslant & 5\kappa \left\| \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1} \\ \widetilde{\boldsymbol{Y}}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix} \boldsymbol{R}^{t} \right\|.\end{aligned}$$

Along with (4.43), there holds $\alpha_3 \leqslant 0.5\eta\sigma_r(\boldsymbol{M})\|\boldsymbol{\Delta}^t\|$. □

4.3.3. Proof of (4.22). For the induction hypodissertation (4.22), without loss of generality, we assume $1 \leqslant l \leqslant n_1$. From (4.7), we have the following decomposition:

$$\begin{aligned}&\left(\begin{bmatrix} \boldsymbol{X}^{t+1,(l)} \\ \boldsymbol{Y}^{t+1,(l)} \end{bmatrix} \boldsymbol{R}^{t+1,(l)} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right)_{l,\cdot} \\ =&(\boldsymbol{X}_{l,\cdot}^{t+1,(l)})^\top \boldsymbol{R}^{t+1,(l)} - \boldsymbol{U}_{l,\cdot}^\top \\ =&(\boldsymbol{X}_{l,\cdot}^{t,(l)})^\top \boldsymbol{R}^{t+1,(l)} - \boldsymbol{U}_{l,\cdot}^\top - \eta \left((\boldsymbol{X}_{l,\cdot}^{t,(l)})^\top (\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}_{l,\cdot}^\top \boldsymbol{V}^\top \right) \boldsymbol{Y}^{t,(l)} \boldsymbol{R}^{t+1,(l)} \\ &- \frac{\eta}{2} (\boldsymbol{X}_{l,\cdot}^{t,(l)})^\top \left((\boldsymbol{X}^{t,(l)})^\top \boldsymbol{X}^{t,(l)} - (\boldsymbol{Y}^{t,(l)})^\top \boldsymbol{Y}^{t,(l)} \right) \boldsymbol{R}^{t+1,(l)} \\ =&\boldsymbol{a}_1 + \boldsymbol{a}_2 - \boldsymbol{a}_3,\end{aligned}$$

where

$$\boldsymbol{a}_1 :=(\boldsymbol{X}_{l,\cdot}^{t,(l)})^\top \boldsymbol{R}^{t,(l)} - \boldsymbol{U}_{l,\cdot}^\top - \eta \left((\boldsymbol{X}_{l,\cdot}^{t,(l)})^\top (\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}_{l,\cdot}^\top \boldsymbol{V}^\top \right) \boldsymbol{Y}^{t,(l)} \boldsymbol{R}^{t,(l)},$$

$$\boldsymbol{a}_2 := \left[(\boldsymbol{X}_{l,\cdot}^{t,(l)})^\top \boldsymbol{R}^{t,(l)} - \eta \left((\boldsymbol{X}_{l,\cdot}^{t,(l)})^\top (\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}_{l,\cdot}^\top \boldsymbol{V}^\top \right) \boldsymbol{Y}^{t,(l)} \boldsymbol{R}^{t,(l)} \right] \left[(\boldsymbol{R}^{t,(l)})^{-1} \boldsymbol{R}^{t+1,(l)} - \boldsymbol{I} \right]$$

and

$$\boldsymbol{a}_3 := \frac{\eta}{2} (\boldsymbol{X}_{l,\cdot}^{t,(l)})^\top \left((\boldsymbol{X}^{t,(l)})^\top \boldsymbol{X}^{t,(l)} - (\boldsymbol{Y}^{t,(l)})^\top \boldsymbol{Y}^{t,(l)} \right) \boldsymbol{R}^{t+1,(l)}.$$

First for $\boldsymbol{a}_1$, denote $\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} := \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{U}$, $\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} := \boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{V}$, then by a decomposition similar to (4.29),

$$
\begin{aligned}
&\|\boldsymbol{a}_1\|_2 \\
=&\left\|(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}^{\top} - \eta\left[(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}^{\top}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)})^{\top} + (\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}^{\top}\boldsymbol{V}^{\top} + \boldsymbol{U}_{l,\cdot}^{\top}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)})^{\top}\right](\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} + \boldsymbol{V})\right\|_2 \\
=&\left\|(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}^{\top} - \eta(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}^{\top}\boldsymbol{V}^{\top}\boldsymbol{V} - \eta\left[(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}^{\top}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)})^{\top} + \boldsymbol{U}_{l,\cdot}^{\top}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)})^{\top}\right]\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} - \eta(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}^{\top}\boldsymbol{V}^{\top}\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\right\|_2 \\
\leqslant&\|\boldsymbol{I} - \eta\boldsymbol{V}^{\top}\boldsymbol{V}\|\|(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}\|_2 + \eta(\|(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}\|_2 + \|\boldsymbol{U}_{l,\cdot}\|_2)\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\|\|\boldsymbol{Y}^{t,(l)}\| + \eta\|(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}\|_2\|\boldsymbol{V}\|\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\|.
\end{aligned}
$$

From (4.21),

$$
\begin{aligned}
\left\|\begin{bmatrix}\boldsymbol{X}^t \\ \boldsymbol{Y}^t\end{bmatrix}\boldsymbol{R}^t - \begin{bmatrix}\boldsymbol{U} \\ \boldsymbol{V}\end{bmatrix}\right\|\left\|\begin{bmatrix}\boldsymbol{U} \\ \boldsymbol{V}\end{bmatrix}\right\| \leqslant& 2C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1 \vee n_2)}{(n_1 \wedge n_2)p}}\sigma_1(\boldsymbol{M}) \\
\leqslant&\frac{\sigma_r(\boldsymbol{M})}{2}
\end{aligned}
$$

holds since

$$
p \geqslant 16C_I^2\frac{\mu r\kappa^8\log(n_1 \vee n_2)}{n_1 \wedge n_2}.
$$

Also from (4.23),

$$
\begin{aligned}
\left\|\begin{bmatrix}\boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)}\end{bmatrix}\boldsymbol{T}^{t,(l)} - \begin{bmatrix}\boldsymbol{X}^t \\ \boldsymbol{Y}^t\end{bmatrix}\boldsymbol{R}^t\right\|\left\|\begin{bmatrix}\boldsymbol{U} \\ \boldsymbol{V}\end{bmatrix}\right\| \leqslant& \left\|\begin{bmatrix}\boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)}\end{bmatrix}\boldsymbol{T}^{t,(l)} - \begin{bmatrix}\boldsymbol{X}^t \\ \boldsymbol{Y}^t\end{bmatrix}\boldsymbol{R}^t\right\|_F\left\|\begin{bmatrix}\boldsymbol{U} \\ \boldsymbol{V}\end{bmatrix}\right\| \\
\leqslant&2C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^{10}\log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2p}}\sigma_1(\boldsymbol{M}) \\
\leqslant&\frac{\sigma_r(\boldsymbol{M})}{4},
\end{aligned}
$$

where the last inequality holds since

$$
p \geqslant 64C_I^2\frac{\mu^2r^2\kappa^{12}\log(n_1 \vee n_2)}{n_1 \wedge n_2}.
$$

Applying Lemma 4.3.1 with

$$
\boldsymbol{X}_0 := \begin{bmatrix}\boldsymbol{U} \\ \boldsymbol{V}\end{bmatrix}, \ \boldsymbol{X}_1 := \begin{bmatrix}\boldsymbol{X}^t \\ \boldsymbol{Y}^t\end{bmatrix}\boldsymbol{R}^t,
$$

$$\boldsymbol{X}_2 := \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)},$$

since we define $\boldsymbol{U}$ by $\widetilde{\boldsymbol{U}}\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{V}$ by $\widetilde{\boldsymbol{V}}\boldsymbol{\Sigma}^{1/2}$, we have $\sigma_1(\boldsymbol{X}_0) = \sqrt{2\sigma_1(\boldsymbol{M})}$, $\sigma_2(\boldsymbol{X}_0) = \sqrt{2\sigma_2(\boldsymbol{M})}$, $\cdots$, $\sigma_r(\boldsymbol{X}_0) = \sqrt{2\sigma_r(\boldsymbol{M})}$, and $\sigma_1(\boldsymbol{X}_0)/\sigma_r(\boldsymbol{X}_0) = \sqrt{\kappa}$. We have

$$\begin{aligned} &\left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{R}^{t,(l)} - \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^{t} \right\|_F \\ \leqslant & 5\kappa \left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} - \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^{t} \right\|_F . \end{aligned}$$

Therefore, by triangle inequality we have

$$\begin{aligned} &\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\| \\ \leqslant & \left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{R}^{t,(l)} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \\ \leqslant & \left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{R}^{t,(l)} - \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^{t} \right\|_F + \left\| \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^{t} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \\ \leqslant & 5\kappa \left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} - \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^{t} \right\|_F + \left\| \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^{t} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \\ \leqslant & 5\kappa C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})} + C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} \sqrt{\sigma_1(\boldsymbol{M})} \\ \leqslant & 2 C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} \sqrt{\sigma_1(\boldsymbol{M})}. \end{aligned} \tag{4.44}$$

For the last inequality, we use the fact that

$$\frac{25\mu r \kappa^6}{n_1 \wedge n_2} \leqslant p \leqslant 1.$$

Equipped with (4.44), and combining with the fact that $\|\boldsymbol{Y}^{t,(l)}\| \leqslant \|\boldsymbol{V}\| + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\|$, we have

$$\begin{aligned}
&\|\boldsymbol{a}_1\|_2 \\
\leqslant&(1-\eta\sigma_r(\boldsymbol{M}))\|(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}\|_2 \\
&+\eta\|(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}\|_2 2C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})} \\
&\times\left(2\sqrt{\sigma_1(\boldsymbol{M})}+2C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})}\right) \\
&+\eta\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}2C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})} \\
&\times\left(\sqrt{\sigma_1(\boldsymbol{M})}+2C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})}\right).
\end{aligned}$$

Given

$$p \geqslant 4C_I^2\frac{\mu r\kappa^6\log(n_1\vee n_2)}{n_1\wedge n_2},$$

we have

$$2C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})} \leqslant \sqrt{\sigma_1(\boldsymbol{M})}.$$

Therefore,

$$\begin{aligned}
&\|\boldsymbol{a}_1\|_2 \\
\leqslant&(1-\eta\sigma_r(\boldsymbol{M}))\|(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}\|_2+\eta\|(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}\|_2 6C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sigma_1(\boldsymbol{M}) \\
&+\eta\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}4C_I\rho^t\times\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sigma_1(\boldsymbol{M}).
\end{aligned}$$

Given

$$p \geqslant 576C_I^2\frac{\mu r\kappa^8\log(n_1\vee n_2)}{n_1\wedge n_2},$$

on the event E^t_{gd},

$$\begin{aligned}
&\|\boldsymbol{a}_1\|_2 \\
\leqslant&(1-\eta\sigma_r(\boldsymbol{M}))\|(\boldsymbol{\Delta}^{t,(l)}_{\boldsymbol{X}})_{l,\cdot}\|_2 + 0.25\eta\sigma_r(\boldsymbol{M})\|(\boldsymbol{\Delta}^{t,(l)}_{\boldsymbol{X}})_{l,\cdot}\|_2 \\
&+\eta\sigma_r(\boldsymbol{M})4C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^9\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})} \qquad (4.45)\\
=&(1-0.75\eta\sigma_r(\boldsymbol{M}))\|(\boldsymbol{\Delta}^{t,(l)}_{\boldsymbol{X}})_{l,\cdot}\|_2 + 4C_I\eta\sigma_r(\boldsymbol{M})\rho^t\sqrt{\frac{\mu^2r^2\kappa^9\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})}
\end{aligned}$$

At the same time from (4.22) we have

$$\begin{aligned}
\|\boldsymbol{a}_1\|_2 \leqslant&(1-0.75\eta\sigma_r(\boldsymbol{M}))\times 100C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^{10}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})} \\
&+4\eta\sigma_r(\boldsymbol{M})C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^9\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})} \\
\leqslant&\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}
\end{aligned}$$

since

$$p\geqslant 10^4C_I^2\frac{\mu r\kappa^9\log(n_1\vee n_2)}{n_1\wedge n_2}$$

and

$$\eta\leqslant\frac{\sigma_r(\boldsymbol{M})}{200\sigma_r^2(\boldsymbol{M})}.$$

For $\boldsymbol{a}_2$, note

$$\begin{aligned}
&\|\boldsymbol{a}_2\|_2 \\
=&\left\|\left[(\boldsymbol{X}^{t,(l)}_{l,\cdot})^\top\boldsymbol{R}^{t,(l)}-\eta\left((\boldsymbol{X}^{t,(l)}_{l,\cdot})^\top(\boldsymbol{Y}^{t,(l)})^\top-\boldsymbol{U}^\top_{l,\cdot}\boldsymbol{V}^\top\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\right]\left[(\boldsymbol{R}^{t,(l)})^{-1}\boldsymbol{R}^{t+1,(l)}-\boldsymbol{I}\right]\right\| \qquad (4.46)\\
\leqslant&\|\boldsymbol{a}_1+\boldsymbol{U}_{l,\cdot}\|_2\|(\boldsymbol{R}^{t,(l)})^{-1}\boldsymbol{R}^{t+1,(l)}-\boldsymbol{I}\| \\
\leqslant&2\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}\|(\boldsymbol{R}^{t,(l)})^{-1}\boldsymbol{R}^{t+1,(l)}-\boldsymbol{I}\|.
\end{aligned}$$

Here we want to use Lemma 4.3.3 to control $\|(\boldsymbol{R}^{t,(l)})^{-1}\boldsymbol{R}^{t+1,(l)}-\boldsymbol{I}\|$. In order to proceed, we first assume the following claim is valid:

Claim 4.3.5. *Under the setup of Lemma 4.2.4, assume $1 \leqslant l \leqslant n_1$. Lemma 4.3.3 can be applied and on the event E_{gd}^t,*

$$\|(\boldsymbol{R}^{t,(l)})^{-1}\boldsymbol{R}^{t+1,(l)} - \boldsymbol{I}\| \leqslant 76C_I^2\frac{\sigma_1^2(\boldsymbol{M})}{\sigma_r(\boldsymbol{M})}\eta\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log^2(n_1\vee n_2)}{(n_1\wedge n_2)^2p^2}} \tag{4.47}$$

holds.

The proof of this claim mainly relies on Lemma 4.3.3, and the verification of conditions required by Lemma 4.3.3 is very similar to the way we handle $\alpha_1, \alpha_2, \alpha_3$ defined in (4.28). For the purpose of self-containedness, we include the proof of the claim in Appendix C.3.

Plugging (4.47) back to (4.46) we have

$$\begin{aligned}
&\|\boldsymbol{a}_2\|_2\\
\leqslant&2\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}\times 76C_I^2\frac{\sigma_1^2(\boldsymbol{M})}{\sigma_r(\boldsymbol{M})}\eta\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log^2(n_1\vee n_2)}{(n_1\wedge n_2)^2p^2}}\\
\leqslant&152C_I^2\eta\rho^t\frac{\sigma_1^2(\boldsymbol{M})}{\sigma_r(\boldsymbol{M})}\sqrt{\frac{\mu^3r^3\kappa^{13}\log^2(n_1\vee n_2)}{(n_1\wedge n_2)^3p^2}}\sqrt{\sigma_1(\boldsymbol{M})}\\
\leqslant&25C_I\eta\sigma_r(\boldsymbol{M})\rho^t\sqrt{\frac{\mu^2r^2\kappa^{10}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})},
\end{aligned} \tag{4.48}$$

where the last inequality uses the fact that

$$p \geqslant 37C_I^2\frac{\mu r\kappa^7\log(n_1\vee n_2)}{n_1\wedge n_2}.$$

Finally, for $\boldsymbol{a}_3$, note the fact that $\boldsymbol{R}^{t+1,(l)}$ and $\boldsymbol{R}^{t,(l)}$ are all orthogonal matrices. And replacing $\boldsymbol{X}$ and $\boldsymbol{Y}$ with $\boldsymbol{X}^{t,(l)}$ and $\boldsymbol{Y}^{t,(l)}$ in (4.31),

$$\begin{aligned}
&\|\boldsymbol{a}_3\|_2\\
=&\frac{\eta}{2}\left\|(\boldsymbol{X}_{l,\cdot}^{t,(l)})^\top\left((\boldsymbol{X}^{t,(l)})^\top\boldsymbol{X}^{t,(l)} - (\boldsymbol{Y}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)}\right)\boldsymbol{R}^{t+1,(l)}\right\|_2\\
=&\frac{\eta}{2}\left\|(\boldsymbol{X}_{l,\cdot}^{t,(l)})^\top\left((\boldsymbol{X}^{t,(l)})^\top\boldsymbol{X}^{t,(l)} - (\boldsymbol{Y}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)}\right)\boldsymbol{R}^{t,(l)}\right\|_2\\
\leqslant&\frac{\eta}{2}\|\boldsymbol{X}_{l,\cdot}^{t,(l)}\|_2\left\|(\boldsymbol{R}^{t,(l)})^\top\left((\boldsymbol{X}^{t,(l)})^\top\boldsymbol{X}^{t,(l)} - (\boldsymbol{Y}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)}\right)\boldsymbol{R}^{t,(l)}\right\|\\
\leqslant&\frac{\eta}{2}\|\boldsymbol{X}_{l,\cdot}^{t,(l)}\|_2\left(2\|\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}\|\|\boldsymbol{U}\| + \|\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}\|^2 + 2\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\|\|\boldsymbol{V}\| + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\|^2\right).
\end{aligned} \tag{4.49}$$

From (4.22), we have

$$\begin{aligned}\|\boldsymbol{X}_{l,\cdot}^{t,(l)}\|_2 \leqslant & \|\boldsymbol{U}_{l,\cdot}\|_2 + \|(\boldsymbol{X}_{l,\cdot}^{t,(l)})^\top \boldsymbol{R}^{t,(l)} - \boldsymbol{U}_{l,\cdot}^\top\|_2 \\ \leqslant & \sqrt{\frac{\mu r \kappa}{n_1 \wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})} + 100 C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\sqrt{\sigma_1(\boldsymbol{M})} \\ \leqslant & 2\sqrt{\frac{\mu r \kappa}{n_1 \wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}.\end{aligned} \tag{4.50}$$

The last line holds since

$$p \geqslant 10^4 C_I^2 \frac{\mu r \kappa^9 \log(n_1 \vee n_2)}{n_1 \wedge n_2}.$$

From (4.44) and given

$$p \geqslant 4 C_I^2 \frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

we have $\|\boldsymbol{\Delta}^{t,(l)}\| \leqslant \sqrt{\sigma_1(\boldsymbol{M})}$. Combining with (4.44), (4.49) and (4.50), we have

$$\begin{aligned}\|\boldsymbol{a}_3\|_2 \leqslant & \eta \sqrt{\frac{\mu r \kappa}{n_1 \wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})} \times 12 C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}}\sigma_1(\boldsymbol{M}) \\ = & 12 C_I \eta \sigma_r(\boldsymbol{M}) \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^9 \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\sqrt{\sigma_1(\boldsymbol{M})}.\end{aligned} \tag{4.51}$$

Putting the estimations on $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$ together, i.e., (4.45), (4.48) and (4.51), we have

$$\begin{aligned}& \left\| \left(\begin{bmatrix} \boldsymbol{X}^{t+1,(l)} \\ \boldsymbol{Y}^{t+1,(l)} \end{bmatrix} \boldsymbol{R}^{t+1,(l)} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right)_{l,\cdot} \right\|_2 \\ \leqslant & \|\boldsymbol{a}_1\|_2 + \|\boldsymbol{a}_2\|_2 + \|\boldsymbol{a}_3\|_2 \\ \leqslant & (1 - 0.75 \eta \sigma_r(\boldsymbol{M})) \|(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}\|_2 + 4 C_I \eta \sigma_r(\boldsymbol{M}) \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^9 \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\sqrt{\sigma_1(\boldsymbol{M})} \\ & + 25 C_I \eta \sigma_r(\boldsymbol{M}) \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\sqrt{\sigma_1(\boldsymbol{M})} \\ & + 12 C_I \eta \sigma_r(\boldsymbol{M}) \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^9 \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\sqrt{\sigma_1(\boldsymbol{M})} \\ \leqslant & 100 C_I \rho^{t+1} \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\sqrt{\sigma_1(\boldsymbol{M})},\end{aligned}$$

with $\rho = 1 - 0.05\eta\sigma_r(\boldsymbol{M})$ on the event E_{gd}^{t+1}, the last inequality uses (4.22). Notice this is the proof for the case of l satisfying $1 \leqslant l \leqslant n_1$, the proof for l satisfying $n_1 + 1 \leqslant l \leqslant n_1 + n_2$ is almost the same.

4.3.4. Proof of (4.23). For (4.23), by the choice of $\boldsymbol{T}^{t+1,(l)}$ in (4.20), we have

$$\left\| \begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix} \boldsymbol{R}^{t+1} - \begin{bmatrix} \boldsymbol{X}^{t+1,(l)} \\ \boldsymbol{Y}^{t+1,(l)} \end{bmatrix} \boldsymbol{T}^{t+1,(l)} \right\|_F^2 \leqslant \left\| \begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix} \boldsymbol{R}^{t} - \begin{bmatrix} \boldsymbol{X}^{t+1,(l)} \\ \boldsymbol{Y}^{t+1,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} \right\|_F^2 .$$

Without loss of generality, we first consider the case that l satisfying $1 \leqslant l \leqslant n_1$. First, by plugging in the definition of $\begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix}$ and $\begin{bmatrix} \boldsymbol{X}^{t+1,(l)} \\ \boldsymbol{Y}^{t+1,(l)} \end{bmatrix}$, we have

$$\begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix} \boldsymbol{R}^{t} - \begin{bmatrix} \boldsymbol{X}^{t+1,(l)} \\ \boldsymbol{Y}^{t+1,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} = \boldsymbol{A}_1 + \eta \begin{bmatrix} \boldsymbol{A}_2 \\ \boldsymbol{A}_3 \end{bmatrix}, \tag{4.52}$$

where

$$\boldsymbol{A}_1 := \left(\begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} - \eta \nabla f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \right) \boldsymbol{R}^t - \left(\begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} - \eta \nabla f(\boldsymbol{X}^{t,(l)}, \boldsymbol{Y}^{t,(l)}) \right) \boldsymbol{T}^{t,(l)}$$

$$\boldsymbol{A}_2 := \mathcal{P}_{l,\cdot} \left(\boldsymbol{X}^{t,(l)} (\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top \right) \boldsymbol{Y}^{t,(l)} \boldsymbol{T}^{t,(l)} - \frac{1}{p} \mathcal{P}_{\Omega_{l,\cdot}} \left(\boldsymbol{X}^{t,(l)} (\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top \right) \boldsymbol{Y}^{t,(l)} \boldsymbol{T}^{t,(l)}$$

and

$$\begin{aligned} \boldsymbol{A}_3 := & \left[\mathcal{P}_{l,\cdot} \left(\boldsymbol{X}^{t,(l)} (\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top \right) \right]^\top \boldsymbol{X}^{t,(l)} \boldsymbol{T}^{t,(l)} \\ & - \left[\frac{1}{p} \mathcal{P}_{\Omega_{l,\cdot}} \left(\boldsymbol{X}^{t,(l)} (\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top \right) \right]^\top \boldsymbol{X}^{t,(l)} \boldsymbol{T}^{t,(l)}. \end{aligned}$$

For $\boldsymbol{A}_1$, we have

(4.53)

$$\begin{aligned}
&\|\boldsymbol{A}_1\|_F^2 \\
=&\left\| \left(\begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} - \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^t \right) - \eta \left(\nabla f(\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}, \boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)}) - \nabla f(\boldsymbol{X}^t\boldsymbol{R}^t, \boldsymbol{Y}^t\boldsymbol{R}^t) \right) \right\|_F^2 \\
=&\left\| \left(\boldsymbol{I} - \eta \int_0^1 \nabla^2 f(*) d\tau \right) \mathrm{vec}\left(\begin{bmatrix} \boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{X}^t\boldsymbol{R}^t \\ \boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{Y}^t\boldsymbol{R}^t \end{bmatrix} \right) \right\|_2^2 \\
\leqslant& \left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{X}^t\boldsymbol{R}^t \\ \boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{Y}^t\boldsymbol{R}^t \end{bmatrix} \right\|_F^2 + \eta^2 \left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{X}^t\boldsymbol{R}^t \\ \boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{Y}^t\boldsymbol{R}^t \end{bmatrix} \right\|_F^2 \max_{0 \leqslant \tau \leqslant 1} \left\| \nabla^2 f(*) \right\|^2 \\
&- 2\eta \min_{0 \leqslant \tau \leqslant 1} \mathrm{vec}\left(\begin{bmatrix} \boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{X}^t\boldsymbol{R}^t \\ \boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{Y}^t\boldsymbol{R}^t \end{bmatrix} \right)^\top \nabla^2 f(*)\, \mathrm{vec}\left(\begin{bmatrix} \boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{X}^t\boldsymbol{R}^t \\ \boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{Y}^t\boldsymbol{R}^t \end{bmatrix} \right),
\end{aligned}$$

where the first equality uses the fact that $\nabla f(\boldsymbol{X}, \boldsymbol{Y}) = \nabla f(\boldsymbol{X}\boldsymbol{R}, \boldsymbol{Y}\boldsymbol{R})$ for any $\boldsymbol{R} \in \mathsf{O}(r)$, and here

$$\nabla^2 f(*) \coloneqq \nabla^2 f(\tau(\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{X}^t\boldsymbol{R}^t) + \boldsymbol{X}^t\boldsymbol{R}^t, \tau(\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{Y}^t\boldsymbol{R}^t) + \boldsymbol{Y}^t\boldsymbol{R}^t).$$

From (4.23) and (4.24), if

$$p \geqslant 2.42 \times 10^{10} C_{10} C_I^2 \frac{\mu^2 r^2 \kappa^{14} \log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

we have

$$\left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} - \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^t \right\|_{2,\infty} \leqslant \frac{1}{1000\kappa\sqrt{n_1+n_2}} \sqrt{\sigma_1(\boldsymbol{M})}$$

and

$$\left\| \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\|_{2,\infty} \leqslant \frac{1}{1000\kappa\sqrt{n_1+n_2}} \sqrt{\sigma_1(\boldsymbol{M})}.$$

Therefore,

$$\|\tau(\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{X}^t\boldsymbol{R}^t) + \boldsymbol{X}^t\boldsymbol{R}^t - \boldsymbol{U}\|_{2,\infty} \leqslant \frac{1}{500\kappa\sqrt{n_1+n_2}} \sqrt{\sigma_1(\boldsymbol{M})},$$

$$\|\tau(\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{Y}^t\boldsymbol{R}^t) + \boldsymbol{Y}^t\boldsymbol{R}^t - \boldsymbol{V}\|_{2,\infty} \leqslant \frac{1}{500\kappa\sqrt{n_1+n_2}}\sqrt{\sigma_1(\boldsymbol{M})}$$

for any τ satisfying $0 \leqslant \tau \leqslant 1$. And we also have

$$\left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \leqslant \frac{1}{500\kappa}\sqrt{\sigma_1(\boldsymbol{M})}.$$

Therefore, Lemma 4.2.1 can be applied here. Noting $E_{gd}^t \subset E_H$ and

$$p \geqslant C_{S1}\frac{\mu r\kappa\log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

we have (4.11) and (4.12) satisfied. Plugging (4.11) and (4.12) back to the estimation (4.53), we have

$$\begin{aligned}
\|\boldsymbol{A}_1\|_F^2 \leqslant & (1 - \frac{2}{5}\eta\sigma_r(\boldsymbol{M}) + 25\eta^2\sigma_1^2(\boldsymbol{M}))\left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} - \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t \right\|_F^2 \\
\leqslant & (1 - 0.2\eta\sigma_r(\boldsymbol{M}))\left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} - \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t \right\|_F^2,
\end{aligned}$$

where the last inequality holds since

$$\eta \leqslant \frac{\sigma_r(\boldsymbol{M})}{200\sigma_1^2(\boldsymbol{M})}.$$

Therefore,

$$\|\boldsymbol{A}_1\|_F \leqslant (1 - 0.1\eta\sigma_r(\boldsymbol{M}))\left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} - \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t \right\|_F \tag{4.54}$$

holds on the event E_{gd}^t.

For the second term $\begin{bmatrix} \boldsymbol{A}_2 \\ \boldsymbol{A}_3 \end{bmatrix}$ in (4.52), by the definition of $\mathcal{P}_{l,\cdot}$ and $\mathcal{P}_{\Omega_{l,\cdot}}$, we can see that entries of

$$\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right) - \frac{1}{p}\mathcal{P}_{\Omega_{l,\cdot}}\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)$$

are all zero except on the l-th row. Using this fact, we have

$$\boldsymbol{A}_2 = -\begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \sum_j (\frac{1}{p}\delta_{l,j} - 1) \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)_{l,j} (\boldsymbol{Y}^{t,(l)}_{j,\cdot})^\top \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \boldsymbol{T}^{t,(l)}$$

and

$$\boldsymbol{A}_3 = -\begin{bmatrix} (\frac{1}{p}\delta_{l,1} - 1) \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)_{l,1} (\boldsymbol{X}^{t,(l)}_{l,\cdot})^\top \\ \vdots \\ (\frac{1}{p}\delta_{l,j} - 1) \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)_{l,j} (\boldsymbol{X}^{t,(l)}_{l,\cdot})^\top \\ \vdots \\ (\frac{1}{p}\delta_{l,n_2} - 1) \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)_{l,n_2} (\boldsymbol{X}^{t,(l)}_{l,\cdot})^\top \end{bmatrix} \boldsymbol{T}^{t,(l)}.$$

Therefore, by triangle inequality,

$$
\begin{aligned}
&\left\| \begin{bmatrix} \boldsymbol{A}_2 \\ \boldsymbol{A}_3 \end{bmatrix} \right\|_F \\
\leqslant &\|\boldsymbol{A}_2\|_F + \|\boldsymbol{A}_3\|_F \\
\leqslant &\Bigg\| \underbrace{\sum_j (\frac{1}{p}\delta_{l,j} - 1) \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)_{l,j} \boldsymbol{Y}^{t,(l)}_{j,\cdot}}_{\boldsymbol{b}_1} \Bigg\|_2 \\
&+ \left\| \underbrace{\begin{bmatrix} (\frac{1}{p}\delta_{l,1} - 1) \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)_{l,1} \\ \vdots \\ (\frac{1}{p}\delta_{l,j} - 1) \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)_{l,j} \\ \vdots \\ (\frac{1}{p}\delta_{l,n_2} - 1) \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)_{l,n_2} \end{bmatrix}}_{\boldsymbol{b}_2} \right\|_2 \|\boldsymbol{X}^{t,(l)}_{l,\cdot}\|_2,
\end{aligned}
\tag{4.55}
$$

where the last inequality uses the fact that $\boldsymbol{T}^{t,(l)} \in \mathsf{O}(r)$.

For $\boldsymbol{b}_1$, we can write $\boldsymbol{b}_1$ in the following form:

$$
\begin{aligned}
\boldsymbol{b}_1 &= \sum_j (\frac{1}{p}\delta_{l,j} - 1) \left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)_{l,j} \boldsymbol{Y}^{t,(l)}_{j,\cdot} \\
&\coloneqq \sum_j \boldsymbol{s}_{1,j}.
\end{aligned}
$$

By the way we define $\boldsymbol{X}^{t,(l)}$ and $\boldsymbol{Y}^{t,(l)}$ in (4.7), (4.8), (4.9) and (4.10), we can see that $\boldsymbol{X}^{t,(l)}$ and $\boldsymbol{Y}^{t,(l)}$ are independent of $\delta_{l,1}, \cdots, \delta_{l,n_2}$. Therefore, conditioned on $\boldsymbol{X}^{t,(l)}$ and $\boldsymbol{Y}^{t,(l)}$, $\boldsymbol{s}_{1,j}$'s are independent and $\mathbb{E}_{\delta_{l,\cdot}} \boldsymbol{s}_{1,j} = \boldsymbol{0}$. Moreover, since

$$
\begin{aligned}
&\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top \\
=&\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}(\boldsymbol{T}^{t,(l)})^\top \boldsymbol{Y}^{t,(l)} - \boldsymbol{U}\boldsymbol{V}^\top \\
=&(\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{U})\boldsymbol{V}^\top + \boldsymbol{U}(\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{V})^\top + (\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{U})(\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)} - \boldsymbol{V})^\top.
\end{aligned}
\tag{4.56}
$$

Therefore, for all j,

$$\begin{aligned}
&\|\boldsymbol{s}_{1,j}\|_2\\
\leqslant&\frac{1}{p}\left\|\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top-\boldsymbol{U}\boldsymbol{V}^\top\right\|_{\ell_\infty}\|\boldsymbol{Y}^{t,(l)}\|_{2,\infty}\\
\leqslant&\frac{1}{p}\left(\|\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}+\|\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{V}\|_{2,\infty}\|\boldsymbol{U}\|_{2,\infty}\right)\|\boldsymbol{Y}^{t,(l)}\|_{2,\infty}\\
&+\frac{1}{p}\|\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{V}\|_{2,\infty}\|\boldsymbol{Y}^{t,(l)}\|_{2,\infty}\\
:=&L_1^{(l)}(\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)})
\end{aligned}$$

holds. By matrix Bernstein inequality [Tro15, Theorem 6.1.1], we have

$$\begin{aligned}
&\mathbb{P}\left[\|\boldsymbol{b}_1\|_2\geqslant 100\left(\sqrt{\mathbb{E}_{\delta_{l,\cdot}}\sum_j\|\boldsymbol{s}_{1,j}\|_2^2\log(n_1\vee n_2)}+L_1^{(l)}(\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)})\log(n_1\vee n_2)\right)\mid\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)}\right]\\
\leqslant&(n_1+n_2)^{-15}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{P}\left[\|\boldsymbol{b}_1\|_2\geqslant 100\left(\sqrt{\mathbb{E}_{\delta_{l,\cdot}}\sum_j\|\boldsymbol{s}_{1,j}\|_2^2\log(n_1\vee n_2)}+L_1^{(l)}(\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)})\log(n_1\vee n_2)\right)\right]\\
=&\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\|\boldsymbol{b}_1\|_2\geqslant 100\left(\sqrt{\mathbb{E}_{\delta_{l,\cdot}}\sum_j\|\boldsymbol{s}_{1,j}\|_2^2\log(n_1\vee n_2)}+L_1^{(l)}(\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)})\log(n_1\vee n_2)\right)}\mid\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)}\right]\right]\\
\leqslant&(n_1+n_2)^{-15}.
\end{aligned}$$

In other words, on an event $E_B^{t,(l),1}$ with probability $\mathbb{P}[E_B^{t,(l),1}]\geqslant 1-(n_1+n_2)^{-15}$,

$$\|\boldsymbol{b}_1\|_2\leqslant 100\left(\sqrt{\mathbb{E}_{\delta_{l,\cdot}}\sum_j\|\boldsymbol{s}_{1,j}\|_2^2\log(n_1\vee n_2)}+L_1^{(l)}(\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)})\log(n_1\vee n_2)\right)\tag{4.57}$$

holds.

On the event E_{gd}^t, if

$$p\geqslant 111^2C_I^2\frac{\mu r\kappa^{11}\log(n_1\vee n_2)}{n_1\wedge n_2},$$

from (C.52), we have

$$\begin{aligned}\left\|\begin{bmatrix}\boldsymbol{X}^{t,(l)}\\ \boldsymbol{Y}^{t,(l)}\end{bmatrix}\boldsymbol{T}^{t,(l)}-\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\|_{2,\infty}&\leqslant\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})},\\ \left\|\begin{bmatrix}\boldsymbol{X}^{t,(l)}\\ \boldsymbol{Y}^{t,(l)}\end{bmatrix}\right\|_{2,\infty}&\leqslant2\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}.\end{aligned}\tag{4.58}$$

Therefore, from (C.52),

$$\begin{aligned}&L_1^{(l)}(\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)})\\ \leqslant&\frac{3}{p}\left\|\begin{bmatrix}\boldsymbol{X}^{t,(l)}\\ \boldsymbol{Y}^{t,(l)}\end{bmatrix}\boldsymbol{T}^{t,(l)}-\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\|_{2,\infty}\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\times\sqrt{\sigma_1(\boldsymbol{M})}\left\|\begin{bmatrix}\boldsymbol{X}^{t,(l)}\\ \boldsymbol{Y}^{t,(l)}\end{bmatrix}\right\|_{2,\infty}\\ \leqslant&\frac{333C_I}{p}\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sigma_1(\boldsymbol{M})\times2\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}\\ \leqslant&666C_I\rho^t\sqrt{\frac{\mu^4r^4\kappa^{14}\log(n_1\vee n_2)}{(n_1\wedge n_2)^4p^3}}\sqrt{\sigma_1(\boldsymbol{M})}^3.\end{aligned}\tag{4.59}$$

Moreover, for $\mathbb{E}_{\delta_{l,\cdot}}\sum_j\|\boldsymbol{s}_{1,j}\|_2^2$, we have

$$\begin{aligned}&\mathbb{E}_{\delta_{l,\cdot}}\sum_j\|\boldsymbol{s}_{1,j}\|_2^2\\ =&\mathbb{E}_{\delta_{l,\cdot}}\sum_j(\frac{1}{p}\delta_{l,j}-1)^2\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top-\boldsymbol{U}\boldsymbol{V}^\top\right)_{l,j}^2\|\boldsymbol{Y}_{j,\cdot}^{t,(l)}\|_2^2\\ \leqslant&\frac{1}{p}\|\boldsymbol{Y}^{t,(l)}\|_{2,\infty}^2\left\|\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top-\boldsymbol{U}\boldsymbol{V}^\top\right)_{l,\cdot}\right\|_2^2.\end{aligned}\tag{4.60}$$

From (4.21) and (4.23),

$$\begin{aligned}
&\left\| \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \\
\leqslant & \left\| \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^{t} - \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} \right\|_F + \left\| \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^{t} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \\
\leqslant & C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})} + C_I \rho^t \sqrt{\frac{\mu r \kappa^{6} \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} \sqrt{\sigma_1(\boldsymbol{M})} \\
\leqslant & 2 C_I \rho^t \sqrt{\frac{\mu r \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} \sqrt{\sigma_1(\boldsymbol{M})},
\end{aligned} \tag{4.61}$$

where the last inequality holds since

$$\frac{\mu r}{n_1 \wedge n_2} \leqslant p \leqslant 1.$$

By triangle inequality, and recall the decomposition (4.56),

$$\begin{aligned}
&\left\| \left(\boldsymbol{X}^{t,(l)} (\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top \right)_{l,\cdot} \right\|_2 \\
\leqslant & \left\| \boldsymbol{U}_{l,\cdot}^\top (\boldsymbol{Y}^{t,(l)} \boldsymbol{T}^{t,(l)} - \boldsymbol{V})^\top \right\|_2 + \left\| (\boldsymbol{X}^{t,(l)} \boldsymbol{T}^{t,(l)} - \boldsymbol{U})_{l,\cdot}^\top \boldsymbol{V}^\top \right\|_2 \\
& + \left\| (\boldsymbol{X}^{t,(l)} \boldsymbol{T}^{t,(l)} - \boldsymbol{U})_{l,\cdot}^\top (\boldsymbol{Y}^{t,(l)} \boldsymbol{T}^{t,(l)} - \boldsymbol{V})^\top \right\|_2, \\
\leqslant & \|\boldsymbol{U}\|_{2,\infty} \|\boldsymbol{Y}^{t,(l)} \boldsymbol{T}^{t,(l)} - \boldsymbol{V}\| + \|\boldsymbol{X}^{t,(l)} \boldsymbol{T}^{t,(l)} - \boldsymbol{U}\|_{2,\infty} \|\boldsymbol{V}\| \\
& + \|\boldsymbol{X}^{t,(l)} \boldsymbol{T}^{t,(l)} - \boldsymbol{U}\|_{2,\infty} \|\boldsymbol{Y}^{t,(l)} \boldsymbol{T}^{t,(l)} - \boldsymbol{V}\|.
\end{aligned}$$

Combining with (C.52) and (4.61) we have

$$
\begin{aligned}
&\left\|\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)_{l,\cdot}\right\|_2 \\
\leqslant & 2\sqrt{\frac{\mu r\kappa}{n_1 \wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}C_I\rho^t \times \sqrt{\frac{\mu r \kappa^{10}\log(n_1 \vee n_2)}{(n_1 \wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})} \\
& + 111 C_I\rho^t\sqrt{\frac{\mu^2 r^2\kappa^{12}\log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\sigma_1(\boldsymbol{M}) \\
& + 111 C_I\rho^t\sqrt{\frac{\mu^2 r^2\kappa^{12}\log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \times 2C_I\rho^t\sqrt{\frac{\mu r \kappa^{10}\log(n_1 \vee n_2)}{(n_1 \wedge n_2)p}}\sigma_1(\boldsymbol{M}) \\
\leqslant & 115 C_I\rho^t\sqrt{\frac{\mu^2 r^2\kappa^{12}\log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\sigma_1(\boldsymbol{M}),
\end{aligned} \tag{4.62}
$$

where the last inequality use the fact that

$$p \geqslant 111^2 C_I^2 \frac{\mu r \kappa^{10}\log(n_1 \vee n_2)}{n_1 \wedge n_2}$$

and $\rho < 1$.

