Model-free Nonconvex Matrix Completion: Local Minima Analysis and Applications in Memory-efficient Kernel PCA

Ji Chen   LJICHEN@UCDAVIS.EDU
Department of Mathematics
University of California, Davis
Davis, CA 95616, USA

Xiaodong Li   XDGLI@UCDAVIS.EDU
Department of Statistics
University of California, Davis
Davis, CA 95616, USA

Editor: Benjamin Recht

Abstract

This work studies low-rank approximation of a positive semidefinite matrix from partial entries via nonconvex optimization. We characterized how well local-minimum based low-rank factorization approximates a fixed positive semidefinite matrix without any assumptions on the rank-matching, the condition number or eigenspace incoherence parameter. Furthermore, under certain assumptions on rank-matching and well-boundedness of condition numbers and eigenspace incoherence parameters, a corollary of our main theorem improves the state-of-the-art sampling rate results for nonconvex matrix completion with no spurious local minima in Ge et al. (2016, 2017). In addition, we have investigated when the proposed nonconvex optimization results in accurate low-rank approximations even in presence of large condition numbers, large incoherence parameters, or rank mismatching. We also propose to apply the nonconvex optimization to memory-efficient kernel PCA. Compared to the well-known Nyström methods, numerical experiments indicate that the proposed nonconvex optimization approach yields more stable results in both low-rank approximation and clustering.

Keywords: low-rank approximation, matrix completion, nonconvex optimization, model-free analysis, local minimum analysis, kernel PCA

1. Introduction

Let $\boldsymbol{M}$ be an $n \times n$ positive semidefinite 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 $\boldsymbol{M} = \boldsymbol{U}_r \boldsymbol{U}_r^{\top}$. By the well-known Eckart-Young-Mirsky Theorem (Golub and Van Loan, 2012), $\boldsymbol{U}_r$ 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., Burer and Monteiro, 2003).

This paper studies 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 Ge et al. (2016, 2017) for matrix completion. Denote $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \end{bmatrix}^\top \in \mathbb{R}^{n \times r}$. A rank-$r$ approximation of $\boldsymbol{M}$ can be found through

$$
\min_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} f(\boldsymbol{X}) := \frac{1}{2} \sum_{(i,j) \in \Omega} \left(\boldsymbol{x}_i^\top \boldsymbol{x}_j - M_{ij} \right)^2 + \lambda \sum_{i=1}^n [(\|\boldsymbol{x}_i\|_2 - \alpha)_+]^4
$$

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

where $G_\alpha(\boldsymbol{X}) := \sum_{i=1}^n [(\|\boldsymbol{x}_i\|_2 - \alpha)_+]^4$. Following the framework of nonconvex optimization *without initialization* in Ge et al. (2016, 2017), 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).

Let's briefly discuss the memory and computational complexity to solve (1) via gradient descent. If Ω is symmetric and does not contain the diagonal entries as later specified in Definition 1, the updating rule of gradient decent

$$
\boldsymbol{X}^{(t+1)} = \boldsymbol{X}^{(t)} - \eta^{(t)} \nabla f(\boldsymbol{X}^{(t)}) \tag{2}
$$

is equivalent to

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

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 $O(|\Omega|r)$.

1.1. Applications in memory-efficient kernel PCA

Kernel PCA (Schölkopf et al., 1998) 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 $O(n^2 d)$. 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 (Williams and Seeger, 2001), which amounts to generating *random partial columns* of the kernel matrix, then finding a low-rank approximation based on these columns. In order to generate random partial columns, uniform sampling without replacement is employed in Williams and Seeger (2001), and different sampling strategies are proposed later (e.g., Drineas and Mahoney, 2005). 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 3.

Another popular approach is stochastic approximation, e.g., Kernel Hebbian Algorithm (KHA) (Kim et al., 2005), which is memory-efficient and approaches the exact principal component solution as the number of iterations goes to infinity with appropriately chosen learning rate (Kim et al., 2005). 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., Halko et al. (2011), 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 Definition 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 (1) is implemented through gradient descent (2). Any local minimum of (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 (Yi et al., 2016). However, to the best

of our knowledge, our work 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 Yi et al. (2016) while we make no assumptions on the positive semidefinite kernel matrix $\boldsymbol{M}$. Entry-sampling has been proposed in Achlioptas et al. (2002); Achlioptas and McSherry (2007) for scalable low-rank approximation. In particular, it is used to speed up kernel PCA in Achlioptas et al. (2002), but spectral methods are subsequently employed after entries sampling as opposed to nonconvex optimization. Empirical comparisons between spectral methods and nonconvex optimization will be demonstrated in Section 3. It is also noteworthy that matrix completion techniques have been applied to certain kernel matrices when it is costly to generate each single entry (Graepel, 2002; Paisley and Carin, 2010), wherein the proposed methods are not memory-efficient. In contrast, our method is memory-efficient in order to serve a different purpose.

1.2. Related work and our contributions

In recent years, a series of papers have been proposed to study nonconvex matrix completion (see, e.g., Rennie and Srebro, 2005; Keshavan et al., 2010b,a; Jain et al., 2013; Zhao et al., 2015; Sun and Luo, 2016; Chen and Wainwright, 2015; Yi et al., 2016; Zheng and Lafferty, 2016; Ge et al., 2016, 2017). Interested readers are referred to Balcan et al. (2017), where required sampling rates in these papers are summarized in Table 1 therein. Compared to convex approaches for matrix completion (e.g., Candès and Recht, 2009), these nonconvex approaches are not only more computationally efficient, but also more convenient in storing. For the same reason, nonconvex optimization approaches have also been investigated for other low-rank recovery problems including phase retirval (e.g., Candes et al., 2015; Sun et al., 2018; Cai et al., 2016), matrix sensing (e.g., Zheng and Lafferty, 2015; Tu et al., 2015), blind deconvolution (e.g., Li et al., 2018), etc.

Our present work follows the framework of local minimum analysis for nonconvex optimization in the literature. For example, Baldi and Hornik (1989) has described the nonconvex landscape of the quadratic loss for PCA. Loh and Wainwright (2015) studies the local minima of regularized M-estimators. Sun et al. (2018) studies the global geometry of the phase retrieval problem. The conditions for no spurious local minima have been investigated in Bhojanapalli et al. (2016) and Ge et al. (2016) 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 Li et al. (2016a). Global geometry of rectangular matrix factorization and sensing is studied Zhu et al. (2017), where the issues of under-parameterization and over-parameterization have been investigated. Similar analysis has been extended to general low-rank optimization problems in Li et al. (2017). Matrix factorization is further studied in Jin et al. (2017) with a novel geometric characterization of saddle points, and this idea is later extended in Ge et al. (2017), where a unified geometric analysis framework is proposed to study the landscapes of nonconvex matrix sensing, matrix completion and robust PCA.

Among these results, Ge et al. (2016) and Ge et al. (2017) are highly relevant to our work in both methodological and technical terms. In fact, exactly the same nonconvex optimization problem (1) has been studied in Ge et al. (2016, 2017) for matrix completion from missing data. To be specific, these papers show that any local minimum $\widehat{\boldsymbol{X}}$ yields $\boldsymbol{M} =$

$\widehat{\boldsymbol{X}\boldsymbol{X}^{\top}}$, as long as $\boldsymbol{M}$ is exactly rank-r, the condition number $\kappa_r := \sigma_1/\sigma_r$ 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 these quantities. The case with additive stochastic noise has also been discussed in Ge et al. (2016).

In contrast, our paper studies the theoretical properties of $\widehat{\boldsymbol{X}\boldsymbol{X}^{\top}}$ with *no assumptions* on $\boldsymbol{M}$. There are actually two questions of interest: how close $\widehat{\boldsymbol{X}\boldsymbol{X}^{\top}}$ is from $\boldsymbol{M}$, and how close $\widehat{\boldsymbol{X}\boldsymbol{X}^{\top}}$ is from $\boldsymbol{M}_r$ (recall that $\boldsymbol{M}_r$ is the best rank-r approximation of $\boldsymbol{M}$ by spectral truncation). In comparison to Ge et al. (2016, 2017), our main contributions to be introduced in the next section include the following:

- Our main result Theorem 2 that characterizes how well any local-minimum based rank-r factorization $\widehat{\boldsymbol{X}\boldsymbol{X}^{\top}}$ approximates $\boldsymbol{M}$ or $\boldsymbol{M}_r$ requires no assumptions imposed on $\boldsymbol{M}$ regarding its rank, eigenvalues and eigenvectors. The sampling rate is only required to satisfy $p \geqslant C(\log n/n)$ for some absolute constant C. Therefore, for applications such as memory-efficient kernel PCA, our framework provides more suitable guidelines than Ge et al. (2016, 2017). In fact, kernel matrices are in general of full rank and their condition numbers and incoherence parameters may not satisfy the strong assumptions in Ge et al. (2016, 2017).

- When $\boldsymbol{M}$ is assumed to be exactly low-rank as in Ge et al. (2016, 2017), Corollary 3 improves the state-of-the-art no-spurious-local-minima results in Ge et al. (2016, 2017) for exact nonconvex matrix completion in terms of sampling rates. To be specific, assuming both condition numbers and incoherence parameters are on the order of $O(1)$, our result improves the result in Ge et al. (2017) from $\widetilde{O}(r^4/n)$ to $\widetilde{O}(r^2/n)$.

- Theorem 2 also implies the conditions under which the nonconvex optimization (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, our paper benefits from Ge et al. (2016, 2017) in various aspects. In order to characterize the properties of any local minimum $\widehat{\boldsymbol{X}}$, we follow the idea in Ge et al. (2017) to combine the first and second order conditions of local minima linearly to construct an auxiliary function, denoted as $K(\boldsymbol{X})$ in our paper, and consequently all local minima satisfy the inequality $K(\widehat{\boldsymbol{X}}) \geqslant 0$ as illustrated in Figure 1. If $\boldsymbol{M}$ is exactly rank-r and its eigenvalues and eigenvectors satisfy particular properties, Ge et al. (2017) shows 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 is no spurious local minima.

However, $K(\boldsymbol{X}) \leqslant 0$ does not hold for all $\boldsymbol{X}$ if no assumptions are imposed on $\boldsymbol{M}$, so we instead focus on analyzing the inequality $K(\widehat{\boldsymbol{X}}) \geqslant 0$ directly in the model-free manner. Among a few novel technical ideas, the success of such model-free analysis relies crucially on the deterministic inequality (Lemma 8) that controls the difference between the function $K(\boldsymbol{X})$ and its population version $\mathbb{E}[K(\boldsymbol{X})]$.
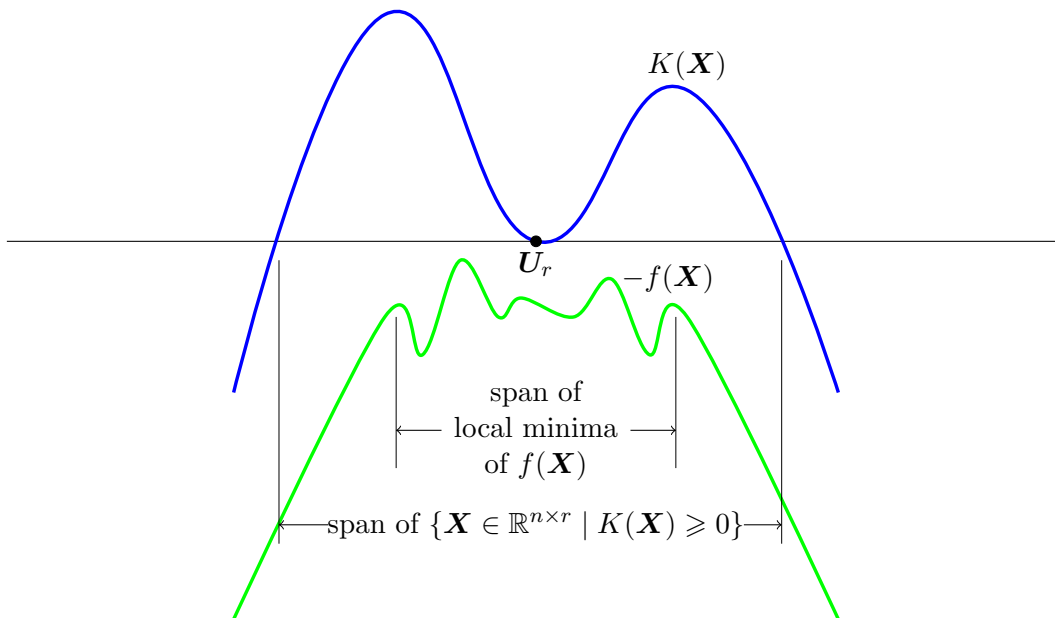
Figure 1: Landscape of $-f(\boldsymbol{X}), K(\boldsymbol{X})$ and $\boldsymbol{U}_r$.

1.3. Organization and notations

The remainder of the paper is organized as follows: Our main theoretical results are stated in Section 2; Numerical simulations and applications in memory-efficient KPCA are given in Section 3. Proofs are deferred to Section 4.

