Noisy Matrix Completion: Understanding Statistical Guarantees for Convex Relaxation via Nonconvex Optimization

Yuxin Chen* Yuejie Chi† Jianqing Fan‡ Cong Ma‡ Yuling Yan‡

October 13, 2019

Abstract

This paper studies noisy low-rank matrix completion: given partial and noisy entries of a large low-rank matrix, the goal is to estimate the underlying matrix faithfully and efficiently. Arguably one of the most popular paradigms to tackle this problem is convex relaxation, which achieves remarkable efficacy in practice. However, the theoretical support of this approach is still far from optimal in the noisy setting, falling short of explaining its empirical success.

We make progress towards demystifying the practical efficacy of convex relaxation vis-à-vis random noise. When the rank and the condition number of the unknown matrix are bounded by a constant, we demonstrate that the convex programming approach achieves near-optimal estimation errors — in terms of the Euclidean loss, the entrywise loss, and the spectral norm loss — for a wide range of noise levels. All of this is enabled by bridging convex relaxation with the nonconvex Burer–Monteiro approach, a seemingly distinct algorithmic paradigm that is provably robust against noise. More specifically, we show that an approximate critical point of the nonconvex formulation serves as an extremely tight approximation of the convex solution, thus allowing us to transfer the desired statistical guarantees of the nonconvex approach to its convex counterpart.

Keywords: matrix completion, minimaxity, stability, convex relaxation, nonconvex optimization, Burer–Monteiro approach.

Contents

Author names are sorted alphabetically.

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

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

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

1 Introduction

Suppose we are interested in a large low-rank data matrix, but only get to observe a highly incomplete subset of its entries. Can we hope to estimate the underlying data matrix in a reliable manner? This problem, often dubbed as *low-rank matrix completion*, spans a diverse array of science and engineering applications (e.g. collaborative filtering [RS05], localization [SY07], system identification [LV09], magnetic resonance parameter mapping [ZPL15], joint alignment [CC18a]), and has inspired a flurry of research activities in the past decade. In the statistics literature, matrix completion also falls under the category of factor models with a large amount of missing data, which finds numerous statistical applications such as controlling false discovery rates for dependence data [Efr07, Efr10, FHG12, FKSZ19], factor-adjusted variable selection [KS11, FKW18], principal component regression [Jol82, BN06, PBHT08, FXY17], and large covariance matrix estimation [FLM13, FWZ19]. Recent years have witnessed the development of many tractable algorithms that come with statistical guarantees, with convex relaxation being one of the most popular paradigms [FHB04, CR09, CT10]. See [DR16, CC18b] for an overview of this topic.

This paper focuses on noisy low-rank matrix completion, assuming that the revealed entries are corrupted by a certain amount of noise. Setting the stage, consider the task of estimating a rank-r data matrix $\boldsymbol{M}^\star = [M^\star_{ij}]_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$,[1] and suppose that this needs to be performed on the basis of a subset of noisy entries

$$M_{ij} = M^\star_{ij} + E_{ij}, \qquad (i,j) \in \Omega, \tag{1}$$

[1] It is straightforward to rephrase our discussions to a general rectangular matrix of size $n_1 \times n_2$. The current paper sets $n = n_1 = n_2$ throughout for simplicity of presentation.

where $\Omega \subseteq \{1, \cdots, n\} \times \{1, \cdots, n\}$ denotes a set of indices, and E_{ij} stands for the additive noise at the location (i, j). As we shall elaborate shortly, solving noisy matrix completion via convex relaxation, while practically exhibiting excellent stability (in terms of the estimation errors against noise), is far less understood theoretically compared to the noiseless setting.

1.1 Convex relaxation: limitations of prior results

Naturally, one would search for a low-rank solution that best fits the observed entries. One choice is the regularized least-squares formulation given by

$$\underset{\boldsymbol{Z} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad \frac{1}{2} \sum_{(i,j) \in \Omega} \big(Z_{ij} - M_{ij}\big)^2 + \lambda\, \mathsf{rank}(\boldsymbol{Z}), \tag{2}$$

where $\lambda > 0$ is some regularization parameter. In words, this approach optimizes certain trade-off between the goodness of fit (through the squared loss expressed in the first term of (2)) and the low-rank structure (through the rank function in the second term of (2)). Due to computational intractability of rank minimization, we often resort to convex relaxation in order to obtain computationally feasible solutions. One notable example is the following convex program:

$$\underset{\boldsymbol{Z} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad g(\boldsymbol{Z}) \triangleq \frac{1}{2} \sum_{(i,j) \in \Omega} \big(Z_{ij} - M_{ij}\big)^2 + \lambda \left\|\boldsymbol{Z}\right\|_*, \tag{3}$$

where $\|\boldsymbol{Z}\|_*$ denotes the nuclear norm (i.e. the sum of singular values) of $\boldsymbol{Z}$ — a convex surrogate for the rank function. A significant portion of existing theory supports the use of this paradigm in the noiseless setting: when E_{ij} vanishes for all $(i, j) \in \Omega$, the solution to (3) is known to be faithful (i.e. the estimation error becomes zero) even under near-minimal sample complexity [CR09, CP10, CT10, Gro11, Rec11, Che15].

By contrast, the performance of convex relaxation remains largely unclear when it comes to noisy settings (which are often more practically relevant). Candès and Plan [CP10] first studied the stability of an equivalent variant[2] of (3) against noise. The estimation error $\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\mathrm{F}}$ derived therein, of the solution $\boldsymbol{Z}_{\mathsf{cvx}}$ to (3), is significantly larger than the oracle lower bound. This does not explain well the effectiveness of (3) in practice. In fact, the numerical experiments reported in [CP10] already indicated that the performance of convex relaxation is far better than their theoretical bounds. This discrepancy between numerical performance and existing theoretical bounds gives rise to the following natural yet challenging questions: *Where does the convex program (3) stand in terms of its stability vis-à-vis additive noise? Can we establish statistical performance guarantees that match its practical effectiveness?*

We note in passing that several other convex relaxation formulations have been thoroughly analyzed for noisy matrix completion, most notably by Negahban and Wainwright [NW12] and by Koltchinskii et al. [KLT11]. These works have significantly advanced our understanding of the power of convex relaxation. However, the estimators studied therein, particularly the one in [KLT11], are quite different from the one (3) considered here; as a consequence, the analysis therein does not lead to improved statistical guarantees of (3). Moreover, the performance guarantees provided for these variants are also suboptimal when restricted to the class of "incoherent" or "de-localized" matrices, unless the magnitudes of the noise are fairly large. See Section 1.4 for more detailed discussions as well as numerical comparisons of these algorithms.

1.2 A detour: nonconvex optimization

While the focus of the current paper is convex relaxation, we take a moment to discuss a seemingly distinct algorithmic paradigm: nonconvex optimization, which turns out to be remarkably helpful in understanding convex relaxation. Inspired by the Burer–Monteiro approach [BM03], the nonconvex scheme starts by representing the rank-r decision matrix (or parameters) $\boldsymbol{Z}$ as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{Y}^\top$ via low-rank factors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n \times r}$,

[2] Technically, [CP10] deals with the constrained version of (3), which is equivalent to the Lagrangian form as in (3) with a proper choice of the regularization parameter.

and proceeds by solving the following nonconvex (regularized) least-squares problem [KMO10a]

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad \frac{1}{2}\sum_{(i,j)\in\Omega}\left[\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)_{ij}-M_{ij}\right]^2+\mathsf{reg}(\boldsymbol{X},\boldsymbol{Y}). \tag{4}$$

Here, $\mathsf{reg}(\cdot,\cdot)$ denotes a certain regularization term that promotes additional structural properties.

To see its intimate connection with the convex program (3), we make the following observation: if the solution to (3) has rank r, then it must coincide with the solution to

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad \frac{1}{2}\sum_{(i,j)\in\Omega}\left[\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)_{ij}-M_{ij}\right]^2+\underbrace{\frac{\lambda}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2+\frac{\lambda}{2}\|\boldsymbol{Y}\|_{\mathrm{F}}^2}_{\mathsf{reg}(\boldsymbol{X},\boldsymbol{Y})}. \tag{5}$$

This can be easily verified by recognizing the elementary fact that

$$\|\boldsymbol{Z}\|_*=\inf_{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}:\boldsymbol{X}\boldsymbol{Y}^\top=\boldsymbol{Z}}\left\{\tfrac{1}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2+\tfrac{1}{2}\|\boldsymbol{Y}\|_{\mathrm{F}}^2\right\} \tag{6}$$

for any rank-r matrix $\boldsymbol{Z}$ [SS05, MHT10]. Note, however, that it is very challenging to predict when the key assumption in establishing this connection — namely, the rank-r assumption of the solution to the convex program (3) — can possibly hold (and in particular, whether it can hold under minimal sample complexity requirement).

Despite the nonconvexity of (4), simple first-order optimization methods, in conjunction with proper initialization, are often effective in solving (4). Partial examples include gradient descent on manifold [KMO10a, KMO10b, WCCL16], gradient descent [SL16, MWCC17], and projected gradient descent [CW15, ZL16]. Apart from their practical efficiency, the nonconvex optimization approach is also appealing in theory. To begin with, algorithms tailored to (4) often enable exact recovery in the noiseless setting. Perhaps more importantly, for a wide range of noise settings, the nonconvex approach achieves appealing estimation accuracy [CW15, MWCC17], which could be significantly better than those bounds derived for convex relaxation discussed earlier. See [CLC19, CC18b] for a summary of recent results. Such intriguing statistical guarantees motivate us to take a closer inspection of the underlying connection between the two contrasting algorithmic frameworks.

1.3 Empirical evidence: convex and nonconvex solutions are often close

In order to obtain a better sense of the relationships between convex and nonconvex approaches, we begin by comparing the estimates returned by the two approaches via numerical experiments. Fix $n = 1000$ and $r = 5$. We generate $\boldsymbol{M}^\star = \boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}$, where $\boldsymbol{X}^\star, \boldsymbol{Y}^\star \in \mathbb{R}^{n\times r}$ are random orthonormal matrices. Each entry $M_{ij}^\star$ of $\boldsymbol{M}^\star$ is observed with probability $p = 0.2$ independently, and then corrupted by an independent Gaussian noise $E_{ij} \sim \mathcal{N}(0, \sigma^2)$. Throughout the experiments, we set $\lambda = 5\sigma\sqrt{np}$. The convex program (3) is solved by the proximal gradient method [PB14], whereas we attempt solving the nonconvex formulation (5) by gradient descent with spectral initialization (see [CLC19] for details). Let $\boldsymbol{Z}_{\mathsf{cvx}}$ (resp. $\boldsymbol{Z}_{\mathsf{ncvx}} = \boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top$) be the solution returned by the convex program (3) (resp. the nonconvex program (5)). Figure 1 displays the relative estimation errors of both methods ($\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\mathrm{F}}/\|\boldsymbol{M}^\star\|_{\mathrm{F}}$ and $\|\boldsymbol{Z}_{\mathsf{ncvx}} - \boldsymbol{M}^\star\|_{\mathrm{F}}/\|\boldsymbol{M}^\star\|_{\mathrm{F}}$) as well as the relative distance $\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{Z}_{\mathsf{ncvx}}\|_{\mathrm{F}}/\|\boldsymbol{M}^\star\|_{\mathrm{F}}$ between the two estimates. The results are averaged over 20 independent trials.

Interestingly, the distance between the convex and the nonconvex solutions seems extremely small (e.g. $\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{Z}_{\mathsf{ncvx}}\|_{\mathrm{F}}/\|\boldsymbol{M}^\star\|_{\mathrm{F}}$ is typically below 10^{-7}); in comparison, the relative estimation errors of both $\boldsymbol{Z}_{\mathsf{cvx}}$ and $\boldsymbol{Z}_{\mathsf{ncvx}}$ are substantially larger. In other words, the estimate returned by the nonconvex approach serves as a remarkably accurate approximation of the convex solution. Given that the nonconvex approach is often guaranteed to achieve intriguing statistical guarantees vis-à-vis random noise [MWCC17], this suggests that the convex program is equally stable — a phenomenon that was not captured by prior theory [CP10]. *Can we leverage existing theory for the nonconvex scheme to improve the statistical analysis of the convex relaxation approach?*

Before continuing, we remark that the above numerical connection between convex relaxation (3) and nonconvex optimization (5) has already been observed multiple times in prior literature [Faz02, SS05, RFP10,

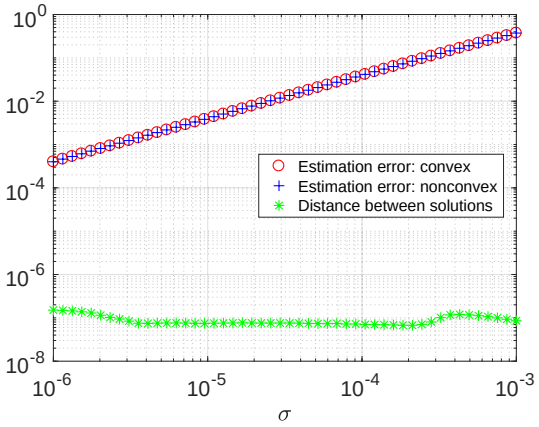


Figure 1: The relative estimation errors of both $\boldsymbol{Z}_{\mathsf{cvx}}$ (the estimate of the convex program (3)) and $\boldsymbol{Z}_{\mathsf{ncvx}}$ (the estimate returned by the nonconvex approach tailored to (5)) and the relative distance between them vs. the standard deviation σ of the noise. The results are reported for $n = 1000$, $r = 5$, $p = 0.2$, $\lambda = 5\sigma\sqrt{np}$ and are averaged over 20 independent trials.

MHT10, KMO10b]. Nevertheless, all prior observations on this connection were either completely empirical, or provided in a way that does not lead to improved statistical error bounds of the convex paradigm (3). In fact, the difficulty in rigorously justifying the above numerical observations has been noted in the literature; see e.g. [KMO10b].[3]

1.4 Models and main results

The numerical experiments reported in Section 1.3 suggest an alternative route for analyzing convex relaxation for noisy matrix completion. If one can formally justify the proximity between the convex and the nonconvex solutions, then it is possible to propagate the appealing stability guarantees from the nonconvex scheme to the convex approach. As it turns out, this simple idea leads to significantly enhanced statistical guarantees for the convex program (3), which we formally present in this subsection.

1.4.1 Models and assumptions

Before proceeding, we introduce a few model assumptions that play a crucial role in our theory.

Assumption 1.

(a) **(Random sampling)** Each index (i, j) belongs to the index set Ω independently with probability p.

(b) **(Random noise)** The noise matrix $\boldsymbol{E} = [E_{ij}]_{1\le i,j\le n}$ is composed of i.i.d. zero-mean sub-Gaussian random variables with sub-Gaussian norm at most $\sigma > 0$, i.e. $\|E_{ij}\|_{\psi_2} \le \sigma$ (see [Ver12, Definition 5.7]).

In addition, let $\boldsymbol{M}^\star = \boldsymbol{U}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}$ be the singular value decomposition (SVD) of $\boldsymbol{M}^\star$, where $\boldsymbol{U}^\star, \boldsymbol{V}^\star \in \mathbb{R}^{n\times r}$ consist of orthonormal columns and $\boldsymbol{\Sigma}^\star = \mathsf{diag}(\sigma_1^\star, \sigma_2^\star, \cdots, \sigma_r^\star) \in \mathbb{R}^{r\times r}$ is a diagonal matrix obeying $\sigma_{\max} \triangleq \sigma_1^\star \ge \sigma_2^\star \ge \cdots \ge \sigma_r^\star \triangleq \sigma_{\min}$. Denote by $\kappa \triangleq \sigma_{\max}/\sigma_{\min}$ the condition number of $\boldsymbol{M}^\star$. We impose the following incoherence condition on $\boldsymbol{M}^\star$, which is known to be crucial for reliable recovery of $\boldsymbol{M}^\star$ [CR09, Che15].

Definition 1. *A rank-r matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$ with SVD $\boldsymbol{M}^\star = \boldsymbol{U}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}$ is said to be μ-incoherent if*

$$\|\boldsymbol{U}^\star\|_{2,\infty} \le \sqrt{\frac{\mu}{n}}\,\|\boldsymbol{U}^\star\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{n}} \qquad \text{and} \qquad \|\boldsymbol{V}^\star\|_{2,\infty} \le \sqrt{\frac{\mu}{n}}\,\|\boldsymbol{V}^\star\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{n}}.$$

[3] The seminal work [KMO10b] by Keshavan, Montanari and Oh stated that "*In view of the identity (6) it might be possible to use the results in this paper to prove stronger guarantees on the nuclear norm minimization approach. Unfortunately this implication is not immediate ... Trying to establish such an implication, and clarifying the relation between the two approaches is nevertheless a promising research direction.*"

Here, $\|\boldsymbol{U}\|_{2,\infty}$ denotes the largest ℓ_2 norm of all rows of a matrix $\boldsymbol{U}$.

Remark 1. It is worth noting that several other conditions on the low-rank matrix have been proposed in the noisy setting. Examples include the spikiness condition [NW12] and the bounded ℓ_∞ norm condition [KLT11]. However, these conditions alone are often unable to ensure identifiability of the true matrix even in the absence of noise.

