Inference and Uncertainty Quantification for Noisy Matrix Completion

Yuxin Chen* Jianqing Fan† Cong Ma† Yuling Yan†

June 2019; Revised: October 2019

Abstract

Noisy matrix completion aims at estimating a low-rank matrix given only partial and corrupted entries. Despite substantial progress in designing efficient estimation algorithms, it remains largely unclear how to assess the uncertainty of the obtained estimates and how to perform statistical inference on the unknown matrix (e.g. constructing a valid and short confidence interval for an unseen entry).

This paper takes a step towards inference and uncertainty quantification for noisy matrix completion. We develop a simple procedure to compensate for the bias of the widely used convex and nonconvex estimators. The resulting de-biased estimators admit nearly precise non-asymptotic distributional characterizations, which in turn enable optimal construction of confidence intervals / regions for, say, the missing entries and the low-rank factors. Our inferential procedures do not rely on sample splitting, thus avoiding unnecessary loss of data efficiency. As a byproduct, we obtain a sharp characterization of the estimation accuracy of our de-biased estimators, which, to the best of our knowledge, are the first tractable algorithms that provably achieve full statistical efficiency (including the preconstant). The analysis herein is built upon the intimate link between convex and nonconvex optimization — an appealing feature recently discovered by [CCF+19].

Keywords: matrix completion, statistical inference, confidence intervals, uncertainty quantification, convex relaxation, nonconvex optimization

Contents

Author names are sorted alphabetically.

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

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

1 Introduction

1.1 Motivation: inference and uncertainty quantification?

Low-rank matrix completion is concerned with recovering a low-rank matrix, when only a small fraction of its entries are revealed to us [Sre04, CR09, KMO10a]. Tackling this problem in large-scale applications is computationally challenging, due to the intrinsic nonconvexity incurred by the low-rank structure. To further complicate matters, another inevitable challenge stems from the imperfectness of data acquisition mechanisms, wherein the acquired samples are contaminated by a certain amount of noise.

Fortunately, if the entries of the unknown matrix are sufficiently de-localized and randomly revealed, this problem may not be as hard as it seems. Substantial progress has been made over the past several years in designing computationally tractable algorithms — including both convex and nonconvex approaches — that allow to fill in unseen entries faithfully given only partial noisy samples [CP10, NW12, KLT11, KMO10b, CW15, MWCC17, CCF$^+$19]. Nevertheless, modern decision making would often require one step further. It not merely anticipates a faithful estimate, but also seeks to quantify the uncertainty or "confidence" of the provided estimate, ideally in a reasonably accurate fashion. For instance, given an estimate returned by the convex approach, how to use it to compute a short interval that is likely to contain a missing entry?

Conducting effective uncertainty quantification for noisy matrix completion is, however, far from straightforward. For the most part, the state-of-the-art matrix completion algorithms require solving highly complex optimization problems, which often do not admit closed-form solutions. Of necessity, it is generally very challenging to pin down the distributions of the estimates returned by these algorithms. The lack of distributional characterizations presents a major roadblock to performing valid, yet efficient, statistical inference on the unknown matrix of interest.

It is worth noting that a number of recent papers have been dedicated to inference and uncertainty quantification for various high-dimensional problems in high-dimensional statistics, including Lasso [ZZ14, vdGBRD14, JM14a], generalized linear models [vdGBRD14, NL17, BFL$^+$18], graphical models [JVDG15, RSZZ15, MLL17]), amongst others. Very little work, however, has looked into noisy matrix completion along this direction. While non-asymptotic statistical guarantees for noisy matrix completion have been derived in prior theory, most, if not all, of the estimation error bounds are supplied only at an order-wise level. Such order-wise error bounds either lose a significant factor relative to the optimal guarantees, or come with an unspecified (but often enormous) pre-constant. Viewed in this light, a confidence region constructed directly based on such results is bound to be overly conservative, resulting in substantial over-coverage.

1.2 A glimpse of our contributions

This paper takes a substantial step towards efficient inference and uncertainty quantification for noisy matrix completion. Specifically, we develop a simple procedure to compensate for the bias of the commonly used convex and nonconvex estimators. The resulting de-biased estimators admit nearly accurate non-asymptotic distributional guarantees. Such distributional characterizations in turn allow us to reason about the uncertainty of the obtained estimates vis-à-vis the unknown matrix. While details of our main findings are postponed to Section 3, we would like to immediately single out a few important merits of the proposed inferential procedures and theory:

1. Our results enable two types of uncertainty assessment, namely, we can construct (i) confidence intervals for each entry — either observed or missing — of the unknown matrix; (ii) confidence regions for the low-rank factors of interest (modulo some unavoidable global ambiguity).

2. Despite the complicated statistical dependency, our procedure and theory do not rely on sample splitting, thus avoiding the unnecessary widening of confidence intervals / regions due to insufficient data usage.

3. The confidence intervals / regions constructed based on the proposed procedures are, in some sense, optimal.

4. We present a unified approach that accommodates both convex and nonconvex estimators seamlessly.

5. As a byproduct, we characterize the Euclidean estimation errors of the proposed de-biased estimators. Such error bounds are sharp and match an oracle lower bound precisely (including the pre-constant). To the best of our knowledge, this is the first theory that demonstrates that a computationally feasible algorithm can achieve the statistical limit including the pre-constant.

All of this is built upon the intimate link between convex and nonconvex estimators [CCF$^+$19], as well as the recent advances in analyzing the stability of nonconvex optimization against random noise [MWCC17].

2 Models and notation

To cast the noisy matrix completion problem in concrete statistical settings, we adopt a model commonly studied in the literature [CR09]. We also introduce some useful notation.

Ground truth. Denote by $\boldsymbol{M}^{\star} \in \mathbb{R}^{n \times n}$ the unknown rank-r matrix of interest,[1] whose (compact) singular value decomposition (SVD) is given by $\boldsymbol{M}^{\star} = \boldsymbol{U}^{\star}\boldsymbol{\Sigma}^{\star}\boldsymbol{V}^{\star\top}$. We set

$$\sigma_{\max} \triangleq \sigma_1(\boldsymbol{M}^{\star}), \quad \sigma_{\min} \triangleq \sigma_r(\boldsymbol{M}^{\star}), \quad \text{and} \quad \kappa \triangleq \sigma_{\max}/\sigma_{\min}, \tag{2.1}$$

where $\sigma_i(\boldsymbol{A})$ denotes the ith largest singular value of a matrix $\boldsymbol{A}$. Further, we let $\boldsymbol{X}^{\star} \triangleq \boldsymbol{U}^{\star}\boldsymbol{\Sigma}^{\star 1/2} \in \mathbb{R}^{n \times r}$ and $\boldsymbol{Y}^{\star} \triangleq \boldsymbol{V}^{\star}\boldsymbol{\Sigma}^{\star 1/2} \in \mathbb{R}^{n \times r}$ stand for the *balanced* low-rank factors of $\boldsymbol{M}^{\star}$, which obey

$$\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star} = \boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star} = \boldsymbol{\Sigma}^{\star} \qquad \text{and} \qquad \boldsymbol{M}^{\star} = \boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top}. \tag{2.2}$$

Observation models. What we observe is a random subset of noisy entries of $\boldsymbol{M}^{\star}$; more specifically, we observe

$$M_{ij} = M_{ij}^{\star} + E_{ij}, \qquad E_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \qquad \text{for all } (i,j) \in \Omega, \tag{2.3}$$

where $\Omega \subseteq \{1, \cdots, n\} \times \{1, \cdots, n\}$ is a subset of indices, and E_{ij} denotes independently generated noise at the location (i,j). From now on, we assume the *random sampling* model where each index (i,j) is included in Ω independently with probability p (i.e. data are missing uniformly at random). We shall use $\mathcal{P}_{\Omega}(\cdot) : \mathbb{R}^{n \times n} \mapsto \mathbb{R}^{n \times n}$ to represent the orthogonal projection onto the subspace of matrices that vanish outside the index set Ω.

Incoherence conditions. Clearly, not all matrices can be reliably estimated from a highly incomplete set of measurements. To address this issue, we impose a standard incoherence condition [CR09, Che15] on the singular subspaces of $\boldsymbol{M}^{\star}$ (i.e. $\boldsymbol{U}^{\star}$ and $\boldsymbol{V}^{\star}$):

$$\max\{\|\boldsymbol{U}^{\star}\|_{2,\infty}, \|\boldsymbol{V}^{\star}\|_{2,\infty}\} \leq \sqrt{\mu r/n}, \tag{2.4}$$

where μ is termed the incoherence parameter and $\|\boldsymbol{A}\|_{2,\infty}$ denotes the largest ℓ_2 norm of all rows in $\boldsymbol{A}$. A small μ implies that the energy of $\boldsymbol{U}^{\star}$ and $\boldsymbol{V}^{\star}$ are reasonably spread out across all of their rows.

Asymptotic notation. Here, $f(n) \lesssim h(n)$ (or $f(n) = O(h(n))$) means $|f(n)| \leq c_1|h(n)|$ for some constant $c_1 > 0$, $f(n) \gtrsim h(n)$ means $|f(n)| \geq c_2|h(n)|$ for some constant $c_2 > 0$, $f(n) \asymp h(n)$ means $c_2|h(n)| \leq |f(n)| \leq c_1|h(n)|$ for some constants $c_1, c_2 > 0$, and $f(n) = o(h(n))$ means $\lim_{n\to\infty} f(n)/h(n) = 0$. We write $f(n) \ll h(n)$ to indicate that $|f(n)| \leq c_1|h(n)|$ for some small constant $c_1 > 0$ (much smaller than 1), and use $f(n) \gg h(n)$ to indicate that $|f(n)| \geq c_2|h(n)|$ for some large constant $c_2 > 0$ (much larger than 1).

3 Inferential procedures and main results

The proposed inferential procedure lays its basis on two of the most popular estimation paradigms — convex relaxation and nonconvex optimization — designed for noisy matrix completion. Recognizing the complicated bias of these two highly nonlinear estimators, we shall first illustrate how to perform bias correction, followed by a distributional theory that establishes the near-Gaussianity and optimality of the proposed de-biased estimators.

3.1 Background: convex and nonconvex estimators

We first review in passing two tractable estimation algorithms that are arguably the most widely used in practice. They serve as the starting point for us to design inferential procedures for noisy low-rank matrix completion. The readers familiar with this literature can proceed directly to Section 3.2.

[1] We restrict our attention to squared matrices for simplicity of presentation. Most findings extend immediately to the more general rectangular case $\boldsymbol{M}^{\star} \in \mathbb{R}^{n_1 \times n_2}$ with different n_1 and n_2.

Algorithm 1 Gradient descent for solving the nonconvex problem (3.4).

```
Suitable initialization: $\boldsymbol{X}^0$, $\boldsymbol{Y}^0$
Gradient updates: for $t = 0, 1, \ldots, t_0 - 1$ do
    $\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \frac{\eta}{p}\big[\mathcal{P}_\Omega(\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M})\boldsymbol{Y}^t + \lambda\boldsymbol{X}^t\big],$   (3.3a)
    $\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \frac{\eta}{p}\big[[\mathcal{P}_\Omega(\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M})]^\top\boldsymbol{X}^t + \lambda\boldsymbol{Y}^t\big],$   (3.3b)
where $\eta > 0$ determines the step size or the learning rate.
```

Convex relaxation. Recall that the rank function $\mathsf{rank}(\cdot)$ is highly nonconvex, which often prevents us from computing a rank-constrained estimator in polynomial time. For the sake of computational feasibility, prior works suggest relaxing the rank function into its convex surrogate [Faz02, RFP10]; for example, one can consider the following penalized least-squares convex program

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\mathsf{minimize}} \qquad \frac{1}{2}\sum_{(i,j)\in\Omega}\left(Z_{ij}-M_{ij}\right)^2 + \lambda\|\boldsymbol{Z}\|_*, \tag{3.1}$$

or using our notation $\mathcal{P}_\Omega$,

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\mathsf{minimize}} \qquad \frac{1}{2}\big\|\mathcal{P}_\Omega\big(\boldsymbol{Z}-\boldsymbol{M}\big)\big\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{Z}\|_*. \tag{3.2}$$

Here, $\|\cdot\|_*$ is the nuclear norm (the sum of singular values, which is a convex surrogate of the rank function), and $\lambda > 0$ is some regularization parameter. Under mild conditions, the solution to the convex program (3.1) provably attains near-optimal estimation accuracy (in an order-wise sense), provided that a proper regularization parameter λ is adopted [CCF+19].

Nonconvex optimization. It is recognized that the convex approach, which typically relies on solving a semidefinite program, is still computationally expensive and not scalable to large dimensions. This motivates an alternative route, which represents the matrix variable via two low-rank factors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}$ and attempts solving the following nonconvex program directly

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\mathsf{minimize}} \qquad \frac{1}{2}\big\|\mathcal{P}_\Omega\big(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}\big)\big\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{Y}\|_{\mathrm{F}}^2. \tag{3.4}$$

Here, we choose a regularizer of the form $0.5\lambda(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2)$ primarily to mimic the nuclear norm $\lambda\|\boldsymbol{Z}\|_*$ (see [SS05, MHT10]). A variety of optimization algorithms have been proposed to tackle the nonconvex program (3.4) or its variants [SL16, CW15, MWCC17]; the readers are referred to [CLC19] for a recent overview. As a prominent example, a two-stage algorithm — gradient descent following suitable initialization — provably enjoys fast convergence and order-wise optimal statistical guarantees for a wide range of scenarios [MWCC17, CCF+19, CLL19]. The current paper focuses on this simple yet powerful algorithm, as documented in Algorithm 1 and detailed in Appendix A.1.

Intimate connections between convex and nonconvex estimates. Denote by $\boldsymbol{Z}^{\mathsf{cvx}}$ any minimizer of the convex program (3.1), and denote by $(\boldsymbol{X}^{\mathsf{ncvx}}, \boldsymbol{Y}^{\mathsf{ncvx}})$ the estimate returned by Algorithm 1 aimed at solving (3.4). As was recently shown in [CCF+19], when the regularization parameter λ is properly chosen, these two estimates obey (see (A.12) in Appendix A.2 for a precise statement)

$$\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} \approx \boldsymbol{Z}^{\mathsf{cvx}} \approx \boldsymbol{Z}^{\mathsf{cvx},r}. \tag{3.5}$$

Here, $\boldsymbol{Z}^{\mathsf{cvx},r} \triangleq \mathcal{P}_{\text{rank-}r}(\boldsymbol{Z}^{\mathsf{cvx}})$ is the best rank-r approximation of the convex estimate $\boldsymbol{Z}^{\mathsf{cvx}}$, where $\mathcal{P}_{\text{rank-}r}(\boldsymbol{B}) \triangleq \arg\min_{\boldsymbol{A}:\mathrm{rank}(\boldsymbol{A})\leq r}\|\boldsymbol{A}-\boldsymbol{B}\|_{\mathrm{F}}$. In truth, the three matrices of interest in (3.5) are exceedingly close to, if not identical with, each other. This salient feature paves the way for a unified treatment of convex and nonconvex approaches: most inferential procedures and guarantees developed for the nonconvex estimate can be readily transferred to perform inference for the convex one, and vice versa.

3.2 Constructing de-biased estimators

We are now well equipped to describe how to construct new estimators based on the convex estimate $\boldsymbol{Z}^{\mathsf{cvx}}$ and the nonconvex estimate $(\boldsymbol{X}^{\mathsf{ncvx}}, \boldsymbol{Y}^{\mathsf{ncvx}})$, so as to enable statistical inference. Motivated by the proximity of the convex and nonconvex estimates and for the sake of conciseness, we shall abuse notation by using the shorthand $\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{Y}$ for both convex and nonconvex estimates; see Table 1 and Appendix B for precise definitions. This allows us to unify the presentation for both convex and nonconvex estimators.

Given that both (3.1) and (3.4) are regularized least-squares problems, they behave effectively like shrinkage estimators, indicating that the provided estimates necessarily suffer from non-negligible bias. In order to enable desired statistical inference, it is natural to first correct the estimation bias.

A de-biased estimator for the matrix. A natural de-biasing strategy that immediately comes to mind is the following simple linear transformation (recall the notation in Table 1):

$$\boldsymbol{Z}^0 \triangleq \boldsymbol{Z} - \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{Z} - \boldsymbol{M}) = \underbrace{\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}^\star)}_{\text{mean: } \boldsymbol{M}^\star} + \underbrace{\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{E})}_{\text{mean: } \boldsymbol{0}} + \underbrace{\boldsymbol{Z} - \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{Z})}_{\text{mean: } \boldsymbol{0} \text{ (heuristically)}}, \tag{3.6}$$

where we identify $\mathcal{P}_\Omega(\boldsymbol{M})$ with $\mathcal{P}_\Omega(\boldsymbol{M}^\star) + \mathcal{P}_\Omega(\boldsymbol{E})$. Heuristically, if Ω and $\boldsymbol{Z}$ are statistically independent, then $\boldsymbol{Z}^0$ serves as an unbiased estimator of $\boldsymbol{M}^\star$, i.e. $\mathbb{E}[\boldsymbol{Z}^0] = \boldsymbol{M}^\star$; this arises since the noise $\boldsymbol{E}$ has zero mean and $\mathbb{E}[\mathcal{P}_\Omega] = p\mathcal{I}$ under the uniform random sampling model, with $\mathcal{I}$ the identity operator. Despite its (near) unbiasedness nature at a heuristic level, however, the matrix $\boldsymbol{Z}^0$ is typically full-rank, with non-negligible energy spread across its entire spectrum. This results in dramatically increased variability in the estimate, which is undesirable for inferential purposes.

To remedy this issue, we propose to further project $\boldsymbol{Z}^0$ onto the set of rank-r matrices, leading to the following de-biased estimator

$$\boldsymbol{M}^{\mathsf{d}} \triangleq \mathcal{P}_{\text{rank-}r}\Big[\boldsymbol{Z} - \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{Z} - \boldsymbol{M})\Big], \tag{3.7}$$

where $\mathcal{P}_{\text{rank-}r}(\boldsymbol{B}) = \arg\min_{\boldsymbol{A}:\text{rank}(\boldsymbol{A})\leq r} \|\boldsymbol{A} - \boldsymbol{B}\|_{\text{F}}$, and $\boldsymbol{Z}$ can again be found in Table 1. This projection step effectively suppresses the variability outside the r-dimensional principal subspace. As we shall see shortly, the proposed estimator (3.7) properly de-biases the provided estimate $\boldsymbol{Z}$, while optimally controlling the extent of uncertainty.

Remark 1. *The estimator (3.7) can be viewed as performing one iteration of singular value projection (SVP) [MJD09, DC18] on the current estimate $\boldsymbol{Z}$.*

Remark 2. *The estimator (3.7) also bears a similarity to the de-biased estimator proposed by [Xia18] for low-rank trace regression; the disparity between them shall be discussed in Section 4.*

Table 1: Notation used to unify the convex estimate $\boldsymbol{Z}^{\mathsf{cvx}}$ and the nonconvex estimate $(\boldsymbol{X}^{\mathsf{ncvx}}, \boldsymbol{Y}^{\mathsf{ncvx}})$. Here, $\boldsymbol{Z}^{\mathsf{cvx},r} = \mathcal{P}_{\text{rank-}r}(\boldsymbol{Z}^{\mathsf{cvx}})$ is the best rank-r approximation of $\boldsymbol{Z}^{\mathsf{cvx}}$. See Appendix B for a complete summary.

$\boldsymbol{Z} \in \mathbb{R}^{n\times n}$	either $\boldsymbol{Z}^{\mathsf{cvx}}$ or $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$.
$\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}$	for the nonconvex case, we take $\boldsymbol{X} = \boldsymbol{X}^{\mathsf{ncvx}}$ and $\boldsymbol{Y} = \boldsymbol{Y}^{\mathsf{ncvx}}$; for the convex case, let $\boldsymbol{X} = \boldsymbol{X}^{\mathsf{cvx}}$ and $\boldsymbol{Y} = \boldsymbol{Y}^{\mathsf{cvx}}$, which are the *balanced* low-rank factors of $\boldsymbol{Z}^{\mathsf{cvx},r}$ obeying $\boldsymbol{Z}^{\mathsf{cvx},r} = \boldsymbol{X}^{\mathsf{cvx}}\boldsymbol{Y}^{\mathsf{cvx}\top}$ and $\boldsymbol{X}^{\mathsf{cvx}\top}\boldsymbol{X}^{\mathsf{cvx}} = \boldsymbol{Y}^{\mathsf{cvx}\top}\boldsymbol{Y}^{\mathsf{cvx}}$.
$\boldsymbol{M}^{\mathsf{d}} \in \mathbb{R}^{n\times n}$	the proposed de-biased estimator as in (3.7).
$\boldsymbol{X}^{\mathsf{d}}, \boldsymbol{Y}^{\mathsf{d}} \in \mathbb{R}^{n\times r}$	the proposed de-shrunken estimator as in (3.8).

An equivalent form: a de-shrunken estimator for the low-rank factors. It turns out that the de-biased estimator (3.7) admits another almost equivalent representation that offers further insights. Specifically, we consider the following *de-shrunken* estimator for the low-rank factors

$$\boldsymbol{X}^{\mathsf{d}} \triangleq \boldsymbol{X}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\Big)^{1/2} \qquad \text{and} \qquad \boldsymbol{Y}^{\mathsf{d}} \triangleq \boldsymbol{Y}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\big(\boldsymbol{Y}^\top\boldsymbol{Y}\big)^{-1}\Big)^{1/2}, \tag{3.8}$$

where we recall the definition of $\boldsymbol{X}$ and $\boldsymbol{Y}$ in Table 1. To develop some intuition regarding why this is called a de-shrunken estimator, let us look at a simple scenario where $\boldsymbol{U\Sigma V}^\top$ is the SVD of $\boldsymbol{XY}^\top$ and $\boldsymbol{X} = \boldsymbol{U\Sigma}^{1/2}$, $\boldsymbol{Y} = \boldsymbol{V\Sigma}^{1/2}$. It is then self-evident that

$$\boldsymbol{X}^{\mathsf{d}} = \boldsymbol{U\Sigma}^{1/2}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\boldsymbol{\Sigma}^{-1}\Big)^{1/2} = \boldsymbol{U}\Big(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r\Big)^{1/2} \qquad \text{and} \qquad \boldsymbol{Y}^{\mathsf{d}} = \boldsymbol{V}\Big(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r\Big)^{1/2}.$$

In words, $\boldsymbol{X}^{\mathsf{d}}$ and $\boldsymbol{Y}^{\mathsf{d}}$ are obtained by de-shrinking the spectrum of $\boldsymbol{X}$ and $\boldsymbol{Y}$ properly.

As we shall formalize in Section 5.1, the de-shrunken estimator (3.8) for the low-rank factors is nearly equivalent to the de-biased estimator (3.7) for the whole matrix, in the sense that

$$\boldsymbol{M}^{\mathsf{d}} \approx \boldsymbol{X}^{\mathsf{d}}\boldsymbol{Y}^{\mathsf{d}\top}. \tag{3.9}$$

Therefore, $\boldsymbol{M}^{\mathsf{d}}$ can be viewed as some sort of de-shrunken estimator as well.

3.3 Main results: distributional guarantees

The proposed estimators admit tractable distributional characterizations in the large-n regime, which facilitates the construction of confidence regions for many quantities of interest. In particular, this paper centers around two types of inferential problems:

1. *Each entry of the matrix* $\boldsymbol{M}^\star$: the entry can be either missing (i.e. predicting an unseen entry) or observed (i.e. de-noising an observed entry). For example, in the Netflix challenge, one would like to infer a user's preference about any movie, given partially revealed ratings [CR09]. Mathematically, this seeks to determine the distribution of

$$M_{ij}^{\mathsf{d}} - M_{ij}^\star, \qquad \text{for all } 1 \le i, j \le n. \tag{3.10}$$

2. *The low-rank factors* $\boldsymbol{X}^\star, \boldsymbol{Y}^\star \in \mathbb{R}^{n\times r}$: the low-rank factors often reveal critical information about the applications of interest (e.g. community memberships of each individual in the community detection problem [AFWZ17], or angles between each object and a global reference point in the angular synchronization problem [Sin11]). Recognizing the global rotational ambiguity issue,[2] we aim to pin down the distributions of $\boldsymbol{X}^{\mathsf{d}}$ and $\boldsymbol{Y}^{\mathsf{d}}$ up to global rotational ambiguity. More precisely, we intend to characterize the distributions of

$$\boldsymbol{X}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{X}^\star \qquad \text{and} \qquad \boldsymbol{Y}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{Y}^\star \tag{3.11}$$

for the global rotation matrix $\boldsymbol{H}^{\mathsf{d}} \in \mathbb{R}^{r\times r}$ that best "aligns" $(\boldsymbol{X}^{\mathsf{d}}, \boldsymbol{Y}^{\mathsf{d}})$ and $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$, i.e.

$$\boldsymbol{H}^{\mathsf{d}} \triangleq \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{X}^{\mathsf{d}}\boldsymbol{R} - \boldsymbol{X}^\star\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y}^{\mathsf{d}}\boldsymbol{R} - \boldsymbol{Y}^\star\right\|_{\mathrm{F}}^2. \tag{3.12}$$

Here and below, $\mathcal{O}^{r\times r}$ denotes the set of orthonormal matrices in $\mathbb{R}^{r\times r}$.

Clearly, the above two inferential problems are tightly related: an accurate distributional characterization for the low-rank factors (3.11) often results in a distributional guarantee for the entries (3.10). As such, we shall begin by presenting our distributional characterizations of the low-rank factors. Here and throughout, $\boldsymbol{e}_i$ represents the ith standard basis vector in $\mathbb{R}^n$.

[2] For any $r \times r$ rotation matrix $\boldsymbol{H}$, we cannot distinguish $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$ from $(\boldsymbol{X}^\star\boldsymbol{H}, \boldsymbol{Y}^\star\boldsymbol{H})$, if only pairwise measurements are available.

Theorem 1 (Distributional guarantees w.r.t. low-rank factors)**.** *Suppose that the sample size and the noise obey*

$$np \gtrsim \kappa^8\mu^3 r^2 \log^3 n \qquad \textit{and} \qquad \sigma/\sigma_{\min} \lesssim \sqrt{p/(\kappa^8\mu n \log^2 n)}. \tag{3.13}$$

Then one has the following decomposition

$$\boldsymbol{X}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{X}^{\star} = \boldsymbol{Z}_{\boldsymbol{X}} + \boldsymbol{\Psi}_{\boldsymbol{X}}, \tag{3.14a}$$

$$\boldsymbol{Y}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{Y}^{\star} = \boldsymbol{Z}_{\boldsymbol{Y}} + \boldsymbol{\Psi}_{\boldsymbol{Y}}. \tag{3.14b}$$

with $(\boldsymbol{X}^{\star}, \boldsymbol{Y}^{\star})$ *defined in* (2.2), $(\boldsymbol{X}^{\mathsf{d}}, \boldsymbol{Y}^{\mathsf{d}})$ *defined in Table* 1, *and* $\boldsymbol{H}^{\mathsf{d}}$ *defined in* (3.12)*. Here, the rows of* $\boldsymbol{Z}_{\boldsymbol{X}} \in \mathbb{R}^{n\times r}$ *(resp.* $\boldsymbol{Z}_{\boldsymbol{Y}} \in \mathbb{R}^{n\times r}$*) are independent and obey*

$$\boldsymbol{Z}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}\Big(\boldsymbol{0}, \frac{\sigma^2}{p}\left(\boldsymbol{\Sigma}^{\star}\right)^{-1}\Big), \qquad \text{for} \quad 1 \leq j \leq n; \tag{3.15a}$$

$$\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\boldsymbol{e}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}\Big(\boldsymbol{0}, \frac{\sigma^2}{p}\left(\boldsymbol{\Sigma}^{\star}\right)^{-1}\Big), \qquad \text{for} \quad 1 \leq j \leq n. \tag{3.15b}$$

In addition, the residual matrices $\boldsymbol{\Psi}_{\boldsymbol{X}}, \boldsymbol{\Psi}_{\boldsymbol{Y}} \in \mathbb{R}^{n\times r}$ *satisfy, with probability at least* $1 - O(n^{-3})$, *that*

$$\max\left\{ \|\boldsymbol{\Psi}_{\boldsymbol{X}}\|_{2,\infty}, \|\boldsymbol{\Psi}_{\boldsymbol{Y}}\|_{2,\infty} \right\} = o\left(\frac{\sigma\sqrt{r}}{\sqrt{p\sigma_{\max}}}\right). \tag{3.16}$$

Remark 3. *A more complete version can be found in Theorem* 5*.*

Remark 4. *Another interesting feature — which we shall make precise in the proof of this theorem — is that: for any given* $1 \leq i, j \leq n$, *the two random vectors* $\boldsymbol{Z}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i$ *and* $\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\boldsymbol{e}_j$ *are nearly statistically independent. This is crucial for deriving inferential guarantees for the entries of the matrix.*

Theorem 1 is a non-asymptotic result. In words, Theorem 1 decomposes the estimation error $\boldsymbol{X}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{X}^{\star}$ (resp. $\boldsymbol{Y}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{Y}^{\star}$) into a Gaussian component $\boldsymbol{Z}_{\boldsymbol{X}}$ (resp. $\boldsymbol{Z}_{\boldsymbol{Y}}$) and a residual term $\boldsymbol{\Psi}_{\boldsymbol{X}}$ (resp. $\boldsymbol{\Psi}_{\boldsymbol{Y}}$). If the sample size is sufficiently large and the noise size is sufficiently small, then the residual terms are much smaller in size compared to $\boldsymbol{Z}_{\boldsymbol{X}}$ and $\boldsymbol{Z}_{\boldsymbol{Y}}$. To see this, it is helpful to leverage the Gaussianity (3.15a) and compute that: for each $1 \leq j \leq n$, the jth row of $\boldsymbol{Z}_{\boldsymbol{X}}$ obeys

$$\mathbb{E}\left[\left\|\boldsymbol{Z}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_j\right\|_2^2\right] = \mathsf{Tr}\Big(\frac{\sigma^2}{p}\left(\boldsymbol{\Sigma}^{\star}\right)^{-1}\Big) \geq \frac{\sigma^2 r}{p\sigma_{\max}};$$

in other words, the typical size of the jth row of $\boldsymbol{Z}_{\boldsymbol{X}}$ is no smaller than the order of $\sigma\sqrt{r/(p\sigma_{\max})}$. In comparison, the size of each row of $\boldsymbol{\Psi}_{\boldsymbol{X}}$ (see (3.16)) is much smaller than $\sigma\sqrt{r/(p\sigma_{\max})}$ (and hence smaller than the size of the corresponding row of $\boldsymbol{Z}_{\boldsymbol{X}}$) with high probability, provided that (3.13) is satisfied.

Equipped with the above master decompositions of the low-rank factors and Remark 4, we are ready to present a similar decomposition for the entry $M_{ij}^{\mathsf{d}} - M_{ij}^{\star}$.

Theorem 2 (Distributional guarantees w.r.t. matrix entries)**.** *For each* $1 \leq i, j \leq n$, *define the variance* $v_{ij}^{\star}$ *as*

$$v_{ij}^{\star} \triangleq \frac{\sigma^2}{p}\left(\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2^2 + \left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2^2\right), \tag{3.17}$$

where $\boldsymbol{U}_{i,\cdot}^{\star}$ *(resp.* $\boldsymbol{V}_{j,\cdot}^{\star}$*) denotes the ith (resp. jth) row of* $\boldsymbol{U}^{\star}$ *(resp.* $\boldsymbol{V}^{\star}$*). Suppose that*

$$np \gtrsim \kappa^8\mu^3 r^3 \log^3 n, \qquad \sigma\sqrt{(\kappa^8\mu r n \log^2 n)/p} \lesssim \sigma_{\min} \qquad \textit{and} \tag{3.18a}$$

$$\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2 + \left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2 \gtrsim \sqrt{\frac{r}{n}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^6\mu^2 r n \log^3 n}{p}}. \tag{3.18b}$$

Then the matrix $\boldsymbol{M}^{\mathsf{d}}$ *defined in Table* 1 *satisfies*

$$M_{ij}^{\mathsf{d}} - M_{ij}^{\star} = g_{ij} + \Delta_{ij}, \tag{3.19}$$

where $g_{ij} \sim \mathcal{N}(0, v_{ij}^{\star})$ *and the residual obeys* $|\Delta_{ij}| = o(\sqrt{v_{ij}^{\star}})$ *with probability exceeding* $1 - O(n^{-3})$*.*

Remark 5 (The symmetric case)**.** *In the symmetric case where the noise $\boldsymbol{E}$, the truth $\boldsymbol{M}^\star$, and the sampling pattern are all symmetric (i.e. $\mathcal{P}_\Omega(\boldsymbol{E}) = \big(\mathcal{P}_\Omega(\boldsymbol{E})\big)^\top$ and $\boldsymbol{M}^\star = \boldsymbol{M}^{\star\top}$), the variance $v_{ii}^\star$ (cf. (3.17)) for the diagonal entries has a different formula; more specifically, it is straightforward to extend our theory to show that*

$$v_{ii}^\star = \frac{4\sigma^2}{p}\left\|\boldsymbol{U}_{i,\cdot}^\star\right\|_2^2 = \frac{2\sigma^2}{p}\left(\left\|\boldsymbol{U}_{i,\cdot}^\star\right\|_2^2 + \left\|\boldsymbol{V}_{i,\cdot}^\star\right\|_2^2\right) \qquad \textit{for the symmetric case.}$$

This additional multiplicative factor of 2 arises since $\boldsymbol{Z}_{\boldsymbol{X}}^\top \boldsymbol{e}_i$ and $\boldsymbol{Z}_{\boldsymbol{Y}}^\top \boldsymbol{e}_i$ are identical (and hence not independent) in this symmetric case. The variance formula for any $v_{ij}^\star$ $(i \neq j)$ remains unchanged.

Several remarks are in order. To begin with, we develop some intuition regarding where the variance $v_{ij}^\star$ comes from. By virtue of Theorem 1, one has the following Gaussian approximation

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H}^{\mathrm{d}} - \boldsymbol{X}^\star \approx \boldsymbol{Z}_{\boldsymbol{X}} \qquad \text{and} \qquad \boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H}^{\mathrm{d}} - \boldsymbol{Y}^\star \approx \boldsymbol{Z}_{\boldsymbol{Y}}.$$

Assuming that the first-order expansion is reasonably tight, one has

$$\begin{aligned} M_{ij}^{\mathrm{d}} - M_{ij}^\star &= \left[\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H}^{\mathrm{d}}\big(\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H}^{\mathrm{d}}\big)^\top - \boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}\right]_{ij} \approx \boldsymbol{e}_i^\top\big(\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H}^{\mathrm{d}} - \boldsymbol{X}^\star\big)\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j + \boldsymbol{e}_i^\top\boldsymbol{X}^\star\big(\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H}^{\mathrm{d}} - \boldsymbol{Y}^\star\big)^\top\boldsymbol{e}_j \\ &\approx \boldsymbol{e}_i^\top\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j + \boldsymbol{e}_i^\top\boldsymbol{X}^\star\boldsymbol{Z}_{\boldsymbol{Y}}^\top\boldsymbol{e}_j. \end{aligned} \tag{3.20}$$

According to Remark 4, $\boldsymbol{Z}_{\boldsymbol{X}}^\top \boldsymbol{e}_i$ and $\boldsymbol{Z}_{\boldsymbol{Y}}^\top \boldsymbol{e}_j$ are nearly independent. It is thus straightforward to compute the variance of (3.20) as

$$\begin{aligned} \mathsf{Var}\left(M_{ij}^{\mathrm{d}} - M_{ij}^\star\right) &\overset{\text{(i)}}{\approx} \mathsf{Var}\left(\boldsymbol{e}_i^\top\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j\right) + \mathsf{Var}\left(\boldsymbol{e}_i^\top\boldsymbol{X}^\star\boldsymbol{Z}_{\boldsymbol{Y}}^\top\boldsymbol{e}_j\right) \\ &\overset{\text{(ii)}}{=} \frac{\sigma^2}{p}\left\{\boldsymbol{e}_j^\top\boldsymbol{Y}^\star\left(\boldsymbol{\Sigma}^\star\right)^{-1}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j + \boldsymbol{e}_i^\top\boldsymbol{X}^\star\left(\boldsymbol{\Sigma}^\star\right)^{-1}\boldsymbol{X}^{\star\top}\boldsymbol{e}_i\right\} \overset{\text{(iii)}}{=} \frac{\sigma^2}{p}\left(\left\|\boldsymbol{U}_{i,\cdot}^\star\right\|_2^2 + \left\|\boldsymbol{V}_{j,\cdot}^\star\right\|_2^2\right) = v_{ij}^\star. \end{aligned}$$

Here, (i) relies on (3.20) and the near independence between $\boldsymbol{Z}_{\boldsymbol{X}}^\top \boldsymbol{e}_i$ and $\boldsymbol{Z}_{\boldsymbol{Y}}^\top \boldsymbol{e}_j$; (ii) uses the variance formula in Theorem 1; (iii) arises from the definitions of $\boldsymbol{X}^\star$ and $\boldsymbol{Y}^\star$ (cf. (2.2)). This computation explains (heuristically) the variance formula $v_{ij}^\star$.

Given that Theorem 2 reveals the tightness of Gaussian approximation under conditions (3.18), it in turn allows us to construct nearly accurate confidence intervals for each matrix entry $M_{ij}^\star$. This is formally summarized in the following corollary, the proof of which is deferred to Appendix F. Here and throughout, we use $[a \pm b]$ to denote the interval $[a-b, a+b]$.

Corollary 1 (Confidence intervals for the entries $\{M_{ij}^\star\}$)**.** *Let $\boldsymbol{X}^{\mathrm{d}}$, $\boldsymbol{Y}^{\mathrm{d}}$ and $\boldsymbol{M}^{\mathrm{d}}$ be as defined in Table 1. For any given $1 \leq i, j \leq n$, suppose that (3.18a) holds and that*

$$\left\|\boldsymbol{U}_{i,\cdot}^\star\right\|_2 + \left\|\boldsymbol{V}_{j,\cdot}^\star\right\|_2 \gtrsim \sqrt{\frac{r}{n}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^{10}\mu^2 rn\log^3 n}{p}}. \tag{3.21}$$

Denote by $\Phi(t)$ the CDF of a standard Gaussian random variable and by $\Phi^{-1}(\cdot)$ its inverse function. Let

$$v_{ij} \triangleq \frac{\sigma^2}{p}\left(\boldsymbol{X}_{i,\cdot}^{\mathsf{d}}\left(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}}\right)^{-1}\left(\boldsymbol{X}_{i,\cdot}^{\mathsf{d}}\right)^\top + \boldsymbol{Y}_{j,\cdot}^{\mathsf{d}}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)^{-1}\left(\boldsymbol{Y}_{j,\cdot}^{\mathsf{d}}\right)^\top\right) \tag{3.22}$$

be the empirical estimate of the theoretical variance $v_{ij}^\star$. Then one has

$$\sup_{0<\alpha<1}\left|\mathbb{P}\left\{M_{ij}^\star \in \left[M_{ij}^{\mathsf{d}} \pm \Phi^{-1}\left(1-\alpha/2\right)\sqrt{v_{ij}}\right]\right\} - (1-\alpha)\right| = o(1).$$

In words, Corollary 1 tells us that for any fixed significance level $0 < \alpha < 1$, the interval

$$\left[M_{ij}^{\mathsf{d}} \pm \Phi^{-1}(1-\alpha/2)\sqrt{v_{ij}}\right] \tag{3.23}$$

is a nearly accurate two-sided $(1-\alpha)$ confidence interval of $M_{ij}^\star$.

In addition, we remark that when $\|\boldsymbol{U}_{i,\cdot}^{\star}\|_2 = \|\boldsymbol{V}_{j,\cdot}^{\star}\|_2 = 0$ (and hence $V_{ij}^{\star} = 0$), the above Gaussian approximation is completely off. In this case, one can still leverage Theorem 1 to show that

$$M_{ij}^{\mathsf{d}} - M_{ij}^{\star} = M_{ij}^{\mathsf{d}} \approx \boldsymbol{u}^{\top}\boldsymbol{v}, \tag{3.24}$$

where $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^r$ are independent and identically distributed according to $\mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{\Sigma}^{\star})^{-1}/p)$. However, it is nontrivial to determine whether $\|\boldsymbol{U}_{i,\cdot}^{\star}\|_2 + \|\boldsymbol{V}_{j,\cdot}^{\star}\|_2$ is vanishingly small or not based on the observed data, which makes it challenging to conduct efficient inference for entries with small (but *a priori* unknown) $\|\boldsymbol{U}_{i,\cdot}^{\star}\|_2 + \|\boldsymbol{V}_{j,\cdot}^{\star}\|_2$.

Last but not least, the careful readers might wonder how to interpret our conditions on the sample complexity and the signal-to-noise ratio. Take the case with $r, \mu, \kappa = O(1)$ for example: our conditions read

$$n^2 p \gtrsim n \log^3 n; \qquad \sigma\sqrt{(n\log^2 n)/p} \lesssim \sigma_{\min}. \tag{3.25}$$

The first condition matches the minimal sample complexity limit (up to some logarithmic factor), while the second one coincides with the regime (up to log factor) in which popular algorithms (like spectral methods or nonconvex algorithms) work better than a random guess [KMO10b, CW15, MWCC17]. The take-away message is this: once we are able to compute a reasonable estimate in an overall ℓ_2 sense, then we can reinforce it to conduct entrywise inference in a statistically efficient fashion. The discussion of the dependency on r and κ is deferred to Section 6.

3.4 Lower bounds and optimality for inference

It is natural to ask how well our inferential procedures perform compared to other algorithms. Encouragingly, the de-biased estimator is optimal in some sense; for instance, it nearly attains the minimum covariance among all unbiased estimators. To formalize this claim, we shall

1. Quantify the performance of two ideal estimators with the assistance of an oracle;
2. Demonstrate that the performance of our de-biased estimators is arbitrarily close to that of the ideal estimators.

In what follows, we denote by $\boldsymbol{X}_{i,\cdot}^{\star}$ (resp. $\boldsymbol{Y}_{i,\cdot}^{\star}$) the ith row of $\boldsymbol{X}^{\star}$ (resp. $\boldsymbol{Y}^{\star}$).

An ideal estimator for $\boldsymbol{X}_{i,\cdot}^{\star}$ ($1 \leq i \leq n$). Suppose that there is an oracle informing us of $\boldsymbol{Y}^{\star}$, and that we observe the same set of data as in (2.3). Under such an idealistic setting and for any given $1 \leq i \leq n$, the following least-squares estimator achieves the minimum covariance among all *unbiased* estimators for the ith row $\boldsymbol{X}_{i,\cdot}^{\star}$ of $\boldsymbol{X}^{\star}$ (see e.g. [Sha03, Theorem 3.7])

$$\boldsymbol{X}_{i,\cdot}^{\mathsf{ideal}} \triangleq \arg\min_{\boldsymbol{u}\in\mathbb{R}^{1\times r}} \sum_{k:(i,k)\in\Omega} \left[M_{ik} - \boldsymbol{u}\big(\boldsymbol{Y}_{k,\cdot}^{\star}\big)^{\top}\right]^2. \tag{3.26}$$

In other words, for any unbiased estimator $\boldsymbol{u}$ of $\boldsymbol{X}_{i,\cdot}^{\star}$ (conditional on Ω), one has

$$\mathsf{Cov}\big(\boldsymbol{u}\,\big|\,\Omega\big) \succeq \mathsf{Cov}\big(\boldsymbol{X}_{i,\cdot}^{\mathsf{ideal}}\,\big|\,\Omega\big) =: \mathsf{CRLB}\big(\boldsymbol{X}_{i,\cdot}^{\star}\mid\Omega\big), \tag{3.27}$$

where $\mathsf{Cov}\big(\boldsymbol{X}_{i,\cdot}^{\mathsf{ideal}}\,\big|\,\Omega\big)$ is precisely the Cramér-Rao lower bound (conditional on Ω) under this ideal setting. As it turns out, with high probability, this lower bound concentrates around $\sigma^2(\boldsymbol{\Sigma}^{\star})^{-1}/p$, as stated in the following lemma. The proof is postponed to Appendix H.1.

Lemma 1. *Fix an arbitrarily small constant $\varepsilon > 0$. Suppose that $n^2 p \geq C_0\varepsilon^{-2}\kappa^4\mu r n$ for some sufficiently large constant $C_0 > 0$ independent of n. Then with probability at least $1 - O(n^{-10})$, one has*

$$\mathsf{CRLB}\big(\boldsymbol{X}_{i,\cdot}^{\star}\mid\Omega\big) \succeq (1-\varepsilon)\,\frac{\sigma^2}{p}\,(\boldsymbol{\Sigma}^{\star})^{-1}.$$

Given that ε can be an arbitrarily small constant, Lemma 1 uncovers that the covariance of the de-shrunken estimator $\boldsymbol{X}_{i,\cdot}^{\mathsf{d}}$ (cf. Theorem 1) matches that of the ideal estimator $\boldsymbol{X}_{i,\cdot}^{\mathsf{ideal}}$, thus achieving the Cramér-Rao lower bound with high probability. The same conclusion applies to $\boldsymbol{Y}_{j,\cdot}^{\mathsf{d}}$ as well.

An ideal estimator for $M^\star_{ij}$ $(1 \leq i,j \leq n)$. Suppose that there is another oracle informing us of $\{\boldsymbol{X}^\star_{k,\cdot}\}_{k:k\neq i}$ and $\{\boldsymbol{Y}^\star_{k,\cdot}\}_{k:k\neq j}$; that is, everything about $\boldsymbol{X}^\star$ except $\boldsymbol{X}^\star_{i,\cdot}$ and everything about $\boldsymbol{Y}^\star$ except $\boldsymbol{Y}^\star_{j,\cdot}$. In addition, we observe the same set of data as in (2.3), except that we do not get to see M_{ij}.[3] Under this idealistic model, the Cramér-Rao lower bound [Sha03, Theorem 3.3] for estimating $M^\star_{ij} = \boldsymbol{X}^\star_{i,\cdot}(\boldsymbol{Y}^\star_{j,\cdot})^\top$ can be computed as

$$\begin{aligned}&\mathsf{CRLB}\left(M^\star_{ij} \mid \Omega\right)\\ &\triangleq \frac{\sigma^2}{p}\cdot\Bigg[\boldsymbol{Y}^\star_{j,\cdot}\Big(\frac{1}{p}\sum_{k:k\neq j,(i,k)\in\Omega}(\boldsymbol{Y}^\star_{k,\cdot})^\top\boldsymbol{Y}^\star_{k,\cdot}\Big)^{-1}(\boldsymbol{Y}^\star_{j,\cdot})^\top + \boldsymbol{X}^\star_{i,\cdot}\Big(\frac{1}{p}\sum_{k:k\neq i,(k,j)\in\Omega}(\boldsymbol{X}^\star_{k,\cdot})^\top\boldsymbol{X}^\star_{k,\cdot}\Big)^{-1}(\boldsymbol{X}^\star_{i,\cdot})^\top\Bigg]. \qquad (3.28)\end{aligned}$$

This means that any unbiased estimator of $M^\star_{ij}$ must have variance no smaller than $\mathsf{CRLB}(M^\star_{ij} \mid \Omega)$. This quantity admits a much simpler lower bound as follows, whose proof can be found in Appendix H.2.