We use bold letters to denote matrices and vectors. For any vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, $\|\boldsymbol{u}\|_2$ denotes its ℓ_2 norm, and $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ their inner product. For any matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$, $M_{i,j}$ denotes its (i,j)-th entry, $\boldsymbol{M}_{i,\cdot} = (M_{i,1}, M_{i,2}, \ldots, M_{i,n})^\top$ its i-th row of $\boldsymbol{M}$, and $\boldsymbol{M}_{\cdot,j} = (M_{1,j}, M_{2,j}, \ldots, M_{n,j})^\top$ its j-th column. Moreover, we use $\|\boldsymbol{M}\|, \|\boldsymbol{M}\|_*, \|\boldsymbol{M}\|_F, \|\boldsymbol{M}\|_{\ell_\infty} := \max_{i,j} |M_{i,j}|, \|\boldsymbol{M}\|_{2,\infty} := \max_i \|\boldsymbol{M}_{i,\cdot}\|_2$ to denote its spectral norm, nuclear norm, Frobenius norm, elementwise max norm and $\ell_{2,\infty}$ norm, respectively. The vectorization of $\boldsymbol{M}$ is represented by $\mathrm{vec}(\boldsymbol{M}) = (M_{1,1}, M_{2,1}, \ldots, M_{1,2}, \ldots, M_{n,n})^\top$. For matrices $\boldsymbol{M}, \boldsymbol{N}$ of the same size, denote $\langle \boldsymbol{M}, \boldsymbol{N} \rangle = \sum_{i,j} M_{i,j} N_{i,j} = \mathrm{trace}\left(\boldsymbol{M}^\top \boldsymbol{N}\right)$. Denote by $\nabla f(\boldsymbol{M}) \in \mathbb{R}^{n \times n}$ and $\nabla^2 f(\boldsymbol{M}) \in \mathbb{R}^{n^2 \times n^2}$ the gradient and Hessian of $f(\boldsymbol{M})$.

Denote $[x]_+ = \max\{x, 0\}$. We use $\boldsymbol{J}$ to denote a matrix whose all entries equal to one. We use $C, C_1, C_2, \ldots$ to denote absolute constants, whose values may change from line to line.

2. Model-free approximation theory

2.1. Main results

The following sampling scheme is employed throughout the paper:

Definition 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 $\circ$ is the Hadamard product.

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

$$\boldsymbol{M} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top + \sum_{i=r+1}^{n} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top := \boldsymbol{M}_r + \boldsymbol{N}, \tag{3}$$

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 $\boldsymbol{M}_r := \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ is the best rank-r approximation of $\boldsymbol{M}$ and $\boldsymbol{N} := \sum_{i=r+1}^{n} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ denotes the residual part. In the case of multiple eigenvalues, the order in the eigenvalue decomposition (3) may not be unique. In this case, we consider the problem for any fixed order in (3) with the fixed $\boldsymbol{M}_r$.

Theorem 2 *Let $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ be a positive semidefinite matrix with the spectral decomposition (3). Let Ω be sampled according to the off-diagonal symmetric Ber(p) model with $p \geqslant C_S \frac{\log n}{n}$ for some absolute constant C_S. Then in an event E with probability $\mathbb{P}[E] \geqslant 1 - 2n^{-3}$, 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 $100\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \leqslant \lambda \leqslant 200\|\boldsymbol{\Omega} - p\boldsymbol{J}\|$, any local minimum $\widehat{\boldsymbol{X}} \in \mathbb{R}^{n \times r}$ of (1) satisfies*

$$\left\| \widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}_r \right\|_F^2 \leqslant C_1 \sum_{i=1}^{r} \left\{ \left[C_2 \left(\sqrt{\frac{n}{p}} + \frac{\log n}{p} \right) \|\boldsymbol{M}_r\|_{\ell_\infty} + C_2 \sigma_{2r+1-i} - \sigma_i \right]_+ \right\}^2 \tag{4}$$
$$+ C_1 \frac{[p(1-p)n + \log n] r \|\boldsymbol{N}\|_{\ell_\infty}^2}{p^2}$$

and

$$\left\| \widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M} \right\|_F^2 \leqslant C_1 \sum_{i=1}^{r} \left\{ \left[C_2 \left(\sqrt{\frac{n}{p}} + \frac{\log n}{p} \right) \|\boldsymbol{M}_r\|_{\ell_\infty} + C_2 \sigma_{2r+1-i} - \sigma_i \right]_+ \right\}^2 \tag{5}$$
$$+ C_1 \frac{[p(1-p)n + \log n] r \|\boldsymbol{N}\|_{\ell_\infty}^2}{p^2} + \|\boldsymbol{N}\|_F^2$$

with C_1, C_2 absolute constants.

Model-free low-rank approximation from partial entries has been studied for for spectral estimators in the literature. For example, under the settings of Theorem 2, the spectral low-rank approximation (denoted as $\boldsymbol{M}_{\text{approx}}$) discussed in Keshavan et al. (2010a, Theorem 1.1) is guaranteed to satisfy

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

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 Ge et al. (2017). Similarly, the SVD-based USVT estimator introduced in Chatterjee (2015) does not imply exact recovery. In contrast, as will be discussed in the next subsection, Theorem 2 implies that any local minimum of (1) yields exact recovery of $\boldsymbol{M}$ with high probability under milder conditions than those in Ge et al. (2017).

2.2. Implications in exact matrix completion

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

$$\boldsymbol{M} = \boldsymbol{M}_r = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top = \boldsymbol{U}_r \boldsymbol{U}_r^\top \tag{6}$$

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 $\mu_r = \frac{n}{r} \max_i \sum_{j=1}^{r} u_{i,j}^2$ (Candès and Recht, 2009) are well-bounded. This is a standard setup in the literature of nonconvex matrix completion (e.g., Keshavan et al., 2010b; Sun and Luo, 2016; Chen and Wainwright, 2015; Zheng and Lafferty, 2016; Ge et al., 2016; Yi et al., 2016; Ge et al., 2017).

Notice that Ge et al. (2016) introduces a slightly different version of incoherence

$$\widetilde{\mu}_r := \frac{\sqrt{n}\|\boldsymbol{U}_r\|_{2,\infty}}{\|\boldsymbol{U}_r\|_F} = \sqrt{\frac{n\|\boldsymbol{M}_r\|_{\ell_\infty}}{\text{trace}(\boldsymbol{M}_r)}} \tag{7}$$

as a measure of spikiness. (Note that this is different from the spikiness defined in Negahban and Wainwright (2012).) By $\|\boldsymbol{M}_r\|_{\ell_\infty} = \|\boldsymbol{U}_r\|_{2,\infty}^2 = \max_i \sum_{j=1}^{r} \sigma_j u_{i,j}^2$, the following relationship between μ and $\widetilde{\mu}$ is straightforward

$$\frac{\widetilde{\mu}_r^2}{\kappa_r} \leqslant \frac{\widetilde{\mu}_r^2 \text{trace}(\boldsymbol{M}_r)}{r\sigma_1} = \frac{n\|\boldsymbol{M}_r\|_{\ell_\infty}}{r\sigma_1} \leqslant \mu_r \leqslant \frac{n\|\boldsymbol{M}_r\|_{\ell_\infty}}{r\sigma_r} = \frac{\widetilde{\mu}_r^2 \text{trace}(\boldsymbol{M}_r)}{r\sigma_r} \leqslant \kappa_r \widetilde{\mu}_r^2. \tag{8}$$

By the fact $\|\boldsymbol{M}\|_{\ell_\infty} \leqslant \frac{r}{n}\sigma_1\mu_r$, Theorem 2 implies the following exact low-rank recovery results:

Corollary 3 *Under the assumptions of Theorem 2, if we further assume* $\text{rank}(\boldsymbol{M}) = r$ *(i.e., $\boldsymbol{M} = \boldsymbol{M}_r$) and*

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

8

or

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

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 objective function $f(\boldsymbol{X})$ defined in (1) satisfies $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top = \boldsymbol{M}$.

The proof is straightforward and deferred to the appendix. Notice that our results are better than the state-of-the-art results for no spurious local minimum in Ge et al. (2017), 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 (8)).

2.3. Examples

Besides improving the state-of-the-art no-spurious-local-minima results in nonconvex matrix completion, Theorem 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.

2.3.1. 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 (6). 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, 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), $\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\|_F^2 \leqslant \frac{1}{100} \sigma_{r-1}^2$ holds. In other words, the relative approximation error satisfies $RE := \frac{\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F} \leqslant \frac{1}{10\sqrt{r-1}}$.

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

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

2.3.2. 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 2.2, 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 (8), there holds $\|\boldsymbol{M}\|_{\ell_\infty} \leqslant \frac{R}{n}\sigma_1\mu_R$.

Case 1: $R < r$. Theorem 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) 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 $RE := \frac{\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F} \leqslant \frac{1}{10}\sqrt{\frac{r-R}{R}}$.

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

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

Theorem 2 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) 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, under the same sampling rate requirement as above, Theorem 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

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

3. Experiments

In the following simulations where the nonconvex optimization (1) is solved, 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) is determined by Armijo's rule (Armijo, 1966). The gradient descent algorithm is implemented with sparse matrix storage in Section 3.2 for the purpose of memory-efficient KPCA, while with full matrix storage in Section 3.1 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.

3.1. Numerical simulations

In this section, we conduct numerical tests on the nonconvex optimization (1) 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 (1) is implemented for 50 times with independent Ω's generated under the off-diagonal symmetric independent $\text{Ber}(p)$ model. To implement the gradient descent algorithm (2), 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 Achlioptas et al. (2002) to obtain an approximate low-rank approximation of $\boldsymbol{M}$ for the purpose of comparison.

3.1.1. FULL RANK CASE

Here $\boldsymbol{M}$ is assumed to have full rank, i.e., $\text{rank}(\boldsymbol{M}) = 500$. 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 (1) 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. One can observe that the relative errors for our nonconvex method (1) 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.

3.1.2. 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 $\text{rank}(\boldsymbol{M}) = 5$ if $\kappa < \infty$ while $\text{rank}(\boldsymbol{M}) = 4$ if $\kappa = \infty$. The selected rank is always assumed to be $r = 5$, while the sampling rate is always $p = 0.2$.

(a) Relative error $\frac{\|M_{\mathrm{approx}} - M_r\|_F}{\|M_r\|_F}$.

(b) Relative error $\frac{\|M_{\mathrm{approx}} - M\|_F}{\|M\|_F}$.
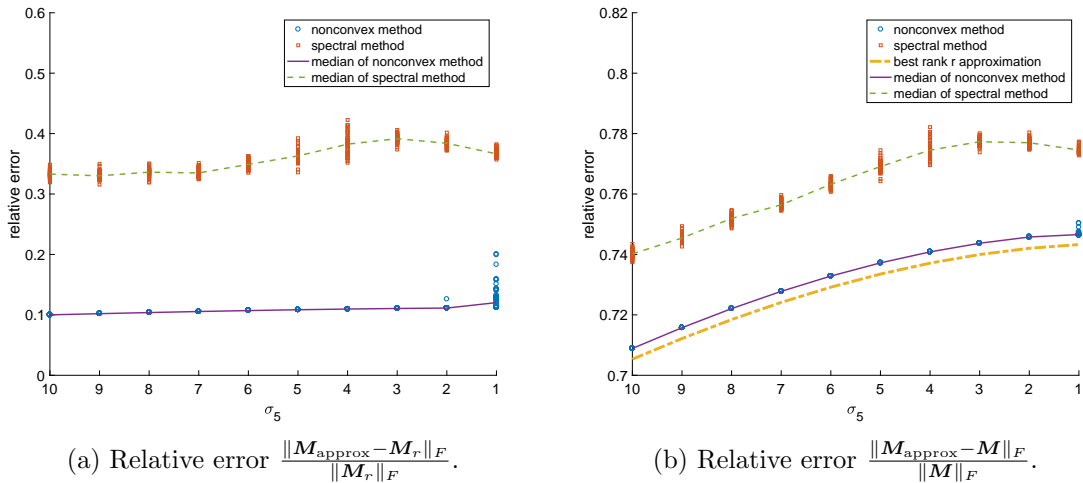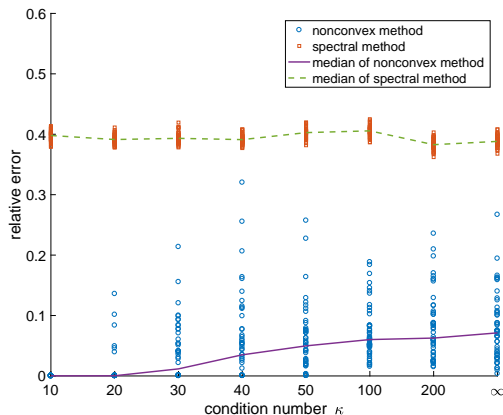
Figure 2: Relative errors for full rank case.

Figure 3: Relative error $\frac{\|M_{\mathrm{approx}} - M\|_F}{\|M\|_F}$ for low-rank matrix with extreme condition numbers.

The performance of our nonconvex approach with various choices of κ is demonstrated in Figure 3. One can observe that our nononvex optimization approach yields exact recovery of M when $\kappa = 10$, while yields accurate low-rank approximation for 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 2.3.1, where we have shown that under certain incoherence conditions, the relative approximation error can be well-bounded even when $\kappa_r = \infty$.