1.4.2 Theoretical guarantees: when both the rank and the condition number are constants

With these in place, we are positioned to present our improved statistical guarantees for convex relaxation. For convenience of presentation, we shall begin with a simple yet fundamentally important class of settings when the rank r and the condition number κ are both fixed constants. As it turns out, this class of problems arises in a variety of engineering applications. For example, in a fundamental problem in cryo-EM called angular synchronization [Sin11], one needs to deal with rank-2 or rank-3 matrices with $\kappa = 1$; in a joint shape mapping problem that arises in computer graphics [HG13, CGH14], the matrix under consideration has low rank and a condition number equal to 1; and in structure from motion in computer vision [TK92], one often seeks to estimate a matrix with $r \le 3$ and a small condition number. Encouragingly, our theory delivers near-optimal statistical guarantees for such practically important scenarios.

Theorem 1. *Let $\boldsymbol{M}^\star$ be rank-r and μ-incoherent with a condition number κ, where the rank and the condition number satisfy $r, \kappa = O(1)$. Suppose that Assumption 1 holds and take $\lambda = C_\lambda \sigma \sqrt{np}$ in (3) for some large enough constant $C_\lambda > 0$. Assume the sample size obeys $n^2 p \ge C\mu^2 n \log^3 n$ for some sufficiently large constant $C > 0$, and the noise satisfies $\sigma \lesssim \sqrt{\frac{np}{\mu^3 \log n}} \|\boldsymbol{M}^\star\|_\infty$ for some sufficiently small constant $c > 0$. Then with probability exceeding $1 - O(n^{-3})$:*

1. *Any minimizer $\boldsymbol{Z}_{\mathsf{cvx}}$ of (3) obeys*

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\|\boldsymbol{M}^\star\right\|_{\mathrm{F}}; \quad \left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\| \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\|\boldsymbol{M}^\star\right\|; \tag{7a}$$

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\infty} \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\mu n \log n}{p}} \left\|\boldsymbol{M}^\star\right\|_{\infty}. \tag{7b}$$

2. *Letting $\boldsymbol{Z}_{\mathsf{cvx},r} \triangleq \arg\min_{\boldsymbol{Z}:\mathsf{rank}(\boldsymbol{Z}) \le r} \|\boldsymbol{Z} - \boldsymbol{Z}_{\mathsf{cvx}}\|_{\mathrm{F}}$ be the best rank-r approximation of $\boldsymbol{Z}_{\mathsf{cvx}}$, we have*

$$\|\boldsymbol{Z}_{\mathsf{cvx},r} - \boldsymbol{Z}_{\mathsf{cvx}}\|_{\mathrm{F}} \le \frac{1}{n^3} \cdot \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\|\boldsymbol{M}^\star\right\|, \tag{8}$$

and the error bounds in (7) continue to hold if $\boldsymbol{Z}_{\mathsf{cvx}}$ is replaced by $\boldsymbol{Z}_{\mathsf{cvx},r}$.

Remark 2. Here and throughout, $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ means $|f(n)|/|g(n)| \le C$ for some constant $C > 0$ when n is sufficiently large; $f(n) \gtrsim g(n)$ means $|f(n)|/|g(n)| \ge C$ for some constant $C > 0$ when n is sufficiently large; and $f(n) \asymp g(n)$ if and only if $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$. In addition, $\|\cdot\|_\infty$ denotes the entrywise ℓ_∞ norm, whereas $\|\cdot\|$ is the spectral norm.

Remark 3. The factor $1/n^3$ in (8) can be replaced by $1/n^c$ for an arbitrarily large fixed constant $c > 0$ (e.g. $c = 100$).

To explain the applicability of the above theorem, we first remark on the conditions required for this theorem to hold; for simplicity, we assume that $\mu = O(1)$.

- *Sample complexity.* To begin with, the sample size needs to exceed the order of $n\mathrm{poly}\log n$, which is information-theoretically optimal up to some logarithmic term [CT10].
- *Noise size.* We then turn attention to the noise requirement, i.e. $\sigma \lesssim \sqrt{\frac{np}{\log n}} \|\boldsymbol{M}^\star\|_\infty$. Note that under the sample size condition $n^2 p \ge Cn \log^3 n$, the size of the noise in each entry is allowed to be substantially larger than the maximum entry in the matrix. In other words, the signal-to-noise ratio w.r.t. each observed

entry could be very small. According to prior literature (e.g. [KMO10b, Theorem 1.1] and [MWCC17, Theorem 2]), such noise conditions are typically required for spectral methods to perform noticeably better than random guessing.

Further, Theorem 1 has several important implications about the power of convex relaxation. The discussions below again concentrate on the case where $\mu = O(1)$.

- *Near-optimal stability guarantees.* Our results reveal that the Euclidean error of any convex optimizer $\boldsymbol{Z}_{\mathsf{cvx}}$ of (3) obeys
$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \sigma\sqrt{n/p}, \tag{9}$$
implying that the performance of convex relaxation degrades gracefully as the signal-to-noise ratio decreases. This result matches the oracle lower bound derived in [CP10, Eq. (III.13)], which also improves upon their statistical guarantee. Specifically, Candès and Plan [CP10] provided a stability guarantee in the presence of arbitrary bounded noise. When applied to the random noise model assumed here, their results yield $\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \sigma n^{3/2}$, which could be $O(\sqrt{n^2 p})$ times more conservative than our bound (9).

- *Nearly low-rank structure of the convex solution.* In light of (8), the optimizer of the convex program (3) is almost, if not exactly, rank-r. When the true rank r is known *a priori*, it is not uncommon for practitioners to return the rank-r approximation of $\boldsymbol{Z}_{\mathsf{cvx}}$. Our theorem formally justifies that there is no loss of statistical accuracy — measured in terms of either $\|\cdot\|_{\mathrm{F}}$ or $\|\cdot\|_\infty$ — when performing the rank-r projection operation.

- *Entrywise and spectral norm error control.* Moving beyond the Euclidean loss, our theory uncovers that the estimation errors of the convex optimizer are fairly spread out across all entries, thus implying near-optimal entrywise error control. This is a stronger form of error bounds, as an optimal Euclidean estimation accuracy alone does not preclude the possibility of the estimation errors being spiky and localized. Furthermore, the spectral norm error of the convex optimizer is also well-controlled. Figure 2 displays the relative estimation errors in both the ℓ_∞ norm and the spectral norm, under the same setting as in Figure 1. As can be seen, both forms of estimation errors scale linearly with the noise level, corroborating our theory.

- *Implicit regularization.* As a byproduct of the entrywise error control, this result indicates that the additional constraint $\|\boldsymbol{Z}\|_\infty \le \alpha$ suggested by [NW12] is automatically satisfied and is hence unnecessary. In other words, the convex approach implicitly controls the spikiness of its entries, without resorting to explicit regularization. This is also confirmed by the numerical experiments reported in Figure 3, where we see that the estimation error of (3) and that of the constrained version considered in [NW12] are nearly identical.

- *Statistical guarantees for fast iterative optimization methods.* Various iterative algorithms have been developed to solve the nuclear norm regularized least-squares problem (3) up to an arbitrarily prescribed accuracy, examples including SVT (or proximal gradient methods) [CCS10], FPC [MGC11], SOFT–IMPUTE [MHT10], FISTA [BT09, TY10], to name just a few. Our theory immediately provides statistical guarantees for these algorithms. As we shall make precise in Section 2, any point $\boldsymbol{Z}$ with $g(\boldsymbol{Z}) \le g(\boldsymbol{Z}_{\mathsf{cvx}}) + \varepsilon$ (where $g(\cdot)$ is defined in (3)) enjoys the same error bounds as in (7) (with $\boldsymbol{Z}_{\mathsf{cvx}}$ replaced by $\boldsymbol{Z}$ in (7)), provided that $\varepsilon > 0$ is sufficiently small. In other words, when these convex optimization algorithms converge w.r.t. the objective value, they are guaranteed to return a statistically reliable estimate.

To better understand our contributions, we take a moment to discuss two important but different convex programs studied in [NW12] and [KLT11]. To begin with, under a spikiness assumption on the low-rank matrix, Negahban and Wainwright [NW12] proposed to enforce an extra entrywise constraint $\|\boldsymbol{Z}\|_\infty \le \alpha$ when solving (3), in order to explicitly control the spikiness of the estimate. When applied to our model with $r, \kappa, \mu \asymp 1$, their results read (up to some logarithmic factor)
$$\left\|\hat{\boldsymbol{Z}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \max\left\{\sigma, \|\boldsymbol{M}^\star\|_\infty\right\}\sqrt{n/p}, \tag{10}$$
where $\hat{\boldsymbol{Z}}$ is the estimate returned by their modified convex algorithm. While this matches the optimal bound when $\sigma \gtrsim \|\boldsymbol{M}^\star\|_\infty$, it becomes suboptimal when $\sigma \ll \|\boldsymbol{M}^\star\|_\infty$ (under our models). Moreover,

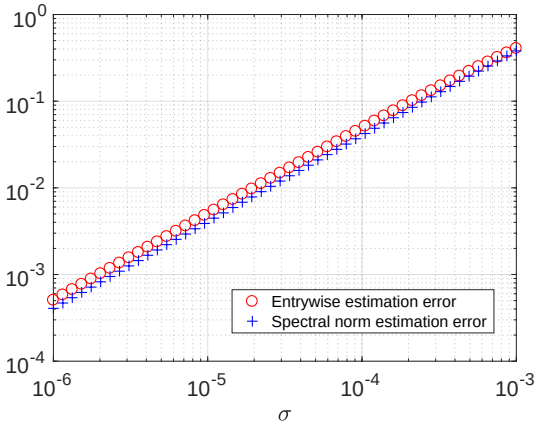


Figure 2: The relative estimation error of $\boldsymbol{Z}_{\mathsf{cvx}}$ measured by both $\|\cdot\|_\infty$ (i.e. $\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{M}^\star\|_\infty/\|\boldsymbol{M}^\star\|_\infty$) and $\|\cdot\|$ (i.e. $\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{M}^\star\|/\|\boldsymbol{M}^\star\|$) vs. the standard deviation σ of the noise. The results are reported for $n=1000$, $r=5$, $p=0.2$, $\lambda=5\sigma\sqrt{np}$ and are averaged over 20 independent trials.

as we have already discussed, the extra spikiness constraint becomes unnecessary in the regime considered herein. This also means that our result complements existing theory about the convex program in [NW12] by demonstrating its minimaxity for an additional range of noise. Another work by Koltchinskii et al. [KLT11] investigated a completely different convex algorithm, which is effectively a spectral method (namely, one round of soft singular value thresholding on a rescaled zero-padded data matrix). The algorithm is shown to be minimax optimal over the class of low-rank matrices with bounded ℓ_∞ norm (note that this is very different from the set of incoherent matrices studied here). When specialized to our model, their error bound is the same as (10) (modulo some log factor), which also becomes suboptimal as σ decreases. As can be seen from the numerical experiments in Figure 3, the estimation error of this thresholding-based spectral algorithm does not decrease as the noise shrinks, and its performance seems uniformly outperformed by that of convex relaxation (3) and the constrained estimator in [NW12]. In fact, this is part of our motivation to pursue an improved theoretical understanding of the formulation (3).

Finally, we make note of a connection between our result and prior theory developed for the noiseless case. Specifically, when the noise vanishes (i.e. $\sigma\to 0$), one can take a diminishing sequence of regularization parameters $\{\lambda_k\}$ with $\lambda_k\to 0$, then the resulting estimation errors associated with this sequence should decrease to 0 as $k\to\infty$ (which implies exact recovery in the limit of k). This parallels the connection between Lasso in sparse linear regression and basis pursuit in compressed sensing.

1.4.3 Theoretical guarantees: extensions to more general settings

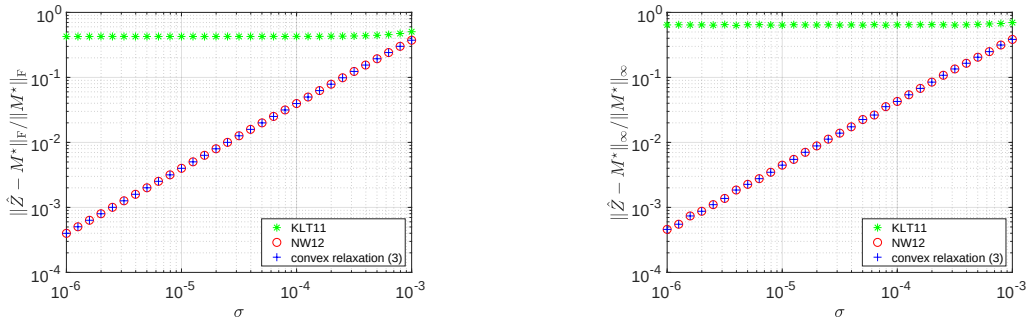
So far we have presented results when the true matrix has bounded rank and condition number, i.e. $r,\kappa=O(1)$. Our theory actually accommodates a significantly broader range of scenarios, where the rank and the condition number are both allowed to grow with the dimension n.

Theorem 2. *Let $\boldsymbol{M}^\star$ be rank-r and μ-incoherent with a condition number κ. Suppose Assumption 1 holds and take $\lambda=C_\lambda\sigma\sqrt{np}$ in (3) for some large enough constant $C_\lambda>0$. Assume the sample size obeys $n^2p\geq C\kappa^4\mu^2r^2n\log^3 n$ for some sufficiently large constant $C>0$, and the noise satisfies $\sigma\sqrt{\frac{n}{p}}\leq c\frac{\sigma_{\min}}{\sqrt{\kappa^4\mu r\log n}}$ for some sufficiently small constant $c>0$. Then with probability exceeding $1-O(n^{-3})$,*

1. *Any minimizer $\boldsymbol{Z}_{\mathsf{cvx}}$ of (3) obeys*

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{M}^\star\right\|_{\mathrm{F}}\lesssim\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{M}^\star\right\|_{\mathrm{F}},\tag{11a}$$

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{M}^\star\right\|_{\infty}\lesssim\sqrt{\kappa^3\mu r}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{M}^\star\right\|_{\infty},\tag{11b}$$



(a) (b)

Figure 3: The relative estimation errors of $\hat{\boldsymbol{Z}}$, measured in terms of ℓ_{F} and ℓ_∞, vs. the standard deviation σ of the noise. Here $\hat{\boldsymbol{Z}}$ can be either the modified convex estimator in [KLT11], the constrained convex estimator in [NW12] or the vanilla convex estimator (3). The results are reported for $n = 1000$, $r = 5$, $p = 0.2$, and are averaged over 20 Monte-Carlo trials. For the modified convex estimator in [KLT11], we choose the regularization parameter λ therein to be $1.5\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{1/(n^3p)}$, as suggested by their theory. For the constrained one in [NW12], the regularization parameter λ is set to be $5\sigma\sqrt{np}$ and the constraint α is set to be $\|\boldsymbol{M}^\star\|_\infty$. Both choices are recommended by [NW12]. As for (3), we set $\lambda = 5\sigma\sqrt{np}$.

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\| \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{M}^\star\right\|; \tag{11c}$$

2. *Letting* $\boldsymbol{Z}_{\mathsf{cvx},r} \triangleq \arg\min_{\boldsymbol{Z}:\mathsf{rank}(\boldsymbol{Z})\le r}\|\boldsymbol{Z} - \boldsymbol{Z}_{\mathsf{cvx}}\|_{\mathrm{F}}$ *be the best rank-r approximation of* $\boldsymbol{Z}_{\mathsf{cvx}}$*, we have*

$$\|\boldsymbol{Z}_{\mathsf{cvx},r} - \boldsymbol{Z}_{\mathsf{cvx}}\|_{\mathrm{F}} \le \frac{1}{n^3}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{M}^\star\right\|, \tag{12}$$

and the error bounds in (11) *continue to hold if* $\boldsymbol{Z}_{\mathsf{cvx}}$ *is replaced by* $\boldsymbol{Z}_{\mathsf{cvx},r}$.

Remark 4 (The noise condition). The incoherence condition (cf. Definition 1) guarantees that the largest entry $\|\boldsymbol{M}^\star\|_\infty$ of the matrix $\boldsymbol{M}^\star$ is no larger than $\kappa\mu r\sigma_{\min}/n$. As a result, the noise condition stated in Theorem 2 covers all scenarios obeying

$$\sigma \lesssim \sqrt{\frac{np}{\kappa^6\mu^3 r^3\log n}}\left\|\boldsymbol{M}^\star\right\|_\infty.$$

Therefore, the typical size of the noise is allowed to be much larger than the size of the largest entry of $\boldsymbol{M}^\star$, provided that $p \gg \frac{\kappa^6\mu^3 r^3\log n}{n}$. In particular, when $r, \kappa = O(1)$, this recovers the noise condition in Theorem 1.

Notably, the sample size condition for noisy matrix completion (i.e. $n^2p \ge C\kappa^4\mu^2r^2n\log^3 n$) is more stringent than that in the noiseless setting (i.e. $n^2p \asymp nr\log^2 n$), and our statistical guarantees are likely suboptimal with respect to the dependency on r and κ. This sub-optimality is mainly due to the analysis of nonconvex optimization, a key ingredient of our analysis of convex relaxation. In fact, the state-of-the-art nonconvex analysis [KMO10b, CW15, MWCC17] requires the sample size to be much larger than the optimal one (e.g. $n^2p \gg n\mathsf{poly}(r)\mathsf{poly}(\kappa)$) even in the noiseless setting. It would certainly be interesting, and in fact important, to see whether it is possible to develop a theory with optimal dependency on r and κ. We leave this for future investigation.