Putting (4.58), (4.60) and (4.62) together we have

$$\mathbb{E}_{\delta_{l,\cdot}} \sum_j \|\boldsymbol{s}_{1,j}\|_2^2 \leqslant 230^2 C_I^2 \rho^{2t}\frac{\mu^3 r^3 \kappa^{13}\log(n_1 \vee n_2)}{(n_1 \wedge n_2)^3 p^2}\sigma_1^3(\boldsymbol{M}). \tag{4.63}$$

So by (4.57), (4.59) and (4.63), on the event $E_B^{t,(l),1} \bigcap E_{gd}^t$, we have

(4.64)

$$\begin{aligned}
&\|\boldsymbol{b}_1\|_2 \\
\leqslant& 100\rho^t \left(230 C_I \sqrt{\frac{\mu^3 r^3 \kappa^{13} \log^2(n_1 \vee n_2)}{(n_1 \wedge n_2)^3 p^2}} + 666 C_I \sqrt{\frac{\mu^4 r^4 \kappa^{14} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^4 p^3}} \log(n_1 \vee n_2) \right) \sqrt{\sigma_1(\boldsymbol{M})}^3 \\
=& 100 C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})} \sigma_r(\boldsymbol{M}) \kappa \\
&\times \left(230 \sqrt{\frac{\mu r \kappa^3 \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} + 666 \sqrt{\frac{\mu^2 r^2 \kappa^4 \log^2(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p^2}} \right) \\
\leqslant& 0.025 \sigma_r(\boldsymbol{M}) C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})},
\end{aligned}$$

where the last inequality holds since

$$p \geqslant 3.3856 \times 10^{12} \frac{\mu r \kappa^5 \log(n_1 \vee n_2)}{n_1 \wedge n_2}.$$

For $\boldsymbol{b}_2$ defined in (4.55), we can use almost the same argument. We can write $\boldsymbol{b}_2$ as

$$\boldsymbol{b}_2 = \sum_j \boldsymbol{e}_j (\frac{1}{p}\delta_{l,j} - 1) \left(\boldsymbol{X}^{t,(l)} (\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top \right)_{l,j} := \sum_j \boldsymbol{s}_{2,j}.$$

By the definition of $\boldsymbol{X}^{t,(l)}$ and $\boldsymbol{Y}^{t,(l)}$, we can see that $\boldsymbol{X}^{t,(l)}$ and $\boldsymbol{Y}^{t,(l)}$ are independent of $\delta_{l,1}, \cdots$, δ_{l,n_2}. Therefore, conditioned on $\boldsymbol{X}^{t,(l)}$ and $\boldsymbol{Y}^{t,(l)}$, $\boldsymbol{s}_{2,j}$'s are independent and $\mathbb{E}_{\delta_{l,\cdot}} \boldsymbol{s}_{2,j} = \boldsymbol{0}$. Note for all j,

$$\begin{aligned}
&\|\boldsymbol{s}_{2,j}\|_2 \\
\leqslant& \frac{1}{p} \left\| \boldsymbol{X}^{t,(l)} (\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top \right\|_{\ell_\infty} \\
\leqslant& \frac{1}{p} \left(\|\boldsymbol{X}^{t,(l)} \boldsymbol{T}^{t,(l)} - \boldsymbol{U}\|_{2,\infty} \|\boldsymbol{V}\|_{2,\infty} + \|\boldsymbol{Y}^{t,(l)} \boldsymbol{T}^{t,(l)} - \boldsymbol{V}\|_{2,\infty} \|\boldsymbol{U}\|_{2,\infty} \right) \qquad (4.65) \\
&+ \frac{1}{p} \|\boldsymbol{X}^{t,(l)} \boldsymbol{T}^{t,(l)} - \boldsymbol{U}\|_{2,\infty} \|\boldsymbol{Y}^{t,(l)} \boldsymbol{T}^{t,(l)} - \boldsymbol{V}\|_{2,\infty} \\
:=& L_2^{(l)} (\boldsymbol{X}^{t,(l)}, \boldsymbol{Y}^{t,(l)}).
\end{aligned}$$

By matrix Bernstein inequality [Tro15, Theorem 6.1.1], we have

$$\mathbb{P}\left[\|\boldsymbol{b}_2\|_2 \geqslant 100\left(\sqrt{\mathbb{E}_{\delta_{l,\cdot}}\sum_j \|\boldsymbol{s}_{2,j}\|_2^2\log(n_1\vee n_2)} + L_2^{(l)}(\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)})\log(n_1\vee n_2)\right) \mid \boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)}\right]$$
$$\leqslant (n_1+n_2)^{-15}.$$

Using the same argument in $\boldsymbol{b}_1$, we have that on an event $E_B^{t,(l),2}$ with probability $\mathbb{P}[E_B^{t,(l),2}] \geqslant 1-(n_1+n_2)^{-15}$,

$$\|\boldsymbol{b}_2\|_2 \leqslant 100\left(\sqrt{\mathbb{E}_{\delta_{l,\cdot}}\sum_j \|\boldsymbol{s}_{2,j}\|_2^2\log(n_1\vee n_2)} + L_2^{(l)}(\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)})\log(n_1\vee n_2)\right) \tag{4.66}$$

holds. Note on the event $E_B^{t,(l),2}\bigcap E_{gd}^t$, the estimation of $\|\boldsymbol{s}_{2,j}\|$ and $\mathbb{E}_{\delta_{l,\cdot}}\sum_j\|\boldsymbol{s}_{2,j}\|_2^2$ are in the same fashion with the one we did on $\boldsymbol{s}_{1,j}$: On the event E_{gd}^t, from (C.52), (4.58) and (4.65),

$$\begin{aligned}
&L_2^{(l)}(\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)})\\
\leqslant&\frac{3}{p}\left\|\begin{bmatrix}\boldsymbol{X}^{t,(l)}\\ \boldsymbol{Y}^{t,(l)}\end{bmatrix}\boldsymbol{T}^{t,(l)}-\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\|_{2,\infty}\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}\\
\leqslant&\frac{1}{p}333C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})}\times\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}\\
=&333C_I\rho^t\sqrt{\frac{\mu^3r^3\kappa^{13}\log(n_1\vee n_2)}{(n_1\wedge n_2)^3p^3}}\sigma_1(\boldsymbol{M}).
\end{aligned}$$

At the same time,

$$\begin{aligned}
\mathbb{E}_{\delta_{l,\cdot}}\sum_j\|\boldsymbol{s}_{2,j}\|_2^2 =&\mathbb{E}_{\delta_{l,\cdot}}\sum_j(\frac{1}{p}\delta_{l,j}-1)^2\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top-\boldsymbol{U}\boldsymbol{V}^\top\right)_{l,j}^2\\
\leqslant&\frac{1}{p}\left\|\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top-\boldsymbol{U}\boldsymbol{V}^\top\right)_{l,\cdot}\right\|_2^2\\
\leqslant&115^2C_I^2\rho^{2t}\frac{\mu^2r^2\kappa^{12}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p^2}\sigma_1^2(\boldsymbol{M}),
\end{aligned}$$

where the last inequality follows from (4.62). Therefore, on the event $E_{gd}^{t} \bigcap E_{B}^{t,(l),2}$,

(4.67)

$$\begin{aligned}
&\|\boldsymbol{A}_3\|_F\\
=&\|\boldsymbol{b}_2\|_2\|\boldsymbol{X}_{l,\cdot}^{t,(l)}\|_2\\
\leqslant&100\left(115C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log^2(n_1\vee n_2)}{(n_1\wedge n_2)^2p^2}}\sigma_1(\boldsymbol{M})+333C_I\rho^t\sqrt{\frac{\mu^3r^3\kappa^{13}\log(n_1\vee n_2)}{(n_1\wedge n_2)^3p^3}}\sigma_1(\boldsymbol{M})\log(n_1\vee n_2)\right)\\
&\times 2\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}\\
=&100C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^{10}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})}\sigma_r(\boldsymbol{M})\kappa\\
&\times\left(230\sqrt{\frac{\mu r\kappa^3\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}+666\sqrt{\frac{\mu^2r^2\kappa^4\log^2(n_1\vee n_2)}{(n_1\wedge n_2)^2p^2}}\right)\\
\leqslant&0.025\sigma_r(\boldsymbol{M})C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^{10}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})},
\end{aligned}$$

where the second inequality uses (4.58) and the last inequality holds since

$$p\geqslant 3.3856\times 10^{12}\frac{\mu r\kappa^5\log(n_1\vee n_2)}{n_1\wedge n_2}.$$

So in summary by (4.55), (4.64) and (4.67), on the event $E_B^{t,(l),1}\bigcap E_B^{t,(l),2}\bigcap E_{gd}^t$ we have

$$\left\|\begin{bmatrix}\boldsymbol{A}_2\\ \boldsymbol{A}_3\end{bmatrix}\right\|_F \leqslant 0.05\sigma_r(\boldsymbol{M})C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^{10}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})}. \tag{4.68}$$

Combining the estimations (4.54) and (4.68) for $\boldsymbol{A}_1$, $\boldsymbol{A}_2$ and $\boldsymbol{A}_3$ together, and using (4.52), we can see that on the event $E_B^{t,(l),1} \bigcap E_B^{t,(l),2} \bigcap E_{gd}^t$,

$$\begin{aligned}
&\left\| \begin{bmatrix} \boldsymbol{X}^{t+1} \\ \boldsymbol{Y}^{t+1} \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{X}^{t+1,(l)} \\ \boldsymbol{Y}^{t+1,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} \right\|_F \\
\leqslant & \|\boldsymbol{A}_1\|_F + \eta \left\| \begin{bmatrix} \boldsymbol{A}_2 \\ \boldsymbol{A}_3 \end{bmatrix} \right\|_F \\
\leqslant & (1 - 0.1\eta\sigma_r(\boldsymbol{M})) \left\| \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} \right\|_F \\
& \quad + 0.05\eta\sigma_r(\boldsymbol{M}) C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})} \\
\leqslant & C_I \rho^{t+1} \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})}
\end{aligned}$$

holds for $\rho = 1 - 0.05\eta\sigma_r(\boldsymbol{M})$ and fixed l satisfying $1 \leqslant l \leqslant n_1$, and the last inequality uses (4.23). The proof is all the same for l satisfying $n_1 + 1 \leqslant l \leqslant n_1 + n_2$. Let

$$E_{gd}^{t+1} = E_{gd}^t \bigcap \left(\bigcap_{l=1}^{n_1+n_2} E_B^{t,(l),1} \right) \bigcap \left(\bigcap_{l=1}^{n_1+n_2} E_B^{t,(l),2} \right),$$

so $E_{gd}^{t+1} \subset E_{gd}^t$, and from union bound, we have $\mathbb{P}[E_{gd}^t \backslash E_{gd}^{t+1}] \leqslant (n_1 + n_2)^{-10}$.

4.3.5. Proof of (4.24). Finally, we want to show that (4.24) can be directly implied by (4.21), (4.22) and (4.23). First, for any l satisfies $1 \leqslant l \leqslant n_1 + n_2$,

$$\begin{aligned}
&\left\| \left(\begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right)_{l,\cdot} \right\|_2 \\
\leqslant & \left\| \left(\begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{R}^{t,(l)} \right)_{l,\cdot} \right\|_2 + \left\| \left(\begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{R}^{t,(l)} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right)_{l,\cdot} \right\|_2 \qquad (4.69) \\
\leqslant & \left\| \begin{bmatrix} \boldsymbol{X}^{t} \\ \boldsymbol{Y}^{t} \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{R}^{t,(l)} \right\|_F + \left\| \left(\begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{R}^{t,(l)} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right)_{l,\cdot} \right\|_2 .
\end{aligned}$$

The second term of the last line is already controlled by (4.22), so our main goal is to control the first term. In order to do so, we want to apply Lemma 4.3.1 with

$$\boldsymbol{X}_0 := \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}, \; \boldsymbol{X}_1 := \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t, \; \boldsymbol{X}_2 := \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)}.$$

Note by the definition of $\boldsymbol{U}$ and $\boldsymbol{V}$, we have $\sigma_1(\boldsymbol{X}_0) = \sqrt{2\sigma_1(\boldsymbol{M})}$, $\sigma_2(\boldsymbol{X}_0) = \sqrt{2\sigma_2(\boldsymbol{M})}$, $\cdots$, $\sigma_r(\boldsymbol{X}_0) = \sqrt{2\sigma_r(\boldsymbol{M})}$, and $\sigma_1(\boldsymbol{X}_0)/\sigma_r(\boldsymbol{X}_0) = \sqrt{\kappa}$. In order to apply the lemma, note from (4.21) we have

$$\left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \left\| \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \leqslant 2C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2)p}} \sigma_1(\boldsymbol{M}).$$

And as long as

$$p \geqslant 16C_I^2 \frac{\mu r \kappa^8 \log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

we have

$$\left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \left\| \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \leqslant \frac{1}{2}\sigma_r(\boldsymbol{M}) \leqslant \frac{1}{2}\sigma_r^2(\boldsymbol{X}_0).$$

And also we have

$$\begin{aligned} & \left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} \right\| \left\| \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \\ \leqslant & \left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} \right\|_F \left\| \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \\ \leqslant & 2C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sigma_1(\boldsymbol{M}) \\ \leqslant & \frac{1}{4}\sigma_r(\boldsymbol{M}) \\ \leqslant & \frac{1}{4}\sigma_r^2(\boldsymbol{X}_0). \end{aligned}$$

Here second inequality we use (4.23) and third inequality holds because we have

$$p \geqslant 64C_I^2 \frac{\mu^2 r^2 \kappa^{12} \log(n_1 \vee n_2)}{n_1 \wedge n_2}.$$

Now by applying Lemma 4.3.1 we have

$$
\begin{aligned}
&\left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{R}^{t,(l)} \right\|_F \\
&\leqslant 5\kappa \left\| \begin{bmatrix} \boldsymbol{X}^t \\ \boldsymbol{Y}^t \end{bmatrix} \boldsymbol{R}^t - \begin{bmatrix} \boldsymbol{X}^{t,(l)} \\ \boldsymbol{Y}^{t,(l)} \end{bmatrix} \boldsymbol{T}^{t,(l)} \right\|_F \\
&\leqslant 10 C_I \rho^t \kappa \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})}.
\end{aligned} \tag{4.70}
$$

Plugging (4.22) and (4.70) into (4.69) we have (4.24).

Finally letting

$$
\begin{aligned}
C_{S3} =& 3.3856 \times 10^{12} + 6600 C_I + 32400 C_I^2 + C_{13} + 333^2 C_{13}^2 \\
&+ 2.42 \times 10^{10} C_{10} C_I^2 + C_{S1}
\end{aligned}
$$

finishes the whole proof of Lemma 4.2.4.

CHAPTER 5

Conclusion

This dissertation focuses on analyzing the nonconvex matrix completion problem, both from geometric perspective and algorithmic perspective. In Chapter 2, based upon the geometric analysis framework introduced in [JGN+17,GJZ17], we propose a model-free framework to analyze the nonconvex matrix completion problem. By introducing novel technologies including a powerful deterministic lemma (Lemma 2.3.5), we are able to characterize how close any local minimum is away from global minimum without assumptions on the underlying matrix. In Chapter 3, we introduce a unified framework analyzing nonconvex matrix completion problem with linear parameterized structures. Finally in Chapter 4, based upon prior work [MWCC18], we show that $\ell_{2,\infty}$-norm regularization is not necessary for nonconvex rectangular matrix completion.

There is still much room for us to explore. For example, is that possible to extend our model-free framework to other problems? In Chapter 3, in order to analyze parameterized matrix completion problem, we made strong assumptions on the underlying matrix to estimate. It is not yet clear if we could establish a model-free framework there, which should be investigated in the near future. In Chapter 3, we consider matrices which can be linearly parameterized. One natural question to consider is whether we could also analyze matrices with other special structures. Finally in Chapter 4, although we could show that the $\ell_{2,\infty}$-norm regularization is not necessary for rectangular matrix completion, the extra balancing penalization $\|\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y}\|_F^2$ is still required for theoretical analysis. It would be an interesting problem to see if we could remove it as in the case of matrix sensing [MLC19].

APPENDIX A

Supporting Proofs of Chapter 2

A.1. Proofs of supporting lemmas in Section 2.3

We present in this section the proofs of lemmas stated in Section 2.3.

A.1.0.1. *A proof of Lemma 2.3.5.*

Proof. First of all, by using the definition of matrix inner product and Hadamard product, we have

$$\begin{aligned} |\langle \mathcal{P}_{\Omega_0}(\boldsymbol{A}\boldsymbol{C}^\top), \mathcal{P}_{\Omega_0}(\boldsymbol{B}\boldsymbol{D}^\top)\rangle - t\langle \boldsymbol{A}\boldsymbol{C}^\top, \boldsymbol{B}\boldsymbol{D}^\top\rangle| =& |\langle \boldsymbol{\Omega}_0 - t\boldsymbol{J}, (\boldsymbol{A}\boldsymbol{C}^\top \circ \boldsymbol{B}\boldsymbol{D}^\top)\rangle| \\ \leqslant& \|\boldsymbol{\Omega}_0 - t\boldsymbol{J}\| \|(\boldsymbol{A}\boldsymbol{C}^\top \circ \boldsymbol{B}\boldsymbol{D}^\top)\|_*, \end{aligned} \tag{A.1}$$

The inequality holds by matrix Hölder's inequality. So the only thing left over is to give a bound of $\|(\boldsymbol{A}\boldsymbol{C}^\top \circ \boldsymbol{B}\boldsymbol{D}^\top)\|_*$. Notice one can decompose the matrix into sum of rank one matrices as following

$$\boldsymbol{A}\boldsymbol{C}^\top \circ \boldsymbol{B}\boldsymbol{D}^\top = \left(\sum_{k=1}^{r_1} \boldsymbol{A}_{\cdot,k}\boldsymbol{C}_{\cdot,k}^\top\right) \circ \left(\sum_{k=1}^{r_2} \boldsymbol{B}_{\cdot,k}\boldsymbol{D}_{\cdot,k}^\top\right) = \sum_{l=1}^{r_1}\sum_{m=1}^{r_2} (\boldsymbol{A}_{\cdot,l} \circ \boldsymbol{B}_{\cdot,m})(\boldsymbol{C}_{\cdot,l} \circ \boldsymbol{D}_{\cdot,m})^\top.$$

Recall $\boldsymbol{M}_{\cdot,j} = (M_{1,j}, M_{2,j}, \ldots, M_{n,j})^\top$ denotes the j-th column of any matrix $\boldsymbol{M} \in \mathbb{R}^{n\times m}$.

Therefore, one can upper bound the nuclear norm via

$$\begin{aligned} \|(\boldsymbol{A}\boldsymbol{C}^\top \circ \boldsymbol{B}\boldsymbol{D}^\top)\|_* \leqslant& \sum_{l=1}^{r_1}\sum_{m=1}^{r_2} \|(\boldsymbol{A}_{\cdot,l} \circ \boldsymbol{B}_{\cdot,m})(\boldsymbol{C}_{\cdot,l} \circ \boldsymbol{D}_{\cdot,m})^\top\|_* \\ =& \sum_{l=1}^{r_1}\sum_{m=1}^{r_2} \|\boldsymbol{A}_{\cdot,l} \circ \boldsymbol{B}_{\cdot,m}\|_2 \|\boldsymbol{C}_{\cdot,l} \circ \boldsymbol{D}_{\cdot,m}\|_2 \\ =& \sum_{l=1}^{r_1}\sum_{m=1}^{r_2} \sqrt{\sum_{k=1}^{n_1} A_{k,l}^2 B_{k,m}^2}\sqrt{\sum_{k=1}^{n_2} C_{k,l}^2 D_{k,m}^2}, \end{aligned}$$

where the first line is by the triangle inequality and we can replace nuclear norm by vector ℓ_2 norms in second line since the summands are all rank one matrices. By applying the Cauchy-Schwarz inequality for twice, we can obtain

$$
\begin{aligned}
\|(\boldsymbol{A}\boldsymbol{C}^\top \circ \boldsymbol{B}\boldsymbol{D}^\top)\|_* \leqslant & \sqrt{\sum_{l=1}^{r_1}\sum_{m=1}^{r_2}\sum_{k=1}^{n_1} A_{k,l}^2 B_{k,m}^2}\sqrt{\sum_{l=1}^{r_1}\sum_{m=1}^{r_2}\sum_{k=1}^{n_2} C_{k,l}^2 D_{k,m}^2} \\
= & \sqrt{\sum_{k=1}^{n_1} \|\boldsymbol{A}_{k,\cdot}\|_2^2 \|\boldsymbol{B}_{k,\cdot}\|_2^2}\sqrt{\sum_{k=1}^{n_2} \|\boldsymbol{C}_{k,\cdot}\|_2^2 \|\boldsymbol{D}_{k,\cdot}\|_2^2}.
\end{aligned} \tag{A.2}
$$

Combining (A.1) and (A.2) together, we have

$$
\begin{aligned}
& |\langle \mathcal{P}_{\Omega_0}(\boldsymbol{A}\boldsymbol{C}^\top), \mathcal{P}_{\Omega_0}(\boldsymbol{B}\boldsymbol{D}^\top)\rangle - t\langle \boldsymbol{A}\boldsymbol{C}^\top, \boldsymbol{B}\boldsymbol{D}^\top\rangle| \\
\leqslant & \|\boldsymbol{\Omega}_0 - t\boldsymbol{J}\| \sqrt{\sum_{k=1}^{n_1} \|\boldsymbol{A}_{k,\cdot}\|_2^2 \|\boldsymbol{B}_{k,\cdot}\|_2^2}\sqrt{\sum_{k=1}^{n_2} \|\boldsymbol{C}_{k,\cdot}\|_2^2 \|\boldsymbol{D}_{k,\cdot}\|_2^2}.
\end{aligned}
$$

□

A.1.0.2. *A proof of Lemma 2.3.9.*

Proof. The proof of Lemma 2.3.9 can be divided into the controls of $K_2(\boldsymbol{X})$, $K_3(\boldsymbol{X})$ and $K_4(\boldsymbol{X})$ separately. In order to combine the controls of $K_2(\boldsymbol{X})$, $K_3(\boldsymbol{X})$ and $K_4(\boldsymbol{X})$ together, von Neumann's trace inequality is employed.

For $K_2(\boldsymbol{X})$, we have

Lemma A.1.1. *In an event E_{Ca1} with probability $\mathbb{P}[E_{Ca1}] \geqslant 1-n^{-3}$, uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n\times r}$ and corresponding $\boldsymbol{\Delta}$ defined as before, we have*

$$
K_2(\boldsymbol{X}) \leqslant \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p}\left[19\sum_{i=1}^{n}\|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 + 18\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 + 9\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=s+1}^{r}\sigma_i\right] + 3\times 10^{-4}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2,
$$

where s is defined by

$$
s := \max\left\{s \leqslant r,\ \sigma_s \geqslant C_{Ca}\frac{\|\boldsymbol{M}_r\|_{\ell_\infty}\log n}{p}\right\} \tag{A.3}
$$

with C_{Ca} an absolute constant defined in Lemma 2.3.3. Set $s = 0$ if $\sigma_1 < C_{Ca}\frac{\|\boldsymbol{M}_r\|_{\ell_\infty}\log n}{p}$.

For $K_3(\boldsymbol{X})$, we use a modified version of [GJZ17, Lemma 11]:

Lemma A.1.2 ([GJZ17, Lemma 11]). *If $\alpha \geqslant 100\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}}$, then uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n\times r}$ and corresponding $\boldsymbol{\Delta}$ defined as before, we have*

$$K_3(\boldsymbol{X}) \leqslant 200\lambda\alpha^2\|\boldsymbol{\Delta}\|_F^2 - 0.3\lambda\sum_{i=1}^{n}\|\boldsymbol{\Delta}_{i,\cdot}\|_2^4.$$

The main modification we have made is that we keep the extra negative term.

For $K_4(\boldsymbol{X})$, we have

Lemma A.1.3. *Uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n\times r}$ and corresponding $\boldsymbol{\Delta}$ defined as before, we have*

$$\begin{aligned}K_4(\boldsymbol{X}) \leqslant & 5\times 10^{-4}\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 + 2\times 10^{-4}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + 10^5\frac{r\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\|^2}{p^2} \\ & + 6\langle\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle.\end{aligned}$$

We can apply Lemma 2.3.1 together with Lemma 2.3.6 to bound $\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\|$ and $\|\boldsymbol{\Omega} - p\boldsymbol{J}\|$ (similar result can also be found in [KMO10a]): As long as $p \geqslant C_v\frac{\log n}{n}$, we have

$$\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\| \leqslant C_v\sqrt{np}\|\boldsymbol{M}_{r+}\|_{\ell_\infty} \tag{A.4}$$

and

$$\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \leqslant C_v\sqrt{np} \tag{A.5}$$

hold in an event E_{v1} with probability $\mathbb{P}[E_{v1}] \geqslant 1 - n^{-3}$.

By putting Lemma A.1.1, Lemma A.1.2 and Lemma A.1.3 together, we have

$$\begin{aligned}&\sum_{i=2}^{4}K_i(\boldsymbol{X}) \\ \leqslant & \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p}\left[19\sum_{i=1}^{n}\|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 + 18\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 + 9\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=s+1}^{r}\sigma_i\right] + 3\times 10^{-4}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 \\ & + 200\lambda\alpha^2\|\boldsymbol{\Delta}\|_F^2 - 0.3\lambda\sum_{i=1}^{n}\|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 + 5\times 10^{-4}\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 \\ & + 2\times 10^{-4}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + 10^5\frac{r\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\|^2}{p^2} + 6\langle\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle.\end{aligned}$$

Replacing α, λ by the assumption $100\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}} \leqslant \alpha \leqslant 200\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}}$ as well as $100\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p} \leqslant \lambda \leqslant 200\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}$, we further have

$$\begin{aligned}
&\sum_{i=2}^{4} K_i(\boldsymbol{X}) \\
\leqslant&\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\left[19\sum_{i=1}^{n}\|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 + 18\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 + 9\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=s+1}^{r}\sigma_i\right] + 3\times10^{-4}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 \\
&+ 1.6\times10^{9}\|\boldsymbol{M}_r\|_{\ell_\infty}\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\|\boldsymbol{\Delta}\|_F^2 - 30\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\sum_{i=1}^{n}\|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 + 5\times10^{-4}\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 \\
&+ 2\times10^{-4}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + 10^5\frac{r\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\|^2}{p^2} + 6\langle\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle.
\end{aligned}$$

Combining with (A.4) and (A.5), and applying union bound,

$$\begin{aligned}
&\sum_{i=2}^{4} K_i(\boldsymbol{X}) \\
\leqslant&(19-30)\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\sum_{i=1}^{n}\|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 + (18+1.6\times10^9)\|\boldsymbol{M}_r\|_{\ell_\infty}\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\|\boldsymbol{\Delta}\|_F^2 \\
&+ 9\|\boldsymbol{M}_r\|_{\ell_\infty}\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\sum_{i=s+1}^{r}\sigma_i + (3+2)\times10^{-4}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 \\
&+ 5\times10^{-4}\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 + 10^5\frac{r\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\|^2}{p^2} + 6\langle\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle \\
\leqslant&5\times10^{-4}\left[\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right] + 10^5C_v^2\frac{nr}{p}\|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2 \\
&+ 2\times10^9C_v\sqrt{\frac{n}{p}}\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 + 9C_v\sqrt{\frac{n}{p}}\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=s+1}^{r}\sigma_i + 6\langle\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle,
\end{aligned} \tag{A.6}$$

holds in an event $E_1 := E_{Ca1} \cap E_{v1}$ with probability $\mathbb{P}[E] \geqslant 1 - 2n^{-3}$.

For $\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F^2$, we have

$$\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F^2 = \langle\boldsymbol{\Delta}^\top\boldsymbol{\Delta}, \boldsymbol{\Delta}^\top\boldsymbol{\Delta}\rangle = \sum_{i=1}^{r}\sigma_i^4(\boldsymbol{\Delta}), \tag{A.7}$$

where $\sigma_i(\boldsymbol{\Delta})$ denotes i-th largest singular value of $\boldsymbol{\Delta}$.

In order to proceed, we need the following von Neumann's trace inequality:

Lemma A.1.4 ([**Bha13**, Problem III.6.14]). *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ be two symmetric matrices, $\lambda_1(\boldsymbol{A}) \geqslant \lambda_2(\boldsymbol{A}) \geqslant \cdots \geqslant \lambda_n(\boldsymbol{A})$ and $\lambda_1(\boldsymbol{B}) \geqslant \lambda_2(\boldsymbol{B}) \geqslant \cdots \geqslant \lambda_n(\boldsymbol{B})$ are eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{B}$. Then the following holds:*

$$\sum_{i=1}^{n} \lambda_i(\boldsymbol{A})\lambda_{n+1-i}(\boldsymbol{B}) \leqslant \langle \boldsymbol{A}, \boldsymbol{B}\rangle \leqslant \sum_{i=1}^{n} \lambda_i(\boldsymbol{A})\lambda_i(\boldsymbol{B}).$$

This result can also be derived from Schur-Horn theorem (see, e.g., [**MOA11**, Theorem 9.B.1, Theorem 9.B.2]) together with Abel's summation formula.

From Lemma A.1.4, we have

$$\begin{aligned} \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 &= \mathrm{trace}(\boldsymbol{\Delta}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Delta}^\top) = \langle \boldsymbol{U}^\top\boldsymbol{U}, \boldsymbol{\Delta}^\top\boldsymbol{\Delta}\rangle \\ &\geqslant \sum_{i=1}^{r} \lambda_{r+1-i}(\boldsymbol{U}^\top\boldsymbol{U})\lambda_i(\boldsymbol{\Delta}^\top\boldsymbol{\Delta}) = \sum_{i=1}^{r} \sigma_i^2(\boldsymbol{\Delta})\sigma_{r+1-i}^2(\boldsymbol{U}), \end{aligned} \tag{A.8}$$

and

$$\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle \leqslant \sum_{i=1}^{n} \lambda_i(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top)\lambda_i(\boldsymbol{M}_{r+}) = \sum_{i=1}^{r} \sigma_i^2(\boldsymbol{\Delta})\sigma_i(\boldsymbol{M}_{r+}). \tag{A.9}$$

Here we use the fact that $\lambda_i(\boldsymbol{U}^\top\boldsymbol{U}) = \sigma_i^2(\boldsymbol{U})$, $\lambda_i(\boldsymbol{\Delta}^\top\boldsymbol{\Delta}) = \sigma_i^2(\boldsymbol{\Delta})$, $\lambda_i(\boldsymbol{M}_{r+}) = \sigma_i(\boldsymbol{M}_{r+})$ and

$$\lambda_i(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top) = \begin{cases} \sigma_i^2(\boldsymbol{\Delta}) & i = 1, \cdots, r \\ 0 & i = r+1, \cdots, n. \end{cases}$$

Putting (A.7), (A.8) and (A.9) together we have

$$\begin{aligned} &-5\times 10^{-4}\left[\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right] + 2\times 10^9 C_v\sqrt{\frac{n}{p}}\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 + 6\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle \\ \leqslant &-5\times 10^{-4}\left[\sum_{i=1}^{r}\sigma_i^4(\boldsymbol{\Delta}) + \sum_{i=1}^{r}\sigma_i^2(\boldsymbol{\Delta})\sigma_{r+1-i}^2(\boldsymbol{U})\right] + 2\times 10^9 C_v\sqrt{\frac{n}{p}}\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=1}^{r}\sigma_i^2(\boldsymbol{\Delta}) \\ &+ 6\sum_{i=1}^{r}\sigma_i^2(\boldsymbol{\Delta})\sigma_i(\boldsymbol{M}_{r+}) \\ \leqslant &5\times 10^{-4}\sum_{i=1}^{r}\left\{-\sigma_i^4(\boldsymbol{\Delta}) + \left[4\times 10^{12}C_v\sqrt{\frac{n}{p}}\|\boldsymbol{M}_r\|_{\ell_\infty} - \sigma_{r+1-i}^2(\boldsymbol{U}) + 1.2\times 10^4\sigma_i(\boldsymbol{M}_{r+})\right]\sigma_i^2(\boldsymbol{\Delta})\right\}. \end{aligned}$$

For the last line, the summands are a series of quadratic functions of $\sigma_i^2(\boldsymbol{\Delta})$. Noticing the fact that for a quadratic function $q(x) = -x^2 + bx$, given the constraint $x \geqslant 0$, the maximum is taken at

$\hat{x} = \frac{1}{2}[b]_+$, and the maximum value is $\frac{1}{4}\{[b]_+\}^2$. Therefore, we have

$$
\begin{aligned}
& -5 \times 10^{-4}\left[\|\boldsymbol{\Delta}^\top \boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right] + 2 \times 10^9 C_v \sqrt{\frac{n}{p}}\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 + 6\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle \\
\leqslant & \frac{5}{4} \times 10^{-4} \sum_{i=1}^{r} \left\{ \left[4 \times 10^{12} C_v \sqrt{\frac{n}{p}} \|\boldsymbol{M}_r\|_{\ell_\infty} - \sigma_{r+1-i}^2(\boldsymbol{U}) + 1.2 \times 10^4 \sigma_i(\boldsymbol{M}_{r+}) \right]_+ \right\}^2 \\
= & 1.25 \times 10^{-4} \sum_{j=1}^{r} \left\{ \left[4 \times 10^{12} C_v \sqrt{\frac{n}{p}} \|\boldsymbol{M}_r\|_{\ell_\infty} - \sigma_j^2(\boldsymbol{U}) + 1.2 \times 10^4 \sigma_{r+1-j}(\boldsymbol{M}_{r+}) \right]_+ \right\}^2 \\
= & 1.25 \times 10^{-4} \sum_{j=1}^{r} \left\{ \left[4 \times 10^{12} C_v \sqrt{\frac{n}{p}} \|\boldsymbol{M}_r\|_{\ell_\infty} + 1.2 \times 10^4 \sigma_{2r+1-j} - \sigma_j \right]_+ \right\}^2 .
\end{aligned} \tag{A.10}
$$

In the second last line, we let $j = r + 1 - i$. In the last line, we use the fact that

$$\sigma_{r+1-j}(\boldsymbol{M}_{r+}) = \sigma_{r+r+1-j}(\boldsymbol{M}) = \sigma_{2r+1-j}$$

and

$$\sigma_j^2(\boldsymbol{U}) = \sigma_j(\boldsymbol{U}\boldsymbol{U}^\top) = \sigma_j(\boldsymbol{M}_r) = \sigma_j(\boldsymbol{M}) = \sigma_j.$$

Finally putting (A.6) and (A.10) together we have

$$
\begin{aligned}
& \sum_{i=2}^{4} K_i(\boldsymbol{X}) \\
\leqslant & 10 \times 10^{-4}\left[\|\boldsymbol{\Delta}^\top \boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right] + 10^5 C_v^2 \frac{nr}{p} \|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2 \\
& - 5 \times 10^{-4}\left[\|\boldsymbol{\Delta}^\top \boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right] + 2 \times 10^9 C_v \sqrt{\frac{n}{p}}\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 \\
& + 9 C_v \sqrt{\frac{n}{p}} \|\boldsymbol{M}_r\|_{\ell_\infty} \sum_{i=s+1}^{r} \sigma_i + 6\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle \\
\leqslant & 10^{-3}\left[\|\boldsymbol{\Delta}^\top \boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right] + 10^5 C_v^2 \frac{nr}{p} \|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2 \\
& + 1.25 \times 10^{-4} \sum_{i=1}^{r} \left\{ \left[4 \times 10^{12} C_v \sqrt{\frac{n}{p}} \|\boldsymbol{M}_r\|_{\ell_\infty} + 1.2 \times 10^4 \sigma_{2r+1-i} - \sigma_i \right]_+ \right\}^2 \\
& + 9 C_v \sqrt{\frac{n}{p}} \|\boldsymbol{M}_r\|_{\ell_\infty} \sum_{i=s+1}^{r} \sigma_i .
\end{aligned} \tag{A.11}
$$

Recall by the definition of s in (A.3), for any $i > s$, we have $\sigma_i < C_{Ca}\frac{\|\boldsymbol{M}_r\|_{\ell_\infty}\log n}{p}$. Therefore, we have for any $i > s$,

$$\begin{aligned}
&2C_{Ca}\left(\sqrt{\frac{n}{p}}+\frac{\log n}{p}\right)\|\boldsymbol{M}_r\|_{\ell_\infty}-\sigma_i\\
=&2C_{Ca}\left(\sqrt{\frac{n}{p}}+\frac{\log n}{p}\right)\|\boldsymbol{M}_r\|_{\ell_\infty}-2\sigma_i+\sigma_i\\
\geqslant&2C_{Ca}\left(\sqrt{\frac{n}{p}}+\frac{\log n}{p}\right)\|\boldsymbol{M}_r\|_{\ell_\infty}-2C_{Ca}\frac{\|\boldsymbol{M}_r\|_{\ell_\infty}\log n}{p}+\sigma_i\\
\geqslant&2C_{Ca}\sqrt{\frac{n}{p}}\|\boldsymbol{M}_r\|_{\ell_\infty}+\sigma_i\\
\geqslant&0.
\end{aligned}$$

Therefore, for all $i > s$,

$$\begin{aligned}
&\left\{\left[2C_{Ca}\left(\sqrt{\frac{n}{p}}+\frac{\log n}{p}\right)\|\boldsymbol{M}_r\|_{\ell_\infty}-\sigma_i\right]_+\right\}^2\\
\geqslant&\left[2C_{Ca}\sqrt{\frac{n}{p}}\|\boldsymbol{M}_r\|_{\ell_\infty}+\sigma_i\right]^2\\
\geqslant&4C_{Ca}\sqrt{\frac{n}{p}}\|\boldsymbol{M}_r\|_{\ell_\infty}\sigma_i.
\end{aligned}$$

Combining with (A.11), we have

$$\begin{aligned}
\sum_{i=2}^{4}K_i(\boldsymbol{X})\leqslant&10^{-3}\left[\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F^2+\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right]+10^5C_v^2\frac{nr}{p}\|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2\\
&+1.25\times10^{-4}\sum_{i=1}^{r}\left\{\left[4\times10^{12}C_v\sqrt{\frac{n}{p}}\|\boldsymbol{M}_r\|_{\ell_\infty}+1.2\times10^4\sigma_{2r+1-i}-\sigma_i\right]_+\right\}^2\\
&+\frac{9C_v}{4C_{Ca}}\sum_{i=1}^{r}\left\{\left[2C_{Ca}\left(\sqrt{\frac{n}{p}}+\frac{\log n}{p}\right)\|\boldsymbol{M}_r\|_{\ell_\infty}-\sigma_i\right]_+\right\}^2.
\end{aligned}$$

By letting

$$C_2=\max\{4\times10^{12}C_v,1.2\times10^4,2C_{Ca}\} \tag{A.12}$$

and

$$C_3 = \max\{10^5 C_v^2, 1.25 \times 10^{-4} + \frac{9C_v}{4C_{Ca}}\} \tag{A.13}$$

we are able to finish the proof.

□

A.1.0.3. *A proof of Lemma A.1.1.*

Proof. Recall that we define $\boldsymbol{\Delta}$ as $\boldsymbol{\Delta} := \boldsymbol{X} - \boldsymbol{U}$, $D_{\Omega,p}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top)$ can be decomposed as following

$$\begin{aligned} & D_{\Omega,p}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top) \\ =& D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top) \\ =& \underbrace{D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top)}_{①} + \underbrace{D_{\Omega,p}(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)}_{②} + \underbrace{4D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)}_{③}. \end{aligned} \tag{A.14}$$

Here we use the fact that Ω is symmetric. Our strategy here is using Lemma 2.3.3 to give a tight bound to as many as possible terms, for those terms that Lemma 2.3.3 cannot handle, we use Lemma 2.3.5 to give a bound. To be more precise, for ② and ③, as Lemma 2.3.3 cannot apply here, we use Lemma 2.3.5 to give a bound. For ①, we need to split it into two parts, the good part we can use Lemma 2.3.3 to control, and the rest part we use Lemma 2.3.5 to give a bound.

First for ② and ③, by applying Lemma 2.3.5,

$$|②| = |D_{\Omega,p}(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)| \leqslant \|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 \tag{A.15}$$

and

$$\begin{aligned} |③| = 4|D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)| \leqslant & 4\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sqrt{\sum_{i=1}^n \|\boldsymbol{U}_{i,\cdot}\|_2^2 \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2} \sqrt{\sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4} \\ \leqslant & 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{\Delta}\|_F^2 + 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4, \end{aligned} \tag{A.16}$$

where for the second inequality we use the fact that $2xy \leqslant x^2 + y^2$.

Finally for $\textcircled{1}$, if $\boldsymbol{U}$ is good enough such that the incoherence $\mu(\boldsymbol{U})$ is well-bounded, then we can apply Lemma 2.3.3 directly and get a tight bound. If $\mu(\boldsymbol{U})$ is not good enough, we want to split $\boldsymbol{U}$ into two parts and hope first few columns have good incoherence. To be more precise, recall that we assume $\boldsymbol{U} = \boldsymbol{U}_r = [\sqrt{\sigma_1}\boldsymbol{u}_1 \ \dots \ \sqrt{\sigma_r}\boldsymbol{u}_r]$, similar to (2.7), for the incoherence of the first k columns, we have

$$\begin{aligned}
&\mu\left(\text{colspan}([\sqrt{\sigma_1}\boldsymbol{u}_1 \ \dots \ \sqrt{\sigma_k}\boldsymbol{u}_k])\right)\\
=&\frac{n}{k}\max_i \sum_{j=1}^{k} u_{i,j}^2 \leqslant \frac{n}{k\sigma_k}\max_i \sum_{j=1}^{k} \sigma_j u_{i,j}^2 \leqslant \frac{n}{k\sigma_k}\max_i \sum_{j=1}^{r} \sigma_j u_{i,j}^2 \leqslant \frac{n\|\boldsymbol{M}_r\|_{\ell_\infty}}{k\sigma_k},
\end{aligned} \tag{A.17}$$

where $\mu(\cdot)$ is defined in (2.8).

For fixed s defined as in (A.3), denote first s columns of $\boldsymbol{U}$ as $\boldsymbol{U}^1$, and remaining part as $\boldsymbol{U}^2$. Decompose $\boldsymbol{U}$ as $\boldsymbol{U} = [\boldsymbol{U}^1 \ \boldsymbol{U}^2]$, and $\boldsymbol{\Delta}$ can also be decomposed as $\boldsymbol{\Delta} = [\boldsymbol{\Delta}^1 \ \boldsymbol{\Delta}^2]$ correspondingly. Note by our assumption that $\boldsymbol{U} = \boldsymbol{U}_r$, we have $(\boldsymbol{U}^1)^\top \boldsymbol{U}^2 = \boldsymbol{0}$. So we can further decompose the first term of (A.14) as

$$\begin{aligned}
\textcircled{1} =&D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top)\\
=&D_{\Omega,p}\Big([\boldsymbol{U}^1 \ \boldsymbol{U}^2][\boldsymbol{\Delta}^1 \ \boldsymbol{\Delta}^2]^\top + [\boldsymbol{\Delta}^1 \ \boldsymbol{\Delta}^2][\boldsymbol{U}^1 \ \boldsymbol{U}^2]^\top, [\boldsymbol{U}^1 \ \boldsymbol{U}^2][\boldsymbol{\Delta}^1 \ \boldsymbol{\Delta}^2]^\top\\
&\qquad + [\boldsymbol{\Delta}^1 \ \boldsymbol{\Delta}^2][\boldsymbol{U}^1 \ \boldsymbol{U}^2]^\top\Big)\\
=&\underbrace{D_{\Omega,p}\left(\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top + \boldsymbol{\Delta}^1(\boldsymbol{U}^1)^\top, \boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top + \boldsymbol{\Delta}^1(\boldsymbol{U}^1)^\top\right)}_{A_1}\\
&+\underbrace{4D_{\Omega,p}\left(\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top, \boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top\right)}_{A_2} + \underbrace{2D_{\Omega,p}\left(\boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top, \boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top\right)}_{A_3}\\
&+\underbrace{2D_{\Omega,p}\left(\boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top, \boldsymbol{\Delta}^2(\boldsymbol{U}^2)^\top\right)}_{A_4} + \underbrace{4D_{\Omega,p}\left(\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top, \boldsymbol{\Delta}^2(\boldsymbol{U}^2)^\top\right)}_{A_5}.
\end{aligned} \tag{A.18}$$

Now we can apply tight approximation Lemma 2.3.3 to the first term of (A.18). By the way we choose s, combining with (A.17),

$$p \geqslant C_{Ca}\frac{\|\boldsymbol{M}_r\|_{\ell_\infty}\log n}{\sigma_s} \geqslant C_{Ca}\frac{\|\boldsymbol{M}_r\|_{\ell_\infty}\log n}{\sigma_s}\cdot\frac{\mu\left(\text{colspan}(\boldsymbol{U}^1)\right)s\sigma_s}{n\|\boldsymbol{M}_r\|_{\ell_\infty}} = C_{Ca}\frac{\mu\left(\text{colspan}(\boldsymbol{U}^1)\right)s\log n}{n}.$$

Therefore, Lemma 2.3.3 ensures that

$$
\begin{aligned}
|A_1| =& \left| D_{\Omega,p}\left(\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top + \boldsymbol{\Delta}^1(\boldsymbol{U}^1)^\top, \boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top + \boldsymbol{\Delta}^1(\boldsymbol{U}^1)^\top\right)\right| \\
\leqslant& 10^{-5}p\|\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top + \boldsymbol{\Delta}^1(\boldsymbol{U}^1)^\top\|_F^2 \\
\leqslant& 2\times 10^{-5}p(\|\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top\|_F^2 + \|\boldsymbol{\Delta}^1(\boldsymbol{U}^1)^\top\|_F^2) \\
\leqslant& 10^{-4}p\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2
\end{aligned}
\tag{A.19}
$$

hold in an event E_{Ca1} with probability $\mathbb{P}[E_{Ca1}] \geqslant 1 - n^{-3}$, where the second inequality uses the fact that $(x+y)^2 \leqslant 2x^2 + 2y^2$, and last inequality uses the fact that $(\boldsymbol{U}^1)^\top \boldsymbol{U}^2 = \boldsymbol{0}$.