Lemma 2. *Fix an arbitrarily small constant $\varepsilon > 0$. Suppose that $n^2p \geq C_0\varepsilon^{-2}\kappa^4\mu rn\log n$ for some sufficiently large constant $C_0 > 0$ independent of n. Then with probability at least $1 - O(n^{-10})$,*

$$\mathsf{CRLB}\big(M^\star_{ij} \mid \Omega\big) \geq (1-\varepsilon)\, v^\star_{ij},$$

where $v^\star_{ij}$ is defined in Theorem 2.

Similar to Lemma 1, Lemma 2 reveals that the variance of our de-biased estimator M^{d}_{ij} (cf. Theorem 2) — which certainly does not have access to the side information provided by the oracle — is arbitrarily close to the Cramér-Rao lower bound aided by an oracle.

All in all, the above lower bounds demonstrate that the degrees of uncertainty underlying our de-shrunken low-rank factors and de-biased matrix are, in some sense, statistically minimal.

3.5 Back to estimation: the de-biased estimator is optimal

While the emphasis of the current paper is on inference, we would nevertheless like to single out an important consequence that informs the estimation step. To be specific, the decompositions and distributional guarantees derived in Theorem 1 and Theorem 2 allow us to track the estimation accuracy of $\boldsymbol{M}^{\mathsf{d}}$, as stated in the following theorem. The proof of this result is postponed to Appendix G.

Theorem 3 (Estimation accuracy of $\boldsymbol{M}^{\mathsf{d}}$)**.** *Let $\boldsymbol{M}^{\mathsf{d}}$ be the de-biased estimator as defined in Table 1. Instate the conditions in (3.18a). Then with probability at least $1 - O(n^{-3})$, one has*

$$\big\|\boldsymbol{M}^{\mathsf{d}} - \boldsymbol{M}^\star\big\|_{\mathrm{F}}^2 = \frac{(2+o(1))nr\sigma^2}{p}. \qquad (3.29)$$

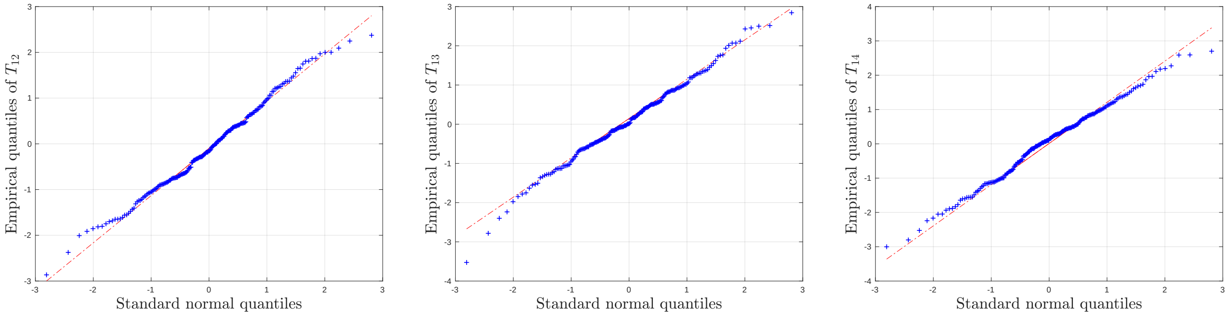
In stark contrast to prior statistical estimation guarantees (e.g. [CP10, NW12, KLT11, CCF+19]), Theorem 3 pins down the estimation error of the proposed de-biased estimator in a sharp manner (namely, even the pre-constant is fully determined). Encouragingly, there is a sense in which the proposed de-biased estimator achieves the best possible statistical estimation accuracy, as revealed by the following result.

Theorem 4 (An oracle lower bound on ℓ_2 estimation errors)**.** *Fix an arbitrarily small constant $\varepsilon > 0$. Suppose that $n^2p \gtrsim \mu rn\log^2 n$, and that $r = o(n)$. Then with probability exceeding $1 - O(n^{-10})$, any unbiased estimator $\widehat{\boldsymbol{M}}$ of $\boldsymbol{M}^\star$ obeys*

$$\mathbb{E}\Big[\big\|\widehat{\boldsymbol{M}} - \boldsymbol{M}^\star\big\|_{\mathrm{F}}^2 \mid \Omega\Big] \geq \frac{(1-\varepsilon)2nr\sigma^2}{p}. \qquad (3.30)$$

Proof. Intuitively, the term $2nr$ reflects approximately the underlying degrees of freedom in the true subspace $T^\star$ of interest (i.e. the tangent space of the rank-r matrices at the truth $\boldsymbol{M}^\star$), whereas the factor $1/p$ captures the effect due to sub-sampling. This result has already been established in [CP10, Section III.B] (together with [CR09, Theorem 4.1]). We thus omit the proof for conciseness. The key idea is to consider an oracle informing us of the true tangent space $T^\star$. □

[3] The exclusion of M_{ij} is merely for ease of presentation. One can consider the model where all M_{ij} with $(i,j) \in \Omega$ are observed with a slightly more complicated argument.



(a) (b) (c)

Figure 1: Q-Q (quantile-quantile) plots of T_{12}, T_{13} and T_{14} vs. the standard normal distribution in (a), (b) and (c), respectively. The results are reported over 200 independent trials for $r = 5$, $p = 0.4$ and $\sigma = 10^{-3}$.

The implication of the above two theorems is remarkable: the de-biasing step not merely facilitates uncertainty assessment, but also proves crucial in minimizing the estimation errors. It achieves optimal statistical efficiency in terms of both the rate and the pre-constant. As far as we know, this is the first theory about a polynomial time algorithm that matches the statistical limit in terms of the pre-constant. This intriguing finding is further corroborated by numerical experiments; see Section 3.6 for details (in particular, Figure 3).

3.6 Numerical experiments

We conduct numerical experiments on synthetic data to verify the distributional characterizations provided in Theorem 1 and Theorem 2. Note that our main results hold for the de-biased estimators built upon $\boldsymbol{Z}^{\mathsf{cvx}}$ and $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$. As we will formalize shortly in Section 5.1, these two de-biased estimators are extremely close to each other; see also Figure 4 for experimental evidence. Therefore, in order to save space, we use the de-biased estimator built upon the convex estimate $\boldsymbol{Z}^{\mathsf{cvx}}$ throughout the experiments.

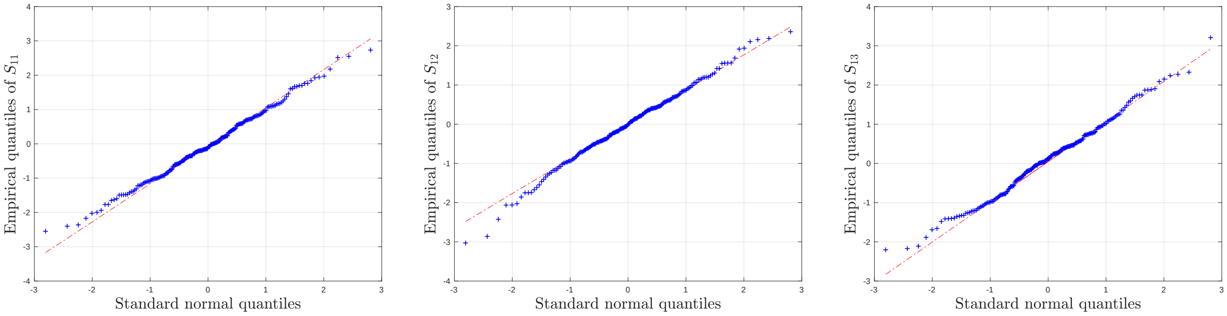
Fix the dimension $n = 1000$ and the regularization parameter $\lambda = 2.5\sigma\sqrt{np}$ throughout the experiments. We generate a rank-r matrix $\boldsymbol{M}^{\star} = \boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top}$, where $\boldsymbol{X}^{\star}, \boldsymbol{Y}^{\star} \in \mathbb{R}^{n\times r}$ are random orthonormal matrices and apply the proximal gradient method [PB14] to solve the convex program (3.1).

We begin by checking the validity of Theorem 1. Suppose that one is interested in estimating the inner product $\boldsymbol{e}_i^{\top}\boldsymbol{X}^{\star}\boldsymbol{X}^{\star\top}\boldsymbol{e}_j$ between $\boldsymbol{X}^{\star\top}\boldsymbol{e}_i$ and $\boldsymbol{X}^{\star\top}\boldsymbol{e}_j$ $(i \neq j)$. In the Netflix challenge, this might correspond to the similarity between the ith user and the jth one. As a straightforward consequence of Theorem 1, the normalized estimation error

$$T_{ij} \triangleq \frac{1}{\sqrt{\rho_{ij}}}\left(\boldsymbol{e}_i^{\top}\boldsymbol{X}^{\mathsf{d}}\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{e}_j - \boldsymbol{e}_i^{\top}\boldsymbol{X}^{\star}\boldsymbol{X}^{\star\top}\boldsymbol{e}_j\right) \tag{3.31}$$

Table 2: Empirical coverage rates of $\boldsymbol{e}_i^{\top}\boldsymbol{X}^{\star}\boldsymbol{X}^{\star\top}\boldsymbol{e}_j$ for different (r, p, σ)'s over 200 Monte Carlo trials.

(r, p, σ)	$\mathsf{Mean}(\widehat{\mathsf{Cov}_{\mathsf{L}}})$	$\mathsf{Std}(\widehat{\mathsf{Cov}_{\mathsf{L}}})$
$(2, 0.2, 10^{-6})$	0.9387	0.0197
$(2, 0.2, 10^{-3})$	0.9400	0.0193
$(2, 0.4, 10^{-6})$	0.9459	0.0161
$(2, 0.4, 10^{-3})$	0.9460	0.0162
$(5, 0.2, 10^{-6})$	0.9227	0.0244
$(5, 0.2, 10^{-3})$	0.9273	0.0226
$(5, 0.4, 10^{-6})$	0.9411	0.0173
$(5, 0.4, 10^{-3})$	0.9418	0.0171



(a) (b) (c)

Figure 2: Q-Q (quantile-quantile) plot of S_{11}, S_{12} and S_{13} vs. the standard normal distribution in (a), (b) and (c) respectively. The results are reported over 200 independent trials for $r = 5$, $p = 0.4$ and $\sigma = 10^{-3}$.

is extremely close to a standard Gaussian random variable. Here, similar to (3.22), we let

$$\rho_{ij} \triangleq \frac{\sigma^2}{p}\left\{\boldsymbol{e}_i^\top \boldsymbol{X}^{\mathsf{d}}(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}})^{-1}\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{e}_i + \boldsymbol{e}_j^\top \boldsymbol{X}^{\mathsf{d}}(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}})^{-1}\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{e}_j\right\} \tag{3.32}$$

be the empirical estimate of the theoretically predicted variance $\sigma^2(\|\boldsymbol{U}^\star_{i,\cdot}\|_2^2 + \|\boldsymbol{U}^\star_{j,\cdot}\|_2^2)/p$. As a result, a 95% confidence interval of $\boldsymbol{e}_i^\top \boldsymbol{X}^\star \boldsymbol{X}^{\star\top}\boldsymbol{e}_j$ would be $[\boldsymbol{e}_i^\top \boldsymbol{X}^{\mathsf{d}} \boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{e}_j \pm 1.96\sqrt{\rho_{ij}}]$. For each (i,j), we define $\widehat{\mathsf{Cov}}_{\mathsf{L},(i,j)}$ to be the empirical coverage rate of $\boldsymbol{e}_i^\top \boldsymbol{X}^\star \boldsymbol{X}^{\star\top}\boldsymbol{e}_j$ over 200 Monte Carlo simulations. Correspondingly, denote by $\mathsf{Mean}(\widehat{\mathsf{Cov}}_{\mathsf{L}})$ (resp. $\mathsf{Std}(\widehat{\mathsf{Cov}}_{\mathsf{L}})$) the average (resp. the standard deviation) of $\widehat{\mathsf{Cov}}_{\mathsf{L},(i,j)}$ over indices $1 \leq i < j \leq n$. Table 2 collects the simulation results for different values of (r, p, σ). As can be seen, the reported empirical coverage rates are reasonably close to the nominal level 95%. In addition, Figure 1 depicts the Q-Q (quantile-quantile) plots of T_{12}, T_{13} and T_{14} vs. the standard Gaussian random variable over 200 Monte Carlo simulations for $r = 5$, $p = 0.4$ and $\sigma = 10^{-3}$. It is clearly seen that all of these are well approximated by a standard Gaussian random variable.

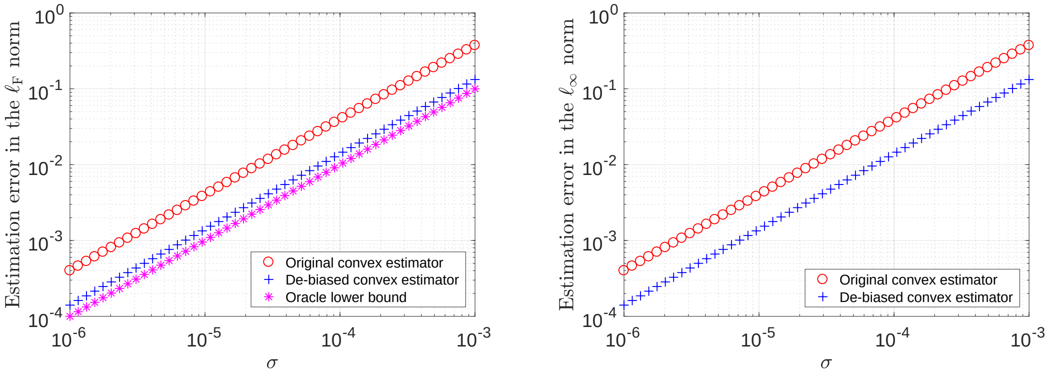
Next, we turn to Theorem 2, namely the distributional guarantee for the entries of the matrix. Denote

$$S_{ij} \triangleq \frac{1}{\sqrt{v_{ij}}}\left(M^{\mathsf{d}}_{ij} - M^\star_{ij}\right), \tag{3.33}$$

where v_{ij} is the empirical variance defined in (3.22). In view of the 95% confidence interval predicted by Corollary 1, and similar to what have done for the low-rank components, for each (i,j), we define $\widehat{\mathsf{Cov}}_{\mathsf{E},(i,j)}$ to be the empirical coverage rate of $M^\star_{ij}$ over 200 Monte Carlo simulations. Correspondingly, denote by $\mathsf{Mean}(\widehat{\mathsf{Cov}}_{\mathsf{E}})$ (resp. $\mathsf{Std}(\widehat{\mathsf{Cov}}_{\mathsf{E}})$) the average (resp. the standard deviation) of $\widehat{\mathsf{Cov}}_{\mathsf{E},(i,j)}$ over indices $1 \leq i, j \leq n$. As before, Table 3 gathers the empirical coverage rates for $M^\star_{ij}$ and Figure 2 displays the Q-Q (quantile-quantile) plots of S_{11}, S_{12} and S_{13} vs. the standard Gaussian random variable over 200 Monte Carlo trials

Table 3: Empirical coverage rates of $M^\star_{ij}$ for different (r, p, σ)'s over 200 Monte Carlo trials.

(r, p, σ)	$\mathsf{Mean}(\widehat{\mathsf{Cov}}_{\mathsf{E}})$	$\mathsf{Std}(\widehat{\mathsf{Cov}}_{\mathsf{E}})$
$(2, 0.2, 10^{-6})$	0.9380	0.0200
$(2, 0.2, 10^{-3})$	0.9392	0.0196
$(2, 0.4, 10^{-6})$	0.9455	0.0164
$(2, 0.4, 10^{-3})$	0.9456	0.0164
$(5, 0.2, 10^{-6})$	0.9226	0.0247
$(5, 0.2, 10^{-3})$	0.9271	0.0228
$(5, 0.4, 10^{-6})$	0.9410	0.0173
$(5, 0.4, 10^{-3})$	0.9417	0.0172



(a)

(b)

Figure 3: (a) Estimation error of $\boldsymbol{Z}^{\mathsf{cvx}}$ vs. $\boldsymbol{M}^{\mathsf{d}}$ measured in the Frobenius norm. (b) Estimation error of $\boldsymbol{Z}^{\mathsf{cvx}}$ vs. $\boldsymbol{M}^{\mathsf{d}}$ measured in the ℓ_∞ norm. The results are averaged over 20 independent trials for $r = 5$, $p = 0.2$ and $n = 1000$.

for $r = 5$, $p = 0.4$ and $\sigma = 10^{-3}$. It is evident that the distribution of S_{ij} matches that of $\mathcal{N}(0, 1)$ reasonably well.

In addition to the tractable distributional guarantees, the de-biased estimator $\boldsymbol{M}^{\mathsf{d}}$ also exhibits superior estimation accuracy compared to the original estimator $\boldsymbol{Z}^{\mathsf{cvx}}$ (cf. Theorem 3). Figure 3 reports the estimation error of $\boldsymbol{M}^{\mathsf{d}}$ vs. $\boldsymbol{Z}^{\mathsf{cvx}}$ measured in both the Frobenius norm and in the ℓ_∞ norm across difference noise levels. The results are averaged over 20 Monte Carlo simulations for $r = 5$, $p = 0.2$. It can be seen that the errors of the de-biased estimator are uniformly smaller than that of the original estimator and are much closer to the oracle lower bound. As a result, we recommend using $\boldsymbol{M}^{\mathsf{d}}$ even for the purpose of estimation.

We conclude this section with experiments on real data. Similar to [CP10], we use the daily temperature data [NCD19] for 1400 stations across the world in 2018, which results in a 1400 × 365 data matrix. Inspection on the singular values reveals that the data matrix is nearly low-rank. We vary the observation probability p from 0.5 to 0.9 and randomly subsample the data accordingly. Based on the observed temperatures, we then apply the proposed methodology to obtain 95% confidence intervals for all the entries. Table 4 reports the empirical coverage probabilities, the average length of the confidence intervals as well as the estimation error of both $\boldsymbol{Z}^{\mathsf{cvx}}$ and $\boldsymbol{M}^{\mathsf{d}}$ over 20 independent experiments. It can be seen that the average coverage probabilities are reasonably close to 95% and the confidence intervals are also quite short. In addition, the estimation error of $\boldsymbol{M}^{\mathsf{d}}$ is smaller than that of $\boldsymbol{Z}^{\mathsf{cvx}}$, which corroborates our theoretical prediction. The discrepancy between the nominal coverage probability and the actual one might arise from the facts that (1) the underlying true temperature matrix is only approximately low-rank, and (2) the noise in the temperature might not be independent.

3.7 A bit of intuition

We pause to develop some intuition behind the distributional guarantees for the proposed estimators. Bearing in mind the intimate link between convex and nonconvex optimization (cf. (3.5)), it suffices to concentrate

Table 4: Empirical coverage rates, average lengths of the confidence intervals of the entries as well as the estimation error vs. observation probability p. The results are averaged over 20 Monte Carlo trials.

p	Coverage		CI Length		$\\|\widehat{\boldsymbol{Z}} - \boldsymbol{M}^\star\\|_{\mathrm{F}} / \\|\boldsymbol{M}^\star\\|_{\mathrm{F}}$	
	Mean	Std	Mean	Std	Convex $\boldsymbol{Z}^{\mathsf{cvx}}$	Debiased $\boldsymbol{M}^{\mathsf{d}}$
0.5	0.8265	0.0016	3.6698	0.0209	0.029	0.028
0.6	0.8268	0.0011	2.8774	0.0098	0.025	0.023
0.7	0.8431	0.0006	2.3426	0.0054	0.022	0.019
0.8	0.8725	0.0003	2.0234	0.0052	0.020	0.015
0.9	0.9093	0.0003	1.8296	0.0072	0.018	0.011

on the nonconvex problem (3.4). For the sake of clarity, we further restrict attention to the rank-1 positive semidefinite case where $\boldsymbol{M}^\star = \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}$ and set $\lambda = 0$, where one can focus on

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\mathsf{minimize}} \qquad f\left(\boldsymbol{x}\right) \triangleq \frac{1}{2}\big\|\mathcal{P}_\Omega\big(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{M}\big)\big\|_\mathrm{F}^2. \tag{3.34}$$

Any optimizer $\widehat{\boldsymbol{x}}$ of (3.34) would necessarily satisfy the first-order optimality condition

$$\mathcal{P}_\Omega\big(\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}^\top - \boldsymbol{M}\big)\widehat{\boldsymbol{x}} = \boldsymbol{0}. \tag{3.35}$$

We shall also assume that $\widehat{\boldsymbol{x}}$ is a reasonably reliable estimate obeying $\widehat{\boldsymbol{x}} \approx \boldsymbol{x}^\star$.

We begin with the no-missing-data case (i.e. $p = 1$), which already conveys the key insight. The condition (3.35) simplifies to

$$\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}^\top\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\widehat{\boldsymbol{x}} = \boldsymbol{E}\widehat{\boldsymbol{x}}, \tag{3.36}$$

which, through a little manipulation, leads to an equivalent decomposition:

$$\|\widehat{\boldsymbol{x}}\|_2^2\left(\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\right) = \underbrace{\boldsymbol{E}\widehat{\boldsymbol{x}}}_{\text{approximately Gaussian}} + \underbrace{\boldsymbol{x}^\star\left(\boldsymbol{x}^\star - \widehat{\boldsymbol{x}}\right)^\top\boldsymbol{x}^\star}_{\text{negligible first-order term}} + \underbrace{\boldsymbol{x}^\star\left(\boldsymbol{x}^\star - \widehat{\boldsymbol{x}}\right)^\top\left(\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\right)}_{\text{second-order term}}. \tag{3.37}$$

Then: (1) the third term of (3.37), which can be viewed as a second-order term (in the sense that it is a quadratic term of $\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star$), becomes vanishingly small when $\widehat{\boldsymbol{x}} \approx \boldsymbol{x}^\star$; (2) while the second term of (3.37) looks like a first-order term, it is natural to conjecture that $\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star$ is sufficiently random and hence $(\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star)^\top\boldsymbol{x}^\star \ll \|\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\|_2\|\boldsymbol{x}^\star\|_2$ (i.e. the estimation error is not aligned with $\boldsymbol{x}^\star$), meaning that this term is also expected to be negligible compared to a typical first-order term (e.g. the term on the left-hand side of 3.37). In summary, these non-rigorous arguments suggest that

$$\|\widehat{\boldsymbol{x}}\|_2^2\left(\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\right) \approx \boldsymbol{E}\widehat{\boldsymbol{x}}. \tag{3.38}$$

If one can be convinced that $\boldsymbol{E}$ and $\widehat{\boldsymbol{x}}$ are only weakly dependent, then this means

$$\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star \approx \frac{1}{\|\widehat{\boldsymbol{x}}\|_2^2}\boldsymbol{E}\widehat{\boldsymbol{x}} \approx \frac{1}{\|\boldsymbol{x}^\star\|_2^2}\boldsymbol{E}\boldsymbol{x}^\star \sim \mathcal{N}\Big(\boldsymbol{0}, \frac{\sigma^2}{\|\boldsymbol{x}^\star\|_2^2}\boldsymbol{I}_n\Big). \tag{3.39}$$

Returning to the missing data scenario with $p < 1$, everything is based on the following approximation

$$\mathcal{P}_\Omega\big(\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}^\top - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\big)\widehat{\boldsymbol{x}} \approx p\big(\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}^\top - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\big)\widehat{\boldsymbol{x}};$$

this is certainly expected — using standard concentration arguments — if we "pretend" that $\mathcal{P}_\Omega$ and $\widehat{\boldsymbol{x}}$ are statistically independent. With this approximation in mind, one can translate (3.35) into

$$p\left(\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}^\top\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\widehat{\boldsymbol{x}}\right) \approx \mathcal{P}_\Omega\left(\boldsymbol{E}\right)\widehat{\boldsymbol{x}}. \tag{3.40}$$

Repeating the above argument then immediately yields

$$\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star \approx \frac{1}{\|\widehat{\boldsymbol{x}}\|_2^2}\cdot\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\widehat{\boldsymbol{x}} \approx \frac{1}{p\|\boldsymbol{x}^\star\|_2^2}\cdot\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{x}^\star \overset{\text{approx.}}{\sim} \mathcal{N}\Big(\boldsymbol{0}, \frac{\sigma^2}{p\|\boldsymbol{x}^\star\|_2^2}\boldsymbol{I}_n\Big). \tag{3.41}$$

The case with $\lambda > 0$ can be intuitively understood in a very similar way by first de-shrinking the estimate; we omit it here for brevity. We note that these hand-waving arguments can all be made rigorous, which is the main content of the proof.

3.8 Inference based on spectral estimates?

One would naturally be curious about whether there are other estimation procedures that also enable reasonable statistical inference. While this is beyond the scope of the current paper, we take a moment to discuss one alternative: the spectral method, as pioneered by [KMO10a, KMO10b] in the matrix completion problem. In a nutshell, this approach consists in computing a rank-r approximation to $\mathcal{P}_\Omega(\boldsymbol{M})/p$, which is precisely the spectral initialization widely used in a two-stage nonconvex algorithm (cf. Algorithm 1) [KMO10b, SL16, CW15, CCF18, MWCC17]. While inference has not been, as far as we know, the focus of prior work on spectral methods,[4] the recent papers [AFWZ17, MWCC17] hinted at the possibility of characterizing the distribution of the spectral estimate. Take a simple symmetric rank-1 case for example (i.e. $\boldsymbol{M}^\star = \boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$ with $\|\boldsymbol{x}^\star\|_2 = 1$): the leading eigenvector $\boldsymbol{u}^{\mathsf{spectral}}$ of $\mathcal{P}_\Omega(\boldsymbol{M})/p$ often admits the following approximation (up to a global sign)

$$\boldsymbol{u}^{\mathsf{spectral}} \approx \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M})\boldsymbol{x}^\star.$$

Expanding $\mathcal{P}_\Omega(\boldsymbol{M}) = p\boldsymbol{x}^\star\boldsymbol{x}^{\star\top} + \mathcal{P}_\Omega(\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}) - p\boldsymbol{x}^\star\boldsymbol{x}^{\star\top} + \mathcal{P}_\Omega(\boldsymbol{E})$, we arrive at

$$\boldsymbol{u}^{\mathsf{spectral}} \approx \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\boldsymbol{x}^\star + \Big(\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}) - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\Big)\boldsymbol{x}^\star + \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{x}^\star,$$

which is equivalent to

$$\boldsymbol{u}^{\mathsf{spectral}} - \boldsymbol{x}^\star \approx \underbrace{\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{x}^\star}_{\text{noise effect}} + \underbrace{\Big(\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}) - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\Big)\boldsymbol{x}^\star}_{\text{effect of random sub-sampling}}. \tag{3.42}$$

In words, two major factors dictate the uncertainty of the spectral estimate: (1) the additive Gaussian noise (cf. the 1st term on the right-hand side of (3.42)), and (2) random sub-sampling (in particular, the randomness incurred by employing the sub-sampled $\mathcal{P}_\Omega(\boldsymbol{x}^\star\boldsymbol{x}^{\star\top})/p$ to approximate the truth $\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}$). Given that the random sub-sampling effect cannot be ignored at all, the spectral estimates often suffer from a much larger estimation error (and hence a higher degree of uncertainty) compared to either the convex or the nonconvex estimates. In truth, this random sub-sampling effect does not go away even when the noise vanishes. Consequently, uncertainty quantification based on the spectral estimates may not be the most desirable option.

4 Prior art

Matrix completion. Low-rank matrix completion, or more broadly, low-rank matrix recovery, is a fundamental task that permeates through a wide spectrum of applications in science, engineering, and finance (e.g. [RS05, SY07, CC14, FSZZ18, CCG15, ZPL15, BN06, CC18b, FWZ19, KS11, CZ16, KX15, FLM13, DR17, CDDD19, DPVW14, SZ12, FS11]). A paper of this length is unable to review all papers motivating and contributing to this enormous subject; interested readers are referred to [DR16, CC18a] for extensive discussions of motivating applications as well as the exciting recent development.

Numerous algorithms have been proposed to solve this problem efficiently, with two paradigms being arguably the most widely used: convex relaxation and nonconvex optimization. We briefly review the literature contributing to these two paradigms.

- Convex relaxation was largely popularized by the seminal works [Faz02, RFP10, CR09]. In the absence of noise, it has been shown that nuclear norm minimization, which can be solved by semidefinite programming, achieves minimal sample complexity under mild conditions [Gro11, Rec11, Che15]. When the observed entries are further corrupted by noise, Candès and Plan [CP10] provided the first theoretical guarantee regarding the estimation accuracy of perhaps the most natural convex relaxation algorithm. While the theory might be tight for certain adversarial scenarios (as shown by the recent work [KS19]), it is loose by some large factor under the natural random noise model. This statistical guarantee has been partially improved later on by two papers [NW12, KLT11] under proper modifications to the convex program (e.g. enforcing an additional spikiness constraint [NW12, Klo14], or modifying the squared loss [KLT11]). Nevertheless, the error bounds provided in these papers (and their follow-ups) remain suboptimal, unless the typical size of the noise is sufficiently large. Our recent work [CCF+19] establishes near-optimal statistical guarantees — when the estimation errors are measured by the Frobenius

[4] We note that inference from spectral estimates has been investigated in other context beyond matrix completion (e.g. the model without missing data [Xia19, FFHL19]).

norm, the spectral norm, and the $\ell_{2,\infty}$ norm — for a wide range of noise levels when $r = O(1)$. All of these estimation guarantees, however, come with a hidden and likely large pre-constant, which do not serve the inferential purpose well.

- Nonconvex optimization algorithms, as pioneered by [KMO10a, Sre04], become increasingly more popular for solving various low-rank factorization problems, due to their appealing computational complexities [JNS13, CLS15, CC17, TBS+16, SL16, ZL16, CCFM19, WZG16, CLL19]. For instance, the gradient-based nonconvex methods have been analyzed for noisy matrix completion [KMO10b, CW15, MWCC17, CCF+19], which are shown to achieve near-optimal statistical accuracy and linear-time convergence guarantees all at once. Going beyond gradient methods, we note that other nonconvex methods (e.g. [RS05, JMD10, WYZ12, JNS13, FRW11, Van13, LXY13, Har14, JKN16, RT11, WCCL16, DC18, ZWL15, ZWYG18, MSL19, CCD+19]) and landscape properties [GLM16, CL17, GJZ17, ZSL19, ZJSL18, SXZ19] have been largely explored as well. The interested readers are referred to [CLC19] for an in-depth discussion. One limitation, however, is that the theoretical guarantees provided for nonconvex algorithms often exhibit sub-optimal dependency in the rank r of the unknown matrix; for instance, most theory requires a sample complexity of at least nr^2 (in fact, often much larger than nr^2). This is outperformed by the convex relaxation approach.

Despite these recent developments, very little work has investigated statistical inference for noisy matrix completion. While [CKLN18, CKL16, CN15, CEGN15] discussed the construction of "honest" confidence regions, the volume of these regions is dependent on some (possibly huge) hidden constants, thus resulting in over-coverage. Perhaps the closest to our paper is the recent work [Xia18], which investigated inference for low-rank trace regression. Employing a closely related de-biased estimator with sample splitting, the paper [Xia18] established asymptotic normality of a certain projected distance between the estimate and the truth. The result therein, however, requires a sampling mechanism obeying the restricted isometry property (e.g. i.i.d. Gaussian designs), which fails to hold for matrix completion. Also, our approach does not require sample splitting — a technique that is convenient for analysis but conservative in constructing confidence regions. Another work by Cai et al. [CLR16] developed a unified approach to provide inference guarantees for linear inverse problems including low-rank matrix estimation. Their results, however, require the sample size to exceed the total dimension n^2 even under the Gaussian design. Finally, a recent line of work [MX17] explored uncertainty quantification under the Bayesian setting, hypothesizing on a special prior regarding the true matrix. This departs drastically from the scenario considered herein.

Inference in high-dimensional problems. Inference in high-dimensional sparse regression has received much attention in the last few years [WR09, ZZ14, BCH11, vdGBRD14, JM14b, DBMM15, CG17, NL17, NNLL18, LSST16, LTTT14, MMB09, DBZ17, ZC17, BFL+18]. Our inferential approach is partly inspired by the recent developments on this topic, particularly with regard to the de-biased / de-sparsified estimators proposed for Lasso. More specifically, recognizing the non-negligible bias of the Lasso estimate

$$\widehat{\boldsymbol{\beta}} \triangleq \arg\min_{\boldsymbol{\beta}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{4.1}$$

A line of work [ZZ14, vdGBRD14, JM14a] came up with a linear transformation of $\widehat{\boldsymbol{\beta}}$ of the form

$$\boldsymbol{\beta}^{\mathsf{d}} \triangleq \widehat{\boldsymbol{\beta}} + \boldsymbol{L}\boldsymbol{X}^{\top}\big(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\big), \tag{4.2}$$

where $\boldsymbol{L}$ is some matrix to be designed, and $\boldsymbol{X}^{\top}\big(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\big)$ corresponds to the negative gradient of the squared loss at $\widehat{\boldsymbol{\beta}}$, or equivalently, the (scaled) sub-gradient of the ℓ_1 norm at $\widehat{\boldsymbol{\beta}}$. If $\boldsymbol{L}$ is properly chosen, then $\boldsymbol{\beta}^{\mathsf{d}}$ is able to correct the bias of this nonlinear estimator $\boldsymbol{\beta}$, while controlling the degree of uncertainty. Many follow-up papers have investigated the design of $\boldsymbol{L}$ as well as the resulting inferential guarantees [ZZ14, vdGBRD14, JM14a, JM15].

Interestingly, our de-biased estimator (3.7) for matrix completion admits a very similar form as (4.2). To see this, recall that our de-biased estimator is given by

$$\boldsymbol{M}^{\mathsf{d}} = \mathcal{P}_{\text{rank-}r}\big(\boldsymbol{Z} - \tfrac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{Z}) + \tfrac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M})\big),$$

where $\boldsymbol{Z}$ can be either $\boldsymbol{Z}^{\mathsf{cvx}}$ or $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$ (see Table 1). Let T be the tangent space of the set of rank-r matrices at $\boldsymbol{Z}^{\mathsf{cvx},r}$ (resp. $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$) in the convex (resp. nonconvex) case, and $\mathcal{P}_T$ be the projection operator onto T. Somewhat surprisingly, replacing $\mathcal{P}_{\text{rank-}r}$ by $\mathcal{P}_T$ does not affect the de-biased estimator by much, in the sense that

$$\boldsymbol{M}^{\mathsf{d}} \approx \mathcal{P}_T\big(\boldsymbol{Z} - \tfrac{1}{p}\mathcal{P}_\Omega(\boldsymbol{Z}) + \tfrac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M})\big). \tag{4.3}$$

In addition, recognizing that $\boldsymbol{Z}$ almost lies within the tangent space T,[5]one can rewrite

$$\boldsymbol{M}^{\mathsf{d}} \approx \boldsymbol{Z} - \tfrac{1}{p}\mathcal{P}_T\mathcal{P}_\Omega(\boldsymbol{Z} - \boldsymbol{M}), \tag{4.4}$$

a fact to be made precise in Section 5.1. This bears a striking resemblance to the de-biasing approach developed for Lasso — the term $\mathcal{P}_\Omega(\boldsymbol{Z}-\boldsymbol{M}^\star)$ represents the gradient of the squared loss $0.5\|\mathcal{P}_\Omega(\boldsymbol{Z}-\boldsymbol{M})\|_{\mathrm{F}}^2$ (or equivalently, the negative sub-gradient of the nuclear norm) at $\boldsymbol{Z}$, and $\mathcal{P}_T$ is the linear operator we pick. To the best of our knowledge, no de-biasing approach — with rigorous theoretical guarantees and without sample splitting — has been proposed and analyzed for matrix completion in prior literature. In addition, we note that our de-biased estimator for matrix completion achieves full statistical efficiency in terms of both the rates and the pre-constant; in comparison, the commonly used de-biased estimators for sparse linear regression typically fall short of achieving the best possible estimation accuracy, unless additional thresholding procedures are enforced.

Finally, de-biased estimators have been put forward to tackle other high-dimensional problems, including but not limited to generalized linear models [vdGBRD14,NL17], graphical models [JVDG15,RSZZ15,MLL17, JvdG17], sparse PCA [JvdG18], treatment effects estimation [CCD+18,AIW18]. These are beyond the scope of the current paper.

5 Architecture of the proof

This section outlines the main steps for establishing Theorem 1 and Theorem 2. Before starting, we introduce some useful notation. For convenience of presentation, we insert the factor $1/p$ into (3.4) and redefine the nonconvex loss function as

$$f(\boldsymbol{X},\boldsymbol{Y}) \triangleq \frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{Y}\|_{\mathrm{F}}^2. \tag{5.1}$$

In addition, for each $1 \le j,k \le n$, we define the indicator $\delta_{jk} \triangleq \mathbb{1}\{(j,k)\in\Omega\}$, which is a Bernoulli random variable with mean p.

We also note that Theorem 1 (resp. Theorem 2) is subsumed by Theorem 5 (resp. Theorem 6). As a result, we shall focus on establishing Theorem 5 (resp. Theorem 6) when it comes to estimating low-rank factors (resp. the entries of the matrix).

Theorem 5. *Suppose that the sample complexity meets $n^2p \ge C\kappa^4\mu^2r^2n\log^3 n$ for some sufficiently large constant $C>0$ and the noise obeys $\sigma\sqrt{(\kappa^4\mu rn\log n)/p} \le c\sigma_{\min}$ for some sufficiently small constant $c>0$. Then the decomposition in Theorem 1 remains valid, except that the residual matrices $\boldsymbol{\Psi}_{\boldsymbol{X}}, \boldsymbol{\Psi}_{\boldsymbol{Y}} \in \mathbb{R}^{n\times r}$ satisfy, with probability at least $1-O(n^{-3})$, that*

$$\max\left\{\|\boldsymbol{\Psi}_{\boldsymbol{X}}\|_{2,\infty}, \|\boldsymbol{\Psi}_{\boldsymbol{Y}}\|_{2,\infty}\right\} \lesssim \frac{\sigma}{\sqrt{p\sigma_{\min}}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^7\mu rn\log n}{p}} + \sqrt{\frac{\kappa^7\mu^3r^3\log^2 n}{np}}\right). \tag{5.2}$$

Theorem 6. *Instate the assumptions of Theorem 5. Recall the definition of $v_{ij}^\star$ in (3.17). Then one has the following decomposition*

$$M_{ij}^{\mathsf{d}} - M_{ij}^\star = g_{ij} + \Delta_{ij}, \tag{5.3}$$

where $g_{ij} \sim \mathcal{N}(0, v_{ij}^\star)$ and the residual obeys — with probability exceeding $1-O(n^{-10})$ — that

$$|\Delta_{ij}| \lesssim \left(\left\|\boldsymbol{U}_{i,\cdot}^\star\right\|_2 + \left\|\boldsymbol{V}_{j,\cdot}^\star\right\|_2\right)\frac{\sigma}{\sqrt{p}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^8\mu rn\log n}{p}} + \sqrt{\frac{\kappa^8\mu^3r^3\log^2 n}{np}}\right) + \left(\frac{\sigma}{\sqrt{\sigma_{\min}}}\sqrt{\frac{\kappa^3\mu r\log n}{p}}\right)^2.$$

[5] More precisely, if $\boldsymbol{Z} = \boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$, then $\boldsymbol{Z}\in T$; if $\boldsymbol{Z} = \boldsymbol{Z}^{\mathsf{cvx}}$, one has $\mathcal{P}_T(\boldsymbol{Z}) \approx \boldsymbol{Z}$.

5.1 Near equivalence between convex and nonconvex estimators

Note that Theorem 5 and Theorem 6 are concerned with the de-biased estimators built upon both convex and nonconvex estimates. At first glance, one needs to establish theoretical guarantees for each of them separately. Fortunately, as alluded to previously (cf. (3.5)), the convex and nonconvex estimates are extremely close — a fact that has been established in [CCF+19]. The proximity of these two estimates naturally extends to the de-biased estimators constructed based on them. As a result, it suffices to concentrate on proving the theorems for any of these estimators; the claims for the other one follow immediately.

The following key lemma formalizes this argument, which will be established in Appendix C (see also Figure 4 for numerical evidence). Before continuing, we remind the readers of the key notation (see Appendix B for precise definitions):

- $(\boldsymbol{X}^{\mathsf{ncvx}}, \boldsymbol{Y}^{\mathsf{ncvx}})$: an approximate solution to the nonconvex problem (3.4) (see Appendix A.1);
- $\boldsymbol{M}^{\mathsf{cvx,d}}, \boldsymbol{X}^{\mathsf{cvx,d}}, \boldsymbol{Y}^{\mathsf{cvx,d}}$: the de-biased estimators built upon the convex optimizer $\boldsymbol{Z}^{\mathsf{cvx}}$;
- $\boldsymbol{M}^{\mathsf{ncvx,d}}, \boldsymbol{X}^{\mathsf{ncvx,d}}, \boldsymbol{Y}^{\mathsf{ncvx,d}}$: the de-biased estimators built upon the nonconvex estimate $(\boldsymbol{X}^{\mathsf{ncvx}}, \boldsymbol{Y}^{\mathsf{ncvx}})$.

Our proximity result is this:

Lemma 3. *Suppose that the sample size obeys $n^2 p \geq C\kappa^4\mu^2 r^2 n \log^3 n$ for some sufficiently large constant $C > 0$ and the noise satisfies $\sigma\sqrt{(\kappa^4 \mu n r \log n)/p} \leq c\sigma_{\min}$ for some sufficiently small constant $c > 0$. Set $\lambda = C_\lambda \sigma \sqrt{np}$ with some large enough constant $C_\lambda > 0$.*

1. *With probability at least $1 - O(n^{-10})$, one has*

$$\max\left\{\left\|\boldsymbol{M}^{\mathsf{cvx,d}} - \boldsymbol{X}^{\mathsf{ncvx,d}}\boldsymbol{Y}^{\mathsf{ncvx,d}\top}\right\|_{\mathrm{F}}, \left\|\boldsymbol{M}^{\mathsf{ncvx,d}} - \boldsymbol{X}^{\mathsf{ncvx,d}}\boldsymbol{Y}^{\mathsf{ncvx,d}\top}\right\|_{\mathrm{F}}\right\} \lesssim \frac{1}{n^4} \cdot \sigma\sqrt{\frac{n}{p}}, \tag{5.4a}$$

$$\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \sqrt{\|\boldsymbol{X}^{\mathsf{cvx,d}}\boldsymbol{R} - \boldsymbol{X}^{\mathsf{ncvx,d}}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{\mathsf{cvx,d}}\boldsymbol{R} - \boldsymbol{Y}^{\mathsf{ncvx,d}}\|_{\mathrm{F}}^2} \lesssim \frac{1}{n^3} \cdot \frac{\sigma}{\sqrt{\sigma_{\min}}}\sqrt{\frac{n}{p}}, \tag{5.4b}$$

   where $\mathcal{O}^{r\times r}$ is the set of $r \times r$ rotation matrices.

2. *With probability exceeding $1 - O(n^{-10})$, one has*

$$\left\|\boldsymbol{M}^{\mathsf{ncvx,d}} - \left[\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} - p^{-1}\mathcal{P}_T\mathcal{P}_\Omega\left(\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} - \boldsymbol{M}\right)\right]\right\|_{\mathrm{F}} \lesssim \frac{1}{n^4} \cdot \sigma\sqrt{\frac{n}{p}}, \tag{5.5}$$

   where T is the tangent space of the set of rank-r matrices at $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$. The same holds true if we replace $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$ with $\boldsymbol{Z}^{\mathsf{cvx}}$ and replace T with the tangent space at $\boldsymbol{Z}^{\mathsf{cvx},r} = \mathcal{P}_{\text{rank-}r}(\boldsymbol{Z}^{\mathsf{cvx}})$.

In short, the first part of Lemma 3 tells us that

$$\boldsymbol{M}^{\mathsf{cvx,d}} \approx \boldsymbol{X}^{\mathsf{ncvx,d}}\boldsymbol{Y}^{\mathsf{ncvx,d}\top} \approx \boldsymbol{M}^{\mathsf{ncvx,d}}, \tag{5.6a}$$

$$\left(\boldsymbol{X}^{\mathsf{cvx,d}}, \boldsymbol{Y}^{\mathsf{cvx,d}}\right) \approx \left(\boldsymbol{X}^{\mathsf{ncvx,d}}, \boldsymbol{Y}^{\mathsf{ncvx,d}}\right) \qquad \text{(up to global rotation)}, \tag{5.6b}$$

whereas the second part of Lemma 3 justifies that the proposed de-biased estimator is closely approximated by a linearized version (cf. (4.4)). Note that this linearized form bears a resemblance to the de-biased estimators developed for sparse linear regression [ZZ14, vdGBRD14, JM14a].

With Lemma 3 in place, we shall, from now on, focus on proving the main theorems for the nonconvex estimators, viz.

1. establishing Theorem 5 for the de-shrunken low-rank factors $(\boldsymbol{X}^{\mathsf{ncvx,d}}, \boldsymbol{Y}^{\mathsf{ncvx,d}})$;
2. establishing Theorem 6 for the de-biased matrix estimator $\boldsymbol{X}^{\mathsf{ncvx,d}}\boldsymbol{Y}^{\mathsf{ncvx,d}\top}$.

To simplify the presentation hereafter, we shall use the following notation throughout the rest of this section:

- $(\boldsymbol{X}, \boldsymbol{Y})$: the nonconvex estimate $(\boldsymbol{X}^{\mathsf{ncvx}}, \boldsymbol{Y}^{\mathsf{ncvx}})$;
- $(\boldsymbol{X}^{\mathsf{d}}, \boldsymbol{Y}^{\mathsf{d}})$: the de-shrunken estimate defined in (3.8) based on $(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{X}^{\mathsf{ncvx}}, \boldsymbol{Y}^{\mathsf{ncvx}})$;
- $\boldsymbol{M}^{\mathsf{d}} \triangleq \boldsymbol{X}^{\mathsf{d}}\boldsymbol{Y}^{\mathsf{d}\top}$.