Despite the above sub-optimality issue, implications similar to those of Theorem 1 hold for this general setting. To begin with, the nearly low-rank structure of the convex solution is preserved (cf. (12)). In addition, the estimation error of the convex estimate is spread out across entries (cf. (11b)), thus uncovering an implicit regularization phenomenon underlying convex relaxation (which implicitly regularizes the spikiness constraint on the solution). Last but not least, the upper bounds (11) and (12) continue to hold for approximate minimizers of the convex program (3), thus yielding statistical guarantees for numerous iterative algorithms aimed at minimizing (3).

2 Strategy and novelty

In this section, we introduce the strategy for proving our main theorem, i.e. Theorem 2. Theorem 1 follows immediately. Informally, the main technical difficulty stems from the lack of closed-form expressions for the primal solution to (3), which in turn makes it difficult to construct a dual certificate. This is in stark contrast to the noiseless setting, where one clearly anticipates the ground truth $\boldsymbol{M}^\star$ to be the primal solution; in fact, this is precisely why the analysis for the noisy case is significantly more challenging. Our strategy, as we shall detail below, mainly entails invoking an iterative nonconvex algorithm to "approximate" such a primal solution.

Before continuing, we introduce a few more notations. Let $\mathcal{P}_\Omega(\cdot) : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ represent the projection onto the subspace of matrices supported on Ω, namely,

$$[\mathcal{P}_\Omega(\boldsymbol{Z})]_{ij} = \begin{cases} Z_{ij}, & \text{for } (i,j)\in\Omega \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

for any matrix $\boldsymbol{Z}\in\mathbb{R}^{n\times n}$. For a rank-$r$ matrix $\boldsymbol{M}$ with singular value decomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, denote by T its tangent space, i.e.

$$T = \left\{\boldsymbol{U}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{V}^\top \mid \boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{n\times r}\right\}. \tag{14}$$

Correspondingly, let $\mathcal{P}_T(\cdot)$ be the orthogonal projection onto the subspace T, that is,

$$\mathcal{P}_T(\boldsymbol{Z}) = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{Z}\boldsymbol{V}\boldsymbol{V}^\top \tag{15}$$

for any matrix $\boldsymbol{Z}\in\mathbb{R}^{n\times n}$. In addition, let $T^\perp$ and $\mathcal{P}_{T^\perp}(\cdot)$ denote the orthogonal complement of T and the projection onto $T^\perp$, respectively. With regards to the ground truth, we denote

$$\boldsymbol{X}^\star = \boldsymbol{U}^\star(\boldsymbol{\Sigma}^\star)^{1/2} \qquad \text{and} \qquad \boldsymbol{Y}^\star = \boldsymbol{V}^\star(\boldsymbol{\Sigma}^\star)^{1/2}. \tag{16}$$

The nonconvex problem (5) is equivalent to

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \qquad f(\boldsymbol{X},\boldsymbol{Y}) \triangleq \frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{Y}\|_{\mathrm{F}}^2, \tag{17}$$

where we have inserted an extra factor $1/p$ (compared to (5)) to simplify the presentation of the analysis later on.

2.1 Exact duality

In order to analyze the convex program (3), it is natural to start with the first-order optimality condition. Specifically, suppose that $\boldsymbol{Z}\in\mathbb{R}^{n\times n}$ is a (primal) solution to (3) with SVD $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$.[4] As before, let T be the tangent space of $\boldsymbol{Z}$, and let $T^\perp$ be the orthogonal complement of T. Then the first-order optimality condition for (3) reads: there exists a matrix $\boldsymbol{W}\in T^\perp$ (called a dual certificate) such that

$$\frac{1}{\lambda}\mathcal{P}_\Omega(\boldsymbol{M}-\boldsymbol{Z}) = \boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W}; \tag{18a}$$

$$\|\boldsymbol{W}\| \le 1. \tag{18b}$$

This condition is not only necessary to certify the optimality of $\boldsymbol{Z}$, but also "almost sufficient" in guaranteeing the uniqueness of the solution $\boldsymbol{Z}$; see Appendix B for in-depth discussions.

The challenge then boils down to identifying such a primal-dual pair $(\boldsymbol{Z},\boldsymbol{W})$ satisfying the optimality condition (18). For the noise-free case, the primal solution is clearly $\boldsymbol{Z} = \boldsymbol{M}^\star$ if exact recovery is to be expected; the dual certificate can then be either constructed exactly by the least-squares solution to a certain underdetermined linear system [CR09, CT10], or produced approximately via a clever golfing scheme pioneered by Gross [Gro11]. For the noisy case, however, it is often difficult to hypothesize on the primal solution $\boldsymbol{Z}$, as it depends on the random noise in a complicated way. In fact, the lack of a suitable guess of $\boldsymbol{Z}$ (and hence $\boldsymbol{W}$) was the major hurdle that prior works faced when carrying out the duality analysis.

[4] Here and below, we use $\boldsymbol{Z}$ (rather than $\boldsymbol{Z}_{\mathsf{cvx}}$) for notational simplicity, whenever it is clear from the context.

2.2 A candidate primal solution via nonconvex optimization

Motivated by the numerical experiment in Section 1.3, we propose to examine whether the optimizer of the nonconvex problem (5) stays close to the solution to the convex program (3). Towards this, suppose that $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}$ form a critical point of (5) with $\mathsf{rank}(\boldsymbol{X}) = \mathsf{rank}(\boldsymbol{Y}) = r$.[5] Then the first-order condition reads

$$\frac{1}{\lambda}\mathcal{P}_\Omega(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top)\boldsymbol{Y} = \boldsymbol{X}; \tag{19a}$$

$$\frac{1}{\lambda}\left[\mathcal{P}_\Omega(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top)\right]^\top \boldsymbol{X} = \boldsymbol{Y}. \tag{19b}$$

To develop some intuition about the connection between (18) and (19), let us take a look at the case with $r = 1$. Denote $\boldsymbol{X} = \boldsymbol{x}$ and $\boldsymbol{Y} = \boldsymbol{y}$ and assume that the two rank-1 factors are "balanced", namely, $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 \neq 0$. It then follows from (19) that $\lambda^{-1}\mathcal{P}_\Omega(\boldsymbol{M} - \boldsymbol{x}\boldsymbol{y}^\top)$ has a singular value 1, whose corresponding left and right singular vectors are $\boldsymbol{x}/\|\boldsymbol{x}\|_2$ and $\boldsymbol{y}/\|\boldsymbol{y}\|_2$, respectively. In other words, one can express

$$\frac{1}{\lambda}\mathcal{P}_\Omega(\boldsymbol{M} - \boldsymbol{x}\boldsymbol{y}^\top) = \frac{1}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2}\boldsymbol{x}\boldsymbol{y}^\top + \boldsymbol{W}, \tag{20}$$

where $\boldsymbol{W}$ is orthogonal to the tangent space of $\boldsymbol{x}\boldsymbol{y}^\top$; this is precisely the condition (18a). It remains to argue that (18b) is valid as well. Towards this end, the first-order condition (19) alone is insufficient, as there might be non-global critical points (e.g. saddle points) that are unable to approximate the convex solution well. Fortunately, as long as the candidate $\boldsymbol{x}\boldsymbol{y}^\top$ is not far away from the ground truth $\boldsymbol{M}^\star$, one can guarantee $\|\boldsymbol{W}\| < 1$ as required in (18b).

The above informal argument about the link between the convex and the nonconvex problems can be rigorized. To begin with, we introduce the following conditions on the regularization parameter λ.

Condition 1 (**Regularization parameter**). The regularization parameter λ satisfies

(a) **(Relative to noise)** $\|\mathcal{P}_\Omega(\boldsymbol{E})\| < \lambda/8$.

(b) **(Relative to nonconvex solution)** $\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star) - p(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star)\| < \lambda/8$.

Remark 5. Condition 1 requires that the regularization parameter λ should dominate a certain norm of the noise, as well as of the deviation of $\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star$ from its mean $p(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star)$; as will be seen shortly, the latter condition can be met when $(\boldsymbol{X}, \boldsymbol{Y})$ is sufficiently close to $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$.

With the above condition in place, the following result demonstrates that a critical point $(\boldsymbol{X}, \boldsymbol{Y})$ of the nonconvex problem (5) readily translates to the unique minimizer of the convex program (3). This lemma is established in Appendix C.1.

Lemma 1 (**Exact nonconvex vs. convex optimizers**). *Suppose that $(\boldsymbol{X}, \boldsymbol{Y})$ is a critical point of (5) satisfying $\mathsf{rank}(\boldsymbol{X}) = \mathsf{rank}(\boldsymbol{Y}) = r$, and the sampling operator $\mathcal{P}_\Omega$ is injective when restricted to the elements of the tangent space T of $\boldsymbol{X}\boldsymbol{Y}^\top$, namely,*

$$\mathcal{P}_\Omega(\boldsymbol{H}) = \boldsymbol{0} \quad \Longleftrightarrow \quad \boldsymbol{H} = \boldsymbol{0}, \quad \text{for all } \boldsymbol{H} \in T. \tag{21}$$

Under Condition 1, the point $\boldsymbol{Z} \triangleq \boldsymbol{X}\boldsymbol{Y}^\top$ is the unique minimizer of (3).

In order to apply Lemma 1, one needs to locate a critical point of (5) that is sufficiently close to the truth, for which one natural candidate is the global optimizer of (5). The caveat, however, is the lack of theory characterizing directly the properties of the optimizer of (5). Instead, what is available in prior theory is the characterization of some iterative sequence (e.g. gradient descent iterates) aimed at solving (5). It is unclear from prior theory whether the iterative algorithm under study (e.g. gradient descent) converges to the global optimizer in the presence of noise. This leads to technical difficulty in justifying the proximity between the nonconvex optimizer and the convex solution via Lemma 1.

[5] Once again, we abuse the notation $(\boldsymbol{X}, \boldsymbol{Y})$ (instead of using $(\boldsymbol{X}_{\mathsf{ncvx}}, \boldsymbol{Y}_{\mathsf{ncvx}})$) for notational simplicity, whenever it is clear from the context.

2.3 Approximate nonconvex optimizers

Fortunately, perfect knowledge of the nonconvex optimizer is not pivotal. Instead, an approximate solution to the nonconvex problem (5) (or equivalently (17)) suffices to serve as a reasonably tight approximation of the convex solution. More precisely, we desire two factors $(\boldsymbol{X}, \boldsymbol{Y})$ that result in nearly zero (rather than exactly zero) gradients:

$$\nabla_{\boldsymbol{X}} f(\boldsymbol{X}, \boldsymbol{Y}) \approx \boldsymbol{0} \qquad \text{and} \qquad \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}, \boldsymbol{Y}) \approx \boldsymbol{0},$$

where $f(\cdot, \cdot)$ is the nonconvex objective function as defined in (17). This relaxes the condition discussed in Lemma 1 (which only applies to critical points of (5) as opposed to approximate critical points). As it turns out, such points can be found via gradient descent tailored to (5). The sufficiency of the near-zero gradient condition is made possible by slightly strengthening the injectivity assumption (21), which is stated below.

Condition 2 (**Injectivity**). Let T be the tangent space of $\boldsymbol{X}\boldsymbol{Y}^\top$. There is a quantity $c_{\text{inj}} > 0$ such that

$$p^{-1} \left\| \mathcal{P}_\Omega(\boldsymbol{H}) \right\|_{\text{F}}^2 \geq c_{\text{inj}} \left\| \boldsymbol{H} \right\|_{\text{F}}^2, \qquad \text{for all } \boldsymbol{H} \in T. \tag{22}$$

The following lemma states quantitatively how an approximate nonconvex optimizer serves as an excellent proxy of the convex solution, which we establish in Appendix C.2.

Lemma 2 (**Approximate nonconvex vs. convex optimizers**). *Suppose that* $(\boldsymbol{X}, \boldsymbol{Y})$ *obeys*

$$\left\| \nabla f(\boldsymbol{X}, \boldsymbol{Y}) \right\|_{\text{F}} \leq c \frac{\sqrt{c_{\text{inj}} p}}{\kappa} \cdot \frac{\lambda}{p} \sqrt{\sigma_{\min}} \tag{23}$$

for some sufficiently small constant $c > 0$. *Further assume that any singular value of* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *lies in* $[\sqrt{\sigma_{\min}/2}, \sqrt{2\sigma_{\max}}]$. *Then under Conditions 1 and 2, any minimizer* $\boldsymbol{Z}_{\mathsf{cvx}}$ *of (3) satisfies*

$$\left\| \boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{Z}_{\mathsf{cvx}} \right\|_{\text{F}} \lesssim \frac{\kappa}{c_{\text{inj}}} \frac{1}{\sqrt{\sigma_{\min}}} \left\| \nabla f(\boldsymbol{X}, \boldsymbol{Y}) \right\|_{\text{F}}. \tag{24}$$

Remark 6. In fact, this lemma continues to hold if $\boldsymbol{Z}_{\mathsf{cvx}}$ is replaced by any $\boldsymbol{Z}$ obeying $g(\boldsymbol{Z}) \leq g(\boldsymbol{X}\boldsymbol{Y}^\top)$, where $g(\cdot)$ is the objective function defined in (3) and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are low-rank factors obeying conditions of Lemma 2. This is important in providing statistical guarantees for iterative methods like SVT [CCS10], FPC [MGC11], SOFT–IMPUTE [MHT10], FISTA [BT09], etc. To be more specific, suppose that $(\boldsymbol{X}, \boldsymbol{Y})$ results in an approximate optimizer of (3), namely, $g(\boldsymbol{X}\boldsymbol{Y}^\top) = g(\boldsymbol{Z}_{\mathsf{cvx}}) + \varepsilon$ for some sufficiently small $\varepsilon > 0$. Then for any $\boldsymbol{Z}$ obeying $g(\boldsymbol{Z}) \leq g(\boldsymbol{X}\boldsymbol{Y}^\top) = g(\boldsymbol{Z}_{\mathsf{cvx}}) + \varepsilon$, one has

$$\left\| \boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{Z} \right\|_{\text{F}} \lesssim \frac{\kappa}{c_{\text{inj}}} \frac{1}{\sqrt{\sigma_{\min}}} \left\| \nabla f(\boldsymbol{X}, \boldsymbol{Y}) \right\|_{\text{F}}. \tag{25}$$

As a result, as long as the above-mentioned algorithms converge in terms of the objective value, they must return a solution obeying (25), which is exceedingly close to $\boldsymbol{X}\boldsymbol{Y}^\top$ if $\|\nabla f(\boldsymbol{X}, \boldsymbol{Y})\|_{\text{F}}$ is small.

It is clear from Lemma 2 that, as the size of the gradient $\nabla f(\boldsymbol{X}, \boldsymbol{Y})$ gets smaller, the nonconvex estimate $\boldsymbol{X}\boldsymbol{Y}^\top$ becomes an increasingly tighter approximation of any convex optimizer of (3), which is consistent with Lemma 1. In contrast to Lemma 1, due to the lack of strong convexity, a nonconvex estimate with a near-zero gradient does not imply the uniqueness of the optimizer of the convex program (3); rather, it indicates that any minimizer of (3) lies within a sufficiently small neighborhood surrounding $\boldsymbol{X}\boldsymbol{Y}^\top$ (cf. (24)).

2.4 Construction of an approximate nonconvex optimizer

So far, Lemmas 1-2 are both deterministic results based on Condition 1. As we will soon see, under Assumption 1, we can derive simpler conditions that — with high probability — guarantee Condition 1. We start with Condition 1(a).

Lemma 3. *Suppose* $n^2 p \geq C n \log^2 n$ *for some sufficiently large constant* $C > 0$. *Then with probability at least* $1 - O(n^{-10})$, *one has* $\|\mathcal{P}_\Omega(\boldsymbol{E})\| \lesssim \sigma\sqrt{np}$. *As a result, Condition 1 holds (i.e.* $\|\mathcal{P}_\Omega(\boldsymbol{E})\| < \lambda/8$*) as long as* $\lambda = C_\lambda \sigma \sqrt{np}$ *for some sufficiently large constant* $C_\lambda > 0$.

Proof. This follows from [CW15, Lemma 11] with a slight and straightforward modification to accommodate the asymmetric noise here. For brevity, we omit the proof. □

Turning attention to Condition 1(b) and Condition 2, we have the following lemma, the proof of which is deferred to Appendix C.3.