For the rest terms in (A.18), by applying Lemma 2.3.5 we have

$$
\begin{aligned}
|A_2| =& 4|D_{\Omega,p}(\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top, \boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top)| \\
\leqslant& 4\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sqrt{\sum_{i=1}^n \|\boldsymbol{U}^1_{i,\cdot}\|_2^2 \|\boldsymbol{U}^2_{i,\cdot}\|_2^2} \sqrt{\sum_{i=1}^n \|\boldsymbol{\Delta}^1_{i,\cdot}\|_2^2 \|\boldsymbol{\Delta}^2_{i,\cdot}\|_2^2} \\
\leqslant& 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \left[\|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{U}^2\|_F^2 + \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 \right]
\end{aligned}
\tag{A.20}
$$

for the second term in (A.18), where the second inequality use the fact that $\|\boldsymbol{U}^1_{i,\cdot}\|_2^2 \leqslant \|\boldsymbol{U}_{i,\cdot}\|_2^2 \leqslant \|\boldsymbol{M}_r\|_{\ell_\infty}, \|\boldsymbol{\Delta}^1_{i,\cdot}\|_2^2 \leqslant \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2, \|\boldsymbol{\Delta}^2_{i,\cdot}\|_2^2 \leqslant \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2$ and $2xy \leqslant x^2 + y^2$. For the third term, applying Lemma 2.3.5 again we have

$$
\begin{aligned}
|A_3| =& 2|D_{\Omega,p}(\boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top, \boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top)| \\
\leqslant& 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sqrt{\sum_{i=1}^n \|\boldsymbol{U}^2_{i,\cdot}\|_2^4} \sqrt{\sum_{i=1}^n \|\boldsymbol{\Delta}^2_{i,\cdot}\|_2^4} \\
\leqslant& \|\boldsymbol{\Omega} - p\boldsymbol{J}\| \left[\|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{U}^2\|_F^2 + \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 \right],
\end{aligned}
\tag{A.21}
$$

where for the second inequality we also use the properties used in bounding second term. For the fourth and last term in (A.18), applying Lemma 2.3.5 and properties listed above, we have

$$\begin{aligned}
|A_4| =& 2|D_{\Omega,p}(\boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top, \boldsymbol{\Delta}^2(\boldsymbol{U}^2)^\top)| \\
\leqslant& 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sum_{i=1}^{n} \|\boldsymbol{U}^2_{i,\cdot}\|_2^2 \|\boldsymbol{\Delta}^2_{i,\cdot}\|_2^2 \\
\leqslant& 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{\Delta}\|_F^2
\end{aligned} \tag{A.22}$$

and

$$\begin{aligned}
|A_5| =& 4|D_{\Omega,p}(\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top, \boldsymbol{\Delta}^2(\boldsymbol{U}^2)^\top)| \\
\leqslant& 4\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sqrt{\sum_{i=1}^{n} \|\boldsymbol{U}^1_{i,\cdot}\|_2^2 \|\boldsymbol{\Delta}^2_{i,\cdot}\|_2^2} \sqrt{\sum_{i=1}^{n} \|\boldsymbol{U}^2_{i,\cdot}\|_2^2 \|\boldsymbol{\Delta}^1_{i,\cdot}\|_2^2} \\
\leqslant& 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{\Delta}^1\|_F^2 + 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{\Delta}^2\|_F^2 \\
=& 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{\Delta}\|_F^2.
\end{aligned} \tag{A.23}$$

Now putting estimations of terms in (A.18) listed above together, i.e., (A.19), (A.20), (A.21), (A.22) and (A.23), we have

$$\begin{aligned}
|\textcircled{1}| =& |D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top)| \\
\leqslant& |A_1| + |A_2| + |A_3| + |A_4| + |A_5| \\
\leqslant& 10^{-4} p \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \left[\|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{U}^2\|_F^2 + \sum_{i=1}^{n} \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 \right] \\
& + \|\boldsymbol{\Omega} - p\boldsymbol{J}\| \left[\|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{U}^2\|_F^2 + \sum_{i=1}^{n} \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 \right] + 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{\Delta}\|_F^2 \\
& + 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{\Delta}\|_F^2 \\
\leqslant& \|\boldsymbol{\Omega} - p\boldsymbol{J}\| \left[3\|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{U}^2\|_F^2 + 3\sum_{i=1}^{n} \|\boldsymbol{\Delta}_i\|_2^4 + 4\|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{\Delta}\|_F^2 \right] + 10^{-4} p \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2.
\end{aligned} \tag{A.24}$$

Plugging estimations (A.15), (A.16) and (A.24) back to (A.14), we have

$$\begin{aligned}
&|D_{\Omega,p}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top)| \\
\leqslant & |\textcircled{1}| + |\textcircled{2}| + |\textcircled{3}| \\
\leqslant & \|\boldsymbol{\Omega} - p\boldsymbol{J}\| \left[3\|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{U}^2\|_F^2 + 3\sum_{i=1}^n \|\boldsymbol{\Delta}_i\|_2^4 + 4\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 \right] + 10^{-4}p\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 \\
& + \|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 + 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 + 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 \\
= & \|\boldsymbol{\Omega} - p\boldsymbol{J}\| \left[3\|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{U}^2\|_F^2 + 6\sum_{i=1}^n \|\boldsymbol{\Delta}_i\|_2^4 + 6\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 \right] + 10^{-4}p\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2.
\end{aligned}$$

Therefore, combining with (A.15), we have

$$\begin{aligned}
K_2(\boldsymbol{X}) \leqslant & \frac{1}{p}|D_{\Omega,p}(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)| + \frac{3}{p}|D_{\Omega,p}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top)| \\
\leqslant & \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p} \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 + 3\frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p} \left[3\|\boldsymbol{M}_r\|_{\ell_\infty} \|\boldsymbol{U}^2\|_F^2 + 6\sum_{i=1}^n \|\boldsymbol{\Delta}_i\|_2^4 + 6\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 \right] \\
& + 3\times 10^{-4}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 \\
\leqslant & \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p} \left[19\sum_{i=1}^n \|\boldsymbol{\Delta}_i\|_2^4 + 18\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 + 9\|\boldsymbol{M}_r\|_{\ell_\infty} \sum_{i=s+1}^r \sigma_i \right] + 3\times 10^{-4}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2.
\end{aligned}$$

The last line uses the fact that $\|\boldsymbol{U}^2\|_F^2 = \sum_{i=s+1}^r \sigma_i$.

□

A.1.0.4. *Proof of Lemma A.1.2.* Here we present a proof of Lemma A.1.2. This proof is exactly the proof in [**GJZ17**] except keeping the extra negative term. We include the proof in [**GJZ17**] here for completeness.

Proof. By [**GJZ17**, Lemma 18], we have

$$\begin{aligned}
& \operatorname{vec}(\boldsymbol{\Delta})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{\Delta}) - 4\langle \nabla G_\alpha(\boldsymbol{X}), \boldsymbol{\Delta}\rangle \\
= & 4\sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^3 \frac{\|\boldsymbol{X}_{i,\cdot}\|_2^2 \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2 - \langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot}\rangle^2}{\|\boldsymbol{X}_{i,\cdot}\|_2^3} \\
& + 12\sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^2 \frac{\langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot}\rangle^2}{\|\boldsymbol{X}_{i,\cdot}\|_2^2} - 16\sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^3 \frac{\langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot}\rangle}{\|\boldsymbol{X}_{i,\cdot}\|_2}.
\end{aligned} \tag{A.25}$$

First of all, since we choose $\alpha \geqslant 100\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}} = 100\|\boldsymbol{U}\|_{2,\infty}$, then for all $\boldsymbol{X}_{i,\cdot}$ satisfying $\|\boldsymbol{X}_{i,\cdot}\|_2 \geqslant \alpha$, we have

$$\langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot}\rangle = \langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{X}_{i,\cdot} - \boldsymbol{U}_{i,\cdot}\rangle \geqslant \|\boldsymbol{X}_{i,\cdot}\|_2^2 - \|\boldsymbol{X}_{i,\cdot}\|_2\|\boldsymbol{U}_{i,\cdot}\|_2 \geqslant (1-0.01)\|\boldsymbol{X}_{i,\cdot}\|_2^2 \geqslant 0.99\|\boldsymbol{X}_{i,\cdot}\|_2^2, \tag{A.26}$$

which gives an lower bound of the inner product between $\boldsymbol{X}_{i,\cdot}$ and $\boldsymbol{\Delta}_{i,\cdot}$. At the same time, we can also upper bound $\|\boldsymbol{\Delta}_{i,\cdot}\|_2$ by $\|\boldsymbol{X}_{i,\cdot}\|_2$:

$$\|\boldsymbol{\Delta}_{i,\cdot}\|_2 \leqslant \|\boldsymbol{X}_{i,\cdot}\|_2 + \|\boldsymbol{U}_{i,\cdot}\|_2 \leqslant 1.01\|\boldsymbol{X}_{i,\cdot}\|_2. \tag{A.27}$$

Plugging the above two estimations (A.26), (A.27) together with the fact that $|\langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot}\rangle|^2 \leqslant \|\boldsymbol{X}_{i,\cdot}\|_2^2\|\boldsymbol{\Delta}_{i,\cdot}\|_2^2$ into (A.25), we have

$$\begin{aligned} &\operatorname{vec}(\boldsymbol{\Delta})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{\Delta}) - 4\langle \nabla G_\alpha(\boldsymbol{X}), \boldsymbol{\Delta}\rangle \\ \leqslant& -15.68\sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^3\|\boldsymbol{X}_{i,\cdot}\|_2 + 12\sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^2\|\boldsymbol{\Delta}_{i,\cdot}\|_2^2. \end{aligned} \tag{A.28}$$

Moreover, for all $\boldsymbol{X}_{i,\cdot}$ satisfies $\|\boldsymbol{X}_{i,\cdot}\|_2 \geqslant 5\alpha$, we can also upper bound $\|\boldsymbol{\Delta}_{i,\cdot}\|_2$ by $\|\boldsymbol{X}_{i,\cdot}\|_2$:

$$\|\boldsymbol{\Delta}_{i,\cdot}\|_2 \leqslant \|\boldsymbol{X}_{i,\cdot}\|_2 + \|\boldsymbol{U}_{i,\cdot}\|_2 \leqslant 1.002\|\boldsymbol{X}_{i,\cdot}\|_2, \tag{A.29}$$

and also lower bound $\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha$ by $\|\boldsymbol{\Delta}_{i,\cdot}\|_2$:

$$\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha \geqslant \left(1 - \frac{1}{5}\right)\|\boldsymbol{X}_{i,\cdot}\|_2 \geqslant \frac{400}{501}\|\boldsymbol{\Delta}_{i,\cdot}\|_2. \tag{A.30}$$

Plugging (A.29) and (A.30) back to (A.28), we have

$$\begin{aligned}
&\operatorname{vec}(\boldsymbol{\Delta})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{\Delta}) - 4\langle \nabla G_\alpha(\boldsymbol{X}), \boldsymbol{\Delta}\rangle \\
\leqslant & 12 \sum_{i, \|\boldsymbol{X}_{i,\cdot}\|_2 < 5\alpha} [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^2 \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2 \\
&+ \left[12 - 15.68 \times \frac{400}{501} \times \frac{1}{1.002}\right] \sum_{i, \|\boldsymbol{X}_{i,\cdot}\|_2 \geqslant 5\alpha} [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^2 \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2 \\
\leqslant & 192\alpha^2 \|\boldsymbol{\Delta}\|_F^2 - 0.3 \sum_{i, \|\boldsymbol{X}_{i,\cdot}\|_2 \geqslant 5\alpha} \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 \\
\leqslant & 200\alpha^2 \|\boldsymbol{\Delta}\|_F^2 - 0.3 \sum_{i=1}^{n} \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4,
\end{aligned}$$

where the last inequality uses the fact that $\|\boldsymbol{\Delta}_{i,\cdot}\|_2 \leqslant \|\boldsymbol{X}_{i,\cdot}\|_2 + \|\boldsymbol{U}_{i,\cdot}\|_2$ and $\alpha \geqslant 100\sqrt{\|\boldsymbol{M}_r\|_{\ell_\infty}}$. □

A.1.0.5. *A proof of Lemma A.1.3.*

Proof. First, by matrix Hölder's inequality,

$$6|\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\rangle| \leqslant 6 \frac{\sqrt{p}\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_*}{\sqrt{r}} \frac{\sqrt{r}\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\|}{\sqrt{p}}.$$

Since $\boldsymbol{\Delta}\boldsymbol{\Delta}^\top$ is at most rank-r, $\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_* \leqslant \sqrt{r}\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F$. Therefore,

$$\begin{aligned}
6|\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\rangle| \leqslant & 6\sqrt{p}\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F \frac{\sqrt{r}\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\|}{\sqrt{p}} \\
\leqslant & 5 \times 10^{-4} p \|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 + 1.8 \times 10^4 \frac{r\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\|^2}{p}.
\end{aligned}$$

For the last inequality, we also use the fact that $2xy \leqslant wx^2 + \frac{y^2}{w}$ for all $w > 0$. Use the same argument we also have

$$8|\langle \boldsymbol{U}\boldsymbol{\Delta}^\top, \mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\rangle| \leqslant 2 \times 10^{-4} p \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + 8 \times 10^4 \frac{r\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\|^2}{p},$$

Therefore, by the way we define $K_4(\boldsymbol{X})$ in (2.14), we have

$$\begin{aligned}K_4(\boldsymbol{X}) \leqslant& \frac{1}{p}|6\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \mathcal{P}_\Omega(\boldsymbol{M}_{r+})\rangle - 6p\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle| + \frac{1}{p}|8\langle \boldsymbol{U}\boldsymbol{\Delta}^\top, \mathcal{P}_\Omega(\boldsymbol{M}_{r+})\rangle - 8p\langle \boldsymbol{U}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle| \\ &+ 6\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle \\ \leqslant& 5\times 10^{-4}\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 + 2\times 10^{-4}\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + 10^5\frac{r\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+}) - p\boldsymbol{M}_{r+}\|^2}{p^2} \\ &+ 6\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle.\end{aligned}$$

□

A.2. Proof of Corollary 2.1.3

PROOF. The inequality (2.7) gives $\|\boldsymbol{M}\|_{\ell_\infty} \leqslant \frac{\mu_r r \sigma_1}{n}$. Therefore, in the case $\mathrm{rank}(\boldsymbol{M}) = r$, the approximation error bound (2.4) becomes

$$\left\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\right\|_F^2 \leqslant C_1 \sum_{i=1}^r \left\{\left[C_2\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\frac{\mu_r r}{n}\sigma_1 - \sigma_i\right]_+\right\}^2.$$

Therefore, if

$$p \geqslant 4C_2 \max\left\{\frac{\mu_r r \kappa_r \log n}{n}, \frac{\mu_r^2 r^2 \kappa_r^2}{n}\right\},$$

we have

$$C_2\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\frac{\mu_r r}{n}\sigma_1 \leqslant \sigma_i, \quad i = 1, \cdots, r.$$

In other words, $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top = \boldsymbol{M}$.

Similarly, by definition (2.6), in the case $\mathrm{rank}(\boldsymbol{M}) = r$, we have

$$\|\boldsymbol{M}\|_{\ell_\infty} = \frac{\widetilde{\mu}_r^2\,\mathrm{trace}(\boldsymbol{M})}{n} \leqslant \frac{\widetilde{\mu}_r^2 r \sigma_1}{n}.$$

Therefore, the approximation error bound (2.4) becomes

$$\left\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\right\|_F^2 \leqslant C_1 \sum_{i=1}^r \left\{\left[C_2\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\frac{\widetilde{\mu}_r^2 r}{n}\sigma_1 - \sigma_i\right]_+\right\}^2.$$

Therefore, if

$$p \geqslant 4C_2 \max\left\{\frac{\widetilde{\mu}_r^2 r \kappa_r \log n}{n}, \frac{\widetilde{\mu}_r^4 r^2 \kappa_r^2}{n}\right\},$$

we have $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top = \boldsymbol{M}$. □

A.3. Proof sketch of Theorem 2.4.1

Recall that the control of K_3 and K_4 in Lemma A.1.2 and Lemma A.1.3 will not be affected except replacing p by t. For K_2, Lemma 2.3.3 is not able to used anymore, therefore, by directly applying Lemma 2.3.5 to all the terms in (A.14), we have

$$K_2(\boldsymbol{X}) \leqslant \phi(t)\left[\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{U}\|_F^2 + 5\sum_{i=1}^n \|\boldsymbol{\Delta}_i\|_2^4 + 4\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2\right].$$

Therefore, we have

$$\begin{aligned}\sum_{i=2}^4 K_i(\boldsymbol{X}) \leqslant & 5\times 10^{-4}\left[\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right] + 10^5\phi^2(t)r\|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2 + \phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=1}^r \sigma_i \\ & + (4 + 1.6\times 10^9)\phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty}\|\boldsymbol{\Delta}\|_F^2 + 6\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{M}_{r+}\rangle.\end{aligned}$$

Similar to what we did in Section A.1.0.2, we have

$$\begin{aligned}&\sum_{i=2}^4 K_i(\boldsymbol{X}) \\ \leqslant & 10^{-3}\left[\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right] + 10^5\phi^2(t)r\|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2 + \phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=1}^r \sigma_i \\ & + 5\times 10^{-4}\sum_{i=1}^r\left\{-\sigma_i^4(\boldsymbol{\Delta}) + \left[4\times 10^{12}\phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty} - \sigma_{r+1-i}^2(\boldsymbol{U}) + 1.2\times 10^4\sigma_i(\boldsymbol{M}_{r+})\right]\sigma_i^2(\boldsymbol{\Delta})\right\} \\ \leqslant & 10^{-3}\left[\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right] + 10^5\phi^2(t)r\|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2 + \phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=1}^r \sigma_i \\ & + 1.25\times 10^{-4}\sum_{i=1}^r\left\{\left[4\times 10^{12}\phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty} + 1.2\times 10^4\sigma_{2r+1-i} - \sigma_i\right]_+\right\}^2 \\ \leqslant & 10^{-3}\left[\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right] + C_3\phi^2(t)r\|\boldsymbol{M}_{r+}\|_{\ell_\infty}^2 + \phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty}\sum_{i=1}^r \sigma_i \\ & + C_3\sum_{i=1}^r\left\{\left[C_2\phi(t)\|\boldsymbol{M}_r\|_{\ell_\infty} + C_2\sigma_{2r+1-i} - \sigma_i\right]_+\right\}^2.\end{aligned}$$

The last inequality uses the definition of C_2, C_3 in (A.12) and (A.13) as well as the fact that the constant $C_v \geqslant 1$. Replace ψ in (2.16) finishes the proof.

A.4. Proof sketch of Theorem 2.5.2

For any $\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}$, let

$$\boldsymbol{Z} \coloneqq \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix}.$$

Suppose $\boldsymbol{Z}^\top \boldsymbol{W}_r$ has SVD $\boldsymbol{Z}^\top \boldsymbol{W}_r = \boldsymbol{A}\boldsymbol{D}\boldsymbol{B}^\top$. Let $\boldsymbol{W} \coloneqq \boldsymbol{W}_r \boldsymbol{B}\boldsymbol{A}^\top, \boldsymbol{U} \coloneqq \boldsymbol{U}_r \boldsymbol{B}\boldsymbol{A}^\top, \boldsymbol{V} \coloneqq \boldsymbol{V}_r \boldsymbol{B}\boldsymbol{A}^\top$. Then $\boldsymbol{Z}^\top \boldsymbol{W} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top$ is a positive semidefinite matrix. It also holds that $\boldsymbol{W}_r \boldsymbol{W}_r^\top = \boldsymbol{W}\boldsymbol{W}^\top$. Similar to what we did in the PSD case, let

$$\boldsymbol{\Delta}_{\boldsymbol{Z}} = \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}} \\ \boldsymbol{\Delta}_{\boldsymbol{Y}} \end{bmatrix} \coloneqq \boldsymbol{Z} - \boldsymbol{W},$$

then we can consider the following auxiliary function:

$$K(\boldsymbol{X}, \boldsymbol{Y}) \coloneqq \operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{Z}})^\top \nabla^2 f(\boldsymbol{X}, \boldsymbol{Y}) \operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{Z}}) - 4\langle \nabla f(\boldsymbol{X}, \boldsymbol{Y}), \boldsymbol{\Delta}_{\boldsymbol{Z}} \rangle.$$

From an elegant lemma developed in [GJZ17], we are able to upper bound the above defined auxiliary function K. More precisely, we have

Lemma A.4.1 ([GJZ17, Lemma 16]). *For any $\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}$, let $\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{V}$, and $\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{\Delta}_{\boldsymbol{X}}, \boldsymbol{\Delta}_{\boldsymbol{Y}}$ be defined as above. Then the auxiliary function $K(\boldsymbol{X}, \boldsymbol{Y})$ can be upper bounded as following:*

$$K(\boldsymbol{X}, \boldsymbol{Y}) \leqslant K_1(\boldsymbol{X}, \boldsymbol{Y}) + K_2(\boldsymbol{X}, \boldsymbol{Y}) + K_3(\boldsymbol{X}, \boldsymbol{Y}) + K_4(\boldsymbol{X}, \boldsymbol{Y}),$$

where

$$\begin{aligned}
K_1(\boldsymbol{X},\boldsymbol{Y}) :=& \frac{1}{4}\left(\left\|\boldsymbol{\Delta}_{\boldsymbol{Z}}\boldsymbol{\Delta}_{\boldsymbol{Z}}^{\top}\right\|_F^2 - 3\left\|\boldsymbol{Z}\boldsymbol{Z}^{\top}-\boldsymbol{W}\boldsymbol{W}^{\top}\right\|_F^2\right),\\
K_2(\boldsymbol{X},\boldsymbol{Y}) :=& \left(\frac{1}{p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\right)\right\|_F^2 - \|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\|_F^2\right)\\
&- \left(\frac{3}{p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)\right\|_F^2 - 3\|\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\|_F^2\right),\\
K_3(\boldsymbol{X},\boldsymbol{Y}) :=& \lambda\left[\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{X}})^{\top}\nabla^2 G_\alpha(\boldsymbol{X})\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{X}}) - 4\left\langle\nabla G_\alpha(\boldsymbol{X}),\boldsymbol{\Delta}_{\boldsymbol{X}}\right\rangle\right]\\
&+ \lambda\left[\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{Y}})^{\top}\nabla^2 G_\alpha(\boldsymbol{Y})\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{Y}}) - 4\left\langle\nabla G_\alpha(\boldsymbol{Y}),\boldsymbol{\Delta}_{\boldsymbol{Y}}\right\rangle\right],\\
K_4(\boldsymbol{X},\boldsymbol{Y}) :=& \frac{6}{p}\left\langle\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top},\mathcal{P}_\Omega(\boldsymbol{M}_{r+})\right\rangle + \frac{4}{p}\left\langle\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}+\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^{\top},\mathcal{P}_\Omega(\boldsymbol{M}_{r+})\right\rangle.
\end{aligned}$$

Similar to the arguments we had in PSD case, without loss of generality, we can assume that $\boldsymbol{W} = \boldsymbol{W}_r$. Given the strong similarities between $K_1(\boldsymbol{X},\boldsymbol{Y}),\ldots,K_4(\boldsymbol{X},\boldsymbol{Y})$ and their counterparts, upper bounds in PSD case, we can derive corresponding upper bounds as following:

First, to get an upper bound of $K_2(\boldsymbol{X},\boldsymbol{Y})$, we need the following lemma.

Lemma A.4.2 ([**CR09**, Theorem 4.1]). *Let Ω be sampled according to Ber(p) model as defined in Model 2.5.1. Define*

$$\mathcal{T} := \{\boldsymbol{M}\in\mathbb{R}^{n_1\times n_2} \mid (\boldsymbol{I}-\boldsymbol{P}_{\mathcal{U}})\boldsymbol{M}(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{V}}) = \boldsymbol{0}\},$$

where $\mathcal{U},\mathcal{V}$ are fixed subspaces of $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$. Let $\mathcal{P}_{\mathcal{T}}$ be the Euclidean projection on to $\mathcal{T}$: For any matrix $\boldsymbol{M}\in\mathbb{R}^{n_1\times n_2}$,

$$\mathcal{P}_{\mathcal{T}}(\boldsymbol{M}) = \boldsymbol{P}_{\mathcal{U}}\boldsymbol{M} + \boldsymbol{M}\boldsymbol{P}_{\mathcal{V}} - \boldsymbol{P}_{\mathcal{U}}\boldsymbol{M}\boldsymbol{P}_{\mathcal{V}}.$$

Then there is an absolute constant C_{Ca}, if $p \geqslant C_{Ca}\frac{[\mu(\mathcal{U})\vee\mu(\mathcal{V})][\dim(\mathcal{U})\vee\dim(\mathcal{V})]\log(n_1\vee n_2)}{n_1\wedge n_2}$ with $\mu(\mathcal{U}),\mu(\mathcal{V})$ defined in (2.8), in an event E_{Ca} with probability $\mathbb{P}[E_{Ca}]\geqslant 1-(n_1+n_2)^{-11}$, we have

$$p^{-1}\|\mathcal{P}_{\mathcal{T}}\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}} - p\mathcal{P}_{\mathcal{T}}\| \leqslant 10^{-5}.$$

Equipped with Lemma A.4.2, and following the proof in controlling $K_2(\boldsymbol{X})$, we get the following lemma corresponding to Lemma A.1.1 in PSD case.

Lemma A.4.3. *In an event E_{Ca2} with probability $\mathbb{P}[E_{Ca2}] \geqslant 1-(n_1+n_2)^{-3}$, uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n_1\times r}, \boldsymbol{Y} \in \mathbb{R}^{n_2\times r}$ and corresponding $\boldsymbol{\Delta_X}, \boldsymbol{\Delta_Y}$ defined as before, we have*

$$K_2(\boldsymbol{X},\boldsymbol{Y}) \leqslant \frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\left[9.5\left(\sum_{i=1}^{n_1}\|(\boldsymbol{\Delta_X})_{i,\cdot}\|_2^4+\sum_{i=1}^{n_2}\|(\boldsymbol{\Delta_Y})_{i,\cdot}\|_2^4\right)+18\|\boldsymbol{W}_r\|_{2,\infty}^2\|\boldsymbol{\Delta_Z}\|_F^2\right]$$
$$+9\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\|\boldsymbol{W}_r\|_{2,\infty}^2\sum_{i=s+1}^{r}\sigma_i+3\times 10^{-4}\|\boldsymbol{W}\boldsymbol{\Delta_Z}^\top\|_F^2,$$

where s is defined by

$$s := \max\left\{s \leqslant r,\ \sigma_s \geqslant C_{Ca}\frac{\|\boldsymbol{M}_r\|_{\ell_\infty}\log(n_1\vee n_2)}{p}\right\}$$

with C_{Ca} an absolute constant defined in Lemma 2.3.3. Set $s=0$ if $\sigma_1 < C_{Ca}\frac{\|\boldsymbol{M}_r\|_{\ell_\infty}\log(n_1\vee n_2)}{p}$.

For $K_3(\boldsymbol{X},\boldsymbol{Y})$, we simply apply Lemma A.1.2 twice and have the following lemma.

Lemma A.4.4. *If $\alpha \geqslant 100\|\boldsymbol{W}_r\|_{2,\infty}$, then uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n_1\times r}, \boldsymbol{Y} \in \mathbb{R}^{n_2\times r}$ and corresponding $\boldsymbol{\Delta_X}, \boldsymbol{\Delta_Y}$ defined as before, we have*

$$K_3(\boldsymbol{X},\boldsymbol{Y}) \leqslant 200\lambda\alpha^2(\|\boldsymbol{\Delta_X}\|_F^2+\|\boldsymbol{\Delta_Y}\|_F^2)-0.3\lambda\left(\sum_{i=1}^{n_1}\|(\boldsymbol{\Delta_X})_{i,\cdot}\|_2^4+\sum_{i=1}^{n_2}\|(\boldsymbol{\Delta_Y})_{i,\cdot}\|_2^4\right).$$

Finally, by replacing $\boldsymbol{\Delta}$ with $\boldsymbol{\Delta_Z}$, we have the following control of $K_4(\boldsymbol{X},\boldsymbol{Y})$.

Lemma A.4.5. *Uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n_1\times r}, \boldsymbol{Y} \in \mathbb{R}^{n_2\times r}$ and corresponding $\boldsymbol{\Delta_X}, \boldsymbol{\Delta_Y}$ defined as before, we have*

$$K_4(\boldsymbol{X},\boldsymbol{Y}) \leqslant 5\times 10^{-4}\|\boldsymbol{\Delta_Z}\boldsymbol{\Delta_Z}^\top\|_F^2+2\times 10^{-4}\|\boldsymbol{W}\boldsymbol{\Delta_Z}^\top\|_F^2+10^5\frac{r\|\mathcal{P}_\Omega(\boldsymbol{M}_{r+})-p\boldsymbol{M}_{r+}\|^2}{p^2}$$
$$+6\langle\boldsymbol{\Delta_X}\boldsymbol{\Delta_Y}^\top,\boldsymbol{M}_{r+}\rangle.$$

Notice the fact that

$$\langle\boldsymbol{\Delta_X}\boldsymbol{\Delta_Y}^\top,\boldsymbol{M}_{r+}\rangle = \frac{1}{2}\langle\boldsymbol{\Delta_Z}\boldsymbol{\Delta_Z}^\top,\overline{\boldsymbol{M}}_{r+}\rangle$$

where

$$\overline{\boldsymbol{M}}_{r+} := \begin{bmatrix}\boldsymbol{0} & \boldsymbol{M}_{r+}\\ \boldsymbol{M}_{r+}^\top & \boldsymbol{0}\end{bmatrix}.$$

Moreover, $\overline{\boldsymbol{M}}_{r+}$ has following eigenvalue decomposition

$$\overline{\boldsymbol{M}}_{r+} = \frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{U}_{r+} & \boldsymbol{U}_{r+} \\ \boldsymbol{V}_{r+} & -\boldsymbol{V}_{r+} \end{bmatrix} \operatorname{diag}(\sigma_{r+1}, \ldots, \sigma_{n_1 \wedge n_2}, -\sigma_{r+1}, \ldots, -\sigma_{n_1 \wedge n_2}) \frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{U}_{r+} & \boldsymbol{U}_{r+} \\ \boldsymbol{V}_{r+} & -\boldsymbol{V}_{r+} \end{bmatrix}^\top$$

where

$$\boldsymbol{U}_{r+} \coloneqq [\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_{n_1 \wedge n_2}], \quad \boldsymbol{V}_{r+} \coloneqq [\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_{n_1 \wedge n_2}].$$

Therefore, by von Neumann's trace inequality Lemma A.1.4,

$$\langle \boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top, \boldsymbol{M}_{r+} \rangle = \frac{1}{2} \langle \boldsymbol{\Delta}_{\boldsymbol{Z}} \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top, \overline{\boldsymbol{M}}_{r+} \rangle \leqslant \frac{1}{2} \sum_{i=1}^{r} \sigma_i^2(\boldsymbol{\Delta}_{\boldsymbol{Z}}) \sigma_{r+i}.$$

Here we use the fact that $\boldsymbol{\Delta}_{\boldsymbol{Z}} \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top$ is a matrix with rank at most r. Therefore, with the following rectangular version of spectral lemma, we are able to copy all the proofs in the PSD case, which finishes the proof.

Lemma A.4.6 ([**Vu18**, **BVH16**]). *There is a constant $C_v > 0$ such that the following holds. If Ω is sampled according to the Ber(p) model with $p \geqslant C_v \frac{\log(n_1 \vee n_2)}{n_1 \wedge n_2}$, then in an event E_v with probability $\mathbb{P}[E_v] \geqslant 1 - (n_1 + n_2)^{-3}$,*

$$\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \leqslant C_v \sqrt{(n_1 \vee n_2) p}.$$

A.4.1. Proof of Lemma A.4.1.

Proof. Lemma A.4.1 is essentially [**GJZ17**, Lemma 16]. Here we give a sketch of the proof for the purpose of self-containedness.

First, denote $f_{\text{clean}}(\boldsymbol{X}, \boldsymbol{Y})$ as

$$\begin{aligned} f_{\text{clean}}(\boldsymbol{X}, \boldsymbol{Y}) = & \frac{1}{2p} \|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}_r)\|_F^2 + \frac{1}{8} \|\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y}\|_F^2 \\ & + \lambda(G_\alpha(\boldsymbol{X}) + G_\alpha(\boldsymbol{Y})). \end{aligned}$$

Comparing with (3.2), We can see

$$f(\boldsymbol{X}, \boldsymbol{Y}) = f_{\text{clean}}(\boldsymbol{X}, \boldsymbol{Y}) - \frac{1}{p} \langle \mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}_r), \mathcal{P}_\Omega(\boldsymbol{M}_{r+}) \rangle + \frac{1}{2p} \|\mathcal{P}_\Omega(\boldsymbol{M}_{r+})\|_F^2.$$

Therefore,

$$\langle \nabla f(\boldsymbol{X},\boldsymbol{Y}), [\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top \rangle = \langle \nabla f_{\text{clean}}(\boldsymbol{X},\boldsymbol{Y}), [\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top \rangle - \frac{1}{p}\langle \mathcal{P}_\Omega(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{Y}^\top + \boldsymbol{X}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top), \mathcal{P}_\Omega(\boldsymbol{M}_{r+})\rangle$$

and

$$\begin{aligned}
&\operatorname{vec}([\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top)^\top \nabla^2 f(\boldsymbol{X},\boldsymbol{Y}) \operatorname{vec}([\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top) \\
=&\operatorname{vec}([\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top)^\top \nabla^2 f_{\text{clean}}(\boldsymbol{X},\boldsymbol{Y}) \operatorname{vec}([\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top) - \frac{2}{p}\langle \mathcal{P}_\Omega(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top), \mathcal{P}_\Omega(\boldsymbol{M}_{r+})\rangle.
\end{aligned}$$

Therefore, we only need to concern about $f_{\text{clean}}(\boldsymbol{X},\boldsymbol{Y})$ now, which has already been discussed in [GJZ17]. Interested readers can refer to [GJZ17] for the detail.

By [GJZ17, Lemma 16], we have

$$\begin{aligned}
&\operatorname{vec}([\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top)^\top \nabla^2 f_{\text{clean}}(\boldsymbol{X},\boldsymbol{Y}) \operatorname{vec}([\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top) - 4\langle \nabla f_{\text{clean}}(\boldsymbol{X},\boldsymbol{Y}), [\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top \rangle \\
\leqslant&\frac{1}{4}\left\{ \left\| \boldsymbol{\Delta}_{\boldsymbol{Z}}\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \right\|_F^2 - 3\left\| \boldsymbol{Z}\boldsymbol{Z}^\top - \boldsymbol{W}\boldsymbol{W}^\top \right\|_F^2 \right\} + \left(\frac{1}{p}\left\| \mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \right) \right\|_F^2 - \|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\|_F^2 \right) \\
&- \left(\frac{3}{p}\left\| \mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top \right) \right\|_F^2 - 3\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top\|_F^2 \right) \\
&+ \lambda\left[\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{X}})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{X}}) - 4\left\langle \nabla G_\alpha(\boldsymbol{X}), \boldsymbol{\Delta}_{\boldsymbol{X}} \right\rangle \right] \\
&+ \lambda\left[\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{Y}})^\top \nabla^2 G_\alpha(\boldsymbol{Y}) \operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{Y}}) - 4\left\langle \nabla G_\alpha(\boldsymbol{Y}), \boldsymbol{\Delta}_{\boldsymbol{Y}} \right\rangle \right].
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\operatorname{vec}([\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top)^\top \nabla^2 f(\boldsymbol{X}, \boldsymbol{Y}) \operatorname{vec}([\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top) - 4\langle \nabla f(\boldsymbol{X}, \boldsymbol{Y}), [\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top \rangle \\
=& \operatorname{vec}([\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top)^\top \nabla^2 f_{\text{clean}}(\boldsymbol{X}, \boldsymbol{Y}) \operatorname{vec}([\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top) - \frac{2}{p}\langle \mathcal{P}_\Omega(\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top), \mathcal{P}_\Omega(\boldsymbol{M}_{r+})\rangle \\
&- 4\langle \nabla f_{\text{clean}}(\boldsymbol{X}, \boldsymbol{Y}), [\boldsymbol{\Delta}_{\boldsymbol{X}}^\top, \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top]^\top \rangle + \frac{4}{p}\langle \mathcal{P}_\Omega(\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{Y}^\top + \boldsymbol{X}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top), \mathcal{P}_\Omega(\boldsymbol{M}_{r+})\rangle \\
\leqslant& \frac{4}{p}\langle \mathcal{P}_\Omega(\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top), \mathcal{P}_\Omega(\boldsymbol{M}_{r+})\rangle + \frac{6}{p}\langle \mathcal{P}_\Omega(\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top), \mathcal{P}_\Omega(\boldsymbol{M}_{r+})\rangle \\
&+ \frac{1}{4}\left\{ \left\| \boldsymbol{\Delta}_{\boldsymbol{Z}} \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \right\|_F^2 - 3\left\| \boldsymbol{Z}\boldsymbol{Z}^\top - \boldsymbol{W}\boldsymbol{W}^\top \right\|_F^2 \right\} \\
&+ \left(\frac{1}{p} \left\| \mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \right) \right\|_F^2 - \|\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\|_F^2 \right) \\
&- \left(\frac{3}{p} \left\| \mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top \right) \right\|_F^2 - 3\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top\|_F^2 \right) \\
&+ \lambda \left[\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{X}})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{X}}) - 4\left\langle \nabla G_\alpha(\boldsymbol{X}), \boldsymbol{\Delta}_{\boldsymbol{X}} \right\rangle \right] \\
&+ \lambda \left[\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{Y}})^\top \nabla^2 G_\alpha(\boldsymbol{Y}) \operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{Y}}) - 4\left\langle \nabla G_\alpha(\boldsymbol{Y}), \boldsymbol{\Delta}_{\boldsymbol{Y}} \right\rangle \right].
\end{aligned}
$$

Combining with Lemma B.1.1 finishes the proof. □

APPENDIX B

Supporting Proofs of Chapter 3

B.1. Proof of Theorem 3.1.3

B.1.1. Auxiliary function. In order to study the properties of local minima of $\tilde{f}(\boldsymbol{\theta})$ as defined in (3.2), similar to what we did in Chapter 2, the first step in our proof of Theorem 3.1.3 is to derive the auxiliary function associated to $\tilde{f}(\boldsymbol{\theta})$. Again, given the smoothness of $\tilde{f}(\boldsymbol{\theta})$, any of its local minima $\hat{\boldsymbol{\xi}}$ satisfies $\nabla \tilde{f}(\hat{\boldsymbol{\xi}}) = \mathbf{0}$ and $\nabla^2 \tilde{f}(\hat{\boldsymbol{\xi}}) \succeq \mathbf{0}$. For any $\boldsymbol{\theta} \in \mathbb{R}^d$, suppose $\boldsymbol{\xi} \in \mathbb{R}^d$ be a vector satisfying (3.1.2) (Recall that there may be multiple vectors satisfying (3.1.2)). Choose $\boldsymbol{\delta_\theta}$ as $\boldsymbol{\delta_\theta} = \boldsymbol{\theta} - \boldsymbol{\xi}$. Therefore, we are now able to define the auxiliary function associated with $\tilde{f}$ as

$$K_{\tilde{f}}(\boldsymbol{\theta}) \coloneqq \boldsymbol{\delta_\theta}^\top \nabla^2 \tilde{f}(\boldsymbol{\theta}) \boldsymbol{\delta_\theta} - 4\boldsymbol{\delta_\theta}^\top \nabla \tilde{f}(\boldsymbol{\theta}). \tag{B.1}$$

For any local minimum $\hat{\boldsymbol{\xi}}$ of $\tilde{f}$, there also holds $K_{\tilde{f}}(\hat{\boldsymbol{\xi}}) \geqslant 0$.

Furthermore, due to the homogeneity and linearity of the parameterization $(\boldsymbol{X}(\boldsymbol{\theta}), \boldsymbol{Y}(\boldsymbol{\theta}))$, there is a strong connection between $K_{\tilde{f}}(\boldsymbol{\theta})$ and the corresponding K of f_{rect} defined in Chapter 2.