3.1.3. RANK MISMATCHING

In this section, we consider rank mismatching, i.e., the rank of M is low but different from the selected rank r. In particular, we consider two settings for simulation: First, we fix M with $\mathrm{rank}(M) = 10$, while the nonconvex optimization is implemented with selected rank $r = 5, 7, 9, 10, 11, 13, 15$; Second, the matrix 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 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 rank($\boldsymbol{M}$) are demonstrated in Figure 4 (constant nonzero eigenvalues) and Figure 6 (decreasing nonzero eigenvalues), while the case of fixed selected rank in Figure 5 (constant nonzero eigenvalues) and Figure 7 (decreasing nonzero eigenvalues). One can observe from these figures that if the selected rank r is less than the actual rank 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 $\frac{\|\boldsymbol{M}_{\text{approx}} - \boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F}$.
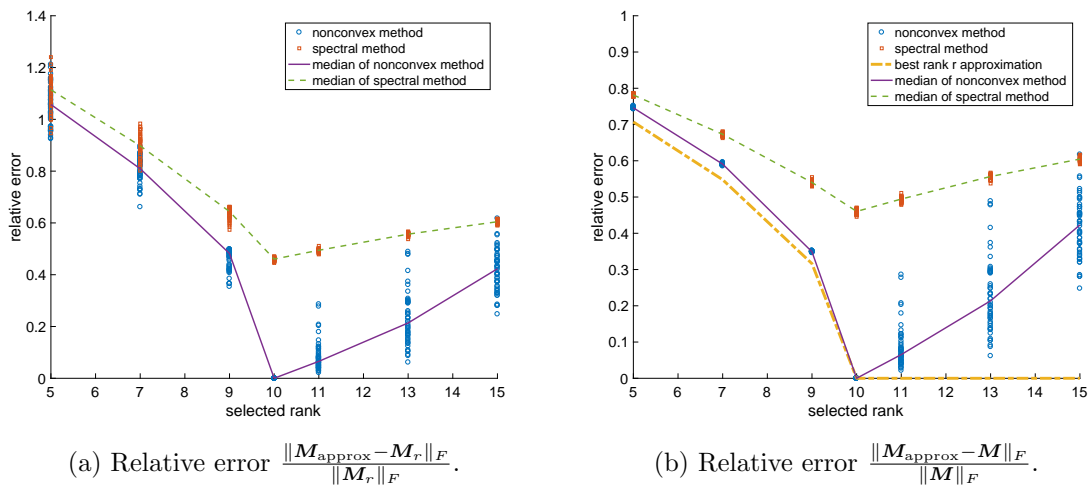
Figure 4: Relative errors for rank mismatching for a fixed $\boldsymbol{M}$ with rank($\boldsymbol{M}$) = 10.

3.2. Memory-efficient kernel PCA

In this section we study the empirical performance of our memory-efficient kernel PCA approach by applying it to the synthetic data set in Wang (2012). 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 $\mathcal{N}(\boldsymbol{0}, \frac{1}{100}\boldsymbol{I}_3)$ 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 (Williams and Seeger, 2001), 50 columns (and corresponding rows) are selected uniformly at random without replacement, then a rank-2

(a) Relative error $\frac{\|M_{\text{approx}}-M_r\|_F}{\|M_r\|_F}$.

(b) Relative error $\frac{\|M_{\text{approx}}-M\|_F}{\|M\|_F}$.
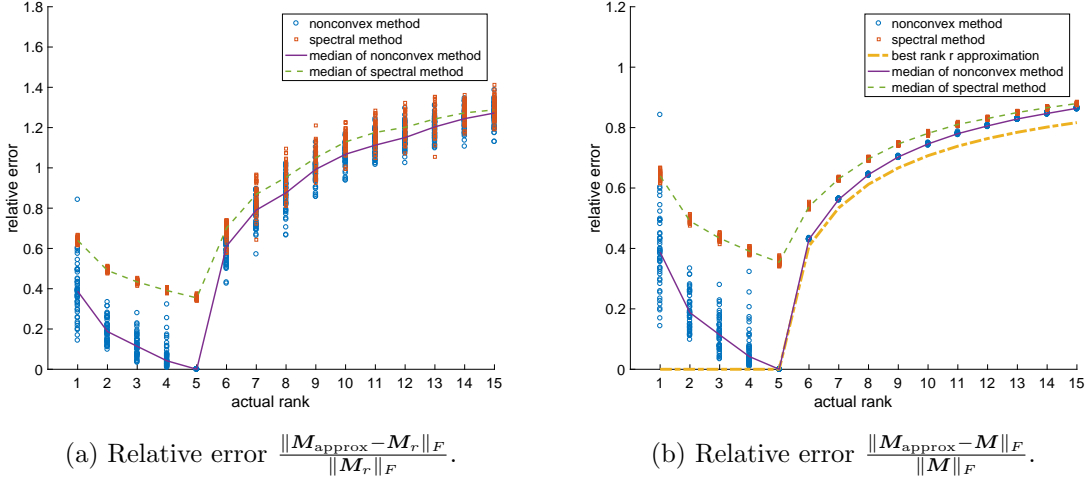
Figure 5: Relative errors for rank mismatching, fixed selected rank.

(a) Relative error $\frac{\|M_{\text{approx}}-M_r\|_F}{\|M_r\|_F}$.

(b) Relative error $\frac{\|M_{\text{approx}}-M\|_F}{\|M\|_F}$.
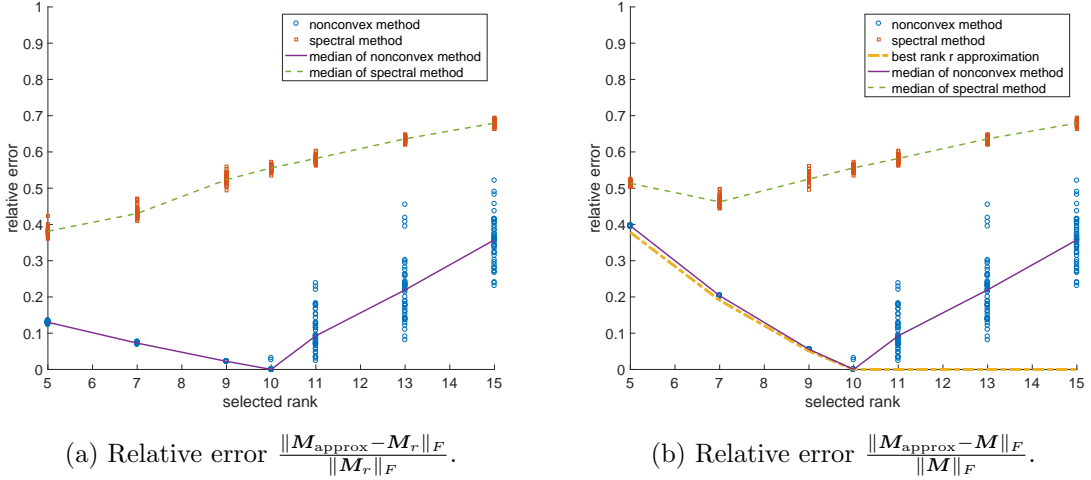
Figure 6: Relative errors for rank mismatching for a fixed M with rank$(M) = 10$.

approximation of the kernel matrix 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 (Saad, 2003), 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 (1) such that the memory consumption is less costly than the Nyström method.

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

(b) Relative error $\frac{\|\boldsymbol{M}_{\mathrm{approx}} - \boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F}$.
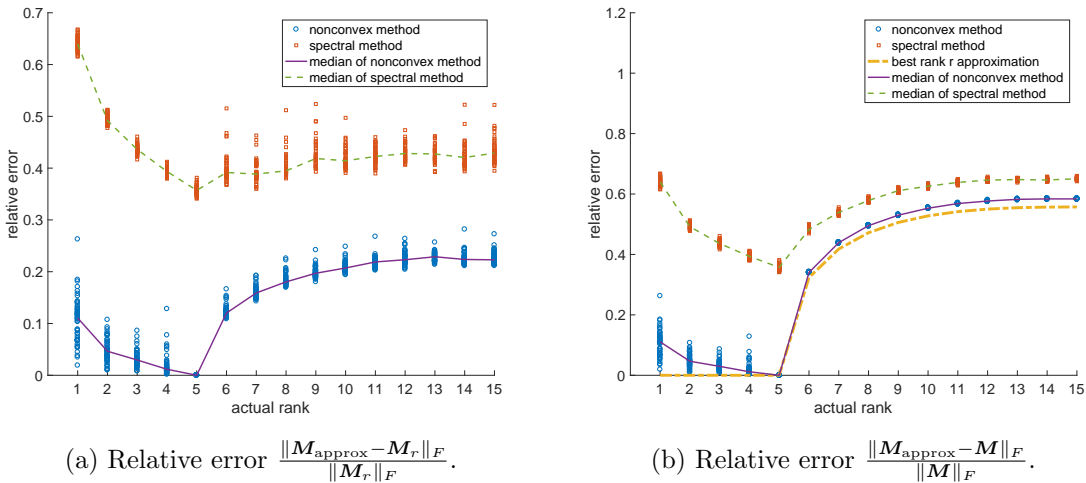
Figure 7: Relative errors for rank mismatching, fixed selected rank.

Fixing such a synthetic data set, we apply both the Nyström method and our approach (with $\alpha = 100\|\boldsymbol{M}\|_{\ell_\infty} = 100$ and $\lambda = 500\sqrt{np_{\mathrm{NCVX}}}$) 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 $\boldsymbol{M}_{\mathrm{approx}}$ the memory efficient rank-2 approximation obtained by Nyström method or our nonconvex optimization. The left and right panels of Figure 8 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 (1) 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 9 compares the two methods in the distribution of clustering accuracies. It clearly shows that our nonconvex optimization (1) 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 $\lambda = 0$.

4. Proofs

In this section, we give a proof for main theorem. In Section 4.1, we will present some useful supporting lemmas; in Section 4.2, we present a proof for our main result Theorem 2; finally in Section 4.3 we give proof of lemmas used in former subsections. Our proof ideas benefit from those in Ge et al. (2017) as well as Zhu et al. (2017), Jin et al. (2017).

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

(b) Relative error $\frac{\|\boldsymbol{M}_{\mathrm{approx}} - \boldsymbol{M}\|_F}{\|\boldsymbol{M}\|_F}$.
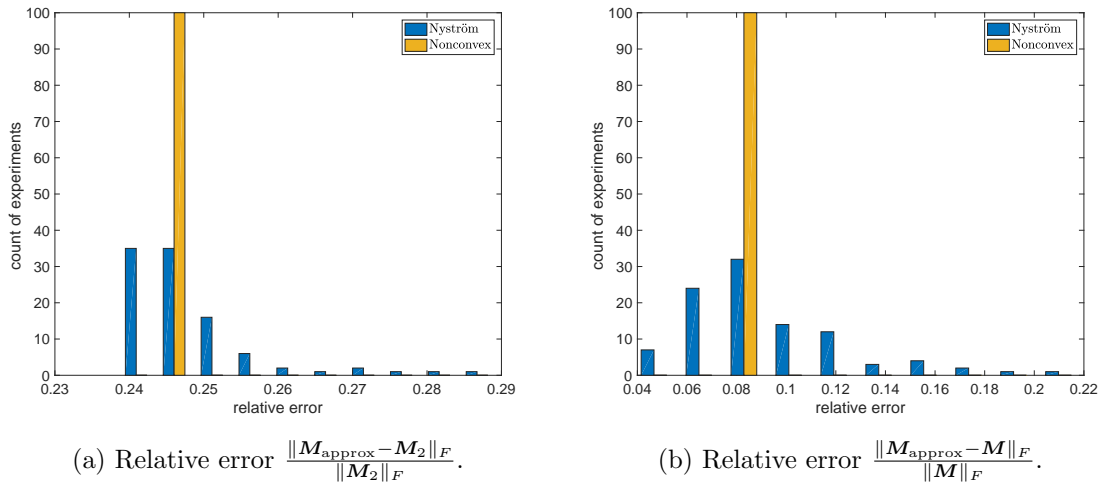
Figure 8: Relative errors for Nyström method with sampling rate $p_{\mathrm{Nys}} \approx 0.01$ and nonconvex method with sampling rate $p_{\mathrm{NCVX}} = \frac{p_{\mathrm{Nys}}}{2.5}$.
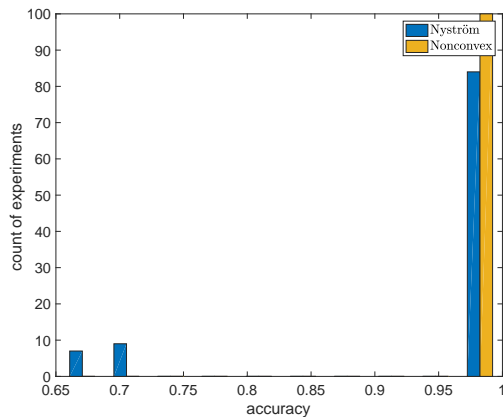
Figure 9: Clustering accuracy for Nyström method with sampling rate $p_{\mathrm{Nys}} \approx 0.01$ and nonconvex method with sampling rate $p_{\mathrm{NCVX}} = \frac{p_{\mathrm{Nys}}}{2.5}$.

4.1. Supporting lemmas

In this section, we give some useful supporting lemmas. The following lemma is well known in the literature, see, e.g., Vu (2018) and Bandeira et al. (2016).

Lemma 4 *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, then*

$$\mathbb{P}\left[\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \geqslant C_v\sqrt{np(1-p)} + C_v\sqrt{\log n}\right] \leqslant n^{-3}.$$

The following eigen-space incoherence parameter has been proposed in Candès and Recht (2009).

Definition 5 (Candès and Recht 2009) *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*

$$\mu(\mathcal{U}) := \frac{n}{r} \max_{1 \leqslant i \leqslant n} \|\boldsymbol{P}_{\mathcal{U}} \boldsymbol{e}_i\|_2^2, \tag{9}$$

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

As with Theorem 4.1 in Candès and Recht (2009), for the off-diagonal symmetric Ber(p) model, we also have:

Lemma 6 *Let Ω be sampled according to the off-diagonal symmetric Ber(p) model. Define*

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

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_c, such that for any $\delta \in (0, 1]$, if $p \geqslant C_c \frac{\mu(\mathcal{U}) \dim(\mathcal{U}) \log n}{\delta^2 n}$ with $\mu(\mathcal{U})$ defined in (9), in an event E_c with probability $\mathbb{P}[E_c] \geqslant 1 - n^{-3}$, we have

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

In Gross (2011) and Gross and Nesme (2010), similar results are given for symmetric uniform sampling with/without replacement. The proof of Lemma 6 is very similar to that in Recht (2011).