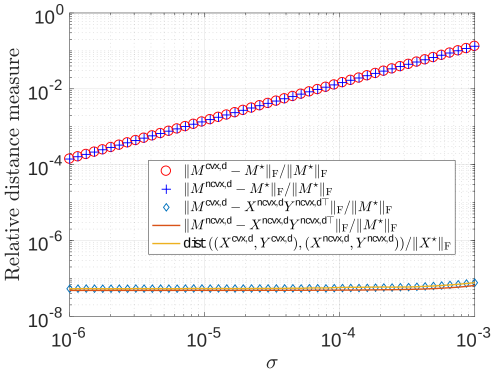


Figure 4: The relative estimation errors of $\boldsymbol{M}^{\mathsf{cvx,d}}$ and $\boldsymbol{M}^{\mathsf{ncvx,d}}$ and related quantities in Lemma 3 vs. the standard deviation σ of the noise. Here, $\mathsf{dist}((\boldsymbol{X}^{\mathsf{cvx,d}}, \boldsymbol{Y}^{\mathsf{cvx,d}}), (\boldsymbol{X}^{\mathsf{ncvx,d}}, \boldsymbol{Y}^{\mathsf{ncvx,d}}))$ is defined to be the left-hand side of (5.4b). The results, which are averaged over 20 trials, are reported for $n = 1000$, $r = 5$, $p = 0.2$, and $\lambda = 5\sigma\sqrt{np}$. As can be seen, the difference between $\boldsymbol{M}^{\mathsf{cvx,d}}$, $\boldsymbol{M}^{\mathsf{ncvx,d}}$ and $\boldsymbol{X}^{\mathsf{ncvx,d}}\boldsymbol{Y}^{\mathsf{ncvx,d}\top}$, as well as the distance $\mathsf{dist}((\boldsymbol{X}^{\mathsf{cvx,d}}, \boldsymbol{Y}^{\mathsf{cvx,d}}), (\boldsymbol{X}^{\mathsf{ncvx,d}}, \boldsymbol{Y}^{\mathsf{ncvx,d}}))$, are all significantly smaller than the estimation errors.

5.2 A precise characterization of the de-shrunken low-rank factors

We start with a precise characterization of the de-shrunken low-rank factors $\boldsymbol{X}^{\mathsf{d}}$ and $\boldsymbol{Y}^{\mathsf{d}}$, which paves the way for demonstrating both Theorem 5 and Theorem 6.

Lemma 4 (Decompositions of low-rank factors). *Denote*

$$\boldsymbol{A} \triangleq \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}\right) - \left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}\right). \tag{5.7}$$

One has the following decompositions for $\boldsymbol{X}^{\mathsf{d}}$ and $\boldsymbol{Y}^{\mathsf{d}}$

$$\begin{aligned}\boldsymbol{X}^{\mathsf{d}} &= \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}^{\mathsf{d}}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)^{-1} + \boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\mathsf{d}}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)^{-1} - \boldsymbol{A}\boldsymbol{Y}^{\mathsf{d}}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)^{-1} \\ &\quad + \nabla_{\boldsymbol{X}} f\left(\boldsymbol{X}, \boldsymbol{Y}\right)\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)^{-1} + \boldsymbol{X}\boldsymbol{\Delta}_{\mathsf{balancing}};\end{aligned} \tag{5.8a}$$

$$\begin{aligned}\boldsymbol{Y}^{\mathsf{d}} &= \frac{1}{p}\left[\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right]^\top\boldsymbol{X}^{\mathsf{d}}\left(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}}\right)^{-1} + \boldsymbol{Y}^\star\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\mathsf{d}}\left(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}}\right)^{-1} - \boldsymbol{A}^\top\boldsymbol{X}^{\mathsf{d}}\left(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}}\right)^{-1} \\ &\quad + \nabla_{\boldsymbol{Y}} f\left(\boldsymbol{X}, \boldsymbol{Y}\right)\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right)^{1/2}\left(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}}\right)^{-1} - \boldsymbol{Y}\boldsymbol{\Delta}_{\mathsf{balancing}}.\end{aligned} \tag{5.8b}$$

Here, we denote

$$\boldsymbol{\Delta}_{\mathsf{balancing}} \triangleq \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right)^{1/2} - \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2}, \tag{5.9}$$

which measures the imbalance between the low-rank factors $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Proof. The claims follow from straightforward algebraic manipulations; see Appendix D.1. □

We make a few observations regarding Lemma 4. Take the decomposition of $\boldsymbol{X}^{\mathsf{d}}$ (5.8a) as an example:

- First, the term $\boldsymbol{A}\boldsymbol{Y}^{\mathsf{d}}(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}})^{-1}$ vanishes when we have full observations, i.e. $p = 1$. Second, the terms involving $\nabla_{\boldsymbol{X}} f(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{\Delta}_{\mathsf{balancing}}$ are both zero if $(\boldsymbol{X}, \boldsymbol{Y})$ is an exact stationary point of $f(\cdot, \cdot)$; to see this, it is not hard to verify that any stationary point of $f(\cdot, \cdot)$ necessarily satisfies $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{Y}^\top\boldsymbol{Y}$, which in turn implies $\boldsymbol{\Delta}_{\mathsf{balancing}} = \boldsymbol{0}$. Consequently, the last three terms in (5.8a) are expected to be small when p is sufficiently large and $(\boldsymbol{X}, \boldsymbol{Y})$ is near a stationary point.

- Turning to the first two terms in (5.8a), we note that the second term of (5.8a) is close to $\boldsymbol{X}^\star$ (up to rotation) if $\boldsymbol{Y}^{\mathsf{d}}$ is a nearly accurate approximation to $\boldsymbol{Y}^\star$. In comparison, the first term $\mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{Y}^{\mathsf{d}}(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}})^{-1}/p$ has to do with a collection of Gaussian random variables, which accounts for the main uncertainty term.

We shall make precise these arguments in subsequent subsections.

5.3 Taking global rotation into account

In order to invoke the decompositions of $\boldsymbol{X}^{\mathsf{d}}$ and $\boldsymbol{Y}^{\mathsf{d}}$ (cf. Lemma 4) to characterize the estimation errors, we still need to incorporate the (unrecoverable) rotation matrix. From now on, we shall focus primarily on the factor $\boldsymbol{X}^{\mathsf{d}}$. The claims on the other factor $\boldsymbol{Y}^{\mathsf{d}}$ can be easily obtained via symmetry.

Denote

$$\overline{\boldsymbol{X}}^{\mathsf{d}} \triangleq \boldsymbol{X}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} \qquad \text{and} \qquad \overline{\boldsymbol{Y}}^{\mathsf{d}} \triangleq \boldsymbol{Y}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}}, \tag{5.10}$$

where we recall that $\boldsymbol{H}^{\mathsf{d}}$ is the rotation matrix that best aligns $(\boldsymbol{X}^{\mathsf{d}}, \boldsymbol{Y}^{\mathsf{d}})$ and $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$ (see (3.12)). Substituting the identity

$$\boldsymbol{Y}^{\mathsf{d}}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)^{-1}\boldsymbol{H}^{\mathsf{d}} = \boldsymbol{Y}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}}\left(\boldsymbol{H}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}}\right)^{-1} = \overline{\boldsymbol{Y}}^{\mathsf{d}}\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1}$$

into the decomposition (5.8a), we arrive at

$$\begin{aligned}
\boldsymbol{X}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{X}^\star &= \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\overline{\boldsymbol{Y}}^{\mathsf{d}}\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1} + \boldsymbol{X}^\star\left[\boldsymbol{Y}^{\star\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1} - \boldsymbol{I}_r\right] - \boldsymbol{A}\overline{\boldsymbol{Y}}^{\mathsf{d}}\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1} \\
&\quad + \nabla_{\boldsymbol{X}} f\left(\boldsymbol{X}, \boldsymbol{Y}\right)\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)^{-1}\boldsymbol{H}^{\mathsf{d}} + \boldsymbol{X}\boldsymbol{\Delta}_{\mathsf{balancing}}\boldsymbol{H}^{\mathsf{d}} \\
&= \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}^\star\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1} + \boldsymbol{\Phi}_{\boldsymbol{X}}.
\end{aligned} \tag{5.11}$$

Here, the term $\boldsymbol{\Phi}_{\boldsymbol{X}} \in \mathbb{R}^{n\times r}$ is defined to be

$$\begin{aligned}
\boldsymbol{\Phi}_{\boldsymbol{X}} \triangleq\ & \underbrace{\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\left[\overline{\boldsymbol{Y}}^{\mathsf{d}}\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1} - \boldsymbol{Y}^\star\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1}\right]}_{:=\boldsymbol{\Phi}_1} + \underbrace{\boldsymbol{X}^\star\left[\boldsymbol{Y}^{\star\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1} - \boldsymbol{I}_r\right]}_{:=\boldsymbol{\Phi}_2} \\
& \underbrace{-\boldsymbol{A}\overline{\boldsymbol{Y}}^{\mathsf{d}}\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1}}_{:=\boldsymbol{\Phi}_3} + \underbrace{\nabla_{\boldsymbol{X}} f\left(\boldsymbol{X}, \boldsymbol{Y}\right)\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)^{-1}\boldsymbol{H}^{\mathsf{d}} + \boldsymbol{X}\boldsymbol{\Delta}_{\mathsf{balancing}}\boldsymbol{H}^{\mathsf{d}}}_{:=\boldsymbol{\Phi}_4},
\end{aligned} \tag{5.12}$$

where $\boldsymbol{A}$ is defined in (5.7). To establish Theorem 5, it remains to (1) demonstrate that $\boldsymbol{\Phi}_{\boldsymbol{X}}$ has small $\ell_{2,\infty}$ norm, and (2) show that $\mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{Y}^\star(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star)^{-1}/p$ is approximately a Gaussian random matrix. These two steps constitute the main content of the next subsection.

5.4 Key lemmas for establishing Theorem 5

We now state five key lemmas. Taking these collectively and substituting them into (5.11) immediately establish Theorem 5.

We shall start by controlling the term $\boldsymbol{\Phi}_1$ as defined in (5.12).

Lemma 5 (Negligibility of $\boldsymbol{\Phi}_1$)**.** *Suppose that the sample complexity obeys $n^2p \geq C\kappa^4\mu^2r^2n\log^3 n$ for some sufficiently large constant $C > 0$ and the noise satisfies $\sigma\sqrt{(\kappa^4\mu rn\log n)/p} \leq c\sigma_{\min}$ for some sufficiently small constant $c < 0$. Then with probability at least $1 - O(n^{-10})$, we have*

$$\left\|\boldsymbol{\Phi}_1\right\|_{2,\infty} \lesssim \frac{\sigma}{\sqrt{p}\sigma_{\min}} \cdot \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^3\mu rn\log n}{p}}.$$

Proof. Fix any $1 \leq j \leq n$. If the de-shrunken estimate $\overline{\boldsymbol{Y}}^{\mathsf{d}}$ were independent of the randomness in the jth row of the matrix, i.e. $\boldsymbol{e}_j^\top \mathcal{P}_\Omega(\boldsymbol{E})$, then $\|\boldsymbol{e}_j^\top \boldsymbol{\Phi}_1\|_2$ would be well controlled. This hypothesis is certainly false, as $\overline{\boldsymbol{Y}}^{\mathsf{d}}$ clearly depends on $\boldsymbol{e}_j^\top \mathcal{P}_\Omega(\boldsymbol{E})$. Nevertheless, by exploiting the leave-one-out technique recently used in [EKBB+13, EK15, AFWZ17, MWCC17, CFMW19, CCF+19, CLL19, DC18], one can properly decouple the dependency and establish the desired bound. See Appendix D.2. □

The next lemma controls the size of $\|\boldsymbol{\Phi}_2\|_{2,\infty}$. In essence, the term $\boldsymbol{\Phi}_2$ measures the difference between the estimate $\overline{\boldsymbol{Y}}^{\mathsf{d}}$ and the true signal $\boldsymbol{Y}^\star$; the closer these two are, the smaller $\|\boldsymbol{\Phi}_2\|_{2,\infty}$ should be. See Appendix D.3 for the proof of the following result.

Lemma 6 (Negligibility of $\boldsymbol{\Phi}_2$)**.** *Suppose that the sample complexity obeys $n^2 p \geq C\kappa^4\mu^2 r^2 n \log^3 n$ for some sufficiently large constant $C > 0$ and the noise satisfies $\sigma\sqrt{(\kappa^4 \mu r n \log n)/p} \leq c\sigma_{\min}$ for some sufficiently small constant $c < 0$. Then with probability exceeding $1 - O(n^{-10})$, one has*

$$\|\boldsymbol{\Phi}_2\|_{2,\infty} \lesssim \frac{\sigma}{\sqrt{p\sigma_{\min}}}\left(\kappa \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^7 \mu r n}{p}} + \sqrt{\frac{\kappa^7\mu^3 r^3 \log n}{np}}\right).$$

Moving on to $\boldsymbol{\Phi}_3$ and $\boldsymbol{\Phi}_4$, one has the following lemmas.

Lemma 7 (Negligibility of $\boldsymbol{\Phi}_3$)**.** *Suppose that the sample complexity obeys $n^2 p \geq C\kappa^4\mu^2 r^2 n \log^3 n$ for some sufficiently large constant $C > 0$ and the noise satisfies $\sigma\sqrt{(\kappa^4 \mu r n \log n)/p} \leq c\sigma_{\min}$ for some sufficiently small constant $c < 0$. Then with probability exceeding $1 - O(n^{-10})$, we have*

$$\|\boldsymbol{\Phi}_3\|_{2,\infty} \lesssim \frac{\sigma}{\sqrt{p\sigma_{\min}}}\sqrt{\frac{\kappa^5\mu^3 r^3 \log^2 n}{np}}.$$

Proof. It is straightforward to check that when $p = 1$, one has $\|\boldsymbol{\Phi}_3\|_{2,\infty} = \|\boldsymbol{A}\| = 0$, where we recall the definition of $\boldsymbol{A}$ in (5.7). Therefore, one expects $\|\boldsymbol{\Phi}_3\|_{2,\infty}$ to be small when p is sufficiently large. See Appendix D.4. □

Lemma 8 (Negligibility of $\boldsymbol{\Phi}_4$)**.** *Suppose that the sample complexity obeys $n^2 p \geq C\kappa^4\mu^2 r^2 n \log^3 n$ for some sufficiently large constant $C > 0$ and the noise satisfies $\sigma\sqrt{(\kappa^4 \mu r n \log n)/p} \leq c\sigma_{\min}$ for some sufficiently small constant $c < 0$. Then with probability at least $1 - O(n^{-10})$, one has*

$$\|\boldsymbol{\Phi}_4\|_{2,\infty} \lesssim \frac{\sigma}{\sqrt{p\sigma_{\min}}} \cdot \frac{1}{n^4}.$$

Proof. It is easily seen that the size of $\boldsymbol{\Phi}_4$ depends on how close $(\boldsymbol{X}, \boldsymbol{Y})$ is to a stationary point of $f(\cdot,\cdot)$. For instance, in the extreme case where $(\boldsymbol{X}, \boldsymbol{Y})$ is an exact stationary point, then one would have $\boldsymbol{\Phi}_4 = \boldsymbol{0}$. See Appendix D.5. □

The last lemma asserts that $\mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{Y}^\star(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star)^{-1}/p$ is, in some sense, close to a zero-mean Gaussian random matrix with the desired covariance.

Lemma 9 (Approximate Gaussianity of $\mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{Y}^\star(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star)^{-1}/p$)**.** *Suppose that the sample size obeys $n^2 p \geq C\kappa^2 \mu r n \log^3 n$ for some sufficiently large constant $C > 0$. Then one has the decomposition*

$$\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}^\star\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1} = \boldsymbol{Z}_{\boldsymbol{X}} + \boldsymbol{\Delta}_{\boldsymbol{X}},$$

where each row of $\boldsymbol{Z}_{\boldsymbol{X}} \in \mathbb{R}^{n\times r}$ is independent and identically distributed according to

$$\boldsymbol{Z}_{\boldsymbol{X}}^\top \boldsymbol{e}_j \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \frac{\sigma^2}{p}\left(\boldsymbol{\Sigma}^\star\right)^{-1}\right), \qquad \text{for} \quad 1 \leq j \leq n.$$

In addition, with probability at least $1 - O(n^{-10})$, the remaining term $\boldsymbol{\Delta}_{\boldsymbol{X}} \in \mathbb{R}^{n\times r}$ obeys

$$\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{2,\infty} \lesssim \frac{\sigma}{\sqrt{p\sigma_{\min}}}\sqrt{\frac{\kappa^2\mu r^2 \log^2 n}{np}}.$$

Proof. Fix any $1 \leq j \leq n$. The jth row, namely $\boldsymbol{e}_j^\top[\mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{Y}^\star(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star)^{-1}/p]$ is conditionally Gaussian in the sense that

$$\boldsymbol{e}_j^\top\left[\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}^\star\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1}\right]\Big|\,\Omega \;\sim\; \mathcal{N}\Big(\boldsymbol{0}, \frac{\sigma^2}{p}\Big(\frac{1}{p}\sum_{k=1}^{n}\delta_{jk}\left(\boldsymbol{\Sigma}^\star\right)^{-1}\left(\boldsymbol{Y}^\star_{k,\cdot}\right)^\top\boldsymbol{Y}^\star_{k,\cdot}\left(\boldsymbol{\Sigma}^\star\right)^{-1}\Big)\Big),$$

where we recall that $\delta_{jk} = \mathbb{1}\{(j,k) \in \Omega\}$. Recognize that the conditional covariance matrix concentrates sharply around its expectation, i.e. $\sigma^2(\boldsymbol{\Sigma}^\star)^{-1}/p$, which is the covariance matrix of $\boldsymbol{Z}_{\boldsymbol{X}}^\top\boldsymbol{e}_j$ that we are after. Hence, one can expect that $\mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{Y}^\star(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star)^{-1}/p$ is, marginally, not too far from a Gaussian random matrix. This argument can be carried out formally; see Appendix D.6. □

5.5 From low-rank factors to matrix entries (Proof of Theorem 6)

We now turn attention to inference on the matrix entries, by establishing Theorem 6. Towards this, we first make the following observation: for any $1 \leq i, j \leq n$,

$$\begin{aligned} M^{\mathsf{d}}_{ij} - M^\star_{ij} &= \boldsymbol{e}_i^\top\overline{\boldsymbol{X}}^{\mathsf{d}}\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\boldsymbol{e}_j - \boldsymbol{e}_i^\top\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j \\ &= \boldsymbol{e}_i^\top\big(\overline{\boldsymbol{X}}^{\mathsf{d}} - \boldsymbol{X}^\star\big)\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j + \boldsymbol{e}_i^\top\boldsymbol{X}^\star\big(\overline{\boldsymbol{Y}}^{\mathsf{d}} - \boldsymbol{Y}^\star\big)^\top\boldsymbol{e}_j + \boldsymbol{e}_i^\top\big(\overline{\boldsymbol{X}}^{\mathsf{d}} - \boldsymbol{X}^\star\big)\big(\overline{\boldsymbol{Y}}^{\mathsf{d}} - \boldsymbol{Y}^\star\big)^\top\boldsymbol{e}_j. \end{aligned} \tag{5.13}$$

One can readily apply the decompositions in Theorem 5 to obtain

$$\boldsymbol{e}_i^\top\big(\overline{\boldsymbol{X}}^{\mathsf{d}} - \boldsymbol{X}^\star\big)\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j = \boldsymbol{e}_i^\top\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j + \boldsymbol{e}_i^\top\boldsymbol{\Psi}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j, \tag{5.14}$$

$$\boldsymbol{e}_i^\top\boldsymbol{X}^\star\big(\overline{\boldsymbol{Y}}^{\mathsf{d}} - \boldsymbol{Y}^\star\big)^\top\boldsymbol{e}_j = \boldsymbol{e}_i^\top\boldsymbol{X}^\star\boldsymbol{Z}_{\boldsymbol{Y}}^\top\boldsymbol{e}_j + \boldsymbol{e}_i^\top\boldsymbol{X}^\star\boldsymbol{\Psi}_{\boldsymbol{Y}}^\top\boldsymbol{e}_j. \tag{5.15}$$

Take the preceding three identities collectively to reach

$$\begin{aligned} M^{\mathsf{d}}_{ij} - M^\star_{ij} = &\underbrace{\boldsymbol{e}_i^\top\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j + \boldsymbol{e}_i^\top\boldsymbol{X}^\star\boldsymbol{Z}_{\boldsymbol{Y}}^\top\boldsymbol{e}_j}_{:=\Theta_{ij}} \\ &+ \underbrace{\boldsymbol{e}_i^\top\boldsymbol{\Psi}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j + \boldsymbol{e}_i^\top\boldsymbol{X}^\star\boldsymbol{\Psi}_{\boldsymbol{Y}}^\top\boldsymbol{e}_j + \boldsymbol{e}_i^\top\big(\overline{\boldsymbol{X}}^{\mathsf{d}} - \boldsymbol{X}^\star\big)\big(\overline{\boldsymbol{Y}}^{\mathsf{d}} - \boldsymbol{Y}^\star\big)^\top\boldsymbol{e}_j}_{:=\Lambda_{ij}}. \end{aligned}$$

Following the same route as in Section 5.4, one can verify that $\Theta_{ij} = \boldsymbol{e}_i^\top\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j + \boldsymbol{e}_i^\top\boldsymbol{X}^\star\boldsymbol{Z}_{\boldsymbol{Y}}^\top\boldsymbol{e}_j$ is approximately Gaussian, whereas the residual term Λ_{ij} is small in magnitude. These claims are formally stated in the next two lemmas, with the proofs deferred to Appendix E.

Lemma 10 (Negligibility of Λ_{ij})**.** *Suppose that the sample complexity obeys $n^2p \geq C\kappa^4\mu^2r^2n\log^3 n$ for some sufficiently large constant $C > 0$ and the noise satisfies $\sigma\sqrt{(\kappa^4\mu rn\log n)/p} \leq c\sigma_{\min}$ for some sufficiently small constant $c < 0$. Then with probability exceeding $1 - O(n^{-10})$, one has*

$$|\Lambda_{ij}| \lesssim \left(\left\|\boldsymbol{U}^\star_{i,\cdot}\right\|_2 + \left\|\boldsymbol{V}^\star_{j,\cdot}\right\|_2\right)\frac{\sigma}{\sqrt{p}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^8\mu rn\log n}{p}} + \sqrt{\frac{\kappa^8\mu^3r^3\log^2 n}{np}}\right) + \left(\frac{\sigma}{\sqrt{\sigma_{\min}}}\sqrt{\frac{\kappa^3\mu r\log n}{p}}\right)^2.$$

Lemma 11 (Approximate Gaussianity of Θ_{ij})**.** *Suppose that $np \geq C\kappa^2\mu r^2\log^2 n$ for some sufficiently large constant $C > 0$. Then we have the decomposition*

$$\Theta_{ij} = \boldsymbol{e}_i^\top\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j + \boldsymbol{e}_i^\top\boldsymbol{X}^\star\boldsymbol{Z}_{\boldsymbol{Y}}^\top\boldsymbol{e}_j = g_{ij} + \theta_{ij},$$

where $g_{ij} \sim \mathcal{N}(0, v^\star_{ij})$ and the remaining term θ_{ij} satisfies — with probability exceeding $1 - O(n^{-10})$ — that

$$|\theta_{ij}| \lesssim \frac{\sigma}{\sqrt{p}}\sqrt{\frac{\kappa^2\mu r\log n}{np}}\min\left\{\left\|\boldsymbol{U}^\star_{i,\cdot}\right\|_2, \left\|\boldsymbol{V}^\star_{j,\cdot}\right\|_2\right\}.$$

Putting the above two lemmas together immediately establishes Theorem 6 and hence Theorem 2.

6 Discussion

The present paper makes progress towards inference and uncertainty quantification for noisy matrix completion, by developing simple de-biased estimators that admit tractable and accurate distributional characterizations. While we have achieved some early success in accomplishing this, our results are likely sub-optimal in the following aspects:

- *Dependency on the rank and the condition number.* To enable valid inference, our sample complexity (cf. (3.13) and (3.18a)) scales sub-optimally with the rank r and the condition number κ. The sub-optimality can be understood through comparisons with the sample size requirement $O(nr\log^2 n)$ in the noise-free settings, which is independent of κ and matches the information limit (up to some log factor). Improving such dependency calls for more refined analysis techniques.

- *Detection of the size of the entries.* On one hand, when the size of the entry $M_{ij}^{\star}$ is moderately large (cf. (3.18b)), Corollary 1 allows us to construct a valid confidence interval for it. On the other hand, when $\|\boldsymbol{U}_{i,\cdot}^{\star}\|_2 + \|\boldsymbol{V}_{j,\cdot}^{\star}\|_2$ vanishes, Theorem 5 tells us that the estimation error $M_{ij}^{\mathsf{d}} - M_{ij}^{\star}$ is better approximated by the inner product of two independent Gaussian random vectors. It remains to be seen how to determine whether $\|\boldsymbol{U}_{i,\cdot}^{\star}\|_2 + \|\boldsymbol{V}_{j,\cdot}^{\star}\|_2$ is too small.

- *Low signal-to-noise (SNR) regime.* Our theory operates under the moderate-to-high SNR regime, where $\sigma_{\min}^2/\sigma^2$ (which is proportional to the SNR) is required to exceed the order of n/p; see the conditions in Theorem 5. It is unclear whether the connection between the convex and the nonconvex estimators hold in the low SNR regime. How to conduct inference in such a scenario is an important future direction.

In addition, our investigation has been dedicated to a natural random model, which by no means covers the most general settings of practical interest. There are numerous possible extensions that merit future investigation:

- *Approximate low-rank structure.* Our current theory is built upon the exact low-rank structure of $\boldsymbol{M}^{\star}$. Realistically, the matrix of interest is often only *approximately* low-rank. It is of great interest to study how to carry out statistical inference under such imperfect structural assumptions.

- *More general sampling patterns.* This paper operates under the uniform random sampling assumption, which might sometimes be off in practical situations. It would be interesting to investigate whether our results in this paper can extend to more general non-uniform sampling patterns (e.g. [NW12]).

- *Extensions to robust PCA, sparse PCA, and 1-bit matrix completion.* A variety of important extensions of matrix completion have been explored in prior literature, including but not limited to the case with sparse outliers (i.e. robust PCA [CLMW11, CSPW11]), the case where the matrix of interest is simultaneously sparse and low-rank (i.e. sparse PCA [ZHT06, CMW13]), and the case where only finite-bit observations are available (i.e. 1-bit matrix completion [DPVW14, CZ13]). Performing valid uncertainty assessment for these scenarios requires non-trivial extensions of the link between convex and nonconvex optimization.

- *Other loss functions.* In the estimation stage, one might sometimes prefer other loss functions beyond the penalized squared loss. This might arise from either statistical considerations (e.g. employing a penalized Poisson log-likelihood to accommodate Poisson noise [CX16]), or computational concerns (e.g. adopting a non-smooth loss to improve convergence [CCD+19]). It would be of fundamental importance to develop a unified inferential framework that covers a broader family of loss functions.

Acknowledgements

Y. Chen is supported in part by the AFOSR YIP award FA9550-19-1-0030, by the ONR grant N00014-19-1-2120, by the ARO grant W911NF-18-1-0303, by the NSF grants CCF-1907661 and IIS-1900140, and by the Princeton SEAS innovation award. J. Fan is supported in part by NSF grants DMS-1662139 and DMS-1712591, ONR grant N00014-19-1-2120, and NIH grant 2R01-GM072611-13. C. Ma is supported in part by Hudson River Trading AI Labs (HAIL) Fellowship. This work was done in part while Y. Chen was visiting the Kavli Institute for Theoretical Physics (supported in part by the NSF grant PHY-1748958). We thank Weijie Su for helpful discussions.

A Preliminaries

In this section, we gather several notation and preliminary facts that are useful throughout the analysis. All the proofs, if needed, are deferred to Appendix I.

A.1 Algorithmic details of nonconvex optimization

To begin with, we make precise the algorithm used to minimize the nonconvex loss function (5.1). Specifically, we describe the following details that are crucial for us to implement Algorithm 1:

- Set the initial point to be $(\boldsymbol{X}^0, \boldsymbol{Y}^0) = (\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$ or the spectral initialization as in [MWCC17, CLL19];
- Set the stepsize $\eta \asymp 1/(n^6\kappa^3\sigma_{\max})$;
- Set the maximum number of iterations to be $t_0 \asymp n^{23}$;
- The returned estimate is $(\boldsymbol{X}^{\mathsf{ncvx}}, \boldsymbol{Y}^{\mathsf{ncvx}}) \triangleq (\boldsymbol{X}^{t_\star}, \boldsymbol{Y}^{t_\star})$, where

$$t_\star \triangleq \min \left\{ 0 \leq t \leq t_0 \,\middle|\, \left\|\nabla f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t\right)\right\|_{\mathrm{F}} \leq \frac{1}{n^5}\frac{\lambda}{p}\sqrt{\sigma_{\min}} \right\}. \tag{A.1}$$

  In words, we run gradient descent in Algorithm 1 until we reach a point whose gradient is exceedingly small.

Remark 6 (Spectral initialization)**.** *Many of the preliminary facts below were established for the case* $(\boldsymbol{X}^0, \boldsymbol{Y}^0) = (\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$ *[CCF⁺19], which is certainly not implementable in practice, however, it serves as a good proxy for studying the convex estimator. Fortunately, the same theoretical guarantees stated in Appendix A.2 can be readily established for spectral initialization using almost the same arguments adopted in [MWCC17, CLL19, CCF⁺19]. We omit this part mainly for the sake of brevity.*

To facilitate analysis, we introduce a set of auxiliary nonconvex loss functions. For any $1 \leq j \leq n$, define

$$f^{(j)}\left(\boldsymbol{X}, \boldsymbol{Y}\right) \triangleq \frac{1}{2p}\left\|\mathcal{P}_{\Omega_{-j,\cdot}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \frac{1}{2}\left\|\mathcal{P}_{j,\cdot}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\left\|\boldsymbol{X}\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\left\|\boldsymbol{Y}\right\|_{\mathrm{F}}^2, \tag{A.2}$$

where $\mathcal{P}_{\Omega_{-j,\cdot}} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ (resp. $\mathcal{P}_{j,\cdot}(\cdot)$) denotes the orthogonal projection onto the subspace of matrices that vanish outside of $\{(i,k) \in \Omega \,|\, i \neq j\}$ (resp. $\{(i,k) \,|\, i = j\}$). Let

$$(\boldsymbol{X}^{(j)}, \boldsymbol{Y}^{(j)}) = (\boldsymbol{X}^{t_\star,(j)}, \boldsymbol{Y}^{t_\star,(j)}) \tag{A.3}$$

be the nonconvex estimate returned by this auxiliary algorithm (i.e. Algorithm 2), which serves as an approximate solution to (A.2).

Algorithm 2 Gradient descent for solving the auxiliary nonconvex problem (A.2).

```
  Suitable initialization: $(\boldsymbol{X}^{0,(j)}, \boldsymbol{Y}^{0,(j)}) = (\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$
  Gradient updates: for $t = 0, 1, \ldots, t_\star - 1$ do
      $\boldsymbol{X}^{t+1,(j)} = \boldsymbol{X}^{t,(j)} - \eta\nabla_{\boldsymbol{X}} f^{(j)}\big(\boldsymbol{X}^{t,(j)}, \boldsymbol{Y}^{t,(j)}\big);$   (A.4a)
      $\boldsymbol{Y}^{t+1,(j)} = \boldsymbol{Y}^{t,(j)} - \eta\nabla_{\boldsymbol{Y}} f^{(j)}\big(\boldsymbol{X}^{t,(j)}, \boldsymbol{Y}^{t,(j)}\big).$   (A.4b)
```

A.2 Properties of approximate nonconvex solutions

This subsection gathers the properties of the (approximate) nonconvex solutions. Throughout this subsection, we use the shorthand

$$(\boldsymbol{X},\boldsymbol{Y})=(\boldsymbol{X}^{\mathsf{ncvx}},\boldsymbol{Y}^{\mathsf{ncvx}}) \tag{A.5}$$

and recall the definition of $(\boldsymbol{X}^{(j)},\boldsymbol{Y}^{(j)})$ in (A.3). The regularization parameter is chosen to satisfy

$$\lambda \asymp \sigma\sqrt{np}. \tag{A.6}$$

To further simplify the presentation, we introduce $\boldsymbol{F}^{\star}$, $\boldsymbol{F}$, $\boldsymbol{F}^{\mathsf{d}}$, $\boldsymbol{F}^{\mathsf{d},(j)}\in\mathbb{R}^{2n\times r}$ as follows

$$\boldsymbol{F}^{\star}\triangleq\begin{bmatrix}\boldsymbol{X}^{\star}\\ \boldsymbol{Y}^{\star}\end{bmatrix};\quad \boldsymbol{F}\triangleq\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix};\quad \boldsymbol{F}^{\mathsf{d}}\triangleq\begin{bmatrix}\boldsymbol{X}^{\mathsf{d}}\\ \boldsymbol{Y}^{\mathsf{d}}\end{bmatrix};\quad \boldsymbol{F}^{\mathsf{d},(j)}\triangleq\begin{bmatrix}\boldsymbol{X}^{\mathsf{d},(j)}\\ \boldsymbol{Y}^{\mathsf{d},(j)}\end{bmatrix}, \tag{A.7}$$

and define

$$\boldsymbol{H}\triangleq\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\|\boldsymbol{F}\boldsymbol{R}-\boldsymbol{F}^{\star}\|_{\mathrm{F}}^{2}=\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\{\|\boldsymbol{X}\boldsymbol{R}-\boldsymbol{X}^{\star}\|_{\mathrm{F}}^{2}+\|\boldsymbol{Y}\boldsymbol{R}-\boldsymbol{Y}^{\star}\|_{\mathrm{F}}^{2}\right\}, \tag{A.8a}$$

$$\boldsymbol{H}^{(j)}\triangleq\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{F}^{(j)}\boldsymbol{R}-\boldsymbol{F}^{\star}\right\|_{\mathrm{F}}^{2}=\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\{\left\|\boldsymbol{X}^{(j)}\boldsymbol{R}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{Y}^{(j)}\boldsymbol{R}-\boldsymbol{Y}^{\star}\right\|_{\mathrm{F}}^{2}\right\}, \tag{A.8b}$$

$$\boldsymbol{R}^{(j)}\triangleq\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{F}^{(j)}\boldsymbol{R}-\boldsymbol{F}\boldsymbol{H}\right\|_{\mathrm{F}}^{2}=\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\{\left\|\boldsymbol{X}^{(j)}\boldsymbol{R}-\boldsymbol{X}\boldsymbol{H}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{Y}^{(j)}\boldsymbol{R}-\boldsymbol{Y}\boldsymbol{H}\right\|_{\mathrm{F}}^{2}\right\}, \tag{A.8c}$$

$$\boldsymbol{H}^{\mathsf{d},(j)}\triangleq\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{F}^{\mathsf{d},(j)}\boldsymbol{R}-\boldsymbol{F}^{\star}\right\|_{\mathrm{F}}^{2}=\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\{\left\|\boldsymbol{X}^{\mathsf{d},(j)}\boldsymbol{R}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{Y}^{\mathsf{d},(j)}\boldsymbol{R}-\boldsymbol{Y}^{\star}\right\|_{\mathrm{F}}^{2}\right\}. \tag{A.8d}$$

The claims stated below hold under the sample complexity and the noise condition presumed in [CCF+19, Theorem 1] (see also Theorem 5 in the current manuscript)

$$n^{2}p\gg\kappa^{4}\mu^{2}r^{2}n\log^{3}n\quad\text{and}\quad\sigma\sqrt{\frac{n}{p}}\ll\frac{\sigma_{\min}}{\sqrt{\kappa^{4}\mu r\log n}}.$$

1. The first set of facts is related to $(\boldsymbol{X},\boldsymbol{Y})$. In view of [CCF+19], $\boldsymbol{F}$ is a faithful estimate[6] of $\boldsymbol{F}^{\star}$, in the sense that

$$\|\boldsymbol{F}\boldsymbol{H}-\boldsymbol{F}^{\star}\|_{\mathrm{F}}\lesssim\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\,\|\boldsymbol{X}^{\star}\|_{\mathrm{F}}\,, \tag{A.9a}$$

$$\|\boldsymbol{F}\boldsymbol{H}-\boldsymbol{F}^{\star}\|\lesssim\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\,\|\boldsymbol{X}^{\star}\|\,, \tag{A.9b}$$

$$\|\boldsymbol{F}\boldsymbol{H}-\boldsymbol{F}^{\star}\|_{2,\infty}\lesssim\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\,\|\boldsymbol{F}^{\star}\|_{2,\infty} \tag{A.9c}$$

hold with probability exceeding $1-O(n^{-10})$. In addition, on the same high-probability event, one has

$$\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\|_{\mathrm{F}}\leq\frac{1}{n^{5}}\frac{\lambda}{p}\sqrt{\sigma_{\min}}; \tag{A.10}$$

$$\left\|\boldsymbol{X}^{\top}\boldsymbol{X}-\boldsymbol{Y}^{\top}\boldsymbol{Y}\right\|_{\mathrm{F}}\leq\frac{1}{n^{5}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\sigma_{\max}\leq\frac{1}{n^{5}}\sigma_{\max}; \tag{A.11}$$

$$\max\left\{\left\|\boldsymbol{Z}^{\mathsf{cvx}}-\boldsymbol{X}\boldsymbol{Y}^{\top}\right\|_{\mathrm{F}},\left\|\boldsymbol{Z}^{\mathsf{cvx},r}-\boldsymbol{X}\boldsymbol{Y}^{\top}\right\|_{\mathrm{F}}\right\}\lesssim\frac{\kappa^{2}}{n^{5}}\frac{\lambda}{p}. \tag{A.12}$$

In words, the first claim ensures that $(\boldsymbol{X},\boldsymbol{Y})$ is an approximate stationary point of $f(\cdot,\cdot)$; the second bound tells us that $(\boldsymbol{X},\boldsymbol{Y})$ is nearly *balanced*, in the sense that $\boldsymbol{X}^{\top}\boldsymbol{X}\approx\boldsymbol{Y}^{\top}\boldsymbol{Y}$; the last one formalizes the proximity between the convex solution and the nonconvex one; see also (3.5).

[6] Technically, the statements in [CCF+19, Lemma 5] are for $\eta\asymp1/(n\kappa^{3}\sigma_{\max})$ and $t_{0}\asymp n^{18}$. Nevertheless, inspecting their proofs reveals that the claims continue to hold for our choices $\eta\asymp1/(n^{6}\kappa^{3}\sigma_{\max})$ and $t_{0}\asymp n^{23}$.

2. We move on to the properties of the de-shrunken estimator $(\boldsymbol{X}^{\mathsf{d}}, \boldsymbol{Y}^{\mathsf{d}})$, which is defined in (3.8). Specifically, we can show that (see Appendix I)

$$\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H} - \boldsymbol{F}^{\star}\right\| \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^{\star}\right\|, \tag{A.13a}$$

$$\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{F}^{\star}\right\| \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^{\star}\right\|, \tag{A.13b}$$

$$\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{F}^{\star}\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}, \tag{A.13c}$$

$$\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{F}^{\star}\right\|_{2,\infty} \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}, \tag{A.13d}$$

$$\left\|\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}} - \boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right\| \lesssim \frac{\kappa}{n^{5}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\sigma_{\max} \tag{A.13e}$$

   hold with probability at least $1 - O(n^{-10})$.

3. As has been shown in [CCF+19], the leave-one-out auxiliary point $(\boldsymbol{X}^{(j)}, \boldsymbol{Y}^{(j)})$ satisfies

$$\left\|\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)} - \boldsymbol{F}\boldsymbol{H}\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}, \tag{A.14a}$$

$$\left\|\boldsymbol{F}^{(j)}\boldsymbol{H}^{(j)} - \boldsymbol{F}\boldsymbol{H}\right\|_{\mathrm{F}} \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}, \tag{A.14b}$$

$$\left\|\boldsymbol{F}^{(j)}\boldsymbol{H}^{(j)} - \boldsymbol{F}^{\star}\right\| \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^{\star}\right\|, \tag{A.14c}$$

$$\left\|\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)} - \boldsymbol{F}^{\star}\right\|_{2,\infty} \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty} \tag{A.14d}$$

   with probability exceeding $1 - O(n^{-10})$.

4. Parallel to the transition from $(\boldsymbol{X}, \boldsymbol{Y})$ to $(\boldsymbol{X}^{\mathsf{d}}, \boldsymbol{Y}^{\mathsf{d}})$, we set

$$\boldsymbol{X}^{\mathsf{d},(j)} \triangleq \boldsymbol{X}^{(j)}\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{(j)\top}\boldsymbol{X}^{(j)}\right)^{-1}\right)^{1/2} \quad \text{and} \quad \boldsymbol{Y}^{\mathsf{d},(j)} \triangleq \boldsymbol{Y}^{(j)}\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\right)^{-1}\right)^{1/2} \tag{A.15}$$

   to be the de-shrunken estimators of $\boldsymbol{X}^{(j)}$ and $\boldsymbol{Y}^{(j)}$, respectively. We shall demonstrate in Appendix I that, with probability at least $1 - O(n^{-10})$,

$$\left\|\boldsymbol{F}^{\mathsf{d},(j)}\boldsymbol{H}^{\mathsf{d},(j)} - \boldsymbol{F}^{\star}\right\| \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^{\star}\right\|, \tag{A.16a}$$

$$\left\|\boldsymbol{F}^{\mathsf{d},(j)}\boldsymbol{H}^{\mathsf{d},(j)} - \boldsymbol{F}^{\star}\right\|_{2,\infty} \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}, \tag{A.16b}$$

$$\left\|\boldsymbol{F}^{\mathsf{d},(j)}\boldsymbol{H}^{\mathsf{d},(j)} - \boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}}\right\| \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}. \tag{A.16c}$$

In addition to these four sets of claims, we have the following immediate consequence of the incoherence condition (2.4)

$$\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty} = \max\left\{\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}, \left\|\boldsymbol{Y}^{\star}\right\|_{2,\infty}\right\} \leq \sqrt{\mu r\sigma_{\max}/n}. \tag{A.17}$$

Moreover, recall that $\boldsymbol{A} = (1/p)\cdot\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top}\right) - \left(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top}\right)$ (cf. (5.7)). We obtain from the proof of [CCF+19, Lemma 8] that

$$\left\|\boldsymbol{A}\right\| \lesssim \sigma\sqrt{\frac{n}{p}}\cdot\sqrt{\frac{\kappa^{4}\mu^{2}r^{2}\log n}{np}}. \tag{A.18}$$

Last but not least, we list a few simple but useful results: the nonconvex solution $\boldsymbol{F}$ satisfies

$$\sigma_r(\boldsymbol{F}) \geq 0.5\sqrt{\sigma_{\min}}, \quad \|\boldsymbol{F}\| \leq 2\|\boldsymbol{X}^{\star}\|, \quad \|\boldsymbol{F}\|_{\mathrm{F}} \leq 2\|\boldsymbol{X}^{\star}\|_{\mathrm{F}}, \quad \|\boldsymbol{F}\|_{2,\infty} \leq 2\|\boldsymbol{F}^{\star}\|_{2,\infty}. \tag{A.19}$$

The same holds true if we replace $\boldsymbol{F}$ by either $\boldsymbol{F}^{\mathsf{d}}$, $\boldsymbol{F}^{(j)}$ $\boldsymbol{F}^{\mathsf{d},(j)}$ or their corresponding low-rank factors. Here j can vary from 1 to n.

B Summary of the proposed estimators

Let $\boldsymbol{Z}^{\mathsf{cvx}}$ be the minimizer of the convex program (3.1), and let $(\boldsymbol{X}^{\mathsf{ncvx}}, \boldsymbol{Y}^{\mathsf{ncvx}})$ be the solution returned by the Algorithm 1 (with algorithmic details specified in Appendix A.1). Recall that $\boldsymbol{Z}^{\mathsf{cvx},r}$ is the best rank-r approximation of $\boldsymbol{Z}^{\mathsf{cvx}}$, viz.

$$\boldsymbol{Z}^{\mathsf{cvx},r} = \mathop{\arg\min}_{\boldsymbol{B}:\,\mathrm{rank}(\boldsymbol{B})\leq r} \|\boldsymbol{B} - \boldsymbol{Z}^{\mathsf{cvx}}\|_{\mathrm{F}}.$$

In addition, we let the matrix estimate obtained by the nonconvex algorithm be $\boldsymbol{Z}^{\mathsf{ncvx}} \triangleq \boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$. We further denote by $(\boldsymbol{X}^{\mathsf{cvx}}, \boldsymbol{Y}^{\mathsf{cvx}})$ the estimate of low-rank factors obtained by convex relaxation; more specifically, we set $(\boldsymbol{X}^{\mathsf{cvx}}, \boldsymbol{Y}^{\mathsf{cvx}})$ to be the balanced rank-r factorization of $\boldsymbol{Z}^{\mathsf{cvx},r}$ obeying $\boldsymbol{X}^{\mathsf{cvx}}\boldsymbol{Y}^{\mathsf{cvx}\top} = \boldsymbol{Z}^{\mathsf{cvx},r}$ and $\boldsymbol{X}^{\mathsf{cvx}\top}\boldsymbol{X}^{\mathsf{cvx}} = \boldsymbol{Y}^{\mathsf{cvx}\top}\boldsymbol{Y}^{\mathsf{cvx}}$. With these notations in place, our de-biased and de-shrunken estimators can be summarized as follows.

- De-biased matrix estimators:

$$\boldsymbol{M}^{\mathsf{cvx,d}} \triangleq \mathcal{P}_{\text{rank-}r}\Big[\boldsymbol{Z}^{\mathsf{cvx}} - \frac{1}{p}\mathcal{P}_{\Omega}\big(\boldsymbol{Z}^{\mathsf{cvx}} - \boldsymbol{M}\big)\Big], \tag{B.1a}$$

$$\boldsymbol{M}^{\mathsf{ncvx,d}} \triangleq \mathcal{P}_{\text{rank-}r}\Big[\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} - \frac{1}{p}\mathcal{P}_{\Omega}\big(\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} - \boldsymbol{M}\big)\Big]. \tag{B.1b}$$

- De-shrunken estimators for low-rank factors:

$$\boldsymbol{X}^{\mathsf{ncvx,d}} \triangleq \boldsymbol{X}^{\mathsf{ncvx}}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\big(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\big)^{-1}\Big)^{1/2}, \tag{B.2a}$$

$$\boldsymbol{Y}^{\mathsf{ncvx,d}} \triangleq \boldsymbol{Y}^{\mathsf{ncvx}}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\big(\boldsymbol{Y}^{\mathsf{ncvx}\top}\boldsymbol{Y}^{\mathsf{ncvx}}\big)^{-1}\Big)^{1/2}, \tag{B.2b}$$

$$\boldsymbol{X}^{\mathsf{cvx,d}} \triangleq \boldsymbol{X}^{\mathsf{cvx}}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\big(\boldsymbol{X}^{\mathsf{cvx}\top}\boldsymbol{X}^{\mathsf{cvx}}\big)^{-1}\Big)^{1/2}, \tag{B.2c}$$

$$\boldsymbol{Y}^{\mathsf{cvx,d}} \triangleq \boldsymbol{Y}^{\mathsf{cvx}}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\big(\boldsymbol{Y}^{\mathsf{cvx}\top}\boldsymbol{Y}^{\mathsf{cvx}}\big)^{-1}\Big)^{1/2}. \tag{B.2d}$$

C Proof of Lemma 3

Throughout this section, let $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$ be the rank-r SVD of the nonconvex estimate $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$ and T the tangent space of the set of rank-r matrices at $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$. Correspondingly, we denote by $\mathcal{P}_T$ the projection operator onto the tangent space T, and let $\mathcal{P}_{T^{\perp}} = \mathcal{I} - \mathcal{P}_T$, where $\mathcal{I}$ is the identity operator.