Lemma 4. *Under the assumptions of Theorem 2, with probability exceeding* $1 - O(n^{-10})$ *we have*

$$\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star) - p(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star)\| < \lambda/8 \qquad \text{(Condition 1(b))}$$

$$\frac{1}{p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{H}\right)\right\|_{\mathrm{F}}^2 \geq \frac{1}{32\kappa}\left\|\boldsymbol{H}\right\|_{\mathrm{F}}^2, \quad \textit{for all } \boldsymbol{H} \in T \qquad \text{(Condition 2 with } c_{\mathrm{inj}} = (32\kappa)^{-1})$$

hold simultaneously for all $(\boldsymbol{X}, \boldsymbol{Y})$ *obeying*

$$\begin{aligned}&\max\left\{\|\boldsymbol{X} - \boldsymbol{X}^\star\|_{2,\infty}, \|\boldsymbol{Y} - \boldsymbol{Y}^\star\|_{2,\infty}\right\} \\ &\qquad \leq C_\infty \kappa \left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\max\left\{\|\boldsymbol{X}^\star\|_{2,\infty}, \|\boldsymbol{Y}^\star\|_{2,\infty}\right\}. \end{aligned} \tag{26}$$

Here, T *denotes the tangent space of* $\boldsymbol{X}\boldsymbol{Y}^\top$*, and* $C_\infty > 0$ *is some absolute constant.*

This lemma is a uniform result, namely, the bounds hold irrespective of the statistical dependency between $(\boldsymbol{X}, \boldsymbol{Y})$ and Ω. As a consequence, to demonstrate the proximity between the convex and the nonconvex solutions (cf. (24)), it remains to identify a point $(\boldsymbol{X}, \boldsymbol{Y})$ with vanishingly small gradient (cf. (23)) that is sufficiently close to the truth (cf. (26)).

As we already alluded to previously, a simple gradient descent algorithm aimed at solving the nonconvex problem (5) might help us produce an approximate nonconvex optimizer. This procedure is summarized in Algorithm 1. Our hope is this: when initialized at the ground truth and run for sufficiently many iterations, the GD trajectory produced by Algorithm 1 will contain at least one approximate stationary point of (5) with the desired properties (23) and (26). We shall note that Algorithm 1 is *not practical* since it starts from the ground truth $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$; this is an auxiliary step mainly to simplify the theoretical analysis. While we can certainly make it practical by adopting spectral initialization as in [MWCC17, CLL19], it requires more lengthy proofs without further improving our statistical guarantees.

Algorithm 1 Construction of an approximate primal solution.

Initialization: $\boldsymbol{X}^0 = \boldsymbol{X}^\star$; $\boldsymbol{Y}^0 = \boldsymbol{Y}^\star$.

Gradient updates: **for** $t = 0, 1, \ldots, t_0 - 1$ **do**

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta\nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) = \boldsymbol{X}^t - \frac{\eta}{p}\Big(\mathcal{P}_\Omega\left(\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}\right)\boldsymbol{Y}^t + \lambda\boldsymbol{X}^t\Big); \tag{27a}$$

$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \eta\nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) = \boldsymbol{Y}^t - \frac{\eta}{p}\Big(\left[\mathcal{P}_\Omega\left(\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}\right)\right]^\top\boldsymbol{X}^t + \lambda\boldsymbol{Y}^t\Big). \tag{27b}$$

Here, $\eta > 0$ is the step size.

2.5 Properties of the nonconvex iterates

In this subsection, we will build upon the literature on nonconvex low-rank matrix completion to justify that the estimates returned by Algorithm 1 satisfy the requirement stated in (26). Our theory will be largely established upon the leave-one-out strategy introduced by Ma et al. [MWCC17], which is an effective analysis technique to control the $\ell_{2,\infty}$ error of the estimates. This strategy has recently been extended by Chen et al. [CLL19] to the more general rectangular case with an improved sample complexity bound.

Before continuing, we introduce several useful notations. Notice that the matrix product of $\boldsymbol{X}^\star$ and $\boldsymbol{Y}^{\star\top}$ is invariant under global orthonormal transformation, namely, for any orthonormal matrix $\boldsymbol{R} \in \mathbb{R}^{r\times r}$ one has

$\boldsymbol{X}^\star \boldsymbol{R}(\boldsymbol{Y}^\star \boldsymbol{R})^\top = \boldsymbol{X}^\star \boldsymbol{Y}^{\star\top}$. Viewed in this light, we shall consider distance metrics modulo global rotation. In particular, the theory relies heavily on a specific global rotation matrix defined as follows

$$\boldsymbol{H}^t \triangleq \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left(\left\|\boldsymbol{X}^t\boldsymbol{R} - \boldsymbol{X}^\star\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y}^t\boldsymbol{R} - \boldsymbol{Y}^\star\right\|_{\mathrm{F}}^2 \right)^{1/2}, \tag{28}$$

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

We are now ready to present the performance guarantees for Algorithm 1.

Lemma 5 (Quality of the nonconvex estimates). *Instate the notation and hypotheses of Theorem 2. With probability at least $1 - O\left(n^{-3}\right)$, the iterates $\{(\boldsymbol{X}^t, \boldsymbol{Y}^t)\}_{0\le t\le t_0}$ of Algorithm 1 satisfy*

$$\max\left\{\left\|\boldsymbol{X}^t\boldsymbol{H}^t - \boldsymbol{X}^\star\right\|_{\mathrm{F}}, \left\|\boldsymbol{Y}^t\boldsymbol{H}^t - \boldsymbol{Y}^\star\right\|_{\mathrm{F}}\right\} \le C_{\mathrm{F}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{X}^\star\|_{\mathrm{F}}, \tag{29a}$$

$$\max\left\{\left\|\boldsymbol{X}^t\boldsymbol{H}^t - \boldsymbol{X}^\star\right\|, \left\|\boldsymbol{Y}^t\boldsymbol{H}^t - \boldsymbol{Y}^\star\right\|\right\} \le C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{X}^\star\|, \tag{29b}$$

$$\begin{aligned}\max\left\{\left\|\boldsymbol{X}^t\boldsymbol{H}^t - \boldsymbol{X}^\star\right\|_{2,\infty}, \left\|\boldsymbol{Y}^t\boldsymbol{H}^t - \boldsymbol{Y}^\star\right\|_{2,\infty}\right\}& \\ \le C_\infty\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)&\max\left\{\|\boldsymbol{X}^\star\|_{2,\infty}, \|\boldsymbol{Y}^\star\|_{2,\infty}\right\},\end{aligned} \tag{29c}$$

$$\min_{0\le t<t_0}\left\|\nabla f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t\right)\right\|_{\mathrm{F}} \le \frac{1}{n^5}\frac{\lambda}{p}\sqrt{\sigma_{\min}}, \tag{30}$$

where $C_{\mathrm{F}}, C_{\mathrm{op}}, C_\infty > 0$ are some absolute constants, provided that $\eta \asymp 1/(n\kappa^3\sigma_{\max})$ and that $t_0 = n^{18}$.

This lemma, which we establish in Appendix D, reveals that for a polynomially large number of iterations, all iterates of the gradient descent sequence — when initialized at the ground truth — remain fairly close to the true low-rank factors. This holds in terms of the estimation errors measured by the Frobenius norm, the spectral norm, and the $\ell_{2,\infty}$ norm. In particular, the proximity in terms of the $\ell_{2,\infty}$ norm error plays a pivotal role in implementing our analysis strategy (particularly Lemmas 2-4) described previously. In addition, this lemma (cf. (30)) guarantees the existence of a small-gradient point within this sequence $\{(\boldsymbol{X}^t, \boldsymbol{Y}^t)\}_{0\le t\le t_0}$, a somewhat straightforward property of GD tailored to smooth problems [Nes12]. This in turn enables us to invoke Lemma 2.

As immediate consequences of Lemma 5, with high probability we have

$$\left\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \le 3\kappa C_{\mathrm{F}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|_{\mathrm{F}} \tag{31a}$$

$$\left\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}^\star\right\|_{\infty} \le 3C_\infty\sqrt{\kappa^3\mu r}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|_{\infty} \tag{31b}$$

$$\left\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}^\star\right\| \le 3C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\| \tag{31c}$$

for all $0 \le t \le t_0$. The proof is deferred to Appendix D.12.

2.6 Proof of Theorem 2

Let $t_* \triangleq \arg\min_{0\le t<t_0}\|\nabla f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t\right)\|_{\mathrm{F}}$, and take $(\boldsymbol{X}_{\mathsf{ncvx}}, \boldsymbol{Y}_{\mathsf{ncvx}}) = (\boldsymbol{X}^{t_*}\boldsymbol{H}^{t_*}, \boldsymbol{Y}^{t_*}\boldsymbol{H}^{t_*})$ (cf. (28)). It is straightforward to verify that $(\boldsymbol{X}_{\mathsf{ncvx}}, \boldsymbol{Y}_{\mathsf{ncvx}})$ obeys (i) the small-gradient condition (23), and (ii) the proximity condition (26). We are now positioned to invoke Lemma 2: for any optimizer $\boldsymbol{Z}_{\mathsf{cvx}}$ of (3), one has

$$\begin{aligned}\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top\right\|_{\mathrm{F}} &\lesssim \frac{\kappa}{c_{\mathrm{inj}}}\frac{1}{\sqrt{\sigma_{\min}}}\left\|\nabla f(\boldsymbol{X}_{\mathsf{ncvx}}, \boldsymbol{Y}_{\mathsf{ncvx}})\right\|_{\mathrm{F}} \lesssim \frac{\kappa^2}{n^5}\frac{\lambda}{p} \\ &= \frac{\kappa}{n^5}\frac{\lambda}{p\,\sigma_{\min}}(\kappa\sigma_{\min}) = \frac{\kappa}{n^5}\frac{\lambda}{p\,\sigma_{\min}}\|\boldsymbol{M}^\star\|\end{aligned}$$

$$\lesssim \frac{1}{n^4}\frac{\lambda}{p\,\sigma_{\min}}\|\boldsymbol{M}^\star\|. \tag{32}$$

The last line arises since $n \gg \kappa$ — a consequence of the sample complexity condition $np \gtrsim \kappa^4\mu^2 r^2 \log^3 n$ (and hence $n \geq np \gtrsim \kappa^4\mu^2 r^2 \log^3 n \gg \kappa^4$). This taken collectively with the property (31) implies that

$$\begin{aligned}
\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} &\leq \left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top\right\|_{\mathrm{F}} + \left\|\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \\
&\lesssim \frac{1}{n^4}\frac{\lambda}{p\,\sigma_{\min}}\|\boldsymbol{M}^\star\| + \kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|_{\mathrm{F}} \\
&\asymp \kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|_{\mathrm{F}}.
\end{aligned}$$

In other words, since $\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top$ and $\boldsymbol{Z}_{\mathsf{ncvx}}$ are exceedingly close, the error $\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star$ is mainly accredited to $\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top - \boldsymbol{M}^\star$. Similar arguments lead to

$$\begin{aligned}
\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\| &\lesssim \left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|, \\
\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_\infty &\lesssim \sqrt{\kappa^3\mu r}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|_\infty.
\end{aligned}$$

We are left with proving the properties of $\boldsymbol{Z}_{\mathsf{cvx},r}$. Since $\boldsymbol{Z}_{\mathsf{cvx},r}$ is defined to be the best rank-r approximation of $\boldsymbol{Z}_{\mathsf{cvx}}$, one can invoke (32) to derive

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{Z}_{\mathsf{cvx},r}\right\|_{\mathrm{F}} \leq \left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^\top\right\|_{\mathrm{F}} \lesssim \frac{1}{n^4}\frac{\lambda}{p\,\sigma_{\min}}\|\boldsymbol{M}^\star\|,$$

from which (12) follows. Repeating the above calculations implies that (11) holds if $\boldsymbol{Z}_{\mathsf{cvx}}$ is replaced by $\boldsymbol{Z}_{\mathsf{cvx},r}$, thus concluding the proof.

3 Prior art

Nuclear norm minimization, pioneered by the seminal works [RFP10, CR09, CT10, Faz02], has been a popular and principled approach to low-rank matrix recovery. In the noiseless setting, i.e. $\boldsymbol{E} = \boldsymbol{0}$, it amounts to solving the following constrained convex program

$$\text{minimize}_{\boldsymbol{Z}\in\mathbb{R}^{n\times n}} \|\boldsymbol{Z}\|_* \qquad \text{subject to} \quad \mathcal{P}_\Omega(\boldsymbol{Z}) = \mathcal{P}_\Omega(\boldsymbol{M}^\star), \tag{33}$$

which enjoys great theoretical success. Informally, this approach enables exact recovery of a rank-r matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$ as soon as the sample size is about the order of nr — the intrinsic degrees of freedom of a rank-r matrix [Gro11, Rec11, Che15]. In particular, Gross [Gro11] blazed a trail by developing an ingenious golfing scheme for dual construction — an analysis technique that has found applications far beyond matrix completion. When it comes to the noisy case, Candès and Plan [CP10] first studied the stability of convex programming when the noise is bounded and possibly adversarial, followed by [NW12] and [KLT11] using two modified convex programs. As we have already discussed, none of these papers provide optimal statistical guarantees under our model when $r = O(1)$. Other related papers such as [Klo14, CZ16] include similar estimation error bounds and suffer from similar sub-optimality issues.

Turning to nonconvex optimization, we note that this approach has recently received much attention for various low-rank matrix factorization problems, owing to its superior computational advantage compared to convex programming (e.g. [KMO10a, JNS13, CLS15, CC17, TBS+16, ZZLC17]). The convergence guarantees for matrix completion have been established for various algorithms such as gradient descent on manifold [KMO10a, KMO10b], alternating minimization [JNS13, Har14], gradient descent [SL16, MWCC17, WZG16, CLL19], and projected gradient descent [CW15], provided that a suitable initialization (like spectral initialization) is available [KMO10a, JNS13, SL16, MWCC17, CCF18]. Our work is mostly related to [MWCC17, CLL19], which studied (vanilla) gradient descent for nonconvex matrix completion. This algorithm was first analyzed by [MWCC17] via a leave-one-out argument — a technique that proves useful in

analyzing various statistical algorithms [EK15,SCC17,ZB18,CFMW19,AFWZ17,LMCC18,DC18,CCFM19]. In the absence of noise and omitting logarithmic factors, [MWCC17] showed that $O(nr^3)$ samples are sufficient for vanilla GD to yield ε accuracy in $O(\log \frac{1}{\varepsilon})$ iterations (without the need of extra regularization procedures); the sample complexity was further improved to $O(nr^2)$ by [CLL19]. Apart from gradient descent, other nonconvex methods (e.g. [RS05, JMD10, WYZ12, JNS13, FRW11, Van13, LXY13, Har14, JKN16, RT$^+$11, WCCL16, DC18, GAGG13, CX16, ZWL15]) and landscape / geometry properties have been investigated [GLM16, CL17, PKCS17, GJZ17, SXZ19]; these are, however, beyond the scope of the current paper.

Another line of works asserted that a large family of SDPs admits low-rank solutions [Bar95], which in turn motivates the Burer-Monteiro approach [BM03, BVB16]. When applied to matrix completion, however, the generic theoretical guarantees therein lead to conservative results. Take the noiseless case (33) for instance: these results revealed the existence of a solution of rank at most $O(\sqrt{n^2p})$, which however is often much larger the true rank (e.g. when $r \asymp 1$ and $p \asymp \mathrm{poly}\log(n)/n$, one has $\sqrt{n^2p} \gg \sqrt{n} \gg r$). Moreover, this line of works does not imply that all solutions to the SDP of interest are (approximately) low-rank.

Finally, the connection between convex and nonconvex optimization has also been explored in line spectral estimation [LT18], although the context therein is drastically different from ours.

4 Discussion

This paper provides an improved statistical analysis for the natural convex program (3), without the need of enforcing additional spikiness constraint. Our theoretical analysis uncovers an intriguing connection between convex relaxation and nonconvex optimization, which we believe is applicable to many other problems beyond matrix completion. Having said that, our current theory leaves open a variety of important directions for future exploration. Here we sample a few interesting ones.

- *Improving dependency on r and κ.* While our theory is optimal when r and κ are both constants, it becomes increasingly looser as either r or κ grows. For instance, in the noiseless setting, it has been shown that the sample complexity for convex relaxation scales as $O(nr)$ — linear in r and independent of κ — which is better than the current results. It is worth noting that existing theory for nonconvex matrix factorization typically falls short of providing optimal scaling in r and κ [KMO10a, SL16, CW15, MWCC17, CLL19]. Thus, tightening the dependency of sample complexity on r and κ might call for new analysis tools.

- *Approximate low-rank structure.* So far our theory is built upon the assumption that the ground-truth matrix $\boldsymbol{M}^\star$ is exactly low-rank, which falls short of accommodating the more realistic scenario where $\boldsymbol{M}^\star$ is only approximately low-rank. For the approximate low-rank case, it is not yet clear whether the nonconvex factorization approach can still serve as a tight proxy. In addition, the landscape of nonconvex optimization for the approximately low-rank case [CL17] might shed light on how to handle this case.

- *Extension to deterministic noise.* Our current theory — in particular, the leave-one-out analysis for the nonconvex approach — relies heavily on the randomness assumption (i.e. i.i.d. sub-Gaussian) of the noise. In order to justify the broad applicability of convex relaxation, it would be interesting to see whether one can generalize the theory to cover deterministic noise with bounded magnitudes.

- *Extension to structured matrix completion.* Many applications involve low-rank matrices that exhibit additional structures, enabling a further reduction of the sample complexity [FHB03, CC14, CWW19]. For instance, if a matrix is Hankel and low-rank, then the sample complexity can be $O(n)$ times smaller than the generic low-rank case. The existing stability guarantee of Hankel matrix completion, however, is overly pessimistic compared to practical performance [CC14]. The analysis framework herein might be amenable to the study of Hankel matrix completion and help close the theory-practice gap.