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

Lemma 7 (Ge et al. 2016, Proposition 4.1) *The first order optimality condition of objective function (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:

$$D_{\Omega_0, t}(\boldsymbol{M}_1, \boldsymbol{M}_2) := \langle \mathcal{P}_{\Omega_0}(\boldsymbol{M}_1), \mathcal{P}_{\Omega_0}(\boldsymbol{M}_2)\rangle - t\langle \boldsymbol{M}_1, \boldsymbol{M}_2 \rangle. \tag{10}$$

Our key lemma is given as follows:

Lemma 8 *Let Ω_0 be any index set in $[n_1] \times [n_2]$, and $\mathbf{\Omega}_0 \in \mathbb{R}^{n_1 \times n_2}$ be defined correspondingly as in Section 2.1. For any $\mathbf{A} \in \mathbb{R}^{n_1 \times r_1}, \mathbf{B} \in \mathbb{R}^{n_1 \times r_2}, \mathbf{C} \in \mathbb{R}^{n_2 \times r_1}, \mathbf{D} \in \mathbb{R}^{n_2 \times r_2}$, and any $t \in \mathbb{R}$, there holds*

$$|D_{\Omega_0,t}(\mathbf{A}\mathbf{C}^\top, \mathbf{B}\mathbf{D}^\top)| \leqslant \|\mathbf{\Omega}_0 - t\mathbf{J}\| \sqrt{\sum_{k=1}^{n_1} \|\mathbf{A}_{k,\cdot}\|_2^2 \|\mathbf{B}_{k,\cdot}\|_2^2} \sqrt{\sum_{k=1}^{n_2} \|\mathbf{C}_{k,\cdot}\|_2^2 \|\mathbf{D}_{k,\cdot}\|_2^2}. \quad (11)$$

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 4.3.1. This result extends the following lemma given in Bhojanapalli and Jain (2014) and Li et al. (2016b):

Lemma 9 (Bhojanapalli and Jain 2014; Li et al. 2016b) *Suppose matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$ can be decomposed as $\mathbf{M} = \mathbf{B}\mathbf{D}^\top$, let $\Omega_0 \subset [n_1] \times [n_2]$ be any index set. Then for any $t \in \mathbb{R}$, we have*

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

Lemma 8 is applied in our proof of Lemma 12 in replace of Theorem D.1 in Ge et al. (2016) to derive tighter control of perturbation terms, i.e., $K_2(\mathbf{X}), K3(\mathbf{X})$ and $K4(\mathbf{X})$ defined in (14). Their result is given here for the purpose of comparison.

Lemma 10 (Ge et al. 2016, Theorem D.1) *With high probability over the choice of Ω, for any two rank-r matrices $\mathbf{W}, \mathbf{Z} \in \mathbb{R}^{n \times n}$, we have*

$$|\langle \mathcal{P}_\Omega(\mathbf{W}), \mathcal{P}_\Omega(\mathbf{Z})\rangle - p\langle \mathbf{W}, \mathbf{Z}\rangle|$$
$$\leqslant O\left(\|\mathbf{W}\|_{\ell_\infty}\|\mathbf{Z}\|_{\ell_\infty} nr \log n + \sqrt{pnr\|\mathbf{W}\|_{\ell_\infty}\|\mathbf{Z}\|_{\ell_\infty}\|\mathbf{W}\|_F\|\mathbf{Z}\|_F \log n}\right).$$

In Sun and Luo (2016), Chen and Wainwright (2015) and Zheng and Lafferty (2016), upper bounds are given to $\|\mathcal{P}_\Omega(\mathbf{H}\mathbf{H}^\top)\|_F^2$ for any $\mathbf{H}$. 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

$$\|\mathcal{P}_\Omega(\mathbf{H}\mathbf{H}^\top)\|_F^2 - p\|\mathbf{H}\|_F^4 \leqslant C\sqrt{np}\sum_{i=1}^n \|\mathbf{H}_{i,\cdot}\|_2^4 \quad (12)$$

with high probability. In contrast, by combining Lemma 4 and Lemma 8, there holds

$$\left|\|\mathcal{P}_\Omega(\mathbf{H}\mathbf{H}^\top)\|_F^2 - p\|\mathbf{H}\mathbf{H}^\top\|_F^2\right| \leqslant C\sqrt{np}\sum_{i=1}^n \|\mathbf{H}_{i,\cdot}\|_2^4 \quad (13)$$

with high probability. This is tighter than (12) in that $\|\mathbf{H}\mathbf{H}^\top\|_F \leqslant \|\mathbf{H}\|_F^2$. Moreover, comparing to (12), our result (13) directly measures the difference between $\|\mathcal{P}_\Omega(\mathbf{H}\mathbf{H}^\top)\|_F^2$ and its expectation $p\|\mathbf{H}\mathbf{H}^\top\|_F^2$, which makes the model-free analysis possible.

4.2. A proof of Theorem 2

This section aims to prove Theorem 2. The proof is basically divided into two parts: In Section 4.2.1, 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 4.2.2, we give a uniform upper bound of $K(\boldsymbol{X})$ in order to control the above superlevel set.

4.2.1. Landscape of objective function f and auxiliary function K

Denote $\boldsymbol{U}_r := [\sqrt{\sigma_1}\boldsymbol{u}_1 \ \ldots \ \sqrt{\sigma_r}\boldsymbol{u}_r]$. 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 Jin et al. (2017) and Ge et al. (2017):

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

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

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

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 7. By repacking terms in Ge et al. (2017, Lemma 7), and given $\langle \boldsymbol{U}\boldsymbol{\Delta}^\top, \boldsymbol{N}\rangle = 0$, due to the definition of $\boldsymbol{U}$ and $\boldsymbol{N}$, $K(\boldsymbol{X})$ can be decomposed as follows:

Lemma 11 (Ge et al. 2017, 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*

$$
\begin{aligned}
K(\boldsymbol{X}) = &\underbrace{p\left(\|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 - 3\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2\right)}_{K_1(\boldsymbol{X})} \\
&+ \underbrace{D_{\Omega,p}(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top) - 3D_{\Omega,p}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top)}_{K_2(\boldsymbol{X})} \\
&+ \underbrace{\lambda\left(\mathrm{vec}(\boldsymbol{\Delta})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \, \mathrm{vec}(\boldsymbol{\Delta}) - 4\langle \nabla G_\alpha(\boldsymbol{X}), \boldsymbol{\Delta}\rangle\right)}_{K_3(\boldsymbol{X})} \\
&+ \underbrace{6D_{\Omega,p}(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{N}) + 8D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top, \boldsymbol{N}) + 6p\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{N}\rangle}_{K_4(\boldsymbol{X})},
\end{aligned}
\tag{14}
$$

where $D_{\Omega,p}(\cdot, \cdot)$ is defined in (10).

Notice that in Theorem 2, we are only concerned about the difference between $\boldsymbol{XX}^\top$ and $\boldsymbol{M}_r$ (or $\boldsymbol{M}$), which remains the same by replacing $\boldsymbol{X}$ with $\widetilde{\boldsymbol{X}} = \boldsymbol{XR}$, 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 $\boldsymbol{R}_{\boldsymbol{XR},\boldsymbol{U}_r} = \boldsymbol{R}_{\boldsymbol{X},\boldsymbol{U}_r}\boldsymbol{R}$ for any $\boldsymbol{R} \in \mathsf{O}(r)$, which implies $\widetilde{\boldsymbol{U}} = \boldsymbol{UR}$ and $\widetilde{\boldsymbol{\Delta}} = \boldsymbol{\Delta R}$. Now we have

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

which means $K_i(\widetilde{\boldsymbol{X}}) = K_i(\boldsymbol{X})$ for $i = 1, 2, 4$. As for K_3, by Ge et al. (2017, 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 (4) and (5) 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., $\boldsymbol{U} = \boldsymbol{U}_r$.

4.2.2. PROOF OF THEOREM 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 $\widehat{\boldsymbol{X}}$. For simplicity of notations, denote $\nu_r := \|\boldsymbol{M}_r\|_{\ell_\infty}$.

Lemma 12 *Assume that tuning parameters α, λ satisfy $100\sqrt{\nu_r} \leqslant \alpha \leqslant 200\sqrt{\nu_r}$, $100\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \leqslant \lambda \leqslant 200\|\boldsymbol{\Omega} - p\boldsymbol{J}\|$, and $p \geqslant C_S \frac{\log n}{n}$ with some absolute constant C_S. Then, in an event E with probability $\mathbb{P}[E] \geqslant 1 - 2n^{-3}$, uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n \times r}$ and corresponding $\boldsymbol{\Delta}$ defined as before, we have*

$$
\begin{aligned}
\sum_{i=2}^4 K_i(\boldsymbol{X}) \leqslant &10^{-3} p \left[\|\boldsymbol{\Delta}^\top \boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U\Delta}^\top\|_F^2 \right] \\
&+ C_3 p \sum_{i=1}^r \left\{ \left[C_4 \left(\sqrt{\frac{n}{p}} + \frac{\log n}{p} \right) \nu_r + C_4 \sigma_{2r+1-i} - \sigma_i \right]_+ \right\}^2 \quad (15) \\
&+ C_3 \frac{[p(1-p)n + \log n] r \|\boldsymbol{N}\|_{\ell_\infty}^2}{p}.
\end{aligned}
$$

Note in our proof of Theorem 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 $\mathbf{\Delta}$,

$$
\begin{aligned}
\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2 &= \|\boldsymbol{U}\mathbf{\Delta}^\top + \mathbf{\Delta}\boldsymbol{U}^\top + \mathbf{\Delta}\mathbf{\Delta}^\top\|_F^2 \\
&= \|\mathbf{\Delta}\mathbf{\Delta}^\top\|_F^2 + 2\|\mathbf{\Delta}\boldsymbol{U}^\top\|_F^2 + 2\langle\mathbf{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\mathbf{\Delta}^\top\rangle + 4\langle\mathbf{\Delta}\mathbf{\Delta}^\top, \boldsymbol{U}\mathbf{\Delta}^\top\rangle.
\end{aligned} \tag{16}
$$

By the definition of matrix inner product, we have

$$
\begin{aligned}
\|\boldsymbol{U}\mathbf{\Delta}^\top\|_F^2 &= \langle\boldsymbol{U}\mathbf{\Delta}^\top, \boldsymbol{U}\mathbf{\Delta}^\top\rangle = \operatorname{trace}(\mathbf{\Delta}\boldsymbol{U}^\top\boldsymbol{U}\mathbf{\Delta}^\top) = \operatorname{trace}(\boldsymbol{U}^\top\boldsymbol{U}\mathbf{\Delta}^\top\mathbf{\Delta}) \\
&= \langle\boldsymbol{U}^\top\boldsymbol{U}, \mathbf{\Delta}^\top\mathbf{\Delta}\rangle,
\end{aligned} \tag{17}
$$

and

$$
\langle\mathbf{\Delta}\mathbf{\Delta}^\top, \boldsymbol{U}\mathbf{\Delta}^\top\rangle = \operatorname{trace}(\mathbf{\Delta}\mathbf{\Delta}^\top\boldsymbol{U}\mathbf{\Delta}^\top) = \operatorname{trace}(\mathbf{\Delta}^\top\mathbf{\Delta}\mathbf{\Delta}^\top\boldsymbol{U}) = \langle\mathbf{\Delta}^\top\mathbf{\Delta}, \mathbf{\Delta}^\top\boldsymbol{U}\rangle. \tag{18}
$$

Here we use the fact that $\operatorname{trace}(\boldsymbol{A}\boldsymbol{B}) = \operatorname{trace}(\boldsymbol{B}\boldsymbol{A})$ 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, $\boldsymbol{U}^\top\mathbf{\Delta} = \mathbf{\Delta}^\top\boldsymbol{U}$ and $\boldsymbol{U}^\top(\mathbf{\Delta} + \boldsymbol{U}) \succeq \mathbf{0}$. Therefore, we also have

$$
\begin{aligned}
\langle\mathbf{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\mathbf{\Delta}^\top\rangle &= \operatorname{trace}(\boldsymbol{U}\mathbf{\Delta}^\top\boldsymbol{U}\mathbf{\Delta}^\top) = \operatorname{trace}(\mathbf{\Delta}\boldsymbol{U}^\top\mathbf{\Delta}\boldsymbol{U}^\top) = \operatorname{trace}(\boldsymbol{U}^\top\mathbf{\Delta}\boldsymbol{U}^\top\mathbf{\Delta}) \\
&= \langle\mathbf{\Delta}^\top\boldsymbol{U}, \boldsymbol{U}^\top\mathbf{\Delta}\rangle = \langle\mathbf{\Delta}^\top\boldsymbol{U}, \mathbf{\Delta}^\top\boldsymbol{U}\rangle = \|\mathbf{\Delta}^\top\boldsymbol{U}\|_F^2
\end{aligned} \tag{19}
$$

and

$$
\langle\mathbf{\Delta}^\top\mathbf{\Delta}, \boldsymbol{U}^\top\boldsymbol{U} + \mathbf{\Delta}^\top\boldsymbol{U}\rangle = \langle\mathbf{\Delta}^\top\mathbf{\Delta}, (\boldsymbol{U} + \mathbf{\Delta})^\top\boldsymbol{U}\rangle \geqslant 0. \tag{20}
$$

Here (20) also uses the fact that inner product of two positive semidefinite matrices is non-negative.