C.1 Proof of the inequality (5.4a)

In essence, we intend to justify that $\boldsymbol{M}^{\mathsf{cvx,d}}$, $\boldsymbol{M}^{\mathsf{ncvx,d}}$ and $\boldsymbol{X}^{\mathsf{ncvx,d}}\boldsymbol{Y}^{\mathsf{ncvx,d}\top}$ are all very close to $\boldsymbol{U}(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r)\boldsymbol{V}^{\top}$.

Recall from the definition of the de-biased estimator $\boldsymbol{M}^{\mathsf{cvx,d}}$ (cf. (B.1a)) that

$$\boldsymbol{M}^{\mathsf{cvx,d}} = \mathcal{P}_{\text{rank-}r}\Big[\boldsymbol{Z}^{\mathsf{cvx}} - \frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{Z}^{\mathsf{cvx}} - \boldsymbol{M}\right)\Big]. \tag{C.1}$$

Replacing $\boldsymbol{Z}^{\mathsf{cvx}}$ by $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$ results in

$$\boldsymbol{Z}^{\mathsf{cvx}} - \frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{Z}^{\mathsf{cvx}} - \boldsymbol{M}\right) = \boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} - \frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} - \boldsymbol{M}\right) + \boldsymbol{\Delta}_{\boldsymbol{Z}}, \tag{C.2}$$

where we denote

$$\boldsymbol{\Delta}_{\boldsymbol{Z}} \triangleq \left(\boldsymbol{Z}^{\mathsf{cvx}} - \boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}\right) + \frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} - \boldsymbol{Z}^{\mathsf{cvx}}\right).$$

Apply the proximity bound (A.12) to obtain (recall that in (A.12), one has $(\boldsymbol{X},\boldsymbol{Y}) = (\boldsymbol{X}^{\mathsf{ncvx}}, \boldsymbol{Y}^{\mathsf{ncvx}})$)

$$\begin{aligned} \|\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_{\mathrm{F}} &\leq \left\|\boldsymbol{Z}^{\mathsf{cvx}} - \boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}\right\|_{\mathrm{F}} + \frac{1}{p}\left\|\boldsymbol{Z}^{\mathsf{cvx}} - \boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}\right\|_{\mathrm{F}} \\ &\leq \frac{2}{p}\left\|\boldsymbol{Z}^{\mathsf{cvx}} - \boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}\right\|_{\mathrm{F}} \lesssim \frac{\kappa^2}{n^5 p}\frac{\lambda}{p} \leq \frac{\lambda}{8p}, \end{aligned} \tag{C.3}$$

as long as $n^5 p \gg \kappa^2$. In addition, in view of [CCF+19, Claim 2], one has the decomposition

$$\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} - \boldsymbol{M}\right) = -\lambda \boldsymbol{U}\boldsymbol{V}^{\top} + \boldsymbol{R}, \tag{C.4}$$

where $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ is a residual matrix obeying

$$\left\|\mathcal{P}_T\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}} \lesssim \kappa \frac{p}{\sqrt{\sigma_{\min}}}\left\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\right\|_{\mathrm{F}} \lesssim \frac{\kappa}{n^5}\lambda \leq \frac{\lambda}{8} \qquad \text{and} \qquad \left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{R}\right)\right\| \leq \frac{\lambda}{2} \tag{C.5}$$

with probability exceeding $1 - O(n^{-10})$. Here we utilize the small-gradient condition $\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\|_{\mathrm{F}} \leq \frac{1}{n^5}\frac{\lambda}{p}\sqrt{\sigma_{\min}}$ (cf. (A.10)). Take (C.1), (C.2) and (C.4) collectively to reach

$$\begin{aligned} \boldsymbol{M}^{\mathsf{cvx,d}} &= \mathcal{P}_{\text{rank-}r}\left[\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} + \frac{\lambda}{p}\boldsymbol{U}\boldsymbol{V}^{\top} - \frac{1}{p}\boldsymbol{R} + \boldsymbol{\Delta}_{\boldsymbol{Z}}\right] \\ &= \mathcal{P}_{\text{rank-}r}\left[\boldsymbol{U}\Big(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^{\top} + \boldsymbol{\Delta}_{\boldsymbol{Z}} - \frac{1}{p}\boldsymbol{R}\right] \\ &= \mathcal{P}_{\text{rank-}r}\Big[\underbrace{\boldsymbol{U}\Big(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^{\top} + \mathcal{P}_{T^{\perp}}\Big(\boldsymbol{\Delta}_{\boldsymbol{Z}} - \frac{1}{p}\boldsymbol{R}\Big)}_{:=\boldsymbol{C}} + \underbrace{\mathcal{P}_T\Big(\boldsymbol{\Delta}_{\boldsymbol{Z}} - \frac{1}{p}\boldsymbol{R}\Big)}_{:=\boldsymbol{\Delta}}\Big], \end{aligned} \tag{C.6}$$

where the middle line follows since $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$ is defined to be the SVD of $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$.

We view $\boldsymbol{\Delta}$ as a perturbation and intend to apply Lemma 14 to control $\|\boldsymbol{M}^{\mathsf{cvx,d}} - \boldsymbol{U}(\boldsymbol{\Sigma} + (\lambda/p)\boldsymbol{I}_r)\boldsymbol{V}^{\top}\|_{\mathrm{F}}$. First, notice that the rth largest singular value obeys $\sigma_r(\boldsymbol{U}(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r)\boldsymbol{V}^{\top}) \geq \frac{\lambda}{p}$, and that

$$\left\|\mathcal{P}_{T^{\perp}}\Big(\boldsymbol{\Delta}_{\boldsymbol{Z}} - \frac{1}{p}\boldsymbol{R}\Big)\right\| \leq \|\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_{\mathrm{F}} + \frac{1}{p}\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}} \leq \frac{5\lambda}{8p}, \tag{C.7}$$

where the last inequality results from (C.3) and (C.5). Combining the above two bounds with the fact that $\boldsymbol{U}(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r)\boldsymbol{V}^{\top}$ and $\mathcal{P}_T(\boldsymbol{\Delta}_{\boldsymbol{Z}} - \frac{1}{p}\boldsymbol{R})$ are orthogonal to each other, we arrive at the conclusion that $\boldsymbol{U}(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r)\boldsymbol{V}^{\top}$ is the top-r SVD of $\boldsymbol{C}$ and

$$\sigma_i\left(\boldsymbol{C}\right) = \sigma_i\Big(\boldsymbol{U}\Big(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^{\top}\Big), \qquad \text{for } 1 \leq i \leq r; \tag{C.8a}$$

$$\sigma_{r+1}\left(\boldsymbol{C}\right) = \left\|\mathcal{P}_{T^{\perp}}\Big(\boldsymbol{\Delta}_{\boldsymbol{Z}} - \frac{1}{p}\boldsymbol{R}\Big)\right\|. \tag{C.8b}$$

Second, let $\hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^{\top}$ be the top-r SVD of $\boldsymbol{C} + \boldsymbol{\Delta}$. By definition, one has $\hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^{\top} = \boldsymbol{M}^{\mathsf{cvx,d}}$. We are left with checking the two conditions in Lemma 14. To begin with, the perturbation term $\boldsymbol{\Delta}$ obeys

$$\|\boldsymbol{\Delta}\|_{\mathrm{F}} \leq \|\boldsymbol{\Delta}_{\boldsymbol{Z}}\|_{\mathrm{F}} + \frac{1}{p}\left\|\mathcal{P}_T\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}} \overset{\text{(i)}}{\lesssim} \frac{\kappa^2}{n^5 p}\frac{\lambda}{p} + \frac{\kappa}{n^5}\frac{\lambda}{p} \overset{\text{(ii)}}{\leq} \frac{1}{2n^4}\frac{\lambda}{p}, \tag{C.9}$$

where (i) comes from (C.3) and (C.5) and the last inequality (ii) arises since $np \gg \kappa^2$. Clearly, the size of the perturbation is much smaller than λ/p and hence $\|\boldsymbol{C}\|$ (cf. (C.8a)). In addition,

$$\sigma_{r+1}\left(\boldsymbol{C} + \boldsymbol{\Delta}\right) \leq \sigma_{r+1}\left(\boldsymbol{C}\right) + \|\boldsymbol{\Delta}\| = \left\|\mathcal{P}_{T^{\perp}}\Big(\boldsymbol{\Delta}_{\boldsymbol{Z}} - \frac{1}{p}\boldsymbol{R}\Big)\right\| + \|\boldsymbol{\Delta}\|_{\mathrm{F}}$$

$$\leq \frac{5\lambda}{8p} + \frac{1}{2n^4}\frac{\lambda}{p} \leq \frac{3\lambda}{4p},$$

where the equality depends on (C.8b) and the last line results from (C.7) and (C.9). Consequently,

$$\sigma_r\left(\boldsymbol{C}\right) - \sigma_{r+1}\left(\boldsymbol{C}+\boldsymbol{\Delta}\right) \geq \sigma_r\left(\boldsymbol{U}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^\top\right) - \frac{3\lambda}{4p} \geq \sigma_r\left(\boldsymbol{\Sigma}\right) + \frac{\lambda}{4p} \geq \frac{\sigma_{\min}}{2}.$$

Here the first relation arises from (C.8a) and the second holds since $\sigma_r(\boldsymbol{\Sigma}) \geq \sigma_{\min}/2$, a simple consequence of (A.19). We are now ready to apply Lemma 14 to obtain

$$\left\|\boldsymbol{M}^{\mathsf{cvx,d}} - \boldsymbol{U}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^\top\right\|_{\mathrm{F}} \leq \left(\frac{12\left\|\boldsymbol{\Sigma}+(\lambda/p)\boldsymbol{I}_r\right\|}{\sigma_{\min}/2}+1\right)\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}} \lesssim \kappa\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}},$$

where we have used the fact that $\|\boldsymbol{\Sigma}+(\lambda/p)\boldsymbol{I}_r\| \lesssim \sigma_{\max}$, which also can be derived from (A.19). The above bound combined with (C.9) yields

$$\left\|\boldsymbol{M}^{\mathsf{cvx,d}} - \boldsymbol{U}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^\top\right\|_{\mathrm{F}} \lesssim \frac{\kappa^3}{n^5 p}\frac{\lambda}{p} + \frac{\kappa^2}{n^5}\frac{\lambda}{p} \leq \frac{1}{2n^4}\frac{\lambda}{p}$$

as long as $np \gg \kappa^3$. We remark that by setting $\boldsymbol{\Delta}_{\boldsymbol{Z}} = \boldsymbol{0}$, one also obtains the bound on $\boldsymbol{M}^{\mathsf{ncvx,d}}$, i.e.

$$\left\|\boldsymbol{M}^{\mathsf{ncvx,d}} - \boldsymbol{U}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^\top\right\|_{\mathrm{F}} \leq \frac{1}{2n^4}\frac{\lambda}{p}. \tag{C.10}$$

We move on to investigating $\|\boldsymbol{X}^{\mathsf{ncvx,d}}\boldsymbol{Y}^{\mathsf{ncvx,d}\top} - \boldsymbol{U}(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r)\boldsymbol{V}^\top\|$, for which we have the following claim.

Claim 1. *One has*

$$\left\|\boldsymbol{X}^{\mathsf{ncvx,d}}\boldsymbol{Y}^{\mathsf{ncvx,d}\top} - \boldsymbol{U}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^\top\right\| \leq \frac{1}{2n^4}\frac{\lambda}{p}. \tag{C.11}$$

Taking the above three bounds collectively and recognizing that $\lambda \lesssim \sigma\sqrt{np}$ yield the advertised bound (5.4a).

Proof of Claim 1. Utilize [CCF+19, Claim 3] to see that

$$\boldsymbol{X}^{\mathsf{ncvx}} = \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q} \qquad \text{and} \qquad \boldsymbol{Y}^{\mathsf{ncvx}} = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}^{-\top} \tag{C.12}$$

hold for some invertible matrix $\boldsymbol{Q} \in \mathbb{R}^{r\times r}$ with SVD $\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^\top$ obeying

$$\left\|\boldsymbol{\Sigma}_{\boldsymbol{Q}} - \boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\right\|_{\mathrm{F}} \leq 8\sqrt{\kappa}\frac{p}{\lambda\sqrt{\sigma_{\min}}}\left\|\nabla f\left(\boldsymbol{X}^{\mathsf{ncvx}},\boldsymbol{Y}^{\mathsf{ncvx}}\right)\right\|_{\mathrm{F}} \leq \frac{8\sqrt{\kappa}}{n^5}. \tag{C.13}$$

The last inequality is the small-gradient condition (see (A.10), in which $(\boldsymbol{X},\boldsymbol{Y}) = (\boldsymbol{X}^{\mathsf{ncvx}},\boldsymbol{Y}^{\mathsf{ncvx}})$). Employ the definitions for $\boldsymbol{X}^{\mathsf{ncvx,d}}$ and $\boldsymbol{Y}^{\mathsf{ncvx,d}}$ (cf. (B.2a) and (B.2b)) to see that

$$\begin{aligned}
\boldsymbol{X}^{\mathsf{ncvx,d}}\boldsymbol{Y}^{\mathsf{ncvx,d}\top} &= \boldsymbol{X}^{\mathsf{ncvx}}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\Big)^{1/2}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^{\mathsf{ncvx}\top}\boldsymbol{Y}^{\mathsf{ncvx}}\right)^{-1}\Big)^{1/2}\boldsymbol{Y}^{\mathsf{ncvx}\top} \\
&= \boldsymbol{X}^{\mathsf{ncvx}}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\Big)^{1/2}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\Big)^{1/2}\boldsymbol{Y}^{\mathsf{ncvx}\top} \\
&\quad - \boldsymbol{X}^{\mathsf{ncvx}}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\Big)\boldsymbol{\Delta}_{\mathsf{balancing}}\boldsymbol{Y}^{\mathsf{ncvx}\top} \\
&= \underbrace{\boldsymbol{X}^{\mathsf{ncvx}}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\Big)\boldsymbol{Y}^{\mathsf{ncvx}\top}}_{:=\boldsymbol{A}_1} - \underbrace{\boldsymbol{X}^{\mathsf{ncvx}}\Big(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\Big)\boldsymbol{\Delta}_{\mathsf{balancing}}\boldsymbol{Y}^{\mathsf{ncvx}\top}}_{:=\boldsymbol{A}_2}.
\end{aligned} \tag{C.14}$$

Here we denote

$$\boldsymbol{\Delta}_{\mathsf{balancing}} \triangleq \Big(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\Big)^{1/2} - \Big(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^{\mathsf{ncvx}\top}\boldsymbol{Y}^{\mathsf{ncvx}}\right)^{-1}\Big)^{1/2}.$$

It then boils down to showing that (i) $\boldsymbol{A}_1$ is very close to $\boldsymbol{U}(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r)\boldsymbol{V}^\top$, and (ii) $\boldsymbol{A}_2$ is small in size.

First, recall that $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, which combined with (C.12) gives

$$\begin{aligned}
\left\|\boldsymbol{A}_1-\boldsymbol{U}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^\top\right\| &= \frac{\lambda}{p}\left\|\boldsymbol{X}^{\mathsf{ncvx}}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\boldsymbol{Y}^{\mathsf{ncvx}\top}-\boldsymbol{U}\boldsymbol{V}^\top\right\| \\
&= \frac{\lambda}{p}\left\|\boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Q}^{-\top}\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}^{1/2}\boldsymbol{V}^\top-\boldsymbol{U}\boldsymbol{V}^\top\right\| \\
&= \frac{\lambda}{p}\left\|\boldsymbol{\Sigma}^{-1/2}\left(\boldsymbol{Q}^{-\top}\boldsymbol{Q}^{-1}-\boldsymbol{I}_r\right)\boldsymbol{\Sigma}^{1/2}\right\| \le \sqrt{\kappa}\frac{\lambda}{p}\|\boldsymbol{Q}^{-\top}\boldsymbol{Q}^{-1}-\boldsymbol{I}_r\| \\
&= \sqrt{\kappa}\frac{\lambda}{p}\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-2}-\boldsymbol{I}_r\| \le \sqrt{\kappa}\frac{\lambda}{p}\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\|\cdot\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}-\boldsymbol{\Sigma}_{\boldsymbol{Q}}\|_{\mathrm{F}} \\
&\lesssim \kappa\frac{\lambda}{p}\frac{1}{n^5}. \qquad\qquad (\text{C.15})
\end{aligned}$$

Here, the last inequality comes from (C.13) and its immediate consequence that $\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}\|\asymp\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\|\asymp 1$.

Second, apply the perturbation bound for matrix square roots (cf. Lemma 13) to obtain

$$\begin{aligned}
\|\boldsymbol{\Delta}_{\mathsf{balancing}}\| &\lesssim \frac{\frac{\lambda}{p}\left\|\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}-\left(\boldsymbol{Y}^{\mathsf{ncvx}\top}\boldsymbol{Y}^{\mathsf{ncvx}}\right)^{-1}\right\|}{\lambda_{\min}\left[\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\right)^{1/2}\right]+\lambda_{\min}\left[\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^{\mathsf{ncvx}\top}\boldsymbol{Y}^{\mathsf{ncvx}}\right)^{-1}\right)^{1/2}\right]} \\
&\overset{\text{(i)}}{\lesssim} \frac{\lambda}{p}\left\|\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}-\left(\boldsymbol{Y}^{\mathsf{ncvx}\top}\boldsymbol{Y}^{\mathsf{ncvx}}\right)^{-1}\right\| \\
&\le \frac{\lambda}{p}\left\|\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\right\|\left\|\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}-\boldsymbol{Y}^{\mathsf{ncvx}\top}\boldsymbol{Y}^{\mathsf{ncvx}}\right\|_{\mathrm{F}}\left\|\left(\boldsymbol{Y}^{\mathsf{ncvx}\top}\boldsymbol{Y}^{\mathsf{ncvx}}\right)^{-1}\right\| \\
&\overset{\text{(ii)}}{\lesssim} \frac{1}{n^5}\frac{\lambda}{p}\frac{\kappa}{\sigma_{\min}}. \qquad\qquad (\text{C.16})
\end{aligned}$$

Here, the inequality (i) depends on the facts that

$$\lambda_{\min}\left[\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\right)^{1/2}\right]\ge 1 \qquad\text{and}\qquad \lambda_{\min}\left[\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^{\mathsf{ncvx}\top}\boldsymbol{Y}^{\mathsf{ncvx}}\right)^{-1}\right)^{1/2}\right]\ge 1,$$

whereas the inequality (ii) holds because of the facts that $\|(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}})^{-1}\|\lesssim 1/\sigma_{\min}$, $\|(\boldsymbol{Y}^{\mathsf{ncvx}\top}\boldsymbol{Y}^{\mathsf{ncvx}})^{-1}\|\lesssim 1/\sigma_{\min}$ and the balancedness condition (A.11)

$$\left\|\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}-\boldsymbol{Y}^{\mathsf{ncvx}\top}\boldsymbol{Y}^{\mathsf{ncvx}}\right\|_{\mathrm{F}}\le\frac{1}{n^5}\sigma_{\max}.$$

As a result, the operator norm of $\boldsymbol{A}_2$ is bounded by

$$\begin{aligned}
\|\boldsymbol{A}_2\| &\le \left\|\boldsymbol{X}^{\mathsf{ncvx}}\Big(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\Big)\right\|\|\boldsymbol{\Delta}_{\mathsf{balancing}}\|\|\boldsymbol{Y}^{\mathsf{ncvx}}\| \\
&\lesssim \sqrt{\sigma_{\max}}\cdot\frac{1}{n^5}\frac{\lambda}{p}\frac{\kappa}{\sigma_{\min}}\cdot\sqrt{\sigma_{\max}}\asymp\frac{\lambda}{p}\frac{\kappa^2}{n^5}. \qquad\qquad (\text{C.17})
\end{aligned}$$

Take (C.14), (C.15) and (C.17) collectively to arrive at

$$\left\|\boldsymbol{X}^{\mathsf{ncvx,d}}\boldsymbol{Y}^{\mathsf{ncvx,d}\top}-\boldsymbol{U}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^\top\right\|\le\left\|\boldsymbol{A}_1-\boldsymbol{U}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^\top\right\|+\|\boldsymbol{A}_2\|\lesssim\frac{\lambda}{p}\frac{\kappa^2}{n^5}\le\frac{1}{2n^4}\frac{\lambda}{p},$$

provided that $n\gg\kappa^2$. □

C.2 Proof of the inequality (5.4b)

Next, we switch attention to the low-rank factors. Our goal is to demonstrate that $(\boldsymbol{X}^{\mathsf{cvx,d}}, \boldsymbol{Y}^{\mathsf{cvx,d}})$ and $(\boldsymbol{X}^{\mathsf{ncvx,d}}, \boldsymbol{Y}^{\mathsf{ncvx,d}})$ are both extremely close to $(\boldsymbol{U}(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r)^{1/2}, \boldsymbol{V}(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r)^{1/2})$ modulo some global rotation, which will be established in (C.19) and (C.20) shortly.

We start by justifying the proximity between $(\boldsymbol{X}^{\mathsf{ncvx,d}}, \boldsymbol{Y}^{\mathsf{ncvx,d}})$ and $(\boldsymbol{U}(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r)^{1/2}, \boldsymbol{V}(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r)^{1/2})$. In view of (C.12), we know that

$$\begin{aligned}
\left\|\boldsymbol{X}^{\mathsf{ncvx}} - \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}\right\| &= \left\|\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top} - \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}\right\| \leq \left\|\boldsymbol{\Sigma}^{1/2}\right\| \left\|\boldsymbol{\Sigma}_{\boldsymbol{Q}} - \boldsymbol{I}_r\right\| \\
&\overset{\text{(i)}}{\lesssim} \sqrt{\sigma_{\max}}\frac{1}{\sigma_{\min}} \left\|\boldsymbol{X}^{\top}\boldsymbol{X} - \boldsymbol{Y}^{\top}\boldsymbol{Y}\right\|_{\mathrm{F}} \\
&\overset{\text{(ii)}}{\lesssim} \sqrt{\sigma_{\max}}\frac{1}{\sigma_{\min}}\frac{1}{n^5}\sigma_{\max}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \overset{\text{(ii)}}{\leq} \frac{1}{n^4}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \cdot \sqrt{\sigma_{\max}}.
\end{aligned} \tag{C.18}$$

Here, (i) depends on the fact that $\|\boldsymbol{\Sigma}_{\boldsymbol{Q}} - \boldsymbol{I}_r\| \lesssim \|\boldsymbol{\Sigma}_{\boldsymbol{Q}} - \boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\|_{\mathrm{F}} \lesssim \|\boldsymbol{X}^{\top}\boldsymbol{X} - \boldsymbol{Y}^{\top}\boldsymbol{Y}\|_{\mathrm{F}}/\sigma_{\min}$ (see [CCF+19, Lemma 20]), (ii) makes use of the balancedness assumption (A.11), whereas (iii) holds if $n \gg \kappa$. Denoting $\tilde{\boldsymbol{X}} \triangleq \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}$, one invokes the triangle inequality to reach

$$\begin{aligned}
&\left\|\boldsymbol{X}^{\mathsf{ncvx,d}} - \tilde{\boldsymbol{X}}\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\tilde{\boldsymbol{X}}^{\top}\tilde{\boldsymbol{X}}\right)^{-1}\right)^{1/2}\right\| \\
&\quad\leq \left\|\boldsymbol{X}^{\mathsf{ncvx}} - \tilde{\boldsymbol{X}}\right\| \left\|\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\right)^{1/2}\right\| \\
&\qquad + \left\|\tilde{\boldsymbol{X}}\right\| \left\|\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\right)^{1/2} - \left(\boldsymbol{I}_r + \frac{\lambda}{p}(\tilde{\boldsymbol{X}}^{\top}\tilde{\boldsymbol{X}})^{-1}\right)^{1/2}\right\| \\
&\quad\leq \frac{1}{n^4}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \cdot \sqrt{\sigma_{\max}}.
\end{aligned}$$

Here the last line arises from (C.18) and the facts $\|\boldsymbol{I}_r + \frac{\lambda}{p}(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}})^{-1}\| \asymp 1$, $\|\tilde{\boldsymbol{X}}\| \lesssim \sqrt{\sigma_{\max}}$ and

$$\left\|\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\mathsf{ncvx}\top}\boldsymbol{X}^{\mathsf{ncvx}}\right)^{-1}\right)^{1/2} - \left(\boldsymbol{I}_r + \frac{\lambda}{p}(\tilde{\boldsymbol{X}}^{\top}\tilde{\boldsymbol{X}})^{-1}\right)^{1/2}\right\| \lesssim \frac{1}{n^4}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}.$$

The latter bound follows from similar derivations as in (C.16). A similar bound holds for $\boldsymbol{Y}^{\mathsf{ncvx,d}}$. Recognizing that

$$\tilde{\boldsymbol{X}}\left(\boldsymbol{I}_r + \frac{\lambda}{p}(\tilde{\boldsymbol{X}}^{\top}\tilde{\boldsymbol{X}})^{-1}\right)^{1/2} = \boldsymbol{U}\left(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r\right)^{1/2}\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top},$$

we have

$$\begin{aligned}
&\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \sqrt{\left\|\boldsymbol{X}^{\mathsf{ncvx,d}}\boldsymbol{R} - \boldsymbol{U}\left(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r\right)^{1/2}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y}^{\mathsf{ncvx,d}}\boldsymbol{R} - \boldsymbol{V}\left(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r\right)^{1/2}\right\|_{\mathrm{F}}^2} \\
&\leq \sqrt{\left\|\boldsymbol{X}^{\mathsf{ncvx,d}} - \boldsymbol{U}\left(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r\right)^{1/2}\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y}^{\mathsf{ncvx,d}} - \boldsymbol{V}\left(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r\right)^{1/2}\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}\right\|_{\mathrm{F}}^2} \\
&\leq \sqrt{r}\sqrt{\left\|\boldsymbol{X}^{\mathsf{ncvx,d}} - \boldsymbol{U}\left(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r\right)^{1/2}\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}\right\|^2 + \left\|\boldsymbol{Y}^{\mathsf{ncvx,d}} - \boldsymbol{V}\left(\boldsymbol{\Sigma} + \frac{\lambda}{p}\boldsymbol{I}_r\right)^{1/2}\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}\right\|^2} \\
&\lesssim \frac{\sqrt{r}}{n^4}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \cdot \sqrt{\sigma_{\max}}.
\end{aligned} \tag{C.19}$$

Next, we establish the connection between $(\boldsymbol{X}^{\mathsf{cvx,d}}, \boldsymbol{Y}^{\mathsf{cvx,d}})$ and $(\boldsymbol{U}(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r)^{1/2}, \boldsymbol{V}(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r)^{1/2})$. To accomplish this, we first study the relationship between $(\boldsymbol{X}^{\mathsf{cvx}}, \boldsymbol{Y}^{\mathsf{cvx}})$ and $(\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}, \boldsymbol{V}\boldsymbol{\Sigma}^{1/2})$. Recall that

$\boldsymbol{X}^{\mathsf{cvx}}$ and $\boldsymbol{Y}^{\mathsf{cvx}}$ constitute a balanced factorization of $\boldsymbol{Z}^{\mathsf{cvx},r}$, while $(\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}, \boldsymbol{V}\boldsymbol{\Sigma}^{1/2})$ is a balanced one of $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$. Hence one can view $\boldsymbol{Z}^{\mathsf{cvx},r}$ as a perturbation of $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$ and investigate the perturbation bounds on the balanced factorizations. Going through the same derivations as in [MWCC17, Appendix B.7] and [CLL19, Appendix B.2.1], one reaches

$$\begin{aligned}\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \sqrt{\left\|\boldsymbol{X}^{\mathsf{cvx}}\boldsymbol{R} - \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\right\|_{\mathrm{F}}^{2} + \left\|\boldsymbol{Y}^{\mathsf{cvx}}\boldsymbol{R} - \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\right\|_{\mathrm{F}}^{2}} &\lesssim \sqrt{r}\cdot\frac{\kappa^{2}}{\sqrt{\sigma_{\min}}}\left\|\boldsymbol{Z}^{\mathsf{cvx},r} - \boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}\right\|_{\mathrm{F}} \\ &\lesssim \sqrt{r}\cdot\frac{\kappa^{4}}{\sqrt{\sigma_{\min}}}\cdot\frac{1}{n^{5}}\frac{\lambda}{p}.\end{aligned}$$

Here the last relation follows from the proximity of the convex estimator and the nonconvex estimator; see (A.12). Repeating the same argument as above to translate the bound between $(\boldsymbol{X}^{\mathsf{cvx}}, \boldsymbol{Y}^{\mathsf{cvx}})$ and $(\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}, \boldsymbol{V}\boldsymbol{\Sigma}^{1/2})$ to that of $(\boldsymbol{X}^{\mathsf{cvx,d}}, \boldsymbol{Y}^{\mathsf{cvx,d}})$ and $(\boldsymbol{U}(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r)^{1/2}, \boldsymbol{V}(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r)^{1/2})$, we conclude that

$$\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \sqrt{\left\|\boldsymbol{X}^{\mathsf{cvx,d}}\boldsymbol{R} - \boldsymbol{U}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)^{1/2}\right\|_{\mathrm{F}}^{2} + \left\|\boldsymbol{Y}^{\mathsf{cvx,d}}\boldsymbol{R} - \boldsymbol{V}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)^{1/2}\right\|_{\mathrm{F}}^{2}} \lesssim \sqrt{r}\cdot\frac{\kappa^{4}}{n^{5}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\cdot\sqrt{\sigma_{\min}}. \tag{C.20}$$

This together with (C.19) and the assumption $n \gg \kappa^4$ concludes the proof.

C.3 Proof of the inequality (5.5)

We shall focus on proving the claim for the nonconvex estimator $\boldsymbol{M}^{\mathsf{ncvx,d}}$ and $\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top}$; the claim for the convex estimator $\boldsymbol{Z}^{\mathsf{cvx}}$ can be treated similarly.

Recall from (C.10) that

$$\left\|\boldsymbol{M}^{\mathsf{ncvx,d}} - \boldsymbol{U}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^{\top}\right\|_{\mathrm{F}} \leq \frac{1}{2n^4}\frac{\lambda}{p}.$$

It then suffices to prove that

$$\left\|\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} - \frac{1}{p}\mathcal{P}_T\mathcal{P}_\Omega\left(\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} - \boldsymbol{M}\right) - \boldsymbol{U}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^{\top}\right\|_{\mathrm{F}} \leq \frac{1}{2n^4}\frac{\lambda}{p}.$$

To see this, it has been established in Appendix C.1 that

$$\begin{aligned}\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} - \frac{1}{p}\mathcal{P}_T\mathcal{P}_\Omega\left(\boldsymbol{X}^{\mathsf{ncvx}}\boldsymbol{Y}^{\mathsf{ncvx}\top} - \boldsymbol{M}\right) &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top} - \frac{1}{p}\mathcal{P}_T\left(-\lambda\boldsymbol{U}\boldsymbol{V}^{\top} + \boldsymbol{R}\right) \\ &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top} + \frac{\lambda}{p}\boldsymbol{U}\boldsymbol{V}^{\top} - \frac{1}{p}\mathcal{P}_T\left(\boldsymbol{R}\right) \\ &= \boldsymbol{U}\Big(\boldsymbol{\Sigma}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)\boldsymbol{V}^{\top} - \frac{1}{p}\mathcal{P}_T\left(\boldsymbol{R}\right).\end{aligned}$$

This together with the fact $\|\mathcal{P}_T(\boldsymbol{R})\|_{\mathrm{F}} \leq \frac{72\kappa}{n^5}\lambda$ (cf. (C.5)) and the assumption $n \gg \kappa$ immediately completes the proof.

D Analysis of the low-rank factors

D.1 Proof of Lemma 4

We concentrate on the factor $\boldsymbol{X}^{\mathsf{d}}$; the other factor $\boldsymbol{Y}^{\mathsf{d}}$ can be treated similarly. By definition of the gradient, one has

$$\nabla_{\boldsymbol{X}} f\left(\boldsymbol{X}, \boldsymbol{Y}\right) = \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M}\right)\boldsymbol{Y} + \frac{\lambda}{p}\boldsymbol{X}. \tag{D.1}$$

Making use of the decomposition

$$\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M}\right) = \boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top} + \boldsymbol{A} - \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right) \tag{D.2}$$

with $\boldsymbol{A}$ defined in (5.7), we can rearrange (D.1) as follows

$$\boldsymbol{X}\Big(\boldsymbol{Y}^\top\boldsymbol{Y}+\frac{\lambda}{p}\boldsymbol{I}_r\Big)=\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}\boldsymbol{Y}+\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{Y}+\nabla_{\boldsymbol{X}}f\left(\boldsymbol{X},\boldsymbol{Y}\right). \tag{D.3}$$

By construction, the de-shrunken estimator $\boldsymbol{Y}^{\mathsf{d}}$ satisfies the following property

$$\begin{aligned}\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}&=\Big(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\Big)^{1/2}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)\Big(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\Big)^{1/2}\\&=\Big(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\Big)^{1/2}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{\frac{1}{2}}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{\frac{1}{2}}\Big(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\Big)^{1/2}\\&=\boldsymbol{Y}^\top\boldsymbol{Y}+\frac{\lambda}{p}\boldsymbol{I}_r,\end{aligned} \tag{D.4}$$

where the last identity follows since $(\boldsymbol{Y}^\top\boldsymbol{Y})^{1/2}$ and $(\boldsymbol{I}_r+\frac{\lambda}{p}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-1})^{1/2}$ commute. Combining (D.3) with the identity (D.4) gives

$$\boldsymbol{X}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)=\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}\boldsymbol{Y}+\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{Y}+\nabla_{\boldsymbol{X}}f\left(\boldsymbol{X},\boldsymbol{Y}\right). \tag{D.5}$$

Multiplying both sides of (D.5) by $(\boldsymbol{I}_r+\frac{\lambda}{p}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-1})^{1/2}$ and recalling the definition of $\boldsymbol{Y}^{\mathsf{d}}$ in (3.8), we have

$$\begin{aligned}&\boldsymbol{X}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)\Big(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\Big)^{1/2}\\&\quad=\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\mathsf{d}}+\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}^{\mathsf{d}}-\boldsymbol{A}\boldsymbol{Y}^{\mathsf{d}}+\nabla_{\boldsymbol{X}}f\left(\boldsymbol{X},\boldsymbol{Y}\right)\Big(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\Big)^{1/2}.\end{aligned} \tag{D.6}$$

Since $\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}$ and $(\boldsymbol{I}_r+\frac{\lambda}{p}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-1})^{1/2}$ also commute, we have

$$\begin{aligned}\boldsymbol{X}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)\Big(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\Big)^{1/2}&=\boldsymbol{X}\Big(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\Big)^{1/2}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)\\&=\boldsymbol{X}\Big(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\Big)^{1/2}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)-\boldsymbol{X}\boldsymbol{\Delta}_{\mathsf{balancing}}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)\\&=\boldsymbol{X}^{\mathsf{d}}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)-\boldsymbol{X}\boldsymbol{\Delta}_{\mathsf{balancing}}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right),\end{aligned} \tag{D.7}$$

where the last relation uses the definition of $\boldsymbol{X}^{\mathsf{d}}$ (see (3.8)).

Substituting the identity (D.7) back into (D.6) and making a few elementary algebraic manipulations yield the desired decomposition (5.8a).

D.2 Proof of Lemma 5

Recall that $\overline{\boldsymbol{Y}}^{\mathsf{d}}=\boldsymbol{Y}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}}$ and similarly define

$$\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\triangleq\boldsymbol{Y}^{\mathsf{d},(j)}\boldsymbol{H}^{\mathsf{d},(j)}.$$

The triangle inequality tells us that for any fixed $1\le j\le n$,

$$\begin{aligned}\left\|\boldsymbol{e}_j^\top\boldsymbol{\Phi}_1\right\|_2\le&\underbrace{\left\|\boldsymbol{e}_j^\top\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\left[\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\left(\overline{\boldsymbol{Y}}^{\mathsf{d},(j)\top}\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\right)^{-1}-\boldsymbol{Y}^\star\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1}\right]\right\|_2}_{:=\alpha_1}\\&+\underbrace{\left\|\boldsymbol{e}_j^\top\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\left[\overline{\boldsymbol{Y}}^{\mathsf{d}}\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1}-\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\left(\overline{\boldsymbol{Y}}^{\mathsf{d},(j)\top}\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\right)^{-1}\right]\right\|_2}_{:=\alpha_2}.\end{aligned}$$

In what follows, we shall control α_1 and α_2 separately.

1. To begin with, denoting $\boldsymbol{\Delta}^{(j)} \triangleq \overline{\boldsymbol{Y}}^{\mathsf{d},(j)}(\overline{\boldsymbol{Y}}^{\mathsf{d},(j)\top}\overline{\boldsymbol{Y}}^{\mathsf{d},(j)})^{-1} - \boldsymbol{Y}^{\star}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}$ results in

$$\alpha_1 = \left\| \boldsymbol{e}_j^{\top} \frac{1}{p} \mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \boldsymbol{\Delta}^{(j)} \right\|_2 = \left\| \frac{1}{p} \sum_{k=1}^{n} E_{jk}\delta_{jk}\boldsymbol{\Delta}_{k,\cdot}^{(j)} \right\|_2. \tag{D.8}$$

Before proceeding, we gather a few useful facts regarding $\boldsymbol{\Delta}^{(j)}$, as summarized in the following claim.

Claim 2. *With probability at least* $1 - O(n^{-11})$, *we have*

$$\begin{aligned} \|\boldsymbol{\Delta}^{(j)}\| &\lesssim \frac{1}{\sqrt{\sigma_{\min}}} \cdot \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\kappa^3 n}{p}}, \\ \|\boldsymbol{\Delta}^{(j)}\|_{2,\infty} &\lesssim \frac{1}{\sqrt{\sigma_{\min}}} \cdot \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\kappa^5 \mu r \log n}{p}}. \end{aligned}$$

With the bounds on $\|\boldsymbol{\Delta}^{(j)}\|$ and $\|\boldsymbol{\Delta}^{(j)}\|_{2,\infty}$ in place, we are ready to control α_1. By construction, $\boldsymbol{\Delta}^{(j)}$ is independent of $\boldsymbol{e}_j^{\top}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)$. Therefore, the vector on the right-hand side of (D.8), $\frac{1}{p}\sum_{k=1}^{n} E_{jk}\delta_{jk}\boldsymbol{\Delta}_{k,\cdot}^{(j)}$, is a sum of conditionally independent random vectors. In particular, conditional on $\boldsymbol{\Delta}^{(j)}$ and $\{\delta_{jk}\}_{k:1\leq k\leq n}$, one has

$$\frac{1}{p} \sum_{k=1}^{n} E_{jk}\delta_{jk}\boldsymbol{\Delta}_{k,\cdot}^{(j)} \,\Big|\, \boldsymbol{\Delta}^{(j)}, \{\delta_{jk}\}_{k:1\leq k\leq n} \;\sim\; \mathcal{N}\Big(\boldsymbol{0}, \underbrace{\frac{\sigma^2}{p^2}\sum_{k=1}^{n}\delta_{jk}\boldsymbol{\Delta}_{k,\cdot}^{(j)\top}\boldsymbol{\Delta}_{k,\cdot}^{(j)}}_{:=\hat{\boldsymbol{\Sigma}}}\Big). \tag{D.9}$$

Invoke the concentration inequality for Gaussian random vectors [HKZ12, Proposition 1.1] to see that

$$\begin{aligned} \alpha_1 &\leq \sqrt{\mathsf{Tr}(\hat{\boldsymbol{\Sigma}}) + 2\sqrt{t}\|\hat{\boldsymbol{\Sigma}}\|_{\mathrm{F}} + 2\|\hat{\boldsymbol{\Sigma}}\|t} \leq \sqrt{r\|\hat{\boldsymbol{\Sigma}}\| + 2\sqrt{rt}\|\hat{\boldsymbol{\Sigma}}\| + 2\|\hat{\boldsymbol{\Sigma}}\|t} \\ &\lesssim \sqrt{\|\hat{\boldsymbol{\Sigma}}\|}\left(\sqrt{r} + \sqrt{t}\right) \end{aligned} \tag{D.10}$$

with probability at least $1 - e^{-t}$. It remains to control $\|\hat{\boldsymbol{\Sigma}}\|$, which we state in the following claim.

Claim 3. *Suppose that* $n^2 p \gg \kappa^2 \mu r n \log^2 n$. *Then with probability exceeding* $1 - O(n^{-11})$,

$$\|\hat{\boldsymbol{\Sigma}}\| \lesssim \frac{\sigma^2}{p}\left(\frac{1}{\sqrt{\sigma_{\min}}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^3 n}{p}}\right)^2.$$

Combine the upper bound on $\|\hat{\boldsymbol{\Sigma}}\|$ with (D.10) and choose $t \asymp \log n$ to arrive at

$$\alpha_1 \lesssim \sqrt{\|\hat{\boldsymbol{\Sigma}}\|}\left(\sqrt{r} + \sqrt{\log n}\right) \lesssim \frac{\sigma}{\sqrt{p}} \frac{1}{\sqrt{\sigma_{\min}}} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\kappa^3 r n \log n}{p}}$$

with probability exceeding $1 - O(n^{-11})$.

2. We move on to bounding α_2, for which we have

$$\begin{aligned} \alpha_2 &\leq \frac{1}{p}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right\| \left\| \overline{\boldsymbol{Y}}^{\mathsf{d}}\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1} - \overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\left(\overline{\boldsymbol{Y}}^{\mathsf{d},(j)\top}\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\right)^{-1} \right\| \\ &\overset{\text{(i)}}{\lesssim} \sigma\sqrt{\frac{n}{p}}\frac{1}{\sigma_{\min}}\left\|\overline{\boldsymbol{Y}}^{\mathsf{d}} - \overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\right\| \\ &\overset{\text{(ii)}}{\lesssim} \sigma\sqrt{\frac{n}{p}}\frac{1}{\sigma_{\min}}\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{Y}^{\star}\right\|_{2,\infty} \end{aligned} \tag{D.11}$$

$$\lesssim \sigma\sqrt{\frac{n}{p}}\frac{1}{\sqrt{\sigma_{\min}}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^3\mu r\log n}{p}}.$$

Here (i) uses the fact that $\|\mathcal{P}_\Omega(\boldsymbol{E})\| \lesssim \sigma\sqrt{np}$ (see [CCF+19, Lemma 3]), the perturbation bounds for pseudo-inverses (see Lemma 12) and (A.19); the penultimate inequality (ii) comes from the fact that $\|\overline{\boldsymbol{Y}}^{\mathsf{d}} - \overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\| \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\|\boldsymbol{Y}^\star\|_{2,\infty}$ (see (A.16c)) and last one uses the incoherence condition $\|\boldsymbol{Y}^\star\|_{2,\infty} \leq \sqrt{\mu r\sigma_{\max}/n}$ (cf. (A.17)).

Combine the bounds on α_1 and α_2 to reach

$$\begin{aligned}\left\|\boldsymbol{e}_j^\top\boldsymbol{\Phi}_1\right\|_2 &\lesssim \frac{\sigma}{\sqrt{p}}\frac{1}{\sqrt{\sigma_{\min}}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^3 rn\log n}{p}} + \sigma\sqrt{\frac{n}{p}}\frac{1}{\sqrt{\sigma_{\min}}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^3\mu r\log n}{p}} \\ &\lesssim \frac{\sigma}{\sqrt{p\sigma_{\min}}}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^3\mu rn\log n}{p}}. \end{aligned} \tag{D.12}$$

Taking the maximum over $1 \leq j \leq n$ establishes our bound on $\|\boldsymbol{\Phi}_1\|_{2,\infty}$.