Lemma B.1.1. *For any $\boldsymbol{\theta} \in \mathbb{R}^d$ and its corresponding $\boldsymbol{\delta_\theta}$, there holds*

$$\begin{aligned} K_{\tilde{f}}(\boldsymbol{\theta}) = &\operatorname{vec}([\boldsymbol{X}(\boldsymbol{\delta_\theta})^\top, \boldsymbol{Y}(\boldsymbol{\delta_\theta})^\top]^\top)^\top \nabla^2 f_{rect}(\boldsymbol{X}(\boldsymbol{\theta}), \boldsymbol{Y}(\boldsymbol{\theta})) \operatorname{vec}([\boldsymbol{X}(\boldsymbol{\delta_\theta})^\top, \boldsymbol{Y}(\boldsymbol{\delta_\theta})^\top]^\top) \\ &- 4\langle \nabla f_{rect}(\boldsymbol{X}(\boldsymbol{\theta}), \boldsymbol{Y}(\boldsymbol{\theta})), [\boldsymbol{X}(\boldsymbol{\delta_\theta})^\top, \boldsymbol{Y}(\boldsymbol{\delta_\theta})^\top]^\top \rangle. \end{aligned} \tag{B.2}$$

Proof. Assumption 3.1.1 implies that both $\boldsymbol{X}(\boldsymbol{\theta})$ and $\boldsymbol{Y}(\boldsymbol{\theta})$ are homogeneous linear functions, so

$$\tilde{f}(\boldsymbol{\theta} + \boldsymbol{\delta_\theta}) = f_{\text{rect}}(\boldsymbol{X}(\boldsymbol{\theta} + \boldsymbol{\delta_\theta}), \boldsymbol{Y}(\boldsymbol{\theta} + \boldsymbol{\delta_\theta})) = f_{\text{rect}}(\boldsymbol{X}(\boldsymbol{\theta}) + \boldsymbol{X}(\boldsymbol{\delta_\theta}), \boldsymbol{Y}(\boldsymbol{\theta}) + \boldsymbol{Y}(\boldsymbol{\delta_\theta})). \tag{B.3}$$

Due to the linear homogeneity of $\boldsymbol{X}(\boldsymbol{\theta})$ and $\boldsymbol{Y}(\boldsymbol{\theta})$ once more, also by considering the Taylor expansions of both sides in (B.3) at $\boldsymbol{\theta}$, we get

$$\boldsymbol{\delta}_{\boldsymbol{\theta}}^\top \nabla \tilde{f}(\boldsymbol{\theta}) = \mathrm{vec}([\boldsymbol{X}(\boldsymbol{\delta}_{\boldsymbol{\theta}})^\top, \boldsymbol{Y}(\boldsymbol{\delta}_{\boldsymbol{\theta}})^\top]^\top)^\top \nabla f_{\mathrm{rect}}(\boldsymbol{X}(\boldsymbol{\theta}), \boldsymbol{Y}(\boldsymbol{\theta})) \tag{B.4}$$

and

$$\boldsymbol{\delta}_{\boldsymbol{\theta}}^\top \nabla^2 \tilde{f}(\boldsymbol{\theta}) \boldsymbol{\delta}_{\boldsymbol{\theta}} = \mathrm{vec}([\boldsymbol{X}(\boldsymbol{\delta}_{\boldsymbol{\theta}})^\top, \boldsymbol{Y}(\boldsymbol{\delta}_{\boldsymbol{\theta}})^\top]^\top)^\top \nabla^2 f_{\mathrm{rect}}(\boldsymbol{X}(\boldsymbol{\theta}), \boldsymbol{Y}(\boldsymbol{\theta}))\, \mathrm{vec}([\boldsymbol{X}(\boldsymbol{\delta}_{\boldsymbol{\theta}})^\top, \boldsymbol{Y}(\boldsymbol{\delta}_{\boldsymbol{\theta}})^\top]^\top) \tag{B.5}$$

The equality (B.2) is obtained through combining (B.4) and (B.5). □

For notation simplicity, we introduce the following abbreviations:

$$\begin{cases} \boldsymbol{X} = \boldsymbol{X}(\boldsymbol{\theta}) \in \mathbb{R}^{n_1 \times r} \\ \boldsymbol{U} = \boldsymbol{X}(\boldsymbol{\xi}) \in \mathbb{R}^{n_1 \times r} \\ \boldsymbol{\Delta}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{\delta}_{\boldsymbol{\theta}}) = \boldsymbol{X} - \boldsymbol{U} \in \mathbb{R}^{n_1 \times r} \\ \boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{\theta}) \in \mathbb{R}^{n_2 \times r} \\ \boldsymbol{V} = \boldsymbol{Y}(\boldsymbol{\xi}) \in \mathbb{R}^{n_2 \times r} \\ \boldsymbol{\Delta}_{\boldsymbol{Y}} = \boldsymbol{Y}(\boldsymbol{\delta}_{\boldsymbol{\theta}}) = \boldsymbol{Y} - \boldsymbol{V} \in \mathbb{R}^{n_2 \times r}. \end{cases} \tag{B.6}$$

In the remaining part of the proof, $\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{\Delta}_{\boldsymbol{X}}, \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{\Delta}_{\boldsymbol{Y}}$ will refer to the matrices defined in (B.6) if not specified. Then (in)equalities in (3.5) are thus abbreviated into

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{V}^\top,\ \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{V}^\top\boldsymbol{V},\ \text{and } \boldsymbol{X}^\top\boldsymbol{U} + \boldsymbol{Y}^\top\boldsymbol{V} \succeq \boldsymbol{0}. \tag{B.7}$$

By applying Lemma B.1.1, analogs to Lemma A.4.1, we are able to upper bound $K_{\tilde{f}}(\boldsymbol{\theta})$ as following.

Lemma B.1.2. *For any $\boldsymbol{\theta} \in \mathbb{R}^d$, with $\boldsymbol{\delta}_{\boldsymbol{\theta}} = \boldsymbol{\theta} - \boldsymbol{\xi}$ where $\boldsymbol{\xi}$ satisfies the conditions in* (3.5), *and $\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{\Delta}_{\boldsymbol{X}}, \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{\Delta}_{\boldsymbol{Y}}$ defined as in* (B.6), *denote*

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix},\ \boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix},\ \textit{and } \boldsymbol{\Delta}_{\boldsymbol{Z}} = \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}} \\ \boldsymbol{\Delta}_{\boldsymbol{Y}} \end{bmatrix} = \boldsymbol{Z} - \boldsymbol{W}.$$

Then the auxiliary function $K_{\tilde{f}}(\boldsymbol{\theta})$ defined in (B.1) can be upper bounded as following:

$$K_{\tilde{f}}(\boldsymbol{\theta}) \leqslant K_1(\boldsymbol{\theta}) + K_2(\boldsymbol{\theta}) + K_3(\boldsymbol{\theta}) + K_4(\boldsymbol{\theta}), \tag{B.8}$$

where

$$\begin{aligned} K_1(\boldsymbol{\theta}) :=& \frac{1}{4}\left(\left\|\boldsymbol{\Delta_Z}\boldsymbol{\Delta_Z}^\top\right\|_F^2 - 3\left\|\boldsymbol{Z}\boldsymbol{Z}^\top - \boldsymbol{W}\boldsymbol{W}^\top\right\|_F^2\right), \\ K_2(\boldsymbol{\theta}) :=& \left(\frac{1}{p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{\Delta_X}\boldsymbol{\Delta_Y}^\top\right)\right\|_F^2 - \|\boldsymbol{\Delta_X}\boldsymbol{\Delta_Y}^\top\|_F^2\right) \\ & - \left(\frac{3}{p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right\|_F^2 - 3\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top\|_F^2\right), \\ K_3(\boldsymbol{\theta}) :=& \lambda\left[\operatorname{vec}(\boldsymbol{\Delta_X})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{\Delta_X}) - 4\left\langle \nabla G_\alpha(\boldsymbol{X}), \boldsymbol{\Delta_X}\right\rangle\right] \\ & + \lambda\left[\operatorname{vec}(\boldsymbol{\Delta_Y})^\top \nabla^2 G_\alpha(\boldsymbol{Y}) \operatorname{vec}(\boldsymbol{\Delta_Y}) - 4\left\langle \nabla G_\alpha(\boldsymbol{Y}), \boldsymbol{\Delta_Y}\right\rangle\right], \\ K_4(\boldsymbol{\theta}) :=& \frac{6}{p}\left\langle \boldsymbol{\Delta_X}\boldsymbol{\Delta_Y}^\top, \mathcal{P}_\Omega(\boldsymbol{N})\right\rangle + \frac{4}{p}\left\langle \boldsymbol{U}\boldsymbol{\Delta_Y}^\top + \boldsymbol{\Delta_X}\boldsymbol{V}^\top, \mathcal{P}_\Omega(\boldsymbol{N})\right\rangle. \end{aligned} \tag{B.9}$$

Comparing to Lemma A.4.1, the only difference is replacing $\boldsymbol{M}_{r+}$ by $\boldsymbol{N}$. Therefore, the proof is omitted here.

B.1.2. Controlling the auxiliary function. This section is meant to control $K_2(\boldsymbol{\theta})$ and $K_3(\boldsymbol{\theta})$, which will further give a bound of right hand side of (B.8).

Before proceed, we here first collect some useful properties of $\boldsymbol{U} = \boldsymbol{X}(\boldsymbol{\xi})$ and $\boldsymbol{V} = \boldsymbol{Y}(\boldsymbol{\xi})$. The proof is left to Section B.2.1.

PROPOSITION B.1.3. *For any $\boldsymbol{\theta}$, the matrices $\boldsymbol{U} = \boldsymbol{X}(\boldsymbol{\xi})$ and $\boldsymbol{V} = \boldsymbol{Y}(\boldsymbol{\xi})$ defined in (B.6) satisfy the following basic properties:*

- $\operatorname{colspan}(\boldsymbol{U}) = \operatorname{colspan}([\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r])$ *and* $\operatorname{colspan}(\boldsymbol{V}) = \operatorname{colspan}([\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r])$*;*
- *The largest singular values of both $\boldsymbol{U}$ and $\boldsymbol{V}$ are $\sqrt{\sigma_1}$;*
- *The r-th singular values of both $\boldsymbol{U}$ and $\boldsymbol{V}$ are $\sqrt{\sigma_r}$.*
- $\|\boldsymbol{U}\|_{2,\infty}^2 \leqslant \frac{\mu r}{n_1}\sigma_1$ *and* $\|\boldsymbol{V}\|_{2,\infty}^2 \leqslant \frac{\mu r}{n_2}\sigma_1$.

B.1.2.1. *Control of $K_2(\boldsymbol{\theta})$.* In this section, we give a control of $K_2(\boldsymbol{\theta})$. Comparing to upper bounding $K_2(\boldsymbol{X}, \boldsymbol{Y})$ as in Section 2.5, here we assume the sampling rate p is sufficiently large, therefore, Lemma A.4.2 can be applied directly here.

By the way we define $\boldsymbol{\Delta}_{\boldsymbol{X}}, \boldsymbol{\Delta}_{\boldsymbol{Y}}$ in (B.6),

$$\begin{aligned}\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top =&(\boldsymbol{U} + \boldsymbol{\Delta}_{\boldsymbol{X}})(\boldsymbol{V} + \boldsymbol{\Delta}_{\boldsymbol{Y}})^\top - \boldsymbol{U}\boldsymbol{V}^\top \\ =&\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top + \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top.\end{aligned}$$

Therefore,

$$\begin{aligned}&\left|\frac{1}{p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right\|_F^2 - \|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top\|_F^2\right| \\ =&\left|\frac{1}{p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top + \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right)\right\|_F^2 - \left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top + \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\|_F^2\right| \\ \leqslant&\underbrace{\left|\frac{1}{p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right)\right\|_F^2 - \left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\|_F^2\right|}_{\textcircled{1}} \\ &+\underbrace{\left|\frac{1}{p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right)\right\|_F^2 - \|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\|_F^2\right|}_{\textcircled{2}} \\ &+\underbrace{\left|\frac{2}{p}\left\langle\mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right)\right\rangle - 2\left\langle\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top, \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\rangle\right|}_{\textcircled{3}} \\ &+\underbrace{\left|\frac{2}{p}\left\langle\mathcal{P}_\Omega\left(\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right)\right\rangle - 2\left\langle\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top, \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\rangle\right|}_{\textcircled{4}}.\end{aligned}$$

By Proposition B.1.3, the matrix $\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top$ belongs to the subspace $\mathcal{T}$ defined in Lemma A.4.2. Therefore, by Lemma A.4.2, in an event E_{Ca3} with probability $\mathbb{P}[E_{Ca3}] \geqslant 1 - (n_1 + n_2)^{-3}$, there holds

$$\textcircled{1} \leqslant 0.0001\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\|_F^2 \leqslant 0.0002\left(\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top\right\|_F^2 + \left\|\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\|_F^2\right).$$

By Lemma 2.3.5, we have

$$\textcircled{2} \leqslant \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{2p}\left(\sum_{k=1}^{n_1}\|(\boldsymbol{\Delta}_{\boldsymbol{X}})_{k,\cdot}\|_2^4 + \sum_{k=1}^{n_2}\|(\boldsymbol{\Delta}_{\boldsymbol{Y}})_{k,\cdot}\|_2^4\right),$$

$$\textcircled{3} \leqslant \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p} \left(\sum_{k=1}^{n_1} \|(\boldsymbol{\Delta}_{\boldsymbol{X}})_{k,\cdot}\|_2^4 + \sum_{k=1}^{n_2} \|\boldsymbol{V}_{k,\cdot}\|_2^2 \|(\boldsymbol{\Delta}_{\boldsymbol{Y}})_{k,\cdot}\|_2^2 \right),$$

and

$$\textcircled{4} \leqslant \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p} \left(\sum_{k=1}^{n_2} \|(\boldsymbol{\Delta}_{\boldsymbol{Y}})_{k,\cdot}\|_2^4 + \sum_{k=1}^{n_1} \|\boldsymbol{U}_{k,\cdot}\|_2^2 \|(\boldsymbol{\Delta}_{\boldsymbol{X}})_{k,\cdot}\|_2^2 \right).$$

By Proposition B.1.3, $\|\boldsymbol{U}\|_{2,\infty}^2 \leqslant \frac{\mu r}{n_1}\sigma_1$ and $\|\boldsymbol{V}\|_{2,\infty}^2 \leqslant \frac{\mu r}{n_2}\sigma_1$. Then

$$\textcircled{3} \leqslant \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p} \left(\sum_{k=1}^{n_1} \|(\boldsymbol{\Delta}_{\boldsymbol{X}})_{k,\cdot}\|_2^4 + \frac{\mu r}{n_2}\sigma_1 \|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_F^2 \right),$$

and

$$\textcircled{4} \leqslant \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p} \left(\sum_{k=1}^{n_2} \|(\boldsymbol{\Delta}_{\boldsymbol{Y}})_{k,\cdot}\|_2^4 + \frac{\mu r}{n_1}\sigma_1 \|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_F^2 \right).$$

Combining the above inequalities together, we have

$$\begin{aligned}
K_2 \leqslant & \left| \frac{1}{p} \left\| \mathcal{P}_\Omega \left(\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \right) \right\|_F^2 - \|\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\|_F^2 \right| \\
& + \left| \frac{3}{p} \left\| \mathcal{P}_\Omega \left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top \right) \right\|_F^2 - 3\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top\|_F^2 \right| \\
\leqslant & \textcircled{2} + 3\left(\textcircled{1} + \textcircled{2} + \textcircled{3} + \textcircled{4} \right) \\
\leqslant & 0.0006 \left(\left\| \boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{V}^\top \right\|_F^2 + \left\| \boldsymbol{U} \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \right\|_F^2 \right) \\
& + \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p} \left(5\sum_{k=1}^{n_1} \|(\boldsymbol{\Delta}_{\boldsymbol{X}})_{k,\cdot}\|_2^4 + 5\sum_{k=1}^{n_2} \|(\boldsymbol{\Delta}_{\boldsymbol{Y}})_{k,\cdot}\|_2^4 \right) \\
& + \frac{\|\boldsymbol{\Omega} - p\boldsymbol{J}\|}{p} \left(3\frac{\mu r}{n_1}\sigma_1 \|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_F^2 + 3\frac{\mu r}{n_2}\sigma_1 \|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_F^2 \right).
\end{aligned}$$

B.1.2.2. *Control of* $K_3(\boldsymbol{\theta})$. For $K_3(\boldsymbol{\theta})$, when $\alpha \geqslant 100\sqrt{\frac{\mu r \sigma_1}{n_1 \wedge n_2}} \geqslant 100\|\boldsymbol{W}\|_{2,\infty}$, by applying Lemma A.1.2 twice, we have

$$K_3(\boldsymbol{\theta}) \leqslant 200\lambda\alpha^2(\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_F^2 + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_F^2) - 0.3\lambda \left(\sum_{k=1}^{n_1} \|(\boldsymbol{\Delta}_{\boldsymbol{X}})_{k,\cdot}\|_2^4 + \sum_{k=1}^{n_2} \|(\boldsymbol{\Delta}_{\boldsymbol{Y}})_{k,\cdot}\|_2^4 \right).$$

B.1.2.3. *Putting $K_2(\boldsymbol{\theta})$ and $K_3(\boldsymbol{\theta})$ together.* Combining the above upper bounds of $K_2(\boldsymbol{\theta})$ and $K_3(\boldsymbol{\theta})$ together, there holds

$$\begin{aligned} K_2+K_3 \leqslant & 0.0006\left(\left\|\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{V}^{\top}\right\|_F^2+\left\|\boldsymbol{U} \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\right\|_F^2\right) \\ & +\left(\frac{3 \mu r \sigma_1\|\boldsymbol{\Omega}-p \boldsymbol{J}\|}{p\left(n_1 \wedge n_2\right)}+200 \lambda \alpha^2\right)\left(\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_F^2\right) \\ & +\left(\frac{5\|\boldsymbol{\Omega}-p \boldsymbol{J}\|}{p}-0.3 \lambda\right)\left(\sum_{k=1}^{n_1}\|(\boldsymbol{\Delta}_{\boldsymbol{X}})_{k, \cdot}\|_2^4+\sum_{k=1}^{n_2}\|(\boldsymbol{\Delta}_{\boldsymbol{Y}})_{k, \cdot}\|_2^4\right) . \end{aligned}$$

By Lemma A.4.6, when $p \geqslant C_v \frac{\log(n_1 \vee n_2)}{n_1 \wedge n_2}$, in an event E_{v3} with probability $\mathbb{P}[E_{v3}] \geqslant 1-(n_1+n_2)^{-3}$, $\|\boldsymbol{\Omega}-p\boldsymbol{J}\| \leqslant C_v\sqrt{(n_1 \vee n_2)p}$. Therefore, combining with assumptions on p, α and λ in (3.6),

$$\begin{aligned} & \frac{3 \mu r \sigma_1\|\boldsymbol{\Omega}-p \boldsymbol{J}\|}{(n_1 \wedge n_2) p}+200 \lambda \alpha^2 \\ \leqslant & \frac{3 \mu r \sigma_1 C_v \sqrt{(n_1 \vee n_2) p}}{(n_1 \wedge n_2) p}+200 \times 10^3 C_5^3 \sqrt{\frac{n_1 \vee n_2}{p}} \frac{\mu r \sigma_1}{n_1 \wedge n_2} \\ \leqslant & \frac{3 \mu r \sigma_1 C_v \sqrt{(n_1 \vee n_2)}}{(n_1 \wedge n_2) \sqrt{C_4 \frac{n_1 \vee n_2}{(n_1 \wedge n_2)^2} \mu^2 r^2 \kappa^2}}+200 \times 10^3 C_5^3 \sqrt{\frac{n_1 \vee n_2}{C_4 \frac{n_1 \vee n_2}{(n_1 \wedge n_2)^2} \mu^2 r^2 \kappa^2}} \frac{\mu r \sigma_1}{n_1 \wedge n_2} \\ = & \left(\frac{3 C_v}{\sqrt{C_4}}+\frac{2 \times 10^5 C_5^3}{\sqrt{C_4}}\right) \sigma_r . \end{aligned}$$

And we also have

$$\frac{5\|\boldsymbol{\Omega}-p \boldsymbol{J}\|}{p}-0.3 \lambda \leqslant \frac{5 \sqrt{n_1 \vee n_2}}{\sqrt{p}}-0.3 C_5 \sqrt{\frac{n_1 \vee n_2}{p}}=(5-0.3 C_5) \sqrt{\frac{n_1 \vee n_2}{p}}$$

Therefore, by choosing $C_5=20$, $C_4=(3C_v+1.6\times 10^9)^2/0.0004^2$, we have

$$\begin{aligned} K_2+K_3 \leqslant & 0.0006\left(\left\|\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{V}^{\top}\right\|_F^2+\left\|\boldsymbol{U} \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\right\|_F^2\right)+0.0004 \sigma_r\left(\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_F^2\right) \\ & +(5-6) \sqrt{\frac{n_1 \vee n_2}{p}}\left(\sum_{k=1}^{n_1}\|(\boldsymbol{\Delta}_{\boldsymbol{X}})_{k, \cdot}\|_2^4+\sum_{k=1}^{n_2}\|(\boldsymbol{\Delta}_{\boldsymbol{Y}})_{k, \cdot}\|_2^4\right) \\ \leqslant & 0.0006\left(\left\|\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{V}^{\top}\right\|_F^2+\left\|\boldsymbol{U} \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\right\|_F^2\right)+0.0004 \sigma_r\left(\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_F^2\right) . \end{aligned}$$

By Proposition B.1.3, there holds

$$\left\|\boldsymbol{U} \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\right\|_F^2 \geqslant \sigma_r^2(\boldsymbol{U})\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\right\|_F^2=\sigma_r\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\right\|_F^2$$

and

$$\left\|\boldsymbol{V}\boldsymbol{\Delta}_{\boldsymbol{X}}^\top\right\|_F^2 \geqslant \sigma_r^2(\boldsymbol{V})\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|_F^2 = \sigma_r\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|_F^2.$$

Therefore

$$K_2 + K_3 \leqslant 0.001\left(\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top\right\|_F^2 + \left\|\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\|_F^2\right). \tag{B.10}$$

B.1.2.4. *Upper bounding right hand side of* (B.8). First of all, we rewrite $K_1(\boldsymbol{\theta})$ in terms of $\boldsymbol{W}$ and $\boldsymbol{\Delta}_{\boldsymbol{Z}}$. Recall

$$\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix}, \ \boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}, \text{ and } \boldsymbol{\Delta}_{\boldsymbol{Z}} = \begin{bmatrix}\boldsymbol{\Delta}_{\boldsymbol{X}}\\ \boldsymbol{\Delta}_{\boldsymbol{Y}}\end{bmatrix} = \boldsymbol{Z} - \boldsymbol{W}.$$

Therefore, we have

$$\begin{aligned}&\left\|\boldsymbol{Z}\boldsymbol{Z}^\top - \boldsymbol{W}\boldsymbol{W}^\top\right\|_F^2\\ =&\left\|\boldsymbol{\Delta}_{\boldsymbol{Z}}\boldsymbol{W}^\top + \boldsymbol{W}\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top + \boldsymbol{\Delta}_{\boldsymbol{Z}}\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\right\|_F^2\\ =&\|\boldsymbol{\Delta}_{\boldsymbol{Z}}\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\|_F^2 + 2\|\boldsymbol{\Delta}_{\boldsymbol{Z}}\boldsymbol{W}^\top\|_F^2 + 2\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}\boldsymbol{W}^\top, \boldsymbol{W}\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\rangle + 4\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}\boldsymbol{W}^\top, \boldsymbol{\Delta}_{\boldsymbol{Z}}\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\rangle\\ =&\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F^2 + 2\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{W}^\top\boldsymbol{W}\rangle + 2\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}, \boldsymbol{W}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\rangle + 4\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\rangle.\end{aligned} \tag{B.11}$$

Here we use the fact that $\langle\boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{trace}(\boldsymbol{A}^\top\boldsymbol{B})$ and trace is invariant under cyclic permutations. By recalling the definition of $K_1(\boldsymbol{\theta})$ in (B.9), (B.11) implies that

$$\begin{aligned}K_1(\boldsymbol{\theta}) = &-\frac{1}{2}\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F^2 - \frac{3}{2}\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{W}^\top\boldsymbol{W}\rangle\\ &-\frac{3}{2}\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}, \boldsymbol{W}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\rangle - 3\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\rangle.\end{aligned} \tag{B.12}$$

Condition (B.7) implies

$$\boldsymbol{Z}^\top\boldsymbol{W} = \boldsymbol{X}^\top\boldsymbol{U} + \boldsymbol{Y}^\top\boldsymbol{V} \succeq \boldsymbol{0}. \tag{B.13}$$

This further implies that $\boldsymbol{W}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}} = \boldsymbol{W}^\top\boldsymbol{Z} - \boldsymbol{Z}^\top\boldsymbol{Z}$ is symmetric (this is a crucial step for the analysis in [JGN+17] and [GJZ17]). This implies that

$$\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}, \boldsymbol{W}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\rangle = \|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\|_F^2. \tag{B.14}$$

Combining (B.14) with (B.12) we have

$$K_1(\boldsymbol{\theta}) = -0.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F^2 - 1.5\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{W}^\top\boldsymbol{W}\rangle - 1.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\|_F^2 - 3\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\rangle.$$

Therefore, based on (B.10), we are able to upper bound the right hand side of (B.8) as following:

$$\begin{aligned}
&K_1(\boldsymbol{\theta}) + K_2(\boldsymbol{\theta}) + K_3(\boldsymbol{\theta}) + K_4(\boldsymbol{\theta})\\
\leqslant& -0.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F^2 - 1.5\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{W}^\top\boldsymbol{W}\rangle - 1.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\|_F^2 - 3\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\rangle\\
&+ 0.001\left(\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top\right\|_F^2 + \left\|\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\|_F^2\right) + |K_4(\boldsymbol{\theta})|.
\end{aligned}$$

Furthermore, there holds

$$\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{W}^\top\boldsymbol{W}\rangle = \mathrm{trace}(\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\boldsymbol{W}^\top\boldsymbol{W}) = \|\boldsymbol{W}\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\|_F^2 \geqslant \left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{V}^\top\right\|_F^2 + \left\|\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\|_F^2. \tag{B.15}$$

Based on (B.15), we further have

$$\begin{aligned}
&K_1(\boldsymbol{\theta}) + K_2(\boldsymbol{\theta}) + K_3(\boldsymbol{\theta}) + K_4(\boldsymbol{\theta})\\
\leqslant& -0.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F^2 - 1.499\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{W}^\top\boldsymbol{W}\rangle - 1.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\|_F^2\\
&- 3\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\rangle + |K_4(\boldsymbol{\theta})|.
\end{aligned}\tag{B.16}$$

The fact $\boldsymbol{Z}^\top\boldsymbol{W} \succeq \boldsymbol{0}$ from (B.13) further implies that

$$\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{W}^\top\boldsymbol{W}\rangle + \langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\rangle = \langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{Z}^\top\boldsymbol{W}\rangle \geqslant 0,$$

in which we use the fact that the inner product of two PSD matrices is nonnegative. Then

$$\begin{aligned}
&K_1(\boldsymbol{\theta}) + K_2(\boldsymbol{\theta}) + K_3(\boldsymbol{\theta}) + K_4(\boldsymbol{\theta})\\
\leqslant& -0.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F^2 - 1.499\left(\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{W}^\top\boldsymbol{W}\rangle + \langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\rangle\right)\\
&- 1.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\|_F^2 - 1.501\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\rangle + |K_4(\boldsymbol{\theta})|\\
\leqslant& -0.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F^2 - 1.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\|_F^2 - 1.501\langle\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\rangle + |K_4(\boldsymbol{\theta})|\\
\leqslant& -0.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F^2 - 1.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\|_F^2 + 1.501\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\boldsymbol{W}\|_F + |K_4(\boldsymbol{\theta})|.
\end{aligned}\tag{B.17}$$

B.1.3. Completing the proof of Theorem 3.1.3. Recall that if $\hat{\boldsymbol{\xi}}$ is a local minimum of $\tilde{f}$, there holds $K_{\tilde{f}}(\hat{\boldsymbol{\xi}}) \geqslant 0$. By Lemma A.4.1, there holds

$$K_1(\boldsymbol{\theta}) + K_2(\boldsymbol{\theta}) + K_3(\boldsymbol{\theta}) + K_4(\boldsymbol{\theta}) \geqslant 0. \tag{B.18}$$

Then (B.17) implies

$$0.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F^2 + 1.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{W}\|_F^2 - 1.501\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F \|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{W}\|_F \leqslant |K_4(\boldsymbol{\theta})|,$$

which gives

$$\|\boldsymbol{\Delta}_{\boldsymbol{Z}} \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\|_F = \|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F \leqslant 3\sqrt{|K_4(\boldsymbol{\theta})|}, \quad \|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{W}\|_F \leqslant 2\sqrt{|K_4(\boldsymbol{\theta})|}. \tag{B.19}$$

By (B.16) as well as (B.18), we have

$$\begin{aligned}
&1.499\langle \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{W}^\top \boldsymbol{W}\rangle \\
\leqslant& -0.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F^2 - 1.5\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{W}\|_F^2 - 3\langle \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{\Delta}_{\boldsymbol{Z}}, \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{W}\rangle + |K_4(\boldsymbol{\theta})| \\
\leqslant& 3\|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{\Delta}_{\boldsymbol{Z}}\|_F \|\boldsymbol{\Delta}_{\boldsymbol{Z}}^\top \boldsymbol{W}\|_F + |K_4(\boldsymbol{\theta})| \\
\leqslant& 19|K_4(\boldsymbol{\theta})|.
\end{aligned}$$

Combining with (B.15) we have

$$\left\|\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{V}^\top\right\|_F^2 + \left\|\boldsymbol{U} \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\|_F^2 \leqslant 13|K_4(\boldsymbol{\theta})|. \tag{B.20}$$

By (B.10) and (B.20), $K_2(\boldsymbol{\theta}) + K_3(\boldsymbol{\theta}) + K_4(\boldsymbol{\theta}) \leqslant 2|K_4(\boldsymbol{\theta})|$. By (B.18) and the definition of $K_1(\boldsymbol{\theta})$ in (B.9),

$$\frac{3}{4}\|\boldsymbol{Z}\boldsymbol{Z}^\top - \boldsymbol{W}\boldsymbol{W}^\top\|_F^2 \leqslant \frac{1}{4}\|\boldsymbol{\Delta}_{\boldsymbol{Z}} \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\|_F^2 + K_2(\boldsymbol{\theta}) + K_3(\boldsymbol{\theta}) + K_4(\boldsymbol{\theta}) \leqslant \frac{17}{4}|K_4(\boldsymbol{\theta})|.$$

The last inequality also use (B.19). Therefore, we are able to upper bound $\|\boldsymbol{Z}\boldsymbol{Z}^\top - \boldsymbol{W}\boldsymbol{W}^\top\|_F^2$ in terms of $|K_4(\boldsymbol{\theta})|$. The only thing left over is to upper bound $|K_4(\boldsymbol{\theta})|$. Recall the fact that $\boldsymbol{\Delta}_{\boldsymbol{Z}} = \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}} \\ \boldsymbol{\Delta}_{\boldsymbol{Y}} \end{bmatrix}$, then $\|\boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\|_F \leqslant \|\boldsymbol{\Delta}_{\boldsymbol{Z}} \boldsymbol{\Delta}_{\boldsymbol{Z}}^\top\|_F$. Therefore, by (B.19) and (B.20), and the definition of $K_4(\boldsymbol{\theta})$ in (B.9), we have

$$\begin{aligned}
&|K_4(\boldsymbol{\theta})| \\
&\leqslant \frac{6}{p}\|\boldsymbol{\Delta_X}\boldsymbol{\Delta_Y}^\top\|_F\|\boldsymbol{P_{\Delta_X}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_{\Delta_Y}}\|_F + \frac{4}{p}\|\boldsymbol{U}\boldsymbol{\Delta_Y}^\top\|_F\|\boldsymbol{P_U}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_{\Delta_Y}}\|_F \\
&\quad + \frac{4}{p}\|\boldsymbol{\Delta_X}\boldsymbol{V}^\top\|_F\|\boldsymbol{P_{\Delta_X}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_V}\|_F \\
&\leqslant \frac{100\sqrt{|K_4(\boldsymbol{\theta})|}}{p}\max\{\|\boldsymbol{P_{\Delta_X}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_{\Delta_Y}}\|_F, \|\boldsymbol{P_U}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_{\Delta_Y}}\|_F, \|\boldsymbol{P_{\Delta_X}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_V}\|_F\}.
\end{aligned} \tag{B.21}$$

Due to the fact that $\boldsymbol{U}, \boldsymbol{\Delta_X} \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{V}, \boldsymbol{\Delta_Y} \in \mathbb{R}^{n_2\times r}$, we can see that $\boldsymbol{P_{\Delta_X}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_{\Delta_Y}}$, $\boldsymbol{P_U}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_{\Delta_Y}}$ and $\boldsymbol{P_{\Delta_X}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_V}$ are matrices with rank at most r. Therefore,

$$\begin{aligned}
&\max\{\|\boldsymbol{P_{\Delta_X}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_{\Delta_Y}}\|_F, \|\boldsymbol{P_U}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_{\Delta_Y}}\|_F, \|\boldsymbol{P_{\Delta_X}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_V}\|_F\} \\
&\leqslant \sqrt{r}\max\{\|\boldsymbol{P_{\Delta_X}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_{\Delta_Y}}\|, \|\boldsymbol{P_U}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_{\Delta_Y}}\|, \|\boldsymbol{P_{\Delta_X}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P_V}\|\} \\
&= \sqrt{r}\varphi.
\end{aligned}$$

Where the last line follows from (B.6) and (3.7). Therefore, (B.21) gives

$$|K_4(\boldsymbol{\theta})| \leqslant \frac{100\sqrt{|K_4(\boldsymbol{\theta})|r}}{p}\varphi.$$

Solve it we have

$$|K_4(\boldsymbol{\theta})| \leqslant \frac{10^4 r}{p^2}\varphi^2.$$

This implies

$$\|\boldsymbol{M} - \widehat{\boldsymbol{M}}\|_F^2 = \|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top\|_F^2 \leqslant \|\boldsymbol{Z}\boldsymbol{Z}^\top - \boldsymbol{W}\boldsymbol{W}^\top\|_F^2 \leqslant \frac{17}{3}|K_4(\boldsymbol{\theta})| \leqslant \frac{6\times 10^4 r}{p^2}\varphi^2.$$

Letting $E_3 = E_{Ca3} \cap E_{v3}$ and $C_6 = 6 \times 10^4$ finishes the proof.

B.2. Supporting proofs of Section B.1

B.2.1. Proof of Proposition B.1.3.

Proof. First, since $\boldsymbol{M}$ has SVD $\boldsymbol{M} = \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, we have

$$\operatorname{colspan}([\boldsymbol{u}_1, \dots, \boldsymbol{u}_r]) = \operatorname{colspan}(\boldsymbol{M}) \quad \text{and} \quad \operatorname{colspan}([\boldsymbol{v}_1, \dots, \boldsymbol{v}_r]) = \operatorname{rowspan}(\boldsymbol{M})$$

as well as

$$\dim(\operatorname{colspan}(\boldsymbol{M})) = \dim(\operatorname{rowspan}(\boldsymbol{M})) = r.$$

From (3.5), we also have

$$\operatorname{colspan}(\boldsymbol{M}) \subset \operatorname{colspan}(\boldsymbol{U}) \quad \text{and} \quad \operatorname{rowspan}(\boldsymbol{M}) \subset \operatorname{colspan}(\boldsymbol{V}).$$

By the way we define $\boldsymbol{U}$ and $\boldsymbol{V}$, we have $\dim(\operatorname{colspan}(\boldsymbol{U})) \leqslant r$ and $\dim(\operatorname{colspan}(\boldsymbol{V})) \leqslant r$. Therefore, $\operatorname{colspan}(\boldsymbol{U}) = \operatorname{colspan}([\boldsymbol{u}_1, \dots, \boldsymbol{u}_r])$ and $\operatorname{colspan}(\boldsymbol{V}) = \operatorname{colspan}([\boldsymbol{v}_1, \dots, \boldsymbol{v}_r])$.

From second equation in (3.5), $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{V}^\top \boldsymbol{V}$, therefore,

$$\sigma_i(\boldsymbol{U}) = \sqrt{\lambda_i(\boldsymbol{U}^\top \boldsymbol{U})} = \sqrt{\lambda_i(\boldsymbol{V}^\top \boldsymbol{V})} = \sigma_i(\boldsymbol{V}),\ i = 1, 2, \dots, r.$$

Moreover, suppose $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{B}\boldsymbol{D}^2\boldsymbol{B}^\top$ be a fixed eigenvalue decomposition of $\boldsymbol{U}^\top \boldsymbol{U}$, with $\boldsymbol{B} \in \mathsf{O}(r)$ and $\boldsymbol{D} \in \mathbb{R}^{r\times r}$ diagonal matrix. Then the reduced SVD of $\boldsymbol{U}$ and $\boldsymbol{V}$ can be written as

$$\boldsymbol{U} = \boldsymbol{A}_{\boldsymbol{U}}\boldsymbol{D}\boldsymbol{B}^\top, \quad \boldsymbol{V} = \boldsymbol{A}_{\boldsymbol{V}}\boldsymbol{D}\boldsymbol{B}^\top$$

with $\boldsymbol{A}_{\boldsymbol{U}} \in \mathbb{R}^{n_1\times r}, \boldsymbol{A}_{\boldsymbol{V}} \in \mathbb{R}^{n_2\times r}$ satisfying $\boldsymbol{A}_{\boldsymbol{U}}^\top \boldsymbol{A}_{\boldsymbol{U}} = \boldsymbol{I}$ and $\boldsymbol{A}_{\boldsymbol{V}}^\top \boldsymbol{A}_{\boldsymbol{V}} = \boldsymbol{I}$. Therefore, $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{V}^\top = \boldsymbol{A}_{\boldsymbol{U}}\boldsymbol{D}^2\boldsymbol{A}_{\boldsymbol{V}}^\top$. It is a reduced SVD of $\boldsymbol{M}$ by the way we define $\boldsymbol{A}_{\boldsymbol{U}}$, $\boldsymbol{A}_{\boldsymbol{V}}$ and $\boldsymbol{D}$. Therefore, $\sigma_1(\boldsymbol{U}) = \sigma_1(\boldsymbol{V}) = \sqrt{\sigma_1}$, $\sigma_r(\boldsymbol{U}) = \sigma_r(\boldsymbol{V}) = \sqrt{\sigma_r}$ and

$$\|\boldsymbol{U}\|_{2,\infty}^2 = \|\boldsymbol{A}_{\boldsymbol{U}}\boldsymbol{D}\boldsymbol{B}^\top\|_{2,\infty}^2 = \|\boldsymbol{A}_{\boldsymbol{U}}\boldsymbol{D}\|_{2,\infty}^2 \leqslant \|\boldsymbol{A}_{\boldsymbol{U}}\|_{2,\infty}^2 \|\boldsymbol{D}\|_{\ell_\infty}^2 = \sigma_1 \|\boldsymbol{A}_{\boldsymbol{U}}\|_{2,\infty}^2.$$

Moreover, there is $\boldsymbol{R}_{\boldsymbol{U}}, \boldsymbol{R}_{\boldsymbol{V}} \in \mathsf{O}(r)$ such that $\boldsymbol{A}_{\boldsymbol{U}} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_r]\boldsymbol{R}_{\boldsymbol{U}}, \boldsymbol{A}_{\boldsymbol{V}} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_r]\boldsymbol{R}_{\boldsymbol{V}}$. Therefore,

$$\|\boldsymbol{U}\|_{2,\infty}^2 \leqslant \sigma_1 \|\boldsymbol{A}_{\boldsymbol{U}}\|_{2,\infty}^2 = \sigma_1 \|[\boldsymbol{u}_1, \dots, \boldsymbol{u}_r]\boldsymbol{R}_{\boldsymbol{U}}\|_{2,\infty}^2 = \sigma_1 \|[\boldsymbol{u}_1, \dots, \boldsymbol{u}_r]\|_{2,\infty}^2 \leqslant \frac{\mu r}{n_1}\sigma_1.$$

Similarly, we also have $\|\boldsymbol{V}\|_{2,\infty}^2 \leqslant \frac{\mu r}{n_2}\sigma_1$.

□

B.3. Proof of Lemma 3.3.1

PROOF. Recall $\widetilde{\boldsymbol{U}}$ and $\widetilde{\boldsymbol{V}}$ are orthonormal basis matrices, $\boldsymbol{P}_{\widetilde{\boldsymbol{U}}} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^\top$, $\boldsymbol{P}_{\widetilde{\boldsymbol{V}}} = \widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^\top$. Therefore,

$$\|\boldsymbol{P}_{\widetilde{\boldsymbol{U}}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P}_{\widetilde{\boldsymbol{V}}}\| = \|\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^\top\mathcal{P}_\Omega(\boldsymbol{N})\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^\top\| = \|\widetilde{\boldsymbol{U}}^\top\mathcal{P}_\Omega(\boldsymbol{N})\widetilde{\boldsymbol{V}}\|.$$

The last equality uses the fact that $\widetilde{\boldsymbol{U}}$ and $\widetilde{\boldsymbol{V}}$ are orthonormal basis matrices, therefore

$$\|\widetilde{\boldsymbol{U}}\boldsymbol{A}\| = \|\boldsymbol{A}\|, \quad \|\boldsymbol{B}\widetilde{\boldsymbol{V}}^\top\| = \|\boldsymbol{B}\|$$

for any $\boldsymbol{A}$, $\boldsymbol{B}$ with suitable size.

Due to the fact that Ω follows from Model 2.5.1, entries of $\mathcal{P}_\Omega(\boldsymbol{N})$ can be written as $[\mathcal{P}_\Omega(\boldsymbol{N})]_{i,j} = \delta_{i,j}N_{i,j}$, where $\delta_{i,j}$'s are i.i.d. Bernoulli random variables such that

$$\delta_{i,j} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases}$$

And $N_{i,j}$'s are i.i.d. centered sub-exponential random variables. Moreover, $\delta_{i,j}$'s and $N_{i,j}$'s are mutually independent. Therefore,

$$\begin{aligned}\|\boldsymbol{P}_{\widetilde{\boldsymbol{U}}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P}_{\widetilde{\boldsymbol{V}}}\| =& \|\widetilde{\boldsymbol{U}}^\top\mathcal{P}_\Omega(\boldsymbol{N})\widetilde{\boldsymbol{V}}\| \\ =& \left\|\widetilde{\boldsymbol{U}}^\top\left(\sum_{i,j}\delta_{i,j}N_{i,j}\boldsymbol{e}_i\boldsymbol{e}_j^\top\right)\widetilde{\boldsymbol{V}}\right\| \\ =& \left\|\sum_{i,j}\delta_{i,j}N_{i,j}\widetilde{\boldsymbol{U}}_{i,\cdot}\widetilde{\boldsymbol{V}}_{j,\cdot}^\top\right\|.\end{aligned}$$

Now let

$$\boldsymbol{Q}_{i,j} := \delta_{i,j}N_{i,j}\begin{bmatrix} \boldsymbol{0} & \widetilde{\boldsymbol{U}}_{i,\cdot}\widetilde{\boldsymbol{V}}_{j,\cdot}^\top \\ \widetilde{\boldsymbol{V}}_{j,\cdot}\widetilde{\boldsymbol{U}}_{i,\cdot}^\top & \boldsymbol{0}\end{bmatrix}.$$

Therefore,

$$\|\boldsymbol{P}_{\widetilde{\boldsymbol{U}}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P}_{\widetilde{\boldsymbol{V}}}\| = \|\widetilde{\boldsymbol{U}}^\top\mathcal{P}_\Omega(\boldsymbol{N})\widetilde{\boldsymbol{V}}\| = \left\|\sum_{i,j}\boldsymbol{Q}_{i,j}\right\|$$

and $\mathbb{E}[\boldsymbol{Q}_{i,j}] = \boldsymbol{0}$. By following the symmetrization argument in [**Wai19**, Example 6.14], without loss of generality, we can assume that $N_{i,j}$'s are symmetric random variable, i.e., $N_{i,j} \overset{d}{=} -N_{i,j}$.