- *Extension to robust PCA and blind deconvolution.* Moving beyond matrix completion, there are other problems that are concerned with recovering low-rank matrices. Notable examples include robust principal component analysis [CLMW11, CSPW11, CJSC13], blind deconvolution [ARR14, LS15] and blind demixing [LS17, JKS17]. The stability analyses of the convex relaxation approaches for these problems [ZLW$^+$10, ARR14, LS17] often adopt a similar approach as [CP10], and consequently are sub-optimal. The insights from the present paper might promise tighter statistical guarantees for such problems.

Finally, we remark that the intimate link between convex and nonconvex optimization enables statistically optimal inference and uncertainty quantification for noisy matrix completion (e.g. construction of optimal confidence intervals for each missing entry). The interested readers are referred to our companion paper [CFMY19] for in-depth discussions.

Acknowledgements

Y. Chen is supported in part by the AFOSR YIP award FA9550-19-1-0030, by the ARO grant W911NF-18-1-0303, by the ONR grant N00014-19-1-2120, by the NSF grants CCF-1907661 and IIS-1900140, and by the Princeton SEAS innovation award. Y. Chi is supported in part by ONR under the grants N00014-18-1-2142 and N00014-19-1-2404, by ARO under the grant W911NF-18-1-0303, and by NSF under the grants CAREER ECCS-1818571 and CCF-1806154. J. Fan is supported in part by NSF Grants DMS-1662139 and DMS-1712591, ONR grant N00014-19-1-2120, and NIH Grant R01-GM072611-12. This work was done in part while Y. Chen was visiting the Kavli Institute for Theoretical Physics (supported in part by NSF grant PHY-1748958). Y. Chen thanks Emmanuel Candès for motivating discussions about noisy matrix completion.

References

[AFWZ17] E. Abbe, J. Fan, K. Wang, and Y. Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

[ARR14] A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

[Bar95] A. I. Barvinok. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13(2):189–202, 1995.

[BM03] S. Burer and R. D. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[BN06] J. Bai and S. Ng. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150, 2006.

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[BVB16] N. Boumal, V. Voroninski, and A. Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *NIPS*, pages 2757–2765, 2016.

[CC14] Y. Chen and Y. Chi. Robust spectral compressed sensing via structured matrix completion. *IEEE Transactions on Information Theory*, 60(10):6576 – 6601, 2014.

[CC17] Y. Chen and E. 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] Y. Chen and E. Candès. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Comm. Pure and Appl. Math.*, 71(8):1648–1714, 2018.

[CC18b] Y. Chen and Y. Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31, July 2018.

[CCF18] Y. Chen, C. Cheng, and J. Fan. Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *arXiv preprint arXiv:1811.12804*, 2018.

[CCFM19] Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2):5–37, July 2019.

[CCS10] J. F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[CFMW19] Y. Chen, J. Fan, C. Ma, and K. Wang. Spectral method and regularized MLE are both optimal for top-K ranking. *Annals of Statistics*, 47(4):2204–2235, August 2019.

[CFMY19] Y. Chen, J. Fan, C. Ma, and Y. Yan. Inference and uncertainty quantification for noisy matrix completion. *arXiv preprint arXiv:1906.04159*, 2019.

[CGH14] Y. Chen, L. J. Guibas, and Q. Huang. Near-optimal joint optimal matching via convex relaxation. *International Conference on Machine Learning (ICML)*, pages 100 – 108, June 2014.

[Che15] Y. Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

[CJSC13] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7):4324–4337, 2013.

[CL17] J. Chen and X. Li. Memory-efficient kernel PCA via partial matrix sampling and nonconvex optimization: a model-free analysis of local minima. *arXiv:1711.01742*, 2017.

[CLC19] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239 – 5269, October 2019.

[CLL19] J. Chen, D. Liu, and X. Li. Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *arXiv:1901.06116v1*, 2019.

[CLMW11] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 58(3):11:1–11:37, Jun 2011.

[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.

[CP10] E. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925 –936, June 2010.

[CR09] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, April 2009.

[CSPW11] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[CT10] E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053 –2080, May 2010.

[CW15] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[CWW19] J.-F. Cai, T. Wang, and K. Wei. Fast and provable algorithms for spectrally sparse signal reconstruction via low-rank hankel matrix completion. *Applied and Computational Harmonic Analysis*, 46(1):94–121, 2019.

[CX16] Y. Cao and Y. Xie. Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing*, 64(6):1609–1620, 2016.

[CZ16] T. T. Cai and W.-X. Zhou. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1):1493–1525, 2016.

[DC18] L. Ding and Y. Chen. The leave-one-out approach for matrix completion: Primal and dual analysis. *arXiv preprint arXiv:1803.07554*, 2018.

[DR16] M. A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[Efr07] B. Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103, 2007.

[Efr10] B. Efron. Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association*, 105(491):1042–1055, 2010.

[EK15] N. El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, pages 1–81, 2015.

[Faz02] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, 2002.

[FHB03] M. Fazel, H. Hindi, and S. P. Boyd. Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. *American Control Conference*, 2003.

[FHB04] M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *American Control Conference*, volume 4, pages 3273–3278, 2004.

[FHG12] J. Fan, X. Han, and W. Gu. Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, 107(499):1019–1035, 2012.

[FKSZ19] J. Fan, Y. Ke, Q. Sun, and W.-X. Zhou. Farmtest: Factor-adjusted robust multiple testing with approximate false discovery control. *Journal of American Statistical Association*, 2019+.

[FKW18] J. Fan, Y. Ke, and K. Wang. Factor-adjusted regularized model selection. *arXiv preprint arXiv:1612.08490*, 2018.

[FLM13] J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 2013.

[FRW11] M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21(4):1614–1640, 2011.

[FWZ19] J. Fan, W. Wang, and Y. Zhong. Robust covariance estimation for approximate factor models. *Journal of econometrics*, 208(1):5–22, 2019.

[FXY17] J. Fan, L. Xue, and J. Yao. Sufficient forecasting using factor models. *Journal of econometrics*, 201(2):292–306, 2017.

[GAGG13] S. Gunasekar, A. Acharya, N. Gaur, and J. Ghosh. Noisy matrix completion using alternating minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 194–209, 2013.

[GJZ17] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

[GLM16] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[Gro11] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, March 2011.

[Har14] M. Hardt. Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS)*, pages 651–660, 2014.

[HG13] Q.-X. Huang and L. Guibas. Consistent shape maps via semidefinite programming. In *Computer Graphics Forum*, volume 32, pages 177–186. Wiley Online Library, 2013.

[JKN16] C. Jin, S. M. Kakade, and P. Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *NIPS*, pages 4520–4528, 2016.

[JKS17] P. Jung, F. Krahmer, and D. Stöger. Blind demixing and deconvolution at near-optimal rate. *IEEE Transactions on Information Theory*, 64(2):704–727, 2017.

[JMD10] P. Jain, R. Meka, and I. 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.

[Jol82] I. T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303, 1982.

[Klo14] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[KLT11] V. Koltchinskii, K. Lounici, and A. 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] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, 2010.

[KS11] A. Kneip and P. Sarda. Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics*, 39(5):2410–2447, 2011.

[Lan93] S. Lang. Real and functional analysis. *Springer-Verlag, New York,*, 10:11–13, 1993.

[LMCC18] Y. Li, C. Ma, Y. Chen, and Y. Chi. Nonconvex matrix factorization from rank-one measurements. *arXiv:1802.06286, accepted to AISTATS*, 2018.

[LS15] S. Ling and T. Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015.

[LS17] S. Ling and T. Strohmer. Blind deconvolution meets blind demixing: Algorithms and performance bounds. *IEEE Transactions on Information Theory*, 63(7):4497–4520, 2017.

[LT18] Q. Li and G. Tang. Approximate support recovery of atomic line spectral estimation: A tale of resolution and precision. *Applied and Computational Harmonic Analysis*, 2018.

[LV09] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.

[LXY13] M.-J. 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.

[MGC11] S. Ma, D. Goldfarb, and L. Chen. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1-2):321–353, 2011.

[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.

[MWCC17] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467, accepted to Foundations of Computational Mathematics*, 2017.

[Nes12] Y. Nesterov. How to make the gradients small. *Optima*, 88:10–11, 2012.

[NW12] S. Negahban and M. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, pages 1665–1697, May 2012.

[PB14] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[PBHT08] D. Paul, E. Bair, T. Hastie, and R. Tibshirani. "Preconditioning" for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, 36(4):1595–1618, 2008.

[PKCS17] D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74, 2017.

[Rec11] B. 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] J. D. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *International conference on Machine learning*, pages 713–719. ACM, 2005.

[RT+11] A. Rohde, A. B. Tsybakov, et al. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

[SCC17] P. Sur, Y. Chen, and E. J. Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *arXiv:1706.01191, accepted to Probability Theory and Related Fields*, 2017.

[Sin11] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20–36, 2011.

[SL16] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[SS05] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005.

[SXZ19] A. Shapiro, Y. Xie, and R. Zhang. Matrix completion with deterministic pattern: A geometric perspective. *IEEE Transactions on Signal Processing*, 67(4):1088–1103, 2019.

[SY07] A. M.-C. So and Y. Ye. Theory of semidefinite programming for sensor network localization. *Mathematical Programming*, 109(2-3):367–384, 2007.

[TBS+16] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973, 2016.

[TK92] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International journal of computer vision*, 9(2):137–154, 1992.

[Tro15] J. A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230, May 2015.

[TY10] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of optimization*, 6(615-640):15, 2010.

[Van13] B. Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

[Ver12] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications*, pages 210 – 268, 2012.

[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.

[WYZ12] Z. Wen, W. Yin, and Y. 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] L. Wang, X. Zhang, and Q. Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. *arXiv preprint arXiv:1610.05275*, 2016.

[YPCC16] X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust PCA via gradient descent. In *NIPS*, pages 4152–4160, 2016.

[ZB18] Y. Zhong and N. Boumal. Near-optimal bound for phase synchronization. *SIAM Journal on Optimization*, 2018.

[ZL16] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv:1605.07051*, 2016.

[ZLW+10] Z. Zhou, X. Li, J. Wright, E. Candès, and Y. Ma. Stable principal component pursuit. In *International Symposium on Information Theory*, pages 1518–1522, 2010.

[ZPL15] T. Zhang, J. M. Pauly, and I. R. Levesque. Accelerating parameter mapping with a locally low rank constraint. *Magnetic resonance in medicine*, 73(2):655–661, 2015.

[ZWL15] T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In *NIPS*, pages 559–567, 2015.

[ZZLC17] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi. A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms. *The Journal of Machine Learning Research*, 18(1):5164–5198, 2017.

A Preliminaries

In this section, we gather a few notations and preliminary facts that are used throughout the proofs.

To begin with, in view of the incoherence assumption (cf. Definition 1), one has

$$\|\boldsymbol{X}^\star\|_{2,\infty} \le \sqrt{\mu r/n}\,\|\boldsymbol{X}^\star\| \qquad \text{and} \qquad \|\boldsymbol{Y}^\star\|_{2,\infty} \le \sqrt{\mu r/n}\,\|\boldsymbol{Y}^\star\|. \tag{34}$$

This follows from

$$\|\boldsymbol{X}^\star\|_{2,\infty} = \left\|\boldsymbol{U}^\star\left(\boldsymbol{\Sigma}^\star\right)^{1/2}\right\|_{2,\infty} \le \|\boldsymbol{U}^\star\|_{2,\infty}\left\|\left(\boldsymbol{\Sigma}^\star\right)^{1/2}\right\| \le \sqrt{\mu r/n}\,\|\boldsymbol{X}^\star\|.$$

The bound for $\boldsymbol{Y}^\star$ follows from the same argument. In addition, we write $A \ll B$ (resp. $A \gg B$) if there exists a sufficiently small (resp. large) constant c such that $A \le cB$ (resp. $A \ge cB$).

Finally, for notational convenience, we shall often denote

$$\mathcal{P}_\Omega^{\mathsf{debias}}(\boldsymbol{B}) \triangleq \mathcal{P}_\Omega(\boldsymbol{B}) - p\boldsymbol{B}, \qquad \text{for all } \boldsymbol{B} \in \mathbb{R}^{n\times n}. \tag{35}$$

B Exact duality analysis

We show in this section that why the first-order optimality condition is almost sufficient in guaranteeing the uniqueness of the optimizer. The argument is standard, see e.g. [CR09].

Lemma 6. *Let $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the SVD of $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$. Denote by T be the tangent space of $\boldsymbol{Z}$ and by $T^\perp$ its orthogonal complement. Suppose that there exists $\boldsymbol{W} \in T^\perp$ such that*

$$\frac{1}{\lambda}\mathcal{P}_\Omega(\boldsymbol{M}-\boldsymbol{Z}) = \boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W}. \tag{36}$$

Then $\boldsymbol{Z}$ is the unique minimizer of (3) if

1. *$\|\boldsymbol{W}\| < 1$;*
2. *The operator $\mathcal{P}_\Omega(\cdot)$ restricted to elements in T is injective, i.e. $\mathcal{P}_\Omega(\boldsymbol{H}) = \boldsymbol{0}$ implies $\boldsymbol{H} = \boldsymbol{0}$ for any $\boldsymbol{H} \in T$.*

Proof of Lemma 6. To begin with, the assumption of this lemma implies that

$$\boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W} \in \partial\|\boldsymbol{Z}\|_*,$$

where $\partial\|\boldsymbol{Z}\|_*$ denotes the subdifferential of $\|\cdot\|_*$ at $\boldsymbol{Z}$. This combined with (36) reveals that

$$\frac{1}{\lambda}\mathcal{P}_\Omega(\boldsymbol{M}-\boldsymbol{Z}) \in \partial\|\boldsymbol{Z}\|_*, \tag{37}$$

thus indicating that $\boldsymbol{Z}$ is a minimizer of the convex program (3).

Next, we justify the uniqueness of $\boldsymbol{Z}$. Before continuing, we record a fact regarding the minimizers of (3).

Claim 1. *Suppose that $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are both minimizers of (3). Then one has $\mathcal{P}_\Omega(\boldsymbol{Z}_1) = \mathcal{P}_\Omega(\boldsymbol{Z}_2)$.*

With this claim at hand, every minimizer of (3) can be written as $\boldsymbol{Z}+\boldsymbol{H}$ for some $\boldsymbol{H}$ obeying $\mathcal{P}_\Omega(\boldsymbol{H}) = \boldsymbol{0}$. It then suffices to prove that for any $\boldsymbol{H} \neq \boldsymbol{0}$, one has $g(\boldsymbol{Z}+\boldsymbol{H}) > g(\boldsymbol{Z})$, where $g(\cdot)$ is the objective function in (3). To this end, we note that

$$\begin{aligned} g(\boldsymbol{Z}+\boldsymbol{H}) &= \tfrac{1}{2}\|\mathcal{P}_\Omega(\boldsymbol{Z}+\boldsymbol{H}-\boldsymbol{M})\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{Z}+\boldsymbol{H}\|_* \\ &= \tfrac{1}{2}\|\mathcal{P}_\Omega(\boldsymbol{Z}-\boldsymbol{M})\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{Z}+\boldsymbol{H}\|_*, \end{aligned} \tag{38}$$

where the last relation follows from Claim 1 (i.e. $\mathcal{P}_\Omega(\boldsymbol{H}) = \boldsymbol{0}$). Let $\boldsymbol{S}$ be a subgradient of $\|\cdot\|_*$ at point $\boldsymbol{Z}$ obeying

$$\mathcal{P}_T(\boldsymbol{S}) = \boldsymbol{U}\boldsymbol{V}^\top, \quad \|\mathcal{P}_{T^\perp}(\boldsymbol{S})\| \le 1 \quad \text{and} \quad \langle\mathcal{P}_{T^\perp}(\boldsymbol{S}), \mathcal{P}_{T^\perp}(\boldsymbol{H})\rangle = \|\mathcal{P}_{T^\perp}(\boldsymbol{H})\|_*. \tag{39}$$

Using the convexity of $\|\cdot\|_*$, one can further lower bound (38) by

$$\begin{aligned} g(\boldsymbol{Z}+\boldsymbol{H}) &\geq \tfrac{1}{2}\left\|\mathcal{P}_\Omega(\boldsymbol{Z}-\boldsymbol{M})\right\|_{\mathrm{F}}^2+\lambda\left(\|\boldsymbol{Z}\|_*+\langle \boldsymbol{S},\boldsymbol{H}\rangle\right) \\ &= g(\boldsymbol{Z})+\lambda\langle \boldsymbol{S},\boldsymbol{H}\rangle \\ &= g(\boldsymbol{Z})+\lambda\left\langle \boldsymbol{U}\boldsymbol{V}^\top+\boldsymbol{W},\boldsymbol{H}\right\rangle+\lambda\left\langle \boldsymbol{S}-\boldsymbol{U}\boldsymbol{V}^\top-\boldsymbol{W},\boldsymbol{H}\right\rangle \\ &\overset{\text{(i)}}{=} g(\boldsymbol{Z})+\lambda\left\langle \boldsymbol{S}-\boldsymbol{U}\boldsymbol{V}^\top-\boldsymbol{W},\boldsymbol{H}\right\rangle \\ &\overset{\text{(ii)}}{=} g(\boldsymbol{Z})+\lambda\left\langle \mathcal{P}_{T^\perp}(\boldsymbol{S})-\boldsymbol{W},\boldsymbol{H}\right\rangle. \end{aligned}$$