Now denote $a := \|\mathbf{\Delta}^\top\mathbf{\Delta}\|_F = \|\mathbf{\Delta}\mathbf{\Delta}^\top\|_F, b := \|\mathbf{\Delta}^\top\boldsymbol{U}\|_F$ and

$$
\psi := C_3\left\{\sum_{i=1}^r\left\{\left[C_4\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\nu_r + C_4\sigma_{2r+1-i} - \sigma_i\right]_+\right\}^2 + \frac{[p(1-p)n + \log n]r\|\boldsymbol{N}\|_{\ell_\infty}^2}{p^2}\right\}.
$$

Putting Lemma 11 and Lemma 12 together, and using (16), we have

$$
\begin{aligned}
\frac{K(\boldsymbol{X})}{p} &\leqslant 1.001\|\mathbf{\Delta}\mathbf{\Delta}^\top\|_F^2 - 3\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2 + 10^{-3}\|\boldsymbol{U}\mathbf{\Delta}^\top\|_F^2 + \psi \\
&= 1.001a^2 - 3\left[\|\mathbf{\Delta}\mathbf{\Delta}^\top\|_F^2 + 2\|\mathbf{\Delta}\boldsymbol{U}^\top\|_F^2 + 2\langle\mathbf{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\mathbf{\Delta}^\top\rangle + 4\langle\mathbf{\Delta}\mathbf{\Delta}^\top, \boldsymbol{U}\mathbf{\Delta}^\top\rangle\right] \\
&\quad + 10^{-3}\|\boldsymbol{U}\mathbf{\Delta}^\top\|_F^2 + \psi,
\end{aligned} \tag{21}
$$

By putting (17), (18), (19), (21) together,

$$
\begin{aligned}
\frac{K(\boldsymbol{X})}{p} &\leqslant 1.001a^2 - 3\|\mathbf{\Delta}\mathbf{\Delta}^\top\|_F^2 - 6\|\mathbf{\Delta}\boldsymbol{U}^\top\|_F^2 - 6\langle\mathbf{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\mathbf{\Delta}^\top\rangle - 12\langle\mathbf{\Delta}\mathbf{\Delta}^\top, \boldsymbol{U}\mathbf{\Delta}^\top\rangle \\
&\quad + 10^{-3}\|\boldsymbol{U}\mathbf{\Delta}^\top\|_F^2 + \psi \\
&= -1.999a^2 - 6\langle\boldsymbol{U}^\top\boldsymbol{U}, \mathbf{\Delta}^\top\mathbf{\Delta}\rangle - 6\|\mathbf{\Delta}^\top\boldsymbol{U}\|_F^2 - 12\langle\mathbf{\Delta}^\top\mathbf{\Delta}, \mathbf{\Delta}^\top\boldsymbol{U}\rangle \\
&\quad + 10^{-3}\langle\boldsymbol{U}^\top\boldsymbol{U}, \mathbf{\Delta}^\top\mathbf{\Delta}\rangle + \psi \\
&= -1.999a^2 - \langle\mathbf{\Delta}^\top\mathbf{\Delta}, 5.999\boldsymbol{U}^\top\boldsymbol{U} + 12\mathbf{\Delta}^\top\boldsymbol{U}\rangle - 6b^2 + \psi.
\end{aligned} \tag{22}
$$

Therefore, combining with (20),

$$\frac{K(\boldsymbol{X})}{p} \leqslant - 1.999a^2 - 6.001\langle \boldsymbol{\Delta}^\top\boldsymbol{\Delta}, \boldsymbol{\Delta}^\top\boldsymbol{U}\rangle - 6b^2 + \psi \tag{23}$$
$$\leqslant - 1.999a^2 + 6.001ab - 6b^2 + \psi$$

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

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

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 (23), there holds

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

which further implies

$$0 \leqslant a \leqslant C_5\sqrt{\psi}, \qquad 0 \leqslant b \leqslant C_5\sqrt{\psi}. \tag{24}$$

From (22), we have

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

Recall from (17), $\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 = \langle \boldsymbol{U}^\top\boldsymbol{U}, \boldsymbol{\Delta}^\top\boldsymbol{\Delta}\rangle$, and $K(\widehat{\boldsymbol{X}}) \geqslant 0$. Therefore, combining with (24),

$$\begin{aligned}
5.999\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 &\leqslant - 1.999a^2 - \langle \boldsymbol{\Delta}^\top\boldsymbol{\Delta}, 12\boldsymbol{\Delta}^\top\boldsymbol{U}\rangle - 6b^2 + \psi \\
&\leqslant - 1.999a^2 + 12\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F\|\boldsymbol{\Delta}^\top\boldsymbol{U}\|_F - 6b^2 + \psi \\
&\leqslant - 1.999a^2 + 12ab - 6b^2 + \psi \\
&\leqslant C_6\psi.
\end{aligned} \tag{25}$$

From (21),

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

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 (24), (25), we futher have

$$3\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2 \leqslant 1.001a^2 + C_7\psi + \psi \leqslant C_8\psi. \tag{26}$$

Therefore, (4) is directly implied by (26). Notice that

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

where the inequality holds since $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top \succeq \boldsymbol{0}$ and $\boldsymbol{N} \succeq \boldsymbol{0}$. Therefore, (5) is implied by (4).

4.3. Proofs of supporting lemmas

We present in this section the proofs of lemmas stated in previous sections.

4.3.1. A proof of Lemma 8

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

$$
\begin{aligned}
|\langle \mathcal{P}_{\Omega_0}(\boldsymbol{AC}^\top), \mathcal{P}_{\Omega_0}(\boldsymbol{BD}^\top)\rangle - t\langle \boldsymbol{AC}^\top, \boldsymbol{BD}^\top\rangle| &= |\langle \boldsymbol{\Omega}_0 - t\boldsymbol{J}, (\boldsymbol{AC}^\top \circ \boldsymbol{BD}^\top)\rangle| \\
&\leqslant \|\boldsymbol{\Omega}_0 - t\boldsymbol{J}\|\|(\boldsymbol{AC}^\top \circ \boldsymbol{BD}^\top)\|_*,
\end{aligned}
\tag{27}
$$

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

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

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

$$
\begin{aligned}
\|(\boldsymbol{AC}^\top \circ \boldsymbol{BD}^\top)\|_* &\leqslant \sum_{l=1}^{r_1}\sum_{m=1}^{r_2} \|(\boldsymbol{A}_{\cdot,l} \circ \boldsymbol{B}_{\cdot,m})(\boldsymbol{C}_{\cdot,l} \circ \boldsymbol{D}_{\cdot,m})^\top\|_* \\
&= \sum_{l=1}^{r_1}\sum_{m=1}^{r_2} \|\boldsymbol{A}_{\cdot,l} \circ \boldsymbol{B}_{\cdot,m}\|_2\|\boldsymbol{C}_{\cdot,l} \circ \boldsymbol{D}_{\cdot,m}\|_2 \\
&= \sum_{l=1}^{r_1}\sum_{m=1}^{r_2} \sqrt{\sum_{k=1}^{n_1} A_{k,l}^2 B_{k,m}^2}\sqrt{\sum_{k=1}^{n_2} C_{k,l}^2 D_{k,m}^2},
\end{aligned}
$$

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

$$
\begin{aligned}
\|(\boldsymbol{AC}^\top \circ \boldsymbol{BD}^\top)\|_* &\leqslant \sqrt{\sum_{l=1}^{r_1}\sum_{m=1}^{r_2}\sum_{k=1}^{n_1} A_{k,l}^2 B_{k,m}^2}\sqrt{\sum_{l=1}^{r_1}\sum_{m=1}^{r_2}\sum_{k=1}^{n_2} C_{k,l}^2 D_{k,m}^2} \\
&= \sqrt{\sum_{k=1}^{n_1} \|\boldsymbol{A}_{k,\cdot}\|_2^2\|\boldsymbol{B}_{k,\cdot}\|_2^2}\sqrt{\sum_{k=1}^{n_2} \|\boldsymbol{C}_{k,\cdot}\|_2^2\|\boldsymbol{D}_{k,\cdot}\|_2^2}.
\end{aligned}
\tag{28}
$$

Combining (27) and (28) together, we have

$$
|\langle \mathcal{P}_{\Omega_0}(\boldsymbol{AC}^\top), \mathcal{P}_{\Omega_0}(\boldsymbol{BD}^\top)\rangle - t\langle \boldsymbol{AC}^\top, \boldsymbol{BD}^\top\rangle|
$$

$$
\leqslant \|\boldsymbol{\Omega}_0 - t\boldsymbol{J}\|\sqrt{\sum_{k=1}^{n_1} \|\boldsymbol{A}_{k,\cdot}\|_2^2\|\boldsymbol{B}_{k,\cdot}\|_2^2}\sqrt{\sum_{k=1}^{n_2} \|\boldsymbol{C}_{k,\cdot}\|_2^2\|\boldsymbol{D}_{k,\cdot}\|_2^2}.
$$

■

4.3.2. A PROOF OF LEMMA 12

Proof The proof of Lemma 12 can be divided into the controls of $K_2(\boldsymbol{X})$, $K_3(\boldsymbol{X})$ and $K_4(\boldsymbol{X})$ separately.

For $K_2(\boldsymbol{X})$, we have

Lemma 13 *In an event E_a with probability $\mathbb{P}[E_a] \geqslant 1 - n^{-3}$, uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n \times r}$ and corresponding $\boldsymbol{\Delta}$ defined as before, we have*

$$K_2(\boldsymbol{X}) \leqslant \|\boldsymbol{\Omega} - p\boldsymbol{J}\| \left[19 \sum_{i=1}^{n} \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 + 18\nu_r \|\boldsymbol{\Delta}\|_F^2 + 9\nu_r \sum_{i=s+1}^{r} \sigma_i \right] + 3 \times 10^{-4} p \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2,$$

where s is defined by

$$s := \max \left\{ s \leqslant r, \ \sigma_s \geqslant C_p \frac{\nu_r \log n}{p} \right\} \tag{29}$$

with C_p an absolute constant. Set $s = 0$ if $\sigma_1 < C_p \frac{\nu_r \log n}{p}$.

For $K_3(\boldsymbol{X})$, we use a modified version of Ge et al. (2017, Lemma 11):

Lemma 14 (Ge et al. 2017, Lemma 11) *If $\alpha \geqslant 100\sqrt{\nu_r}$, then uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n \times r}$ and corresponding $\boldsymbol{\Delta}$ defined as before, we have*

$$K_3(\boldsymbol{X}) \leqslant 199.54\lambda\alpha^2 \|\boldsymbol{\Delta}\|_F^2 - 0.3\lambda \sum_{i=1}^{n} \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4.$$

The main modification we have made is that we keep the extra negative term. We will give a proof in the appendix for completeness.

For $K_4(\boldsymbol{X})$, we have

Lemma 15 *Uniformly for all $\boldsymbol{X} \in \mathbb{R}^{n \times r}$ and corresponding $\boldsymbol{\Delta}$ defined as before, we have*

$$\begin{aligned} K_4(\boldsymbol{X}) \leqslant &5 \times 10^{-4} p \|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\|_F^2 + 2 \times 10^{-4} p \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + C_{10} \frac{r\|\mathcal{P}_\Omega(\boldsymbol{N}) - p\boldsymbol{N}\|^2}{p} \\ &+ 6p\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{N} \rangle. \end{aligned}$$

We can apply Lemma 4 together with Lemma 9 to bound $\|\mathcal{P}_\Omega(\boldsymbol{N}) - p\boldsymbol{N}\|$ and $\|\boldsymbol{\Omega} - p\boldsymbol{J}\|$ (similar result can also be found in Keshavan et al. (2010b)): As long as $p \geqslant C_S \frac{\log n}{n}$ with some absolute constant C_S, there is an absolute constant C_9, such that

$$\|\mathcal{P}_\Omega(\boldsymbol{N}) - p\boldsymbol{N}\| \leqslant \left(C_9 \sqrt{np(1-p)} + C_9 \sqrt{\log n} \right) \|\boldsymbol{N}\|_{\ell_\infty} \tag{30}$$

and

$$\|\mathbf{\Omega} - p\mathbf{J}\| \leqslant C_9\sqrt{np} \tag{31}$$

hold in an event E_b with probability $\mathbb{P}[E_b] \geqslant 1 - n^{-3}$.