Now we want to verify the Bernstein's condition [Wai19, Definition 6.10] for $\boldsymbol{Q}_{i,j}$'s. For $k \geqslant 3$,

$$\mathbb{E}\left[\boldsymbol{Q}_{i,j}^k\right] = \mathbb{E}\left[\delta_{i,j}^k N_{i,j}^k \begin{bmatrix} \mathbf{0} & \widetilde{\boldsymbol{U}}_{i,\cdot}\widetilde{\boldsymbol{V}}_{j,\cdot}^\top \\ \widetilde{\boldsymbol{V}}_{j,\cdot}\widetilde{\boldsymbol{U}}_{i,\cdot}^\top & \mathbf{0} \end{bmatrix}^k\right] = p\mathbb{E}[N_{i,j}^k] \begin{bmatrix} \mathbf{0} & \widetilde{\boldsymbol{U}}_{i,\cdot}\widetilde{\boldsymbol{V}}_{j,\cdot}^\top \\ \widetilde{\boldsymbol{V}}_{j,\cdot}\widetilde{\boldsymbol{U}}_{i,\cdot}^\top & \mathbf{0} \end{bmatrix}^k.$$

Due to the symmetry of $N_{i,j}$, $\mathbb{E}[N_{i,j}^k] = 0$ when $k \geqslant 3$ is odd, therefore, $\mathbb{E}[\boldsymbol{Q}_{i,j}^k] = 0$. For $k \geqslant 2$ even, we have

$$\begin{aligned} \begin{bmatrix} \mathbf{0} & \widetilde{\boldsymbol{U}}_{i,\cdot}\widetilde{\boldsymbol{V}}_{j,\cdot}^\top \\ \widetilde{\boldsymbol{V}}_{j,\cdot}\widetilde{\boldsymbol{U}}_{i,\cdot}^\top & \mathbf{0} \end{bmatrix}^k &= \begin{bmatrix} (\widetilde{\boldsymbol{U}}_{i,\cdot}\widetilde{\boldsymbol{V}}_{j,\cdot}^\top\widetilde{\boldsymbol{V}}_{j,\cdot}\widetilde{\boldsymbol{U}}_{i,\cdot}^\top)^{k/2} & \mathbf{0} \\ \mathbf{0} & (\widetilde{\boldsymbol{V}}_{j,\cdot}\widetilde{\boldsymbol{U}}_{i,\cdot}^\top\widetilde{\boldsymbol{U}}_{i,\cdot}\widetilde{\boldsymbol{V}}_{j,\cdot}^\top)^{k/2} \end{bmatrix} \\ &= \|\widetilde{\boldsymbol{U}}_{i,\cdot}\|_2^k \|\widetilde{\boldsymbol{V}}_{j,\cdot}\|_2^k \begin{bmatrix} \frac{1}{\|\widetilde{\boldsymbol{U}}_{i,\cdot}\|_2^2}\widetilde{\boldsymbol{U}}_{i,\cdot}\widetilde{\boldsymbol{U}}_{i,\cdot}^\top & \mathbf{0} \\ \mathbf{0} & \frac{1}{\|\widetilde{\boldsymbol{V}}_{j,\cdot}\|_2^2}\widetilde{\boldsymbol{V}}_{j,\cdot}\widetilde{\boldsymbol{V}}_{j,\cdot}^\top \end{bmatrix}, \end{aligned}$$

which is a positive semidefinite matrix. And due to the fact that $N_{i,j}$'s satisfy the Bernstein condition, for $k \geqslant 2$,

$$\mathbb{E}[N_{i,j}^k] \leqslant \frac{1}{2}k!\nu^2 b^{k-2}.$$

Therefore, for $k \geqslant 3$ even,

$$\mathbb{E}\left[\boldsymbol{Q}_{i,j}^k\right] \preceq \frac{1}{2}k!\nu^2 b^{k-2} p \|\widetilde{\boldsymbol{U}}_{i,\cdot}\|_2^k \|\widetilde{\boldsymbol{V}}_{j,\cdot}\|_2^k \begin{bmatrix} \frac{1}{\|\widetilde{\boldsymbol{U}}_{i,\cdot}\|_2^2}\widetilde{\boldsymbol{U}}_{i,\cdot}\widetilde{\boldsymbol{U}}_{i,\cdot}^\top & \mathbf{0} \\ \mathbf{0} & \frac{1}{\|\widetilde{\boldsymbol{V}}_{j,\cdot}\|_2^2}\widetilde{\boldsymbol{V}}_{j,\cdot}\widetilde{\boldsymbol{V}}_{j,\cdot}^\top \end{bmatrix}.$$

And we also have

$$\begin{aligned} \mathbb{V}\left[\boldsymbol{Q}_{i,j}\right] =& \mathbb{E}\left[\boldsymbol{Q}_{i,j}^2\right] \\ =& p\mathbb{E}[N_{i,j}^2] \begin{bmatrix} \mathbf{0} & \widetilde{\boldsymbol{U}}_{i,\cdot}\widetilde{\boldsymbol{V}}_{j,\cdot}^\top \\ \widetilde{\boldsymbol{V}}_{j,\cdot}\widetilde{\boldsymbol{U}}_{i,\cdot}^\top & \mathbf{0} \end{bmatrix}^2 \\ =& p\nu^2 \|\widetilde{\boldsymbol{U}}_{i,\cdot}\|_2^2 \|\widetilde{\boldsymbol{V}}_{j,\cdot}\|_2^2 \begin{bmatrix} \frac{1}{\|\widetilde{\boldsymbol{U}}_{i,\cdot}\|_2^2}\widetilde{\boldsymbol{U}}_{i,\cdot}\widetilde{\boldsymbol{U}}_{i,\cdot}^\top & \mathbf{0} \\ \mathbf{0} & \frac{1}{\|\widetilde{\boldsymbol{V}}_{j,\cdot}\|_2^2}\widetilde{\boldsymbol{V}}_{j,\cdot}\widetilde{\boldsymbol{V}}_{j,\cdot}^\top \end{bmatrix}. \end{aligned}$$

Therefore, for $k \geqslant 3$,

$$\mathbb{E}\left[\boldsymbol{Q}_{i,j}^k\right] \preceq \frac{1}{2}k! b^{k-2} \|\widetilde{\boldsymbol{U}}_{i,\cdot}\|_2^{k-2} \|\widetilde{\boldsymbol{V}}_{j,\cdot}\|_2^{k-2} \mathbb{V}\left[\boldsymbol{Q}_{i,j}\right].$$

Therefore, $\boldsymbol{Q}_{i,j}$ satisfies Bernstein condition with parameter $b\|\widetilde{\boldsymbol{U}}_{i,\cdot}\|_2\|\widetilde{\boldsymbol{V}}_{j,\cdot}\|_2 \leqslant b\sqrt{\frac{\mu_{\widetilde{U}}\mu_{\widetilde{V}}s_1s_2}{n_1n_2}}$. Furthermore,

$$
\begin{aligned}
\frac{1}{n_1n_2}\left\|\sum_{(i,j)\in[n_1]\times[n_2]}\mathbb{V}\left[\boldsymbol{Q}_{i,j}\right]\right\| =& \frac{1}{n_1n_2}p\nu^2\left\|\sum_{(i,j)\in[n_1]\times[n_2]}\begin{bmatrix} \|\widetilde{\boldsymbol{V}}_{j,\cdot}\|_2^2\widetilde{\boldsymbol{U}}_{i,\cdot}\widetilde{\boldsymbol{U}}_{i,\cdot}^\top & \mathbf{0} \\ \mathbf{0} & \|\widetilde{\boldsymbol{U}}_{i,\cdot}\|_2^2\widetilde{\boldsymbol{V}}_{j,\cdot}\widetilde{\boldsymbol{V}}_{j,\cdot}^\top \end{bmatrix}\right\| \\
=& \frac{1}{n_1n_2}p\nu^2\left\|\begin{bmatrix} \|\widetilde{\boldsymbol{V}}\|_F^2\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}} & \mathbf{0} \\ \mathbf{0} & \|\widetilde{\boldsymbol{U}}\|_F^2\widetilde{\boldsymbol{V}}^\top\widetilde{\boldsymbol{V}} \end{bmatrix}\right\| \\
\leqslant& \frac{1}{n_1n_2}p\nu^2(s_1+s_2).
\end{aligned}
$$

Where the last equality uses the fact that $\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}} = \boldsymbol{I}$, $\widetilde{\boldsymbol{V}}^\top\widetilde{\boldsymbol{V}} = \boldsymbol{I}$. Then by [Wai19, Theorem 6.17], for all $t > 0$,

$$
\mathbb{P}\left[\frac{1}{n_1n_2}\left\|\sum_{i,j}\boldsymbol{Q}_{i,j}\right\| \geqslant t\right] \leqslant 2(n_1+n_2)\exp\left(-\frac{n_1n_2t^2}{2\left(\frac{1}{n_1n_2}p\nu^2(s_1+s_2) + b\sqrt{\frac{\mu_{\widetilde{U}}\mu_{\widetilde{V}}s_1s_2}{n_1n_2}}t\right)}\right).
$$

Therefore, by choosing t as

$$
t = C_w\frac{1}{n_1n_2}\left(\sqrt{p\nu^2(s_1+s_2)\log(n_1+n_2)} + b\sqrt{\frac{\mu_{\widetilde{U}}\mu_{\widetilde{V}}s_1s_2}{n_1n_2}}\log(n_1+n_2)\right)
$$

with absolute constant C_w sufficiently large, say $C_w = 10$, then

$$
\begin{aligned}
&\mathbb{P}\left[\left\|\sum_{i,j}\boldsymbol{Q}_{i,j}\right\| \geqslant C_w\left(\sqrt{p\nu^2(s_1+s_2)\log(n_1+n_2)} + b\sqrt{\frac{\mu_{\widetilde{U}}\mu_{\widetilde{V}}s_1s_2}{n_1n_2}}\log(n_1+n_2)\right)\right] \\
\leqslant&(n_1+n_2)^{-3}.
\end{aligned}
$$

Using the fact that

$$
\|\boldsymbol{P}_{\widetilde{\boldsymbol{U}}}\mathcal{P}_\Omega(\boldsymbol{N})\boldsymbol{P}_{\widetilde{\boldsymbol{V}}}\| = \left\|\sum_{i,j}\boldsymbol{Q}_{i,j}\right\|
$$

finishes the proof. □

APPENDIX C

Supporting Proofs of Chapter 4

C.1. Proof of Lemma 4.2.1

For the proof, we mainly follow the technical framework introduced by [MWCC18] and extend their result to the rectangular case. Within the proof, we employ Lemma 4.4 from [CL19] as well as Lemma 9 from [ZL16] (Lemma C.1.1 here) to simplify the proof, and get a weaker assumption (4.13) (here) comparing to equation (63a) in [MWCC18, Lemma 7] by a factor of $\log(n_1 \vee n_2)$.

Proof. For the Hessian, we can compute as [GLM16, GJZ17, ZLTW17] did and have

$$\begin{aligned}
&\operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right)^{\top} \nabla^2 f(\boldsymbol{X}, \boldsymbol{Y}) \operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right) \\
=&\frac{2}{p}\langle \mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{U}\boldsymbol{V}^\top), \mathcal{P}_\Omega(\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{D}_{\boldsymbol{Y}}^\top)\rangle + \frac{1}{p}\left\|\mathcal{P}_\Omega(\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{Y}^\top + \boldsymbol{X}\boldsymbol{D}_{\boldsymbol{Y}}^\top)\right\|_F^2 \\
&+ \frac{1}{2}\langle \boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}, \boldsymbol{D}_{\boldsymbol{X}}^\top\boldsymbol{D}_{\boldsymbol{X}} - \boldsymbol{D}_{\boldsymbol{Y}}^\top\boldsymbol{D}_{\boldsymbol{Y}}\rangle + \frac{1}{4}\left\|\boldsymbol{D}_{\boldsymbol{X}}^\top\boldsymbol{X} + \boldsymbol{X}^\top\boldsymbol{D}_{\boldsymbol{X}} - \boldsymbol{Y}^\top\boldsymbol{D}_{\boldsymbol{Y}} - \boldsymbol{D}_{\boldsymbol{Y}}^\top\boldsymbol{Y}\right\|_F^2.
\end{aligned}$$

First we consider the population level, i.e.,

$$\mathbb{E}\left[\operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right)^{\top} \nabla^2 f(\boldsymbol{X}, \boldsymbol{Y}) \operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right)\right].$$

Denoting $\boldsymbol{\Delta}_{\boldsymbol{X}} \coloneqq \boldsymbol{X} - \boldsymbol{U}, \boldsymbol{\Delta}_{\boldsymbol{Y}} \coloneqq \boldsymbol{Y} - \boldsymbol{V}$, and using similar decomposition as in (4.29) and (4.31), we have

$$
\begin{aligned}
&\mathbb{E}\left[\operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right)^{\top} \nabla^2 f(\boldsymbol{X}, \boldsymbol{Y}) \operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right)\right] \\
=& 2\langle \boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{V}^{\top} + \boldsymbol{U} \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} + \boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}, \boldsymbol{D}_{\boldsymbol{X}} \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \rangle + \left\| \boldsymbol{D}_{\boldsymbol{X}} \boldsymbol{V}^{\top} + \boldsymbol{D}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} + \boldsymbol{U} \boldsymbol{D}_{\boldsymbol{Y}}^{\top} + \boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \right\|_F^2 \\
&+ \frac{1}{2} \langle \boldsymbol{U}^{\top} \boldsymbol{\Delta}_{\boldsymbol{X}} + \boldsymbol{\Delta}_{\boldsymbol{X}}^{\top} \boldsymbol{U} + \boldsymbol{\Delta}_{\boldsymbol{X}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{X}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} \boldsymbol{V} - \boldsymbol{V}^{\top} \boldsymbol{\Delta}_{\boldsymbol{Y}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{Y}}, \boldsymbol{D}_{\boldsymbol{X}}^{\top} \boldsymbol{D}_{\boldsymbol{X}} - \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \boldsymbol{D}_{\boldsymbol{Y}} \rangle \\
&+ \frac{1}{4} \left\| \boldsymbol{D}_{\boldsymbol{X}}^{\top} \boldsymbol{U} + \boldsymbol{D}_{\boldsymbol{X}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{X}} + \boldsymbol{U}^{\top} \boldsymbol{D}_{\boldsymbol{X}} + \boldsymbol{\Delta}_{\boldsymbol{X}}^{\top} \boldsymbol{D}_{\boldsymbol{X}} - \boldsymbol{V}^{\top} \boldsymbol{D}_{\boldsymbol{Y}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} \boldsymbol{D}_{\boldsymbol{Y}} - \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \boldsymbol{V} - \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{Y}} \right\|_F^2 \\
=& \left\| \boldsymbol{D}_{\boldsymbol{X}} \boldsymbol{V}^{\top} + \boldsymbol{U} \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \right\|_F^2 + \frac{1}{4} \left\| \boldsymbol{D}_{\boldsymbol{X}}^{\top} \boldsymbol{U} + \boldsymbol{U}^{\top} \boldsymbol{D}_{\boldsymbol{X}} - \boldsymbol{V}^{\top} \boldsymbol{D}_{\boldsymbol{Y}} - \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \boldsymbol{V} \right\|_F^2 + \mathcal{E}_1.
\end{aligned}
\tag{C.1}
$$

Here we use the fact that $\boldsymbol{U}^{\top}\boldsymbol{U} = \boldsymbol{V}^{\top}\boldsymbol{V}$, and $\mathcal{E}_1$ contains terms with $\boldsymbol{\Delta}_{\boldsymbol{X}}$'s and $\boldsymbol{\Delta}_{\boldsymbol{Y}}$'s, i.e.,

$$
\begin{aligned}
&\mathcal{E}_1 \\
=& 2\langle \boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{V}^{\top} + \boldsymbol{U} \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} + \boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}, \boldsymbol{D}_{\boldsymbol{X}} \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \rangle + \left\| \boldsymbol{D}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} + \boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \right\|_F^2 \\
&+ 2\langle \boldsymbol{\Delta}_{\boldsymbol{X}} \boldsymbol{D}_{\boldsymbol{Y}}^{\top} + \boldsymbol{D}_{\boldsymbol{X}} \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}, \boldsymbol{D}_{\boldsymbol{X}} \boldsymbol{V}^{\top} + \boldsymbol{U} \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \rangle \\
&+ \frac{1}{2} \langle \boldsymbol{U}^{\top} \boldsymbol{\Delta}_{\boldsymbol{X}} + \boldsymbol{\Delta}_{\boldsymbol{X}}^{\top} \boldsymbol{U} + \boldsymbol{\Delta}_{\boldsymbol{X}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{X}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} \boldsymbol{V} - \boldsymbol{V}^{\top} \boldsymbol{\Delta}_{\boldsymbol{Y}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{Y}}, \boldsymbol{D}_{\boldsymbol{X}}^{\top} \boldsymbol{D}_{\boldsymbol{X}} - \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \boldsymbol{D}_{\boldsymbol{Y}} \rangle \\
&+ \frac{1}{2} \langle \boldsymbol{D}_{\boldsymbol{X}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{X}} + \boldsymbol{\Delta}_{\boldsymbol{X}}^{\top} \boldsymbol{D}_{\boldsymbol{X}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} \boldsymbol{D}_{\boldsymbol{Y}} - \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{Y}}, \boldsymbol{D}_{\boldsymbol{X}}^{\top} \boldsymbol{U} + \boldsymbol{U}^{\top} \boldsymbol{D}_{\boldsymbol{X}} - \boldsymbol{V}^{\top} \boldsymbol{D}_{\boldsymbol{Y}} - \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \boldsymbol{V} \rangle \\
&+ \frac{1}{4} \left\| \boldsymbol{D}_{\boldsymbol{X}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{X}} + \boldsymbol{\Delta}_{\boldsymbol{X}}^{\top} \boldsymbol{D}_{\boldsymbol{X}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} \boldsymbol{D}_{\boldsymbol{Y}} - \boldsymbol{D}_{\boldsymbol{Y}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{Y}} \right\|_F^2 .
\end{aligned}
$$

Multiplying terms through we have

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{vec}\left(\left[\begin{array}{c}\boldsymbol{D}_{\boldsymbol{X}}\\ \boldsymbol{D}_{\boldsymbol{Y}}\end{array}\right]\right)^{\top}\nabla^2 f(\boldsymbol{X},\boldsymbol{Y})\operatorname{vec}\left(\left[\begin{array}{c}\boldsymbol{D}_{\boldsymbol{X}}\\ \boldsymbol{D}_{\boldsymbol{Y}}\end{array}\right]\right)\right]\\
=&\left\|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^{\top}\right\|_F^2+\left\|\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\right\|_F^2+\frac{1}{2}\left\|\boldsymbol{D}_{\boldsymbol{X}}^{\top}\boldsymbol{U}\right\|_F^2\\
&+\frac{1}{2}\left\|\boldsymbol{V}^{\top}\boldsymbol{D}_{\boldsymbol{Y}}\right\|_F^2-\langle\boldsymbol{D}_{\boldsymbol{X}}^{\top}\boldsymbol{U},\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\boldsymbol{V}\rangle+\frac{1}{2}\langle\boldsymbol{D}_{\boldsymbol{X}}^{\top}\boldsymbol{U},\boldsymbol{U}^{\top}\boldsymbol{D}_{\boldsymbol{X}}\rangle\\
&+\frac{1}{2}\langle\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\boldsymbol{V},\boldsymbol{V}^{\top}\boldsymbol{D}_{\boldsymbol{Y}}\rangle+\langle\boldsymbol{D}_{\boldsymbol{X}}^{\top}\boldsymbol{U},\boldsymbol{V}^{\top}\boldsymbol{D}_{\boldsymbol{Y}}\rangle+\mathcal{E}_1\\
=&\left\|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^{\top}\right\|_F^2+\left\|\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\right\|_F^2+\frac{1}{2}\left\|\boldsymbol{D}_{\boldsymbol{X}}^{\top}\boldsymbol{U}-\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\boldsymbol{V}\right\|_F^2\\
&+\frac{1}{2}\langle\boldsymbol{U}^{\top}\boldsymbol{D}_{\boldsymbol{X}}+\boldsymbol{V}^{\top}\boldsymbol{D}_{\boldsymbol{Y}},\boldsymbol{D}_{\boldsymbol{X}}^{\top}\boldsymbol{U}+\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\boldsymbol{V}\rangle+\mathcal{E}_1.
\end{aligned}$$

Now for the fourth term, we split $\boldsymbol{U}$ as $\boldsymbol{U}-\boldsymbol{X}_2+\boldsymbol{X}_2$, $\boldsymbol{V}$ as $\boldsymbol{V}-\boldsymbol{Y}_2+\boldsymbol{Y}_2$, and plug it back. Then we have

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{vec}\left(\left[\begin{array}{c}\boldsymbol{D}_{\boldsymbol{X}}\\ \boldsymbol{D}_{\boldsymbol{Y}}\end{array}\right]\right)^{\top}\nabla^2 f(\boldsymbol{X},\boldsymbol{Y})\operatorname{vec}\left(\left[\begin{array}{c}\boldsymbol{D}_{\boldsymbol{X}}\\ \boldsymbol{D}_{\boldsymbol{Y}}\end{array}\right]\right)\right]\\
=&\left\|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^{\top}\right\|_F^2+\left\|\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\right\|_F^2+\frac{1}{2}\left\|\boldsymbol{D}_{\boldsymbol{X}}^{\top}\boldsymbol{U}-\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\boldsymbol{V}\right\|_F^2\\
&+\frac{1}{2}\langle\boldsymbol{X}_2^{\top}\boldsymbol{D}_{\boldsymbol{X}}+\boldsymbol{Y}_2^{\top}\boldsymbol{D}_{\boldsymbol{Y}},\boldsymbol{D}_{\boldsymbol{X}}^{\top}\boldsymbol{X}_2+\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y}_2\rangle+\mathcal{E}_1+\mathcal{E}_2,
\end{aligned}$$

where $\mathcal{E}_2$ contains terms with $\boldsymbol{U}-\boldsymbol{X}_2$'s and $\boldsymbol{V}-\boldsymbol{Y}_2$'s, i.e.,

$$\begin{aligned}
&\mathcal{E}_2\\
=&\frac{1}{2}\langle(\boldsymbol{U}-\boldsymbol{X}_2)^{\top}\boldsymbol{D}_{\boldsymbol{X}}+(\boldsymbol{V}-\boldsymbol{Y}_2)^{\top}\boldsymbol{D}_{\boldsymbol{Y}},\boldsymbol{D}_{\boldsymbol{X}}^{\top}\boldsymbol{X}_2+\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y}_2\rangle\\
&+\frac{1}{2}\langle\boldsymbol{X}_2^{\top}\boldsymbol{D}_{\boldsymbol{X}}+\boldsymbol{Y}_2^{\top}\boldsymbol{D}_{\boldsymbol{Y}},\boldsymbol{D}_{\boldsymbol{X}}^{\top}(\boldsymbol{U}-\boldsymbol{X}_2)+\boldsymbol{D}_{\boldsymbol{Y}}^{\top}(\boldsymbol{V}-\boldsymbol{Y}_2)\rangle\\
&+\frac{1}{2}\langle(\boldsymbol{U}-\boldsymbol{X}_2)^{\top}\boldsymbol{D}_{\boldsymbol{X}}+(\boldsymbol{V}-\boldsymbol{Y}_2)^{\top}\boldsymbol{D}_{\boldsymbol{Y}},\boldsymbol{D}_{\boldsymbol{X}}^{\top}(\boldsymbol{U}-\boldsymbol{X}_2)+\boldsymbol{D}_{\boldsymbol{Y}}^{\top}(\boldsymbol{V}-\boldsymbol{Y}_2)\rangle.
\end{aligned}$$

By the way we define $\widehat{\boldsymbol{R}}$ in (4.14), $\begin{bmatrix} \boldsymbol{X}_2 \\ \boldsymbol{Y}_2 \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}$ is symmetric. Using this fact we have

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right)^\top \nabla^2 f(\boldsymbol{X},\boldsymbol{Y}) \operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right)\right] \\
&= \left\|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top\right\|_F^2 + \left\|\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^\top\right\|_F^2 + \frac{1}{2}\left\|\boldsymbol{D}_{\boldsymbol{X}}^\top\boldsymbol{U} - \boldsymbol{D}_{\boldsymbol{Y}}^\top\boldsymbol{V}\right\|_F^2 + \frac{1}{2}\left\|\boldsymbol{X}_2^\top\boldsymbol{D}_{\boldsymbol{X}} + \boldsymbol{Y}_2^\top\boldsymbol{D}_{\boldsymbol{Y}}\right\|_F^2 + \mathcal{E}_1 + \mathcal{E}_2.
\end{aligned} \tag{C.2}$$

For $\mathcal{E}_1 + \mathcal{E}_2$, by the way we define them, we have the following bound:

$$\begin{aligned}
&|\mathcal{E}_1 + \mathcal{E}_2| \\
\leqslant& 9[(\|\boldsymbol{U} - \boldsymbol{X}_2\| + \|\boldsymbol{V} - \boldsymbol{Y}_2\|)(\|\boldsymbol{X}_2\| + \|\boldsymbol{Y}_2\|) + (\|\boldsymbol{U} - \boldsymbol{X}_2\| + \|\boldsymbol{V} - \boldsymbol{Y}_2\|)^2 \\
&+ (\|\boldsymbol{\Delta}_{\boldsymbol{X}}\| + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|)(\|\boldsymbol{U}\| + \|\boldsymbol{V}\|) + (\|\boldsymbol{\Delta}_{\boldsymbol{X}}\| + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|)^2] \times (\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2 + \|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2).
\end{aligned}$$

From the assumption,

$$\left\|\begin{bmatrix} \boldsymbol{X}_2 - \boldsymbol{U} \\ \boldsymbol{Y}_2 - \boldsymbol{V} \end{bmatrix}\right\| \leqslant \frac{1}{500\kappa}\sqrt{\sigma_1(\boldsymbol{M})},$$

$$\left\|\begin{bmatrix} \boldsymbol{X} - \boldsymbol{U} \\ \boldsymbol{Y} - \boldsymbol{V} \end{bmatrix}\right\|_{2,\infty} \leqslant \frac{1}{500\kappa\sqrt{n_1+n_2}}\sqrt{\sigma_1(\boldsymbol{M})},$$

and

$$\begin{aligned}
\left\|\begin{bmatrix} \boldsymbol{X} - \boldsymbol{U} \\ \boldsymbol{Y} - \boldsymbol{V} \end{bmatrix}\right\| \leqslant& \left\|\begin{bmatrix} \boldsymbol{X} - \boldsymbol{U} \\ \boldsymbol{Y} - \boldsymbol{V} \end{bmatrix}\right\|_F \\
\leqslant& \sqrt{n_1+n_2}\left\|\begin{bmatrix} \boldsymbol{X} - \boldsymbol{U} \\ \boldsymbol{Y} - \boldsymbol{V} \end{bmatrix}\right\|_{2,\infty} \\
\leqslant& \frac{1}{500\kappa}\sqrt{\sigma_1(\boldsymbol{M})},
\end{aligned}$$

therefore we have

$$|\mathcal{E}_1 + \mathcal{E}_2| \leqslant \frac{1}{5}\sigma_r(\boldsymbol{M})\left\|\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right\|_F^2. \tag{C.3}$$

Now we start to consider the difference between population level and empirical level, comparing with (C.1):

$$\begin{aligned}&\operatorname{vec}\left(\begin{bmatrix}\boldsymbol{D_X}\\ \boldsymbol{D_Y}\end{bmatrix}\right)^\top \nabla^2 f(\boldsymbol{X},\boldsymbol{Y}) \operatorname{vec}\left(\begin{bmatrix}\boldsymbol{D_X}\\ \boldsymbol{D_Y}\end{bmatrix}\right)\\ &-\mathbb{E}\left[\operatorname{vec}\left(\begin{bmatrix}\boldsymbol{D_X}\\ \boldsymbol{D_Y}\end{bmatrix}\right)^\top \nabla^2 f(\boldsymbol{X},\boldsymbol{Y}) \operatorname{vec}\left(\begin{bmatrix}\boldsymbol{D_X}\\ \boldsymbol{D_Y}\end{bmatrix}\right)\right]\\ =&\textcircled{1}+\textcircled{2}+\textcircled{3}+\textcircled{4},\end{aligned}$$

where $D(\cdot,\cdot)$ denotes the difference between population level and empirical level, i.e.,

$$D(\boldsymbol{AC}^\top,\boldsymbol{BD}^\top) := \frac{1}{p}\langle \mathcal{P}_\Omega(\boldsymbol{AC}^\top), \mathcal{P}_\Omega(\boldsymbol{BD}^\top)\rangle - \langle \boldsymbol{AC}^\top, \boldsymbol{BD}^\top\rangle. \tag{C.4}$$

And

$$\begin{aligned}\textcircled{1} :=& 2D(\boldsymbol{\Delta_X V}^\top, \boldsymbol{D_X D_Y^\top}) + 2D(\boldsymbol{U\Delta_Y^\top}, \boldsymbol{D_X D_Y^\top}) + 2D(\boldsymbol{\Delta_X \Delta_Y^\top}, \boldsymbol{D_X D_Y^\top})\\ &+ 2D(\boldsymbol{D_X\Delta_Y^\top}, \boldsymbol{UD_Y^\top}) + +2D(\boldsymbol{D_X V}^\top, \boldsymbol{\Delta_X D_Y^\top})2D(\boldsymbol{D_X \Delta_Y^\top}, \boldsymbol{\Delta_X D_Y^\top}),\end{aligned}$$

$$\textcircled{2} := D(\boldsymbol{D_X V}^\top, \boldsymbol{D_X V}^\top) + D(\boldsymbol{UD_Y^\top}, \boldsymbol{UD_Y^\top}) + 2D(\boldsymbol{D_X V}^\top, \boldsymbol{UD_Y^\top}),$$

$$\textcircled{3} := D(\boldsymbol{D_X \Delta_Y^\top}, \boldsymbol{D_X \Delta_Y^\top}) + D(\boldsymbol{\Delta_X D_Y^\top}, \boldsymbol{\Delta_X D_Y^\top}),$$

$$\textcircled{4} := 2D(\boldsymbol{D_X V}^\top, \boldsymbol{D_X \Delta_Y^\top}) + 2D(\boldsymbol{UD_Y^\top}, \boldsymbol{\Delta_X D_Y^\top}).$$

Now for terms with different circled numbers, we deal with them with different bounds. First, for $\textcircled{1}$, we apply Lemma 2.3.5. Therefore,

$$\begin{aligned}&|\textcircled{1}|\\ \leqslant&\frac{2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\|\boldsymbol{\Delta_X}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}\|\boldsymbol{D_X}\|_F\|\boldsymbol{D_Y}\|_F + \frac{2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\|\boldsymbol{\Delta_Y}\|_{2,\infty}\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{D_X}\|_F\|\boldsymbol{D_Y}\|_F\\ &+\frac{2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\|\boldsymbol{\Delta_X}\|_{2,\infty}\|\boldsymbol{\Delta_Y}\|_{2,\infty}\|\boldsymbol{D_X}\|_F\|\boldsymbol{D_Y}\|_F + \frac{2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\|\boldsymbol{\Delta_X}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}\|\boldsymbol{D_X}\|_F\|\boldsymbol{D_Y}\|_F\\ &+\frac{2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\|\boldsymbol{\Delta_Y}\|_{2,\infty}\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{D_X}\|_F\|\boldsymbol{D_Y}\|_F + \frac{2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\|\boldsymbol{\Delta_X}\|_{2,\infty}\|\boldsymbol{\Delta_Y}\|_{2,\infty}\|\boldsymbol{D_X}\|_F\|\boldsymbol{D_Y}\|_F.\end{aligned}$$

Using Lemma 4.3.2 and using the fact that

$$\left\| \begin{bmatrix} \boldsymbol{X} - \boldsymbol{U} \\ \boldsymbol{Y} - \boldsymbol{V} \end{bmatrix} \right\|_{2,\infty} \leqslant \frac{1}{500\kappa\sqrt{n_1 + n_2}} \sqrt{\sigma_1(\boldsymbol{M})},$$

if

$$p \geqslant C_{13} \frac{\mu r \log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

we have

$$|\textcircled{1}| \leqslant 12 C_{13} \sqrt{\frac{\mu r \kappa}{p}} \frac{1}{\kappa\sqrt{n_1 + n_2}} \sigma_1(\boldsymbol{M}) (\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2 + \|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2). \tag{C.5}$$

For $\textcircled{2}$, we apply Lemma A.4.2. Therefore,

$$\begin{aligned} |\textcircled{2}| =& |D(\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top, \boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top) + D(\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^\top, \boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^\top) + 2D(\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top, \boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^\top)| \\ =& |D(\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^\top, \boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^\top)| \\ \leqslant& 0.1 \|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^\top\|_F^2 \end{aligned} \tag{C.6}$$

given

$$p \geqslant C_{Ca} \frac{\mu r \log(n_1 \vee n_2)}{n_1 \wedge n_2}.$$

For $\textcircled{3}$, we need the following lemma:

Lemma C.1.1 ([**ZL16**, Lemma 9]). *If* $p \geqslant C_{12} \frac{\log(n_1 \vee n_2)}{n_1 \wedge n_2}$ *for some absolute constant* C_{12}*, then on an event* E_Z *with probability* $\mathbb{P}[E_Z] \geqslant 1 - (n_1 + n_2)^{-11}$*, uniformly for all matrices* $\boldsymbol{A} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{B} \in \mathbb{R}^{n_2 \times r}$*,*

$$p^{-1} \left\| \mathcal{P}_\Omega(\boldsymbol{A}\boldsymbol{B}^\top) \right\|_F^2 \leqslant 2(n_1 \vee n_2) \min\left\{ \|\boldsymbol{A}\|_F^2 \|\boldsymbol{B}\|_{2,\infty}^2, \|\boldsymbol{A}\|_{2,\infty}^2 \|\boldsymbol{B}\|_F^2 \right\}$$

holds.

In order to apply Lemma C.1.1 in our case, note

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{B}^\top\|_F^2 =&\sum_{i,j}\langle \boldsymbol{A}_{i,\cdot}, \boldsymbol{B}_{j,\cdot}\rangle^2\\ \leqslant&\sum_{i,j}\|\boldsymbol{A}_{i,\cdot}\|_2^2\|\boldsymbol{B}_{j,\cdot}\|_2^2\\ \leqslant&(n_1 \vee n_2)\min\left\{\|\boldsymbol{A}\|_F^2\|\boldsymbol{B}\|_{2,\infty}^2, \|\boldsymbol{A}\|_{2,\infty}^2\|\boldsymbol{B}\|_F^2\right\}.\end{aligned}$$

Therefore, by triangle inequality,

$$|D(\boldsymbol{A}\boldsymbol{B}^\top, \boldsymbol{A}\boldsymbol{B}^\top)| \leqslant 3(n_1 \vee n_2)\min\left\{\|\boldsymbol{A}\|_F^2\|\boldsymbol{B}\|_{2,\infty}^2, \|\boldsymbol{A}\|_{2,\infty}^2\|\boldsymbol{B}\|_F^2\right\}.$$

So we have

$$|\textcircled{3}| \leqslant 3(n_1 \vee n_2)\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{2,\infty}^2 + 3(n_1 \vee n_2)\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{2,\infty}^2.$$

Using the fact that

$$\left\|\begin{bmatrix}\boldsymbol{X}-\boldsymbol{U}\\ \boldsymbol{Y}-\boldsymbol{V}\end{bmatrix}\right\|_{2,\infty} \leqslant \frac{1}{500\kappa\sqrt{n_1+n_2}}\sqrt{\sigma_1(\boldsymbol{M})},$$

we further have

$$|\textcircled{3}| \leqslant 3(n_1 \vee n_2)\frac{1}{250000\kappa^2(n_1+n_2)} \times \sigma_1(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2 + \|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2). \tag{C.7}$$

Finally, for $\textcircled{4}$, by triangle inequality,

$$\begin{aligned}&|D(\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top, \boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top)|\\ =&|p^{-1}\langle \mathcal{P}_\Omega(\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top), \mathcal{P}_\Omega(\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top)\rangle - \langle \boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top, \boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\rangle|\\ \leqslant&\sqrt{p^{-1}\|\mathcal{P}_\Omega(\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top)\|_F^2}\sqrt{p^{-1}\|\mathcal{P}_\Omega(\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top)\|_F^2} + |\langle \boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top, \boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\rangle|.\end{aligned}$$

Now by applying Lemma C.1.1 and Lemma A.4.2 we have

$$\begin{aligned}&|D(\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top, \boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top)|\\ \leqslant&\sqrt{2(n_1 \vee n_2)\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{2,\infty}^2}\sqrt{(1+0.1)\|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^\top\|_F^2} + \|\boldsymbol{V}\|\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2\\ \leqslant&\sqrt{3(n_1 \vee n_2)}\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{2,\infty}\|\boldsymbol{V}\|\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2 + \|\boldsymbol{V}\|\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2.\end{aligned}$$

Similarly, we also have

$$|D(\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^{\top}, \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{D}_{\boldsymbol{Y}}^{\top})| \leqslant \sqrt{3(n_1 \vee n_2)}\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{2,\infty}\|\boldsymbol{U}\|\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2 + \|\boldsymbol{U}\|\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2.$$

Using the fact that

$$\left\|\begin{bmatrix} \boldsymbol{X} - \boldsymbol{U} \\ \boldsymbol{Y} - \boldsymbol{V} \end{bmatrix}\right\| \leqslant \frac{1}{500\kappa}\sqrt{\sigma_1(\boldsymbol{M})}.$$

Therefore,

$$\begin{aligned}
&|\textcircled{4}| \\
&\leqslant 2\sqrt{3(n_1 \vee n_2)}\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{2,\infty}\|\boldsymbol{V}\|\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2 + 2\|\boldsymbol{V}\|\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2 \\
&\quad + 2\sqrt{3(n_1 \vee n_2)}\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{2,\infty}\|\boldsymbol{U}\|\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2 + 2\|\boldsymbol{U}\|\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2 \\
&\leqslant 2\sqrt{3(n_1 \vee n_2)}\frac{1}{500\kappa\sqrt{n_1 + n_2}} \times \sigma_1(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2 + \|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2) \\
&\quad + \frac{2}{500\kappa}\sigma_1(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2 + \|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2).
\end{aligned} \tag{C.8}$$

Putting the estimation for $\textcircled{1}$, $\textcircled{2}$, $\textcircled{3}$ and $\textcircled{4}$ together, i.e., (C.5), (C.6), (C.7), (C.8), if

$$p \geqslant (C_{Ca} + C_{12} + C_{13})\frac{\mu r \log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

then

$$\begin{aligned}
&\left|\operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right)^{\top} \nabla^2 f(\boldsymbol{X}, \boldsymbol{Y}) \operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right) - \mathbb{E}\left[\operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right)^{\top} \nabla^2 f(\boldsymbol{X}, \boldsymbol{Y}) \operatorname{vec}\left(\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{X}} \\ \boldsymbol{D}_{\boldsymbol{Y}} \end{bmatrix}\right)\right]\right| \\
&\leqslant 12C_{13}\sqrt{\frac{\mu r \kappa}{p}}\frac{1}{\kappa\sqrt{n_1 + n_2}}\sigma_1(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2 + \|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2) + 0.1\|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^{\top} + \boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\|_F^2 \\
&\quad + 3(n_1 \vee n_2)\frac{1}{250000\kappa^2(n_1 + n_2)}\sigma_1(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2 + \|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2) \\
&\quad + 2\sqrt{3(n_1 \vee n_2)}\frac{1}{500\kappa\sqrt{n_1 + n_2}}\sigma_1(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2 + \|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2) + \frac{2}{500\kappa}\sigma_1(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2 + \|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2)
\end{aligned}$$

holds on an event $E_H = E_S \bigcap E_{Ca} \bigcap E_Z$ with probability $\mathbb{P}[E_H] = \mathbb{P}[E_S \bigcap E_{Ca} \bigcap E_Z] \geqslant 1 - 3(n_1 + n_2)^{-11}$. If in addition

$$p \geqslant 14400C_{13}^2\frac{\mu r \kappa}{n_1 \wedge n_2},$$

then

(C.9)

$$\left|\operatorname{vec}\left(\begin{bmatrix}\boldsymbol{D}_{\boldsymbol{X}}\\ \boldsymbol{D}_{\boldsymbol{Y}}\end{bmatrix}\right)^{\top}\nabla^2 f(\boldsymbol{X},\boldsymbol{Y})\operatorname{vec}\left(\begin{bmatrix}\boldsymbol{D}_{\boldsymbol{X}}\\ \boldsymbol{D}_{\boldsymbol{Y}}\end{bmatrix}\right)-\mathbb{E}\left[\operatorname{vec}\left(\begin{bmatrix}\boldsymbol{D}_{\boldsymbol{X}}\\ \boldsymbol{D}_{\boldsymbol{Y}}\end{bmatrix}\right)^{\top}\nabla^2 f(\boldsymbol{X},\boldsymbol{Y})\operatorname{vec}\left(\begin{bmatrix}\boldsymbol{D}_{\boldsymbol{X}}\\ \boldsymbol{D}_{\boldsymbol{Y}}\end{bmatrix}\right)\right]\right|$$
$$\leqslant\frac{1}{5}\sigma_r(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2)+\frac{1}{5}(\|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^{\top}\|_F^2+\|\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\|_F^2).$$

Now by putting (C.2), (C.3), (C.9) together, we have

$$\begin{aligned}&\operatorname{vec}\left(\begin{bmatrix}\boldsymbol{D}_{\boldsymbol{X}}\\ \boldsymbol{D}_{\boldsymbol{Y}}\end{bmatrix}\right)^{\top}\nabla^2 f(\boldsymbol{X},\boldsymbol{Y})\operatorname{vec}\left(\begin{bmatrix}\boldsymbol{D}_{\boldsymbol{X}}\\ \boldsymbol{D}_{\boldsymbol{Y}}\end{bmatrix}\right)\\
\geqslant&\left\|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^{\top}\right\|_F^2+\left\|\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\right\|_F^2-\frac{1}{5}\sigma_r(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2)\\
&-\frac{1}{5}\sigma_r(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2)-\frac{1}{5}(\|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^{\top}\|_F^2+\|\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\|_F^2)\\
\geqslant&\frac{1}{5}\sigma_r(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2),\end{aligned}$$

where the last inequality we use the fact that $\|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^{\top}\|_F^2\geqslant\sigma_r^2(\boldsymbol{V})\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2=\sigma_r(\boldsymbol{M})\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2$ and also $\|\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\|_F^2\geqslant\sigma_r(\boldsymbol{M})\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2$. For the upper bound, we also have

$$\begin{aligned}&\operatorname{vec}\left(\begin{bmatrix}\boldsymbol{D}_{\boldsymbol{X}}\\ \boldsymbol{D}_{\boldsymbol{Y}}\end{bmatrix}\right)^{\top}\nabla^2 f(\boldsymbol{X},\boldsymbol{Y})\operatorname{vec}\left(\begin{bmatrix}\boldsymbol{D}_{\boldsymbol{X}}\\ \boldsymbol{D}_{\boldsymbol{Y}}\end{bmatrix}\right)\\
\leqslant&\left\|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^{\top}\right\|_F^2+\left\|\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\right\|_F^2+\frac{1}{2}\left\|\boldsymbol{D}_{\boldsymbol{X}}^{\top}\boldsymbol{U}-\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\boldsymbol{V}\right\|_F^2+\frac{1}{2}\left\|\boldsymbol{X}_2^{\top}\boldsymbol{D}_{\boldsymbol{X}}+\boldsymbol{Y}_2^{\top}\boldsymbol{D}_{\boldsymbol{Y}}\right\|_F^2\\
&+\frac{1}{5}\sigma_r(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2)+\frac{1}{5}\sigma_r(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2)+\frac{1}{5}(\|\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}^{\top}\|_F^2+\|\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\|_F^2)\\
\leqslant&\frac{6}{5}\sigma_1(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2)+\|\boldsymbol{D}_{\boldsymbol{X}}^{\top}\boldsymbol{U}\|_F^2+\|\boldsymbol{D}_{\boldsymbol{Y}}^{\top}\boldsymbol{V}\|_F^2+\|\boldsymbol{X}_2^{\top}\boldsymbol{D}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{Y}_2^{\top}\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2\\
&+\frac{2}{5}\sigma_r(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2)\\
\leqslant&\frac{13}{5}\sigma_1(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2)+\|\boldsymbol{X}_2\|^2\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{Y}_2\|^2\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2\\
\leqslant&5\sigma_1(\boldsymbol{M})(\|\boldsymbol{D}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{D}_{\boldsymbol{Y}}\|_F^2),\end{aligned}$$

where the last inequality we use the fact that

$$\left\| \begin{bmatrix} \boldsymbol{X}_2 - \boldsymbol{U} \\ \boldsymbol{Y}_2 - \boldsymbol{V} \end{bmatrix} \right\| \leqslant \frac{1}{500\kappa}\sqrt{\sigma_1(\boldsymbol{M})}.$$

Choosing $C_{S1} = C_{Ca} + C_{12} + C_{13} + 14400C_{13}^2$ finishes the proof. □

C.2. Proof of Lemma 4.2.2

In this section we first summarize some useful lemmas from [MWCC18]. We then follow the technical framework in [MWCC18] but replace [MWCC18, Lemma 39] with [Che15, Lemma 2] (Lemma 4.2.3 here) to get a better initialization guarantee.