Here, (i) follows from our assumption that $\boldsymbol{U}\boldsymbol{V}^\top+\boldsymbol{W}$ is supported on Ω (cf. (36)) and the fact that $\mathcal{P}_\Omega(\boldsymbol{H})=\boldsymbol{0}$, and (ii) holds since $\mathcal{P}_T(\boldsymbol{S})=\boldsymbol{U}\boldsymbol{V}^\top$ (cf. (39)). We can now expand the above expression as

$$\begin{aligned} g(\boldsymbol{Z}+\boldsymbol{H}) &\geq g(\boldsymbol{Z})+\lambda\left\langle \mathcal{P}_{T^\perp}(\boldsymbol{S}),\mathcal{P}_{T^\perp}(\boldsymbol{H})\right\rangle-\lambda\left\langle \boldsymbol{W},\mathcal{P}_{T^\perp}(\boldsymbol{H})\right\rangle \\ &\geq g(\boldsymbol{Z})+\lambda\left(1-\|\boldsymbol{W}\|\right)\left\|\mathcal{P}_{T^\perp}(\boldsymbol{H})\right\|_*, \end{aligned} \tag{40}$$

where the last inequality holds by using the last property of (39) and invoking the elementary inequality

$$\left\langle \boldsymbol{W},\mathcal{P}_{T^\perp}(\boldsymbol{H})\right\rangle\leq\|\boldsymbol{W}\|\left\|\mathcal{P}_{T^\perp}(\boldsymbol{H})\right\|_*.$$

Given that $\boldsymbol{W}$ is assumed to obey $\|\boldsymbol{W}\|<1$, one has $g(\boldsymbol{Z}+\boldsymbol{H})>g(\boldsymbol{Z})$ unless $\mathcal{P}_{T^\perp}(\boldsymbol{H})=\boldsymbol{0}$. However, if $\mathcal{P}_{T^\perp}(\boldsymbol{H})=\boldsymbol{0}$ (and hence $\boldsymbol{H}\in T$), then the injectivity assumption together with the fact that $\mathcal{P}_\Omega(\boldsymbol{H})=\boldsymbol{0}$ forces $\boldsymbol{H}=\boldsymbol{0}$. Consequently, any minimizer $\boldsymbol{Z}+\boldsymbol{H}$ with $\boldsymbol{H}\neq\boldsymbol{0}$ must satisfy $g(\boldsymbol{Z}+\boldsymbol{H})>g(\boldsymbol{Z})$, which results in contradiction. This concludes the proof. □

Proof of Claim 1. Consider any minimizers $\boldsymbol{Z}_1\neq\boldsymbol{Z}_2$, and suppose instead that $\mathcal{P}_\Omega(\boldsymbol{Z}_1-\boldsymbol{Z}_2)\neq\boldsymbol{0}$. For any $0<\alpha<1$, define

$$\boldsymbol{Z}_\alpha\triangleq\alpha\boldsymbol{Z}_1+(1-\alpha)\boldsymbol{Z}_2.$$

Since $\|\cdot\|_*$ is convex, we have

$$\begin{aligned} g(\boldsymbol{Z}_\alpha) &= \tfrac{1}{2}\left\|\mathcal{P}_\Omega(\alpha\boldsymbol{Z}_1+(1-\alpha)\boldsymbol{Z}_2-\boldsymbol{M})\right\|_{\mathrm{F}}^2+\lambda\left\|\alpha\boldsymbol{Z}_1+(1-\alpha)\boldsymbol{Z}_2\right\|_* \\ &\leq \tfrac{1}{2}\left\|\mathcal{P}_\Omega(\alpha\boldsymbol{Z}_1+(1-\alpha)\boldsymbol{Z}_2-\boldsymbol{M})\right\|_{\mathrm{F}}^2+\alpha\lambda\left\|\boldsymbol{Z}_1\right\|_*+(1-\alpha)\lambda\left\|\boldsymbol{Z}_2\right\|_*. \end{aligned} \tag{41}$$

Furthermore, by the strong convexity of $\|\cdot\|_{\mathrm{F}}^2$ we have

$$\begin{aligned} g(\boldsymbol{Z}_\alpha) &< \tfrac{1}{2}\big(\alpha\left\|\mathcal{P}_\Omega(\boldsymbol{Z}_1-\boldsymbol{M})\right\|_{\mathrm{F}}^2+(1-\alpha)\left\|\mathcal{P}_\Omega(\boldsymbol{Z}_2-\boldsymbol{M})\right\|_{\mathrm{F}}^2\big)+\alpha\lambda\left\|\boldsymbol{Z}_1\right\|_*+(1-\alpha)\lambda\left\|\boldsymbol{Z}_2\right\|_* \\ &= \alpha g(\boldsymbol{Z}_1)+(1-\alpha)g(\boldsymbol{Z}_2)=g(\boldsymbol{Z}_1). \end{aligned}$$

This contradicts the fact that $\boldsymbol{Z}_1$ is a minimizer of (3), thus completing the proof. □

C Connections between convex and nonconvex solutions

C.1 Proof of Lemma 1

First of all, since $(\boldsymbol{X},\boldsymbol{Y})$ is a stationary point of (5), we have the first-order optimality conditions

$$\mathcal{P}_\Omega\left(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^\top\right)\boldsymbol{Y}=\lambda\boldsymbol{X}; \tag{42a}$$

$$\left[\mathcal{P}_\Omega\left(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^\top\right)\right]^\top\boldsymbol{X}=\lambda\boldsymbol{Y}. \tag{42b}$$

As an immediate consequence, one has

$$\boldsymbol{X}^\top\boldsymbol{X}=\lambda^{-1}\boldsymbol{X}^\top\mathcal{P}_\Omega\left(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^\top\right)\boldsymbol{Y}=\boldsymbol{Y}^\top\boldsymbol{Y}. \tag{43}$$

In words, any stationary point $(\boldsymbol{X},\boldsymbol{Y})$ has "balanced" scale.

Let $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the singular value decomposition of $\boldsymbol{X}\boldsymbol{Y}^\top$ with $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n\times r}$ orthonormal and $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$ diagonal. In view of the balanced scale of $(\boldsymbol{X}, \boldsymbol{Y})$ (namely, (43)) and Lemma 20, we can write

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{R} \qquad \text{and} \qquad \boldsymbol{Y} = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\boldsymbol{R} \tag{44}$$

for some orthonormal matrix $\boldsymbol{R} \in \mathbb{R}^{r\times r}$. Substitution into (42) results in

$$\mathcal{P}_\Omega\left(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top\right)\boldsymbol{V} = \lambda \boldsymbol{U}; \tag{45a}$$

$$\left[\mathcal{P}_\Omega\left(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top\right)\right]^\top \boldsymbol{U} = \lambda \boldsymbol{V}, \tag{45b}$$

implying that the columns of $\boldsymbol{U}$ (resp. $\boldsymbol{V}$) are the left (resp. right) singular vectors of the matrix $\mathcal{P}_\Omega(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top)$. We can therefore write

$$\frac{1}{\lambda}\mathcal{P}_\Omega\left(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top\right) = \boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W}, \tag{46}$$

where $\boldsymbol{W} \in T^\perp$; recall that T is the tangent space of $\boldsymbol{X}\boldsymbol{Y}^\top$ and also $\boldsymbol{U}\boldsymbol{V}^\top$. In view of Lemma 6, it suffices to show that $\|\boldsymbol{W}\| < 1$, which is the content of the rest of the proof.

One can rewrite $\mathcal{P}_\Omega(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top)$ as

$$\mathcal{P}_\Omega\left(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top\right) = p\left(\boldsymbol{M}^\star - \boldsymbol{X}\boldsymbol{Y}^\top\right) + \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{M}^\star - \boldsymbol{X}\boldsymbol{Y}^\top\right) + \mathcal{P}_\Omega\left(\boldsymbol{E}\right).$$

Substitute this identity into (45) and rearrange terms to obtain

$$\left[p\boldsymbol{M}^\star + \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{M}^\star - \boldsymbol{X}\boldsymbol{Y}^\top\right) + \mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right]\boldsymbol{V} = \boldsymbol{U}\left(p\boldsymbol{\Sigma} + \lambda\boldsymbol{I}_r\right);$$

$$\left[p\boldsymbol{M}^\star + \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{M}^\star - \boldsymbol{X}\boldsymbol{Y}^\top\right) + \mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right]^\top \boldsymbol{U} = \boldsymbol{V}\left(p\boldsymbol{\Sigma} + \lambda\boldsymbol{I}_r\right).$$

These tell us that the columns of $\boldsymbol{U}$ (resp. $\boldsymbol{V}$) are the left (resp. right) singular vectors of the matrix

$$p\boldsymbol{M}^\star + \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{M}^\star - \boldsymbol{X}\boldsymbol{Y}^\top\right) + \mathcal{P}_\Omega\left(\boldsymbol{E}\right),$$

which is equivalent to saying that[6]

$$p\boldsymbol{M}^\star + \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{M}^\star - \boldsymbol{X}\boldsymbol{Y}^\top\right) + \mathcal{P}_\Omega\left(\boldsymbol{E}\right) = \boldsymbol{U}\left(p\boldsymbol{\Sigma} + \lambda\boldsymbol{I}_r\right)\boldsymbol{V}^\top + \lambda\boldsymbol{W}_2, \tag{47}$$

for some $\boldsymbol{W}_2 \in T^\perp$. One can then derive from (46) that

$$\begin{aligned}
\boldsymbol{W} &\overset{\text{(i)}}{=} \tfrac{1}{\lambda}\mathcal{P}_{T^\perp}\left[\mathcal{P}_\Omega\left(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top\right)\right] \\
&= \tfrac{1}{\lambda}\mathcal{P}_{T^\perp}\left[p\boldsymbol{M}^\star - p\boldsymbol{X}\boldsymbol{Y}^\top + \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{M}^\star - \boldsymbol{X}\boldsymbol{Y}^\top\right) + \mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right] \\
&\overset{\text{(ii)}}{=} \tfrac{1}{\lambda}\mathcal{P}_{T^\perp}\left[p\boldsymbol{M}^\star + \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{M}^\star - \boldsymbol{X}\boldsymbol{Y}^\top\right) + \mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right] \\
&\overset{\text{(iii)}}{=} \tfrac{1}{\lambda}\mathcal{P}_{T^\perp}\left[\boldsymbol{U}\left(p\boldsymbol{\Sigma} + \lambda\boldsymbol{I}_r\right)\boldsymbol{V}^\top + \lambda\boldsymbol{W}_2\right] \\
&\overset{\text{(iv)}}{=} \boldsymbol{W}_2,
\end{aligned}$$

where (i), (ii) and (iv) arise from the facts that $\boldsymbol{U}\boldsymbol{V}^\top \in T$, $\boldsymbol{X}\boldsymbol{Y}^\top \in T$ and $\boldsymbol{U}(p\boldsymbol{\Sigma}+\lambda\boldsymbol{I}_r)\boldsymbol{V}^\top \in T$, respectively, and (iii) relies on the identity (47).

It then suffices to control $\|\boldsymbol{W}_2\|$. To this end, apply Weyl's inequality to (47) to obtain that: for $r+1 \le i \le n$, the ith largest singular value of $\boldsymbol{U}(p\boldsymbol{\Sigma} + \lambda\boldsymbol{I}_r)\boldsymbol{V}^\top + \lambda\boldsymbol{W}_2$ obeys

$$\begin{aligned}
\sigma_i\left(\boldsymbol{U}\left(p\boldsymbol{\Sigma} + \lambda\boldsymbol{I}_r\right)\boldsymbol{V}^\top + \lambda\boldsymbol{W}_2\right) &\le p\sigma_i\left(\boldsymbol{M}^\star\right) + \left\|\mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{M}^\star - \boldsymbol{X}\boldsymbol{Y}^\top\right) + \mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right\| \\
&\le \left\|\mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{M}^\star - \boldsymbol{X}\boldsymbol{Y}^\top\right)\right\| + \left\|\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right\| \\
&< \lambda,
\end{aligned}$$

where the second inequality comes from the fact that $\boldsymbol{M}^\star$ has rank r (so that $\sigma_i(\boldsymbol{M}^\star) = 0$ for $r+1 \le i \le n$) as well as the triangle inequality, and the last inequality follows from the assumptions of the lemma. Furthermore, it is seen that $\boldsymbol{U}(p\boldsymbol{\Sigma} + \lambda\boldsymbol{I}_r)\boldsymbol{V}^\top$ has rank r and all of its singular values are at least λ. These facts taken collectively demonstrate that

$$\|\boldsymbol{W}\| = \|\boldsymbol{W}_2\| = \tfrac{1}{\lambda}\max_{r<i\le n}\sigma_i\left(\boldsymbol{U}\left(p\boldsymbol{\Sigma} + \lambda\boldsymbol{I}_r\right)\boldsymbol{V}^\top + \lambda\boldsymbol{W}_2\right) < 1.$$

This together with Lemma 6 completes the proof.

[6]Here, the pre-factor λ is chosen to simplify the analysis later on.

C.2 Proof of Lemma 2

We begin by collecting a few simple properties resulting from our assumptions. By definition, the gradient of $f(\cdot,\cdot)$ in (17) is given by

$$\nabla f(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{p}\begin{bmatrix} \mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\boldsymbol{Y} + \lambda\boldsymbol{X} \\ \left[\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right]^\top \boldsymbol{X} + \lambda\boldsymbol{Y} \end{bmatrix},$$

which together with the small-gradient assumption $\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\|_{\mathrm{F}} \le c\lambda\sqrt{c_{\mathrm{inj}}\, p\sigma_{\min}/\kappa^2}/p$ implies that

$$\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\boldsymbol{Y} + \lambda\boldsymbol{X}\right\|_{\mathrm{F}} \le p\left\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\right\|_{\mathrm{F}} \le c\lambda\sqrt{c_{\mathrm{inj}}\, p\sigma_{\min}/\kappa^2}; \tag{48a}$$

$$\left\|\left(\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M})\right)^\top \boldsymbol{X} + \lambda\boldsymbol{Y}\right\|_{\mathrm{F}} \le p\left\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\right\|_{\mathrm{F}} \le c\lambda\sqrt{c_{\mathrm{inj}}\, p\sigma_{\min}/\kappa^2}. \tag{48b}$$

Throughout the proof, we let the SVD of $\boldsymbol{X}\boldsymbol{Y}^\top$ be $\boldsymbol{X}\boldsymbol{Y}^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, and denote by T the tangent space of $\boldsymbol{X}\boldsymbol{Y}^\top$ and by $T^\perp$ its orthogonal complement. Additionally, our assumption regarding the singular values of $\boldsymbol{X}$ and $\boldsymbol{Y}$ implies that

$$\sigma_{\min}/2 \le \sigma_{\min}(\boldsymbol{\Sigma}) \le \sigma_{\max}(\boldsymbol{\Sigma}) \le 2\sigma_{\max}. \tag{49}$$

This can be easily seen from the following two inequalities

$$\begin{aligned}
\sigma_{\max}(\boldsymbol{\Sigma}) &= \left\|\boldsymbol{X}\boldsymbol{Y}^\top\right\| \le \|\boldsymbol{X}\|\,\|\boldsymbol{Y}\| \le 2\sigma_{\max}; \\
\sigma_{\min}(\boldsymbol{\Sigma}) &= \sigma_{\min}\left(\boldsymbol{X}\boldsymbol{Y}^\top\right) \ge \sigma_{\min}(\boldsymbol{X})\,\sigma_{\min}(\boldsymbol{Y}) \ge \sigma_{\min}/2.
\end{aligned}$$

Before proceeding, we record a claim that will prove useful in the subsequent analysis.

Claim 2. *Under the notations and assumptions of Lemma 2, one has*

$$\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right) = -\lambda\boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{R}, \tag{50}$$

where $\boldsymbol{R}$ is some residual matrix satisfying

$$\|\mathcal{P}_T(\boldsymbol{R})\|_{\mathrm{F}} \le 72\kappa\frac{p}{\sqrt{\sigma_{\min}}}\left\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\right\|_{\mathrm{F}} \quad \text{and} \quad \|\mathcal{P}_{T^\perp}(\boldsymbol{R})\| < \lambda/2. \tag{51}$$

With Claim 2 in place, we are ready to prove Lemma 2. Let $\boldsymbol{Z}_{\mathsf{cvx}}$ be any minimizer of (3) and denote $\boldsymbol{\Delta} \triangleq \boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{X}\boldsymbol{Y}^\top$. The proof can be divided into the following steps.

- First, show that the difference $\boldsymbol{\Delta}$ primarily lies in the tangent space of $\boldsymbol{X}\boldsymbol{Y}^\top$; see (58).
- Next, utilize this property to connect $\|\mathcal{P}_\Omega(\boldsymbol{\Delta})\|_{\mathrm{F}}^2$ with the size of the gradient $\nabla f(\boldsymbol{X},\boldsymbol{Y})$; see (60).
- In the end, obtain a lower bound on $\|\mathcal{P}_\Omega(\boldsymbol{\Delta})\|_{\mathrm{F}}^2$ in terms of $\|\boldsymbol{\Delta}\|_{\mathrm{F}}$ using the injectivity property; see (61).