By putting Lemma 13, Lemma 14 and Lemma 15 together, we have

$$\sum_{i=2}^{4} K_i(\mathbf{X}) \leqslant \|\mathbf{\Omega} - p\mathbf{J}\| \left[19 \sum_{i=1}^{n} \|\mathbf{\Delta}_{i,\cdot}\|_2^4 + 18\nu_r\|\mathbf{\Delta}\|_F^2 + 9\nu_r \sum_{i=s+1}^{r} \sigma_i \right] + 3 \times 10^{-4} p\|\mathbf{U}\mathbf{\Delta}^\top\|_F^2$$

$$+ 199.54\lambda\alpha^2\|\mathbf{\Delta}\|_F^2 - 0.3\lambda \sum_{i=1}^{n} \|\mathbf{\Delta}_{i,\cdot}\|_2^4 + 5 \times 10^{-4} p\|\mathbf{\Delta}\mathbf{\Delta}^\top\|_F^2$$

$$+ 2 \times 10^{-4} p\|\mathbf{U}\mathbf{\Delta}^\top\|_F^2 + C_{10}\frac{r\|\mathcal{P}_\Omega(\mathbf{N}) - p\mathbf{N}\|^2}{p} + 6p\langle\mathbf{\Delta}\mathbf{\Delta}^\top, \mathbf{N}\rangle.$$

Replacing α, λ by the assumption $100\sqrt{\nu_r} \leqslant \alpha \leqslant 200\sqrt{\nu_r}, 100\|\mathbf{\Omega} - p\mathbf{J}\| \leqslant \lambda \leqslant 200\|\mathbf{\Omega} - p\mathbf{J}\|$, we futher have

$$\sum_{i=2}^{4} K_i(\mathbf{X}) \leqslant \|\mathbf{\Omega} - p\mathbf{J}\| \left[19 \sum_{i=1}^{n} \|\mathbf{\Delta}_{i,\cdot}\|_2^4 + 18\nu_r\|\mathbf{\Delta}\|_F^2 + 9\nu_r \sum_{i=s+1}^{r} \sigma_i \right] + 3 \times 10^{-4} p\|\mathbf{U}\mathbf{\Delta}^\top\|_F^2$$

$$+ 7.9816 \times 10^8 \nu_r\|\mathbf{\Omega} - p\mathbf{J}\|\|\mathbf{\Delta}\|_F^2 - 30\|\mathbf{\Omega} - p\mathbf{J}\| \sum_{i=1}^{n} \|\mathbf{\Delta}_{i,\cdot}\|_2^4 + 5 \times 10^{-4} p\|\mathbf{\Delta}\mathbf{\Delta}^\top\|_F^2$$

$$+ 2 \times 10^{-4} p\|\mathbf{U}\mathbf{\Delta}^\top\|_F^2 + C_{10}\frac{r\|\mathcal{P}_\Omega(\mathbf{N}) - p\mathbf{N}\|^2}{p} + 6p\langle\mathbf{\Delta}\mathbf{\Delta}^\top, \mathbf{N}\rangle.$$

Combining with (30) and (31), and applying union bound,

$$\sum_{i=2}^{4} K_i(\mathbf{X}) \leqslant (19 - 30)\|\mathbf{\Omega} - p\mathbf{J}\| \sum_{i=1}^{n} \|\mathbf{\Delta}_{i,\cdot}\|_2^4 + (18 + 7.9816 \times 10^8)\nu_r\|\mathbf{\Omega} - p\mathbf{J}\|\|\mathbf{\Delta}\|_F^2$$

$$+ 9\nu_r\|\mathbf{\Omega} - p\mathbf{J}\| \sum_{i=s+1}^{r} \sigma_i + (3 + 2) \times 10^{-4} p\|\mathbf{U}\mathbf{\Delta}^\top\|_F^2$$

$$+ 5 \times 10^{-4} p\|\mathbf{\Delta}\mathbf{\Delta}^\top\|_F^2 + C_{10}\frac{r\|\mathcal{P}_\Omega(\mathbf{N}) - p\mathbf{N}\|^2}{p} + 6p\langle\mathbf{\Delta}\mathbf{\Delta}^\top, \mathbf{N}\rangle \tag{32}$$

$$\leqslant 5 \times 10^{-4} p\left[\|\mathbf{\Delta}^\top\mathbf{\Delta}\|_F^2 + \|\mathbf{U}\mathbf{\Delta}^\top\|_F^2\right] + C_3\frac{p(1-p)n + \log n}{p}r\|\mathbf{N}\|_{\ell_\infty}^2$$

$$+ C_{11}\sqrt{np}\nu_r\|\mathbf{\Delta}\|_F^2 + C_{12}\sqrt{np}\nu_r \sum_{i=s+1}^{r} \sigma_i + 6p\langle\mathbf{\Delta}\mathbf{\Delta}^\top, \mathbf{N}\rangle,$$

holds in an event E with probability $\mathbb{P}[E] \geqslant 1 - 2n^{-3}$.

For $\|\mathbf{\Delta}^\top\mathbf{\Delta}\|_F^2$, we have

$$\|\mathbf{\Delta}^\top\mathbf{\Delta}\|_F^2 = \langle\mathbf{\Delta}^\top\mathbf{\Delta}, \mathbf{\Delta}^\top\mathbf{\Delta}\rangle = \sum_{i=1}^{r} \sigma_i^4(\mathbf{\Delta}), \tag{33}$$

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 16 (Bhatia 2013, 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., Marshall et al. (2011, Theorem 9.B.1, Theorem 9.B.2)) together with Abel's summation formula.

From Lemma 16, we have

$$
\begin{aligned}
\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 = \mathrm{trace}(\boldsymbol{\Delta}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Delta}^\top) &= \langle \boldsymbol{U}^\top\boldsymbol{U}, \boldsymbol{\Delta}^\top\boldsymbol{\Delta} \rangle \\
&\geqslant \sum_{i=1}^{r} \lambda_{r+1-i}(\boldsymbol{U}^\top\boldsymbol{U}) \lambda_i(\boldsymbol{\Delta}^\top\boldsymbol{\Delta}) = \sum_{i=1}^{r} \sigma_i^2(\boldsymbol{\Delta}) \sigma_{r+1-i}^2(\boldsymbol{U}),
\end{aligned}
\tag{34}
$$

and

$$\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{N} \rangle \leqslant \sum_{i=1}^{n} \lambda_i(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top) \lambda_i(\boldsymbol{N}) = \sum_{i=1}^{r} \sigma_i^2(\boldsymbol{\Delta}) \sigma_i(\boldsymbol{N}). \tag{35}$$

Here we use the fact that $\lambda_i(\boldsymbol{U}^\top\boldsymbol{U}) = \sigma_i^2(\boldsymbol{U})$, $\lambda_i(\boldsymbol{\Delta}^\top\boldsymbol{\Delta}) = \sigma_i^2(\boldsymbol{\Delta})$, $\lambda_i(\boldsymbol{N}) = \sigma_i(\boldsymbol{N})$ 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 (33), (34) and (35) together we have

$$
\begin{aligned}
&- 5 \times 10^{-4} p \left[\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 \right] + C_{11}\sqrt{np}\nu_r\|\boldsymbol{\Delta}\|_F^2 + 6p\langle \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{N} \rangle \\
&\leqslant - 5 \times 10^{-4} p \left[\sum_{i=1}^{r} \sigma_i^4(\boldsymbol{\Delta}) + \sum_{i=1}^{r} \sigma_i^2(\boldsymbol{\Delta})\sigma_{r+1-i}^2(\boldsymbol{U}) \right] + C_{11}\sqrt{np}\nu_r \sum_{i=1}^{r} \sigma_i^2(\boldsymbol{\Delta}) \\
&\quad + 6p \sum_{i=1}^{r} \sigma_i^2(\boldsymbol{\Delta})\sigma_i(\boldsymbol{N}) \\
&\leqslant 5 \times 10^{-4} p \sum_{i=1}^{r} \left\{ -\sigma_i^4(\boldsymbol{\Delta}) + \left[C_{13}\sqrt{\frac{n}{p}}\nu_r - \sigma_{r+1-i}^2(\boldsymbol{U}) + C_{13}\sigma_i(\boldsymbol{N}) \right] \sigma_i^2(\boldsymbol{\Delta}) \right\}.
\end{aligned}
$$

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 over $\hat{x} = \frac{1}{2}[b]_+$, and the maximum value is $\frac{1}{4}\{[b]_+\}^2$. Therefore, we have

$$
\begin{aligned}
&-5 \times 10^{-4} p \left[\|\mathbf{\Delta}^\top \mathbf{\Delta}\|_F^2 + \|\mathbf{U}\mathbf{\Delta}^\top\|_F^2 \right] + C_{11}\sqrt{np}\nu_r \|\mathbf{\Delta}\|_F^2 + 6p\langle \mathbf{\Delta}\mathbf{\Delta}^\top, \mathbf{N}\rangle \\
&\leqslant C_{14}p \sum_{i=1}^{r} \left\{ \left[C_{13}\sqrt{\frac{n}{p}}\nu_r - \sigma_{r+1-i}^2(\mathbf{U}) + C_{13}\sigma_i(\mathbf{N}) \right]_+ \right\}^2 \\
&= C_{14}p \sum_{j=1}^{r} \left\{ \left[C_{13}\sqrt{\frac{n}{p}}\nu_r - \sigma_j^2(\mathbf{U}) + C_{13}\sigma_{r+1-j}(\mathbf{N}) \right]_+ \right\}^2 \\
&= C_{14}p \sum_{j=1}^{r} \left\{ \left[C_{13}\sqrt{\frac{n}{p}}\nu_r + C_{13}\sigma_{2r+1-j} - \sigma_j \right]_+ \right\}^2 .
\end{aligned}
\tag{36}
$$

In the second last line, we let $j = r + 1 - i$. In the last line, we use the fact that

$$
\sigma_{r+1-j}(\mathbf{N}) = \sigma_{r+r+1-j}(\mathbf{M}) = \sigma_{2r+1-j}
$$

and

$$
\sigma_j^2(\mathbf{U}) = \sigma_j(\mathbf{U}\mathbf{U}^\top) = \sigma_j(\mathbf{M}_r) = \sigma_j(\mathbf{M}) = \sigma_j.
$$

Finally putting (32) and (36) together we have

$$
\begin{aligned}
\sum_{i=2}^{4} K_i(\mathbf{X}) &\leqslant 10 \times 10^{-4} p \left[\|\mathbf{\Delta}^\top \mathbf{\Delta}\|_F^2 + \|\mathbf{U}\mathbf{\Delta}^\top\|_F^2 \right] + C_3 \frac{p(1-p)n + \log n}{p} r \|\mathbf{N}\|_{\ell_\infty}^2 \\
&\quad - 5 \times 10^{-4} p \left[\|\mathbf{\Delta}^\top \mathbf{\Delta}\|_F^2 + \|\mathbf{U}\mathbf{\Delta}^\top\|_F^2 \right] + C_{11}\sqrt{np}\nu_r \|\mathbf{\Delta}\|_F^2 \\
&\quad + C_{12}\sqrt{np}\nu_r \sum_{i=s+1}^{r} \sigma_i + 6p\langle \mathbf{\Delta}\mathbf{\Delta}^\top, \mathbf{N}\rangle \\
&\leqslant 10^{-3} p \left[\|\mathbf{\Delta}^\top \mathbf{\Delta}\|_F^2 + \|\mathbf{U}\mathbf{\Delta}^\top\|_F^2 \right] + C_3 \frac{p(1-p)n + \log n}{p} r \|\mathbf{N}\|_{\ell_\infty}^2 \\
&\quad + C_{14}p \sum_{i=1}^{r} \left\{ \left[C_{13}\sqrt{\frac{n}{p}}\nu_r + C_{13}\sigma_{2r+1-i} - \sigma_i \right]_+ \right\}^2 + C_{12}\sqrt{np}\nu_r \sum_{i=s+1}^{r} \sigma_i.
\end{aligned}
\tag{37}
$$

Recall by the definition of s in (29), for any $i > s$, we have $\sigma_i < C_p \frac{\nu_r \log n}{p}$. By choosing C_{13} sufficient large, i.e., $C_{13} \geqslant 2C_p$, we have

$$
\begin{aligned}
C_{13}\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\nu_r + C_{13}\sigma_{2r+1-i} - \sigma_i &\geqslant C_{13}\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\nu_r - 2\sigma_i + \sigma_i \\
&\geqslant C_{13}\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\nu_r - 2C_p\frac{\nu_r \log n}{p} + \sigma_i \\
&= C_{13}\sqrt{\frac{n}{p}}\nu_r + \sigma_i + (C_{13} - 2C_p)\frac{\nu_r \log n}{p} \\
&\geqslant C_{13}\sqrt{\frac{n}{p}}\nu_r + \sigma_i \geqslant 0.
\end{aligned}
$$

27

Therefore, for all $i > s$,

$$\left\{\left[C_{13}\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\nu_r + C_{13}\sigma_{2r+1-i} - \sigma_i\right]_+\right\}^2 \geqslant \left[C_{13}\sqrt{\frac{n}{p}}\nu_r + \sigma_i\right]^2 \geqslant \sqrt{\frac{n}{p}}\nu_r\sigma_i.$$

Combining with (37), we have

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

$$+ C_{14}p\sum_{i=1}^r\left\{\left[C_{13}\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\nu_r + C_{13}\sigma_{2r+1-i} - \sigma_i\right]_+\right\}^2$$

$$+ C_{12}p\sum_{i=s+1}^r\left\{\left[C_{13}\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\nu_r + C_{13}\sigma_{2r+1-i} - \sigma_i\right]_+\right\}^2$$

$$+ C_3\frac{p(1-p)n + \log n}{p}r\|\boldsymbol{N}\|_{\ell_\infty}^2$$

$$\leqslant 10^{-3}p\left[\|\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2\right]$$

$$+ C_3p\sum_{i=1}^r\left\{\left[C_4\left(\sqrt{\frac{n}{p}} + \frac{\log n}{p}\right)\nu_r + C_4\sigma_{2r+1-i} - \sigma_i\right]_+\right\}^2$$

$$+ C_3\frac{p(1-p)n + \log n}{p}r\|\boldsymbol{N}\|_{\ell_\infty}^2$$

which finishes the proof. $\blacksquare$

4.3.3. A PROOF OF LEMMA 13

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

$$\begin{aligned}
&D_{\Omega,p}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top)\\
=&D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)\\
=&\underbrace{D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top)}_{\text{①}} + \underbrace{D_{\Omega,p}(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)}_{\text{②}} + \underbrace{4D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)}_{\text{③}}.
\end{aligned} \tag{38}$$

Here we use the fact that Ω is symmetric. Our strategy here is using Lemma 6 to give a tight bound to as many as possible terms, for those terms that Lemma 6 cannot handle, we use Lemma 8 to give a bound. To be more precise, for ② and ③, as Lemma 6 cannot apply here, we use Lemma 8 to give a bound. For ①, we need to split it into two parts, the good part we can use Lemma 6 to control, and the rest part we use Lemma 8 to give a bound.