C.2.1. Useful lemmas. Here we summarize some useful lemmas in [AFWZ17] as well as [MWCC18]. We relax the PSD assmptions on $\boldsymbol{M}_1$ in Lemma C.2.2, Lemma C.2.3 and Lemma C.2.4 to symmetric assumptions by following the proof framework introduced in [MWCC18]. In fact, lemmas listed in this section can be derived from Davis-Kahan SinΘ theorem [DK70]. We summarize lemmas here since they are intensively used throughout the proof. Moreover, for the simplicity of the expression, we made some additional assumptions on the eignevalues of $\boldsymbol{M}_1$ within the following lemmas (i.e., $\lambda_r(\boldsymbol{M}_1) > 0$, $\lambda_r(\boldsymbol{M}_1) > \lambda_{r+1}(\boldsymbol{M}_1)$, $\lambda_{r+1}(\boldsymbol{M}_1) = 0$ and $\lambda_1(\boldsymbol{M}_1) = -\lambda_n(\boldsymbol{M}_1)$), the results still hold (with a more complicated expression) without those extra assumptions. Recall that here $\lambda_1(\boldsymbol{A}) \geqslant \lambda_2(\boldsymbol{A}) \geqslant \cdots \geqslant \lambda_n(\boldsymbol{A})$ stands for eigenvalues of symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

First, we need a specified version of [AFWZ17, Lemma 3]:

Lemma C.2.1 ([AFWZ17, Lemma 3]). *Let $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{n\times n}$ be two symmetric matrices with top-r eigenvalue decomposition $\widetilde{\boldsymbol{U}}_1\boldsymbol{\Lambda}_1\widetilde{\boldsymbol{U}}_1^\top$ and $\widetilde{\boldsymbol{U}}_2\boldsymbol{\Lambda}_2\widetilde{\boldsymbol{U}}_2^\top$ correspondingly. Then if $\lambda_r(\boldsymbol{M}_1) > 0$, $\lambda_r(\boldsymbol{M}_1) > \lambda_{r+1}(\boldsymbol{M}_1)$ and*

$$\|\boldsymbol{M}_1 - \boldsymbol{M}_2\| \leqslant \frac{1}{4}\min(\lambda_r(\boldsymbol{M}_1), \lambda_r(\boldsymbol{M}_1) - \lambda_{r+1}(\boldsymbol{M}_1)),$$

we have

$$\|\widetilde{\boldsymbol{U}}_1^\top\widetilde{\boldsymbol{U}}_2 - \mathrm{sgn}(\widetilde{\boldsymbol{U}}_1^\top\widetilde{\boldsymbol{U}}_2)\| \leqslant 4\frac{\|\boldsymbol{M}_1 - \boldsymbol{M}_2\|^2}{\min\{\lambda_r(\boldsymbol{M}_1), \lambda_r(\boldsymbol{M}_1) - \lambda_{r+1}(\boldsymbol{M}_1)\}^2}$$

and

$$\|(\widetilde{\boldsymbol{U}}_1^\top \widetilde{\boldsymbol{U}}_2)^{-1}\| \leqslant 2.$$

And we also need some useful lemmas from [MWCC18]:

Lemma C.2.2 ([MWCC18, Lemma 45]). *Let $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{n\times n}$ be symmetric matrices with top-r eigenvalue decomposition $\widetilde{\boldsymbol{U}}_1\boldsymbol{\Lambda}_1\widetilde{\boldsymbol{U}}_1^\top$ and $\widetilde{\boldsymbol{U}}_2\boldsymbol{\Lambda}_2\widetilde{\boldsymbol{U}}_2^\top$ correspondingly. Assume $\lambda_r(\boldsymbol{M}_1) > 0$, $\lambda_{r+1}(\boldsymbol{M}_1) = 0$ and $\|\boldsymbol{M}_1 - \boldsymbol{M}_2\| \leqslant \frac{1}{4}\lambda_r(\boldsymbol{M}_1)$. Denote*

$$\widetilde{\boldsymbol{Q}} \coloneqq \operatorname*{argmin}_{\boldsymbol{R}\in \mathsf{O}(r)} \|\widetilde{\boldsymbol{U}}_2\boldsymbol{R} - \widetilde{\boldsymbol{U}}_1\|_F.$$

Then

$$\|\widetilde{\boldsymbol{U}}_2\widetilde{\boldsymbol{Q}} - \widetilde{\boldsymbol{U}}_1\| \leqslant \frac{3}{\lambda_r(\boldsymbol{M}_1)}\|\boldsymbol{M}_1 - \boldsymbol{M}_2\|.$$

Lemma C.2.3 ([MWCC18, Lemma 46]). *Let $\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3 \in \mathbb{R}^{n\times n}$ be symmetric matrices with top-r eigenvalue decomposition $\widetilde{\boldsymbol{U}}_1\boldsymbol{\Lambda}_1\widetilde{\boldsymbol{U}}_1^\top$, $\widetilde{\boldsymbol{U}}_2\boldsymbol{\Lambda}_2\widetilde{\boldsymbol{U}}_2^\top$ and $\widetilde{\boldsymbol{U}}_3\boldsymbol{\Lambda}_3\widetilde{\boldsymbol{U}}_3^\top$ correspondingly. Assume $\lambda_1(\boldsymbol{M}_1) = -\lambda_n(\boldsymbol{M}_1), \lambda_r(\boldsymbol{M}_1) > 0, \lambda_{r+1}(\boldsymbol{M}_1) = 0$ and $\|\boldsymbol{M}_1 - \boldsymbol{M}_2\| \leqslant \frac{1}{4}\lambda_r(\boldsymbol{M}_1)$, $\|\boldsymbol{M}_1 - \boldsymbol{M}_3\| \leqslant \frac{1}{4}\lambda_r(\boldsymbol{M}_1)$. Denote*

$$\widetilde{\boldsymbol{Q}} \coloneqq \operatorname*{argmin}_{\boldsymbol{R}\in \mathsf{O}(r)} \|\widetilde{\boldsymbol{U}}_2\boldsymbol{R} - \widetilde{\boldsymbol{U}}_3\|_F.$$

Then

$$\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{Q}} - \widetilde{\boldsymbol{Q}}\boldsymbol{\Lambda}_3^{1/2}\| \leqslant 15\frac{\lambda_1(\boldsymbol{M}_1)}{\lambda_r^{3/2}(\boldsymbol{M}_1)}\|\boldsymbol{M}_2 - \boldsymbol{M}_3\|$$

and

$$\|\boldsymbol{\Lambda}_2^{1/2}\widetilde{\boldsymbol{Q}} - \widetilde{\boldsymbol{Q}}\boldsymbol{\Lambda}_3^{1/2}\|_F \leqslant 15\frac{\lambda_1(\boldsymbol{M}_1)}{\lambda_r^{3/2}(\boldsymbol{M}_1)}\|(\boldsymbol{M}_2 - \boldsymbol{M}_3)\widetilde{\boldsymbol{U}}_2\|_F.$$

Lemma C.2.4 ([MWCC18, Lemma 47]). *Let $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{n\times n}$ be symmetric matrices with top-r eigenvalue decomposition $\widetilde{\boldsymbol{U}}_1\boldsymbol{\Lambda}_1\widetilde{\boldsymbol{U}}_1^\top$ and $\widetilde{\boldsymbol{U}}_2\boldsymbol{\Lambda}_2\widetilde{\boldsymbol{U}}_2^\top$ correspondingly. Assume $\lambda_1(\boldsymbol{M}_1) = -\lambda_n(\boldsymbol{M}_1)$, $\lambda_r(\boldsymbol{M}_1) > 0, \lambda_{r+1}(\boldsymbol{M}_1) = 0$ and*

$$\|\boldsymbol{M}_1 - \boldsymbol{M}_2\| \leqslant \frac{1}{40}\frac{\lambda_r^{5/2}(\boldsymbol{M}_1)}{\lambda_1^{3/2}(\boldsymbol{M}_1)}.$$

Denote $\boldsymbol{X}_1 = \widetilde{\boldsymbol{U}}_1 \boldsymbol{\Lambda}_1^{1/2}$ *and* $\boldsymbol{X}_2 = \widetilde{\boldsymbol{U}}_2 \boldsymbol{\Lambda}_2^{1/2}$ *and define*

$$\widetilde{\boldsymbol{Q}} \coloneqq \operatorname*{argmin}_{\boldsymbol{R} \in \mathsf{O}(r)} \|\widetilde{\boldsymbol{U}}_2 \boldsymbol{R} - \widetilde{\boldsymbol{U}}_1\|_F$$

and

$$\boldsymbol{H} \coloneqq \operatorname*{argmin}_{\boldsymbol{R} \in \mathsf{O}(r)} \|\boldsymbol{X}_2 \boldsymbol{R} - \boldsymbol{X}_1\|_F.$$

Then

$$\|\widetilde{\boldsymbol{Q}} - \boldsymbol{H}\| \leqslant 15 \frac{\lambda_1^{3/2}(\boldsymbol{M}_1)}{\lambda_r^{5/2}(\boldsymbol{M}_1)} \|\boldsymbol{M}_1 - \boldsymbol{M}_2\|$$

holds.

C.2.2. Proof. In this subsection, we will follow the technical framework in [**MWCC18**]: First we give an upper bound of $\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}) - \boldsymbol{M}\|$, and then prove Lemma 4.2.2 by applying the lemmas introduced in Section C.2.1. As claimed before, here we replace [**MWCC18**, Lemma 39] with [**Che15**, Lemma 2] to give an upper bound of $\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}) - \boldsymbol{M}\|$ and obtain a tighter error bound of the initializations.

Define the symmetric matrix

$$\overline{\boldsymbol{M}} \coloneqq \begin{bmatrix} \boldsymbol{0} & \boldsymbol{M} \\ \boldsymbol{M}^\top & \boldsymbol{0} \end{bmatrix}. \tag{C.10}$$

The SVD $\boldsymbol{M} = \widetilde{\boldsymbol{U}} \boldsymbol{\Sigma} \widetilde{\boldsymbol{V}}^\top$ implies the following eigenvalue decomposition of $\overline{\boldsymbol{M}}$:

$$\overline{\boldsymbol{M}} = \frac{1}{\sqrt{2}} \begin{bmatrix} \widetilde{\boldsymbol{U}} & \widetilde{\boldsymbol{U}} \\ \widetilde{\boldsymbol{V}} & -\widetilde{\boldsymbol{V}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{\Sigma} \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} \widetilde{\boldsymbol{U}} & \widetilde{\boldsymbol{U}} \\ \widetilde{\boldsymbol{V}} & -\widetilde{\boldsymbol{V}} \end{bmatrix}^\top.$$

From the eigenvalue decomposition, we can see $\lambda_1(\overline{\boldsymbol{M}}) = \sigma_1(\boldsymbol{M}), \cdots, \lambda_r(\overline{\boldsymbol{M}}) = \sigma_r(\boldsymbol{M})$, $\lambda_{r+1}(\overline{\boldsymbol{M}}) = 0, \cdots, \lambda_{n_1+n_2-r}(\overline{\boldsymbol{M}}) = 0$, $\lambda_{n_1+n_2-r+1}(\overline{\boldsymbol{M}}) = -\sigma_r(\boldsymbol{M}), \cdots, \lambda_{n_1+n_2}(\overline{\boldsymbol{M}}) = -\sigma_1(\boldsymbol{M})$. At the same time, we define

$$\frac{1}{p}\mathcal{P}_{\overline{\Omega}}(\overline{\boldsymbol{M}}) = \begin{bmatrix} \boldsymbol{0} & \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}) \\ \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M})^\top & \boldsymbol{0} \end{bmatrix}$$

with

$$\overline{\Omega} \coloneqq \{(i,j)\,|1 \leqslant i,j \leqslant n_1+n_2,\ (i,j-n_1)\in\Omega \text{ or } (j,i-n_1)\in\Omega\}.$$

Applying Lemma 4.2.3 on $\overline{\boldsymbol{M}}$ here, then

$$\begin{aligned}
&\left\|\frac{1}{p}\mathcal{P}_{\overline{\Omega}}(\overline{\boldsymbol{M}})-\overline{\boldsymbol{M}}\right\| \\
\leqslant& C_{14}\left(\frac{\log(n_1+n_2)}{p}\|\overline{\boldsymbol{M}}\|_{\ell_\infty}+\sqrt{\frac{\log(n_1+n_2)}{p}}\|\overline{\boldsymbol{M}}\|_{2,\infty}\right)\\
\leqslant& C_{14}\left(\frac{\log(n_1+n_2)}{p}\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}+\sqrt{\frac{\log(n_1+n_2)}{p}}(\|\boldsymbol{U}\|\|\boldsymbol{V}\|_{2,\infty}\vee\|\boldsymbol{V}\|\|\boldsymbol{U}\|_{2,\infty})\right) \qquad \text{(C.11)}\\
\leqslant& 2C_{14}\left(\frac{\mu r\kappa\log(n_1\vee n_2)}{(n_1\wedge n_2)p}+\sqrt{\frac{\mu r\kappa\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\right)\sigma_1(\boldsymbol{M})\\
\leqslant& 4C_{14}\sqrt{\frac{\mu r\kappa\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sigma_1(\boldsymbol{M})
\end{aligned}$$

holds on an event E_{Ch1} with probability $\mathbb{P}[E_{Ch1}]\geqslant 1-(n_1+n_2)^{-11}$. The last inequality holds given

$$p\geqslant\frac{\mu r\kappa\log(n_1\vee n_2)}{n_1\wedge n_2}.$$

In addition if

$$p\geqslant 25600C_{14}^2\frac{\mu r\kappa^6\log(n_1\vee n_2)}{n_1\wedge n_2},$$

we have

$$\left\|\frac{1}{p}\mathcal{P}_{\overline{\Omega}}(\overline{\boldsymbol{M}})-\overline{\boldsymbol{M}}\right\|\leqslant\frac{1}{40\sqrt{\kappa}^3}\sigma_r(\boldsymbol{M})\leqslant\frac{1}{4}\sigma_r(\boldsymbol{M}) \qquad \text{(C.12)}$$

holds on an event E_{Ch1}.

For the simplicity of notations, we denote $\overline{\boldsymbol{M}}^0$ as

$$\overline{\boldsymbol{M}}^0\coloneqq\begin{bmatrix}\mathbf{0} & \boldsymbol{M}^0\\ (\boldsymbol{M}^0)^\top & \mathbf{0}\end{bmatrix}, \qquad \text{(C.13)}$$

and denote $\overline{\boldsymbol{M}}^{0,(l)}$ as

$$\overline{\boldsymbol{M}}^{0,(l)} := \begin{bmatrix} \mathbf{0} & \boldsymbol{M}^{0,(l)} \\ (\boldsymbol{M}^{0,(l)})^\top & \mathbf{0} \end{bmatrix}. \tag{C.14}$$

$\boldsymbol{M}^0$ and $\boldsymbol{M}^{0,(l)}$ are defined in (4.2) and (4.5), correspondingly.

Again by Lemma 4.2.3, we can see on an event E_{Ch1}, for all $1 \leqslant l \leqslant n_1 + n_2$,

$$\left\|\overline{\boldsymbol{M}}^{0,(l)} - \overline{\boldsymbol{M}}\right\| \leqslant 4C_{14}\sqrt{\frac{\mu r\kappa \log(n_1 \vee n_2)}{(n_1 \wedge n_2)p}}\sigma_1(\boldsymbol{M}).$$

If

$$p \geqslant 25600C_{14}^2 \frac{\mu r\kappa^6 \log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

we also have

$$\left\|\overline{\boldsymbol{M}}^{0,(l)} - \overline{\boldsymbol{M}}\right\| \leqslant \frac{1}{40\sqrt{\kappa}^3}\sigma_r(\boldsymbol{M}) \leqslant \frac{1}{4}\sigma_r(\boldsymbol{M}). \tag{C.15}$$

Now assume $\boldsymbol{M}^0$ has SVD $\boldsymbol{A}\boldsymbol{D}\boldsymbol{B}^\top$, then by construction, $\overline{\boldsymbol{M}}^0$ have following eigendecomposition:

$$\overline{\boldsymbol{M}}^0 = \frac{1}{\sqrt{2}}\begin{bmatrix} \boldsymbol{A} & \boldsymbol{A} \\ \boldsymbol{B} & -\boldsymbol{B} \end{bmatrix}\begin{bmatrix} \boldsymbol{D} & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{D} \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} \boldsymbol{A} & \boldsymbol{A} \\ \boldsymbol{B} & -\boldsymbol{B} \end{bmatrix}^\top.$$

So if $\widetilde{\boldsymbol{X}}^0\boldsymbol{\Sigma}^0(\widetilde{\boldsymbol{Y}}^0)^\top$ is the top-r singular value decomposition of $\boldsymbol{M}^0$, we can also have

$$\frac{1}{\sqrt{2}}\begin{bmatrix} \widetilde{\boldsymbol{X}}^0 \\ \widetilde{\boldsymbol{Y}}^0 \end{bmatrix}\boldsymbol{\Sigma}^0\frac{1}{\sqrt{2}}\begin{bmatrix} \widetilde{\boldsymbol{X}}^0 \\ \widetilde{\boldsymbol{Y}}^0 \end{bmatrix}^\top$$

to be the top-r eigenvalue decomposition of $\overline{\boldsymbol{M}}^0$. So by Weyl's inequality and (C.12), we have

$$\frac{3}{4}\sigma_r(\boldsymbol{M}) \leqslant \sigma_r(\boldsymbol{\Sigma}^0) \leqslant \sigma_1(\boldsymbol{\Sigma}^0) \leqslant 2\sigma_1(\boldsymbol{M}). \tag{C.16}$$

Similarly, the same arguments also applies for $\overline{\boldsymbol{M}}^{0,(l)}$. From Weyl's inequality and (C.15), we have

$$\frac{3}{4}\sigma_r(\boldsymbol{M}) \leqslant \sigma_r(\boldsymbol{\Sigma}^{0,(l)}) \leqslant \sigma_1(\boldsymbol{\Sigma}^{0,(l)}) \leqslant 2\sigma_1(\boldsymbol{M}). \tag{C.17}$$

Now let $\boldsymbol{X}^0 := \widetilde{\boldsymbol{X}}^0(\boldsymbol{\Sigma}^0)^{1/2}, \boldsymbol{Y}^0 := \widetilde{\boldsymbol{Y}}^0(\boldsymbol{\Sigma}^0)^{1/2}$, $\boldsymbol{X}^{0,(l)} := \widetilde{\boldsymbol{X}}^{0,(l)}(\boldsymbol{\Sigma}^{0,(l)})^{1/2}$, and

$$\boldsymbol{Y}^{0,(l)} := \widetilde{\boldsymbol{Y}}^{0,(l)}(\boldsymbol{\Sigma}^{0,(l)})^{1/2},$$

where $\boldsymbol{M}^{0,(l)}$ has top-r singular value decomposition $\widetilde{\boldsymbol{X}}^{0,(l)}\boldsymbol{\Sigma}^{0,(l)}(\widetilde{\boldsymbol{Y}}^{0,(l)})^\top$. Let

$$\widetilde{\boldsymbol{W}} := \frac{1}{\sqrt{2}}\begin{bmatrix} \widetilde{\boldsymbol{U}} \\ \widetilde{\boldsymbol{V}} \end{bmatrix}, \quad \boldsymbol{W} := \frac{1}{\sqrt{2}}\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix},$$

$$\widetilde{\boldsymbol{Z}}^0 := \frac{1}{\sqrt{2}}\begin{bmatrix} \widetilde{\boldsymbol{X}}^0 \\ \widetilde{\boldsymbol{Y}}^0 \end{bmatrix}, \quad \boldsymbol{Z}^0 := \frac{1}{\sqrt{2}}\begin{bmatrix} \boldsymbol{X}^0 \\ \boldsymbol{Y}^0 \end{bmatrix},$$

and also we can denote

$$\widetilde{\boldsymbol{Z}}^{0,(l)} := \frac{1}{\sqrt{2}}\begin{bmatrix} \widetilde{\boldsymbol{X}}^{0,(l)} \\ \widetilde{\boldsymbol{Y}}^{0,(l)} \end{bmatrix}, \quad \boldsymbol{Z}^{0,(l)} := \frac{1}{\sqrt{2}}\begin{bmatrix} \boldsymbol{X}^{0,(l)} \\ \boldsymbol{Y}^{0,(l)} \end{bmatrix}. \tag{C.18}$$

Moreover, define

$$\boldsymbol{Q}^0 := \operatorname*{argmin}_{\boldsymbol{R}\in\mathsf{O}(r)} \left\|\widetilde{\boldsymbol{Z}}^0\boldsymbol{R} - \widetilde{\boldsymbol{W}}\right\|_F,$$
$$\boldsymbol{Q}^{0,(l)} := \operatorname*{argmin}_{\boldsymbol{R}\in\mathsf{O}(r)} \left\|\widetilde{\boldsymbol{Z}}^{0,(l)}\boldsymbol{R} - \widetilde{\boldsymbol{W}}\right\|_F.$$

C.2.2.1. *Proof for* (4.16). For spectral norm, by triangle inequality, we have

$$\begin{aligned} &\|\boldsymbol{Z}^0\boldsymbol{R}^0 - \boldsymbol{W}\| \\ =&\left\|\widetilde{\boldsymbol{Z}}^0(\boldsymbol{\Sigma}^0)^{1/2}(\boldsymbol{R}^0 - \boldsymbol{Q}^0) + \widetilde{\boldsymbol{Z}}^0\left((\boldsymbol{\Sigma}^0)^{1/2}\boldsymbol{Q}^0 - \boldsymbol{Q}^0\boldsymbol{\Sigma}^{1/2}\right) + \left(\widetilde{\boldsymbol{Z}}^0\boldsymbol{Q}^0 - \widetilde{\boldsymbol{W}}\right)\boldsymbol{\Sigma}^{1/2}\right\| \\ \leqslant&\|(\boldsymbol{\Sigma}^0)^{1/2}\|\left\|\boldsymbol{R}^0 - \boldsymbol{Q}^0\right\| + \|(\boldsymbol{\Sigma}^0)^{1/2}\boldsymbol{Q}^0 - \boldsymbol{Q}^0\boldsymbol{\Sigma}^{1/2}\| + \|\boldsymbol{\Sigma}^{1/2}\|\left\|\widetilde{\boldsymbol{Z}}^0\boldsymbol{Q}^0 - \widetilde{\boldsymbol{W}}\right\|. \end{aligned} \tag{C.19}$$

Now applying Lemma C.2.4 with $\boldsymbol{M}_1 = \overline{\boldsymbol{M}}, \boldsymbol{M}_2 = \overline{\boldsymbol{M}}^0$, we have

$$\|\boldsymbol{R}^0 - \boldsymbol{Q}^0\| \leqslant 15\frac{\sqrt{\kappa}^3}{\sigma_r(\boldsymbol{M})}\|\overline{\boldsymbol{M}} - \overline{\boldsymbol{M}}^0\|; \tag{C.20}$$

applying Lemma C.2.3 with $\boldsymbol{M}_1 = \boldsymbol{M}_2 = \overline{\boldsymbol{M}}, \boldsymbol{M}_3 = \overline{\boldsymbol{M}}^0$, we have

$$\|(\boldsymbol{\Sigma}^0)^{1/2}\boldsymbol{Q}^0 - \boldsymbol{Q}^0\boldsymbol{\Sigma}^{1/2}\| \leqslant 15\frac{\kappa}{\sqrt{\sigma_r(\boldsymbol{M})}}\|\overline{\boldsymbol{M}} - \overline{\boldsymbol{M}}^0\|; \tag{C.21}$$

finally, applying Lemma C.2.2 with $\boldsymbol{M}_1 = \overline{\boldsymbol{M}}, \boldsymbol{M}_2 = \overline{\boldsymbol{M}}^0$, we have

$$\left\|\widetilde{\boldsymbol{Z}}^0\boldsymbol{Q}^0 - \widetilde{\boldsymbol{W}}\right\| \leqslant \frac{3}{\sigma_r(\boldsymbol{M})}\|\overline{\boldsymbol{M}} - \overline{\boldsymbol{M}}^0\|. \tag{C.22}$$

Plugging the estimations (C.20), (C.21) and (C.22) back to (C.19), and using (C.16) and (C.17),

$$\begin{aligned}
&\left\|\begin{bmatrix}\boldsymbol{X}^0\\ \boldsymbol{Y}^0\end{bmatrix}\boldsymbol{R}^0 - \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\| \\
=&\sqrt{2}\left\|\boldsymbol{Z}^0\boldsymbol{R}^0 - \boldsymbol{W}\right\| \\
\leqslant&30\left(\frac{\sqrt{\sigma_1(\boldsymbol{M})\kappa^3}}{\sigma_r(\boldsymbol{M})} + \frac{\kappa}{\sqrt{\sigma_r(\boldsymbol{M})}} + \frac{\sqrt{\sigma_1(\boldsymbol{M})}}{\sigma_r(\boldsymbol{M})}\right)\|\overline{\boldsymbol{M}} - \overline{\boldsymbol{M}}^0\| \\
\leqslant&360C_{14}\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})}
\end{aligned} \tag{C.23}$$

holds. For the last inequality we use the estimation (C.11).

C.2.2.2. *Proof for* (4.17). Now we start to consider the bound of

$$\left\|\left(\begin{bmatrix}\boldsymbol{X}^{0,(l)}\\ \boldsymbol{Y}^{0,(l)}\end{bmatrix}\boldsymbol{R}^{0,(l)} - \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right)_{l,\cdot}\right\|_2.$$

By triangle inequality,

$$\begin{aligned}
&\left\|\left(\boldsymbol{Z}^{0,(l)}\boldsymbol{R}^{0,(l)} - \boldsymbol{W}\right)_{l,\cdot}\right\|_2 \\
=&\left\|\left(\boldsymbol{Z}^{0,(l)}\boldsymbol{R}^{0,(l)} - \boldsymbol{Z}^{0,(l)}\boldsymbol{Q}^{0,(l)} + \boldsymbol{Z}^{0,(l)}\boldsymbol{Q}^{0,(l)} - \boldsymbol{W}\right)_{l,\cdot}\right\|_2 \\
\leqslant&\left\|\left(\boldsymbol{Z}^{0,(l)}\boldsymbol{Q}^{0,(l)} - \boldsymbol{W}\right)_{l,\cdot}\right\|_2 + \left\|(\boldsymbol{Z}^{0,(l)}_{l,\cdot})^\top(\boldsymbol{R}^{0,(l)} - \boldsymbol{Q}^{0,(l)})\right\|_2.
\end{aligned} \tag{C.24}$$

First we give a bound of the first term. Note

$$\boldsymbol{W} = \widetilde{\boldsymbol{W}}\boldsymbol{\Sigma}^{1/2} = \widetilde{\boldsymbol{W}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{W}}^\top\widetilde{\boldsymbol{W}}\boldsymbol{\Sigma}^{-1/2} = \overline{\boldsymbol{M}}\widetilde{\boldsymbol{W}}\boldsymbol{\Sigma}^{-1/2},$$

where the last equality holds since

$$
\begin{aligned}
&\overline{\boldsymbol{M}}\widetilde{\boldsymbol{W}}\\
=&\frac{1}{\sqrt{2}}\begin{bmatrix}\widetilde{\boldsymbol{U}} & \widetilde{\boldsymbol{U}}\\ \widetilde{\boldsymbol{V}} & -\widetilde{\boldsymbol{V}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & -\boldsymbol{\Sigma}\end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix}\widetilde{\boldsymbol{U}} & \widetilde{\boldsymbol{U}}\\ \widetilde{\boldsymbol{V}} & -\widetilde{\boldsymbol{V}}\end{bmatrix}^{\top}\frac{1}{\sqrt{2}}\begin{bmatrix}\widetilde{\boldsymbol{U}}\\ \widetilde{\boldsymbol{V}}\end{bmatrix}\\
=&\frac{1}{\sqrt{2}}\begin{bmatrix}\widetilde{\boldsymbol{U}}\\ \widetilde{\boldsymbol{V}}\end{bmatrix}\boldsymbol{\Sigma}\frac{1}{\sqrt{2}}\begin{bmatrix}\widetilde{\boldsymbol{U}}\\ \widetilde{\boldsymbol{V}}\end{bmatrix}^{\top}\frac{1}{\sqrt{2}}\begin{bmatrix}\widetilde{\boldsymbol{U}}\\ \widetilde{\boldsymbol{V}}\end{bmatrix}+\frac{1}{\sqrt{2}}\begin{bmatrix}\widetilde{\boldsymbol{U}}\\ -\widetilde{\boldsymbol{V}}\end{bmatrix}(-\boldsymbol{\Sigma})\frac{1}{\sqrt{2}}\begin{bmatrix}\widetilde{\boldsymbol{U}}\\ -\widetilde{\boldsymbol{V}}\end{bmatrix}^{\top}\frac{1}{\sqrt{2}}\begin{bmatrix}\widetilde{\boldsymbol{U}}\\ \widetilde{\boldsymbol{V}}\end{bmatrix}\\
=&\widetilde{\boldsymbol{W}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{W}}^{\top}\widetilde{\boldsymbol{W}},
\end{aligned}
$$

the last equality uses the fact that $\widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}}=\boldsymbol{I}=\widetilde{\boldsymbol{V}}^{\top}\widetilde{\boldsymbol{V}}$. Similarly, we also have

$$
\boldsymbol{Z}^{0,(l)}=\widetilde{\boldsymbol{Z}}^{0,(l)}(\boldsymbol{\Sigma}^{0,(l)})^{1/2}=\overline{\boldsymbol{M}}^{0,(l)}\widetilde{\boldsymbol{Z}}^{0,(l)}(\boldsymbol{\Sigma}^{0,(l)})^{-1/2}.
$$

By the way we define $\overline{\boldsymbol{M}}^{0,(l)}$ and $\overline{\boldsymbol{M}}$ in (C.14) and (C.10), $\overline{\boldsymbol{M}}^{0,(l)}_{l,\cdot}=\overline{\boldsymbol{M}}_{l,\cdot}$. By triangle inequality we have

$$
\begin{aligned}
&\left\|\left(\boldsymbol{Z}^{0,(l)}\boldsymbol{Q}^{0,(l)}-\boldsymbol{W}\right)_{l,\cdot}\right\|_2\\
=&\left\|\left(\overline{\boldsymbol{M}}^{0,(l)}\widetilde{\boldsymbol{Z}}^{0,(l)}(\boldsymbol{\Sigma}^{0,(l)})^{-1/2}\boldsymbol{Q}^{0,(l)}-\overline{\boldsymbol{M}}\widetilde{\boldsymbol{W}}\boldsymbol{\Sigma}^{-1/2}\right)_{l,\cdot}\right\|_2\\
=&\left\|(\overline{\boldsymbol{M}}_{l,\cdot})^{\top}\left(\widetilde{\boldsymbol{Z}}^{0,(l)}(\boldsymbol{\Sigma}^{0,(l)})^{-1/2}\boldsymbol{Q}^{0,(l)}-\widetilde{\boldsymbol{W}}\boldsymbol{\Sigma}^{-1/2}\right)\right\|_2 \qquad \text{(C.25)}\\
=&\left\|(\overline{\boldsymbol{M}}_{l,\cdot})^{\top}\left(\widetilde{\boldsymbol{Z}}^{0,(l)}\left[(\boldsymbol{\Sigma}^{0,(l)})^{-1/2}\boldsymbol{Q}^{0,(l)}-\boldsymbol{Q}^{0,(l)}\boldsymbol{\Sigma}^{-1/2}\right]+\left[\widetilde{\boldsymbol{Z}}^{0,(l)}\boldsymbol{Q}^{0,(l)}-\widetilde{\boldsymbol{W}}\right]\boldsymbol{\Sigma}^{-1/2}\right)\right\|_2\\
\leqslant&\|\overline{\boldsymbol{M}}_{l,\cdot}\|_2\left(\|(\boldsymbol{\Sigma}^{0,(l)})^{-1/2}\boldsymbol{Q}^{0,(l)}-\boldsymbol{Q}^{0,(l)}\boldsymbol{\Sigma}^{-1/2}\|+\|\widetilde{\boldsymbol{Z}}^{0,(l)}\boldsymbol{Q}^{0,(l)}-\widetilde{\boldsymbol{W}}\|\frac{1}{\sqrt{\sigma_r(\boldsymbol{M})}}\right).
\end{aligned}
$$

By Lemma C.2.2 with $\boldsymbol{M}_1=\overline{\boldsymbol{M}},\boldsymbol{M}_2=\overline{\boldsymbol{M}}^{0,(l)}$, we have

$$
\|\widetilde{\boldsymbol{Z}}^{0,(l)}\boldsymbol{Q}^{0,(l)}-\widetilde{\boldsymbol{W}}\|\leqslant\frac{3}{\sigma_r(\boldsymbol{M})}\|\overline{\boldsymbol{M}}-\overline{\boldsymbol{M}}^{0,(l)}\|. \tag{C.26}
$$

By Lemma C.2.3 with $\boldsymbol{M}_1 = \boldsymbol{M}_3 = \overline{\boldsymbol{M}}, \boldsymbol{M}_2 = \overline{\boldsymbol{M}}^{0,(l)}$, we have

$$
\begin{aligned}
&\|(\boldsymbol{\Sigma}^{0,(l)})^{-1/2}\boldsymbol{Q}^{0,(l)} - \boldsymbol{Q}^{0,(l)}\boldsymbol{\Sigma}^{-1/2}\| \\
=&\left\|(\boldsymbol{\Sigma}^{0,(l)})^{-1/2}\left(\boldsymbol{Q}^{0,(l)}\boldsymbol{\Sigma}^{1/2} - (\boldsymbol{\Sigma}^{0,(l)})^{1/2}\boldsymbol{Q}^{0,(l)}\right)\boldsymbol{\Sigma}^{-1/2}\right\| \\
\leqslant&\|(\boldsymbol{\Sigma}^{0,(l)})^{-1/2}\|\|\boldsymbol{\Sigma}^{-1/2}\|\|\boldsymbol{Q}^{0,(l)}\boldsymbol{\Sigma}^{1/2} - (\boldsymbol{\Sigma}^{0,(l)})^{1/2}\boldsymbol{Q}^{0,(l)}\| \\
\leqslant&\frac{20}{\sigma_r(\boldsymbol{M})}\frac{\kappa}{\sqrt{\sigma_r(\boldsymbol{M})}}\|\overline{\boldsymbol{M}} - \overline{\boldsymbol{M}}^{0,(l)}\|.
\end{aligned}
\tag{C.27}
$$

The last inequality uses the fact that $\sigma_r(\boldsymbol{\Sigma}^{0,(l)}) \geqslant \frac{3}{4}\sigma_r(\boldsymbol{M})$.

Putting estimations (C.26) and (C.27) together and plugging back to (C.25) we have

$$
\begin{aligned}
&\left\|\left(\boldsymbol{Z}^{0,(l)}\boldsymbol{Q}^{0,(l)} - \boldsymbol{W}\right)_{l,\cdot}\right\|_2 \\
\leqslant&\|\overline{\boldsymbol{M}}_{l,\cdot}\|_2\frac{23\kappa}{\sqrt{\sigma_r(\boldsymbol{M})}^3}\|\overline{\boldsymbol{M}} - \overline{\boldsymbol{M}}^{0,(l)}\| \\
\leqslant&\max(\|\boldsymbol{U}\|\|\boldsymbol{V}\|_{2,\infty}, \|\boldsymbol{V}\|\|\boldsymbol{U}\|_{2,\infty}) \times \frac{92C_{14}\kappa}{\sqrt{\sigma_r(\boldsymbol{M})}^3}\sqrt{\frac{\mu r\kappa\log(n_1 \vee n_2)}{(n_1 \wedge n_2)p}}\sigma_1(\boldsymbol{M}) \\
\leqslant&92C_{14}\sqrt{\frac{\mu^2 r^2\kappa^7\log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\sqrt{\sigma_1(\boldsymbol{M})}.
\end{aligned}
\tag{C.28}
$$

In order to control the second term in (C.24), note from (C.28),

$$
\begin{aligned}
&\|\boldsymbol{Z}^{0,(l)}_{l,\cdot}\|_2 \\
\leqslant&\|\boldsymbol{W}_{l,\cdot}\|_2 + \left\|\left(\boldsymbol{Z}^{0,(l)}\boldsymbol{Q}^{0,(l)} - \boldsymbol{W}\right)_{l,\cdot}\right\|_2 \\
\leqslant&\|\boldsymbol{W}\|_{2,\infty} + \left\|\left(\boldsymbol{Z}^{0,(l)}\boldsymbol{Q}^{0,(l)} - \boldsymbol{W}\right)_{l,\cdot}\right\|_2 \\
\leqslant&\left(\sqrt{\frac{\mu r\kappa}{n_1 \wedge n_2}} + 92C_{14}\sqrt{\frac{\mu^2 r^2\kappa^7\log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\right)\sqrt{\sigma_1(\boldsymbol{M})}.
\end{aligned}
$$

Then by Lemma C.2.4 with $\boldsymbol{M}_1 = \overline{\boldsymbol{M}}, \boldsymbol{M}_2 = \overline{\boldsymbol{M}}^{0,(l)}$, we have

$$\begin{aligned}
&\left\|(\boldsymbol{Z}_{l,\cdot}^{0,(l)})^\top(\boldsymbol{R}^{0,(l)} - \boldsymbol{Q}^{0,(l)})\right\|_2 \\
\leqslant& \|\boldsymbol{Z}_{l,\cdot}^{0,(l)}\|_2\|\boldsymbol{R}^{0,(l)} - \boldsymbol{Q}^{0,(l)}\| \\
\leqslant& \|\boldsymbol{Z}_{l,\cdot}^{0,(l)}\|_2 15\frac{\sqrt{\kappa}^3}{\sigma_r(\boldsymbol{M})}\|\overline{\boldsymbol{M}} - \overline{\boldsymbol{M}}^{0,(l)}\| \\
\leqslant& 60C_{14}\frac{\sqrt{\kappa}^3}{\sigma_r(\boldsymbol{M})}\sqrt{\frac{\mu r\kappa\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sigma_1(\boldsymbol{M}) \times \left(\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}} + 92C_{14}\sqrt{\frac{\mu^2r^2\kappa^7\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\right)\sqrt{\sigma_1(\boldsymbol{M})}.
\end{aligned}$$

So as long as we have

$$p \geqslant 92^2C_{14}^2\frac{\mu r\kappa^6\log(n_1\vee n_2)}{n_1\wedge n_2},$$

then

$$\left\|(\boldsymbol{Z}_{l,\cdot}^{0,(l)})^\top(\boldsymbol{R}^{0,(l)} - \boldsymbol{Q}^{0,(l)})\right\|_2 \leqslant 120C_{14}\sqrt{\frac{\mu^2r^2\kappa^7\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})}. \tag{C.29}$$

Putting estimation (C.28) and (C.29) together we have

$$\begin{aligned}
&\left\|\left(\begin{bmatrix}\boldsymbol{X}^{0,(l)}\\ \boldsymbol{Y}^{0,(l)}\end{bmatrix}\boldsymbol{R}^{0,(l)} - \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right)_{l,\cdot}\right\|_2 \\
=&\sqrt{2}\left\|\left(\boldsymbol{Z}^{0,(l)}\boldsymbol{R}^{0,(l)} - \boldsymbol{W}\right)_{l,\cdot}\right\|_2 \\
\leqslant& 212\sqrt{2}C_{14}\sqrt{\frac{\mu^2r^2\kappa^7\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})}.
\end{aligned} \tag{C.30}$$

C.2.2.3. *Proof for* (4.18). Finally, we want to give a bound for

$$\left\|\begin{bmatrix}\boldsymbol{X}^{0}\\ \boldsymbol{Y}^{0}\end{bmatrix}\boldsymbol{R}^{0} - \begin{bmatrix}\boldsymbol{X}^{0,(l)}\\ \boldsymbol{Y}^{0,(l)}\end{bmatrix}\boldsymbol{T}^{0,(l)}\right\|_F.$$

Without loss of generality, assume that l satisfies $1 \leqslant l \leqslant n_1$. First denote

$$\boldsymbol{B} := \underset{\boldsymbol{R}\in\mathsf{O}(r)}{\operatorname{argmin}}\|\widetilde{\boldsymbol{Z}}^{0,(l)}\boldsymbol{R} - \widetilde{\boldsymbol{Z}}^{0}\|_F.$$

From the choice of $\boldsymbol{T}^{0,(l)}$ in (4.15), we have

$$\left\|\boldsymbol{Z}^0\boldsymbol{R}^0 - \boldsymbol{Z}^{0,(l)}\boldsymbol{T}^{0,(l)}\right\|_F \leqslant \|\boldsymbol{Z}^{0,(l)}\boldsymbol{B} - \boldsymbol{Z}^0\|_F. \tag{C.31}$$