The desired upper bound on $\|\boldsymbol{\Delta}\|_{\mathrm{F}}$ advertised in the lemma then follows by combining these results. In what follows, we shall carry out these steps one by one.

1. The optimality of $\boldsymbol{Z}_{\mathsf{cvx}} = \boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{\Delta}$ reveals that

$$\tfrac{1}{2}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{\Delta} - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \lambda\left\|\boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{\Delta}\right\|_* \le \tfrac{1}{2}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \lambda\left\|\boldsymbol{X}\boldsymbol{Y}^\top\right\|_*.$$

   A little algebra allows us to rearrange terms as follows

$$\tfrac{1}{2}\left\|\mathcal{P}_\Omega(\boldsymbol{\Delta})\right\|_{\mathrm{F}}^2 \le -\left\langle\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right), \boldsymbol{\Delta}\right\rangle + \lambda\left\|\boldsymbol{X}\boldsymbol{Y}^\top\right\|_* - \lambda\left\|\boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{\Delta}\right\|_*. \tag{52}$$

   In addition, it follows from the convexity of $\|\cdot\|_*$ that

$$\left\|\boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{\Delta}\right\|_* \ge \left\|\boldsymbol{X}\boldsymbol{Y}^\top\right\|_* + \left\langle\boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W}, \boldsymbol{\Delta}\right\rangle \tag{53}$$

for any $\boldsymbol{W} \in T^{\perp}$ obeying $\|\boldsymbol{W}\| \leq 1$, where $\boldsymbol{U}\boldsymbol{V}^{\top} + \boldsymbol{W}$ serves as a subgradient of $\|\cdot\|_*$ at $\boldsymbol{X}\boldsymbol{Y}^{\top}$. In what follows, we shall pick $\boldsymbol{W}$ such that $\langle \boldsymbol{W}, \boldsymbol{\Delta} \rangle = \|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\|_*$. Combining this with (52) and (53), we reach

$$\begin{aligned} \tfrac{1}{2}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}}^2 &\leq -\left\langle \mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M}\right), \boldsymbol{\Delta}\right\rangle - \lambda\left\langle \boldsymbol{U}\boldsymbol{V}^{\top}, \boldsymbol{\Delta}\right\rangle - \lambda\left\langle \boldsymbol{W}, \boldsymbol{\Delta}\right\rangle \\ &= -\left\langle \mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M}\right), \boldsymbol{\Delta}\right\rangle - \lambda\left\langle \boldsymbol{U}\boldsymbol{V}^{\top}, \boldsymbol{\Delta}\right\rangle - \lambda\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_*. \end{aligned} \tag{54}$$

This together with the decomposition (50) leads to

$$\begin{aligned} 0 \leq \tfrac{1}{2}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}}^2 &\leq -\left\langle \boldsymbol{R}, \boldsymbol{\Delta}\right\rangle - \lambda\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \\ &= -\left\langle \mathcal{P}_T(\boldsymbol{R}), \boldsymbol{\Delta}\right\rangle - \left\langle \mathcal{P}_{T^{\perp}}(\boldsymbol{R}), \boldsymbol{\Delta}\right\rangle - \lambda\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_*, \end{aligned} \tag{55}$$

and therefore

$$\left\langle \mathcal{P}_T(\boldsymbol{R}), \boldsymbol{\Delta}\right\rangle + \left\langle \mathcal{P}_{T^{\perp}}(\boldsymbol{R}), \boldsymbol{\Delta}\right\rangle + \lambda\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \leq 0. \tag{56}$$

In addition, elementary inequalities give

$$\begin{aligned} &-\left\|\mathcal{P}_T(\boldsymbol{R})\right\|_{\mathrm{F}}\left\|\mathcal{P}_T(\boldsymbol{\Delta})\right\|_{\mathrm{F}} - \left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{R})\right\|\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_* + \lambda\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_* \\ &\quad\leq \left\langle \mathcal{P}_T(\boldsymbol{R}), \boldsymbol{\Delta}\right\rangle + \left\langle \mathcal{P}_{T^{\perp}}(\boldsymbol{R}), \boldsymbol{\Delta}\right\rangle + \lambda\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \leq 0. \end{aligned}$$

From the condition (51) we have $\|\mathcal{P}_{T^{\perp}}(\boldsymbol{R})\| \leq \lambda/2$, and hence the above inequality gives

$$\left\|\mathcal{P}_T(\boldsymbol{R})\right\|_{\mathrm{F}}\left\|\mathcal{P}_T(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \geq -\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{R})\right\|\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_* + \lambda\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_* \geq \tfrac{\lambda}{2}\left\|\mathcal{P}_{T\perp}(\boldsymbol{\Delta})\right\|_*, \tag{57}$$

which together with the condition (51) on $\|\mathcal{P}_T(\boldsymbol{R})\|_{\mathrm{F}}$ and the small gradient assumption (23) yields

$$\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_* \leq 144\kappa\frac{p}{\lambda\sqrt{\sigma_{\min}}}\left\|\nabla f\left(\boldsymbol{X}, \boldsymbol{Y}\right)\right\|_{\mathrm{F}}\left\|\mathcal{P}_T(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \leq 144c\sqrt{c_{\mathrm{inj}}p}\left\|\mathcal{P}_T(\boldsymbol{\Delta})\right\|_{\mathrm{F}}. \tag{58}$$

This essentially means that $\boldsymbol{\Delta}$ lies primarily in the tangent space of $\boldsymbol{X}\boldsymbol{Y}^{\top}$ for c sufficiently small. As an immediate consequence,

$$\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \leq \left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_* \leq 144c\sqrt{c_{\mathrm{inj}}p}\left\|\mathcal{P}_T(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \leq \left\|\mathcal{P}_T(\boldsymbol{\Delta})\right\|_{\mathrm{F}}, \tag{59}$$

as long as c is sufficiently small. Note that we also use the elementary fact that $c_{\mathrm{inj}} \leq 1/p$ (otherwise we will have the contradictory inequality $p^{-1}\|\mathcal{P}_{\Omega}(\boldsymbol{H})\|_{\mathrm{F}}^2 \geq c_{\mathrm{inj}}\|\boldsymbol{H}\|_{\mathrm{F}}^2 > p^{-1}\|\boldsymbol{H}\|_{\mathrm{F}}^2$).

2. Continue the upper bound in (55) to obtain

$$\begin{aligned} \tfrac{1}{2}\left\|\mathcal{P}_{\Omega}(\boldsymbol{\Delta})\right\|_{\mathrm{F}}^2 &\leq -\left\langle \mathcal{P}_T(\boldsymbol{R}), \boldsymbol{\Delta}\right\rangle - \left\langle \mathcal{P}_{T^{\perp}}(\boldsymbol{R}), \boldsymbol{\Delta}\right\rangle - \lambda\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \\ &\leq \left\|\mathcal{P}_T(\boldsymbol{R})\right\|_{\mathrm{F}}\left\|\mathcal{P}_T(\boldsymbol{\Delta})\right\|_{\mathrm{F}} - \tfrac{\lambda}{2}\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_*. \end{aligned}$$

Here, the last line uses the fact that $-\left\langle \mathcal{P}_{T^{\perp}}(\boldsymbol{R}), \boldsymbol{\Delta}\right\rangle \leq \|\mathcal{P}_{T^{\perp}}(\boldsymbol{R})\| \cdot \|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\|_* \leq \frac{\lambda}{2}\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\|_*$, which follows from (51). Therefore, using the condition (51) we reach

$$\frac{1}{2}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}}^2 \leq \left\|\mathcal{P}_T(\boldsymbol{R})\right\|_{\mathrm{F}}\left\|\mathcal{P}_T(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \leq 72\kappa\frac{p}{\sqrt{\sigma_{\min}}}\left\|\nabla f\left(\boldsymbol{X}, \boldsymbol{Y}\right)\right\|_{\mathrm{F}}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}. \tag{60}$$

3. We are left with lower bounding $\|\mathcal{P}_{\Omega}(\boldsymbol{\Delta})\|_{\mathrm{F}}^2$. Using the decomposition $\boldsymbol{\Delta} = \mathcal{P}_T(\boldsymbol{\Delta}) + \mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})$, we obtain

$$\begin{aligned} \tfrac{1}{\sqrt{p}}\left\|\mathcal{P}_{\Omega}(\boldsymbol{\Delta})\right\|_{\mathrm{F}} &= \tfrac{1}{\sqrt{p}}\left\|\mathcal{P}_{\Omega}\mathcal{P}_T(\boldsymbol{\Delta}) + \mathcal{P}_{\Omega}\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \geq \tfrac{1}{\sqrt{p}}\left\|\mathcal{P}_{\Omega}\mathcal{P}_T(\boldsymbol{\Delta})\right\|_{\mathrm{F}} - \tfrac{1}{\sqrt{p}}\left\|\mathcal{P}_{\Omega}\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \\ &\geq \sqrt{c_{\mathrm{inj}}}\left\|\mathcal{P}_T(\boldsymbol{\Delta})\right\|_{\mathrm{F}} - \tfrac{1}{\sqrt{p}}\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_{\mathrm{F}}, \end{aligned}$$

where the last inequality follows from the injectivity assumption (22). In addition, (58) implies

$$\tfrac{1}{\sqrt{p}}\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \leq \tfrac{1}{\sqrt{p}}\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_* \leq \tfrac{1}{\sqrt{p}}144c\sqrt{c_{\mathrm{inj}}p}\left\|\mathcal{P}_T(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \leq \tfrac{\sqrt{c_{\mathrm{inj}}}}{2}\left\|\mathcal{P}_T(\boldsymbol{\Delta})\right\|_{\mathrm{F}}$$

as long as c is sufficiently small. As a result,

$$\frac{1}{\sqrt{p}}\left\|\mathcal{P}_{\Omega}(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \geq \frac{\sqrt{c_{\mathrm{inj}}}}{2}\left\|\mathcal{P}_{T}(\boldsymbol{\Delta})\right\|_{\mathrm{F}}.$$

In addition, by (59) we have

$$\|\boldsymbol{\Delta}\|_{\mathrm{F}} \leq \left\|\mathcal{P}_{T}(\boldsymbol{\Delta})\right\|_{\mathrm{F}} + \left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \leq 2\left\|\mathcal{P}_{T}(\boldsymbol{\Delta})\right\|_{\mathrm{F}},$$

and therefore

$$\frac{1}{\sqrt{p}}\left\|\mathcal{P}_{\Omega}(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \geq \frac{\sqrt{c_{\mathrm{inj}}}}{2}\left\|\mathcal{P}_{T}(\boldsymbol{\Delta})\right\|_{\mathrm{F}} \geq \frac{\sqrt{c_{\mathrm{inj}}}}{4}\|\boldsymbol{\Delta}\|_{\mathrm{F}}. \tag{61}$$

Taking (60) and (61) collectively yields

$$\frac{c_{\mathrm{inj}}}{32}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^{2} \leq \frac{1}{2p}\left\|\mathcal{P}_{\Omega}(\boldsymbol{\Delta})\right\|_{\mathrm{F}}^{2} \leq 72\kappa\frac{1}{\sqrt{\sigma_{\min}}}\left\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\right\|_{\mathrm{F}}\|\boldsymbol{\Delta}\|_{\mathrm{F}},$$

thus indicating that

$$\|\boldsymbol{\Delta}\|_{\mathrm{F}} \lesssim \frac{\kappa}{c_{\mathrm{inj}}}\frac{1}{\sqrt{\sigma_{\min}}}\left\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\right\|_{\mathrm{F}}.$$

C.2.1 Proof of Claim 2

Before proceeding to the proof of Claim 2, we state a useful fact; the proof is deferred to Appendix C.2.2.

Claim 3. *Instate the notations and assumptions in Lemma 2. Let $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$ be the SVD of $\boldsymbol{X}\boldsymbol{Y}^{\top}$. There exists an invertible matrix $\boldsymbol{Q}\in\mathbb{R}^{r\times r}$ such that $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}$, $\boldsymbol{Y}=\boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}^{-\top}$ and*

$$\left\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}-\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\right\|_{\mathrm{F}} \leq 8\sqrt{\kappa}\frac{p}{\lambda\sqrt{\sigma_{\min}}}\left\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\right\|_{\mathrm{F}} \leq 8c\sqrt{c_{\mathrm{inj}}p/\kappa}, \tag{62}$$

where $\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}$ is the SVD of $\boldsymbol{Q}$.

In light of the assumptions (48), one has

$$\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{M}\right)\boldsymbol{Y}=-\lambda\boldsymbol{X}+\boldsymbol{B}_{1} \qquad \text{and} \qquad \left[\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{M}\right)\right]^{\top}\boldsymbol{X}=-\lambda\boldsymbol{Y}+\boldsymbol{B}_{2} \tag{63}$$

for some $\boldsymbol{B}_{1}\in\mathbb{R}^{n\times r}$ and $\boldsymbol{B}_{2}\in\mathbb{R}^{n\times r}$, where $\max\left\{\|\boldsymbol{B}_{1}\|_{\mathrm{F}},\|\boldsymbol{B}_{2}\|_{\mathrm{F}}\right\}\leq p\left\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\right\|_{\mathrm{F}}$. Recall that

$$\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{M}\right)=-\lambda\boldsymbol{U}\boldsymbol{V}^{\top}+\boldsymbol{R}. \tag{64}$$

In the sequel, we shall prove the upper bounds on both $\left\|\mathcal{P}_{T}(\boldsymbol{R})\right\|_{\mathrm{F}}$ and $\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{R})\right\|$ separately.

1. From the definition of $\mathcal{P}_{T}(\cdot)$ (see (15)), we have

$$\begin{aligned}\left\|\mathcal{P}_{T}(\boldsymbol{R})\right\|_{\mathrm{F}} &= \left\|\boldsymbol{U}\boldsymbol{U}^{\top}\boldsymbol{R}(\boldsymbol{I}-\boldsymbol{V}\boldsymbol{V}^{\top})+\boldsymbol{R}\boldsymbol{V}\boldsymbol{V}^{\top}\right\|_{\mathrm{F}}\\ &\leq \left\|\boldsymbol{U}^{\top}\boldsymbol{R}(\boldsymbol{I}-\boldsymbol{V}\boldsymbol{V}^{\top})\right\|_{\mathrm{F}}+\|\boldsymbol{R}\boldsymbol{V}\|_{\mathrm{F}}\\ &\leq \left\|\boldsymbol{U}^{\top}\boldsymbol{R}\right\|_{\mathrm{F}}+\|\boldsymbol{R}\boldsymbol{V}\|_{\mathrm{F}}.\end{aligned} \tag{65}$$

In addition, invoke Claim 3 to obtain

$$\boldsymbol{X}=\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q} \qquad \text{and} \qquad \boldsymbol{Y}=\boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}^{-\top} \tag{66}$$

for some invertible matrix $\boldsymbol{Q}\in\mathbb{R}^{r\times r}$, whose SVD $\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}$ obeys (62). Combine (63) and (64) to see

$$-\lambda\boldsymbol{U}\boldsymbol{V}^{\top}\boldsymbol{Y}+\boldsymbol{R}\boldsymbol{Y}=-\lambda\boldsymbol{X}+\boldsymbol{B}_{1},$$

which together with (66) yields

$$\boldsymbol{R}\boldsymbol{V}=\lambda\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{I}_{r}-\boldsymbol{Q}\boldsymbol{Q}^{\top}\right)\boldsymbol{\Sigma}^{-1/2}+\boldsymbol{B}_{1}\boldsymbol{Q}^{\top}\boldsymbol{\Sigma}^{-1/2}.$$

Apply the triangle inequality to get

$$\begin{aligned}\|\boldsymbol{R}\boldsymbol{V}\|_{\mathrm{F}} &\leq \|\lambda \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{I}_r - \boldsymbol{Q}\boldsymbol{Q}^\top\right)\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{F}} + \|\boldsymbol{B}_1\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{F}} \\ &\leq \lambda\left\|\boldsymbol{\Sigma}^{1/2}\right\|\left\|\boldsymbol{\Sigma}^{-1/2}\right\|\left\|\boldsymbol{Q}\boldsymbol{Q}^\top - \boldsymbol{I}_r\right\|_{\mathrm{F}} + \|\boldsymbol{Q}\|\left\|\boldsymbol{\Sigma}^{-1/2}\right\|\|\boldsymbol{B}_1\|_{\mathrm{F}}. \end{aligned} \tag{67}$$

In order to further upper bound (67), we first recognize that (49) implies

$$\left\|\boldsymbol{\Sigma}^{1/2}\right\| \leq \sqrt{2\sigma_{\max}}, \qquad \text{and} \qquad \left\|\boldsymbol{\Sigma}^{-1/2}\right\| = 1/\sqrt{\sigma_{\min}\left(\boldsymbol{\Sigma}\right)} \leq \sqrt{2/\sigma_{\min}}.$$