Proof of Claim 2. Apply the perturbation bound for pseudo-inverses (see Lemma 12) to obtain

$$\begin{aligned}\left\|\boldsymbol{\Delta}^{(j)}\right\| &\lesssim \max\left\{\left\|\boldsymbol{Y}^\star\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1}\right\|^2, \left\|\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\left(\overline{\boldsymbol{Y}}^{\mathsf{d},(j)\top}\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\right)^{-1}\right\|^2\right\}\left\|\overline{\boldsymbol{Y}}^{\mathsf{d},(j)} - \boldsymbol{Y}^\star\right\| \\ &\lesssim \frac{1}{\sigma_{\min}}\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\| \asymp \frac{1}{\sqrt{\sigma_{\min}}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^3 n}{p}}, \end{aligned} \tag{D.13}$$

Here we have utilized the facts that $\|\overline{\boldsymbol{Y}}^{\mathsf{d},(j)} - \boldsymbol{Y}^\star\| \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\|\boldsymbol{X}^\star\|$ (see (A.16a)) and a simple consequence of (A.19), viz.

$$\max\left\{\left\|\boldsymbol{Y}^\star\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1}\right\|^2, \left\|\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\left(\overline{\boldsymbol{Y}}^{\mathsf{d},(j)\top}\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\right)^{-1}\right\|^2\right\} \lesssim \frac{1}{\sigma_{\min}}.$$

Moreover, the triangle inequality tells us that

$$\begin{aligned}\left\|\boldsymbol{\Delta}^{(j)}\right\|_{2,\infty} &\leq \left\|\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\left[\left(\overline{\boldsymbol{Y}}^{\mathsf{d},(j)\top}\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\right)^{-1} - \left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1}\right]\right\|_{2,\infty} + \left\|\left(\overline{\boldsymbol{Y}}^{\mathsf{d},(j)} - \boldsymbol{Y}^\star\right)\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1}\right\|_{2,\infty} \\ &\leq \left\|\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\right\|_{2,\infty}\left\|\left(\overline{\boldsymbol{Y}}^{\mathsf{d},(j)\top}\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\right)^{-1} - \left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1}\right\| + \left\|\overline{\boldsymbol{Y}}^{\mathsf{d},(j)} - \boldsymbol{Y}^\star\right\|_{2,\infty}\left\|\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1}\right\| \\ &\lesssim \frac{1}{\sigma_{\min}}\kappa^2\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{F}^\star\right\|_{2,\infty} + \frac{1}{\sigma_{\min}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^2 n\log n}{p}}\left\|\boldsymbol{F}^\star\right\|_{2,\infty} \\ &\lesssim \frac{1}{\sqrt{\sigma_{\min}}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^5\mu r\log n}{p}}, \end{aligned}$$

where the penultimate inequality follows from the facts that $\|\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\|_{2,\infty} \leq 2\|\boldsymbol{F}^\star\|_{2,\infty}$, $\|\overline{\boldsymbol{Y}}^{\mathsf{d},(j)} - \boldsymbol{Y}^\star\|_{2,\infty} \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\|\boldsymbol{F}^\star\|_{2,\infty}$ (see (A.16b)) and that

$$\begin{aligned}\left\|\left(\overline{\boldsymbol{Y}}^{\mathsf{d},(j)\top}\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\right)^{-1} - \left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1}\right\| &\leq \left\|\left(\overline{\boldsymbol{Y}}^{\mathsf{d},(j)\top}\overline{\boldsymbol{Y}}^{\mathsf{d},(j)}\right)^{-1}\right\|\left\|\overline{\boldsymbol{Y}}^{\mathsf{d},(j)\top}\overline{\boldsymbol{Y}}^{\mathsf{d},(j)} - \boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right\|\left\|\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right)^{-1}\right\| \\ &\lesssim \frac{1}{\sigma_{\min}^2}\left\|\boldsymbol{F}^{\mathsf{d},(j)}\boldsymbol{H}^{\mathsf{d},(j)} - \boldsymbol{F}^\star\right\|\left\|\boldsymbol{F}^\star\right\| \\ &\lesssim \frac{1}{\sigma_{\min}}\kappa^2\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}. \end{aligned}$$

Here the penultimate inequality follows from (A.19). The proof of the claim is then complete. □

Proof of Claim 3. Conditional on $\boldsymbol{\Delta}^{(j)}$, using Bernstein's inequality and the fact that $\boldsymbol{\Delta}^{(j)}$ and $\{\delta_{jk}\}_{k:1\leq k\leq n}$ are independent, we arrive at that with probability exceeding $1-O(n^{-11})$,

$$\left\|\hat{\boldsymbol{\Sigma}}-\frac{\sigma^2}{p}\boldsymbol{\Delta}^{(j)\top}\boldsymbol{\Delta}^{(j)}\right\|\lesssim\frac{\sigma^2}{p^2}\left(\sqrt{V\log n}+B\log n\right),$$

where

$$\begin{aligned}
B &\triangleq \max_{1\leq k\leq n}\left\|\left(\delta_{jk}-p\right)\boldsymbol{\Delta}_{k,\cdot}^{(j)\top}\boldsymbol{\Delta}_{k,\cdot}^{(j)}\right\|\leq\left\|\boldsymbol{\Delta}^{(j)}\right\|_{2,\infty}^2,\\
V &\triangleq \left\|\sum_{k=1}^{n}\mathbb{E}\left(\delta_{jk}-p\right)^2\boldsymbol{\Delta}_{k,\cdot}^{(j)\top}\boldsymbol{\Delta}_{k,\cdot}^{(j)}\boldsymbol{\Delta}_{k,\cdot}^{(j)\top}\boldsymbol{\Delta}_{k,\cdot}^{(j)}\right\|\leq p\left\|\boldsymbol{\Delta}^{(j)}\right\|_{2,\infty}^2\left\|\boldsymbol{\Delta}^{(j)}\right\|^2.
\end{aligned}$$

As a result, with probability at least $1-O(n^{-11})$, we have

$$\begin{aligned}
\left\|\hat{\boldsymbol{\Sigma}}-\frac{\sigma^2}{p}\boldsymbol{\Delta}^{(j)\top}\boldsymbol{\Delta}^{(j)}\right\| &\lesssim\frac{\sigma^2}{p^2}\left\|\boldsymbol{\Delta}^{(j)}\right\|_{2,\infty}\left(\sqrt{p\log n}\left\|\boldsymbol{\Delta}^{(j)}\right\|+\left\|\boldsymbol{\Delta}^{(j)}\right\|_{2,\infty}\log n\right)\\
&\lesssim\frac{\sigma^2}{p^2}\left\|\boldsymbol{\Delta}^{(j)}\right\|_{2,\infty}\left(\sqrt{p\log n}\frac{1}{\sqrt{\sigma_{\min}}}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^3 n}{p}}+\frac{1}{\sqrt{\sigma_{\min}}}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^5\mu r\log n}{p}}\log n\right)\\
&\lesssim\frac{\sigma^2}{p^2}\left\|\boldsymbol{\Delta}^{(j)}\right\|_{2,\infty}\frac{1}{\sqrt{\sigma_{\min}}}\frac{\sigma}{\sigma_{\min}}\sqrt{\kappa^3 n\log n},
\end{aligned}$$

as long as $np\gg\kappa^2\mu r\log^2 n$. Here the middle inequality uses Claim 2. In view of the triangle inequality,

$$\begin{aligned}
\left\|\hat{\boldsymbol{\Sigma}}\right\| &\leq\left\|\frac{\sigma^2}{p}\boldsymbol{\Delta}^{(j)\top}\boldsymbol{\Delta}^{(j)}\right\|+O\left(\frac{\sigma^2}{p^2}\left\|\boldsymbol{\Delta}^{(j)}\right\|_{2,\infty}\frac{1}{\sqrt{\sigma_{\min}}}\frac{\sigma}{\sigma_{\min}}\sqrt{\kappa^3 n\log n}\right)\\
&\lesssim\frac{\sigma^2}{p}\left(\left\|\boldsymbol{\Delta}^{(j)}\right\|^2+\frac{1}{p}\left\|\boldsymbol{\Delta}^{(j)}\right\|_{2,\infty}\frac{1}{\sqrt{\sigma_{\min}}}\frac{\sigma}{\sigma_{\min}}\sqrt{\kappa^3 n\log n}\right)\\
&\lesssim\frac{\sigma^2}{p}\left(\left\|\boldsymbol{\Delta}^{(j)}\right\|^2+\frac{1}{p}\left\|\boldsymbol{\Delta}^{(j)}\right\|_{2,\infty}\frac{1}{\sqrt{\sigma_{\min}}}\frac{\sigma}{\sigma_{\min}}\sqrt{\kappa^3 n\log n}\right)\\
&\lesssim\frac{\sigma^2}{p}\left(\frac{1}{\sqrt{\sigma_{\min}}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^3 n}{p}}\right)^2,
\end{aligned}$$

with the proviso that $n^2p\gg\kappa^2\mu rn\log^2 n$. Again, the last line makes use of Claim 2. This concludes the proof of the claim. □

D.3 Proof of Lemma 6

Recall that $\overline{\boldsymbol{Y}}^{\mathsf{d}}=\boldsymbol{Y}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}}$. The sub-multiplicativity of the operator norm gives that for any $1\leq j\leq n$,

$$\begin{aligned}
\left\|\boldsymbol{e}_j^\top\boldsymbol{\Phi}_2\right\|_2 &=\left\|\boldsymbol{e}_j^\top\boldsymbol{X}^\star\left[\boldsymbol{Y}^{\star\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1}-\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1}\right]\right\|_2\\
&\leq\left\|\boldsymbol{e}_j^\top\boldsymbol{X}^\star\right\|_2\left\|\left(\boldsymbol{Y}^\star-\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^\top\overline{\boldsymbol{Y}}^{\mathsf{d}}\right\|\left\|\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1}\right\| \qquad\qquad (\text{D.14})\\
&\lesssim\sqrt{\frac{\mu r\sigma_{\max}}{n}}\frac{1}{\sigma_{\min}}\left\|\left(\boldsymbol{Y}^\star-\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^\top\overline{\boldsymbol{Y}}^{\mathsf{d}}\right\|\\
&\asymp\sqrt{\frac{\kappa\mu r}{n}}\frac{1}{\sqrt{\sigma_{\min}}}\left\|\left(\boldsymbol{Y}^\star-\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^\top\overline{\boldsymbol{Y}}^{\mathsf{d}}\right\|, \qquad\qquad (\text{D.15})
\end{aligned}$$

where the second inequality follows from the incoherence assumption that $\|\boldsymbol{e}_j^\top\boldsymbol{X}^\star\|_2\leq\|\boldsymbol{X}^\star\|_{2,\infty}\leq\sqrt{\mu r\sigma_{\max}/n}$ (cf. (A.17)) and the fact that $\|(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}})^{-1}\|\lesssim 1/\sigma_{\min}$, a simple consequence of (A.19).

It remains to control $\|(\overline{\boldsymbol{Y}}^{\mathsf{d}} - \boldsymbol{Y}^{\star})^{\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\|$. To simplify notation hereafter, define $\boldsymbol{\Delta}_{\boldsymbol{X}} \triangleq \overline{\boldsymbol{X}}^{\mathsf{d}} - \boldsymbol{X}^{\star}$ and $\boldsymbol{\Delta}_{\boldsymbol{Y}} \triangleq \overline{\boldsymbol{Y}}^{\mathsf{d}} - \boldsymbol{Y}^{\star}$. First, observe that

$$\big(\boldsymbol{Y}^{\star} - \overline{\boldsymbol{Y}}^{\mathsf{d}}\big)^{\top}\overline{\boldsymbol{Y}}^{\mathsf{d}} = \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y}^{\star} + \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{Y}}. \tag{D.16}$$

Second, in view of the decomposition of $\boldsymbol{Y}^{\mathsf{d}}$ given in (5.8b), we have

$$\begin{aligned}\overline{\boldsymbol{Y}}^{\mathsf{d}} &= \boldsymbol{Y}^{\star}\boldsymbol{X}^{\star\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1} + \frac{1}{p}\left[\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right]^{\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1} - \boldsymbol{A}^{\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1} \\ &\quad + \nabla_{\boldsymbol{Y}}f\left(\boldsymbol{X},\boldsymbol{Y}\right)\Big(\boldsymbol{I}_r + \frac{\lambda}{p}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\Big)^{1/2}(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}})^{-1}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{Y}\boldsymbol{\Delta}_{\mathsf{balancing}}\boldsymbol{H}^{\mathsf{d}}. \end{aligned} \tag{D.17}$$

As a result, one obtains

$$\begin{aligned}\boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y}^{\star} &= \left\{\boldsymbol{Y}^{\star}\left(\boldsymbol{X}^{\star\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1} - \boldsymbol{I}_r\right) + \frac{1}{p}\left[\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right]^{\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\left(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\right)^{-1} - \boldsymbol{A}^{\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\left(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\right)^{-1}\right\}^{\top}\boldsymbol{Y}^{\star} \\ &\quad + \left\{\nabla_{\boldsymbol{Y}}f\left(\boldsymbol{X},\boldsymbol{Y}\right)\Big(\boldsymbol{I}_r + \frac{\lambda}{p}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\Big)^{1/2}(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}})^{-1}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{Y}\boldsymbol{\Delta}_{\mathsf{balancing}}\boldsymbol{H}^{\mathsf{d}}\right\}^{\top}\boldsymbol{Y}^{\star} \\ &= -\big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1}\overline{\boldsymbol{X}}^{\mathsf{d}\top}\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Sigma}^{\star} + \big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1}\overline{\boldsymbol{X}}^{\mathsf{d}\top}\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{Y}^{\star} - \big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1}\overline{\boldsymbol{X}}^{\mathsf{d}\top}\boldsymbol{A}\boldsymbol{Y}^{\star} \\ &\quad + \left\{\nabla_{\boldsymbol{Y}}f\left(\boldsymbol{X},\boldsymbol{Y}\right)\Big(\boldsymbol{I}_r + \frac{\lambda}{p}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\Big)^{1/2}(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}})^{-1}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{Y}\boldsymbol{\Delta}_{\mathsf{balancing}}\boldsymbol{H}^{\mathsf{d}}\right\}^{\top}\boldsymbol{Y}^{\star} \\ &= -\big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1}\boldsymbol{X}^{\star\top}\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Sigma}^{\star} + \boldsymbol{S}, \end{aligned} \tag{D.18}$$

where we have used

$$\boldsymbol{X}^{\star\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1} - \boldsymbol{I}_r = \boldsymbol{X}^{\star\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1} - \overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1} = -\boldsymbol{\Delta}_{\boldsymbol{X}}^{\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1}.$$

Here, we define $\boldsymbol{S}$ to be

$$\begin{aligned}\boldsymbol{S} &\triangleq -\big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1}\boldsymbol{\Delta}_{\boldsymbol{X}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Sigma}^{\star} + \big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1}\overline{\boldsymbol{X}}^{\mathsf{d}\top}\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{Y}^{\star} - \big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1}\overline{\boldsymbol{X}}^{\mathsf{d}\top}\boldsymbol{A}\boldsymbol{Y}^{\star} \\ &\quad + \left\{\nabla_{\boldsymbol{Y}}f\left(\boldsymbol{X},\boldsymbol{Y}\right)\Big(\boldsymbol{I}_r + \frac{\lambda}{p}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\Big)^{1/2}(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}})^{-1}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{Y}\boldsymbol{\Delta}_{\mathsf{balancing}}\boldsymbol{H}^{\mathsf{d}}\right\}^{\top}\boldsymbol{Y}^{\star}. \end{aligned} \tag{D.19}$$

The following claim connects $\boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y}^{\star}$ with $\boldsymbol{X}^{\star\top}\boldsymbol{\Delta}_{\boldsymbol{X}}$.

Claim 4. *The following identity holds true:*

$$\boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y}^{\star} - \boldsymbol{X}^{\star\top}\boldsymbol{\Delta}_{\boldsymbol{X}} = \frac{1}{2}\left(\boldsymbol{\Delta}_{\boldsymbol{X}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{X}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{Y}}\right) + \underbrace{\frac{1}{2}\boldsymbol{H}^{\mathsf{d}\top}(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}} - \boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}})\boldsymbol{H}^{\mathsf{d}}}_{:=\boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}}^{\mathsf{d}}}.$$

This relation together with (D.18) yields

$$\boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y}^{\star} = -\big(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\big)^{-1}\left[\boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y}^{\star} - \frac{1}{2}\left(\boldsymbol{\Delta}_{\boldsymbol{X}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{X}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{Y}}\right) - \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}}^{\mathsf{d}}\right]\boldsymbol{\Sigma}^{\star} + \boldsymbol{S}.$$

A little algebraic manipulation then gives

$$\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y}^{\star} + \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y}^{\star}\boldsymbol{\Sigma}^{\star} = \overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\boldsymbol{S} + \frac{1}{2}\left(\boldsymbol{\Delta}_{\boldsymbol{X}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{X}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{Y}}\right)\boldsymbol{\Sigma}^{\star} + \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}}^{\mathsf{d}}\boldsymbol{\Sigma}^{\star}.$$

It is easy to check from (A.19) that $0.25\sigma_{\min}\boldsymbol{I}_r \preceq \overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}, \boldsymbol{\Sigma}^{\star} \preceq 4\sigma_{\max}\boldsymbol{I}_r$. Hence one can invoke Lemma 15 with $\boldsymbol{X} = \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y}^{\star}$, $\boldsymbol{A} = \boldsymbol{\Sigma}^{\star}$, $\boldsymbol{B} = \overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}$ and $\boldsymbol{C} = \overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\boldsymbol{S} + 0.5(\boldsymbol{\Delta}_{\boldsymbol{X}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{X}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{Y}})\boldsymbol{\Sigma}^{\star} + \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}}^{\mathsf{d}}\boldsymbol{\Sigma}^{\star}$ to obtain

$$\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{Y}^{\star}\| \lesssim \frac{1}{\sigma_{\min}}\left\|\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\boldsymbol{S} + \frac{1}{2}\left(\boldsymbol{\Delta}_{\boldsymbol{X}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{X}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{Y}}\right)\boldsymbol{\Sigma}^{\star} + \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}}^{\mathsf{d}}\boldsymbol{\Sigma}^{\star}\right\|$$

$$
\begin{aligned}
&\le \frac{1}{\sigma_{\min}} \left\| \overline{\boldsymbol{X}}^{\mathsf{d}\top} \frac{1}{p} \mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \boldsymbol{Y}^{\star} - \overline{\boldsymbol{X}}^{\mathsf{d}\top} \boldsymbol{A} \boldsymbol{Y}^{\star} - \frac{1}{2}\left(\boldsymbol{\Delta}_{\boldsymbol{X}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{X}} + \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top}\boldsymbol{\Delta}_{\boldsymbol{Y}}\right)\boldsymbol{\Sigma}^{\star} \right\| \\
&\quad + \frac{1}{\sigma_{\min}} \left\| \boldsymbol{H}^{\mathsf{d}\top} \left(\boldsymbol{I}_r + \frac{\lambda}{p}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\right)^{1/2} \left[\nabla_{\boldsymbol{Y}} f\left(\boldsymbol{X},\boldsymbol{Y}\right)\right]^{\top} \boldsymbol{Y}^{\star} + \overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}\top} \boldsymbol{\Delta}_{\mathsf{balancing}} \boldsymbol{Y}^{\top}\boldsymbol{Y}^{\star} + \boldsymbol{\Delta}_{\boldsymbol{XY}}^{\mathsf{d}} \boldsymbol{\Sigma}^{\star} \right\|,
\end{aligned}
$$

where we have plugged in the definition of $\boldsymbol{S}$ (see (D.19)) and used the identity $\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}\top} (\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}})^{-1} = \boldsymbol{H}^{\mathsf{d}\top}$. Combine the above inequality with (D.16) to obtain

$$
\begin{aligned}
&\left\| \left(\overline{\boldsymbol{Y}}^{\mathsf{d}} - \boldsymbol{Y}^{\star}\right)^{\top} \overline{\boldsymbol{Y}}^{\mathsf{d}} \right\| \le \left\| \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} \boldsymbol{Y}^{\star} \right\| + \left\| \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{Y}} \right\| \\
&\quad \lesssim \frac{1}{\sigma_{\min}} \underbrace{\left\| \overline{\boldsymbol{X}}^{\mathsf{d}\top} \frac{1}{p} \mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \boldsymbol{Y}^{\star} \right\|}_{:=\alpha_1} + \frac{1}{\sigma_{\min}} \underbrace{\left\| \overline{\boldsymbol{X}}^{\mathsf{d}\top} \boldsymbol{A} \boldsymbol{Y}^{\star} \right\|}_{:=\alpha_2} + \kappa \underbrace{\left(\left\| \boldsymbol{\Delta}_{\boldsymbol{X}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{X}} \right\| + \left\| \boldsymbol{\Delta}_{\boldsymbol{Y}}^{\top} \boldsymbol{\Delta}_{\boldsymbol{Y}} \right\| \right)}_{:=\alpha_3} \\
&\qquad + \frac{1}{\sigma_{\min}} \underbrace{\left\| \boldsymbol{H}^{\mathsf{d}\top} \left(\boldsymbol{I}_r + \frac{\lambda}{p}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\right)^{1/2} \left[\nabla_{\boldsymbol{Y}} f\left(\boldsymbol{X},\boldsymbol{Y}\right)\right]^{\top} \boldsymbol{Y}^{\star} - \overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}\top} \boldsymbol{\Delta}_{\mathsf{balancing}} \boldsymbol{Y}^{\top}\boldsymbol{Y}^{\star} + \boldsymbol{\Delta}_{\boldsymbol{XY}}^{\mathsf{d}} \boldsymbol{\Sigma}^{\star} \right\|}_{:=\alpha_4}.
\end{aligned}
\tag{D.20}
$$

It then boils down to controlling the above terms $\alpha_1, \alpha_2, \alpha_3$ and α_4.

1. First, the term α_4 can be upper bounded by

$$
\begin{aligned}
\alpha_4 &\le \left\| \left(\boldsymbol{I}_r + \frac{\lambda}{p}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\right)^{1/2} \right\| \left\| \nabla_{\boldsymbol{Y}} f\left(\boldsymbol{X},\boldsymbol{Y}\right) \right\|_{\mathrm{F}} \left\| \boldsymbol{Y}^{\star} \right\| + \left\| \overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}} \right\| \left\| \boldsymbol{\Delta}_{\mathsf{balancing}} \right\| \left\| \boldsymbol{Y}^{\top}\boldsymbol{Y}^{\star} \right\| + \left\| \boldsymbol{\Delta}_{\boldsymbol{XY}}^{\mathsf{d}} \right\| \left\| \boldsymbol{\Sigma}^{\star} \right\| \\
&\lesssim \frac{1}{n^5}\frac{\lambda}{p}\sqrt{\sigma_{\min}}\sqrt{\sigma_{\max}} + \sigma_{\max}^2 \frac{1}{n^5}\frac{\lambda}{p}\frac{\kappa}{\sigma_{\min}} + \sigma_{\max}\frac{\kappa}{n^5}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\sigma_{\max} \\
&\asymp \frac{\kappa}{n^5}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\sigma_{\max}^2.
\end{aligned}
$$

Here, the second line utilizes the facts that $\|(\boldsymbol{I}_r + \lambda(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}/p)^{1/2}\| \lesssim 1$, $\|\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\| \asymp \|\boldsymbol{Y}^{\top}\boldsymbol{Y}^{\star}\| \lesssim \sigma_{\max}$ and the results in (A.10), (A.13e) and (C.16).

2. Moving on to α_3, we recall from (A.13b) that

$$
\max\left\{ \left\| \boldsymbol{\Delta}_{\boldsymbol{X}} \right\|, \left\| \boldsymbol{\Delta}_{\boldsymbol{Y}} \right\| \right\} \lesssim \kappa \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\| \boldsymbol{X}^{\star} \right\|.
$$

Therefore one arrives at

$$
\alpha_3 \lesssim \left(\kappa \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \right)^2 \sigma_{\max}.
$$

3. Regarding the term α_2, we have

$$
\alpha_2 \lesssim \left\| \overline{\boldsymbol{X}}^{\mathsf{d}} \right\| \left\| \boldsymbol{A} \right\| \left\| \boldsymbol{X}^{\star} \right\| \lesssim \sigma_{\max} \sigma \sqrt{\frac{n}{p}} \cdot \sqrt{\frac{\kappa^4 \mu^2 r^2 \log n}{np}},
$$

where we utilize the bound in (A.18)

$$
\left\| \boldsymbol{A} \right\| \lesssim \sigma \sqrt{\frac{n}{p}} \cdot \sqrt{\frac{\kappa^4 \mu^2 r^2 \log n}{np}}.
$$

4. Finally, for the term α_1, by the triangle inequality one has

$$
\alpha_1 \le \left\| \boldsymbol{X}^{\star\top} \frac{1}{p} \mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \boldsymbol{Y}^{\star} \right\| + \left\| \boldsymbol{\Delta}_{\boldsymbol{X}}^{\top} \frac{1}{p} \mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \boldsymbol{Y}^{\star} \right\|.
$$

Note that

$$\left\|\boldsymbol{X}^{\star\top}\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{Y}^{\star}\right\| \leq \left\|\boldsymbol{X}^{\star\top}\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{Y}^{\star}\right\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{r}\left|\left(\boldsymbol{X}_{\cdot,i}^{\star}\right)^{\top}\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{Y}_{\cdot,j}^{\star}\right|^{2}}, \tag{D.21}$$

Observe that conditional on $\{\delta_{jk}\}_{1\leq j,k\leq n}$ one has

$$\left(\boldsymbol{X}_{\cdot,i}^{\star}\right)^{\top}\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{Y}_{\cdot,j}^{\star} = \left\langle \boldsymbol{E},\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}_{\cdot,i}^{\star}\left(\boldsymbol{Y}_{\cdot,j}^{\star}\right)^{\top}\right)\right\rangle \sim\mathcal{N}\left(0,\sigma^{2}\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}_{\cdot,i}^{\star}\left(\boldsymbol{Y}_{\cdot,j}^{\star}\right)^{\top}\right)\right\|_{\mathrm{F}}^{2}\right)$$

As a result, we obtain that with probability at least $1-O(n^{-10})$,

$$\begin{aligned}\left|\left(\boldsymbol{X}_{\cdot,i}^{\star}\right)^{\top}\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{Y}_{\cdot,j}^{\star}\right| &\lesssim \sigma\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}_{\cdot,i}^{\star}\left(\boldsymbol{Y}_{\cdot,j}^{\star}\right)^{\top}\right)\right\|_{\mathrm{F}}\sqrt{\log n}\\ &\lesssim \sigma\sqrt{\frac{\log n}{p}}\left\|\boldsymbol{X}_{\cdot,i}^{\star}\left(\boldsymbol{Y}_{\cdot,j}^{\star}\right)^{\top}\right\|_{\mathrm{F}}.\end{aligned} \tag{D.22}$$

Here, the second relation uses the fact that

$$\left\|\frac{1}{\sqrt{p}}\mathcal{P}_{\Omega}\left(\boldsymbol{X}_{\cdot,i}^{\star}\left(\boldsymbol{Y}_{\cdot,j}^{\star}\right)^{\top}\right)\right\|_{\mathrm{F}} \asymp \|\boldsymbol{X}_{\cdot,i}^{\star}\left(\boldsymbol{Y}_{\cdot,j}^{\star}\right)^{\top}\|_{\mathrm{F}}$$

with probability at least $1-O(n^{-10})$ as long as $n^2p \gg \mu rn\log n$, which follows from [MWCC17, Lemma 38] or [CR09, Section 4.2] by observing that $\boldsymbol{X}_{\cdot,i}^{\star}(\boldsymbol{Y}_{\cdot,j}^{\star})^{\top}$ lies in the tangent space of $\boldsymbol{M}^{\star}$. Take (D.21) and (D.22) collectively to reach

$$\left\|\boldsymbol{X}^{\star\top}\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{Y}^{\star}\right\| \lesssim \sigma\sqrt{\frac{\log n}{p}}\sqrt{\sum_{i=1}^{r}\sum_{j=1}^{r}\left\|\boldsymbol{X}_{\cdot,i}^{\star}\left(\boldsymbol{Y}_{\cdot,j}^{\star}\right)^{\top}\right\|_{\mathrm{F}}^{2}} \leq \sigma\sqrt{\frac{\log n}{p}}\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}\left\|\boldsymbol{Y}^{\star}\right\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{\log n}{p}}r\sigma_{\max}.$$

In addition, we have

$$\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}^{\top}\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{Y}^{\star}\right\| \leq \left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right\|\left\|\boldsymbol{Y}^{\star}\right\| \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^{\star}\right\|\sigma\sqrt{\frac{n}{p}}\left\|\boldsymbol{Y}^{\star}\right\| \lesssim \left(\kappa\sigma\sqrt{\frac{n}{p}}\right)^{2}.$$

Combine these two bounds to reach

$$\alpha_{1} \lesssim \sigma\sqrt{\frac{r^{2}\log n}{p}}\sigma_{\max} + \left(\kappa\sigma\sqrt{\frac{n}{p}}\right)^{2}.$$

Substituting the bounds on α_1, α_2, α_3 and α_4 back to (D.20) results in

$$\begin{aligned}\left\|\left(\overline{\boldsymbol{Y}}^{\mathsf{d}}-\boldsymbol{Y}^{\star}\right)^{\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right\| &\lesssim \frac{1}{\sigma_{\min}}\alpha_{1}+\frac{1}{\sigma_{\min}}\alpha_{2}+\kappa\alpha_{3}+\frac{1}{\sigma_{\min}}\alpha_{4}\\ &\lesssim \frac{1}{\sigma_{\min}}\left(\sigma\sqrt{\frac{r^{2}\log n}{p}}\sigma_{\max}+\left(\kappa\sigma\sqrt{\frac{n}{p}}\right)^{2}+\sigma_{\max}\sigma\sqrt{\frac{n}{p}}\cdot\sqrt{\frac{\kappa^{4}\mu^{2}r^{2}\log n}{np}}\right)\\ &\quad+\kappa\left(\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)^{2}\sigma_{\max}+\frac{1}{\sigma_{\min}}\frac{\kappa}{n^{5}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\sigma_{\max}^{2}\\ &\asymp \kappa\sigma_{\max}\left(\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)^{2}+\sigma_{\max}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\cdot\sqrt{\frac{\kappa^{4}\mu^{2}r^{2}\log n}{np}},\end{aligned}$$

which together with (D.15) yields

$$\left\| \boldsymbol{e}_j^\top \boldsymbol{X}^\star \left[\boldsymbol{Y}^{\star\top} \overline{\boldsymbol{Y}}^{\mathsf{d}} \left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top} \overline{\boldsymbol{Y}}^{\mathsf{d}} \right)^{-1} - \boldsymbol{I}_r \right] \right\|_2 \lesssim \frac{\sigma}{\sqrt{p\sigma_{\min}}} \left(\kappa \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\kappa^7 \mu r n}{p}} + \sqrt{\frac{\kappa^7 \mu^3 r^3 \log n}{np}} \right).$$

Taking the maximum over $1 \leq j \leq n$ leads to the desired result.

Finally, we are left with proving Claim 4.

Proof of Claim 4. First, by $\boldsymbol{X}^{\star\top}\boldsymbol{X}^\star = \boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star$, one can obtain

$$\begin{aligned} \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \boldsymbol{Y}^\star - \boldsymbol{X}^{\star\top} \boldsymbol{\Delta}_{\boldsymbol{X}} &= \left(\boldsymbol{Y}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} - \boldsymbol{Y}^\star \right)^\top \boldsymbol{Y}^\star - \boldsymbol{X}^{\star\top} \left(\boldsymbol{X}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} - \boldsymbol{X}^\star \right) \\ &= \left(\boldsymbol{Y}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} \right)^\top \boldsymbol{Y}^\star - \boldsymbol{X}^{\star\top} \left(\boldsymbol{X}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} \right) \\ &= \left(\boldsymbol{Y}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} \right)^\top \left(\boldsymbol{Y}^\star - \boldsymbol{Y}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} \right) + \boldsymbol{H}^{\mathsf{d}\top} \boldsymbol{Y}^{\mathsf{d}\top} \boldsymbol{Y}^{\mathsf{d}} \boldsymbol{H} - \boldsymbol{X}^{\star\top} \left(\boldsymbol{X}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} \right) \\ &= - \left(\boldsymbol{Y}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} \right)^\top \boldsymbol{\Delta}_{\boldsymbol{Y}} + \boldsymbol{\Delta}_{\boldsymbol{X}}^\top \left(\boldsymbol{X}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} \right) + \boldsymbol{H}^{\mathsf{d}\top} \left(\boldsymbol{Y}^{\mathsf{d}\top} \boldsymbol{Y}^{\mathsf{d}} - \boldsymbol{X}^{\mathsf{d}\top} \boldsymbol{X}^{\mathsf{d}} \right) \boldsymbol{H}^{\mathsf{d}}. \end{aligned}$$

We can further decompose it as

$$\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \boldsymbol{Y}^\star - \boldsymbol{X}^{\star\top} \boldsymbol{\Delta}_{\boldsymbol{X}} = \boldsymbol{\Delta}_{\boldsymbol{X}}^\top \boldsymbol{X}^\star + \boldsymbol{\Delta}_{\boldsymbol{X}}^\top \boldsymbol{\Delta}_{\boldsymbol{X}} - \boldsymbol{Y}^{\star\top} \boldsymbol{\Delta}_{\boldsymbol{Y}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \boldsymbol{\Delta}_{\boldsymbol{Y}} + \boldsymbol{H}^{\mathsf{d}\top} \left(\boldsymbol{Y}^{\mathsf{d}\top} \boldsymbol{Y}^{\mathsf{d}} - \boldsymbol{X}^{\mathsf{d}\top} \boldsymbol{X}^{\mathsf{d}} \right) \boldsymbol{H}^{\mathsf{d}}. \tag{D.23}$$

Second, since $\boldsymbol{H}^{\mathsf{d}}$ is the best rotation matrix to align $(\boldsymbol{X}^{\mathsf{d}}, \boldsymbol{Y}^{\mathsf{d}})$ and $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$, we know from [MWCC17, Lemma 35] that

$$\left(\boldsymbol{X}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} \right)^\top \boldsymbol{X}^\star + \left(\boldsymbol{Y}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} \right)^\top \boldsymbol{Y}^\star \succeq \boldsymbol{0},$$

which implies

$$\left(\boldsymbol{X}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} - \boldsymbol{X}^\star \right)^\top \boldsymbol{X}^\star + \left(\boldsymbol{Y}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}} - \boldsymbol{Y}^\star \right)^\top \boldsymbol{Y}^\star = \boldsymbol{\Delta}_{\boldsymbol{X}}^\top \boldsymbol{X}^\star + \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \boldsymbol{Y}^\star$$

is a symmetric matrix, i.e.

$$\boldsymbol{\Delta}_{\boldsymbol{X}}^\top \boldsymbol{X}^\star + \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \boldsymbol{Y}^\star = \boldsymbol{X}^{\star\top} \boldsymbol{\Delta}_{\boldsymbol{X}} + \boldsymbol{Y}^{\star\top} \boldsymbol{\Delta}_{\boldsymbol{Y}}.$$

This is equivalent to

$$\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \boldsymbol{Y}^\star - \boldsymbol{X}^{\star\top} \boldsymbol{\Delta}_{\boldsymbol{X}} = \boldsymbol{Y}^{\star\top} \boldsymbol{\Delta}_{\boldsymbol{Y}} - \boldsymbol{\Delta}_{\boldsymbol{X}}^\top \boldsymbol{X}^\star. \tag{D.24}$$

Combine (D.23) and (D.24) to arrive at

$$\boldsymbol{\Delta}_{\boldsymbol{X}}^\top \boldsymbol{X}^\star + \boldsymbol{\Delta}_{\boldsymbol{X}}^\top \boldsymbol{\Delta}_{\boldsymbol{X}} - \boldsymbol{Y}^{\star\top} \boldsymbol{\Delta}_{\boldsymbol{Y}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \boldsymbol{\Delta}_{\boldsymbol{Y}} + \boldsymbol{H}^{\mathsf{d}\top} \left(\boldsymbol{Y}^{\mathsf{d}\top} \boldsymbol{Y}^{\mathsf{d}} - \boldsymbol{X}^{\mathsf{d}\top} \boldsymbol{X}^{\mathsf{d}} \right) \boldsymbol{H}^{\mathsf{d}} = \boldsymbol{Y}^{\star\top} \boldsymbol{\Delta}_{\boldsymbol{Y}} - \boldsymbol{\Delta}_{\boldsymbol{X}}^\top \boldsymbol{X}^\star,$$

which results in

$$\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \boldsymbol{Y}^\star - \boldsymbol{X}^{\star\top} \boldsymbol{\Delta}_{\boldsymbol{X}} = \boldsymbol{Y}^{\star\top} \boldsymbol{\Delta}_{\boldsymbol{Y}} - \boldsymbol{\Delta}_{\boldsymbol{X}}^\top \boldsymbol{X}^\star = \frac{1}{2} \left(\boldsymbol{\Delta}_{\boldsymbol{X}}^\top \boldsymbol{\Delta}_{\boldsymbol{X}} - \boldsymbol{\Delta}_{\boldsymbol{Y}}^\top \boldsymbol{\Delta}_{\boldsymbol{Y}} \right) + \frac{1}{2} \boldsymbol{H}^{\mathsf{d}\top} \left(\boldsymbol{Y}^{\mathsf{d}\top} \boldsymbol{Y}^{\mathsf{d}} - \boldsymbol{X}^{\mathsf{d}\top} \boldsymbol{X}^{\mathsf{d}} \right) \boldsymbol{H}^{\mathsf{d}}.$$

This completes the proof of the claim. □

D.4 Proof of Lemma 7

Recall that

$$\boldsymbol{A} = \frac{1}{p} \mathcal{P}_\Omega \left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star \right) - \left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star \right) \qquad \text{and} \qquad \boldsymbol{\Phi}_3 = -\boldsymbol{A}\overline{\boldsymbol{Y}}^{\mathsf{d}} \left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top} \overline{\boldsymbol{Y}}^{\mathsf{d}} \right)^{-1}$$

with $\overline{\boldsymbol{Y}}^{\mathsf{d}} = \boldsymbol{Y}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}}$. For any $1 \leq j \leq n$, we have

$$\begin{aligned} \left\| \boldsymbol{e}_j^\top \boldsymbol{A} \overline{\boldsymbol{Y}}^{\mathsf{d}} \left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top} \overline{\boldsymbol{Y}}^{\mathsf{d}} \right)^{-1} \right\|_2 &\leq \left\| \boldsymbol{e}_j^\top \boldsymbol{A} \overline{\boldsymbol{Y}}^{\mathsf{d}} \right\|_2 \left\| \left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top} \overline{\boldsymbol{Y}}^{\mathsf{d}} \right)^{-1} \right\| \\ &\overset{\text{(i)}}{=} \left\| \boldsymbol{e}_j^\top \boldsymbol{A} \boldsymbol{Y}^{\mathsf{d}} \right\|_2 \left\| \left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top} \overline{\boldsymbol{Y}}^{\mathsf{d}} \right)^{-1} \right\| \\ &\overset{\text{(ii)}}{=} \left\| \boldsymbol{e}_j^\top \boldsymbol{A} \boldsymbol{Y} \left(\boldsymbol{I}_r + \frac{\lambda}{p} \left(\boldsymbol{Y}^\top \boldsymbol{Y} \right)^{-1} \right)^{1/2} \right\|_2 \left\| \left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top} \overline{\boldsymbol{Y}}^{\mathsf{d}} \right)^{-1} \right\| \end{aligned}$$

$$
\begin{aligned}
&\leq \left\|\boldsymbol{e}_j^\top \boldsymbol{A}\boldsymbol{Y}\right\|_2 \left\|\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2}\right\| \left\|\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1}\right\| \\
&\overset{\text{(iii)}}{\lesssim} \frac{1}{\sigma_{\min}}\left\|\boldsymbol{e}_j^\top \boldsymbol{A}\boldsymbol{Y}\right\|_2 \overset{\text{(iv)}}{=} \frac{1}{\sigma_{\min}}\left\|\boldsymbol{e}_j^\top \boldsymbol{A}\boldsymbol{Y}\boldsymbol{H}\right\|_2 .
\end{aligned}
$$

Here (i) and (iv) rely on the unitary invariance of the operator norm, (ii) uses the definition of $\boldsymbol{Y}^{\mathsf{d}}$ (see (3.8)) and (iii) follows from the choice $\lambda \lesssim \sigma\sqrt{np}$ and immediate consequences of (A.19)

$$
\left\|\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2}\right\| \asymp 1 \qquad \text{and} \qquad \left\|\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1}\right\| \lesssim \frac{1}{\sigma_{\min}}.
$$

Therefore, it suffices to control $\|\boldsymbol{e}_j^\top \boldsymbol{A}\boldsymbol{Y}\boldsymbol{H}\|_2$. To this end, we have the following decomposition

$$
\begin{aligned}
\boldsymbol{e}_j^\top \boldsymbol{A}\boldsymbol{Y}\boldsymbol{H} &= \boldsymbol{e}_j^\top \left[\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right) - \left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\right]\boldsymbol{Y}\boldsymbol{H} \\
&= \boldsymbol{e}_j^\top \left[\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^{(j)}\boldsymbol{Y}^{(j)\top} - \boldsymbol{M}^\star\right) - \left(\boldsymbol{X}^{(j)}\boldsymbol{Y}^{(j)\top} - \boldsymbol{M}^\star\right)\right]\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)} + \boldsymbol{\Delta}_2,
\end{aligned}
\tag{D.25}
$$

where we define

$$
\begin{aligned}
\boldsymbol{\Delta}_2 \triangleq\ & \boldsymbol{e}_j^\top \left[\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right) - \left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\right]\boldsymbol{Y}\boldsymbol{H} \\
& - \boldsymbol{e}_j^\top \left[\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^{(j)}\boldsymbol{Y}^{(j)\top} - \boldsymbol{M}^\star\right) - \left(\boldsymbol{X}^{(j)}\boldsymbol{Y}^{(j)\top} - \boldsymbol{M}^\star\right)\right]\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}.
\end{aligned}
$$

Denoting

$$
\boldsymbol{v} = [v_1, \cdots, v_n] \triangleq \boldsymbol{e}_j^\top\left(\boldsymbol{X}^{(j)}\boldsymbol{Y}^{(j)\top} - \boldsymbol{M}^\star\right),
$$

we can rewrite the first term of (D.25) as

$$
\boldsymbol{e}_j^\top \left[\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^{(j)}\boldsymbol{Y}^{(j)\top} - \boldsymbol{M}^\star\right) - \left(\boldsymbol{X}^{(j)}\boldsymbol{Y}^{(j)\top} - \boldsymbol{M}^\star\right)\right]\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)} = \frac{1}{p}\sum_{k=1}^{n}\left(\delta_{jk} - p\right)v_k\left[\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right]_{k,\cdot}.
$$

Since $(\boldsymbol{X}^{(j)}, \boldsymbol{Y}^{(j)})$ is independent of $\{\delta_{jk}\}_{1\leq k\leq n}$, the right hand side of the above equation can be viewed as a sum of independent random vectors, conditional on $(\boldsymbol{X}^{(j)}, \boldsymbol{Y}^{(j)})$. Invoke Bernstein's inequality to see that

$$
\left\|\frac{1}{p}\sum_{k=1}^{n}\left(\delta_{jk} - p\right)v_k\left[\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right]_{k,\cdot}\right\|_2 \lesssim \frac{1}{p}\left(\sqrt{V\log n} + B\log n\right)
$$

holds with probability at least $1 - O(n^{-10})$. Here, we denote

$$
\begin{aligned}
V &\triangleq \left\|\sum\nolimits_{k=1}^{n}\mathbb{E}\left[\left(\delta_{jk} - p\right)^2\right]v_k^2\left[\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right]_{k,\cdot}\left[\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right]_{k,\cdot}^\top\right\| \leq p\left\|\boldsymbol{v}\right\|_\infty^2\left\|\boldsymbol{Y}^{(j)}\right\|_{\mathrm{F}}^2, \\
B &\triangleq \max_{1\leq k\leq n}\left\|\left(\delta_{jk} - p\right)v_k\left[\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right]_{k,\cdot}\right\|_2 \leq \left\|\boldsymbol{v}\right\|_\infty\left\|\boldsymbol{Y}^{(j)}\right\|_{2,\infty}.
\end{aligned}
$$

As a result, we obtain

$$
\begin{aligned}
\left\|\frac{1}{p}\sum_{k=1}^{n}\left(\delta_{jk} - p\right)v_k\left[\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right]_{k,\cdot}\right\|_2 &\lesssim \frac{1}{p}\left(\sqrt{p\log n}\left\|\boldsymbol{v}\right\|_\infty\left\|\boldsymbol{Y}^{(j)}\right\|_{\mathrm{F}} + \left\|\boldsymbol{v}\right\|_\infty\left\|\boldsymbol{Y}^{(j)}\right\|_{2,\infty}\log n\right) \\
&\lesssim \frac{\left\|\boldsymbol{v}\right\|_\infty}{p}\left(\sqrt{pr\sigma_{\max}\log n} + \sqrt{\frac{\mu r}{n}\sigma_{\max}\log^2 n}\right) \\
&\asymp \left\|\boldsymbol{v}\right\|_\infty\sqrt{\frac{r\log n}{p}\sigma_{\max}}
\end{aligned}
$$

with the proviso that $np \gg \mu \log n$. Here the middle line depends on $\|\boldsymbol{Y}^{(j)}\|_{\mathrm{F}} \lesssim \sqrt{r\sigma_{\max}}$ and $\|\boldsymbol{Y}^{(j)}\|_{2,\infty} \lesssim \sqrt{\mu r \sigma_{\max}/n}$. Additionally,

$$\begin{aligned}
\|\boldsymbol{v}\|_{\infty} &\le \left\|\boldsymbol{X}^{(j)}\boldsymbol{Y}^{(j)\top} - \boldsymbol{M}^{\star}\right\|_{\infty} \le \left\|\left(\boldsymbol{X}^{(j)}\boldsymbol{R}^{(j)} - \boldsymbol{X}^{\star}\right)\boldsymbol{R}^{(j)\top}\boldsymbol{Y}^{(j)\top} + \boldsymbol{X}^{\star}\left(\boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)} - \boldsymbol{Y}^{\star}\right)^{\top}\right\|_{\infty} \\
&\le \left\|\boldsymbol{X}^{(j)}\boldsymbol{R}^{(j)} - \boldsymbol{X}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{Y}^{(j)}\right\|_{2,\infty} + \|\boldsymbol{X}^{\star}\|_{2,\infty}\left\|\boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)} - \boldsymbol{Y}^{\star}\right\|_{2,\infty} \\
&\lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\frac{\mu r}{n}\sigma_{\max} \lesssim \kappa^{2}\sigma\sqrt{\frac{\mu^{2}r^{2}\log n}{np}}.
\end{aligned}$$

Here the penultimate inequality uses (A.14d) and the bound $\|\boldsymbol{Y}^{(j)}\|_{2,\infty} \lesssim \sqrt{\mu r\sigma_{\max}/n}$. We arrive at the conclusion that: with probability exceeding $1 - O(n^{-10})$,

$$\left\|\frac{1}{p}\sum_{k=1}^{n}(\delta_{jk} - p)\,v_{k}\big(\boldsymbol{X}^{(j)}\boldsymbol{H}^{(j)}\big)_{k,\cdot}\right\|_{2} \lesssim \sigma\sqrt{\frac{\sigma_{\max}}{p}}\cdot\sqrt{\frac{\kappa^{4}\mu^{2}r^{3}\log^{2}n}{np}}.$$

Next, we move on to the second term $\boldsymbol{\Delta}_2$ of (D.25), which can be further decomposed as follows

$$\begin{aligned}
\boldsymbol{\Delta}_2 = &\underbrace{\boldsymbol{e}_j^{\top}\left[\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M}^{\star}\right) - \left(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M}^{\star}\right)\right]\left(\boldsymbol{Y}\boldsymbol{H} - \boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)}_{:=\boldsymbol{\theta}_1} \\
&+ \underbrace{\boldsymbol{e}_j^{\top}\left[\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{X}^{(j)}\boldsymbol{Y}^{(j)\top}\right) - \left(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{X}^{(j)}\boldsymbol{Y}^{(j)\top}\right)\right]\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}}_{:=\boldsymbol{\theta}_2}.
\end{aligned}$$

In what follows, we bound $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ sequentially.

1. Regarding $\boldsymbol{\theta}_1$, using the definition of $\boldsymbol{A}$ we obtain

$$\begin{aligned}
\|\boldsymbol{\theta}_1\|_2 &\le \|\boldsymbol{A}\|\left\|\boldsymbol{Y}\boldsymbol{H} - \boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}\cdot\sqrt{\frac{\kappa^{4}\mu^{2}r^{2}\log n}{np}}\cdot\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\mu r\log n}{p}}\sqrt{\sigma_{\max}} \\
&\asymp \sigma\sqrt{\frac{\sigma_{\max}}{p}}\cdot\sqrt{\frac{\kappa^{4}\mu^{2}r^{2}\log n}{np}}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^{2}\mu r n\log n}{p}},
\end{aligned}$$

where the second relation holds due to (A.14b) and the fact that $\|\boldsymbol{A}\| \lesssim \sigma\sqrt{\frac{n}{p}}\sqrt{\frac{\kappa^{4}\mu^{2}r^{2}\log n}{np}}$ (cf. (A.18)).