By triangle inequality,

$$\begin{aligned}
&\|\boldsymbol{Z}^{0,(l)}\boldsymbol{B} - \boldsymbol{Z}^0\|_F \\
=&\left\|\widetilde{\boldsymbol{Z}}^{0,(l)}(\boldsymbol{\Sigma}^{0,(l)})^{1/2}\boldsymbol{B} - \widetilde{\boldsymbol{Z}}^0(\boldsymbol{\Sigma}^0)^{1/2}\right\|_F \\
=&\left\|\widetilde{\boldsymbol{Z}}^{0,(l)}\left[(\boldsymbol{\Sigma}^{0,(l)})^{1/2}\boldsymbol{B} - \boldsymbol{B}(\boldsymbol{\Sigma}^0)^{1/2}\right] + (\widetilde{\boldsymbol{Z}}^{0,(l)}\boldsymbol{B} - \widetilde{\boldsymbol{Z}}^0)(\boldsymbol{\Sigma}^0)^{1/2}\right\|_F \\
\leqslant&\left\|\widetilde{\boldsymbol{Z}}^{0,(l)}\left[(\boldsymbol{\Sigma}^{0,(l)})^{1/2}\boldsymbol{B} - \boldsymbol{B}(\boldsymbol{\Sigma}^0)^{1/2}\right]\right\|_F + \left\|(\widetilde{\boldsymbol{Z}}^{0,(l}\boldsymbol{B} - \widetilde{\boldsymbol{Z}}^0)(\boldsymbol{\Sigma}^0)^{1/2}\right\|_F \\
\leqslant&\left\|(\boldsymbol{\Sigma}^{0,(l)})^{1/2}\boldsymbol{B} - \boldsymbol{B}(\boldsymbol{\Sigma}^0)^{1/2}\right\|_F + \left\|\widetilde{\boldsymbol{Z}}^{0,(l)}\boldsymbol{B} - \widetilde{\boldsymbol{Z}}^0\right\|_F \|(\boldsymbol{\Sigma}^0)^{1/2}\|.
\end{aligned} \tag{C.32}$$

By Lemma C.2.3 with $\boldsymbol{M}_1 = \overline{\boldsymbol{M}}, \boldsymbol{M}_2 = \overline{\boldsymbol{M}}^{0,(l)}, \boldsymbol{M}_3 = \overline{\boldsymbol{M}}^0$, we have

$$\left\|(\boldsymbol{\Sigma}^{0,(l)})^{1/2}\boldsymbol{B} - \boldsymbol{B}(\boldsymbol{\Sigma}^0)^{1/2}\right\|_F \leqslant 15\frac{\kappa}{\sqrt{\sigma_r(\boldsymbol{M})}}\left\|\left(\overline{\boldsymbol{M}}^0 - \overline{\boldsymbol{M}}^{0,(l)}\right)\widetilde{\boldsymbol{Z}}^{0,(l)}\right\|_F. \tag{C.33}$$

Moreover, by Davis-Kahan SinΘ theorem [DK70], we have

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{Z}}^{0,(l)}\boldsymbol{B} - \widetilde{\boldsymbol{Z}}^0\right\|_F \leqslant&\sqrt{2}\left\|\left(\boldsymbol{I} - \widetilde{\boldsymbol{Z}}^0(\widetilde{\boldsymbol{Z}}^0)^\top\right)\widetilde{\boldsymbol{Z}}^{0,(l)}\right\|_F \\
\leqslant&\frac{2\sqrt{2}}{\sigma_r(\boldsymbol{M})}\left\|\left(\overline{\boldsymbol{M}}^0 - \overline{\boldsymbol{M}}^{0,(l)}\right)\widetilde{\boldsymbol{Z}}^{0,(l)}\right\|_F.
\end{aligned} \tag{C.34}$$

So putting the estimations (C.32), (C.33) and (C.34) together we have

$$\begin{aligned}
&\|\boldsymbol{Z}^{0,(l)}\boldsymbol{B} - \boldsymbol{Z}^0\|_F \\
\leqslant&15\frac{\kappa}{\sqrt{\sigma_r(\boldsymbol{M})}}\left\|\left(\overline{\boldsymbol{M}}^0 - \overline{\boldsymbol{M}}^{0,(l)}\right)\widetilde{\boldsymbol{Z}}^{0,(l)}\right\|_F + 4\frac{\sqrt{\kappa}}{\sqrt{\sigma_r(\boldsymbol{M})}}\left\|\left(\overline{\boldsymbol{M}}^0 - \overline{\boldsymbol{M}}^{0,(l)}\right)\widetilde{\boldsymbol{Z}}^{0,(l)}\right\|_F \\
\leqslant&20\frac{\kappa}{\sqrt{\sigma_r(\boldsymbol{M})}}\left\|\left(\overline{\boldsymbol{M}}^0 - \overline{\boldsymbol{M}}^{0,(l)}\right)\widetilde{\boldsymbol{Z}}^{0,(l)}\right\|_F.
\end{aligned} \tag{C.35}$$

By the way we define $\overline{\boldsymbol{M}}^0$ and $\overline{\boldsymbol{M}}^{0,(l)}$ in (C.13) and (C.14),

$$\left(\overline{\boldsymbol{M}}^0 - \overline{\boldsymbol{M}}^{0,(l)}\right)\widetilde{\boldsymbol{Z}}^{0,(l)} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \sum_j \left(\frac{1}{p}\delta_{l,j} - 1\right)\overline{M}_{l,n_1+j}(\widetilde{\boldsymbol{Z}}^{0,(l)}_{n_1+j,\cdot})^\top \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \left(\frac{1}{p}\delta_{l,1} - 1\right)\overline{M}_{n_1+1,l}(\widetilde{\boldsymbol{Z}}^{0,(l)}_{l,\cdot})^\top \\ \vdots \\ \left(\frac{1}{p}\delta_{l,j} - 1\right)\overline{M}_{n_1+j,l}(\widetilde{\boldsymbol{Z}}^{0,(l)}_{l,\cdot})^\top \\ \vdots \\ \left(\frac{1}{p}\delta_{l,n_2} - 1\right)\overline{M}_{n_1+n_2,l}(\widetilde{\boldsymbol{Z}}^{0,(l)}_{l,\cdot})^\top \end{bmatrix}.$$

Recall that here we assume $1 \leqslant l \leqslant n_1$. Therefore by triangle inequality,

$$\begin{aligned} &\left\|\left(\overline{\boldsymbol{M}}^0 - \overline{\boldsymbol{M}}^{0,(l)}\right)\widetilde{\boldsymbol{Z}}^{0,(l)}\right\|_F \\ &\leqslant \left\|\sum_j \left(\frac{1}{p}\delta_{l,j} - 1\right)\overline{M}_{l,n_1+j}\widetilde{\boldsymbol{Z}}^{0,(l)}_{n_1+j,\cdot}\right\|_2 + \left\|\begin{bmatrix} \left(\frac{1}{p}\delta_{l,1} - 1\right)\overline{M}_{n_1+1,l}(\widetilde{\boldsymbol{Z}}^{0,(l)}_{l,\cdot})^\top \\ \vdots \\ \left(\frac{1}{p}\delta_{l,j} - 1\right)\overline{M}_{n_1+j,l}(\widetilde{\boldsymbol{Z}}^{0,(l)}_{l,\cdot})^\top \\ \vdots \\ \left(\frac{1}{p}\delta_{l,n_2} - 1\right)\overline{M}_{n_1+n_2,l}(\widetilde{\boldsymbol{Z}}^{0,(l)}_{l,\cdot})^\top \end{bmatrix}\right\|_F \\ &= \left\|\underbrace{\sum_j \left(\frac{1}{p}\delta_{l,j} - 1\right)\overline{M}_{l,n_1+j}\widetilde{\boldsymbol{Z}}^{0,(l)}_{n_1+j,\cdot}}_{\boldsymbol{a}_1}\right\|_2 + \left\|\underbrace{\begin{bmatrix} \left(\frac{1}{p}\delta_{l,1} - 1\right)\overline{M}_{n_1+1,l} \\ \vdots \\ \left(\frac{1}{p}\delta_{l,j} - 1\right)\overline{M}_{n_1+j,l} \\ \vdots \\ \left(\frac{1}{p}\delta_{l,n_2} - 1\right)\overline{M}_{n_1+n_2,l} \end{bmatrix}}_{\boldsymbol{a}_2}\right\|_2 \|\widetilde{\boldsymbol{Z}}^{0,(l)}_{l,\cdot}\|_2. \end{aligned} \tag{C.36}$$

Note by (C.18) and the fact that $\boldsymbol{M}^{0,(l)}$ has top-r singular value decomposition

$$\widetilde{\boldsymbol{X}}^{0,(l)}\boldsymbol{\Sigma}^{0,(l)}(\widetilde{\boldsymbol{Y}}^{0,(l)})^\top,$$

$\widetilde{\boldsymbol{Z}}^{0,(l)}_{n_1+j,\cdot}$ is independent of $\delta_{l,j}$'s. For $\boldsymbol{a}_1$,

$$\boldsymbol{a}_1 = \sum_j \left(\frac{1}{p}\delta_{l,j} - 1\right) \overline{M}_{l,n_1+j}\widetilde{\boldsymbol{Z}}^{0,(l)}_{n_1+j,\cdot} := \sum_j \boldsymbol{s}_{1,j}.$$

Conditioned on $\widetilde{\boldsymbol{Z}}^{0,(l)}_{n_1+j,\cdot}$, $\boldsymbol{s}_{1,j}$'s are independent, and $\mathbb{E}_{\delta_{l,\cdot}}\boldsymbol{s}_{1,j} = \boldsymbol{0}$. We also have

$$\begin{aligned}\|\boldsymbol{s}_{1,j}\|_2 \leqslant & \frac{1}{p}\|\overline{\boldsymbol{M}}\|_{\ell_\infty}\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty} \\ \leqslant & \frac{1}{p}\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty},\end{aligned}$$

and

$$\begin{aligned}&\left\|\mathbb{E}_{\delta_{l,\cdot}}\sum_j \boldsymbol{s}_{1,j}^\top \boldsymbol{s}_{1,j}\right\| \\ =&\sum_j \mathbb{E}_{\delta_{l,\cdot}}\left(\frac{1}{p}\delta_{l,j} - 1\right)^2 \overline{M}^2_{l,n_1+j}\|\widetilde{\boldsymbol{Z}}^{0,(l)}_{n_1+j,\cdot}\|_2^2 \\ \leqslant&\frac{1}{p}\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty}^2\|\overline{\boldsymbol{M}}_{l,\cdot}\|_2^2 \\ \leqslant&\frac{1}{p}\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty}^2 \max\left(\|\boldsymbol{U}\|\|\boldsymbol{V}\|_{2,\infty}, \|\boldsymbol{V}\|\|\boldsymbol{U}\|_{2,\infty}\right)^2.\end{aligned}$$

For $\left\|\sum_j \mathbb{E}_{\delta_{l,\cdot}}\boldsymbol{s}_{1,j}\boldsymbol{s}_{1,j}^\top\right\|$ we have the same bound. Then by matrix Bernstein inequality [Tro15, Theorem 6.1.1],

$$\begin{aligned}&\mathbb{P}\left[\|\boldsymbol{a}_1\|_2 \geqslant 100\left(\sqrt{\frac{\log(n_1\vee n_2)}{p}}\max\left(\|\boldsymbol{U}\|\|\boldsymbol{V}\|_{2,\infty}, \|\boldsymbol{V}\|\|\boldsymbol{U}\|_{2,\infty}\right) + \frac{\log(n_1 \vee n_2)}{p}\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}\right)\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty} \mid \widetilde{\boldsymbol{Z}}^{0,(l)}\right] \\ \leqslant&(n_1+n_2)^{-15}.\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left[\|\boldsymbol{a}_1\|_2 \geqslant 100\left(\sqrt{\frac{\log(n_1 \vee n_2)}{p}}\max\left(\|\boldsymbol{U}\|\|\boldsymbol{V}\|_{2,\infty}, \|\boldsymbol{V}\|\|\boldsymbol{U}\|_{2,\infty}\right) + \frac{\log(n_1 \vee n_2)}{p}\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}\right)\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty}\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\|\boldsymbol{a}_1\|_2\geqslant 100\left(\sqrt{\frac{\log(n_1\vee n_2)}{p}}\max\left(\|\boldsymbol{U}\|\|\boldsymbol{V}\|_{2,\infty},\|\boldsymbol{V}\|\|\boldsymbol{U}\|_{2,\infty}\right)+\frac{\log(n_1\vee n_2)}{p}\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}\right)\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty}} \mid \widetilde{\boldsymbol{Z}}^{0,(l)}\right]\right] \\
\leqslant&(n_1+n_2)^{-15}.
\end{aligned}$$

In other words, on an event $E_B^{0,(l),1}$ with probability $\mathbb{P}[E_B^{0,(l),1}] \geqslant 1-(n_1+n_2)^{-15}$, we have

$$\begin{aligned}
&\|\boldsymbol{a}_1\|_2 \\
\leqslant&100\sqrt{\frac{\log(n_1 \vee n_2)}{p}}\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty}\max\left(\|\boldsymbol{U}\|\|\boldsymbol{V}\|_{2,\infty}, \|\boldsymbol{V}\|\|\boldsymbol{U}\|_{2,\infty}\right) \\
&+100\frac{\log(n_1 \vee n_2)}{p}\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty} \\
\leqslant&100\left(\sqrt{\frac{\mu r\kappa\log(n_1\vee n_2)}{(n_1\wedge n_2)p}} + \frac{\mu r\kappa\log(n_1\vee n_2)}{(n_1\wedge n_2)p}\right)\times\sigma_1(\boldsymbol{M})\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty}.
\end{aligned} \tag{C.37}$$

For $\boldsymbol{a}_2$, we can decompose it as

$$\begin{aligned}
\boldsymbol{a}_2 = &\begin{bmatrix} \left(\frac{1}{p}\delta_{l,1}-1\right)\overline{M}_{n_1+1,l} \\ \vdots \\ \left(\frac{1}{p}\delta_{l,j}-1\right)\overline{M}_{n_1+j,l} \\ \vdots \\ \left(\frac{1}{p}\delta_{l,n_2}-1\right)\overline{M}_{n_1+n_2,l} \end{bmatrix} \\
=&\sum_j\left(\frac{1}{p}\delta_{l,j}-1\right)\overline{M}_{n_1+j,l}\boldsymbol{e}_j \\
=&\sum_j \boldsymbol{s}_{2,j}.
\end{aligned}$$

Then we have $\mathbb{E}\boldsymbol{s}_{2,j}=\boldsymbol{0}$,

$$\|\boldsymbol{s}_{2,j}\|_2 \leqslant \frac{1}{p}\|\overline{\boldsymbol{M}}\|_{\ell_\infty} \leqslant \frac{1}{p}\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}$$

and

$$
\begin{aligned}
\|\mathbb{E}\sum_j \boldsymbol{s}_{2,j}\boldsymbol{s}_{2,j}^\top\| \leqslant & \sum_j \mathbb{E}\|\boldsymbol{s}_{2,j}\|_2^2 \\
= & \sum_j \mathbb{E}\left(\frac{1}{p}\delta_{l,j}-1\right)^2 \overline{M}_{n_1+j,l}^2 \\
\leqslant & \sum_j \frac{1}{p}\overline{M}_{n_1+j,l}^2 \\
\leqslant & \frac{1}{p}\max\left(\|\boldsymbol{U}\|\|\boldsymbol{V}\|_{2,\infty}, \|\boldsymbol{V}\|\|\boldsymbol{U}\|_{2,\infty}\right)^2.
\end{aligned}
$$

Therefore by matrix Bernstein inequality [Tro15, Theorem 6.1.1] again, on an event $E_B^{0,2}$ with probability $\mathbb{P}[E_B^{0,2}] \geqslant 1-(n_1+n_2)^{-15}$, we have

$$
\|\boldsymbol{a}_2\|_2 \leqslant 100\left(\sqrt{\frac{\mu r\kappa\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}+\frac{\mu r\kappa\log(n_1\vee n_2)}{(n_1\wedge n_2)p}\right)\sigma_1(\boldsymbol{M}). \tag{C.38}
$$

So putting (C.36), (C.37) and (C.38) together we have

$$
\begin{aligned}
&\left\|\left(\overline{\boldsymbol{M}}^0-\overline{\boldsymbol{M}}^{0,(l)}\right)\widetilde{\boldsymbol{Z}}^{0,(l)}\right\|_F \\
\leqslant & \|\boldsymbol{a}_1\|_2+\|\boldsymbol{a}_2\|_2\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty} \\
\leqslant & 200\left(\sqrt{\frac{\mu r\kappa\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}+\frac{\mu r\kappa\log(n_1\vee n_2)}{(n_1\wedge n_2)p}\right)\sigma_1(\boldsymbol{M})\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty}
\end{aligned} \tag{C.39}
$$

on an event $E_B^0=\left(\bigcap_{l=1}^{n_1+n_2}E_B^{0,(l),1}\right)\bigcap E_B^{0,2}$. Moreover, by applying union bound we have $\mathbb{P}[E_B^0] \geqslant 1-(n_1+n_2)^{-11}$.

Now we need to bound $\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty}$. We have the following claim:

Claim C.2.5. *Under the setup of Lemma 4.2.2, on an event E_{Claim} with probability $\mathbb{P}[E_{Claim}] \geqslant 1-3(n_1+n_2)^{-11}$, the following inequality*

$$
\begin{aligned}
\|\widetilde{\boldsymbol{Z}}^{0,(l)}\|_{2,\infty} \leqslant & (4+4\kappa+9C_{15}\kappa^2)\|\widetilde{\boldsymbol{W}}\|_{2,\infty} \\
\leqslant & (8+9C_{15})\kappa^2\frac{1}{\sqrt{\sigma_r(\boldsymbol{M})}}\|\boldsymbol{W}\|_{2,\infty}
\end{aligned} \tag{C.40}
$$

holds with the absolute constant C_{15} defined in Lemma C.2.6.

If the claim is true, from (C.31), (C.35), (C.39) and (C.40) and if

$$p \geqslant \frac{\mu r \kappa \log(n_1 \vee n_2)}{n_1 \wedge n_2},$$

then

$$\begin{aligned}
&\left\| \begin{bmatrix} \boldsymbol{X}^0 \\ \boldsymbol{Y}^0 \end{bmatrix} \boldsymbol{R}^0 - \begin{bmatrix} \boldsymbol{X}^{0,(l)} \\ \boldsymbol{Y}^{0,(l)} \end{bmatrix} \boldsymbol{T}^{0,(l)} \right\|_F \\
=&\sqrt{2} \left\| \boldsymbol{Z}^0 \boldsymbol{R}^0 - \boldsymbol{Z}^{0,(l)} \boldsymbol{T}^{0,(l)} \right\|_F \\
\leqslant& 20\sqrt{2} \frac{\kappa}{\sqrt{\sigma_r(\boldsymbol{M})}} \left\| \left(\overline{\boldsymbol{M}}^0 - \overline{\boldsymbol{M}}^{0,(l)} \right) \widetilde{\boldsymbol{Z}}^{0,(l)} \right\|_F \\
\leqslant& (64000\sqrt{2} + 72000\sqrt{2} C_{15}) \sqrt{\frac{\mu^2 r^2 \kappa^{10} \log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})}
\end{aligned} \tag{C.41}$$

holds for any l satisfying $1 \leqslant l \leqslant n_1$. For the case $n_1 + 1 \leqslant l \leqslant n_1 + n_2$, we can use the same argument.

Note on an event

$$\begin{aligned}
E_{init} &= E_{Ch1} \bigcap E_{Claim} \bigcap E_H \bigcap E_B^0 \\
&= E_S \bigcap E_{Ca} \bigcap E_Z \bigcap E_{Ch1} \bigcap E_{Ch2} \bigcap E_A \bigcap E_B^0,
\end{aligned}$$

(C.23), (C.30) and (C.41) hold. Choosing C_I to be

$$C_I = 64000\sqrt{2} + 212\sqrt{2} C_{14} + 72000\sqrt{2} C_{15}$$

and C_{S2} to be

$$C_{S2} = 256 + 25600 C_{14}^2 + C_{15},$$

using union bound $\mathbb{P}[E_{init}] \geqslant 1 - 7(n_1+n_2)^{-11} \geqslant 1 - (n_1+n_2)^{-10}$, which finishes the proof.

Proof of Claim C.2.5. Follow the way people did in [**MWCC18**], let $\overline{\boldsymbol{M}}^{0,(l),\text{zero}}$ be the matrix derived by zeroing out the l-th row and column of $\overline{\boldsymbol{M}}^{0,(l)}$, and $\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}} \in \mathbb{R}^{(n_1+n_2)\times r}$

containing the leading r eigenvectors of $\overline{\boldsymbol{M}}^{0,(l),\text{zero}}$. Notice

$$
\begin{aligned}
&\left\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\operatorname{sgn}\left((\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top\widetilde{\boldsymbol{W}}\right)-\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}(\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top\widetilde{\boldsymbol{W}}\right\|_{2,\infty} \\
&=\left\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}(\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top\widetilde{\boldsymbol{W}}\left((\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top\widetilde{\boldsymbol{W}}\right)^{-1}\cdot\left(\operatorname{sgn}\left((\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top\widetilde{\boldsymbol{W}}\right)-(\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top\widetilde{\boldsymbol{W}}\right)\right\|_{2,\infty} \\
&\leqslant\left\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}(\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top\widetilde{\boldsymbol{W}}\right\|_{2,\infty}\left\|\left((\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top\widetilde{\boldsymbol{W}}\right)^{-1}\right\|\left\|\operatorname{sgn}\left((\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top\widetilde{\boldsymbol{W}}\right)-(\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top\widetilde{\boldsymbol{W}}\right\|.
\end{aligned}
\tag{C.42}
$$

By triangle inequality,

$$
\begin{aligned}
&\left\|\overline{\boldsymbol{M}}^{0,(l),\text{zero}}-\overline{\boldsymbol{M}}\right\| \\
&\leqslant\left\|\overline{\boldsymbol{M}}^{0,(l),\text{zero}}-\overline{\boldsymbol{M}}^{(l),\text{zero}}\right\|+\left\|\begin{bmatrix}\boldsymbol{0} & & \overline{M}_{1,l} & & \boldsymbol{0} \\ & & \vdots & & \\ \overline{M}_{l,1} & \cdots & \overline{M}_{l,l} & \cdots & \overline{M}_{l,n_1+n_2} \\ & & \vdots & & \\ \boldsymbol{0} & & \overline{M}_{n_1+n_2,l} & & \boldsymbol{0}\end{bmatrix}\right\|,
\end{aligned}
\tag{C.43}
$$

where here we define $\overline{\boldsymbol{M}}^{(l),\text{zero}}$ as $\overline{\boldsymbol{M}}$ zeroing out the l-th row and column of $\overline{\boldsymbol{M}}$. The first part we can again apply Lemma 4.2.3 on $\overline{\boldsymbol{M}}^{(l),\text{zero}}$ to see

$$\left\|\overline{\boldsymbol{M}}^{0,(l),\text{zero}}-\overline{\boldsymbol{M}}^{(l),\text{zero}}\right\|\leqslant 4C_{14}\sqrt{\frac{\mu r\kappa\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sigma_1(\boldsymbol{M})$$

holds on an event E_{Ch2} with probability $\mathbb{P}[E_{Ch2}]\geqslant 1-(n_1+n_2)^{-11}$. Therefore since

$$p\geqslant 1024C_{14}^2\frac{\mu r\kappa^3\log(n_1\vee n_2)}{n_1\wedge n_2},$$

we have

$$\left\|\overline{\boldsymbol{M}}^{0,(l),\text{zero}}-\overline{\boldsymbol{M}}^{(l),\text{zero}}\right\|\leqslant\frac{1}{8}\sigma_r(\boldsymbol{M}). \tag{C.44}$$

Moreover, for the second part of the right hand side of (C.43), we have

$$
\begin{aligned}
&\left\|\begin{bmatrix} \mathbf{0} & & \overline{M}_{1,l} & & \mathbf{0} \\ & & \vdots & & \\ \overline{M}_{l,1} & \cdots & \overline{M}_{l,l} & \cdots & \overline{M}_{l,n_1+n_2} \\ & & \vdots & & \\ \mathbf{0} & & \overline{M}_{n_1+n_2,l} & & \mathbf{0} \end{bmatrix}\right\| \\
\leqslant &\left\|\begin{bmatrix} \overline{M}_{l,1} & \cdots & \overline{M}_{l,l} & \cdots & \overline{M}_{l,n_1+n_2} \end{bmatrix}\right\| + \left\|\begin{bmatrix} \overline{M}_{1,l} \\ \vdots \\ \overline{M}_{l-1,l} \\ 0 \\ \overline{M}_{l+1,l} \\ \vdots \\ \overline{M}_{n_1+n_2,l} \end{bmatrix}\right\| \tag{C.45} \\
\leqslant &\|\overline{\boldsymbol{M}}_{l,\cdot}\|_2 + \|\overline{\boldsymbol{M}}_{\cdot,l}\|_2 \\
\leqslant &2\max\{\|\boldsymbol{U}\|\|\boldsymbol{V}\|_{2,\infty}, \|\boldsymbol{V}\|\|\boldsymbol{U}\|_{2,\infty}\} \\
\leqslant &2\sqrt{\frac{\mu r\kappa}{n_1 \wedge n_2}}\sigma_1(\boldsymbol{M}).
\end{aligned}
$$

As long as

$$
256\frac{\mu r\kappa^3}{n_1 \wedge n_2} \leqslant p \leqslant 1,
$$

plugging back to (C.45) we have

$$
\left\|\begin{bmatrix} \mathbf{0} & & \overline{M}_{1,l} & & \mathbf{0} \\ & & \vdots & & \\ \overline{M}_{l,1} & \cdots & \overline{M}_{l,l} & \cdots & \overline{M}_{l,n_1+n_2} \\ & & \vdots & & \\ \mathbf{0} & & \overline{M}_{n_1+n_2,l} & & \mathbf{0} \end{bmatrix}\right\| \leqslant \frac{1}{8}\sigma_r(\boldsymbol{M}). \tag{C.46}
$$

Combining the estimation (C.44) and (C.46) together we have

$$\left\|\overline{\boldsymbol{M}}^{0,(l),\text{zero}} - \overline{\boldsymbol{M}}\right\| \leqslant \frac{1}{4}\sigma_r(\boldsymbol{M}). \tag{C.47}$$

Applying Lemma C.2.1 here, we have

$$\left\|\left((\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right)^{-1}\right\| \leqslant 2$$

and

$$\left\|\operatorname{sgn}\left((\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right) - (\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right\| \leqslant \frac{1}{4}.$$

Therefore from (C.42) we have

$$\left\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}} \operatorname{sgn}\left((\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right) - \widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}(\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right\|_{2,\infty} \leqslant \frac{1}{2}\left\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}(\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right\|_{2,\infty}$$

and

$$\begin{aligned}
&\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\|_{2,\infty} \\
=&\left\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}} \operatorname{sgn}\left((\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right)\right\|_{2,\infty} \\
\leqslant&\left\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}(\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right\|_{2,\infty} \\
&+\left\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}} \operatorname{sgn}\left((\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right) - \widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}(\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right\|_{2,\infty} \\
\leqslant&2\left\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}(\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right\|_{2,\infty}.
\end{aligned}$$

In order to give a control of

$$\left\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}(\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right\|_{2,\infty},$$

we need Lemma 4 and Lemma 14 in [**AFWZ17**]. For the purpose of simplicity we combine those two lemmas together and only include those useful bounds in our case:

Lemma C.2.6 ([**AFWZ17**, Lemma 4 and Lemma 14 rewrited]). *Under our setup, there is some absolute constant C_{15}, if $p \geqslant C_{15}\frac{\mu^2 r^2 \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2)}$, then on an event E_A with probability $\mathbb{P}[E_A] \geqslant$*

$1-(n_1+n_2)^{-11}$,

$$\max_l \|\widetilde{\boldsymbol{Z}}^{0,(l),zero}(\widetilde{\boldsymbol{Z}}^{0,(l),zero})^\top \widetilde{\boldsymbol{W}} - \widetilde{\boldsymbol{W}}\|_{2,\infty} \leqslant 4\kappa\|\widetilde{\boldsymbol{Z}}^0(\widetilde{\boldsymbol{Z}}^0)^\top \widetilde{\boldsymbol{W}}\|_{2,\infty} + \|\widetilde{\boldsymbol{W}}\|_{2,\infty}$$

and

$$\|\widetilde{\boldsymbol{Z}}^0\|_{2,\infty} \leqslant C_{15}\left(\kappa\|\widetilde{\boldsymbol{W}}\|_{2,\infty} + \sqrt{\frac{n_1 \wedge n_2}{p}}\frac{\|\overline{\boldsymbol{M}}\|_{\ell_\infty}\|\overline{\boldsymbol{M}}\|_{2,\infty}}{\sigma_r^2(\boldsymbol{M})}\right)$$

holds.

By the lemma we have

$$\begin{aligned}\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\|_{2,\infty} \leqslant& 2\left\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}(\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}})^\top \widetilde{\boldsymbol{W}}\right\|_{2,\infty}\\ \leqslant& 4\|\widetilde{\boldsymbol{W}}\|_{2,\infty} + 8\kappa\|\widetilde{\boldsymbol{Z}}^0(\widetilde{\boldsymbol{Z}}^0)^\top \widetilde{\boldsymbol{W}}\|_{2,\infty}\\ \leqslant& 4\|\widetilde{\boldsymbol{W}}\|_{2,\infty} + 8\kappa\|\widetilde{\boldsymbol{Z}}^0\|_{2,\infty}\|(\widetilde{\boldsymbol{Z}}^0)^\top \widetilde{\boldsymbol{W}}\|\\ \leqslant& 4\|\widetilde{\boldsymbol{W}}\|_{2,\infty} + 8\kappa\|\widetilde{\boldsymbol{Z}}^0\|_{2,\infty}\\ \leqslant& \left(4 + 8C_{15}\kappa^2 + 8\sqrt{2}C_{15}\sqrt{\frac{\mu^2 r^2\kappa^6}{(n_1\wedge n_2)p}}\right)\|\widetilde{\boldsymbol{W}}\|_{2,\infty}.\end{aligned}$$

The fourth inequality uses the fact that $\|(\widetilde{\boldsymbol{Z}}^0)^\top\widetilde{\boldsymbol{W}}\| \leqslant 1$ since $\widetilde{\boldsymbol{Z}}^0$ and $\widetilde{\boldsymbol{W}}$ both have orthonormal columns, and the last inequality uses the fact that

$$\begin{aligned}\|\overline{\boldsymbol{M}}\|_{2,\infty} \leqslant& \max(\|\boldsymbol{U}\|\|\boldsymbol{V}\|_{2,\infty}, \|\boldsymbol{V}\|\|\boldsymbol{U}\|_{2,\infty})\\ \leqslant& \sqrt{\sigma_1(\boldsymbol{M})}\sqrt{2}\|\boldsymbol{W}\|_{2,\infty}\\ \leqslant& \sqrt{2}\sigma_1(\boldsymbol{M})\|\widetilde{\boldsymbol{W}}\|_{2,\infty}.\end{aligned}$$

So as long as

$$p \geqslant 128\frac{\mu^2 r^2\kappa^2}{n_1\wedge n_2},$$

we have

$$\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\|_{2,\infty} \leqslant (4 + 9C_{15}\kappa^2)\|\widetilde{\boldsymbol{W}}\|_{2,\infty}. \tag{C.48}$$

Recall that in (C.47) and (C.15), we have already shown

$$\left\|\overline{\boldsymbol{M}}^{0,(l),\text{zero}}-\overline{\boldsymbol{M}}\right\| \leqslant \frac{1}{4}\sigma_r(\boldsymbol{M})$$

and

$$\left\|\overline{\boldsymbol{M}}^{0,(l)}-\overline{\boldsymbol{M}}\right\| \leqslant \frac{1}{4}\sigma_r(\boldsymbol{M})$$

hold on the events E_{Ch2} and E_{Ch1}, respectively. Therefore, by the Davis-Kahan SinΘ theorem [**DK70**], we have

$$\begin{aligned}&\left\|\widetilde{\boldsymbol{Z}}^{0,(l)}\operatorname{sgn}\left((\widetilde{\boldsymbol{Z}}^{0,(l)})^\top\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\right)-\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\right\|_F\\ \leqslant&\frac{2\sqrt{2}}{\sigma_r(\boldsymbol{M})}\left\|\left(\overline{\boldsymbol{M}}^{0,(l),\text{zero}}-\overline{\boldsymbol{M}}^{0,(l)}\right)\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\right\|_F.\end{aligned}$$

For $i\neq l$, we have

$$\begin{aligned}&\left(\overline{\boldsymbol{M}}^{0,(l)}-\overline{\boldsymbol{M}}^{0,(l),\text{zero}}\right)_{i,\cdot}^\top\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\\ =&\left(\overline{\boldsymbol{M}}^{0,(l)}-\overline{\boldsymbol{M}}^{0,(l),\text{zero}}\right)_{i,l}(\widetilde{\boldsymbol{Z}}_{l,\cdot}^{0,(l),\text{zero}})^\top\\ =&\mathbf{0}.\end{aligned}$$

The last equation holds since by construction we have $\widetilde{\boldsymbol{Z}}_{l,\cdot}^{0,(l),\text{zero}}=\mathbf{0}$. In order to see this, note the fact that by definition, entries on l-th row of $\overline{\boldsymbol{M}}^{0,(l),\text{zero}}$ are identical zeros, so if there is an eigenvector $\boldsymbol{v}$ with $v_l\neq 0$, the corresponding eigenvalue must be zero. Since $\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}$ is the collection of top-r eigenvectors. By Weyl's inequality and $\left\|\overline{\boldsymbol{M}}^{0,(l),\text{zero}}-\overline{\boldsymbol{M}}\right\|\leqslant\frac{1}{4}\sigma_r(\boldsymbol{M})$ we have the corresponding eigenvalues are all positive. Therefore we have $\widetilde{\boldsymbol{Z}}_{l,\cdot}^{0,(l),\text{zero}}=\mathbf{0}$.

So we have

$$\begin{aligned}
&\left\|\left(\overline{\boldsymbol{M}}^{0,(l),\text{zero}}-\overline{\boldsymbol{M}}^{0,(l)}\right)\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\right\|_F\\
=&\left\|\left(\overline{\boldsymbol{M}}^{0,(l),\text{zero}}-\overline{\boldsymbol{M}}^{0,(l)}\right)_{l,\cdot}^{\top}\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\right\|_2\\
=&\left\|\overline{\boldsymbol{M}}_{l,\cdot}^{\top}\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\right\|_2\\
\leqslant&\|\overline{\boldsymbol{M}}\|_{2,\infty}\\
\leqslant&\sigma_1(\boldsymbol{M})\max\{\|\widetilde{\boldsymbol{U}}\|_{2,\infty},\|\widetilde{\boldsymbol{V}}\|_{2,\infty}\}\\
\leqslant&\sqrt{2}\sigma_1(\boldsymbol{M})\|\widetilde{\boldsymbol{W}}\|_{2,\infty}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{Z}}^{0,(l)}\operatorname{sgn}\left((\widetilde{\boldsymbol{Z}}^{0,(l)})^{\top}\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\right)-\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\right\|_F\\
\leqslant&\frac{4}{\sigma_r(\boldsymbol{M})}\sigma_1(\boldsymbol{M})\|\widetilde{\boldsymbol{W}}\|_{2,\infty}\\
=&4\kappa\|\widetilde{\boldsymbol{W}}\|_{2,\infty}.
\end{aligned}\tag{C.49}$$

Putting (C.48) and (C.49) together we have

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{Z}}^{0,(l)}\right\|_{2,\infty}\\
=&\|\widetilde{\boldsymbol{Z}}^{0,(l)}\operatorname{sgn}\left((\widetilde{\boldsymbol{Z}}^{0,(l)})^{\top}\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\right)\|_{2,\infty}\\
\leqslant&\|\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\|_{2,\infty}\\
&+\left\|\widetilde{\boldsymbol{Z}}^{0,(l)}\operatorname{sgn}\left((\widetilde{\boldsymbol{Z}}^{0,(l)})^{\top}\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\right)-\widetilde{\boldsymbol{Z}}^{0,(l),\text{zero}}\right\|_F\\
\leqslant&(4+4\kappa+9C_{15}\kappa^2)\|\widetilde{\boldsymbol{W}}\|_{2,\infty},
\end{aligned}$$

holds on an event $E_{Claim}=E_{Ch1}\bigcap E_{Ch2}\bigcap E_A$, using union bound we have $\mathbb{P}[E_{Claim}]\geqslant 1-3(n_1+n_2)^{-11}$, which proves the claim. □

C.3. Proof of Claim 4.3.5

Proof. Similar to what we did in the control of spectral norm, define the auxiliary iteration as

$$\begin{aligned}\widetilde{\boldsymbol{X}}^{t+1,(l)} :=& \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} - \frac{\eta}{p}\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{V} - \eta\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{V}\\ &- \frac{\eta}{2}\boldsymbol{U}(\boldsymbol{R}^{t,(l)})^\top\left(\left(\boldsymbol{X}^{t,(l)}\right)^\top\boldsymbol{X}^{t,(l)} - \left(\boldsymbol{Y}^{t,(l)}\right)^\top\boldsymbol{Y}^{t,(l)}\right)\boldsymbol{R}^{t,(l)},\end{aligned}$$

$$\begin{aligned}\widetilde{\boldsymbol{Y}}^{t+1,(l)} :=& \boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} - \frac{\eta}{p}\left[\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right]^\top\boldsymbol{U} - \eta\left[\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right]^\top\boldsymbol{U}\\ &- \frac{\eta}{2}\boldsymbol{V}(\boldsymbol{R}^{t,(l)})^\top\left(\left(\boldsymbol{Y}^{t,(l)}\right)^\top\boldsymbol{Y}^{t,(l)} - \left(\boldsymbol{X}^{t,(l)}\right)^\top\boldsymbol{X}^{t,(l)}\right)\boldsymbol{R}^{t,(l)}.\end{aligned}$$

Here we want apply Lemma 4.3.3 with

$$\boldsymbol{C} = \begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1,(l)}\\ \widetilde{\boldsymbol{Y}}^{t+1,(l)}\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix},$$

$$\boldsymbol{E} = \begin{bmatrix}\boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)}\\ \boldsymbol{Y}^{t+1,(l)}\boldsymbol{R}^{t,(l)} - \widetilde{\boldsymbol{Y}}^{t+1,(l)}\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}.$$

By definition of $\boldsymbol{R}^{t+1,(l)}$ we have

$$\begin{aligned}&(\boldsymbol{R}^{t,(l)})^{-1}\boldsymbol{R}^{t+1,(l)}\\ =&\operatorname*{argmin}_{\boldsymbol{R}}\left\|\begin{bmatrix}\boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t,(l)}\\ \boldsymbol{Y}^{t+1,(l)}\boldsymbol{R}^{t,(l)}\end{bmatrix}\boldsymbol{R} - \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\|_F\\ =&\operatorname{sgn}(\boldsymbol{C}+\boldsymbol{E}).\end{aligned}$$

If $\boldsymbol{C}$ is a positive definite matrix, then $\operatorname{sgn}(\boldsymbol{C}) = \boldsymbol{I}$, and we have

$$\begin{aligned}&\|(\boldsymbol{R}^{t,(l)})^{-1}\boldsymbol{R}^{t+1,(l)} - \boldsymbol{I}\|\\ =&\|\operatorname{sgn}(\boldsymbol{C}+\boldsymbol{E}) - \operatorname{sgn}(\boldsymbol{C})\|\\ \leqslant&\frac{1}{\sigma_r(\boldsymbol{P})}\left\|\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)}\\ \boldsymbol{Y}^{t+1,(l)}\boldsymbol{R}^{t,(l)} - \widetilde{\boldsymbol{Y}}^{t+1,(l)}\end{bmatrix}\right\|.\end{aligned}$$

The remaining part are devoted to verifying the required conditions of Lemma 4.3.3, $\boldsymbol{C}$ is a positive definite matrix and upper bounding

$$\left\| \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)} \\ \boldsymbol{Y}^{t+1,(l)}\boldsymbol{R}^{t,(l)} - \widetilde{\boldsymbol{Y}}^{t+1,(l)} \end{bmatrix} \right\|.$$

Let $\boldsymbol{P} := \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}^\top \begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1,(l)} \\ \widetilde{\boldsymbol{Y}}^{t+1,(l)} \end{bmatrix}$, we have

$$\begin{aligned}
\boldsymbol{P} =& \boldsymbol{U}^\top \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} - \frac{\eta}{p}\boldsymbol{U}^\top \mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{V} \\
&- \eta\boldsymbol{U}^\top \mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{V} \\
&- \frac{\eta}{2}\boldsymbol{U}^\top\boldsymbol{U}(\boldsymbol{R}^{t,(l)})^\top\left(\left(\boldsymbol{X}^{t,(l)}\right)^\top\boldsymbol{X}^{t,(l)} - \left(\boldsymbol{Y}^{t,(l)}\right)^\top\boldsymbol{Y}^{t,(l)}\right)\boldsymbol{R}^{t,(l)} \\
&+ \boldsymbol{V}^\top\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} - \frac{\eta}{p}\boldsymbol{V}^\top\left[\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right]^\top\boldsymbol{U} \\
&- \eta\boldsymbol{V}^\top\left[\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right]^\top\boldsymbol{U} \\
&- \frac{\eta}{2}\boldsymbol{V}^\top\boldsymbol{V}(\boldsymbol{R}^{t,(l)})^\top\left(\left(\boldsymbol{Y}^{t,(l)}\right)^\top\boldsymbol{Y}^{t,(l)} - \left(\boldsymbol{X}^{t,(l)}\right)^\top\boldsymbol{X}^{t,(l)}\right)\boldsymbol{R}^{t,(l)} \\
=& \boldsymbol{U}^\top \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} - \frac{\eta}{p}\boldsymbol{U}^\top \mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{V} \\
&- \eta\boldsymbol{U}^\top \mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{V} \\
&+ \boldsymbol{V}^\top\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} - \frac{\eta}{p}\boldsymbol{V}^\top\left[\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right]^\top\boldsymbol{U} \\
&- \eta\boldsymbol{V}^\top\left[\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right]^\top\boldsymbol{U},
\end{aligned}$$

here the last equality use the fact that $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{V}^\top\boldsymbol{V}$. By the choice of $\boldsymbol{R}^{t,(l)}$, we also have $\boldsymbol{U}^\top\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} + \boldsymbol{V}^\top\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}$ is symmetric, therefore $\boldsymbol{P}$ is symmetric.