First for ② and ③, by applying Lemma 8,

$$|②| = |D_{\Omega,p}(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)| \leqslant \|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 \tag{39}$$

and

$$|③| = 4|D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)| \leqslant 4\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sqrt{\sum_{i=1}^n \|\boldsymbol{U}_{i,\cdot}\|_2^2 \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2} \sqrt{\sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4}$$

$$\leqslant 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\|\nu_r\|\boldsymbol{\Delta}\|_F^2 + 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4, \tag{40}$$

where for the second inequality we use the fact that $2xy \leqslant x^2 + y^2$.

Finally for ①, if $\boldsymbol{U}$ is good enough such that the incoherence $\mu(\boldsymbol{U})$ is well-bounded, then we can apply Lemma 6 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 (8), for the incoherence of the first k columns, we have

$$\mu\left(\mathrm{colspan}([\sqrt{\sigma_1}\boldsymbol{u}_1 \ \dots \ \sqrt{\sigma_k}\boldsymbol{u}_k])\right)$$

$$= \frac{n}{k} \max_i \sum_{j=1}^k u_{i,j}^2 \leqslant \frac{n}{k\sigma_k} \max_i \sum_{j=1}^k \sigma_j u_{i,j}^2 \leqslant \frac{n}{k\sigma_k} \max_i \sum_{j=1}^r \sigma_j u_{i,j}^2 \leqslant \frac{n\nu_r}{k\sigma_k}, \tag{41}$$

where $\mu(\cdot)$ is defined in (9).

For fixed s defined as in (29), 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 (38) as

$$\begin{aligned}
① &= D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top) \\
&= D_{\Omega,p}\left([\boldsymbol{U}^1 \ \boldsymbol{U}^2][\boldsymbol{\Delta}^1 \ \boldsymbol{\Delta}^2]^\top + [\boldsymbol{\Delta}^1 \ \boldsymbol{\Delta}^2][\boldsymbol{U}^1 \ \boldsymbol{U}^2]^\top, [\boldsymbol{U}^1 \ \boldsymbol{U}^2][\boldsymbol{\Delta}^1 \ \boldsymbol{\Delta}^2]^\top \right.\\
&\qquad \left. + [\boldsymbol{\Delta}^1 \ \boldsymbol{\Delta}^2][\boldsymbol{U}^1 \ \boldsymbol{U}^2]^\top\right) \\
&= \underbrace{D_{\Omega,p}\left(\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top + \boldsymbol{\Delta}^1(\boldsymbol{U}^1)^\top, \boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top + \boldsymbol{\Delta}^1(\boldsymbol{U}^1)^\top\right)}_{A_1} \\
&\quad + \underbrace{4D_{\Omega,p}\left(\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top, \boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top\right)}_{A_2} + \underbrace{2D_{\Omega,p}\left(\boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top, \boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top\right)}_{A_3} \\
&\quad + \underbrace{2D_{\Omega,p}\left(\boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top, \boldsymbol{\Delta}^2(\boldsymbol{U}^2)^\top\right)}_{A_4} + \underbrace{4D_{\Omega,p}\left(\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top, \boldsymbol{\Delta}^2(\boldsymbol{U}^2)^\top\right)}_{A_5}.
\end{aligned} \tag{42}$$

Now we can apply tight approximation Lemma 6 to the first term of (42). By the way we choose s, combining with (41),

$$p \geqslant C_p \frac{\nu_r \log n}{\sigma_s} \geqslant C_p \frac{\nu_r \log n}{\sigma_s} \cdot \frac{\mu\left(\mathrm{colspan}(\boldsymbol{U}^1)\right) s \sigma_s}{n \nu_r} = C_p \frac{\mu\left(\mathrm{colspan}(\boldsymbol{U}^1)\right) s \log n}{n}.$$

By choosing C_p sufficient large, Lemma 6 ensures that

$$
\begin{aligned}
|A_1| &= \left| D_{\Omega,p}\left(\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top + \boldsymbol{\Delta}^1(\boldsymbol{U}^1)^\top, \boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top + \boldsymbol{\Delta}^1(\boldsymbol{U}^1)^\top \right) \right| \\
&\leqslant 2.5 \times 10^{-5} p \| \boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top + \boldsymbol{\Delta}^1(\boldsymbol{U}^1)^\top \|_F^2 \\
&\leqslant 5 \times 10^{-5} p (\| \boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top \|_F^2 + \| \boldsymbol{\Delta}^1(\boldsymbol{U}^1)^\top \|_F^2) \\
&\leqslant 10^{-4} p \| \boldsymbol{U}\boldsymbol{\Delta}^\top \|_F^2
\end{aligned}
\tag{43}
$$

hold in an event E_a with probability $\mathbb{P}[E_a] \geqslant 1 - n^{-3}$, 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 (42), by applying Lemma 8 we have

$$
\begin{aligned}
|A_2| &= 4|D_{\Omega,p}(\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top, \boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top)| \\
&\leqslant 4\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sqrt{\sum_{i=1}^n \|\boldsymbol{U}_{i,\cdot}^1\|_2^2 \|\boldsymbol{U}_{i,\cdot}^2\|_2^2} \sqrt{\sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}^1\|_2^2 \|\boldsymbol{\Delta}_{i,\cdot}^2\|_2^2} \\
&\leqslant 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \left[\nu_r \|\boldsymbol{U}^2\|_F^2 + \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 \right]
\end{aligned}
\tag{44}
$$

for the second term in (42), where the second inequality use the fact that $\|\boldsymbol{U}_{i,\cdot}^1\|_2^2 \leqslant \|\boldsymbol{U}_{i,\cdot}\|_2^2 \leqslant \nu_r, \|\boldsymbol{\Delta}_{i,\cdot}^1\|_2^2 \leqslant \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2, \|\boldsymbol{\Delta}_{i,\cdot}^2\|_2^2 \leqslant \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2$ and $2xy \leqslant x^2 + y^2$. For the third term, applying Lemma 8 again we have

$$
\begin{aligned}
|A_3| &= 2|D_{\Omega,p}(\boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top, \boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top)| \\
&\leqslant 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sqrt{\sum_{i=1}^n \|\boldsymbol{U}_{i,\cdot}^2\|_2^4} \sqrt{\sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}^2\|_2^4} \\
&\leqslant \|\boldsymbol{\Omega} - p\boldsymbol{J}\| \left[\nu_r \|\boldsymbol{U}^2\|_F^2 + \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 \right],
\end{aligned}
\tag{45}
$$

where for the second inequality we also use the properties used in bounding second term. For the fourth and last term in (42), applying Lemma 8 and properties listed above, we have

$$
\begin{aligned}
|A_4| &= 2|D_{\Omega,p}(\boldsymbol{U}^2(\boldsymbol{\Delta}^2)^\top, \boldsymbol{\Delta}^2(\boldsymbol{U}^2)^\top)| \\
&\leqslant 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sum_{i=1}^n \|\boldsymbol{U}_{i,\cdot}^2\|_2^2 \|\boldsymbol{\Delta}_{i,\cdot}^2\|_2^2 \leqslant 2\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \nu_r \|\boldsymbol{\Delta}\|_F^2
\end{aligned}
\tag{46}
$$

and

$$
\begin{aligned}
|A_5| =&4|D_{\Omega,p}(\boldsymbol{U}^1(\boldsymbol{\Delta}^1)^\top, \boldsymbol{\Delta}^2(\boldsymbol{U}^2)^\top)| \\
\leqslant&4\|\boldsymbol{\Omega}-p\boldsymbol{J}\|\sqrt{\sum_{i=1}^n \|\boldsymbol{U}_{i,\cdot}^1\|_2^2\|\boldsymbol{\Delta}_{i,\cdot}^2\|_2^2}\sqrt{\sum_{i=1}^n \|\boldsymbol{U}_{i,\cdot}^2\|_2^2\|\boldsymbol{\Delta}_{i,\cdot}^1\|_2^2} \\
\leqslant&2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|\nu_r\|\boldsymbol{\Delta}^1\|_F^2 + 2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|\nu_r\|\boldsymbol{\Delta}^2\|_F^2 \\
=&2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|\nu_r\|\boldsymbol{\Delta}\|_F^2.
\end{aligned}
\tag{47}
$$

Now putting estimations of terms in (42) listed above together, i.e., (43), (44), (45), (46) and (47), we have

$$
\begin{aligned}
|①| =&|D_{\Omega,p}(\boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top, \boldsymbol{U}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}^\top)| \\
\leqslant&|A_1| + |A_2| + |A_3| + |A_4| + |A_5| \\
\leqslant&10^{-4}p\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 + 2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|\left[\nu_r\|\boldsymbol{U}^2\|_F^2 + \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4\right] \\
&+ \|\boldsymbol{\Omega}-p\boldsymbol{J}\|\left[\nu_r\|\boldsymbol{U}^2\|_F^2 + \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4\right] + 2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|\nu_r\|\boldsymbol{\Delta}\|_F^2 \\
&+ 2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|\nu_r\|\boldsymbol{\Delta}\|_F^2 \\
\leqslant&\|\boldsymbol{\Omega}-p\boldsymbol{J}\|\left[3\nu_r\|\boldsymbol{U}^2\|_F^2 + 3\sum_{i=1}^n \|\boldsymbol{\Delta}_i\|_2^4 + 4\nu_r\|\boldsymbol{\Delta}\|_F^2\right] + 10^{-4}p\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2.
\end{aligned}
\tag{48}
$$

Plugging estimations (39), (40) and (48) back to (38), we have

$$
\begin{aligned}
&|D_{\Omega,p}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top)| \\
\leqslant&|①| + |②| + |③| \\
\leqslant&\|\boldsymbol{\Omega}-p\boldsymbol{J}\|\left[3\nu_r\|\boldsymbol{U}^2\|_F^2 + 3\sum_{i=1}^n \|\boldsymbol{\Delta}_i\|_2^4 + 4\nu_r\|\boldsymbol{\Delta}\|_F^2\right] + 10^{-4}p\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2 \\
&+ \|\boldsymbol{\Omega}-p\boldsymbol{J}\|\sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 + 2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|\nu_r\|\boldsymbol{\Delta}\|_F^2 + 2\|\boldsymbol{\Omega}-p\boldsymbol{J}\|\sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 \\
=&\|\boldsymbol{\Omega}-p\boldsymbol{J}\|\left[3\nu_r\|\boldsymbol{U}^2\|_F^2 + 6\sum_{i=1}^n \|\boldsymbol{\Delta}_i\|_2^4 + 6\nu_r\|\boldsymbol{\Delta}\|_F^2\right] + 10^{-4}p\|\boldsymbol{U}\boldsymbol{\Delta}^\top\|_F^2.
\end{aligned}
$$

31

Therefore, combining with (39), we have

$$
\begin{aligned}
K_2(\boldsymbol{X}) \leqslant & |D_{\Omega,p}(\boldsymbol{\Delta\Delta}^\top, \boldsymbol{\Delta\Delta}^\top)| + 3|D_{\Omega,p}(\boldsymbol{XX}^\top - \boldsymbol{UU}^\top, \boldsymbol{XX}^\top - \boldsymbol{UU}^\top)| \\
\leqslant & \|\boldsymbol{\Omega} - p\boldsymbol{J}\| \sum_{i=1}^{n} \|\boldsymbol{\Delta}_{i,.}\|_2^4 + 3\|\boldsymbol{\Omega} - p\boldsymbol{J}\| \left[3\nu_r \|\boldsymbol{U}^2\|_F^2 + 6\sum_{i=1}^{n} \|\boldsymbol{\Delta}_i\|_2^4 + 6\nu_r\|\boldsymbol{\Delta}\|_F^2 \right] \\
& + 3\times 10^{-4} p \|\boldsymbol{U\Delta}^\top\|_F^2 \\
\leqslant & \|\boldsymbol{\Omega} - p\boldsymbol{J}\| \left[19 \sum_{i=1}^{n} \|\boldsymbol{\Delta}_i\|_2^4 + 18\nu_r \|\boldsymbol{\Delta}\|_F^2 + 9\nu_r \sum_{i=s+1}^{r} \sigma_i \right] + 3\times 10^{-4} p \|\boldsymbol{U\Delta}^\top\|_F^2.
\end{aligned}
$$

The last line uses the fact that $\|\boldsymbol{U}^2\|_F^2 = \sum_{i=s+1}^{r} \sigma_i$. ∎

4.3.4. A PROOF OF LEMMA 15

Proof First, by matrix Hölder's inequality,

$$
6|\langle \boldsymbol{\Delta\Delta}^\top, \mathcal{P}_\Omega(\boldsymbol{N}) - p\boldsymbol{N}\rangle| \leqslant 6 \frac{\sqrt{p}\|\boldsymbol{\Delta\Delta}^\top\|_*}{\sqrt{r}} \frac{\sqrt{r}\|\mathcal{P}_\Omega(\boldsymbol{N}) - p\boldsymbol{N}\|}{\sqrt{p}}.
$$