2. Moving on to $\boldsymbol{\theta}_2$, we can utilize the identity

$$\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{X}^{(j)}\boldsymbol{Y}^{(j)\top} = \big(\boldsymbol{X}\boldsymbol{H} - \boldsymbol{X}^{(j)}\boldsymbol{H}^{(j)}\big)\big(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\big)^{\top} + \boldsymbol{X}\boldsymbol{H}\big(\boldsymbol{Y}\boldsymbol{H} - \boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\big)^{\top}$$

to deduce that

$$\begin{aligned}
\|\boldsymbol{\theta}_2\|_2 \le &\left\|\boldsymbol{e}_j^{\top}\left[\frac{1}{p}\mathcal{P}_{\Omega}\left[\big(\boldsymbol{X}\boldsymbol{H} - \boldsymbol{X}^{(j)}\boldsymbol{H}^{(j)}\big)\big(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\big)^{\top}\right] - \left(\boldsymbol{X}\boldsymbol{H} - \boldsymbol{X}^{(j)}\boldsymbol{H}^{(j)}\right)\left(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)^{\top}\right]\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right\|_2 \\
&+ \left\|\boldsymbol{e}_j^{\top}\left[\frac{1}{p}\mathcal{P}_{\Omega}\left[\boldsymbol{X}\boldsymbol{H}\left(\boldsymbol{Y}\boldsymbol{H} - \boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)^{\top}\right] - \boldsymbol{X}\boldsymbol{H}\left(\boldsymbol{Y}\boldsymbol{H} - \boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)^{\top}\right]\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right\|_2 \\
= &\underbrace{\left\|\big(\boldsymbol{X}\boldsymbol{H} - \boldsymbol{X}^{(j)}\boldsymbol{H}^{(j)}\big)_{j,\cdot}\frac{1}{p}\sum_{k=1}^{n}(\delta_{jk} - p)\,\big(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\big)_{k,\cdot}^{\top}\big(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\big)_{k,\cdot}\right\|_2}_{:=\alpha_1} \\
&+ \underbrace{\left\|\big(\boldsymbol{X}\boldsymbol{H}\big)_{j,\cdot}\frac{1}{p}\sum_{k=1}^{n}(\delta_{jk} - p)\left(\boldsymbol{Y}\boldsymbol{H} - \boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k,\cdot}^{\top}\left(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k,\cdot}\right\|_2}_{:=\alpha_2}.
\end{aligned}$$

With regards to α_1, we have by Bernstein's inequality and (A.14b) that

$$\begin{aligned}\alpha_1 &\leq \left\|\boldsymbol{X}\boldsymbol{H} - \boldsymbol{X}^{(j)}\boldsymbol{H}^{(j)}\right\|_{\mathrm{F}} \left\|\frac{1}{p}\sum_{k=1}^{n}\left(\delta_{jk}-p\right)\left(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k,\cdot}^{\top}\left(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k,\cdot}\right\| \\ &\lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\sqrt{\frac{\mu r}{n}\sigma_{\max}}\cdot\frac{1}{p}\left(\sqrt{V_2\log n}+B_2\log n\right)\end{aligned}$$

holds with probability exceeding $1-O(n^{-10})$. Here, we define

$$\begin{aligned}V_2 &\triangleq \left\|\sum_{k=1}^{n}\mathbb{E}\left(\delta_{jk}-p\right)^2\left(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k,\cdot}^{\top}\left(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k,\cdot}\left(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k,\cdot}^{\top}\left(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k,\cdot}\right\| \\ &\leq p\left\|\boldsymbol{Y}^{(j)}\right\|_{2,\infty}^{2}\left\|\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\right\|, \\ B_2 &\triangleq \max_{1\leq k\leq n}\left\|\left(\delta_{jk}-p\right)\left(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k,\cdot}^{\top}\left(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k,\cdot}\right\| \leq \left\|\boldsymbol{Y}^{(j)}\right\|_{2,\infty}^{2}.\end{aligned}$$

As a result, we can obtain

$$\begin{aligned}\alpha_1 &\lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\mu r\log n}{p}}\sigma_{\max}\cdot\frac{1}{p}\left(\sqrt{p\sigma_{\max}\log n}\left\|\boldsymbol{Y}^{(j)}\right\|_{2,\infty}+\left\|\boldsymbol{Y}^{(j)}\right\|_{2,\infty}^{2}\log n\right) \\ &\lesssim \sigma\sqrt{\frac{\sigma_{\max}}{p}}\cdot\sqrt{\frac{\kappa^4\mu^2r^2\log^2 n}{np}},\end{aligned}$$

provided that $np \gg \mu r\log n$. Here we apply the bounds $\|\boldsymbol{Y}^{(j)}\| \lesssim \sqrt{\sigma_{\max}}$ and $\|\boldsymbol{Y}^{(j)}\|_{2,\infty} \lesssim \sqrt{\mu r\sigma_{\max}/n}$ (see (A.19) and the following remarks). In the end, we turn to the term α_2, which obeys

$$\begin{aligned}\alpha_2 &\leq \frac{1}{p}\left\|\boldsymbol{X}\right\|_{2,\infty}\sum_{k=1}^{n}\left|\delta_{jk}-p\right|\left\|\left(\boldsymbol{Y}\boldsymbol{H}-\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k\cdot}\right\|_2\left\|\left(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k\cdot}\right\|_2 \\ &\leq \frac{1}{p}\sqrt{\frac{\mu r}{n}}\sqrt{\sigma_{\max}}\cdot\sqrt{\sum_{k=1}^{n}\left(\delta_{jk}-p\right)^2}\cdot\sqrt{\sum_{k=1}^{n}\left\|\left(\boldsymbol{Y}\boldsymbol{H}-\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k\cdot}\right\|_2^2\left\|\left(\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right)_{k\cdot}\right\|_2^2} \\ &\lesssim \frac{1}{p}\sqrt{\frac{\mu r}{n}}\sqrt{\sigma_{\max}}\cdot\sqrt{np}\cdot\left\|\boldsymbol{Y}\boldsymbol{H}-\boldsymbol{Y}^{(j)}\boldsymbol{H}^{(j)}\right\|_{\mathrm{F}}\left\|\boldsymbol{Y}^{(j)}\right\|_{2,\infty} \\ &\lesssim \sqrt{\frac{\mu r}{p}}\sqrt{\sigma_{\max}}\cdot\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\frac{\mu r}{n}\sigma_{\max} \asymp \sigma\sqrt{\frac{\sigma_{\max}}{p}}\cdot\sqrt{\frac{\kappa^4\mu^3r^3\log n}{np}},\end{aligned}$$

where the second line arises from the Cauchy-Schwarz inequality.

Take the previous bounds collectively to arrive at

$$\begin{aligned}\left\|\boldsymbol{\Delta}_2\right\|_2 &\lesssim \sigma\sqrt{\frac{\sigma_{\max}}{p}}\left\{\sqrt{\frac{\kappa^4\mu^2r^2\log n}{np}}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^2\mu rn\log n}{p}}+\sqrt{\frac{\kappa^4\mu^2r^2\log^2 n}{np}}+\sqrt{\frac{\kappa^4\mu^3r^3\log n}{np}}\right\} \\ &\lesssim \sigma\sqrt{\frac{\sigma_{\max}}{p}}\cdot\sqrt{\frac{\kappa^4\mu^3r^3\log^2 n}{np}}\end{aligned}$$

as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^2 n\log n}{p}} \ll 1$. Finally, we conclude that

$$\left\|\boldsymbol{e}_j^{\top}\boldsymbol{A}\overline{\boldsymbol{Y}}^{\mathsf{d}}\left(\overline{\boldsymbol{Y}}^{\mathsf{d}\top}\overline{\boldsymbol{Y}}^{\mathsf{d}}\right)^{-1}\right\|_2 \lesssim \frac{1}{\sigma_{\min}}\left\|\boldsymbol{e}_j^{\top}\boldsymbol{A}\boldsymbol{Y}\boldsymbol{H}\right\|_2$$

$$\begin{aligned}
&\lesssim \frac{1}{\sigma_{\min}}\left(\sigma\sqrt{\frac{\sigma_{\max}}{p}}\sqrt{\frac{\kappa^4\mu^2 r^3\log^2 n}{np}}+\sigma\sqrt{\frac{\sigma_{\max}}{p}}\sqrt{\frac{\kappa^4\mu^3 r^3\log^2 n}{np}}\right)\\
&\asymp \frac{\sigma}{\sqrt{p\sigma_{\min}}}\cdot\sqrt{\frac{\kappa^5\mu^3 r^3\log^2 n}{np}},
\end{aligned}\tag{D.26}$$

thus concluding the proof.

D.5 Proof of Lemma 8

First, it is straightforward to verify that

$$\begin{aligned}
&\left\|\nabla_{\boldsymbol{X}} f\left(\boldsymbol{X},\boldsymbol{Y}\right)\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)^{-1}\boldsymbol{H}^{\mathsf{d}}\right\|_{2,\infty}\\
&\quad\le\left\|\nabla_{\boldsymbol{X}} f\left(\boldsymbol{X},\boldsymbol{Y}\right)\right\|_{\mathrm{F}}\left\|\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2}\right\|\left\|\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)^{-1}\right\|\\
&\quad\lesssim\frac{1}{n^5}\frac{\lambda}{p}\sqrt{\sigma_{\min}}\cdot\frac{1}{\sigma_{\min}}\lesssim\frac{\sigma}{\sqrt{p\sigma_{\min}}}\cdot\frac{1}{n^4},
\end{aligned}\tag{D.27}$$

where the last line arises from (A.10), the choice $\lambda\lesssim\sigma\sqrt{np}$ (cf. (A.6)), and the bounds

$$\left\|\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2}\right\|\asymp 1\qquad\text{and}\qquad\left\|\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)^{-1}\right\|\lesssim\frac{1}{\sigma_{\min}}.$$

Here the latter two are immediate consequences of (A.19). Second, with regards to the term involving $\boldsymbol{\Delta}_{\mathsf{balancing}}$, we have

$$\begin{aligned}
\left\|\boldsymbol{X}\boldsymbol{\Delta}_{\mathsf{balancing}}\boldsymbol{H}^{\mathsf{d}}\right\|_{2,\infty}&\le\left\|\boldsymbol{X}\right\|_{2,\infty}\left\|\boldsymbol{\Delta}_{\mathsf{balancing}}\right\|\\
&\lesssim\sqrt{\frac{\mu r}{n}\sigma_{\max}}\left\|\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right)^{1/2}-\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2}\right\|\\
&\lesssim\sqrt{\frac{\mu r}{n}\sigma_{\max}}\cdot\frac{1}{n^5}\frac{\lambda}{p}\frac{\kappa}{\sigma_{\min}}\asymp\frac{\sigma}{\sqrt{p\sigma_{\min}}}\sqrt{\frac{\kappa^3\mu r}{n^{10}}},
\end{aligned}\tag{D.28}$$

where the middle line uses (A.19) and the last one follows from (C.16).

Combine (D.27), (D.28) and the triangle inequality to establish the advertised result, with the proviso that $n^2\gg\kappa^3\mu r$.

D.6 Proof of Lemma 9

We invoke the identity $\boldsymbol{Y}^\star(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star)^{-1}=\boldsymbol{V}^\star(\boldsymbol{\Sigma}^\star)^{-1/2}$ (since $\boldsymbol{Y}^\star=\boldsymbol{V}^\star(\boldsymbol{\Sigma}^\star)^{1/2}$) to see that for any $1\le i\le n$,

$$\left(\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}^\star(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star)^{-1}\right)^\top\boldsymbol{e}_i=\left(\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{V}^\star\left(\boldsymbol{\Sigma}^\star\right)^{-1/2}\right)^\top\boldsymbol{e}_i=\sum_{k=1}^n\frac{1}{p}E_{ik}\delta_{ik}\left(\boldsymbol{\Sigma}^\star\right)^{-1/2}\left(\boldsymbol{V}_{k,\cdot}^\star\right)^\top\tag{D.29}$$

consists of a sum of independent random vectors, where we recall that $\delta_{ik}=\mathbb{1}\{(i,k)\in\Omega\}$. In addition, the right-hand side of the above formula is conditionally Gaussian, namely,

$$\sum_{k=1}^n\frac{1}{p}E_{ik}\delta_{ik}\left(\boldsymbol{\Sigma}^\star\right)^{-1/2}\left(\boldsymbol{V}_{k,\cdot}^\star\right)^\top\;\Big|\;\{\delta_{ik}\}_{k:1\le k\le n}\sim\mathcal{N}\Bigg(\boldsymbol{0},\underbrace{\frac{\sigma^2}{p^2}\sum_{k=1}^n\delta_{ik}\left(\boldsymbol{\Sigma}^\star\right)^{-1/2}\left(\boldsymbol{V}_{k,\cdot}^\star\right)^\top\boldsymbol{V}_{k,\cdot}^\star\left(\boldsymbol{\Sigma}^\star\right)^{-1/2}}_{:=\boldsymbol{S}}\Bigg).$$

Note that $\boldsymbol{S}$ depends on the index i through $\{\delta_{ik}\}_{k:1\le k\le n}$. Denote by $\boldsymbol{S}^\star$ the expectation of $\boldsymbol{S}$, that is,

$$\boldsymbol{S}^\star\triangleq\mathbb{E}\big[\boldsymbol{S}\big]=p^{-1}\sigma^2\left(\boldsymbol{\Sigma}^\star\right)^{-1}\succeq\sigma^2/(p\sigma_{\max})\cdot\boldsymbol{I}_r,$$

and introduce the following event

$$\mathcal{E} \triangleq \left\{ \|\boldsymbol{S} - \boldsymbol{S}^{\star}\| \lesssim \frac{\sigma^2}{p\sigma_{\min}} \sqrt{\frac{\mu r \log n}{np}} \right\}.$$

Clearly, when $np \gg \kappa^2 \mu r \log n$, one has $\boldsymbol{S} \succ \boldsymbol{0}$ on the event $\mathcal{E}$ and hence $\boldsymbol{S}^{-1/2}$ is well-defined. As a result, on the event $\mathcal{E}$, we have

$$\left(\boldsymbol{S}^{\star}\right)^{1/2} \boldsymbol{S}^{-1/2} \sum_{k=1}^{n} \frac{1}{p} E_{ik}\delta_{ik} \left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2} \left(\boldsymbol{V}_{k,\cdot}^{\star}\right)^{\top} \;\Big|\; \{\delta_{ik}\}_{k:1\leq k\leq n} \;\sim\; \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{S}^{\star}\right). \tag{D.30}$$

In view of this relation, we can define the ith row of $\boldsymbol{Z_X} \in \mathbb{R}^{n\times r}$ to be

$$\boldsymbol{e}_i^{\top} \boldsymbol{Z_X} \triangleq \begin{cases} \frac{1}{p} \boldsymbol{e}_i^{\top} \mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \boldsymbol{Y}^{\star} (\boldsymbol{Y}^{\star\top} \boldsymbol{Y}^{\star})^{-1} \boldsymbol{S}^{-1/2} \left(\boldsymbol{S}^{\star}\right)^{1/2}, & \text{on the event } \mathcal{E}, \\ \boldsymbol{e}_i^{\top} \boldsymbol{G_X}, & \text{on the event } \mathcal{E}^{c}, \end{cases} \tag{D.31}$$

where $\boldsymbol{G_X} \in \mathbb{R}^{n\times r}$ is an independently generated random matrix satisfying

$$\boldsymbol{G}_{\boldsymbol{X}}^{\top} \boldsymbol{e}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \frac{\sigma^2}{p} \left(\boldsymbol{\Sigma}^{\star}\right)^{-1}\right) \qquad \text{for} \quad 1 \leq i \leq n.$$

As can be easily seen from (D.29) and (D.30), each row of $\boldsymbol{Z_X}$ follows the Gaussian distribution

$$\boldsymbol{Z}_{\boldsymbol{X}}^{\top} \boldsymbol{e}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \frac{\sigma^2}{p} \left(\boldsymbol{\Sigma}^{\star}\right)^{-1}\right) \qquad \text{for} \quad 1 \leq i \leq n.$$

It remains to show that, with high probability,

$$\boldsymbol{\Delta_X} \triangleq \frac{1}{p} \mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \boldsymbol{Y}^{\star} (\boldsymbol{Y}^{\star\top} \boldsymbol{Y}^{\star})^{-1} - \boldsymbol{Z_X} = \frac{1}{p} \mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \boldsymbol{V}^{\star} \left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2} - \boldsymbol{Z_X}$$

is small when measured by the $\ell_{2,\infty}$ norm. To this end, observe that on the event $\mathcal{E}$,

$$\begin{aligned} \boldsymbol{e}_i^{\top} \boldsymbol{\Delta_X} &= \frac{1}{p} \boldsymbol{e}_i^{\top} \mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \boldsymbol{V}^{\star} \left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2} - \frac{1}{p} \boldsymbol{e}_i^{\top} \mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \boldsymbol{V}^{\star} \left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2} \boldsymbol{S}^{-1/2} \left(\boldsymbol{S}^{\star}\right)^{1/2} \\ &= \frac{1}{p} \boldsymbol{e}_i^{\top} \mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \boldsymbol{V}^{\star} \left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2} \left[\boldsymbol{I}_r - \boldsymbol{S}^{-1/2} \left(\boldsymbol{S}^{\star}\right)^{1/2}\right], \end{aligned}$$

and therefore, we have

$$\begin{aligned} \left\|\boldsymbol{e}_i^{\top} \boldsymbol{\Delta_X}\right\|_2 &\leq \frac{1}{p} \left\|\boldsymbol{e}_i^{\top} \mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \boldsymbol{V}^{\star}\right\|_2 \left\|\left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2}\right\| \left\|\boldsymbol{I}_r - \boldsymbol{S}^{-1/2} \left(\boldsymbol{S}^{\star}\right)^{1/2}\right\| \\ &= \frac{1}{p\sqrt{\sigma_{\min}}} \left\|\sum\nolimits_k E_{ik}\delta_{ik} \boldsymbol{V}_{k,\cdot}^{\star}\right\|_2 \left\|\boldsymbol{I}_r - \boldsymbol{S}^{-1/2} \left(\boldsymbol{S}^{\star}\right)^{1/2}\right\|. \end{aligned}$$

In what follows, we shall bound the two terms on the right-hand side of the above display sequentially.

1. First, observe that $\sum_{k=1}^{n} E_{ik}\delta_{ik} \boldsymbol{V}_{k,\cdot}^{\star}$ involves a sum of independent random vectors with

$$\left\|\left\|E_{ik}\delta_{ik} \boldsymbol{V}_{k,\cdot}^{\star}\right\|_2\right\|_{\psi_1} \leq \left\|\boldsymbol{V}_{k,\cdot}^{\star}\right\|_2 \left\|E_{jk}\delta_{jk}\right\|_{\psi_1} \lesssim \sigma\sqrt{\mu r/n},$$

where $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm [Ver17]. One can then apply the matrix Bernstein inequality [Kol11, Theorem 2.7] to conclude that with probability at least $1 - O(n^{-20})$,

$$\left\|\sum\nolimits_k E_{ik}\delta_{ik} \boldsymbol{V}_{k,\cdot}^{\star}\right\|_2 \lesssim \sqrt{V_1 \log n} + \max_{1\leq k\leq n} \left\|\left\|E_{ik}\delta_{ik} \boldsymbol{V}_{k,\cdot}^{\star}\right\|_2\right\|_{\psi_1} \log^2 n,$$

where we denote

$$V_1 \triangleq \left\| \mathbb{E}\left[\sum\nolimits_{k=1}^{n} E_{jk}^2 \delta_{jk}^2 \boldsymbol{V}_{k,\cdot}^{\star} \left(\boldsymbol{V}_{k,\cdot}^{\star}\right)^{\top} \right] \right\| = \sigma^2 p \left\|\boldsymbol{V}^{\star}\right\|_{\mathrm{F}}^2 = \sigma^2 pr.$$

As a result, we arrive at

$$\left\| \sum\nolimits_{k} E_{ik}\delta_{ik}\boldsymbol{V}_{k,\cdot}^{\star} \right\|_2 \lesssim \sqrt{\sigma^2 pr \log n} + \sigma\sqrt{\frac{\mu r}{n}}\log^2 n \lesssim \sigma\sqrt{pr\log n} \tag{D.32}$$

as long as $np \gg \mu \log^3 n$.

2. Next, we move on to $\|\boldsymbol{I}_r - \boldsymbol{S}^{-1/2}(\boldsymbol{S}^{\star})^{1/2}\|$. Recall that on the event $\mathcal{E}$, one has

$$\left\|\boldsymbol{S} - \boldsymbol{S}^{\star}\right\| \lesssim \frac{\sigma^2}{p\sigma_{\min}}\sqrt{\frac{\mu r \log n}{np}}.$$

This together with the fact that $\sigma^2/(p\sigma_{\max}) \leq \lambda_{\min}(\boldsymbol{S}^{\star}) \leq \lambda_{\max}(\boldsymbol{S}^{\star}) \leq \sigma^2/(p\sigma_{\min})$ gives

$$\frac{\sigma^2}{2p\sigma_{\max}} \leq \lambda_{\min}\left(\boldsymbol{S}\right) \leq \lambda_{\max}\left(\boldsymbol{S}\right) \leq \frac{2\sigma^2}{p\sigma_{\min}}, \quad \sqrt{\frac{\sigma^2}{2p\sigma_{\max}}} \leq \lambda_{\min}\left(\boldsymbol{S}^{1/2}\right) \leq \lambda_{\max}\left(\boldsymbol{S}^{1/2}\right) \leq \sqrt{\frac{2\sigma^2}{p\sigma_{\min}}}, \tag{D.33}$$

with the proviso that $np \gg \kappa^2 \mu r \log n$. Therefore, straightforward calculations yield

$$\begin{aligned} \left\|\boldsymbol{I}_r - \boldsymbol{S}^{-1/2}\left(\boldsymbol{S}^{\star}\right)^{1/2}\right\| &\leq \left\|\boldsymbol{S}^{-1/2}\right\| \cdot \left\|\boldsymbol{S}^{1/2} - \left(\boldsymbol{S}^{\star}\right)^{1/2}\right\| \\ &\leq \left\|\boldsymbol{S}^{-1/2}\right\| \frac{1}{\lambda_{\min}\left(\boldsymbol{S}^{1/2}\right) + \lambda_{\min}\left(\left(\boldsymbol{S}^{\star}\right)^{1/2}\right)} \left\|\boldsymbol{S} - \boldsymbol{S}^{\star}\right\| \\ &\lesssim \sqrt{\frac{p\sigma_{\max}}{\sigma^2}} \cdot \frac{1}{\sqrt{\frac{\sigma^2}{p\sigma_{\max}}}} \cdot \frac{\sigma^2}{p\sigma_{\min}}\sqrt{\frac{\mu r \log n}{np}} \asymp \sqrt{\frac{\kappa^2 \mu r \log n}{np}}. \end{aligned}$$

Here the second relation is the perturbation bound for the matrix square roots (see Lemma 13).

Combine the above two bounds to conclude that

$$\begin{aligned} \left\|\boldsymbol{e}_i^{\top}\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|_2 &= \frac{1}{p}\left\| \boldsymbol{e}_i^{\top}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{V}^{\star}\left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2}\left[\boldsymbol{I}_r - \boldsymbol{S}^{-1/2}\left(\boldsymbol{S}^{\star}\right)^{1/2}\right]\right\|_2 \\ &\lesssim \frac{1}{p}\cdot\sigma\sqrt{pr\log n}\cdot\frac{1}{\sqrt{\sigma_{\min}}}\cdot\sqrt{\frac{\kappa^2\mu r\log n}{np}} \asymp \frac{\sigma}{\sqrt{p\sigma_{\min}}}\cdot\sqrt{\frac{\kappa^2\mu r^2\log^2 n}{np}}. \end{aligned}$$

Finally, we are left with demonstrating that $\mathbb{P}(\mathcal{E}^c) = O(n^{-10})$. To see this, by definition one has

$$\begin{aligned} \left\|\boldsymbol{S} - \boldsymbol{S}^{\star}\right\| &= \frac{\sigma^2}{p}\left\| \frac{1}{p}\sum_{k=1}^{n}\delta_{ik}\left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2}\left(\boldsymbol{V}_{k,\cdot}^{\star}\right)^{\top}\boldsymbol{V}_{k,\cdot}^{\star}\left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2} - \left(\boldsymbol{\Sigma}^{\star}\right)^{-1}\right\| \\ &\leq \frac{\sigma^2}{p^2\sigma_{\min}}\left\|\sum\nolimits_{k}\delta_{ik}\left(\boldsymbol{V}_{k,\cdot}^{\star}\right)^{\top}\boldsymbol{V}_{k,\cdot}^{\star} - p\boldsymbol{I}_r\right\| \\ &\lesssim \frac{\sigma^2}{p^2\sigma_{\min}}\left(\sqrt{V_2\log n} + B_2\log n\right) \end{aligned}$$

with probability at least $1 - O(n^{-10})$. Here the last line utilizes the matrix Bernstein inequality, where

$$\begin{aligned} B_2 &\triangleq \max_{1\leq k\leq n}\left\|\left(\delta_{jk} - p\right)\left(\boldsymbol{V}_{k,\cdot}^{\star}\right)^{\top}\boldsymbol{V}_{k,\cdot}^{\star}\right\| \leq \frac{\mu r}{n}, \\ V_2 &\triangleq \left\|\mathbb{E}\left[\sum\nolimits_{k}\left(\delta_{jk} - p\right)^2\left(\boldsymbol{V}_{k,\cdot}^{\star}\right)^{\top}\boldsymbol{V}_{k,\cdot}^{\star}\left(\boldsymbol{V}_{k,\cdot}^{\star}\right)^{\top}\boldsymbol{V}_{k,\cdot}^{\star}\right]\right\| \leq p\frac{\mu r}{n}\left\|\boldsymbol{V}^{\star\top}\boldsymbol{V}^{\star}\right\| = \frac{\mu r p}{n}. \end{aligned}$$

Consequently with probability exceeding $1 - O(n^{-10})$ one has

$$\left\|\boldsymbol{S} - \boldsymbol{S}^{\star}\right\| \lesssim \frac{\sigma^2}{p^2\sigma_{\min}}\left(\sqrt{\frac{\mu r p \log n}{n}} + \frac{\mu r}{n}\log n\right) \asymp \frac{\sigma^2}{p\sigma_{\min}}\sqrt{\frac{\mu r \log n}{np}}$$

as long as $np \gtrsim \mu r \log n$. This means that $\mathbb{P}(\mathcal{E}^c) = O(n^{-10})$ and taking the union bounds over $1 \leq i \leq n$ concludes the proof.

E Analysis of the entries of the matrix

E.1 Proof of Lemma 10

The term Λ_{ij} can be naturally split into two terms, namely

$$\boldsymbol{e}_i^\top \boldsymbol{\Psi}_{\boldsymbol{X}} \boldsymbol{Y}^{\star\top} \boldsymbol{e}_j + \boldsymbol{e}_i^\top \boldsymbol{X}^\star \boldsymbol{\Psi}_{\boldsymbol{Y}}^\top \boldsymbol{e}_j \qquad \text{and} \qquad \boldsymbol{e}_i^\top \big(\overline{\boldsymbol{X}}^{\mathsf{d}} - \boldsymbol{X}^\star\big)\big(\overline{\boldsymbol{Y}}^{\mathsf{d}} - \boldsymbol{Y}^\star\big)^\top \boldsymbol{e}_j.$$

In what follows, we shall bound each term individually.

1. Regarding the first term, one sees from Theorem 5 that with probability exceeding $1 - O(n^{-10})$

$$\max\big\{\|\boldsymbol{\Psi}_{\boldsymbol{X}}\|_{2,\infty}, \|\boldsymbol{\Psi}_{\boldsymbol{Y}}\|_{2,\infty}\big\} \lesssim \frac{\sigma}{\sqrt{p\sigma_{\min}}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^7 \mu r n \log n}{p}} + \sqrt{\frac{\kappa^7 \mu^3 r^3 \log^2 n}{np}}\right).$$

   As a result, we obtain

$$\begin{aligned}\big|\boldsymbol{e}_i^\top \boldsymbol{\Psi}_{\boldsymbol{X}} \boldsymbol{Y}^{\star\top} \boldsymbol{e}_j + \boldsymbol{e}_i^\top \boldsymbol{X}^\star \boldsymbol{\Psi}_{\boldsymbol{Y}}^\top \boldsymbol{e}_j\big| &\le \|\boldsymbol{\Psi}_{\boldsymbol{X}}\|_{2,\infty} \big\|\boldsymbol{Y}_{j,\cdot}^\star\big\|_2 + \big\|\boldsymbol{X}_{i,\cdot}^\star\big\|_2 \|\boldsymbol{\Psi}_{\boldsymbol{Y}}\|_{2,\infty} \\ &\lesssim \left(\big\|\boldsymbol{U}_{i,\cdot}^\star\big\|_2 + \big\|\boldsymbol{V}_{j,\cdot}^\star\big\|_2\right)\frac{\sigma}{\sqrt{p}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^8 \mu r n \log n}{p}} + \sqrt{\frac{\kappa^8 \mu^3 r^3 \log^2 n}{np}}\right),\end{aligned}$$

   where the last line follows since $\|\boldsymbol{X}_{i,\cdot}^\star\|_2 \le \sqrt{\sigma_{\max}}\|\boldsymbol{U}_{i,\cdot}^\star\|_2$ and $\|\boldsymbol{Y}_{j,\cdot}^\star\|_2 \le \sqrt{\sigma_{\max}}\|\boldsymbol{V}_{j,\cdot}^\star\|_2$.

2. Turning to the second term, we have by the Cauchy-Schwarz inequality that

$$\begin{aligned}\left|\boldsymbol{e}_i^\top \big(\overline{\boldsymbol{X}}^{\mathsf{d}} - \boldsymbol{X}^\star\big)\big(\overline{\boldsymbol{Y}}^{\mathsf{d}} - \boldsymbol{Y}^\star\big)^\top \boldsymbol{e}_j\right| \le \big\|\overline{\boldsymbol{X}}^{\mathsf{d}} - \boldsymbol{X}^\star\big\|_{2,\infty}\big\|\overline{\boldsymbol{Y}}^{\mathsf{d}} - \boldsymbol{Y}^\star\big\|_{2,\infty} &\lesssim \left(\kappa \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\|\boldsymbol{F}^\star\|_{2,\infty}\right)^2 \\ &\lesssim \left(\frac{\sigma}{\sqrt{\sigma_{\min}}}\sqrt{\frac{\kappa^3 \mu r \log n}{p}}\right)^2,\end{aligned}$$

   where the penultimate inequality uses (A.13d) and the last one depends on the incoherence assumption that $\|\boldsymbol{F}^\star\|_{2,\infty} \le \sqrt{\mu r \sigma_{\max}/n}$ (see (A.17)).

Take collectively the above two bounds to complete the proof.

E.2 Proof of Lemma 11

If $\boldsymbol{Z}_{\boldsymbol{X}}^\top \boldsymbol{e}_i$ and $\boldsymbol{Z}_{\boldsymbol{Y}}^\top \boldsymbol{e}_j$ were independent, then clearly one would have

$$\boldsymbol{e}_i^\top \boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{Y}^{\star\top} \boldsymbol{e}_j + \boldsymbol{e}_i^\top \boldsymbol{X}^\star \boldsymbol{Z}_{\boldsymbol{Y}}^\top \boldsymbol{e}_j \sim \mathcal{N}\left(0, v_{ij}^\star\right).$$

As such, the main ingredient of the proof boils down to demonstrating that $\boldsymbol{Z}_{\boldsymbol{X}}^\top \boldsymbol{e}_i$ and $\boldsymbol{Z}_{\boldsymbol{Y}}^\top \boldsymbol{e}_j$ are nearly independent.

To begin with, we remind the readers of the way we construct $\boldsymbol{e}_i^\top \boldsymbol{Z}_{\boldsymbol{X}}$ and $\boldsymbol{e}_j^\top \boldsymbol{Z}_{\boldsymbol{Y}}$ in Appendix D.6: there exist events $\mathcal{E}$ and $\widetilde{\mathcal{E}}$ with $\mathbb{P}(\mathcal{E}^{\mathrm{c}} \cup \widetilde{\mathcal{E}}^{\mathrm{c}}) \lesssim n^{-10}$ such that

$$\begin{aligned}\boldsymbol{e}_i^\top \boldsymbol{Z}_{\boldsymbol{X}} &\triangleq \frac{1}{p}\boldsymbol{e}_i^\top \mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}^\star (\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star)^{-1}\boldsymbol{S}^{-1/2}\left(\boldsymbol{S}^\star\right)^{1/2} \qquad &&\text{on the event } \mathcal{E} \\ \boldsymbol{e}_j^\top \boldsymbol{Z}_{\boldsymbol{Y}} &\triangleq \frac{1}{p}\boldsymbol{e}_j^\top \left(\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right)^\top \boldsymbol{X}^\star (\boldsymbol{X}^{\star\top}\boldsymbol{X}^\star)^{-1}\tilde{\boldsymbol{S}}^{-1/2}\left(\boldsymbol{S}^\star\right)^{1/2} \qquad &&\text{on the event } \widetilde{\mathcal{E}}\end{aligned}$$

where the randomness of $\boldsymbol{S}$ only comes from $\{\delta_{ik}\}_{k:1\le k\le n}$, and the randomness of $\tilde{\boldsymbol{S}}$ only comes from $\{\delta_{kj}\}_{k:1\le k\le n}$. In addition, the events $\mathcal{E}$ and $\widetilde{\mathcal{E}}$ depend only on $\{\delta_{ik}\}_{k:1\le k\le n}$ and $\{\delta_{kj}\}_{k:1\le k\le n}$, respectively. As a result, $\boldsymbol{Z}_{\boldsymbol{X}}^\top \boldsymbol{e}_i$ depends only on $\{\delta_{ik}, E_{ik}\}_{k:1\le k\le n}$ and $\boldsymbol{Z}_{\boldsymbol{Y}}^\top \boldsymbol{e}_j$ relies only on $\{\delta_{kj}, E_{kj}\}_{k:1\le k\le n}$. This tells us that: the only common randomness underlying $\boldsymbol{Z}_{\boldsymbol{X}}^\top \boldsymbol{e}_i$ and $\boldsymbol{Z}_{\boldsymbol{Y}}^\top \boldsymbol{e}_j$ lies in δ_{ij} and E_{ij}.

Fortunately, this weak dependency can be easily decoupled, for which we have the following claim.

Claim 5. *Suppose that $np \gg \kappa^2 \mu r^2 \log^2 n$. One has the decomposition*

$$\boldsymbol{Z}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i = \widetilde{\boldsymbol{Z}}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i + \boldsymbol{\Delta}_i,$$

where $\widetilde{\boldsymbol{Z}}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{\Sigma}^{\star})^{-1}/p)$ and is independent of $\{\delta_{kj}, E_{k}j\}_{k:1\leq k\leq n}$ and hence of $\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\boldsymbol{e}_j$. In addition, with probability at least $1 - O(n^{-10})$ one has

$$\|\boldsymbol{\Delta}_i\|_2 \lesssim \frac{\sigma}{\sqrt{p\sigma_{\min}}}\sqrt{\frac{\kappa\mu r\log n}{np}}.$$

The desired result follows immediately from Claim 5, since

$$\boldsymbol{e}_i^{\top}\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j + \boldsymbol{e}_i^{\top}\boldsymbol{X}^{\star}\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\boldsymbol{e}_j = \underbrace{\boldsymbol{e}_i^{\top}\widetilde{\boldsymbol{Z}}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j + \boldsymbol{e}_i^{\top}\boldsymbol{X}^{\star}\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\boldsymbol{e}_j}_{\sim\mathcal{N}\left(0, v_{ij}^{\star}\right)} + \boldsymbol{\Delta}_i^{\top}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j,$$

where

$$\left|\boldsymbol{\Delta}_i^{\top}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j\right| \leq \|\boldsymbol{\Delta}_i\|_2\|\boldsymbol{Y}_{j,\cdot}^{\star}\|_{2,\infty} \lesssim \frac{\sigma}{\sqrt{p\sigma_{\min}}}\sqrt{\frac{\kappa\mu r\log n}{np}}\sqrt{\sigma_{\max}}\|\boldsymbol{V}_{j,\cdot}^{\star}\|_{2,\infty} \asymp \frac{\sigma}{\sqrt{p}}\sqrt{\frac{\kappa^2\mu r\log n}{np}}\|\boldsymbol{V}_{j,\cdot}^{\star}\|_{2,\infty}.$$

Similarly, repeating the same argument above, we can also show that $\boldsymbol{e}_i^{\top}\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{e}_j + \boldsymbol{e}_i^{\top}\boldsymbol{X}^{\star}\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\boldsymbol{e}_j$ can be decomposed as a Gaussian random variable $\mathcal{N}\left(0, v_{ij}^{\star}\right)$ as well as a residual term bounded above by $(\sigma/\sqrt{p})\sqrt{(\kappa^2\mu r\log n)/(np)}\|\boldsymbol{U}_{i,\cdot}^{\star}\|_{2,\infty}$ with high probability. These together finish the proof.

Proof of Claim 5. Instate the notation used in Appendix D.6. Recall that

$$\boldsymbol{Z}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i = \begin{cases} \left(\boldsymbol{S}^{\star}\right)^{1/2}\boldsymbol{S}^{-1/2}\sum_{k=1}^{n}\frac{1}{p}E_{ik}\delta_{ik}\left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2}\left(\boldsymbol{V}_{k,\cdot}^{\star}\right)^{\top}, & \text{on the event } \mathcal{E}, \\ \boldsymbol{G}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i, & \text{on the event } \mathcal{E}^c. \end{cases}$$

To remove the effect of δ_{ij}, E_{ij} on $\boldsymbol{Z}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i$, we construct an auxiliary random matrix $\widetilde{\boldsymbol{Z}}_{\boldsymbol{X}}$ as follows

$$\widetilde{\boldsymbol{Z}}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i = \begin{cases} \left(\boldsymbol{S}^{\star}\right)^{1/2}\boldsymbol{S}_{-j}^{-1/2}\sum_{k:k\neq j}\frac{1}{p}E_{ik}\delta_{ik}\left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2}\left(\boldsymbol{V}_{k,\cdot}^{\star}\right)^{\top}, & \text{on the event } \mathcal{E}_{-j}, \\ \boldsymbol{G}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i, & \text{on the event } \mathcal{E}_{-j}^c, \end{cases}$$

where $\boldsymbol{S}^{\star} = p^{-1}\sigma^2\left(\boldsymbol{\Sigma}^{\star}\right)^{-1}$,

$$\boldsymbol{S}_{-j} \triangleq \frac{\sigma^2}{p^2}\sum_{k:k\neq j}\delta_{ik}\left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2}\left(\boldsymbol{V}_{k,\cdot}^{\star}\right)^{\top}\boldsymbol{V}_{k,\cdot}^{\star}\left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2} \qquad \text{and} \qquad \mathcal{E}_{-j} \triangleq \left\{\|\boldsymbol{S}_{-j} - \boldsymbol{S}^{\star}\| \lesssim \frac{\sigma^2}{p\sigma_{\min}}\sqrt{\frac{\mu r\log n}{np}}\right\}.$$

It is easily seen that $\widetilde{\boldsymbol{Z}}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{\Sigma}^{\star})^{-1}/p)$; more importantly $\widetilde{\boldsymbol{Z}}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i$ is independent of $\{\delta_{kj}, E_{kj}\}_{1\leq k\leq n}$ and hence of $\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\boldsymbol{e}_j$.

We still need to verify the closeness between $\widetilde{\boldsymbol{Z}}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i$ and $\boldsymbol{Z}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i$. Towards this, we first repeat the proof in Appendix D.6 to obtain $\mathbb{P}(\mathcal{E}_{-j}) \geq 1 - O(n^{-10})$. Therefore on the high probability event $\mathcal{E}\cap\mathcal{E}_{-j}$, one has

$$\left\|\widetilde{\boldsymbol{Z}}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i - \boldsymbol{Z}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i\right\|_2 \leq \left\|\left(\boldsymbol{S}^{\star}\right)^{1/2}\right\|\left\|\boldsymbol{S}^{-1/2}\left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2}\sum_{k=1}^{n}\frac{1}{p}E_{ik}\delta_{ik}\left(\boldsymbol{V}_{k,\cdot}^{\star}\right)^{\top} - \boldsymbol{S}_{-j}^{-1/2}\left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2}\sum_{k:k\neq j}\frac{1}{p}E_{ik}\delta_{ik}\left(\boldsymbol{V}_{k,\cdot}^{\star}\right)^{\top}\right\|,$$

which together with the triangle inequality and the fact $\|\boldsymbol{S}^{\star}\| = \sigma^2/(p\sigma_{\min})$ yields

$$\begin{aligned}\sqrt{\frac{p\sigma_{\min}}{\sigma^2}}\left\|\widetilde{\boldsymbol{Z}}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i - \boldsymbol{Z}_{\boldsymbol{X}}^{\top}\boldsymbol{e}_i\right\|_2 &\leq \left\|\boldsymbol{S}^{-1/2} - \boldsymbol{S}_{-j}^{-1/2}\right\|\left\|\left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2}\right\|\left\|\sum_{k:k\neq j}\frac{1}{p}E_{ik}\delta_{ik}\boldsymbol{V}_{k,\cdot}^{\star}\right\|_2 \\ &\quad + \left\|\boldsymbol{S}^{-1/2}\right\|\left\|\left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2}\right\|\left\|\frac{1}{p}E_{ij}\delta_{ij}\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2\end{aligned}$$

$$\lesssim \left\| \boldsymbol{S}^{-1/2} - \boldsymbol{S}_{-j}^{-1/2} \right\| \frac{\sigma}{\sqrt{\sigma_{\min}}} \sqrt{\frac{r \log n}{p}} + \sqrt{\frac{p\sigma_{\max}}{\sigma^2}} \frac{1}{\sqrt{\sigma_{\min}}} \frac{\sigma \sqrt{\log n}}{p} \sqrt{\frac{\mu r}{n}}.$$

Here we have used the results in (D.32) and (D.33). We are left with bounding $\|\boldsymbol{S}^{-1/2} - \boldsymbol{S}_{-j}^{-1/2}\|$, for which we have

$$\|\boldsymbol{S} - \boldsymbol{S}_{-j}\| = \frac{\sigma^2}{p^2} \left\| \delta_{ij} \left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2} \left(\boldsymbol{V}_{j,\cdot}^{\star}\right)^{\top} \boldsymbol{V}_{j,\cdot}^{\star} \left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2} \right\| \leq \frac{\sigma^2}{p^2 \sigma_{\min}} \frac{\mu r}{n}.$$

Take the above bound collectively with (D.33) to yield

$$\left\| \boldsymbol{S}_{-j}^{-1/2} \right\| \lesssim \sqrt{\frac{p\sigma_{\max}}{\sigma^2}},$$

as long as $np \gg \kappa \mu r$. As a result, we have

$$\begin{aligned} \left\| \boldsymbol{S}^{-1/2} - \boldsymbol{S}_{-j}^{-1/2} \right\| &\leq \left\| \boldsymbol{S}^{-1/2} \right\| \left\| \boldsymbol{S}^{1/2} - \boldsymbol{S}_{-j}^{1/2} \right\| \left\| \boldsymbol{S}_{-j}^{-1/2} \right\| \\ &\lesssim \frac{p\sigma_{\max}}{\sigma^2} \cdot \frac{1}{\lambda_{\min}(\boldsymbol{S}^{1/2}) + \lambda_{\min}(\boldsymbol{S}_{-j}^{1/2})} \|\boldsymbol{S} - \boldsymbol{S}_{-j}\| \\ &\lesssim \frac{p\sigma_{\max}}{\sigma^2} \cdot \frac{1}{\sqrt{\frac{\sigma^2}{p\sigma_{\max}}}} \cdot \frac{\sigma^2}{p^2 \sigma_{\min}} \frac{\mu r}{n} \asymp \frac{\kappa \mu r}{np} \sqrt{\frac{p\sigma_{\max}}{\sigma^2}}, \end{aligned}$$

where the middle line relies on the perturbation of matrix square roots; see Lemma 13. Combining all, we arrive at

$$\sqrt{\frac{p\sigma_{\min}}{\sigma^2}} \left\| \widetilde{\boldsymbol{Z}}_{\boldsymbol{X}}^{\top} \boldsymbol{e}_i - \boldsymbol{Z}_{\boldsymbol{X}}^{\top} \boldsymbol{e}_i \right\|_2 \lesssim \frac{\kappa \mu r}{np} \cdot \sqrt{\kappa r \log n} + \sqrt{\frac{\kappa \mu r \log n}{np}} \asymp \sqrt{\frac{\kappa \mu r \log n}{np}},$$

with the proviso that $np \gg \kappa^2 \mu r^2 \log^2 n$. This finishes the proof. □

F Proof of Corollary 1

This section is dedicated to establishing the following result, which subsumes Corollary 1 as a special case.