Denote

$$\begin{aligned}&\widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1,(l)}\\ :=&\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\eta\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)\boldsymbol{V}\\ &-\frac{\eta}{2}\boldsymbol{U}(\boldsymbol{R}^{t,(l)})^{\top}\left(\left(\boldsymbol{X}^{t,(l)}\right)^{\top}\boldsymbol{X}^{t,(l)}-\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}\boldsymbol{Y}^{t,(l)}\right)\boldsymbol{R}^{t,(l)},\end{aligned}$$

and

$$\begin{aligned}&\widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1,(l)}\\ :=&\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}-\eta\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)^{\top}\boldsymbol{U}\\ &-\frac{\eta}{2}\boldsymbol{V}(\boldsymbol{R}^{t,(l)})^{\top}\left(\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}\boldsymbol{Y}^{t,(l)}-\left(\boldsymbol{X}^{t,(l)}\right)^{\top}\boldsymbol{X}^{t,(l)}\right)\boldsymbol{R}^{t,(l)}.\end{aligned}$$

In order to see all the eigenvalues of $\boldsymbol{P}$ are positive, first by triangle inequality,

$$\begin{aligned}&\left\|\begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1,(l)}-\boldsymbol{U}\\ \widetilde{\boldsymbol{Y}}^{t+1,(l)}-\boldsymbol{V}\end{bmatrix}\right\|\\ \leqslant&\left\|\begin{bmatrix}\widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1,(l)}\\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1,(l)}\end{bmatrix}-\begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1,(l)}\\ \widetilde{\boldsymbol{Y}}^{t+1,(l)}\end{bmatrix}\right\|+\left\|\begin{bmatrix}\widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1,(l)}\\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1,(l)}\end{bmatrix}-\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\|.\end{aligned}\tag{C.50}$$

For the first term of the right hand side of (C.50), note

$$\begin{aligned}&\left\|\begin{bmatrix}\widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1,(l)}\\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1,(l)}\end{bmatrix}-\begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1,(l)}\\ \widetilde{\boldsymbol{Y}}^{t+1,(l)}\end{bmatrix}\right\|\\ =&\eta\left\|\begin{bmatrix}-\mathcal{P}_{-l,\cdot}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)\boldsymbol{V}+\frac{1}{p}\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)\boldsymbol{V}\\ -\left[\mathcal{P}_{-l,\cdot}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)\right]^{\top}\boldsymbol{U}+\frac{1}{p}\left[\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)\right]^{\top}\boldsymbol{U}\end{bmatrix}\right\|\\ \leqslant&2\eta\|\boldsymbol{U}\|\left\|\frac{1}{p}\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)-\mathcal{P}_{-l,\cdot}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)\right\|\\ \leqslant&2\eta\|\boldsymbol{U}\|\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)-\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)\right\|.\end{aligned}\tag{C.51}$$

The last line uses the fact that

$$\frac{1}{p}\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)-\mathcal{P}_{-l,\cdot}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)$$

is a matrix with l-th row all zero and

$$\left\|\begin{bmatrix}\boldsymbol{A}\\ \boldsymbol{0}\end{bmatrix}\right\|\leqslant\left\|\begin{bmatrix}\boldsymbol{A}\\ \boldsymbol{b}^{\top}\end{bmatrix}\right\|$$

for any matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$ with suitable shape. Using Lemma 2.3.6, we have

$$\begin{aligned}
&\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)-\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)\right\|\\
\leqslant&\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\left(\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{U}\right)\boldsymbol{V}^{\top}\right)-\left(\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{U}\right)\boldsymbol{V}^{\top}\right\|\\
&+\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{U}\left(\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{V}\right)^{\top}\right)-\boldsymbol{U}\left(\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{V}\right)^{\top}\right\|\\
&+\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\left(\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{U}\right)\left(\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{V}\right)^{\top}\right)-\left(\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{U}\right)\left(\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{V}\right)^{\top}\right\|\\
\leqslant&\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\left(\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{U}\right\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}+\|\boldsymbol{U}\|_{2,\infty}\left\|\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{V}\right\|_{2,\infty}\right)\\
&+\frac{\|\boldsymbol{\Omega}-p\boldsymbol{J}\|}{p}\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{U}\right\|_{2,\infty}\left\|\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{V}\right\|_{2,\infty}.
\end{aligned}$$

Here we use the fact that

$$\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}=\left(\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}\right)\left(\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)}\right)^{\top}.$$

On the event E_{gd}^{t}, from (4.23) and (4.24), we have

$$\begin{aligned}
&\left\|\begin{bmatrix}\boldsymbol{X}^{t,(l)}\\ \boldsymbol{Y}^{t,(l)}\end{bmatrix}\boldsymbol{T}^{t,(l)}-\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\|_{2,\infty}\\
\leqslant&\left\|\begin{bmatrix}\boldsymbol{X}^{t}\\ \boldsymbol{Y}^{t}\end{bmatrix}\boldsymbol{R}^{t}-\begin{bmatrix}\boldsymbol{X}^{t,(l)}\\ \boldsymbol{Y}^{t,(l)}\end{bmatrix}\boldsymbol{T}^{t,(l)}\right\|_{F}+\left\|\begin{bmatrix}\boldsymbol{X}^{t}\\ \boldsymbol{Y}^{t}\end{bmatrix}\boldsymbol{R}^{t}-\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}\right\|_{2,\infty}\\
\leqslant&111C_{I}\rho^{t}\sqrt{\frac{\mu^{2}r^{2}\kappa^{12}\log(n_1\vee n_2)}{(n_1\wedge n_2)^{2}p}}\sqrt{\sigma_1(\boldsymbol{M})}.
\end{aligned}\tag{C.52}$$

From Lemma 4.3.2 and (C.52),

$$
\begin{aligned}
&\left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right) - \left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\right\| \\
&\leqslant \sqrt{\frac{n_1 \wedge n_2}{p}} \times 111 C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{12}\log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}} \sqrt{\sigma_1(\boldsymbol{M})} \\
&\quad \times \left(2\sqrt{\frac{\mu r \kappa}{n_1 \wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})} + 111 C_I \rho^t \sqrt{\frac{\mu^2 r^2 \kappa^{12}\log(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p}}\sqrt{\sigma_1(\boldsymbol{M})}\right).
\end{aligned}
\tag{C.53}
$$

For the second term of the right hand side of (C.50), we deal with it very similar to the way we deal with α_2 defined in (4.28): Note

$$
\begin{aligned}
&\begin{bmatrix} \widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1,(l)} \\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1,(l)} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} - \eta\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{V} \\ \boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} - \eta\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top - \boldsymbol{U}\boldsymbol{V}^\top\right)^\top\boldsymbol{U} \end{bmatrix} \\
&\quad - \begin{bmatrix} \frac{\eta}{2}\boldsymbol{U}(\boldsymbol{R}^{t,(l)})^\top\left(\left(\boldsymbol{X}^{t,(l)}\right)^\top\boldsymbol{X}^{t,(l)} - \left(\boldsymbol{Y}^{t,(l)}\right)^\top\boldsymbol{Y}^{t,(l)}\right)\boldsymbol{R}^{t,(l)} \\ \frac{\eta}{2}\boldsymbol{V}(\boldsymbol{R}^{t,(l)})^\top\left(\left(\boldsymbol{Y}^{t,(l)}\right)^\top\boldsymbol{Y}^{t,(l)} - \left(\boldsymbol{X}^{t,(l)}\right)^\top\boldsymbol{X}^{t,(l)}\right)\boldsymbol{R}^{t,(l)} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} - \eta\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}\boldsymbol{V}^\top\boldsymbol{V} - \eta\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)})^\top\boldsymbol{V} \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} - \eta\boldsymbol{V}(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})^\top\boldsymbol{U} - \eta\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\boldsymbol{U}^\top\boldsymbol{U} \end{bmatrix} \\
&\quad + \begin{bmatrix} -\frac{\eta}{2}\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})^\top\boldsymbol{U} - \frac{\eta}{2}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} + \frac{\eta}{2}\boldsymbol{U}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)})^\top\boldsymbol{V} \\ -\frac{\eta}{2}\boldsymbol{V}(\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)})^\top\boldsymbol{V} - \frac{\eta}{2}\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} + \frac{\eta}{2}\boldsymbol{V}(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})^\top\boldsymbol{U} \end{bmatrix} + \begin{bmatrix} \frac{\eta}{2}\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} + \eta\boldsymbol{\mathcal{E}}_1 \\ \frac{\eta}{2}\boldsymbol{V}\boldsymbol{U}^\top\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} + \eta\boldsymbol{\mathcal{E}}_2 \end{bmatrix},
\end{aligned}
$$

where $\boldsymbol{\mathcal{E}}_1, \boldsymbol{\mathcal{E}}_2$ denote those terms with at least two $\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}$'s and $\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}$'s (the expression of $\boldsymbol{\mathcal{E}}_1$ and $\boldsymbol{\mathcal{E}}_2$ one can refer to (4.33) and (4.34), replacing $\boldsymbol{\Delta}_{\boldsymbol{X}}^{t}$ and $\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t}$ by $\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}$ and $\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}$). Again by the way we define $\boldsymbol{R}^{t,(l)}$,

$$
\begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} = (\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})^\top\boldsymbol{U} + (\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)})^\top\boldsymbol{V}
$$

is symmetric. Plugging back we have

$$\begin{aligned}
&\begin{bmatrix} \widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1,(l)} \\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1,(l)} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \\
=&\begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} - \eta \boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} \boldsymbol{V}^\top \boldsymbol{V} - \eta \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} + \eta \boldsymbol{\mathcal{E}}_1 \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} - \eta \boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} \boldsymbol{U}^\top \boldsymbol{U} - \eta \boldsymbol{V}\boldsymbol{V}^\top \boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} + \eta \boldsymbol{\mathcal{E}}_2 \end{bmatrix} \\
=&\frac{1}{2} \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} \end{bmatrix} (\boldsymbol{I} - 2\eta \boldsymbol{U}^\top \boldsymbol{U}) + \frac{1}{2} \left(\boldsymbol{I} - 2\eta \begin{bmatrix} \boldsymbol{U}\boldsymbol{U}^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}\boldsymbol{V}^\top \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} \end{bmatrix} + \eta \boldsymbol{\mathcal{E}},
\end{aligned}$$

where the last line we use the fact that $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{V}^\top \boldsymbol{V}$, and $\boldsymbol{\mathcal{E}} := \begin{bmatrix} \boldsymbol{\mathcal{E}}_1 \\ \boldsymbol{\mathcal{E}}_2 \end{bmatrix}$. Since $\boldsymbol{U}\boldsymbol{U}^\top$ and $\boldsymbol{V}\boldsymbol{V}^\top$ sharing the same eigenvalues, we have

$$\begin{aligned}
&\left\| \begin{bmatrix} \widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1,(l)} \\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1,(l)} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \\
\leqslant& \frac{1}{2} \|\boldsymbol{I} - 2\eta \boldsymbol{U}^\top \boldsymbol{U}\| \|\boldsymbol{\Delta}^{t,(l)}\| + \frac{1}{2} \|\boldsymbol{\Delta}^{t,(l)}\| \left\| \boldsymbol{I} - 2\eta \begin{bmatrix} \boldsymbol{U}\boldsymbol{U}^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}\boldsymbol{V}^\top \end{bmatrix} \right\| + \eta \|\boldsymbol{\mathcal{E}}\| \\
\leqslant& (1 - \eta \sigma_r(\boldsymbol{M})) \|\boldsymbol{\Delta}^{t,(l)}\| + \eta \|\boldsymbol{\mathcal{E}}\|.
\end{aligned}$$

By the definition of $\boldsymbol{\mathcal{E}}$, we have

$$\|\boldsymbol{\mathcal{E}}\| \leqslant 4 \|\boldsymbol{\Delta}^{t,(l)}\|^2 \|\boldsymbol{U}\|.$$

From (4.44),

$$\begin{aligned}
&\left\| \begin{bmatrix} \widetilde{\mathbb{E}}\widetilde{\boldsymbol{X}}^{t+1,(l)} \\ \widetilde{\mathbb{E}}\widetilde{\boldsymbol{Y}}^{t+1,(l)} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\| \\
\leqslant& (1 - \eta \sigma_r(\boldsymbol{M})) \times 2 C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} \sqrt{\sigma_1(\boldsymbol{M})} \\
&+ 4\eta \left(2 C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} \sqrt{\sigma_1(\boldsymbol{M})} \right)^2 \sqrt{\sigma_1(\boldsymbol{M})}
\end{aligned} \tag{C.54}$$

holds. Combining (C.50), (C.51), (C.53) and (C.54) together, we have

$$\begin{aligned}
&\left\|\begin{bmatrix} \widetilde{\boldsymbol{X}}^{t+1,(l)}-\boldsymbol{U} \\ \widetilde{\boldsymbol{Y}}^{t+1,(l)}-\boldsymbol{V} \end{bmatrix}\right\| \\
\leqslant & 2\eta\sqrt{\sigma_1(\boldsymbol{M})}\sqrt{\frac{n_1\wedge n_2}{p}}\times 111C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})} \\
&\times\left(2\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})}+111C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})}\right) \\
&+(1-\eta\sigma_r(\boldsymbol{M}))\times 2C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})} \\
&+4\eta\left(2C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})}\right)^2\times\sqrt{\sigma_1(\boldsymbol{M})} \\
\leqslant & \eta\sigma_r(\boldsymbol{M})C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})}+(1-\eta\sigma_r(\boldsymbol{M}))\times 2C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})} \\
&+\eta\sigma_r(\boldsymbol{M})C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})} \\
= & 2C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})} \\
\leqslant & \frac{1}{4\kappa}\sqrt{\sigma_1(\boldsymbol{M})},
\end{aligned}$$

where the second inequality holds since

$$p\geqslant(666^2+111^2C_I^2)\frac{\mu^2r^2\kappa^{11}\log(n_1\vee n_2)}{n_1\wedge n_2}$$

and the last line holds since

$$p\geqslant 64C_I^2\frac{\mu r\kappa^8\log(n_1\vee n_2)}{n_1\wedge n_2}.$$

Therefore,

$$\begin{aligned}&\left\|\begin{bmatrix}\boldsymbol{U}\\\boldsymbol{V}\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{U}\\\boldsymbol{V}\end{bmatrix}-\begin{bmatrix}\boldsymbol{U}\\\boldsymbol{V}\end{bmatrix}^\top\begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1,(l)}\\\widetilde{\boldsymbol{Y}}^{t+1,(l)}\end{bmatrix}\right\|\\\leqslant&\left\|\begin{bmatrix}\boldsymbol{U}\\\boldsymbol{V}\end{bmatrix}\right\|\left\|\begin{bmatrix}\widetilde{\boldsymbol{X}}^{t+1,(l)}-\boldsymbol{U}\\\widetilde{\boldsymbol{Y}}^{t+1,(l)}-\boldsymbol{V}\end{bmatrix}\right\|\leqslant 0.5\sigma_r(\boldsymbol{M}).\end{aligned}\tag{C.55}$$

By Weyl's inequality, we see eigenvalues of $\boldsymbol{P}$ are all nonnegative. Combining with the fact that $\boldsymbol{P}$ is symmetric, we can see $\boldsymbol{P}$ is positive definite. And also from Weyl's inequality, $\sigma_r(\boldsymbol{P})\geqslant 1.5\sigma_r(\boldsymbol{M})$.

Moreover, by the definition of $\boldsymbol{X}^{t,(l)}$ and $\boldsymbol{Y}^{t,(l)}$, as well as the assumption that $1\leqslant l\leqslant n_1$,

$$\begin{aligned}&\boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t,(l)}\\=&\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\frac{\eta}{p}\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top-\boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\\&-\eta\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top-\boldsymbol{U}\boldsymbol{V}^\top\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\\&-\frac{\eta}{2}\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}(\boldsymbol{R}^{t,(l)})^\top\left((\boldsymbol{X}^{t,(l)})^\top\boldsymbol{X}^{t,(l)}-(\boldsymbol{Y}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)}\right)\boldsymbol{R}^{t,(l)},\end{aligned}$$

$$\begin{aligned}&\boldsymbol{Y}^{t+1,(l)}\boldsymbol{R}^{t,(l)}\\=&\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}-\frac{\eta}{p}\left[\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top-\boldsymbol{U}\boldsymbol{V}^\top\right)\right]^\top\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\\&-\eta\left[\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}(\boldsymbol{Y}^{t,(l)})^\top-\boldsymbol{U}\boldsymbol{V}^\top\right)\right]^\top\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\\&-\frac{\eta}{2}\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}(\boldsymbol{R}^{t,(l)})^\top\left((\boldsymbol{Y}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)}-(\boldsymbol{X}^{t,(l)})^\top\boldsymbol{X}^{t,(l)}\right)\boldsymbol{R}^{t,(l)}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\begin{bmatrix}\boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t,(l)}-\widetilde{\boldsymbol{X}}^{t+1,(l)}\\\boldsymbol{Y}^{t+1,(l)}\boldsymbol{R}^{t,(l)}-\widetilde{\boldsymbol{Y}}^{t+1,(l)}\end{bmatrix}\\=&\begin{bmatrix}\boldsymbol{0}&\eta\boldsymbol{A}\\\eta\boldsymbol{A}^\top&\boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}\\\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\end{bmatrix}\\&+\begin{bmatrix}-\frac{\eta}{2}\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}(\boldsymbol{R}^{t,(l)})^\top\left(\left(\boldsymbol{X}^{t,(l)}\right)^\top\boldsymbol{X}^{t,(l)}-\left(\boldsymbol{Y}^{t,(l)}\right)^\top\boldsymbol{Y}^{t,(l)}\right)\boldsymbol{R}^{t,(l)}\\-\frac{\eta}{2}\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}(\boldsymbol{R}^{t,(l)})^\top\left(\left(\boldsymbol{Y}^{t,(l)}\right)^\top\boldsymbol{Y}^{t,(l)}-\left(\boldsymbol{X}^{t,(l)}\right)^\top\boldsymbol{X}^{t,(l)}\right)\boldsymbol{R}^{t,(l)}\end{bmatrix}\end{aligned}\tag{C.56}$$

with

$$\boldsymbol{A} \coloneqq -\frac{1}{p}\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)-\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right).$$

First in order to give a bound of $\|\boldsymbol{A}\|$, we can first decompose $\boldsymbol{A}$ as

$$\begin{aligned}\boldsymbol{A} =& \underbrace{-\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)}_{\boldsymbol{A}_1}\\ &+\underbrace{\frac{1}{p}\mathcal{P}_{\Omega_{l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)-\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)}_{\boldsymbol{A}_2}.\end{aligned} \tag{C.57}$$

From Lemma 2.3.6 and Lemma 4.3.2,

$$\begin{aligned}\|\boldsymbol{A}_1\| \leqslant& \left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)-\left(\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right)\right\|\\ &+\left\|\boldsymbol{X}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\right)^{\top}-\boldsymbol{U}\boldsymbol{V}^{\top}\right\|\\ \leqslant& C_{13}\sqrt{\frac{n_1\wedge n_2}{p}}\left(\|\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}+\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{V}\|_{2,\infty}\right)\\ &+C_{13}\sqrt{\frac{n_1\wedge n_2}{p}}\|\boldsymbol{X}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{Y}^{t,(l)}\boldsymbol{T}^{t,(l)}-\boldsymbol{V}\|_{2,\infty}\\ &+\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{U}\|\|\boldsymbol{V}\|+\|\boldsymbol{U}\|\|\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{V}\|+\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{U}\|\|\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{V}\|\end{aligned}$$

holds on the event $E_{gd}^t \subset E_S$.

From (C.52) and (4.44),

$$
\begin{aligned}
&\|\boldsymbol{A}_1\| \\
\leqslant & C_{13}\sqrt{\frac{n_1\wedge n_2}{p}}\times 222C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})}\times\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})} \\
&+C_{13}\sqrt{\frac{n_1\wedge n_2}{p}}\left(111C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})}\right)^2 \\
&+4C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sigma_1(\boldsymbol{M})+\left(2C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})}\right)^2 \qquad \text{(C.58)} \\
\leqslant & C_{13}\sqrt{\frac{n_1\wedge n_2}{p}}\times 333C_I\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log(n_1\vee n_2)}{(n_1\wedge n_2)^2p}}\sqrt{\sigma_1(\boldsymbol{M})}\times\sqrt{\frac{\mu r\kappa}{n_1\wedge n_2}}\sqrt{\sigma_1(\boldsymbol{M})} \\
&+5C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sigma_1(\boldsymbol{M}) \\
\leqslant & 6C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sigma_1(\boldsymbol{M})
\end{aligned}
$$

where the second inequality holds since

$$p\geqslant 111^2C_I^2\frac{\mu r\kappa^{11}\log(n_1\vee n_2)}{n_1\wedge n_2}$$

and the last inequality holds since

$$p\geqslant 333^2C_{13}^2\frac{\mu^2r^2\kappa^7}{n_1\wedge n_2}.$$

Note by the definition of $\boldsymbol{A}_2$, we have $\|\boldsymbol{A}_2\|=\|(\boldsymbol{A}_2)_{l,\cdot}\|_2$, and note $(\boldsymbol{A}_2)_{l,\cdot}$ here is exactly $\boldsymbol{b}_2$ we define in (4.55), therefore we directly use the result (4.66) and (4.67):

$$
\begin{aligned}
&\|\boldsymbol{A}_2\| \\
=&\|\boldsymbol{b}_2\|_2 \\
\leqslant & 100\rho^t\left(115C_I\sqrt{\frac{\mu^2r^2\kappa^{12}\log^2(n_1\vee n_2)}{(n_1\wedge n_2)^2p^2}}+333C_I\sqrt{\frac{\mu^3r^3\kappa^{13}\log^3(n_1\vee n_2)}{(n_1\wedge n_2)^3p^3}}\right)\sigma_1(\boldsymbol{M}) \qquad \text{(C.59)} \\
\leqslant & C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sigma_1(\boldsymbol{M})
\end{aligned}
$$

holds on the event E_{gd}^{t+1}, where the last inequality holds since

$$p \geqslant 5.29 \times 10^8 \frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{n_1 \wedge n_2}.$$

By putting (C.58), (C.59) and (C.57) together we have

$$\|\boldsymbol{A}\| \leqslant \|\boldsymbol{A}_1\| + \|\boldsymbol{A}_2\| \leqslant 7 C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} \sigma_1(\boldsymbol{M}) \tag{C.60}$$

holds on the event E_{gd}^{t+1}. Moreover,

$$\begin{aligned}
&\left\| \begin{bmatrix} -\frac{\eta}{2} \boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} (\boldsymbol{R}^{t,(l)})^\top \left(\left(\boldsymbol{X}^{t,(l)}\right)^\top \boldsymbol{X}^{t,(l)} - \left(\boldsymbol{Y}^{t,(l)}\right)^\top \boldsymbol{Y}^{t,(l)} \right) \boldsymbol{R}^{t,(l)} \\ -\frac{\eta}{2} \boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} (\boldsymbol{R}^{t,(l)})^\top \left(\left(\boldsymbol{Y}^{t,(l)}\right)^\top \boldsymbol{Y}^{t,(l)} - \left(\boldsymbol{X}^{t,(l)}\right)^\top \boldsymbol{X}^{t,(l)} \right) \boldsymbol{R}^{t,(l)} \end{bmatrix} \right\| \\
\leqslant & \frac{\eta}{2} \left\| \boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} (\boldsymbol{R}^{t,(l)})^\top \left(\left(\boldsymbol{X}^{t,(l)}\right)^\top \boldsymbol{X}^{t,(l)} - \left(\boldsymbol{Y}^{t,(l)}\right)^\top \boldsymbol{Y}^{t,(l)} \right) \boldsymbol{R}^{t,(l)} \right\| \\
& + \frac{\eta}{2} \left\| \boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} (\boldsymbol{R}^{t,(l)})^\top \left(\left(\boldsymbol{Y}^{t,(l)}\right)^\top \boldsymbol{Y}^{t,(l)} - \left(\boldsymbol{X}^{t,(l)}\right)^\top \boldsymbol{X}^{t,(l)} \right) \boldsymbol{R}^{t,(l)} \right\| \\
\leqslant & \frac{\eta}{2} \left(\|\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}\| + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\| \right) \left(2 \|\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}\| \|\boldsymbol{U}\| + \|\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}\|^2 + 2 \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\| \|\boldsymbol{V}\| + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\|^2 \right) \\
\leqslant & \frac{\eta}{2} \times 4 C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} \sqrt{\sigma_1(\boldsymbol{M})} \times 12 C_I \rho^t \sqrt{\frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{(n_1 \wedge n_2) p}} \sigma_1(\boldsymbol{M}) \\
\leqslant & 24 C_I^2 \eta \rho^{2t} \sqrt{\frac{\mu^2 r^2 \kappa^{12} \log^2(n_1 \vee n_2)}{(n_1 \wedge n_2)^2 p^2}} \sqrt{\sigma_1(\boldsymbol{M})}^3,
\end{aligned} \tag{C.61}$$

where the third inequality uses (4.44) and

$$p \geqslant 4 C_I^2 \frac{\mu r \kappa^6 \log(n_1 \vee n_2)}{n_1 \wedge n_2}.$$

Now from (C.56), (C.60), (C.61) and also (4.44) we can see that

$$\begin{aligned}
&\left\|\begin{bmatrix} \boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)} \\ \boldsymbol{Y}^{t+1,(l)}\boldsymbol{R}^{t,(l)} - \widetilde{\boldsymbol{Y}}^{t+1,(l)} \end{bmatrix}\right\| \\
&\leqslant \eta \times 7C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sigma_1(\boldsymbol{M}) \times 2C_I\rho^t\sqrt{\frac{\mu r\kappa^6\log(n_1\vee n_2)}{(n_1\wedge n_2)p}}\sqrt{\sigma_1(\boldsymbol{M})} \\
&\quad + 24C_I^2\eta\rho^{2t}\sqrt{\frac{\mu^2r^2\kappa^{12}\log^2(n_1\vee n_2)}{(n_1\wedge n_2)^2p^2}}\sqrt{\sigma_1(\boldsymbol{M})}^3 \\
&\leqslant 38C_I^2\eta\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log^2(n_1\vee n_2)}{(n_1\wedge n_2)^2p^2}}\sqrt{\sigma_1(\boldsymbol{M})}^3.
\end{aligned} \tag{C.62}$$

Therefore,

$$\begin{aligned}
&\left\|\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)} \\ \boldsymbol{Y}^{t+1,(l)}\boldsymbol{R}^{t,(l)} - \widetilde{\boldsymbol{Y}}^{t+1,(l)} \end{bmatrix}\right\| \\
&\leqslant \left\|\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}\right\| \left\|\begin{bmatrix} \boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)} \\ \boldsymbol{Y}^{t+1,(l)}\boldsymbol{R}^{t,(l)} - \widetilde{\boldsymbol{Y}}^{t+1,(l)} \end{bmatrix}\right\| \\
&\leqslant \sigma_r(\boldsymbol{M}) \\
&\leqslant \sigma_r(\boldsymbol{P}),
\end{aligned}$$

where the second last inequality uses the fact that

$$p \geqslant 76C_I^2\frac{\mu r\kappa^6\log(n_1\vee n_2)}{n_1\wedge n_2}$$

and

$$\eta \leqslant \frac{\sigma_r(\boldsymbol{M})}{200\sigma_1^2(\boldsymbol{M})}.$$

Therefore, we have

$$\begin{aligned}
&\|(\boldsymbol{R}^{t,(l)})^{-1}\boldsymbol{R}^{t+1,(l)}-\boldsymbol{I}\| \\
=&\|\operatorname{sgn}(\boldsymbol{C}+\boldsymbol{E})-\operatorname{sgn}(\boldsymbol{C})\| \\
\leqslant&\frac{1}{\sigma_r(\boldsymbol{P})}\left\|\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}^{\top}\begin{bmatrix}\boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t,(l)}-\widetilde{\boldsymbol{X}}^{t+1,(l)}\\ \boldsymbol{Y}^{t+1,(l)}\boldsymbol{R}^{t,(l)}-\widetilde{\boldsymbol{Y}}^{t+1,(l)}\end{bmatrix}\right\| \\
\leqslant&2\frac{\sqrt{\sigma_1(\boldsymbol{M})}}{\sigma_r(\boldsymbol{M})}\times 38C_I^2\eta\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log^2(n_1\vee n_2)}{(n_1\wedge n_2)^2p^2}}\sqrt{\sigma_1(\boldsymbol{M})}^3 \\
\leqslant&76C_I^2\frac{\sigma_1^2(\boldsymbol{M})}{\sigma_r(\boldsymbol{M})}\eta\rho^t\sqrt{\frac{\mu^2r^2\kappa^{12}\log^2(n_1\vee n_2)}{(n_1\wedge n_2)^2p^2}},
\end{aligned}$$

where the second inequality uses the fact that $\sigma_r(\boldsymbol{P})\geqslant 1.5\sigma_r(\boldsymbol{M})$ and the third one uses (C.62).

□

Bibliography

[AFWZ17] E. Abbe, J. Fan, K. Wang, and Y. Zhong, *Entrywise eigenvector analysis of random matrices with low expected rank*, arXiv preprint arXiv:1709.09565 (2017).

[AM07] D. Achlioptas and F. McSherry, *Fast computation of low-rank matrix approximations*, Journal of the ACM (JACM) **54** (2007), no. 2, 9–es.

[AMS02] D. Achlioptas, F. McSherry, and B. Schölkopf, *Sampling techniques for kernel methods*, Advances in Neural Information Processing Systems, 2002, pp. 335–342.

[Arm66] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific Journal of Mathematics **16** (1966), no. 1, 1–3.

[BH89] P. Baldi and K. Hornik, *Neural networks and principal component analysis: Learning from examples without local minima*, Neural Networks **2** (1989), no. 1, 53–58.

[Bha13] R. Bhatia, *Matrix Analysis*, Graduate Texts in Mathematics, Springer New York, 2013.

[BJ14] S. Bhojanapalli and P. Jain, *Universal matrix completion*, International Conference on Machine Learning, 2014, pp. 1881–1889.

[BM03] S. Burer and R. D. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Mathematical Programming **95** (2003), no. 2, 329–357.

[BNS16] S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Global optimality of local search for low rank matrix recovery*, Advances in Neural Information Processing Systems, 2016, pp. 3873 – 3881.

[BVH16] A. S. Bandeira and R. Van Handel, *Sharp nonasymptotic bounds on the norm of random matrices with independent entries*, The Annals of Probability **44** (2016), no. 4, 2479–2506.

[CCFM19] Y. Chen, Y. Chi, J. Fan, and C. Ma, *Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval*, Mathematical Programming **176** (2019), no. 1-2, 5–37.

[CG18] Y. Cheng and R. Ge, *Non-convex matrix completion against a semi-random adversary*, Conference On Learning Theory, 2018, pp. 1362–1394.

[Cha15] S. Chatterjee, *Matrix estimation by universal singular value thresholding*, The Annals of Statistics **43** (2015), no. 1, 177–214.

[Che15] Y. Chen, *Incoherence-optimal matrix completion*, IEEE Transactions on Information Theory **61** (2015), no. 5, 2909–2923.

[CL19] J. Chen and X. Li, *Model-free nonconvex matrix completion: Local minima analysis and applications in memory-efficient kernel PCA*, Journal of Machine Learning Research **20** (2019), no. 142, 1–39.

[CLL19] 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).

[CLM16] 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 **44** (2016), no. 5, 2221–2251.

[CLM20] J. Chen, X. Li, and Z. Ma, *Nonconvex matrix completion with linearly parameterized factors*, arXiv preprint arXiv:2003.13153 (2020).

[CLMW11] E. J. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?*, Journal of the ACM (JACM) **58** (2011), no. 3, 1–37.

[CLS15] E. J. Candès, X. Li, and M. Soltanolkotabi, *Phase retrieval via wirtinger flow: Theory and algorithms*, IEEE Transactions on Information Theory **61** (2015), no. 4, 1985–2007.

[CR09] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Foundations of Computational Mathematics **9** (2009), no. 6, 717–772.

[CT10] E. J. Candès and T. Tao, *The power of convex relaxation: Near-optimal matrix completion*, IEEE Transactions on Information Theory **56** (2010), no. 5, 2053–2080.

[CW15] 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).

[DC20] L. Ding and Y. Chen, *Leave-one-out approach for matrix completion: Primal and dual analysis*, IEEE Transactions on Information Theory (2020).

[DK70] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation. iii*, SIAM Journal on Numerical Analysis **7** (1970), no. 1, 1–46.

[DM05] P. Drineas and M. W. Mahoney, *On the Nyström method for approximating a gram matrix for improved kernel-based learning*, Journal of Machine Learning Research **6** (2005), no. Dec, 2153–2175.

[EKBB$^+$13] 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 **110** (2013), no. 36, 14557–14562.

[EYW18] A. Eftekhari, D. Yang, and M. B. Wakin, *Weighted matrix completion and recovery with prior subspace information*, IEEE Transactions on Information Theory **64** (2018), no. 6, 4044–4071.

[GJZ17] R. Ge, C. Jin, and Y. Zheng, *No spurious local minima in nonconvex low rank problems: A unified geometric analysis*, International Conference on Machine Learning, 2017, pp. 1233–1242.

[GL11] D. F. Gleich and L.-h. Lim, *Rank aggregation via nuclear norm minimization*, Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 60–68.

[GLF+10] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, *Quantum state tomography via compressed sensing*, Physical Review Letters **105** (2010), no. 15, 150401.

[GLM16] R. Ge, J. D. Lee, and T. Ma, *Matrix completion has no spurious local minimum*, Advances in Neural Information Processing Systems, 2016, pp. 2973–2981.

[GN10] D. Gross and V. Nesme, *Note on sampling without replacing from a finite collection of matrices*, arXiv preprint arXiv:1001.2738 (2010).

[Gra02] T. Graepel, *Kernel matrix completion by semidefinite programming*, International Conference on Artificial Neural Networks, Springer, 2002, pp. 694–699.

[Gro11] D. Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Transactions on Information Theory **57** (2011), no. 3, 1548–1566.

[GVL12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, vol. 3, JHU Press, 2012.

[GWB+17] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Implicit regularization in matrix factorization*, Advances in Neural Information Processing Systems, 2017, pp. 6151–6159.

[Har14] M. Hardt, *Understanding alternating minimization for matrix completion*, 2014 IEEE 55th Annual Symposium on Foundations of Computer Science, IEEE, 2014, pp. 651–660.

[HKZ11] D. Hsu, S. Kakade, and T. Zhang., *Robust matrix decomposition with sparse corruptions*, IEEE Transactions on Information Theory **57(11)** (2011), 7221–7234.

[HMT11] N. Halko, P.-G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Review **53** (2011), no. 2, 217–288.

[HW14] M. Hardt and M. Wootters, *Fast matrix completion without the condition number*, Conference on Learning Theory, 2014, pp. 638–678.

[JD13] P. Jain and I. S. Dhillon, *Provable inductive matrix completion*, arXiv preprint arXiv:1306.0626 (2013).

[JGN+17] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, *How to escape saddle points efficiently*, International Conference on Machine Learning, 2017, pp. 1724–1732.

[JLYY11] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, *Statistical ranking and combinatorial hodge theory*, Mathematical Programming **127** (2011), no. 1, 203–244.

[JM13] A. Javanmard and A. Montanari, *Localization from incomplete noisy distance measurements*, Foundations of Computational Mathematics **13** (2013), no. 3, 297–345.

[JMD10] P. Jain, R. Meka, and I. S. Dhillon, *Guaranteed rank minimization via singular value projection*, Advances in Neural Information Processing Systems, 2010, pp. 937–945.

[JNS13] P. Jain, P. Netrapalli, and S. Sanghavi, *Low-rank matrix completion using alternating minimization*, Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing, ACM, 2013, pp. 665–674.

[KFS05] K. I. Kim, M. O. Franz, and B. Schölkopf, *Iterative kernel principal component analysis for image modeling*, IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005), no. 9, 1351–1366.

[KLT11] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, *Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion*, The Annals of Statistics **39** (2011), no. 5, 2302–2329.

[KMO10a] R. H. Keshavan, A. Montanari, and S. Oh, *Matrix completion from a few entries*, IEEE Transactions on Information Theory **56** (2010), no. 6, 2980–2998.

[KMO10b] ———, *Matrix completion from noisy entries*, Journal of Machine Learning Research **11** (2010), no. Jul, 2057–2078.

[LHV13] Z. Liu, A. Hansson, and L. Vandenberghe, *Nuclear norm system identification with missing inputs and outputs*, Systems & Control Letters **62** (2013), no. 8, 605–612.

[LLR16] Y. Li, Y. Liang, and A. Risteski, *Recovery guarantee of weighted low-rank approximation via alternating minimization*, International Conference on Machine Learning, 2016, pp. 2358–2367.

[LLSW19] X. Li, S. Ling, T. Strohmer, and K. Wei, *Rapid, robust, and reliable blind deconvolution via nonconvex optimization*, Applied and Computational Harmonic Analysis **47** (2019), no. 3, 893–934.

[LLZ$^+$20] S. Li, Q. Li, Z. Zhu, G. Tang, and M. B. Wakin, *The global geometry of centralized and distributed low-rank matrix recovery without regularization*, arXiv preprint arXiv:2003.10981 (2020).

[LMZ18] Y. Li, T. Ma, and H. Zhang, *Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations*, Conference On Learning Theory, 2018, pp. 2–47.

[LPP$^+$17] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, *First-order methods almost always avoid saddle points*, arXiv preprint arXiv:1710.07406 (2017).

[LSJR16] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, *Gradient descent only converges to minimizers*, Conference on Learning Theory, 2016, pp. 1246–1257.

[LV10] Z. Liu and L. Vandenberghe, *Interior-point method for nuclear norm approximation with application to system identification*, SIAM Journal on Matrix Analysis and Applications **31** (2010), no. 3, 1235–1256.

[LW15] P.-L. Loh and M. J. Wainwright, *Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima*, Journal of Machine Learning Research **16** (2015), no. 19, 559–616.

[LWL$^+$16] X. Li, Z. Wang, J. Lu, R. Arora, J. Haupt, H. Liu, and T. Zhao, *Symmetry, saddle points, and global geometry of nonconvex matrix factorization*, arXiv preprint arXiv:1612.09296 (2016).

[LZT17] Q. Li, Z. Zhu, and G. Tang, *Geometry of factored nuclear norm regularization*, arXiv preprint arXiv:1704.01265 (2017).

[Mat93] R. Mathias, *Perturbation bounds for the polar decomposition*, SIAM Journal on Matrix Analysis and Applications **14** (1993), no. 2, 588–597.

[MHT10] R. Mazumder, T. Hastie, and R. Tibshirani, *Spectral regularization algorithms for learning large incomplete matrices*, Journal of Machine Learning Research **11** (2010), no. Aug, 2287–2322.

[MLC19] C. Ma, Y. Li, and Y. Chi, *Beyond procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing*, 2019 53rd Asilomar Conference on Signals, Systems, and Computers, IEEE, 2019, pp. 721–725.

[MOA11] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, Springer, 2011.

[MWCC18] 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*, Foundations of Computational Mathematics (2018), 1–182.

[NW12] S. Negahban and M. J. Wainwright, *Restricted strong convexity and weighted matrix completion: Optimal bounds with noise*, Journal of Machine Learning Research **13** (2012), no. May, 1665–1697.

[OMK10] S. Oh, A. Montanari, and A. Karbasi, *Sensor network localization from local connectivity: Performance analysis for the mds-map algorithm*, 2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo), IEEE, 2010, pp. 1–5.

[Pau02] V. Paulsen, *Completely Bounded Maps and Operator Algebras*, vol. 78, Cambridge University Press, 2002.

[PC10] J. Paisley and L. Carin, *A nonparametric bayesian model for kernel matrix completion*, 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2010, pp. 2090–2093.

[PP17] I. Panageas and G. Piliouras, *Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions*, Innovations in Theoretical Computer Science, 2017.

[Rec11] B. Recht, *A simpler approach to matrix completion*, Journal of Machine Learning Research **12** (2011), no. Dec, 3413–3430.

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM review **52** (2010), no. 3, 471–501.

[RS05] J. D. Rennie and N. Srebro, *Fast maximum margin matrix factorization for collaborative prediction*, Proceedings of the 22nd International Conference on Machine Learning, ACM, 2005, pp. 713–719.

[Saa03] Y. Saad, *Iterative Methods for Sparse Linear Systems*, vol. 82, SIAM, 2003.

[SCH$^+$16] S. Si, K.-Y. Chiang, C.-J. Hsieh, N. Rao, and I. S. Dhillon, *Goal-directed inductive matrix completion*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1165–1174.

[Sin08] A. Singer, *A remark on global positioning from local distances*, Proceedings of the National Academy of Sciences **105** (2008), no. 28, 9507–9511.

[SL16] R. Sun and Z.-Q. Luo, *Guaranteed matrix completion via non-convex factorization*, IEEE Transactions on Information Theory **62** (2016), no. 11, 6535–6579.

[SQW18] J. Sun, Q. Qu, and J. Wright, *A geometric analysis of phase retrieval*, Foundations of Computational Mathematics **18** (2018), no. 5, 1131–1198.

[SSM98] B. Schölkopf, A. Smola, and K.-R. Müller, *Nonlinear component analysis as a kernel eigenvalue problem*, Neural Computation **10** (1998), no. 5, 1299–1319.

[Sun15] R. Sun, *Matrix completion via nonconvex factorization: Algorithms and theory*, PhD dissertation, University of Minnesota, 2015.

[SY07] A. M.-C. So and Y. Ye, *Theory of semidefinite programming for sensor network localization*, Mathematical Programming **109** (2007), no. 2-3, 367–384.

[SZ12] T. Sun and C.-H. Zhang, *Calibrated elastic regularization in matrix completion*, Advances in Neural Information Processing Systems, 2012, pp. 863–871.

[TBS$^+$15] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, *Low-rank solutions of linear matrix equations via procrustes flow*, arXiv preprint arXiv:1507.03566 (2015).

[Tro15] J. A. Tropp, *An introduction to matrix concentration inequalities*, Foundations and Trends® in Machine Learning **8** (2015), no. 1-2, 1–230.

[Vu18] V. Vu, *A simple SVD algorithm for finding hidden partitions*, Combinatorics, Probability and Computing **27** (2018), no. 1, 124–140.

[Wai19] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48, Cambridge University Press, 2019.

[Wan12] Q. Wang, *Kernel principal component analysis and its applications in face recognition and active shape models*, arXiv preprint arXiv:1207.3538 (2012).

[WS01] C. K. Williams and M. Seeger, *Using the Nyström method to speed up kernel machines*, Advances in Neural Information Processing Systems, 2001, pp. 682–688.

[WZG17] L. Wang, X. Zhang, and Q. Gu, *A Unified Computational and Statistical Framework for Nonconvex Low-rank Matrix Estimation*, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Fort Lauderdale, FL, USA), Proceedings of Machine Learning Research, vol. 54, PMLR, 20–22 Apr 2017, pp. 981–990.

[XJZ13] M. Xu, R. Jin, and Z.-H. Zhou, *Speedup matrix completion with side information: Application to multi-label learning*, Advances in Neural Information Processing Systems, 2013, pp. 2301–2309.

[You61] D. Youla, *A normal form for a matrix under the unitary congruence group*, Canadian Journal of Mathematics **13** (1961), 694–704.

[YPCC16] X. Yi, D. Park, Y. Chen, and C. Caramanis, *Fast algorithms for robust PCA via gradient descent*, Advances in Neural Information Processing Systems, 2016, pp. 4152–4160.

[YZJ$^+$13] J. Yi, L. Zhang, R. Jin, Q. Qian, and A. Jain, *Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion*, International Conference on Machine Learning, 2013, pp. 1400–1408.

[ZL15] Q. Zheng and J. Lafferty, *A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements*, Advances in Neural Information Processing Systems, 2015, pp. 109–117.

[ZL16] ———, *Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent*, arXiv preprint arXiv:1605.07051 (2016).

[ZLTW17] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, *The global optimization geometry of nonsymmetric matrix factorization and sensing*, arXiv preprint arXiv:1703.01256 (2017).

[ZWL15] T. Zhao, Z. Wang, and H. Liu, *A nonconvex optimization framework for low rank matrix estimation*, Advances in Neural Information Processing Systems, 2015, pp. 559–567.