Since $\boldsymbol{\Delta\Delta}^\top$ is at most rank-r, $\|\boldsymbol{\Delta\Delta}^\top\|_* \leqslant \sqrt{r}\|\boldsymbol{\Delta\Delta}^\top\|_F$. Therefore,

$$
\begin{aligned}
6|\langle \boldsymbol{\Delta\Delta}^\top, \mathcal{P}_\Omega(\boldsymbol{N}) - p\boldsymbol{N}\rangle| \leqslant & 6\sqrt{p}\|\boldsymbol{\Delta\Delta}^\top\|_F \frac{\sqrt{r}\|\mathcal{P}_\Omega(\boldsymbol{N}) - p\boldsymbol{N}\|}{\sqrt{p}} \\
\leqslant & 5\times 10^{-4} p \|\boldsymbol{\Delta\Delta}^\top\|_F^2 + C_{15} \frac{r\|\mathcal{P}_\Omega(\boldsymbol{N}) - p\boldsymbol{N}\|^2}{p}.
\end{aligned}
$$

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

$$
8|\langle \boldsymbol{U\Delta}^\top, \mathcal{P}_\Omega(\boldsymbol{N}) - p\boldsymbol{N}\rangle| \leqslant 2\times 10^{-4} p \|\boldsymbol{U\Delta}^\top\|_F^2 + C_{15} \frac{r\|\mathcal{P}_\Omega(\boldsymbol{N}) - p\boldsymbol{N}\|^2}{p},
$$

Therefore, by the way we define $K_4(\boldsymbol{X})$ in (14), we have

$$
\begin{aligned}
K_4(\boldsymbol{X}) \leqslant & |6\langle \boldsymbol{\Delta\Delta}^\top, \mathcal{P}_\Omega(\boldsymbol{N})\rangle - 6p\langle \boldsymbol{\Delta\Delta}^\top, \boldsymbol{N}\rangle| + |8\langle \boldsymbol{U\Delta}^\top, \mathcal{P}_\Omega(\boldsymbol{N})\rangle - 8p\langle \boldsymbol{U\Delta}^\top, \boldsymbol{N}\rangle| \\
& + 6p\langle \boldsymbol{\Delta\Delta}^\top, \boldsymbol{N}\rangle \\
\leqslant & 5\times 10^{-4} p \|\boldsymbol{\Delta\Delta}^\top\|_F^2 + 2\times 10^{-4} p \|\boldsymbol{U\Delta}^\top\|_F^2 + C_{10} \frac{r\|\mathcal{P}_\Omega(\boldsymbol{N}) - p\boldsymbol{N}\|^2}{p} \\
& + 6p\langle \boldsymbol{\Delta\Delta}^\top, \boldsymbol{N}\rangle.
\end{aligned}
$$

∎

5. Discussions

This paper studies low-rank approximation of a positive semidefinite matrix from partial entries via nonconvex optimization. We established a model-free theory for local-minimum based low-rank approximation without any assumptions on its rank, condition number or eigenspace incoherence parameter. We have also improved the state-of-the-art sampling rate results for nonconvex matrix completion with no spurious local minima in Ge et al. (2016, 2017), and have investigated the performance of the proposed nonconvex optimization in presence of large condition numbers, large incoherence parameters, or rank mismatching. The nonconvex optimization is further applied to the problem of memory-efficient kernel PCA. Compared to the well-known Nyström methods, numerical experiments illustrate that the proposed nonconvex optimization approach yields more stable results in both low-rank approximation and clustering.

For future research, we are interested in understanding whether and how fast first-order methods converge to a neighborhood of the set of local minima with theoretical guarantees. In fact, a series of recent works in nonconvex optimization have discussed why and when first-order iterative algorithms can avoid strict saddle points almost surely. For example, in a very recent work by Lee et al. (2017), the authors show that under mild conditions of the nonconvex objective function, a variety of first order algorithms can avoid strict saddle points with almost all initialization, which extends the previous results in Lee et al. (2016) and Panageas and Piliouras (2017). We are particularly interested in the robust version of the strict saddle points condition discussed in Ge et al. (2015) and Jin et al. (2017), referred to as (θ, γ, ζ)-strict saddle, under which noisy stochastic/deterministic gradient descent methods are proven to converge to a neighborhood of the local minima. In fact, Ge et al. (2017, Theorem 12) shows that the nonconvex optimization (1) satisfies certain (θ, γ, ζ)-strict saddle conditions as long as $\boldsymbol{M}$ is exactly of rank r, its condition number and eigenspace incoherence parameter are well-bounded, and the sampling rate is sufficiently large, but their argument cannot be straightforwardly extended to the model-free settings. We plan to explore the (θ, γ, ζ)-strict saddle conditions for (1) under a model-free framework in future.

Acknowledgments

References

Dimitris Achlioptas and Frank McSherry. Fast computation of low-rank matrix approximations. *Journal of the ACM (JACM)*, 54(2):9, 2007.

Dimitris Achlioptas, Frank McSherry, and Bernhard Schölkopf. Sampling techniques for kernel methods. In *Advances in Neural Information Processing Systems*, pages 335–342, 2002.

Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

Maria-Florina Balcan, Yingyu Liang, David P Woodruff, and Hongyang Zhang. Optimal sample complexity for matrix completion and related problems via ℓ_2-regularization. *arXiv preprint arXiv:1704.08683*, 2017.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

Afonso S Bandeira, Ramon Van Handel, et al. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, 2016.

Rajendra Bhatia. *Matrix Analysis*. Graduate Texts in Mathematics. Springer New York, 2013. ISBN 9781461206538.

Srinadh Bhojanapalli and Prateek Jain. Universal matrix completion. In *International Conference on Machine Learning*, pages 1881–1889, 2014.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873 – 3881, 2016.

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

T Tony Cai, Xiaodong Li, Zongming Ma, et al. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *The Annals of Statistics*, 44(5):2221–2251, 2016.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Petros Drineas and Michael W Mahoney. On the Nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6(Dec):2153–2175, 2005.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points–online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242, 2017.

Gene H Golub and Charles F Van Loan. *Matrix Computations*, volume 3. JHU Press, 2012.

Thore Graepel. Kernel matrix completion by semidefinite programming. In *International Conference on Artificial Neural Networks*, pages 694–699. Springer, 2002.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

David Gross and Vincent Nesme. Note on sampling without replacing from a finite collection of matrices. *arXiv preprint arXiv:1001.2738*, 2010.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732, 2017.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010a.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010b.

Kwang In Kim, Matthias O Franz, and Bernhard Schölkopf. Iterative kernel principal component analysis for image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1351–1366, 2005.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

Qiuwei Li, Zhihui Zhu, and Gongguo Tang. Geometry of factored nuclear norm regularization. *arXiv preprint arXiv:1704.01265*, 2017.

Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and computational harmonic analysis*, 2018.

Xingguo Li, Zhaoran Wang, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, and Tuo Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296*, 2016a.

Yuanzhi Li, Yingyu Liang, and Andrej Risteski. Recovery guarantee of weighted low-rank approximation via alternating minimization. In *International Conference on Machine Learning*, pages 2358–2367, 2016b.

Po-Ling Loh and Martin J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(19):559–616, 2015.

Albert W Marshall, Ingram Olkin, and Barry C Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, 2011.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May):1665–1697, 2012.

John Paisley and Lawrence Carin. A nonparametric bayesian model for kernel matrix completion. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2090–2093. IEEE, 2010.

Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *Innovations in Theoretical Computer Science*, 2017.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.

Yousef Saad. *Iterative Methods for Sparse Linear Systems*, volume 82. SIAM, 2003.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

Van Vu. A Simple SVD Algorithm for Finding Hidden Partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018.

Quan Wang. Kernel principal component analysis and its applications in face recognition and active shape models. *arXiv preprint arXiv:1207.3538*, 2012.

Christopher KI Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688, 2001.

Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. In *Advances in Neural Information Processing Systems*, pages 4152–4160, 2016.

Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015.

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. The global optimization geometry of nonsymmetric matrix factorization and sensing. *arXiv preprint arXiv:1703.01256*, 2017.

Appendix A. Proof of Corollary 3

Proof The inequality (8) gives $\|M\|_{\ell_\infty} \leqslant \frac{\mu_r r \sigma_1}{n}$. Therefore, in the case $\text{rank}(M) = r$, the approximation error bound (5) becomes

$$\left\|\widehat{X}\widehat{X}^\top - M\right\|_F^2 \leqslant C_1 \sum_{i=1}^r \left\{ \left[C_2 \left(\sqrt{\frac{n}{p}} + \frac{\log n}{p} \right) \frac{\mu_r r}{n} \sigma_1 - \sigma_i \right]_+ \right\}^2.$$

Therefore, if

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

with absolute constant C sufficient large, 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 (7), 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 (5) 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

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

with absolute constant C sufficient large, we have $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top = \boldsymbol{M}$. ■

Appendix B. Proof of Lemma 14

Here we present a proof of Lemma 14. This proof is exactly the proof in Ge et al. (2017) except keeping the extra negative term. We include the proof in Ge et al. (2017) here for completeness.

Proof By Ge et al. (2017, Lemma 18), we have

$$\begin{aligned}
&\mathrm{vec}(\boldsymbol{\Delta})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \mathrm{vec}(\boldsymbol{\Delta}) - 4\langle \nabla G_\alpha(\boldsymbol{X}), \boldsymbol{\Delta} \rangle \\
=&4 \sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^3 \frac{\|\boldsymbol{X}_{i,\cdot}\|_2^2 \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2 - \langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot} \rangle^2}{\|\boldsymbol{X}_{i,\cdot}\|_2^3} \\
&+ 12 \sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^2 \frac{\langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot} \rangle^2}{\|\boldsymbol{X}_{i,\cdot}\|_2^2} - 16 \sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^3 \frac{\langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot} \rangle}{\|\boldsymbol{X}_{i,\cdot}\|_2}.
\end{aligned} \tag{49}$$

First of all, since we choose $\alpha \geqslant 100\sqrt{\nu_r} = 100\|\boldsymbol{U}\|_{2,\infty}$, then for all $\boldsymbol{X}_{i,\cdot}$ satisfying $\|\boldsymbol{X}_{i,\cdot}\|_2 \geqslant \alpha$, we have

$$\langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{\Delta}_{i,\cdot} \rangle = \langle \boldsymbol{X}_{i,\cdot}, \boldsymbol{X}_{i,\cdot} - \boldsymbol{U}_{i,\cdot} \rangle \geqslant \|\boldsymbol{X}_{i,\cdot}\|_2^2 - \|\boldsymbol{X}_{i,\cdot}\|_2 \|\boldsymbol{U}_{i,\cdot}\|_2 \geqslant (1 - 0.01)\|\boldsymbol{X}_{i,\cdot}\|_2^2 \geqslant 0.99\|\boldsymbol{X}_{i,\cdot}\|_2^2, \tag{50}$$

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 $\|\boldsymbol{X}_{i,\cdot}\|_2$:

$$\|\boldsymbol{\Delta}_{i,\cdot}\|_2 \leqslant \|\boldsymbol{X}_{i,\cdot}\|_2 + \|\boldsymbol{U}_{i,\cdot}\|_2 \leqslant 1.01\|\boldsymbol{X}_{i,\cdot}\|_2. \tag{51}$$

Plugging the above two estimations (50), (51) 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 (49), we have

$$\begin{aligned}
&\mathrm{vec}(\boldsymbol{\Delta})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \mathrm{vec}(\boldsymbol{\Delta}) - 4\langle \nabla G_\alpha(\boldsymbol{X}), \boldsymbol{\Delta} \rangle \\
\leqslant & -15.68 \sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^3 \|\boldsymbol{X}_{i,\cdot}\|_2 + 12 \sum_{i=1}^n [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^2 \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2.
\end{aligned} \tag{52}$$

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 $\|\boldsymbol{X}_{i,\cdot}\|_2$:

$$\|\boldsymbol{\Delta}_{i,\cdot}\|_2 \leqslant \|\boldsymbol{X}_{i,\cdot}\|_2 + \|\boldsymbol{U}_{i,\cdot}\|_2 \leqslant 1.002\|\boldsymbol{X}_{i,\cdot}\|_2, \tag{53}$$

and also lower bound $\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha$ by $\|\boldsymbol{\Delta}_{i,\cdot}\|_2$:

$$\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha \geqslant \left(1 - \frac{1}{5}\right)\|\boldsymbol{X}_{i,\cdot}\|_2 \geqslant \frac{400}{501}\|\boldsymbol{\Delta}_{i,\cdot}\|_2. \tag{54}$$

Plugging (53) and (54) back to (52), we have

$$\begin{aligned}
&\operatorname{vec}(\boldsymbol{\Delta})^\top \nabla^2 G_\alpha(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{\Delta}) - 4\langle \nabla G_\alpha(\boldsymbol{X}), \boldsymbol{\Delta}\rangle \\
&\leqslant 12 \sum_{i, \|\boldsymbol{X}_{i,\cdot}\|_2 < 5\alpha} [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^2 \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2 \\
&\quad + \left[12 - 15.68 \times \frac{400}{501} \times \frac{1}{1.002}\right] \sum_{i, \|\boldsymbol{X}_{i,\cdot}\|_2 \geqslant 5\alpha} [(\|\boldsymbol{X}_{i,\cdot}\|_2 - \alpha)_+]^2 \|\boldsymbol{\Delta}_{i,\cdot}\|_2^2 \\
&\leqslant 192\alpha^2 \|\boldsymbol{\Delta}\|_F^2 - 0.3 \sum_{i, \|\boldsymbol{X}_{i,\cdot}\|_2 \geqslant 5\alpha} \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4 \\
&\leqslant 199.54\alpha^2 \|\boldsymbol{\Delta}\|_F^2 - 0.3 \sum_{i=1}^n \|\boldsymbol{\Delta}_{i,\cdot}\|_2^4,
\end{aligned}$$

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 $\alpha \geqslant 100\sqrt{\nu_r}$. $\blacksquare$