Corollary 2. *Suppose that the conditions (3.18) hold, and recall the notation in Corollary 1. Then one has*

$$\begin{aligned} &\sup_{0<\alpha<1} \left| \mathbb{P}\left\{ M_{ij}^{\star} \in \left[M_{ij}^{\mathsf{d}} \pm \Phi^{-1}(1-\alpha/2) \sqrt{v_{ij}} \right] \right\} - (1-\alpha) \right| \\ &\quad \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\kappa^8 \mu r n \log n}{p}} + \sqrt{\frac{\kappa^8 \mu^3 r^3 \log^2 n}{np}} + \left(\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2 + \left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2 \right)^{-1} \sqrt{\frac{r}{n}} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\kappa^{10} \mu^2 r n \log^2 n}{p}}. \end{aligned}$$

Before entering the main proof of Corollary 2, we make a simple observation that

$$\max\left\{ \left\|\overline{\boldsymbol{X}}_{i,\cdot}^{\mathsf{d}} - \boldsymbol{X}_{i,\cdot}^{\star}\right\|_2, \left\|\overline{\boldsymbol{Y}}_{j,\cdot}^{\mathsf{d}} - \boldsymbol{Y}_{j,\cdot}^{\star}\right\|_2 \right\} \lesssim \kappa \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} \sqrt{\frac{\mu r \sigma_{\max}}{n}} \leq \frac{\sqrt{\sigma_{\max}}}{\kappa^2} \left(\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2 + \left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2 \right), \tag{F.1}$$

where we recall that $\overline{\boldsymbol{X}}^{\mathsf{d}} = \boldsymbol{X}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}}$ and $\overline{\boldsymbol{Y}}^{\mathsf{d}} = \boldsymbol{Y}^{\mathsf{d}} \boldsymbol{H}^{\mathsf{d}}$. Here, the first inequality arises from (A.13d) and the second one uses the assumption on $\|\boldsymbol{U}_{i,\cdot}^{\star}\|_2 + \|\boldsymbol{V}_{j,\cdot}^{\star}\|_2$ (i.e. (3.18b)). A simple consequence of (F.1) is that

$$\max\left\{ \left\|\overline{\boldsymbol{X}}_{i,\cdot}^{\mathsf{d}}\right\|_2, \left\|\overline{\boldsymbol{Y}}_{j,\cdot}^{\mathsf{d}}\right\|_2 \right\} \leq 2\sqrt{\sigma_{\max}} \left(\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2 + \left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2 \right). \tag{F.2}$$

Turning to the main proof, we define

$$\Delta_V \triangleq \frac{M_{ij}^{\mathsf{d}} - M_{ij}^{\star}}{\sqrt{v_{ij}}} - \frac{M_{ij}^{\mathsf{d}} - M_{ij}^{\star}}{\sqrt{v_{ij}^{\star}}}, \tag{F.3}$$

which in conjunction with Theorem 6 yields the following decomposition

$$\frac{M_{ij}^{\mathsf{d}}-M_{ij}^{\star}}{\sqrt{v_{ij}}}=\frac{M_{ij}^{\mathsf{d}}-M_{ij}^{\star}}{\sqrt{v_{ij}^{\star}}}+\Delta_{V}=\frac{g_{ij}}{\sqrt{v_{ij}^{\star}}}+\frac{\Delta_{ij}}{\sqrt{v_{ij}^{\star}}}+\Delta_{V}.$$

With this decomposition at hand, we have that for any $\varepsilon>0$,

$$\begin{aligned}\mathbb{P}\left(\frac{M_{ij}^{\mathsf{d}}-M_{ij}^{\star}}{\sqrt{v_{ij}}}\leq t\right)-\Phi\left(t\right)&=\mathbb{P}\left(\frac{g_{ij}}{\sqrt{v_{ij}^{\star}}}+\frac{\Delta_{ij}}{\sqrt{v_{ij}^{\star}}}+\Delta_{V}\leq t\right)-\Phi\left(t\right)\\&\leq\mathbb{P}\left(\frac{g_{ij}}{\sqrt{v_{ij}^{\star}}}\leq t+\varepsilon\right)+\mathbb{P}\left(\frac{\left|\Delta_{ij}\right|}{\sqrt{v_{ij}^{\star}}}+\left|\Delta_{V}\right|\geq\varepsilon\right)-\Phi\left(t\right)\\&\overset{\text{(i)}}{=}\Phi\left(t+\varepsilon\right)-\Phi\left(t\right)+\mathbb{P}\left(\left|\Delta_{ij}\right|+\left|\Delta_{V}\right|\sqrt{v_{ij}^{\star}}\geq\varepsilon\sqrt{v_{ij}^{\star}}\right)\\&\leq\varepsilon+\mathbb{P}\left(\left|\Delta_{ij}\right|+\left|\Delta_{V}\right|\sqrt{v_{ij}^{\star}}\geq\varepsilon\sqrt{v_{ij}^{\star}}\right),\end{aligned}$$

where $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$. Here, the relation (i) uses the fact that $g_{ij}\sim\mathcal{N}(0,v_{ij}^{\star})$. It then suffices to upper bound the right-hand side $\varepsilon+\mathbb{P}(|\Delta_{ij}|+|\Delta_{V}|\sqrt{v_{ij}^{\star}}\geq\varepsilon\sqrt{v_{ij}^{\star}})$. Our goal is to demonstrate that for a particular choice of $\varepsilon>0$, this quantity is well controlled. In view of Theorem 6, we know that $|\Delta_{ij}|$ is small with high probability. We are still in need of a high probability bound on the term $|\Delta_{V}|$, which we obtain through the following claim.

Claim 6. *With probability exceeding* $1-O(n^{-10})$, *the term* Δ_{V} *obeys*

$$|\Delta_{V}|\lesssim\left(\left\Vert \boldsymbol{U}_{i,\cdot}^{\star}\right\Vert_{2}+\left\Vert \boldsymbol{V}_{j,\cdot}^{\star}\right\Vert_{2}\right)^{-1}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^{10}\mu^{2}rn\log^{2}n}{p}}\sqrt{\frac{r}{n}}.$$

With Claim 6 at hand, we are ready to take

$$\varepsilon\asymp\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^{8}\mu rn\log n}{p}}+\sqrt{\frac{\kappa^{8}\mu^{3}r^{3}\log^{2}n}{np}}+\left(\left\Vert \boldsymbol{U}_{i,\cdot}^{\star}\right\Vert_{2}+\left\Vert \boldsymbol{V}_{j,\cdot}^{\star}\right\Vert_{2}\right)^{-1}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^{10}\mu^{2}rn\log^{2}n}{p}}\sqrt{\frac{r}{n}}$$

and arrive at the upper bound

$$\mathbb{P}\left(\frac{M_{ij}^{\mathsf{d}}-M_{ij}^{\star}}{\sqrt{v_{ij}}}\leq t\right)-\Phi\left(t\right)\leq\ \varepsilon+n^{-3}.$$

A similar argument yields the lower bound on $\mathbb{P}(M_{ij}^{\mathsf{d}}-M_{ij}^{\star}\leq t\sqrt{v_{ij}})-\Phi(t)$. As a result, one has

$$\begin{aligned}&\left|\mathbb{P}\left(\frac{M_{ij}^{\mathsf{d}}-M_{ij}^{\star}}{\sqrt{v_{ij}}}\leq t\right)-\Phi\left(t\right)\right|\lesssim\varepsilon+n^{-3}\\&\quad\asymp\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^{8}\mu rn\log n}{p}}+\sqrt{\frac{\kappa^{8}\mu^{3}r^{3}\log^{2}n}{np}}+\left(\left\Vert \boldsymbol{U}_{i,\cdot}^{\star}\right\Vert_{2}+\left\Vert \boldsymbol{V}_{j,\cdot}^{\star}\right\Vert_{2}\right)^{-1}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^{10}\mu^{2}rn\log^{2}n}{p}}\sqrt{\frac{r}{n}}\end{aligned}$$

for any t. This immediately establishes Corollary 2.

Proof of Claim 6. Recall that

$$\Delta_{V}=\left(M_{ij}^{\mathsf{d}}-M_{ij}^{\star}\right)\left[\left(v_{ij}\right)^{-1/2}-\left(v_{ij}^{\star}\right)^{-1/2}\right]=\left(M_{ij}^{\mathsf{d}}-M_{ij}^{\star}\right)\frac{v_{ij}^{\star}-v_{ij}}{\sqrt{v_{ij}^{\star}}\sqrt{v_{ij}}}\frac{1}{\sqrt{v_{ij}^{\star}}+\sqrt{v_{ij}}}.$$

Suppose for the moment that $|v_{ij}-v_{ij}^{\star}|\leq cv_{ij}^{\star}$ for some $c\leq1/2$. Then it follows immediately that

$$|\Delta_{V}|\lesssim c\frac{\left|M_{ij}^{\mathsf{d}}-M_{ij}^{\star}\right|}{\sqrt{v_{ij}^{\star}}}.$$

Therefore if suffices to control $|M_{ij}^{\mathsf{d}}-M_{ij}^{\star}|$ and $|v_{ij}^{\star}-v_{ij}|$ (i.e. obtaining the quantity c).

- First, expand M_{ij}^{d} and $M_{ij}^{\star}$ to see

$$\begin{aligned}\left|M_{ij}^{\mathsf{d}}-M_{ij}^{\star}\right| &= \left|\overline{\boldsymbol{X}}_{i,\cdot}^{\mathsf{d}}(\overline{\boldsymbol{Y}}_{j,\cdot}^{\mathsf{d}})^{\top}-\boldsymbol{X}_{i,\cdot}^{\star}(\boldsymbol{Y}_{j,\cdot}^{\star})^{\top}\right| \leq \|\overline{\boldsymbol{X}}_{i,\cdot}^{\mathsf{d}}-\boldsymbol{X}_{i,\cdot}^{\star}\|_2\|\overline{\boldsymbol{Y}}_{j,\cdot}^{\mathsf{d}}\|_2+\left\|\boldsymbol{X}_{i,\cdot}^{\star}\right\|_2\left\|\overline{\boldsymbol{Y}}_{j,\cdot}^{\mathsf{d}}-\boldsymbol{Y}_{j,\cdot}^{\star}\right\|_2\\ &\lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\sqrt{\frac{\mu r}{n}}\sigma_{\max}\left(\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2+\left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2\right)\\ &\lesssim \kappa^2\sqrt{\mu r\log n}\sqrt{v_{ij}^{\star}},\end{aligned}$$

where the middle line depends on (F.1) and (F.2), and the last inequality arises since $\sigma(\|\boldsymbol{U}_{i,\cdot}^{\star}\|_2+\|\boldsymbol{V}_{j,\cdot}^{\star}\|_2)/\sqrt{p}\lesssim\sqrt{v_{ij}^{\star}}$.

- Now we move on to $|v_{ij}^{\star}-v_{ij}|$. By the definition of v_{ij}, one has

$$\begin{aligned}\left|v_{ij}^{\star}-v_{ij}\right| \leq{}& \frac{\sigma^2}{p}\left|\boldsymbol{X}_{i,\cdot}^{\star}\left(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}\right)^{-1}(\boldsymbol{X}_{i,\cdot}^{\star})^{\top}-\boldsymbol{X}_{i,\cdot}^{\mathsf{d}}\left(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}}\right)^{-1}(\boldsymbol{X}_{i,\cdot}^{\mathsf{d}})^{\top}\right|\\ &+\frac{\sigma^2}{p}\left|\boldsymbol{Y}_{j,\cdot}^{\star}\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star}\right)^{-1}(\boldsymbol{Y}_{j,\cdot}^{\star})^{\top}-\boldsymbol{Y}_{j,\cdot}^{\mathsf{d}}\left(\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}\right)^{-1}(\boldsymbol{Y}_{j,\cdot}^{\mathsf{d}})^{\top}\right|.\end{aligned}$$

Focusing on the $\boldsymbol{X}$ factor, we have — with probability at least $1-O(n^{-10})$ — that

$$\begin{aligned}&\left|\boldsymbol{X}_{i,\cdot}^{\star}\left(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}\right)^{-1}(\boldsymbol{X}_{i,\cdot}^{\star})^{\top}-\boldsymbol{X}_{i,\cdot}^{\mathsf{d}}\left(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}}\right)^{-1}(\boldsymbol{X}_{i,\cdot}^{\mathsf{d}})^{\top}\right|\\ &\quad=\left|\boldsymbol{X}_{i,\cdot}^{\star}\left(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}\right)^{-1}(\boldsymbol{X}_{i,\cdot}^{\star})^{\top}-\overline{\boldsymbol{X}}_{i,\cdot}^{\mathsf{d}}(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}})^{-1}(\overline{\boldsymbol{X}}_{i,\cdot}^{\mathsf{d}})^{\top}\right|\\ &\quad\leq\left\|\boldsymbol{X}_{i,\cdot}^{\star}\left(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}\right)^{-1}\right\|_2\|\boldsymbol{X}_{i,\cdot}^{\star}-\overline{\boldsymbol{X}}_{i,\cdot}^{\mathsf{d}}\|_2+\|\boldsymbol{X}_{i,\cdot}^{\star}\|_2\left\|\left(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}\right)^{-1}-(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}})^{-1}\right\|\|\overline{\boldsymbol{X}}_{i,\cdot}^{\mathsf{d}}\|_2\\ &\qquad+\|\boldsymbol{X}_{i,\cdot}^{\star}-\overline{\boldsymbol{X}}_{i,\cdot}^{\mathsf{d}}\|_2\|(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}})^{-1}\|\|\overline{\boldsymbol{X}}_{i,\cdot}^{\mathsf{d}}\|_2.\end{aligned}\tag{F.4}$$

Here, the first relation comes from the identity $\boldsymbol{X}_{i,\cdot}^{\mathsf{d}}(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}})^{-1}(\boldsymbol{X}_{i,\cdot}^{\mathsf{d}})^{\top}=\overline{\boldsymbol{X}}_{i,\cdot}^{\mathsf{d}}(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}})^{-1}(\overline{\boldsymbol{X}}_{i,\cdot}^{\mathsf{d}})^{\top}$, and the inequality arises from the triangle inequality. Notice that $\|(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}})^{-1}\|\lesssim 1/\sigma_{\min}$ and that

$$\begin{aligned}\left\|\left(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}\right)^{-1}-(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}})^{-1}\right\| &\leq\left\|\left(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}\right)^{-1}\right\|\left\|\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}-\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}}\right\|\left\|(\overline{\boldsymbol{X}}^{\mathsf{d}\top}\overline{\boldsymbol{X}}^{\mathsf{d}})^{-1}\right\|\\ &\lesssim\frac{1}{\sigma_{\min}^2}\|\overline{\boldsymbol{X}}^{\mathsf{d}}-\boldsymbol{X}^{\star}\|\,\|\boldsymbol{X}^{\star}\|\lesssim\kappa^2\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\cdot\frac{1}{\sigma_{\min}},\end{aligned}$$

where the last inequality follows from (A.13b). Using the bounds (F.1) and (F.2), we continue the upper bound in (F.4) as follows

$$\begin{aligned}&\left|\boldsymbol{X}_{i,\cdot}^{\star}\left(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}\right)^{-1}(\boldsymbol{X}_{i,\cdot}^{\star})^{\top}-\boldsymbol{X}_{i,\cdot}^{\mathsf{d}}\left(\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}}\right)^{-1}(\boldsymbol{X}_{i,\cdot}^{\mathsf{d}})^{\top}\right|\\ &\quad\lesssim\frac{1}{\sqrt{\sigma_{\min}}}\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2\cdot\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\sqrt{\frac{\mu r\sigma_{\max}}{n}}\\ &\qquad+\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2\sqrt{\sigma_{\max}}\cdot\kappa^2\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\frac{1}{\sigma_{\min}}\cdot\sqrt{\sigma_{\max}}\left(\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2+\left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2\right)\\ &\qquad+\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\sqrt{\frac{\mu r\sigma_{\max}}{n}}\cdot\frac{1}{\sigma_{\min}}\cdot\sqrt{\sigma_{\max}}\left(\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2+\left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2\right)\\ &\quad\lesssim\kappa^3\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\sqrt{\frac{\mu r}{n}}\left(\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2+\left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2\right).\end{aligned}$$

A similar bound holds for the factor $\boldsymbol{Y}$. Therefore, with high probability we have

$$\left|v_{ij}^{\star}-v_{ij}\right|\lesssim\frac{\sigma^2}{p}\kappa^3\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\sqrt{\frac{\mu r}{n}}\left(\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2+\left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2\right)$$

$$\lesssim \left(\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2 + \left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2\right)^{-1} \kappa^3 \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n\log n}{p}} \sqrt{\frac{\mu r}{n}} V_{ij}^{\star} \le \frac{1}{2} v_{ij}^{\star},$$

where the last relation results from the condition on $\|\boldsymbol{U}_{i,\cdot}^{\star}\|_2 + \|\boldsymbol{V}_{j,\cdot}^{\star}\|_2$ (cf. (3.18b)).

Combine the bounds on $|M_{ij}^{\mathsf{d}} - M_{ij}^{\star}|$ and $|v_{ij}^{\star} - v_{ij}|$ to see that with probability exceeding $1 - O(n^{-10})$,

$$\begin{aligned} |\Delta_V| &\lesssim \left(\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2 + \left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2\right)^{-1} \kappa^3 \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n\log n}{p}} \sqrt{\frac{\mu r}{n}} \cdot \kappa^2 \sqrt{\mu r \log n} \\ &\lesssim \left(\left\|\boldsymbol{U}_{i,\cdot}^{\star}\right\|_2 + \left\|\boldsymbol{V}_{j,\cdot}^{\star}\right\|_2\right)^{-1} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\kappa^{10} n \mu^2 r \log^2 n}{p}} \sqrt{\frac{r}{n}}. \end{aligned}$$

This establishes the desired upper bound on $|\Delta_V|$. □

G Proof of Theorem 3

As we have argued in Section 5.1, it suffices to prove the claim for $\boldsymbol{M}^{\mathsf{d}} = \boldsymbol{X}^{\mathsf{d}} \boldsymbol{X}^{\mathsf{d}\top} = \overline{\boldsymbol{X}}^{\mathsf{d}} \overline{\boldsymbol{Y}}^{\mathsf{d}\top}$. For simplicity of notation, we define

$$\boldsymbol{\Gamma}_{\boldsymbol{X}} \triangleq \overline{\boldsymbol{X}}^{\mathsf{d}} - \boldsymbol{X}^{\star} \qquad \text{and} \qquad \boldsymbol{\Gamma}_{\boldsymbol{Y}} \triangleq \overline{\boldsymbol{Y}}^{\mathsf{d}} - \boldsymbol{Y}^{\star}.$$

Apply the decompositions in Theorem 5 to obtain

$$\begin{aligned} \boldsymbol{M}^{\mathsf{d}} - \boldsymbol{M}^{\star} &= \overline{\boldsymbol{X}}^{\mathsf{d}} \overline{\boldsymbol{Y}}^{\mathsf{d}\top} - \boldsymbol{X}^{\star} \boldsymbol{Y}^{\star\top} \\ &= \boldsymbol{\Gamma}_{\boldsymbol{X}} \boldsymbol{Y}^{\star\top} + \boldsymbol{X}^{\star} \boldsymbol{\Gamma}_{\boldsymbol{Y}}^{\top} + \boldsymbol{\Gamma}_{\boldsymbol{X}} \boldsymbol{\Gamma}_{\boldsymbol{Y}}^{\top} \\ &= \boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{Y}^{\star\top} + \boldsymbol{X}^{\star} \boldsymbol{Z}_{\boldsymbol{Y}}^{\top} + \underbrace{\boldsymbol{\Psi}_{\boldsymbol{X}} \boldsymbol{Y}^{\star\top} + \boldsymbol{X}^{\star} \boldsymbol{\Psi}_{\boldsymbol{Y}}^{\top} + \boldsymbol{\Gamma}_{\boldsymbol{X}} \boldsymbol{\Gamma}_{\boldsymbol{Y}}^{\top}}_{\triangleq \boldsymbol{\Theta}}, \end{aligned} \tag{G.1}$$

where $\boldsymbol{\Psi}_{\boldsymbol{X}}$ and $\boldsymbol{\Psi}_{\boldsymbol{Y}}$ are defined in Theorem 5. Further, expand $\|\boldsymbol{M}^{\mathsf{d}} - \boldsymbol{M}^{\star}\|_{\mathrm{F}}^2$ to obtain

$$\left\|\boldsymbol{M}^{\mathsf{d}} - \boldsymbol{M}^{\star}\right\|_{\mathrm{F}}^2 = \left\|\boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{X}^{\star} \boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\right\|_{\mathrm{F}}^2 + \mathsf{rem},$$

where we define the remainder term as

$$\mathsf{rem} \triangleq 2\mathsf{Tr}\left(\boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{Y}^{\star\top} \boldsymbol{Z}_{\boldsymbol{Y}} \boldsymbol{X}^{\star\top}\right) + \|\boldsymbol{\Theta}\|_{\mathrm{F}}^2 + 2\mathsf{Tr}\left(\boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{Y}^{\star\top} \boldsymbol{\Theta}^{\top}\right) + 2\mathsf{Tr}\left(\boldsymbol{X}^{\star} \boldsymbol{Z}_{\boldsymbol{Y}}^{\top} \boldsymbol{\Theta}^{\top}\right).$$

In what follows, we aim to demonstrate that $\|\boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{Y}^{\star}\|_{\mathrm{F}}^2 + \|\boldsymbol{X}^{\star} \boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\|_{\mathrm{F}}^2$, which can be shown to sharply concentrate around its mean, is the dominant term, and the remainder term rem is much smaller in magnitude with high probability.

- We begin with the term $\|\boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{Y}^{\star}\|_{\mathrm{F}}^2 + \|\boldsymbol{X}^{\star} \boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\|_{\mathrm{F}}^2$. We shall focus on bounding $\|\boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{Y}^{\star}\|_{\mathrm{F}}^2$ since the other term $\|\boldsymbol{X}^{\star} \boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\|_{\mathrm{F}}^2$ can be treated analogously. To this end, we first have the identity

$$\frac{p}{\sigma^2} \left\|\boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}^2 = \frac{p}{\sigma^2} \mathsf{Tr}\left(\boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{Y}^{\star\top} \boldsymbol{Y}^{\star} \boldsymbol{Z}_{\boldsymbol{X}}^{\top}\right) = \frac{p}{\sigma^2} \mathsf{Tr}\left(\boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{\Sigma}^{\star} \boldsymbol{Z}_{\boldsymbol{X}}^{\top}\right) = \sum_{i=1}^{n} \left\| \frac{\sqrt{p}}{\sigma} \left(\boldsymbol{\Sigma}^{\star}\right)^{1/2} \boldsymbol{Z}_{\boldsymbol{X}}^{\top} \boldsymbol{e}_i \right\|_2^2,$$

where we use the fact that $\boldsymbol{Y}^{\star\top} \boldsymbol{Y}^{\star} = \boldsymbol{\Sigma}^{\star}$. Theorem 5 tells us that $\boldsymbol{Z}_{\boldsymbol{X}}^{\top} \boldsymbol{e}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2 (\boldsymbol{\Sigma}^{\star})^{-1}/p)$, which further implies

$$\frac{\sqrt{p}}{\sigma} \left(\boldsymbol{\Sigma}^{\star}\right)^{1/2} \boldsymbol{Z}_{\boldsymbol{X}}^{\top} \boldsymbol{e}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_r\right).$$

Therefore, the quantity $p\|\boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{Y}^{\star}\|_{\mathrm{F}}^2/\sigma^2$ follows the chi-squared distribution with nr degrees of freedom. Standard concentration inequalities [Wai19, Equation (2.19)] reveals that with probability at least $1 - O(n^{-10})$,

$$\left| \frac{p}{\sigma^2} \left\|\boldsymbol{Z}_{\boldsymbol{X}} \boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}^2 - nr \right| \lesssim \sqrt{nr \log n}.$$

Repeating the above argument for $\|\boldsymbol{X}^{\star}\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\|_{\mathrm{F}}^{2}$, we conclude that with probability at least $1-O(n^{-10})$,

$$\left\|\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{X}^{\star}\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\right\|_{\mathrm{F}}^{2}=2\frac{\sigma^{2}nr}{p}+O\left(\frac{\sigma^{2}}{p}\sqrt{nr\log n}\right)=(2+o(1))\frac{\sigma^{2}nr}{p}.$$

- Now we turn to the term rem, for which we have the following two claims.

  Claim 7. *With probability at least $1-O(n^{-10})$, one has*

  $$\left|\|\boldsymbol{\Theta}\|_{\mathrm{F}}^{2}+2\mathsf{Tr}\left(\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{\Theta}^{\top}\right)+2\mathsf{Tr}\left(\boldsymbol{X}^{\star}\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\boldsymbol{\Theta}^{\top}\right)\right|=o\left(\frac{\sigma^{2}nr}{p}\right).$$

  Claim 8. *With probability exceeding $1-O(n^{-10})$, we have*

  $$\left|\mathsf{Tr}\left(\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{Z}_{\boldsymbol{Y}}\boldsymbol{X}^{\star\top}\right)\right|=o\left(\frac{\sigma^{2}nr}{p}\right).$$

Combine all of the above bounds to yield the desired result.

Proof of Claim 7. Use triangle inequality and the bound $|\mathsf{Tr}(\boldsymbol{A}\boldsymbol{B})|\leq\|\boldsymbol{A}\|_{\mathrm{F}}\|\boldsymbol{B}\|_{\mathrm{F}}$ to obtain

$$\begin{aligned}\left|\|\boldsymbol{\Theta}\|_{\mathrm{F}}^{2}+2\mathsf{Tr}\left(\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{\Theta}^{\top}\right)+2\mathsf{Tr}\left(\boldsymbol{X}^{\star}\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\boldsymbol{\Theta}^{\top}\right)\right| &\leq\|\boldsymbol{\Theta}\|_{\mathrm{F}}^{2}+2\left\|\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}\|\boldsymbol{\Theta}\|_{\mathrm{F}}+2\left\|\boldsymbol{X}^{\star}\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\right\|_{\mathrm{F}}\|\boldsymbol{\Theta}\|_{\mathrm{F}}\\ &=\left(\|\boldsymbol{\Theta}\|_{\mathrm{F}}+2\left\|\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}+2\left\|\boldsymbol{X}^{\star}\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\right\|_{\mathrm{F}}\right)\|\boldsymbol{\Theta}\|_{\mathrm{F}}. \qquad \text{(G.2)}\end{aligned}$$

Plug in the definition of $\boldsymbol{\Theta}$ (cf. (G.1)) and invoke the triangle inequality again to see that

$$\begin{aligned}\|\boldsymbol{\Theta}\|_{\mathrm{F}}&\leq\left\|\boldsymbol{\Psi}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}+\left\|\boldsymbol{X}^{\star}\boldsymbol{\Psi}_{\boldsymbol{Y}}^{\top}\right\|_{\mathrm{F}}+\left\|\boldsymbol{\Gamma}_{\boldsymbol{X}}\boldsymbol{\Gamma}_{\boldsymbol{Y}}^{\top}\right\|_{\mathrm{F}}\\ &\leq\|\boldsymbol{\Psi}_{\boldsymbol{X}}\|_{\mathrm{F}}\sqrt{\sigma_{\max}}+\sqrt{\sigma_{\max}}\|\boldsymbol{\Psi}_{\boldsymbol{Y}}\|_{\mathrm{F}}+\|\boldsymbol{\Gamma}_{\boldsymbol{X}}\|_{\mathrm{F}}\|\boldsymbol{\Gamma}_{\boldsymbol{Y}}\|_{\mathrm{F}}\\ &\leq\sqrt{n\sigma_{\max}}(\|\boldsymbol{\Psi}_{\boldsymbol{X}}\|_{2,\infty}+\|\boldsymbol{\Psi}_{\boldsymbol{Y}}\|_{2,\infty})+\|\boldsymbol{\Gamma}_{\boldsymbol{X}}\|_{\mathrm{F}}\|\boldsymbol{\Gamma}_{\boldsymbol{Y}}\|_{\mathrm{F}}.\end{aligned}$$

Combine Theorem 5 and the fact $\max\{\|\boldsymbol{\Gamma}_{\boldsymbol{X}}\|_{\mathrm{F}},\|\boldsymbol{\Gamma}_{\boldsymbol{Y}}\|_{\mathrm{F}}\}\lesssim(\sigma/\sigma_{\min})\sqrt{n/p}\|\boldsymbol{X}^{\star}\|_{\mathrm{F}}$ (see (A.13c)) to conclude that with probability at least $1-O(n^{-3})$

$$\begin{aligned}\|\boldsymbol{\Theta}\|_{\mathrm{F}}&\lesssim\sqrt{n\sigma_{\max}}\frac{\sigma}{\sqrt{p\sigma_{\min}}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{\kappa^{7}\mu rn\log n}{p}}+\sqrt{\frac{\kappa^{7}\mu^{3}r^{3}\log^{2}n}{np}}\right)+\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\sqrt{r\sigma_{\max}}\right)^{2}\\ &=o\left(\sigma\sqrt{\frac{nr}{p}}\right).\end{aligned}$$

Here the last relation depends on the assumption (3.18a). Second, we have already established in this section that

$$\|\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star}\|_{\mathrm{F}}+\|\boldsymbol{X}^{\star}\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\|_{\mathrm{F}}=O(\sigma\sqrt{nr/p})$$

with probability exceeding $1-O(n^{-10})$. Substitute the above two facts into (G.2) to arrive at

$$\left|\|\boldsymbol{\Theta}\|_{\mathrm{F}}^{2}+2\mathsf{Tr}\left(\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{\Theta}^{\top}\right)+2\mathsf{Tr}\left(\boldsymbol{X}^{\star}\boldsymbol{Z}_{\boldsymbol{Y}}^{\top}\boldsymbol{\Theta}^{\top}\right)\right|\lesssim\sigma\sqrt{\frac{nr}{p}}\|\boldsymbol{\Theta}\|_{\mathrm{F}}=o\left(\frac{\sigma^{2}nr}{p}\right).$$

This concludes the proof. □

Proof of Claim 8. According to Lemma 9, one can write

$$\boldsymbol{Z}_{\boldsymbol{X}}=\underbrace{\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{Y}^{\star}\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star}\right)^{-1}}_{\triangleq\boldsymbol{Z}_{\boldsymbol{X},\boldsymbol{E}}}-\boldsymbol{\Delta}_{\boldsymbol{X}},\qquad\boldsymbol{Z}_{\boldsymbol{Y}}=\underbrace{\frac{1}{p}\left[\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right]^{\top}\boldsymbol{X}^{\star}\left(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}\right)^{-1}}_{\triangleq\boldsymbol{Z}_{\boldsymbol{Y},\boldsymbol{E}}}-\boldsymbol{\Delta}_{\boldsymbol{Y}},$$

where $\max\left\{\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{2,\infty},\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{2,\infty}\right\}\lesssim\frac{\sigma}{\sqrt{p\sigma_{\min}}}\sqrt{\frac{\kappa^2\mu r^2\log^2 n}{np}}$ and hence

$$\max\left\{\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{\mathrm{F}},\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{\mathrm{F}}\right\}\lesssim\sqrt{n}\max\left\{\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{2,\infty},\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{2,\infty}\right\}\lesssim\frac{\sigma}{\sqrt{p\sigma_{\min}}}\sqrt{\frac{\kappa^2\mu r^2\log^2 n}{p}}. \tag{G.3}$$

Consequently, use the triangle inequality and Cauchy-Schwarz to verify that

$$\begin{aligned}
&\left|\mathsf{Tr}\big(\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{Z}_{\boldsymbol{Y}}\boldsymbol{X}^{\star\top}\big)-\mathsf{Tr}\big(\boldsymbol{Z}_{\boldsymbol{X},\boldsymbol{E}}\boldsymbol{Y}^{\star\top}\boldsymbol{Z}_{\boldsymbol{Y},\boldsymbol{E}}\boldsymbol{X}^{\star\top}\big)\right|\\
&\quad\leq\left|\mathsf{Tr}\big(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{Z}_{\boldsymbol{Y}}\boldsymbol{X}^{\star\top}\big)\right|+\left|\mathsf{Tr}\big(\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{\Delta}_{\boldsymbol{Y}}\boldsymbol{X}^{\star\top}\big)\right|+\left|\mathsf{Tr}\big(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{\Delta}_{\boldsymbol{Y}}\boldsymbol{X}^{\star\top}\big)\right|\\
&\quad\leq\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{\mathrm{F}}\|\boldsymbol{Y}^{\star}\|\|\boldsymbol{Z}_{\boldsymbol{Y}}\boldsymbol{X}^{\star\top}\|_{\mathrm{F}}+\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{\mathrm{F}}\|\boldsymbol{X}^{\star}\|\|\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\|_{\mathrm{F}}+\|\boldsymbol{X}^{\star}\|\|\boldsymbol{Y}^{\star}\|\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{\mathrm{F}}\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{\mathrm{F}}\\
&\quad\overset{\text{(i)}}{\leq}\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{\mathrm{F}}\|\boldsymbol{Y}^{\star}\|\|\boldsymbol{Z}_{\boldsymbol{Y}}\boldsymbol{\Sigma}^{\star1/2}\|_{\mathrm{F}}+\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{\mathrm{F}}\|\boldsymbol{X}^{\star}\|\|\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{\Sigma}^{\star1/2}\|_{\mathrm{F}}+\|\boldsymbol{X}^{\star}\|\|\boldsymbol{Y}^{\star}\|\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{\mathrm{F}}\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{\mathrm{F}} &&\text{(G.4)}\\
&\quad\overset{\text{(ii)}}{\lesssim}\frac{\sigma}{\sqrt{p}}\sqrt{\frac{\kappa^3\mu r^2\log^2 n}{p}}\|\boldsymbol{Z}_{\boldsymbol{Y}}\boldsymbol{\Sigma}^{\star1/2}\|_{\mathrm{F}}+\frac{\sigma}{\sqrt{p}}\sqrt{\frac{\kappa^3\mu r^2\log^2 n}{p}}\|\boldsymbol{Z}_{\boldsymbol{Y}}\boldsymbol{\Sigma}^{\star1/2}\|_{\mathrm{F}}+\frac{\sigma^2}{p}\cdot\frac{\kappa^3\mu r^2\log^2 n}{p}, &&\text{(G.5)}
\end{aligned}$$

where (i) follows since $\boldsymbol{X}^{\star}=\boldsymbol{U}^{\star}\boldsymbol{\Sigma}^{\star1/2}$ and $\|\boldsymbol{U}^{\star}\|=1$, and (ii) makes use of (G.3) as well as the facts $\|\boldsymbol{Y}^{\star}\|,\|\boldsymbol{X}^{\star}\|=\sqrt{\sigma_{\max}}$. In addition, invoke Lemma 9 to see that $\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{\Sigma}^{\star1/2}$ and $\boldsymbol{Z}_{\boldsymbol{Y}}\boldsymbol{\Sigma}^{\star1/2}$ are both Gaussian matrices with i.i.d. $\mathcal{N}(0,\sigma^2/p)$ entries, which together with standard concentration results implies that

$$\|\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{\Sigma}^{\star1/2}\|_{\mathrm{F}}=(1+o(1))\sigma\sqrt{nr/p};\qquad\|\boldsymbol{Z}_{\boldsymbol{Y}}\boldsymbol{\Sigma}^{\star1/2}\|_{\mathrm{F}}=(1+o(1))\sigma\sqrt{nr/p}.$$

Substituting it into (G.5) gives

$$\left|\mathsf{Tr}\big(\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{Z}_{\boldsymbol{Y}}\boldsymbol{X}^{\star\top}\big)-\mathsf{Tr}\big(\boldsymbol{Z}_{\boldsymbol{X},\boldsymbol{E}}\boldsymbol{Y}^{\star\top}\boldsymbol{Z}_{\boldsymbol{Y},\boldsymbol{E}}\boldsymbol{X}^{\star\top}\big)\right|\lesssim\frac{\sigma^2}{p}\sqrt{\frac{\kappa^3\mu nr^3\log^2 n}{p}}+\frac{\sigma^2}{p}\frac{\kappa^3\mu r^2\log^2 n}{p}\asymp o\left(\frac{\sigma^2 nr}{p}\right),$$

with the proviso that $np\gtrsim\kappa^3\mu r\log^3 n$. This means that, with high probability,

$$\mathsf{Tr}\big(\boldsymbol{Z}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\boldsymbol{Z}_{\boldsymbol{Y}}\boldsymbol{X}^{\star\top}\big)=\mathsf{Tr}\big(\boldsymbol{Z}_{\boldsymbol{X},\boldsymbol{E}}\boldsymbol{Y}^{\star\top}\boldsymbol{Z}_{\boldsymbol{Y},\boldsymbol{E}}\boldsymbol{X}^{\star\top}\big)+o\left(\sigma^2 nr/p\right). \tag{G.6}$$

Everything then boils down to controlling $\mathsf{Tr}\big(\boldsymbol{Z}_{\boldsymbol{X},\boldsymbol{E}}\boldsymbol{Y}^{\star\top}\boldsymbol{Z}_{\boldsymbol{Y},\boldsymbol{E}}\boldsymbol{X}^{\star\top}\big)$. Towards this end, we first note that

$$\boldsymbol{Z}_{\boldsymbol{X},\boldsymbol{E}}\boldsymbol{Y}^{\star\top}=p^{-1}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{Y}^{\star}\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star}\right)^{-1}\boldsymbol{Y}^{\star\top}=p^{-1}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{V}^{\star}\boldsymbol{V}^{\star\top}.$$

Similarly, $\boldsymbol{Z}_{\boldsymbol{Y},\boldsymbol{E}}\boldsymbol{X}^{\star\top}=[\mathcal{P}_{\Omega}(\boldsymbol{E})]^{\top}\boldsymbol{U}^{\star}\boldsymbol{U}^{\star\top}/p$. These identities allow us to derive

$$\begin{aligned}
\mathsf{Tr}\left(\boldsymbol{Z}_{\boldsymbol{X},\boldsymbol{E}}\boldsymbol{Y}^{\star\top}\boldsymbol{Z}_{\boldsymbol{Y},\boldsymbol{E}}\boldsymbol{X}^{\star\top}\right)&=\frac{1}{p^2}\mathsf{Tr}\left(\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{V}^{\star}\boldsymbol{V}^{\star\top}\left[\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right]^{\top}\boldsymbol{U}^{\star}\boldsymbol{U}^{\star\top}\right)\\
&=\frac{1}{p^2}\mathsf{Tr}\left(\boldsymbol{U}^{\star\top}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{V}^{\star}\boldsymbol{V}^{\star\top}\left[\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right]^{\top}\boldsymbol{U}^{\star}\right)\\
&=\left\|\boldsymbol{U}^{\star\top}\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{V}^{\star}\right\|_{\mathrm{F}}^2.
\end{aligned}$$

Apply the same arguments in controlling (D.21) to obtain that with probability at least $1-O(n^{-10})$,

$$\left\|\boldsymbol{U}^{\star\top}\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{V}^{\star}\right\|_{\mathrm{F}}^2\lesssim\sigma^2\frac{\log n}{p}\left\|\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}^2\left\|\boldsymbol{V}^{\star}\right\|_{\mathrm{F}}^2\asymp\frac{\sigma^2 r^2\log n}{p}=o\left(\frac{\sigma^2 nr}{p}\right),$$

as long as $n\gtrsim r\log^2 n$. This combined with (G.6) yields the desired claim. □

H Proof of lower bounds

H.1 Proof of Lemma 1

Fix any $\varepsilon > 0$. It suffices to prove that the matrix $\mathsf{CRLB}(\boldsymbol{X}^\star_{i,\cdot} \mid \Omega)$ defined in (3.27) satisfies

$$\left\| \frac{p}{\sigma^2}\mathsf{CRLB}(\boldsymbol{X}^\star_{i,\cdot} \mid \Omega) - (\boldsymbol{\Sigma}^\star)^{-1} \right\| \le \frac{\varepsilon}{\sigma_{\max}} \tag{H.1}$$

with probability at least $1 - O(n^{-10})$, provided that $np \ge C_0 \varepsilon^{-2}\kappa^4 \mu r$. Towards this end, we first compute

$$\mathsf{CRLB}(\boldsymbol{X}^\star_{i,\cdot} \mid \Omega) = \sigma^2 \Bigg(\sum_{k:(i,k)\in\Omega} (\boldsymbol{Y}^\star_{k,\cdot})^\top \boldsymbol{Y}^\star_{k,\cdot} \Bigg)^{-1} = \frac{\sigma^2}{p} \Bigg(\underbrace{\frac{1}{p}\sum_{k=1}^{n} \delta_{ik} (\boldsymbol{Y}^\star_{k,\cdot})^\top \boldsymbol{Y}^\star_{k,\cdot}}_{:=\boldsymbol{A}} \Bigg)^{-1},$$

where we recall that $\delta_{ik} = \mathbb{1}\{(i,k) \in \Omega\}$. Next, define the following event

$$\mathcal{E} \triangleq \left\{ \left\| \boldsymbol{A} - \boldsymbol{\Sigma}^\star \right\| \le C\sqrt{\frac{\mu r \log n}{np}}\sigma_{\max} \right\},$$

where $C > 0$ is some large absolute constant. On the event $\mathcal{E}$, in view of the fact $\sigma_{\min}\boldsymbol{I}_r \preceq \boldsymbol{\Sigma}^\star \preceq \sigma_{\max}\boldsymbol{I}_r$, one has

$$0.5\sigma_{\min}\boldsymbol{I}_r \preceq \boldsymbol{A} \preceq 2\sigma_{\max}\boldsymbol{I}_r,$$

with the proviso that $np \ge 4C^2\kappa^2\mu r \log n$. This further implies that

$$\begin{aligned}
\left\| \frac{p}{\sigma^2}\mathsf{CRLB}(\boldsymbol{X}^\star_{i,\cdot} \mid \Omega) - (\boldsymbol{\Sigma}^\star)^{-1} \right\| &= \left\| \boldsymbol{A}^{-1} - (\boldsymbol{\Sigma}^\star)^{-1} \right\| \le \left\| \boldsymbol{A} - \boldsymbol{\Sigma}^\star \right\| \cdot \left\| \boldsymbol{A}^{-1} \right\| \cdot \left\| (\boldsymbol{\Sigma}^\star)^{-1} \right\| \\
&\le \frac{2C}{\sigma_{\min}} \sqrt{\frac{\kappa^2 \mu r \log n}{np}}
\end{aligned}$$

on the event $\mathcal{E}$. Clearly, the requirement (H.1) holds true if $np \ge C_0 \varepsilon^{-2}\kappa^4 \mu r \log n$ with $C_0 = 4C^2$.

To finish up, we are left with proving that $\mathcal{E}$ occurs with probability at least $1 - O(n^{-10})$. Invoke the matrix Bernstein inequality to show that

$$\left\| \boldsymbol{A} - \boldsymbol{\Sigma}^\star \right\| = \frac{1}{p}\left\| \sum_{k=1}^{n} (\delta_{ik} - p)\,(\boldsymbol{Y}^\star_{k,\cdot})^\top \boldsymbol{Y}^\star_{k,\cdot} \right\| \lesssim \frac{1}{p}\left(\sqrt{V \log n} + B \log n \right)$$

holds with probability at least $1 - O(n^{-10})$, where we define

$$\begin{aligned}
B &\triangleq \max_{1\le k\le n} \left\| (\delta_{ik} - p)\,(\boldsymbol{Y}^\star_{k,\cdot})^\top \boldsymbol{Y}^\star_{k,\cdot} \right\| \le \left\| \boldsymbol{Y}^\star \right\|_{2,\infty}^2 \le \mu r \sigma_{\max}/n, \\
V &\triangleq \left\| \sum_{k=1}^{n} \mathbb{E}\left[(\delta_{ik} - p)^2\,(\boldsymbol{Y}^\star_{k,\cdot})^\top \boldsymbol{Y}^\star_{k,\cdot} (\boldsymbol{Y}^\star_{k,\cdot})^\top \boldsymbol{Y}^\star_{k,\cdot} \right] \right\| \le p \left\| \sum_{k=1}^{n} (\boldsymbol{Y}^\star_{k,\cdot})^\top \boldsymbol{Y}^\star_{k,\cdot} (\boldsymbol{Y}^\star_{k,\cdot})^\top \boldsymbol{Y}^\star_{k,\cdot} \right\| \\
&\le p \left\| \boldsymbol{Y}^\star \right\|_{2,\infty}^2 \left\| \boldsymbol{Y}^{\star\top} \boldsymbol{Y}^\star \right\| \le \mu r p \sigma_{\max}^2/n.
\end{aligned}$$

Here we have used the incoherence condition (A.17). Consequently, one reaches the conclusion that with probability exceeding $1 - O(n^{-10})$,

$$\left\| \boldsymbol{A} - \boldsymbol{\Sigma}^\star \right\| \lesssim \frac{1}{p}\left(\sqrt{\frac{\mu r p \sigma_{\max}^2}{n} \log n} + \frac{\mu r \sigma_{\max}}{n} \log n \right) \asymp \sqrt{\frac{\mu r \log n}{np}}\sigma_{\max}$$

as long as $np \gg \mu r \log n$, thus concluding the proof.

H.2 Proof of Lemma 2

The proof strategy is similar to the one used in proving Lemma 1 (cf. Appendix H.1). Fix any $\varepsilon > 0$. It is sufficient to establish the following inequality

$$\frac{p}{\sigma^2}\left|\mathsf{CRLB}(M_{ij}^\star \mid \Omega) - v_{ij}^\star\right| \le \varepsilon \frac{p}{\sigma^2} v_{ij}^\star, \tag{H.2}$$

where the scalar $\mathsf{CRLB}(M_{ij}^\star \mid \Omega)$ is defined in (3.28) and $v_{ij}^\star$ is defined in Theorem 2. Expand the left-hand side to reach

$$\begin{aligned}\frac{p}{\sigma^2}\left|\mathsf{CRLB}(M_{ij}^\star \mid \Omega) - v_{ij}^\star\right| &\le \Bigg|\boldsymbol{Y}_{j,\cdot}^\star \underbrace{\Bigg(\frac{1}{p}\sum_{k:k\neq j,(i,k)\in\Omega}(\boldsymbol{Y}_{k,\cdot}^\star)^\top \boldsymbol{Y}_{k,\cdot}^\star\Bigg)}_{:=\boldsymbol{A}_Y}{}^{-1}(\boldsymbol{Y}_{j,\cdot}^\star)^\top - \boldsymbol{Y}_{j,\cdot}^\star(\boldsymbol{\Sigma}^\star)^{-1}(\boldsymbol{Y}_{j,\cdot}^\star)^\top\Bigg| \\ &\quad + \Bigg|\boldsymbol{X}_{i,\cdot}^\star \underbrace{\Bigg(\frac{1}{p}\sum_{k:k\neq i,(k,j)\in\Omega}(\boldsymbol{X}_{k,\cdot}^\star)^\top \boldsymbol{X}_{k,\cdot}^\star\Bigg)}_{:=\boldsymbol{A}_X}{}^{-1}(\boldsymbol{X}_{i,\cdot}^\star)^\top - \boldsymbol{X}_{j,\cdot}^\star(\boldsymbol{\Sigma}^\star)^{-1}(\boldsymbol{X}_{j,\cdot}^\star)^\top\Bigg| \\ &\le \left\|\boldsymbol{V}_{j,\cdot}^\star\right\|_2^2 \sigma_{\max}\left\|\boldsymbol{A}_Y^{-1} - (\boldsymbol{\Sigma}^\star)^{-1}\right\| + \left\|\boldsymbol{U}_{i,\cdot}^\star\right\|_2^2 \sigma_{\max}\left\|\boldsymbol{A}_X^{-1} - (\boldsymbol{\Sigma}^\star)^{-1}\right\|,\end{aligned}$$

where the last line follows from the observations that $\|\boldsymbol{Y}_{j,\cdot}^\star\|_2 \le \sqrt{\sigma_{\max}}\|\boldsymbol{V}_{j,\cdot}^\star\|_2$ and $\|\boldsymbol{X}_{i,\cdot}^\star\|_2 \le \sqrt{\sigma_{\max}}\|\boldsymbol{U}_{i,\cdot}^\star\|_2$.

Define the following event

$$\mathcal{E}_2 \triangleq \left\{\max\left\{\left\|\boldsymbol{A}_Y - \boldsymbol{\Sigma}^\star\right\|, \left\|\boldsymbol{A}_X - \boldsymbol{\Sigma}^\star\right\|\right\} \le C\sqrt{\frac{\mu r \log n}{np}}\sigma_{\max}\right\},$$

where $C > 0$ is some large universal constant. Two observations are sufficient to derive the desired the result (H.2). First, the event $\mathcal{E}_2$ happens with probability at least $1 - O(n^{-10})$ — an easy consequence of the proof of Lemma 1 (cf. Appendix H.1). Second, on the event $\mathcal{E}_2$, repeating the same proof of Lemma 1 (cf. Appendix H.1), one can deduce that

$$\frac{p}{\sigma^2}\left|\mathsf{CRLB}(M_{ij}^\star \mid \Omega) - v_{ij}^\star\right| \le \left(\left\|\boldsymbol{U}_{i,\cdot}^\star\right\|_2^2 + \left\|\boldsymbol{V}_{j,\cdot}^\star\right\|_2^2\right)\sigma_{\max}\cdot\frac{2C}{\sigma_{\min}}\sqrt{\frac{\kappa^2\mu r\log n}{np}}. \tag{H.3}$$

Comparing (H.2) and (H.3), one arrives at the desired result as long as $np \ge 4C^2\varepsilon^{-2}\kappa^4\mu r\log n$.

I Proofs in Section A

I.1 Proof of the inequalities (A.13)

We start with (A.13a). Invoke the triangle inequality to get

$$\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H} - \boldsymbol{F}^\star\right\| \le \left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H} - \boldsymbol{F}\boldsymbol{H}\right\| + \left\|\boldsymbol{F}\boldsymbol{H} - \boldsymbol{F}^\star\right\| = \left\|\boldsymbol{F}^{\mathsf{d}} - \boldsymbol{F}\right\| + O\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|\right), \tag{I.1}$$

where the last relation depends on the unitary invariance of the operator norm and (A.9b). It then boils down to controlling $\|\boldsymbol{F}^{\mathsf{d}} - \boldsymbol{F}\|$. Notice that

$$\begin{aligned}\left\|\boldsymbol{F}^{\mathsf{d}} - \boldsymbol{F}\right\| &\le \left\|\boldsymbol{F}\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right)^{1/2} - \boldsymbol{F}\right\| + \left\|\boldsymbol{Y}\left[\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2} - \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right)^{1/2}\right]\right\| \\ &\le \left\|\boldsymbol{F}\right\|\left\|\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right)^{1/2} - \boldsymbol{I}_r\right\| + \left\|\boldsymbol{Y}\right\|\left\|\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\right)^{1/2} - \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right)^{1/2}\right\| \\ &\le \left\|\boldsymbol{F}\right\|\left\|\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right)^{1/2} - \boldsymbol{I}_r\right\| + O\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|\right),\end{aligned}$$

where the last inequality uses $\|\boldsymbol{Y}\| \le \|\boldsymbol{F}\| \le 2\|\boldsymbol{X}^\star\|$ (cf. (A.19)), the fact that $\lambda \lesssim \sigma\sqrt{np}$ (see (A.6)), the bound (C.16) and the condition $n^5 \gg \kappa$. Apply the perturbation bound for matrix square roots (see Lemma 13) to obtain that

$$\begin{aligned}\left\|\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right)^{1/2} - \boldsymbol{I}_r\right\| &\le \frac{\lambda/p}{\lambda_{\min}\left(\boldsymbol{I}_r\right) + \lambda_{\min}\left[\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right)^{1/2}\right]}\left\|\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right\| \\ &\overset{\text{(i)}}{\lesssim} \frac{\lambda}{p\sigma_{\min}} \overset{\text{(ii)}}{\lesssim} \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}.\end{aligned}$$

Here, (i) uses the facts that $\|(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\| \lesssim 1/\sigma_{\min}$ and that $\lambda_{\min}[(\boldsymbol{I}_r + \lambda/p(\boldsymbol{X}^\top\boldsymbol{X})^{-1})^{1/2}] \ge 1$, and (ii) follows from the condition that $\lambda \lesssim \sigma\sqrt{np}$ (see (A.6)). Combine the above two bounds with $\|\boldsymbol{F}\| \le 2\|\boldsymbol{X}^\star\|$ (cf. (A.19)) to reach

$$\left\|\boldsymbol{F}^{\mathsf{d}} - \boldsymbol{F}\right\| \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|. \tag{I.2}$$

Substitution into (I.1) gives

$$\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H} - \boldsymbol{F}^\star\right\| \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|. \tag{I.3}$$

Analogous arguments yield

$$\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{F}^\star\right\|_{\mathrm{F}} \le \left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H} - \boldsymbol{F}^\star\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}},$$

which is the claim in (A.13c).

Moving on to (A.13b), we apply the triangle inequality and (I.3) to see that

$$\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{F}^\star\right\| \le \left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}\right\| + \left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H} - \boldsymbol{F}^\star\right\| \le \left\|\boldsymbol{F}^{\mathsf{d}}\right\|\left\|\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{H}\right\| + O\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|\right).$$

In order to control $\|\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{H}\|$, we leverage [MWCC17, Lemma 36] to get

$$\left\|\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{H}\right\| \le \frac{1}{\sigma_{\min}\left(\boldsymbol{F}^\top\boldsymbol{F}^\star\right)}\left\|\boldsymbol{F}^{\mathsf{d}\top}\boldsymbol{F}^\star - \boldsymbol{F}^\top\boldsymbol{F}^\star\right\| \lesssim \frac{1}{\sigma_{\min}}\left\|\boldsymbol{F}^{\mathsf{d}} - \boldsymbol{F}\right\|\left\|\boldsymbol{F}^\star\right\| \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}, \tag{I.4}$$

where the last relation uses (I.2) and $\|\boldsymbol{F}^\star\| \asymp \|\boldsymbol{X}^\star\| \asymp \sqrt{\sigma_{\max}}$. Taking these bounds collectively yields

$$\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{F}^\star\right\| \lesssim \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|.$$

Now we turn attention to (A.13d). Observe that

$$\begin{aligned}\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{F}^\star\right\|_{2,\infty} &\le \left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}\right\|_{2,\infty} + \left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H} - \boldsymbol{F}\boldsymbol{H}\right\|_{2,\infty} + \left\|\boldsymbol{F}\boldsymbol{H} - \boldsymbol{F}^\star\right\|_{2,\infty} \\ &\le \left\|\boldsymbol{F}^{\mathsf{d}}\right\|_{2,\infty}\left\|\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{H}\right\| + \left\|\boldsymbol{F}^{\mathsf{d}} - \boldsymbol{F}\right\|_{2,\infty} + O\Big(\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^\star\right\|_{2,\infty}\Big),\end{aligned} \tag{I.5}$$

where the last bound arises from (A.9c). Going through the same calculation as in bounding $\|\boldsymbol{F}^{\mathsf{d}} - \boldsymbol{F}\|$, we arrive at

$$\left\|\boldsymbol{F}^{\mathsf{d}} - \boldsymbol{F}\right\|_{2,\infty} \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{F}^\star\right\|_{2,\infty} \qquad \text{and} \qquad \left\|\boldsymbol{F}^{\mathsf{d}}\right\|_{2,\infty} \le 2\left\|\boldsymbol{F}^\star\right\|_{2,\infty}$$

as long as $\sigma\sqrt{n/p} \ll \sigma_{\min}$. We can thus continue the upper bound in (I.5) to derive

$$\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{F}^\star\right\|_{2,\infty} \lesssim \left\|\boldsymbol{F}^\star\right\|_{2,\infty}\left\|\boldsymbol{H}^{\mathsf{d}} - \boldsymbol{H}\right\| + \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^\star\right\|_{2,\infty}$$

$$
\begin{aligned}
&\lesssim \kappa \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\| \boldsymbol{F}^{\star} \right\|_{2,\infty} + \kappa \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} \left\| \boldsymbol{F}^{\star} \right\|_{2,\infty} \\
&\asymp \kappa \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} \left\| \boldsymbol{F}^{\star} \right\|_{2,\infty}.
\end{aligned}
$$

Here, the second line results from (I.4).

Finally, we deal with (A.13e). From the definition of the de-shrunken estimator (3.8), we have

$$
\begin{aligned}
\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}} - \boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}} = &\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\right)^{1/2} \boldsymbol{X}^{\top}\boldsymbol{X} \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\right)^{1/2} \\
&- \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^{\top}\boldsymbol{Y}\right)^{-1}\right)^{1/2} \boldsymbol{Y}^{\top}\boldsymbol{Y} \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^{\top}\boldsymbol{Y}\right)^{-1}\right)^{1/2}.
\end{aligned}
$$

This combined with the triangle inequality reveals that

$$
\begin{aligned}
&\left\| \boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}} - \boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}} \right\| \\
&\quad \leq \left\| \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\right)^{1/2} \right\| \left\| \boldsymbol{X}^{\top}\boldsymbol{X} - \boldsymbol{Y}^{\top}\boldsymbol{Y} \right\| \left\| \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\right)^{1/2} \right\| \\
&\qquad + \left\| \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\right)^{1/2} - \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^{\top}\boldsymbol{Y}\right)^{-1}\right)^{1/2} \right\| \left\| \boldsymbol{Y}^{\top}\boldsymbol{Y} \right\| \left\| \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\right)^{1/2} \right\| \\
&\qquad + \left\| \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^{\top}\boldsymbol{Y}\right)^{-1}\right)^{1/2} \right\| \left\| \boldsymbol{Y}^{\top}\boldsymbol{Y} \right\| \left\| \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\right)^{1/2} - \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^{\top}\boldsymbol{Y}\right)^{-1}\right)^{1/2} \right\|.
\end{aligned}
$$

Making use of (A.11) and (C.16) allows us to establish the claim.

I.2 Proof of the inequalities (A.16)

The proofs of (A.16a) and (A.16b) are the same as those of (A.13b) and (A.13d), and are hence omitted for conciseness. We are left with (A.16c). Denoting

$$
\boldsymbol{F}_0 \triangleq \boldsymbol{F}^{\star}, \qquad \boldsymbol{F}_1 \triangleq \boldsymbol{F}^{\mathsf{d}}\boldsymbol{H} \qquad \text{and} \qquad \boldsymbol{F}_2 \triangleq \boldsymbol{F}^{\mathsf{d},(j)}\boldsymbol{R}^{(j)},
$$

one has

$$
\left\| \boldsymbol{F}_1 - \boldsymbol{F}_0 \right\| \left\| \boldsymbol{F}_0 \right\| = \left\| \boldsymbol{F}^{\mathsf{d}}\boldsymbol{H} - \boldsymbol{F}^{\star} \right\| \left\| \boldsymbol{F}^{\star} \right\| \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \sigma_{\max} \leq \sigma_{\min} = \frac{\sigma_r^2\left(\boldsymbol{F}_0\right)}{2},
$$

as long as $\sigma\sqrt{n/p} \ll \sigma_{\min}/\kappa$. Here the first inequality follows from (A.13a). In addition, we have

$$
\begin{aligned}
&\left\| \boldsymbol{F}_1 - \boldsymbol{F}_2 \right\| \left\| \boldsymbol{F}_0 \right\| = \left\| \boldsymbol{F}^{\mathsf{d}}\boldsymbol{H} - \boldsymbol{F}^{\mathsf{d},(j)}\boldsymbol{R}^{(j)} \right\| \left\| \boldsymbol{F}^{\star} \right\| \\
&\quad \leq \left\| \boldsymbol{F}\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^{\top}\boldsymbol{Y}\right)^{-1}\right)^{1/2} \boldsymbol{H} - \boldsymbol{F}^{(j)}\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\right)^{-1}\right)^{1/2} \boldsymbol{R}^{(j)} \right\| \left\| \boldsymbol{F}^{\star} \right\| + \theta, \\
&\quad = \left\| \boldsymbol{F}\boldsymbol{H}\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{H}^{\top}\boldsymbol{Y}^{\top}\boldsymbol{Y}\boldsymbol{H}\right)^{-1}\right)^{1/2} - \boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)}\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{R}^{(j)\top}\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)}\right)^{-1}\right)^{1/2} \right\| \left\| \boldsymbol{F}^{\star} \right\| + \theta,
\end{aligned} \tag{I.6}
$$

where θ is defined to be

$$
\begin{aligned}
\theta \triangleq &\left\| \boldsymbol{X}\left[\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^{\top}\boldsymbol{Y}\right)^{-1}\right)^{1/2} - \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\right)^{1/2}\right] \right\| \left\| \boldsymbol{F}^{\star} \right\| \\
&+ \left\| \boldsymbol{X}^{(j)}\left[\left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\right)^{-1}\right)^{1/2} - \left(\boldsymbol{I}_r + \frac{\lambda}{p}\left(\boldsymbol{X}^{(j)\top}\boldsymbol{X}^{(j)}\right)^{-1}\right)^{1/2}\right] \right\| \left\| \boldsymbol{F}^{\star} \right\|.
\end{aligned}
$$

Regarding θ, one can apply the bound (C.16) for $(\boldsymbol{X},\boldsymbol{Y})$ and a similar bound for $(\boldsymbol{X}^{(j)},\boldsymbol{Y}^{(j)})$ to obtain

$$\theta \lesssim \sigma_{\max}\cdot\frac{\kappa}{n^5}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}.$$

Returning to (I.6), one has by the triangle inequality that

$$\begin{aligned}
&\left\|\boldsymbol{F}\boldsymbol{H}\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{H}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{H}\right)^{-1}\right)^{1/2}-\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)}\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{R}^{(j)\top}\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)}\right)^{-1}\right)^{1/2}\right\|\\
&\quad\leq\left\|\left(\boldsymbol{F}\boldsymbol{H}-\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)}\right)\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{H}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{H}\right)^{-1}\right)^{1/2}\right\|\\
&\qquad+\left\|\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)}\left[\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{H}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{H}\right)^{-1}\right)^{1/2}-\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{R}^{(j)\top}\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)}\right)^{-1}\right)^{1/2}\right]\right\|\\
&\quad\leq\left\|\boldsymbol{F}\boldsymbol{H}-\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)}\right\|_{\mathrm{F}}\left\|\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{H}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{H}\right)^{-1}\right)^{1/2}\right\|\\
&\qquad+\left\|\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)}\right\|\left\|\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{H}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{H}\right)^{-1}\right)^{1/2}-\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{R}^{(j)\top}\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)}\right)^{-1}\right)^{1/2}\right\|.
\end{aligned}$$

Recognizing that

$$\lambda_{\min}\left[\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{H}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{H}\right)^{-1}\right)^{1/2}\right]\geq 1\qquad\text{and}\qquad\lambda_{\min}\left[\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{R}^{(j)\top}\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)}\right)^{-1}\right)^{1/2}\right]\geq 1,$$

we can apply the perturbation bound for matrix square roots (see Lemma 13) to obtain

$$\begin{aligned}
&\left\|\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{H}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{H}\right)^{-1}\right)^{1/2}-\left(\boldsymbol{I}_r+\frac{\lambda}{p}\left(\boldsymbol{R}^{(j)\top}\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)}\right)^{-1}\right)^{1/2}\right\|\\
&\quad\lesssim\frac{\lambda}{p}\left\|\left(\boldsymbol{H}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{H}\right)^{-1}-\left(\boldsymbol{R}^{(j)\top}\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)}\right)^{-1}\right\|\\
&\quad\lesssim\frac{\lambda}{p}\left\|\left(\boldsymbol{H}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{H}\right)^{-1}\right\|\left\|\boldsymbol{H}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{H}-\boldsymbol{R}^{(j)\top}\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)}\right\|\left\|\left(\boldsymbol{R}^{(j)\top}\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)}\right)^{-1}\right\|\\
&\quad\lesssim\frac{\lambda}{p}\frac{1}{\sigma_{\min}^2}\left\|\boldsymbol{H}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{H}-\boldsymbol{R}^{(j)\top}\boldsymbol{Y}^{(j)\top}\boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)}\right\|\lesssim\frac{\lambda}{p}\frac{1}{\sigma_{\min}^2}\sqrt{\sigma_{\max}}\left\|\boldsymbol{F}\boldsymbol{H}-\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)}\right\|_{\mathrm{F}}\\
&\quad\lesssim\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\frac{\sqrt{\sigma_{\max}}}{\sigma_{\min}}\left\|\boldsymbol{F}\boldsymbol{H}-\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)}\right\|_{\mathrm{F}}.
\end{aligned}$$

Collect the pieces to arrive at

$$\begin{aligned}
\left\|\boldsymbol{F}_1-\boldsymbol{F}_2\right\|\left\|\boldsymbol{F}_0\right\|&\lesssim\sqrt{\sigma_{\max}}\left(\left\|\boldsymbol{F}\boldsymbol{H}-\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)}\right\|_{\mathrm{F}}+\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{F}\boldsymbol{H}-\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)}\right\|_{\mathrm{F}}\right)+\sigma_{\max}\cdot\frac{\kappa}{n^5}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\\
&\lesssim\sqrt{\sigma_{\max}}\left\|\boldsymbol{F}\boldsymbol{H}-\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)}\right\|_{\mathrm{F}}+\sigma_{\max}\cdot\frac{\kappa}{n^5}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\\
&\lesssim\sqrt{\sigma_{\max}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^\star\right\|_{2,\infty}\ll\frac{\sigma_r^2\left(\boldsymbol{F}_0\right)}{4},
\end{aligned}$$

where the penultimate relation uses (A.14a) as well as the fact that $\|\boldsymbol{F}^\star\|_{2,\infty}\geq\sqrt{\sigma_{\min}r/n}$.

With the above bound in place, we are ready to invoke [CCF+19, Lemma 22] to obtain

$$\begin{aligned}
\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}^{\mathsf{d}}-\boldsymbol{F}^{\mathsf{d},(j)}\boldsymbol{H}^{\mathsf{d},(j)}\right\|\lesssim\kappa\left\|\boldsymbol{F}^{\mathsf{d}}\boldsymbol{H}-\boldsymbol{F}^{\mathsf{d},(j)}\boldsymbol{R}^{(j)}\right\|&\lesssim\kappa\left\|\boldsymbol{F}\boldsymbol{H}-\boldsymbol{F}^{(j)}\boldsymbol{R}^{(j)}\right\|_{\mathrm{F}}\\
&\lesssim\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^\star\right\|_{2,\infty},
\end{aligned}$$

where the last line comes from (A.14a). This concludes the proof.

J Technical lemmas

This section collects a few useful matrix perturbation bounds. The first one is concerned with the perturbation of pseudo-inverses.

Lemma 12 (Perturbation of pseudo-inverses)**.** *Let $\boldsymbol{A}^{\dagger}$ (resp. $\boldsymbol{B}^{\dagger}$) be the pseudo-inverse (i.e. Moore–Penrose inverse) of $\boldsymbol{A}$ (resp. $\boldsymbol{B}$). Then we have*

$$\|\boldsymbol{B}^{\dagger}-\boldsymbol{A}^{\dagger}\| \leq 3\max\left\{\|\boldsymbol{A}^{\dagger}\|^2, \|\boldsymbol{B}^{\dagger}\|^2\right\}\|\boldsymbol{B}-\boldsymbol{A}\|.$$

Proof. See [Ste77, Theorem 3.3]. □

The next lemma focuses on the perturbation bound for matrix square roots.

Lemma 13 (Perturbation of matrix square roots)**.** *Consider two symmetric matrices obeying $\boldsymbol{A}_1 \succeq \mu_1 \boldsymbol{I}$ and $\boldsymbol{A}_2 \succeq \mu_2 \boldsymbol{I}$ for some $\mu_1, \mu_2 > 0$. Let $\boldsymbol{R}_1 \succeq \boldsymbol{0}$ (resp. $\boldsymbol{R}_2 \succeq \boldsymbol{0}$) be the (principal) matrix square root of $\boldsymbol{A}_1$ (resp. $\boldsymbol{A}_2$). Then one has*

$$\|\boldsymbol{R}_1-\boldsymbol{R}_2\| \leq \frac{1}{\sqrt{\mu_1}+\sqrt{\mu_2}}\|\boldsymbol{A}_1-\boldsymbol{A}_2\|.$$

Proof. See [Sch92, Lemma 2.1]. □

The following lemma concerns the perturbation of top-r components of matrices.

Lemma 14 (Perturbation of top-r components)**.** *Consider two matrices $\boldsymbol{M}, \boldsymbol{M}+\boldsymbol{E} \in \mathbb{R}^{n\times n}$. Suppose that $\|\boldsymbol{E}\| \leq \|\boldsymbol{M}\|$ and $\sigma_r(\boldsymbol{M}) > \sigma_{r+1}(\boldsymbol{M}+\boldsymbol{E})$. Let $\boldsymbol{U\Sigma V}^{\top}$ (resp. $\hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^{\top}$) be the rank-r SVD of $\boldsymbol{M}$ (resp. $\boldsymbol{M}+\boldsymbol{E}$). Then one has*

$$\left\|\boldsymbol{U\Sigma V}^{\top}-\hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^{\top}\right\|_{\mathrm{F}} \leq \left(\frac{12\,\|\boldsymbol{\Sigma}\|}{\sigma_r(\boldsymbol{M})-\sigma_{r+1}(\boldsymbol{M}+\boldsymbol{E})}+1\right)\|\boldsymbol{E}\|_{\mathrm{F}}.$$

Proof. From Wedin's $\sin\boldsymbol{\Theta}$ theorem [Wed72], there exist orthonormal matrices $\boldsymbol{R}_1, \boldsymbol{R}_2 \in \mathcal{O}^{r\times r}$ such that

$$\max\left\{\|\hat{\boldsymbol{U}}\boldsymbol{R}_1-\boldsymbol{U}\|_{\mathrm{F}}, \|\hat{\boldsymbol{V}}\boldsymbol{R}_2-\boldsymbol{V}\|_{\mathrm{F}}\right\} \leq \frac{2}{\sigma_r(\boldsymbol{M})-\sigma_{r+1}(\boldsymbol{M}+\boldsymbol{E})}\|\boldsymbol{E}\|_{\mathrm{F}}. \tag{J.1}$$

In addition, Weyl's inequality tells us that

$$\left\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\right\| \leq \|\boldsymbol{E}\| \qquad \text{and hence} \qquad \left\|\hat{\boldsymbol{\Sigma}}\right\| \leq 2\left\|\boldsymbol{\Sigma}\right\|. \tag{J.2}$$

Here, the second inequality follows from the triangle inequality and the assumption that $\|\boldsymbol{E}\| \leq \|\boldsymbol{M}\| = \|\boldsymbol{\Sigma}\|$. Expand $\boldsymbol{U\Sigma V}^{\top}-\hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^{\top}$ and apply the triangle inequality to obtain

$$\begin{aligned}\left\|\boldsymbol{U\Sigma V}^{\top}-\hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^{\top}\right\|_{\mathrm{F}} &= \left\|\boldsymbol{U\Sigma V}^{\top}-\hat{\boldsymbol{U}}\boldsymbol{R}_1\boldsymbol{R}_1^{\top}\hat{\boldsymbol{\Sigma}}\boldsymbol{R}_2\boldsymbol{R}_2^{\top}\hat{\boldsymbol{V}}^{\top}\right\|_{\mathrm{F}} \\ &\leq \left\|(\boldsymbol{U}-\hat{\boldsymbol{U}}\boldsymbol{R}_1)\boldsymbol{\Sigma V}^{\top}\right\|_{\mathrm{F}} + \left\|\hat{\boldsymbol{U}}\boldsymbol{R}_1(\boldsymbol{\Sigma}-\boldsymbol{R}_1^{\top}\hat{\boldsymbol{\Sigma}}\boldsymbol{R}_2)\boldsymbol{V}^{\top}\right\|_{\mathrm{F}} \\ &\quad + \left\|\hat{\boldsymbol{U}}\boldsymbol{R}_1\boldsymbol{R}_1^{\top}\hat{\boldsymbol{\Sigma}}\boldsymbol{R}_2(\boldsymbol{V}-\hat{\boldsymbol{V}}\boldsymbol{R}_2)^{\top}\right\|_{\mathrm{F}},\end{aligned}$$

which further implies that

$$\begin{aligned}\left\|\boldsymbol{U\Sigma V}^{\top}-\hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^{\top}\right\|_{\mathrm{F}} &\leq \left\|\boldsymbol{U}-\hat{\boldsymbol{U}}\boldsymbol{R}_1\right\|_{\mathrm{F}}\left\|\boldsymbol{\Sigma}\right\| + \left\|\boldsymbol{\Sigma}-\boldsymbol{R}_1^{\top}\hat{\boldsymbol{\Sigma}}\boldsymbol{R}_2\right\|_{\mathrm{F}} + \left\|\hat{\boldsymbol{\Sigma}}\right\|\left\|\boldsymbol{V}-\hat{\boldsymbol{V}}\boldsymbol{R}_2\right\|_{\mathrm{F}} \\ &\leq \frac{6\,\|\boldsymbol{\Sigma}\|}{\sigma_r(\boldsymbol{M})-\sigma_{r+1}(\boldsymbol{M}+\boldsymbol{E})}\|\boldsymbol{E}\|_{\mathrm{F}} + \left\|\boldsymbol{\Sigma}-\boldsymbol{R}_1^{\top}\hat{\boldsymbol{\Sigma}}\boldsymbol{R}_2\right\|_{\mathrm{F}}.\end{aligned} \tag{J.3}$$

Here, the last line arises from (J.1) and (J.2). It then boils down to controlling $\|\boldsymbol{\Sigma}-\boldsymbol{R}_1^{\top}\hat{\boldsymbol{\Sigma}}\boldsymbol{R}_2\|_{\mathrm{F}}$. Recognizing that $\boldsymbol{\Sigma}=\boldsymbol{U}^{\top}\boldsymbol{MV}$ and $\hat{\boldsymbol{\Sigma}}=\hat{\boldsymbol{U}}^{\top}(\boldsymbol{M}+\boldsymbol{E})\hat{\boldsymbol{V}}$, we obtain

$$\left\|\boldsymbol{\Sigma}-\boldsymbol{R}_1^{\top}\hat{\boldsymbol{\Sigma}}\boldsymbol{R}_2\right\|_{\mathrm{F}} = \left\|\boldsymbol{U}^{\top}\boldsymbol{MV}-\boldsymbol{R}_1^{\top}\hat{\boldsymbol{U}}^{\top}(\boldsymbol{M}+\boldsymbol{E})\hat{\boldsymbol{V}}\boldsymbol{R}_2\right\|_{\mathrm{F}}$$

$$\begin{aligned}&\leq \left\|\left(\boldsymbol{U}-\hat{\boldsymbol{U}}\boldsymbol{R}_1\right)^\top \boldsymbol{M}\boldsymbol{V}\right\|_{\mathrm{F}} + \left\|\boldsymbol{R}_1^\top \hat{\boldsymbol{U}}^\top \boldsymbol{E}\boldsymbol{V}\right\|_{\mathrm{F}} + \left\|\boldsymbol{R}_1^\top \hat{\boldsymbol{U}}^\top \left(\boldsymbol{M}+\boldsymbol{E}\right)\left(\boldsymbol{V}-\hat{\boldsymbol{V}}\boldsymbol{R}_2\right)\right\|_{\mathrm{F}}\\ &\leq \left\|\boldsymbol{U}-\hat{\boldsymbol{U}}\boldsymbol{R}_1\right\|_{\mathrm{F}} \|\boldsymbol{\Sigma}\| + \|\boldsymbol{E}\|_{\mathrm{F}} + \|\hat{\boldsymbol{\Sigma}}\|\left\|\boldsymbol{V}-\hat{\boldsymbol{V}}\boldsymbol{R}_2\right\|_{\mathrm{F}}.\end{aligned}$$

Once again, employ (J.1) and (J.2) to arrive at

$$\left\|\boldsymbol{\Sigma}-\boldsymbol{R}_1^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{R}_2\right\|_{\mathrm{F}} \leq \frac{6\|\boldsymbol{\Sigma}\|}{\sigma_r(\boldsymbol{M})-\sigma_{r+1}(\boldsymbol{M}+\boldsymbol{E})}\|\boldsymbol{E}\|_{\mathrm{F}} + \|\boldsymbol{E}\|_{\mathrm{F}}. \tag{J.4}$$

Combining (J.3) and (J.4), we reach

$$\left\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^\top\right\|_{\mathrm{F}} \leq \left(\frac{12\|\boldsymbol{\Sigma}\|}{\sigma_r(\boldsymbol{M})-\sigma_{r+1}(\boldsymbol{M}+\boldsymbol{E})}+1\right)\|\boldsymbol{E}\|_{\mathrm{F}}$$

as claimed. □

The last bound centers around the well-known Sylvester equation $\boldsymbol{X}\boldsymbol{A}+\boldsymbol{B}\boldsymbol{X}=\boldsymbol{C}$.

Lemma 15 (The Sylvester equation)**.** *Suppose $\boldsymbol{X}\in\mathbb{R}^{r\times r}$ satisfies the matrix equation $\boldsymbol{X}\boldsymbol{A}+\boldsymbol{B}\boldsymbol{X}=\boldsymbol{C}$ for some matrices $\boldsymbol{A}\in\mathbb{R}^{r\times r}$,$\boldsymbol{B}\in\mathbb{R}^{r\times r}$ and $\boldsymbol{C}\in\mathbb{R}^{r\times r}$. Then one has*

$$\|\boldsymbol{X}\| \leq (2\lambda_{\min})^{-1}\|\boldsymbol{C}\|,$$

as long as $\lambda_{\min}\boldsymbol{I}_r \preceq \boldsymbol{A} \preceq \lambda_{\max}\boldsymbol{I}_r$ and $\lambda_{\min}\boldsymbol{I}_r \preceq \boldsymbol{B} \preceq \lambda_{\max}\boldsymbol{I}_r$ for some $\lambda_{\max}\geq\lambda_{\min}>0$.

Proof. To begin with, we intend to show that under the condition $\lambda_{\min}\boldsymbol{I}_r \preceq \boldsymbol{A},\boldsymbol{B} \preceq \lambda_{\max}\boldsymbol{I}_r$ for some $\lambda_{\max}\geq\lambda_{\min}>0$, there is a unique solution to the matrix equation $\boldsymbol{X}\boldsymbol{A}+\boldsymbol{B}\boldsymbol{X}=\boldsymbol{C}$. Use the notation of Kronecker product to obtain an equivalent form of $\boldsymbol{X}\boldsymbol{A}+\boldsymbol{B}\boldsymbol{X}=\boldsymbol{C}$ as follows

$$\mathsf{vec}\left(\boldsymbol{X}\boldsymbol{A}+\boldsymbol{B}\boldsymbol{X}\right) = \left(\boldsymbol{A}^\top\otimes\boldsymbol{I}_r + \boldsymbol{I}_r\otimes\boldsymbol{B}\right)\cdot\mathsf{vec}\left(\boldsymbol{X}\right) = \mathsf{vec}\left(\boldsymbol{C}\right),$$

where $\otimes$ denotes the Kronecker product and $\mathsf{vec}(\boldsymbol{A})$ stands for the vectorization of the matrix $\boldsymbol{A}$. Given that $\boldsymbol{A}\succ\boldsymbol{0}$ and $\boldsymbol{B}\succ\boldsymbol{0}$, it is straightforward to see that $\boldsymbol{A}^\top\otimes\boldsymbol{I}_r+\boldsymbol{I}_r\otimes\boldsymbol{B}$ is invertible, thus justifying the uniqueness of $\boldsymbol{X}$.

The next step is to characterize $\boldsymbol{X}$ explicitly. The argument herein is adapted from [Smi68] and [Sch92]. Specifically, it has been shown in [Smi68] that the equation $\boldsymbol{X}\boldsymbol{A}+\boldsymbol{B}\boldsymbol{X}=\boldsymbol{C}$ is equivalent to

$$\boldsymbol{X}-\boldsymbol{U}\boldsymbol{X}\boldsymbol{V}=\boldsymbol{W},$$

where $\boldsymbol{U}=(q\boldsymbol{I}_r+\boldsymbol{B})^{-1}(q\boldsymbol{I}_r-\boldsymbol{B})$, $\boldsymbol{V}=(q\boldsymbol{I}_r-\boldsymbol{A})(q\boldsymbol{I}_r+\boldsymbol{A})^{-1}$ and $\boldsymbol{W}=2q(q\boldsymbol{I}_r+\boldsymbol{B})^{-1}\boldsymbol{C}(q\boldsymbol{I}_r+\boldsymbol{A})^{-1}$, for any $q>0$. In particular, when $q>\lambda_{\min}$, the matrix

$$\boldsymbol{X}=\sum_{k=1}^{\infty}\boldsymbol{U}^{k-1}\boldsymbol{W}\boldsymbol{V}^{k-1} \tag{J.5}$$

is the unique solution to $\boldsymbol{X}-\boldsymbol{U}\boldsymbol{X}\boldsymbol{V}=\boldsymbol{W}$ and hence to $\boldsymbol{X}\boldsymbol{A}+\boldsymbol{B}\boldsymbol{X}=\boldsymbol{C}$. To show this, it suffices to verify that the matrix series is convergent. Note that when $q>\lambda_{\min}$, one has

$$\|\boldsymbol{U}\| \leq \left\|(q\boldsymbol{I}_r+\boldsymbol{B})^{-1}\right\|\|q\boldsymbol{I}_r-\boldsymbol{B}\| \leq \frac{q-\lambda_{\min}}{q+\lambda_{\min}} < 1,$$

and similarly $\|\boldsymbol{V}\|\leq(q-\lambda_{\min})/(q+\lambda_{\max})<1$. These two bounds taken together immediately establish the convergence of the matrix series (J.5).

In the end, the explicit representation (J.5) allows us to upper bound $\|\boldsymbol{X}\|$. A little algebra reveals that

$$\|\boldsymbol{X}\| \leq \sum_{k=1}^{\infty}\left\|\boldsymbol{U}^{k-1}\boldsymbol{W}\boldsymbol{V}^{k-1}\right\| \leq \|\boldsymbol{W}\|\sum_{k=1}^{\infty}\|\boldsymbol{U}\|^{k-1}\|\boldsymbol{V}\|^{k-1} \leq \frac{\|\boldsymbol{W}\|}{1-\|\boldsymbol{U}\|\|\boldsymbol{V}\|},$$

where we make use of the fact $\|\boldsymbol{U}\|\|\boldsymbol{V}\| < 1$. In addition, from the definition of $\boldsymbol{W}$ we know that

$$\|\boldsymbol{W}\| \leq 2q\big\| \left(q\boldsymbol{I}_r + \boldsymbol{B}\right)^{-1} \big\| \|\boldsymbol{C}\| \big\| \left(q\boldsymbol{I}_r + \boldsymbol{A}\right)^{-1} \big\| \leq \|\boldsymbol{C}\| \frac{2q}{\left(q+\lambda_{\min}\right)^2},$$

provided that $q > 0$. Combine this with the bounds on $\|\boldsymbol{U}\|$ and $\|\boldsymbol{V}\|$ to reach

$$\|\boldsymbol{X}\| \leq \frac{\|\boldsymbol{C}\| \frac{2q}{(q+\lambda_{\min})^2}}{1-\left(\frac{q-\lambda_{\min}}{q+\lambda_{\min}}\right)^2} = \frac{2q\,\|\boldsymbol{C}\|}{\left(q+\lambda_{\min}\right)^2 - \left(q-\lambda_{\min}\right)^2} = \frac{\|\boldsymbol{C}\|}{2\lambda_{\min}}$$

as claimed. □

References

[AFWZ17] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv:1709.09565*, 2017.

[AIW18] Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

[BCH11] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*, 2011.

[BFL+18] Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of statistics*, 46(3):1352, 2018.

[BN06] Jushan Bai and Serena Ng. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150, 2006.

[CC14] Yuxin Chen and Yuejie Chi. Robust spectral compressed sensing via structured matrix completion. *IEEE Transactions on Information Theory*, 60(10):6576 – 6601, 2014.

[CC17] Yuxin Chen and Emmanuel J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Comm. Pure Appl. Math.*, 70(5):822–883, 2017.

[CC18a] Yudong Chen and Yuejie Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31, July 2018.

[CC18b] Yuxin Chen and Emmanuel Candès. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Communications on Pure and Applied Mathematics*, 71(8):1648–1714, 2018.

[CCD+18] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[CCD+19] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv preprint arXiv:1904.10020*, 2019.

[CCF18] Yuxin Chen, Chen Cheng, and Jianqing Fan. Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *arXiv preprint arXiv:1811.12804*, 2018.

[CCF$^+$19] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *arXiv:1902.07698*, 2019.

[CCFM19] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2):5–37, July 2019.

[CCG15] Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[CDDD19] Vasileios Charisopoulos, Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Composite optimization for robust blind deconvolution. *arXiv preprint arXiv:1901.01624*, 2019.

[CEGN15] Alexandra Carpentier, Jens Eisert, David Gross, and Richard Nickl. Uncertainty quantification for matrix compressed sensing and quantum tomography problems. *arXiv preprint arXiv:1504.03234*, 2015.

[CFMW19] Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized MLE are both optimal for top-K ranking. *Annals of Statistics*, 47(4):2204–2235, August 2019.

[CG17] T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics*, 45(2):615–646, 2017.

[Che15] Yudong Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

[CKL16] Alexandra Carpentier, Olga Klopp, and Matthias Löffler. Constructing confidence sets for the matrix completion problem. In *Conference of the International Society for Non-Parametric Statistics*, pages 103–118. Springer, 2016.

[CKLN18] Alexandra Carpentier, Olga Klopp, Matthias Löffler, and Richard Nickl. Adaptive confidence sets for matrix completion. *Bernoulli*, 24(4A):2429–2460, 2018.

[CL17] Ji Chen and Xiaodong Li. Memory-efficient kernel PCA via partial matrix sampling and nonconvex optimization: a model-free analysis of local minima. *arXiv:1711.01742*, 2017.

[CLC19] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239 – 5269, October 2019.

[CLL19] Ji Chen, Dekai Liu, and Xiaodong Li. Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *arXiv:1901.06116v1*, 2019.

[CLMW11] Emmanuel Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of ACM*, 58(3):11:1–11:37, Jun 2011.

[CLR16] T Tony Cai, Tengyuan Liang, and Alexander Rakhlin. Geometric inference for general high-dimensional linear inverse problems. *The Annals of Statistics*, 44(4):1536–1563, 2016.

[CLS15] E. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, April 2015.

[CMW13] T Tony Cai, Zongming Ma, and Yihong Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

[CN15] Alexandra Carpentier and Richard Nickl. On signal detection and confidence sets for low rank inference problems. *Electronic Journal of Statistics*, 9(2):2675–2688, 2015.

[CP10] Emmanuel Candès and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925 –936, June 2010.

[CR09] Emmanuel Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, April 2009.

[CSPW11] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[CW15] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv:1509.03025*, 2015.

[CX16] Yang Cao and Yao Xie. Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing*, 64(6):1609–1620, 2016.

[CZ13] Tony Cai and Wen-Xin Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research*, 14(1):3619–3647, 2013.

[CZ16] T Tony Cai and Wen-Xin Zhou. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1):1493–1525, 2016.

[DBMM15] R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558, 2015.

[DBZ17] Ruben Dezeure, Peter Bühlmann, and Cun-Hui Zhang. High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719, 2017.

[DC18] Lijun Ding and Yudong Chen. The leave-one-out approach for matrix completion: Primal and dual analysis. *arXiv preprint arXiv:1803.07554*, 2018.

[DPVW14] Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

[DR16] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[DR17] John C Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *arXiv:1705.02356, Information and Inference*, 2017.

[EK15] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, pages 1–81, 2015.

[EKBB$^+$13] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[Faz02] Maryam Fazel. *Matrix rank minimization with applications.* PhD thesis, 2002.

[FFHL19] Jianqing Fan, Yingying Fan, Xiao Han, and Jinchi Lv. Asymptotic theory of eigenvectors for large random matrices. *arXiv preprint arXiv:1902.06846*, 2019.

[FLM13] J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Stat. Society: Series B*, 75(4):603–680, 2013.

[FRW11] Massimo Fornasier, Holger Rauhut, and Rachel Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21(4):1614–1640, 2011.

[FS11] Rina Foygel and Nathan Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In *Conference on Learning Theory*, pages 315–340, 2011.

[FSZZ18] Jianqing Fan, Qiang Sun, Wen-Xin Zhou, and Ziwei Zhu. Principal component analysis for big data. *arXiv preprint arXiv:1801.01602*, 2018.

[FWZ19] Jianqing Fan, Weichen Wang, and Yiqiao Zhong. Robust covariance estimation for approximate factor models. *Journal of econometrics*, 208(1):5–22, 2019.

[GJZ17] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

[GLM16] 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.

[Gro11] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, March 2011.

[Har14] Moritz Hardt. Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS)*, pages 651–660, 2014.

[HKZ12] Daniel Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17:no. 52, 6, 2012.

[JKN16] Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *NIPS*, pages 4520–4528, 2016.

[JM14a] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[JM14b] Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.

[JM15] Adel Javanmard and Andrea Montanari. De-biasing the lasso: Optimal sample size for Gaussian designs. *arXiv preprint arXiv:1508.02757*, 2015.

[JMD10] Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010.

[JNS13] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *ACM symposium on Theory of computing*, pages 665–674, 2013.

[JVDG15] Jana Jankova and Sara Van De Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015.

[JvdG17] Jana Janková and Sara van de Geer. Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test*, 26(1):143–162, 2017.

[JvdG18] Jana Janková and Sara van de Geer. De-biased sparse pca: Inference and testing for eigenstructure of large covariance matrices. *arXiv preprint arXiv:1801.10567*, 2018.

[Klo14] Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[KLT11] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

[KMO10a] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980 –2998, June 2010.

[KMO10b] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, 2010.

[Kol11] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011.

[KS11] Alois Kneip and Pascal Sarda. Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics*, 39(5):2410–2447, 2011.

[KS19] Felix Krahmer and Dominik Stöger. On the convex geometry of blind deconvolution and matrix completion. *arXiv preprint arXiv:1902.11156*, 2019.

[KX15] Vladimir Koltchinskii and Dong Xia. Optimal estimation of low rank density matrices. *Journal of Machine Learning Research*, 16:1757–1792, 2015.

[LSST16] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[LTTT14] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

[LXY13] M. Lai, Y. Xu, and W. Yin. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM Journal on Numerical Analysis*, 51(2):927–957, 2013.

[MHT10] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

[MJD09] Raghu Meka, Prateek Jain, and Inderjit S. Dhillon. Guaranteed rank minimization via singular value projection. *preprint*, 2009.

[MLL17] Cong Ma, Junwei Lu, and Han Liu. Inter-subject analysis: Inferring sparse interactions with dense intra-graphs. *arXiv preprint arXiv:1709.07036*, 2017.

[MMB09] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

[MSL19] Igor Molybog, Somayeh Sojoudi, and Javad Lavaei. No spurious solutions in non-convex matrix sensing: Structure compensates for isometry. 2019.

[MWCC17] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467, accepted to Foundations of Computational Mathematics*, 2017.

[MX17] Simon Mak and Yao Xie. Active matrix completion with uncertainty quantification. *arXiv preprint arXiv:1706.08037*, 2017.

[NCD19] National climatic data center. https://www.ncdc.noaa.gov/, 2019. Accessed: 2019-08-31.

[NL17] Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

[NNLL18] Matey Neykov, Yang Ning, Jun S Liu, and Han Liu. A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science*, 33(3):427–443, 2018.

[NW12] S. Negahban and M.J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, pages 1665–1697, May 2012.

[PB14] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[Rec11] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[RS05] Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. *International conference on Machine learning*, pages 713–719, 2005.

[RSZZ15] Z. Ren, T. Sun, C. Zhang, and H. Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

[RT11] Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

[Sch92] Bernhard A. Schmitt. Perturbation bounds for matrix square roots and Pythagorean sums. *Linear Algebra Appl.*, 174:215–227, 1992.

[Sha03] J. Shao. *Mathematical Statistics*. Springer Texts in Statistics. Springer, 2003.

[Sin11] Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20–36, 2011.

[SL16] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[Smi68] RA Smith. Matrix equation $XA + BX = C$. *SIAM Journal on Applied Mathematics*, 16(1):198–201, 1968.

[Sre04] Nathan Srebro. Learning with matrix factorizations. *Ph. D. thesis*, 2004.

[SS05] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005.

[Ste77] Gilbert W Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM review*, 19(4):634–662, 1977.

[SXZ19] Alexander Shapiro, Yao Xie, and Rui Zhang. Matrix completion with deterministic pattern: A geometric perspective. *IEEE Transactions on Signal Processing*, 67(4):1088–1103, 2019.

[SY07] Anthony Man-Cho So and Yinyu Ye. Theory of semidefinite programming for sensor network localization. *Mathematical Programming*, 109(2-3):367–384, 2007.

[SZ12] Tingni Sun and Cun-Hui Zhang. Calibrated elastic regularization in matrix completion. In *Advances in Neural Information Processing Systems*, pages 863–871, 2012.

[TBS$^+$16] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. *ICML*, pages 964–973, 2016.

[Van13] Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

[vdGBRD14] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[Ver17] Roman Vershynin. High dimensional probability, 2017.

[Wai19] M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[WCCL16] K. Wei, J.F. Cai, T. Chan, and S. Leung. Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

[Wed72] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[WR09] Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

[WYZ12] Zaiwen Wen, Wotao Yin, and Yin Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012.

[WZG16] Lingxiao Wang, Xiao Zhang, and Quanquan Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. *arXiv preprint arXiv:1610.05275*, 2016.

[Xia18] Dong Xia. Confidence interval of singular vectors for high-dimensional and low-rank matrix regression. *arXiv preprint arXiv:1805.09871*, 2018.

[Xia19] Dong Xia. Data-dependent confidence regions of singular subspaces. *arXiv preprint arXiv:1901.00304*, 2019.

[ZC17] Xianyang Zhang and Guang Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768, 2017.

[ZHT06] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

[ZJSL18] Richard Zhang, Cédric Josz, Somayeh Sojoudi, and Javad Lavaei. How much restricted isometry is needed in nonconvex matrix recovery? pages 5586–5597, 2018.

[ZL16] Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv:1605.07051*, 2016.

[ZPL15] Tao Zhang, John M Pauly, and Ives R Levesque. Accelerating parameter mapping with a locally low rank constraint. *Magnetic resonance in medicine*, 73(2):655–661, 2015.

[ZSL19] R. Zhang, S. Sojoudi, and J. Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *arXiv:1901.01631*, 2019.

[ZWL15] Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. In *NIPS*, pages 559–567, 2015.

[ZWYG18] Xiao Zhang, Lingxiao Wang, Yaodong Yu, and Quanquan Gu. A primal-dual analysis of global optimality in nonconvex low-rank matrix recovery. In *International conference on machine learning*, pages 5857–5866, 2018.

[ZZ14] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76(1):217–242, 2014.