Second, Claim 3 yields

$$\left\|\boldsymbol{\Sigma}_{\boldsymbol{Q}} - \boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\right\|_{\mathrm{F}} \leq 8\sqrt{\kappa}\frac{p}{\lambda\sqrt{\sigma_{\min}}}\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\|_{\mathrm{F}} \leq 8c\sqrt{c_{\mathrm{inj}}p/\kappa} \ll 1,$$

with the proviso that c is sufficiently small. Here we have used the facts that $c_{\mathrm{inj}} \leq 1/p$ and that $\kappa \geq 1$. This in turn implies that $\|\boldsymbol{Q}\| = \left\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}\right\| \leq 2$. Putting the above bounds together yields

$$\begin{aligned}\|\boldsymbol{R}\boldsymbol{V}\|_{\mathrm{F}} &\leq \lambda\sqrt{2\sigma_{\max}}\sqrt{\frac{2}{\sigma_{\min}}}\left\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}^2 - \boldsymbol{I}_r\right\|_{\mathrm{F}} + 2\sqrt{\frac{2}{\sigma_{\min}}}p\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\|_{\mathrm{F}} \\ &\leq \lambda\sqrt{2\sigma_{\max}}\sqrt{\frac{2}{\sigma_{\min}}}\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}\|\left\|\boldsymbol{\Sigma}_{\boldsymbol{Q}} - \boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\right\|_{\mathrm{F}} + 2\sqrt{\frac{2}{\sigma_{\min}}}p\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\|_{\mathrm{F}} \\ &\leq 2\lambda\sqrt{2\sigma_{\max}}\sqrt{\frac{2}{\sigma_{\min}}}8\sqrt{\kappa}\frac{p}{\lambda\sqrt{\sigma_{\min}}}\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\|_{\mathrm{F}} + 2\sqrt{\frac{2}{\sigma_{\min}}}p\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\|_{\mathrm{F}} \\ &\leq 36\kappa\frac{p}{\sqrt{\sigma_{\min}}}\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\|_{\mathrm{F}}. \end{aligned}$$

Similarly we can show that $\left\|\boldsymbol{U}^\top\boldsymbol{R}\right\|_{\mathrm{F}} \leq 36\kappa p\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\|_{\mathrm{F}}/\sqrt{\sigma_{\min}}$. These bounds together with (65) result in

$$\|\mathcal{P}_T(\boldsymbol{R})\|_{\mathrm{F}} \leq 72\kappa\frac{p}{\sqrt{\sigma_{\min}}}\|\nabla f\left(\boldsymbol{X},\boldsymbol{Y}\right)\|_{\mathrm{F}}. \tag{68}$$

2. We now move on to bounding $\|\mathcal{P}_{T^\perp}(\boldsymbol{R})\|$. In view of the definition of $\mathcal{P}_\Omega^{\mathsf{debias}}(\cdot)$ in (35), we can rearrange (63) to derive

$$\begin{aligned}\left[p\boldsymbol{M}^\star + \mathcal{P}_\Omega(\boldsymbol{E}) - \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\right]\boldsymbol{Y} &= p\boldsymbol{X}\boldsymbol{Y}^\top\boldsymbol{Y} + \lambda\boldsymbol{X} - \boldsymbol{B}_1, \\ \left[p\boldsymbol{M}^\star + \mathcal{P}_\Omega(\boldsymbol{E}) - \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\right]^\top\boldsymbol{X} &= p\boldsymbol{Y}\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{Y} - \boldsymbol{B}_2. \end{aligned}$$

In view of the representation $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}$ and $\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}^{-\top}$, the above identities are equivalent to

$$\begin{aligned}\left[p\boldsymbol{M}^\star + \mathcal{P}_\Omega(\boldsymbol{E}) - \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\right]\boldsymbol{V} &= p\boldsymbol{U}\boldsymbol{\Sigma} + \lambda\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{B}_1\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2}, \\ \left[p\boldsymbol{M}^\star + \mathcal{P}_\Omega(\boldsymbol{E}) - \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\right]^\top\boldsymbol{U} &= p\boldsymbol{V}\boldsymbol{\Sigma} + \lambda\boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}^{-\top}\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{B}_2\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}^{-1/2}. \end{aligned}$$

Letting

$$p\boldsymbol{M}^\star + \mathcal{P}_\Omega(\boldsymbol{E}) - \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right) = p\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top + \lambda\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{V}^\top + \tilde{\boldsymbol{R}} \tag{69}$$

for some residual matrix $\tilde{\boldsymbol{R}} \in \mathbb{R}^{n\times n}$, we have

$$\begin{aligned}\mathcal{P}_{T^\perp}\left(\boldsymbol{R}\right) &\overset{\text{(i)}}{=} \mathcal{P}_{T^\perp}\left[\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star - \boldsymbol{E}\right)\right] \\ &\overset{\text{(ii)}}{=} \mathcal{P}_{T^\perp}\left[p\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right) + \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right) - \mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right] \\ &\overset{\text{(iii)}}{=} \mathcal{P}_{T^\perp}\left[p\boldsymbol{M}^\star + \mathcal{P}_\Omega(\boldsymbol{E}) - \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\right] \\ &\overset{\text{(iv)}}{=} \mathcal{P}_{T^\perp}\left(\tilde{\boldsymbol{R}}\right), \end{aligned} \tag{70}$$

where (i) follows from the definition of $\boldsymbol{R}$ and the fact that $\boldsymbol{U}\boldsymbol{V}^\top \in T$, (ii) uses the definition of $\mathcal{P}_\Omega^{\mathsf{debias}}(\cdot)$, (iii) relies on the fact that $\boldsymbol{X}\boldsymbol{Y}^\top \in T$, and (iv) applies (69) and the facts that $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \in T$ and that $\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{V}^\top \in T$. Therefore, it suffices to bound $\|\mathcal{P}_{T^\perp}(\tilde{\boldsymbol{R}})\|$. Rewrite (69) as

$$p\boldsymbol{M}^\star + \mathcal{P}_\Omega(\boldsymbol{E}) - \mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right) - \mathcal{P}_T\big(\tilde{\boldsymbol{R}}\big) = \boldsymbol{U}(p\boldsymbol{\Sigma} + \lambda\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2})\boldsymbol{V}^\top + \mathcal{P}_{T^\perp}\big(\tilde{\boldsymbol{R}}\big). \tag{71}$$

Suppose for the moment that

$$\big\|\mathcal{P}_T\big(\tilde{\boldsymbol{R}}\big)\big\| \le \lambda/4. \tag{72}$$

This together with the assumptions that $\|\mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\| < \lambda/8$ and $\|\mathcal{P}_\Omega(\boldsymbol{E})\| < \lambda/8$ reveals that

$$\big\|\mathcal{P}_\Omega(\boldsymbol{E}) - \mathcal{P}_\Omega^{\mathsf{debias}}(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star) - \mathcal{P}_T(\tilde{\boldsymbol{R}})\big\| < \lambda/2. \tag{73}$$

By Weyl's inequality and the relations (71) and (73), one has

$$\begin{aligned}\sigma_i\left[\boldsymbol{U}\big(p\boldsymbol{\Sigma} + \lambda\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2}\big)\boldsymbol{V}^\top + \mathcal{P}_{T^\perp}\big(\tilde{\boldsymbol{R}}\big)\right] &\le \sigma_i\left(p\boldsymbol{M}^\star\right) + \big\|\mathcal{P}_\Omega(\boldsymbol{E}) - \mathcal{P}_\Omega^{\mathsf{debias}}(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star) - \mathcal{P}_T(\tilde{\boldsymbol{R}})\big\| \\ &< p\sigma_i\left(\boldsymbol{M}^\star\right) + \lambda/2 = \lambda/2 \end{aligned} \tag{74}$$

for any $r+1 \le i \le n$, where $\sigma_i(\boldsymbol{A})$ denotes the ith largest singular value of a matrix $\boldsymbol{A}$. Here, we have used the fact that $\boldsymbol{M}^\star$ has rank r and hence $\sigma_i(\boldsymbol{M}^\star) = 0$ for any $i > r$. In addition, it is seen that

$$\begin{aligned}\big\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I}_r\big\| &= \big\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{Q}\boldsymbol{Q}^\top - \boldsymbol{I}_r)\boldsymbol{\Sigma}^{-1/2}\big\| \\ &\le \big\|\boldsymbol{\Sigma}^{1/2}\big\|\big\|\boldsymbol{\Sigma}^{-1/2}\big\|\,\big\|\boldsymbol{Q}\boldsymbol{Q}^\top - \boldsymbol{I}_r\big\|_{\mathrm{F}}.\end{aligned}$$

Note that in (67), we have obtained

$$\big\|\boldsymbol{\Sigma}^{1/2}\big\|\big\|\boldsymbol{\Sigma}^{-1/2}\big\|\,\big\|\boldsymbol{Q}\boldsymbol{Q}^\top - \boldsymbol{I}_r\big\|_{\mathrm{F}} \le 2\sqrt{2\sigma_{\max}}\sqrt{2/\sigma_{\min}}8c\sqrt{c_{\mathrm{inj}}p/\kappa} \le 1/10$$

as long as c is sufficiently small, and hence $\big\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I}_r\big\| \le 1/10$. Therefore, for any $1 \le i \le r$ we know that

$$\begin{aligned}\sigma_i\left[\boldsymbol{U}\big(p\boldsymbol{\Sigma} + \lambda\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2}\big)\boldsymbol{V}^\top\right] &\ge \sigma_r\left[\boldsymbol{U}\big(p\boldsymbol{\Sigma} + \lambda\boldsymbol{I}_r + \lambda\big(\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I}_r\big)\big)\boldsymbol{V}^\top\right] \\ &\ge \sigma_r\left(p\boldsymbol{\Sigma} + \lambda\boldsymbol{I}_r\right) - \lambda\big\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I}_r\big\| \\ &\ge \lambda - \lambda\big\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I}_r\big\| \\ &\ge \lambda - \lambda/10 > \lambda/2,\end{aligned}$$

where the second inequality results from Weyl's inequality. This combined with (70) and (74) yields

$$\|\mathcal{P}_{T^\perp}(\boldsymbol{R})\| = \big\|\mathcal{P}_{T^\perp}\big(\tilde{\boldsymbol{R}}\big)\big\| < \lambda/2;$$

this happens because at least $n-r$ singular values of $\boldsymbol{U}\left(p\boldsymbol{\Sigma} + \lambda\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2}\right)\boldsymbol{V}^\top + \mathcal{P}_{T^\perp}\big(\tilde{\boldsymbol{R}}\big)$ are no larger than $\lambda/2$ and they cannot correspond to directions simultaneously in the column space spanned by $\boldsymbol{U}$ and the row space spanned by $\boldsymbol{V}^\top$.

The proof is then complete by verifying (72). To this end, observe that

$$\tilde{\boldsymbol{R}}\boldsymbol{V} = -\boldsymbol{B}_1\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1/2}, \quad \tilde{\boldsymbol{R}}^\top\boldsymbol{U} = \lambda\boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}^{-\top}\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}^{-1/2} - \lambda\boldsymbol{V}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{1/2} - \boldsymbol{B}_2\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}^{-1/2}.$$

Then following similar technique used to bound $\|\mathcal{P}_T(\boldsymbol{R})\|$, we have

$$\big\|\mathcal{P}_T(\tilde{\boldsymbol{R}})\big\| \le \big\|\mathcal{P}_T\big(\tilde{\boldsymbol{R}}\big)\big\|_{\mathrm{F}} \le \big\|\boldsymbol{U}^\top\tilde{\boldsymbol{R}}\big\|_{\mathrm{F}} + \big\|\tilde{\boldsymbol{R}}\boldsymbol{V}\big\|_{\mathrm{F}} \lesssim c\sqrt{c_{\mathrm{inj}}p}\lambda < \lambda/4 \tag{75}$$

as long as c is small enough.

C.2.2 Proof of Claim 3

Let

$$\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}\right)\boldsymbol{Y}+\lambda\boldsymbol{X}=\boldsymbol{B}_1 \qquad \text{and} \qquad \left[\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}\right)\right]^\top\boldsymbol{X}+\lambda\boldsymbol{Y}=\boldsymbol{B}_2 \tag{76}$$

for some $\boldsymbol{B}_1,\boldsymbol{B}_2\in\mathbb{R}^{n\times r}$. Clearly, it is seen from the assumption (48) that

$$\max\{\|\boldsymbol{B}_1\|_{\mathrm{F}},\|\boldsymbol{B}_2\|_{\mathrm{F}}\}\le p\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\|_{\mathrm{F}}. \tag{77}$$

In addition, the identities (76) allow us to obtain

$$\begin{aligned}\left\|\boldsymbol{X}^\top\boldsymbol{X}-\boldsymbol{Y}^\top\boldsymbol{Y}\right\|_{\mathrm{F}} &= \tfrac{1}{\lambda}\left\|\boldsymbol{X}^\top(\boldsymbol{B}_1-\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}\right)\boldsymbol{Y})-\left(\boldsymbol{B}_2-\left[\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}\right)\right]^\top\boldsymbol{X}\right)^\top\boldsymbol{Y}\right\|_{\mathrm{F}}\\ &= \tfrac{1}{\lambda}\left\|\boldsymbol{X}^\top\boldsymbol{B}_1-\boldsymbol{B}_2^\top\boldsymbol{Y}\right\|_{\mathrm{F}}\\ &\le \tfrac{1}{\lambda}\|\boldsymbol{X}\|\|\boldsymbol{B}_1\|_{\mathrm{F}}+\tfrac{1}{\lambda}\|\boldsymbol{B}_2\|_{\mathrm{F}}\|\boldsymbol{Y}\|\\ &\le 2\tfrac{p}{\lambda}\sqrt{2\sigma_{\max}}\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\|_{\mathrm{F}}.\end{aligned} \tag{78}$$

Here, the last line makes use of (77) and the assumption that $\|\boldsymbol{X}\|,\|\boldsymbol{Y}\|\le\sqrt{2\sigma_{\max}}$. In view of Lemma 20, one can find an invertible $\boldsymbol{Q}$ such that $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}$, $\boldsymbol{Y}=\boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}^{-\top}$ and

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}-\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\right\|_{\mathrm{F}} &\le \frac{1}{\sigma_{\min}(\boldsymbol{\Sigma})}\left\|\boldsymbol{X}^\top\boldsymbol{X}-\boldsymbol{Y}^\top\boldsymbol{Y}\right\|_{\mathrm{F}}\\ &\overset{\text{(i)}}{\le} \frac{2}{\sigma_{\min}}\cdot 2\frac{p}{\lambda}\sqrt{2\sigma_{\max}}\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\|_{\mathrm{F}}\\ &\le 8\sqrt{\kappa}\frac{p}{\lambda\sqrt{\sigma_{\min}}}\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\|_{\mathrm{F}}\\ &\overset{\text{(ii)}}{\le} 8c\sqrt{c_{\mathrm{inj}}p/\kappa},\end{aligned}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{Q}}$ is a diagonal matrix consisting of all singular values of $\boldsymbol{Q}$. Here, (i) follows from (49) as well as the bound (78), and the last inequality (ii) uses the assumption (23). This completes the proof.

C.3 Proof of Lemma 4

Lemma 4 consists of two parts, which we restate into the following two lemmas, namely Lemmas 7-8.

First of all, Lemma 7 demonstrates that as long as $(\boldsymbol{X},\boldsymbol{Y})$ is sufficiently close to $(\boldsymbol{X}^\star,\boldsymbol{Y}^\star)$, the operator $\mathcal{P}_\Omega(\cdot)$ restricted to the tangent space T of $\boldsymbol{X}\boldsymbol{Y}^\top$ is injective. The proof is deferred to Appendix C.3.1.

Lemma 7. *Suppose that the sample complexity obeys $n^2p\ge C\mu rn\log n$ for some sufficiently large constant $C>0$. Then with probability exceeding $1-O(n^{-10})$,*

$$\frac{1}{p}\|\mathcal{P}_\Omega(\boldsymbol{H})\|_{\mathrm{F}}^2\ge\frac{1}{32\kappa}\|\boldsymbol{H}\|_{\mathrm{F}}^2, \qquad \forall\boldsymbol{H}\in T$$

holds simultaneously for all $(\boldsymbol{X},\boldsymbol{Y})$ obeying

$$\max\left\{\|\boldsymbol{X}-\boldsymbol{X}^\star\|_{2,\infty},\|\boldsymbol{Y}-\boldsymbol{Y}^\star\|_{2,\infty}\right\}\le\frac{c}{\kappa\sqrt{n}}\|\boldsymbol{X}^\star\|. \tag{79}$$

Here, $c>0$ is some sufficiently small constant, and T denotes the tangent space of $\boldsymbol{X}\boldsymbol{Y}^\top$.

Remark 7. In the prior literature, the injectivity of $\mathcal{P}_\Omega(\cdot)$ has been mostly studied when restricted to a *fixed* tangent space independent of Ω (see [CR09, Gro11]). In comparison, this lemma demonstrates that the injectivity property holds uniformly over a large set of tangent spaces. This allows one to handle tangent spaces that are statistically dependent on Ω.

Remark 8. Note that the condition (79) on $(\boldsymbol{X}, \boldsymbol{Y})$ is weaker than (26) under the assumptions of Lemma 4. To see this, if (26) holds, then one necessarily has

$$\begin{aligned}\|\boldsymbol{X}-\boldsymbol{X}^\star\|_{2,\infty} &\le C_\infty\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\max\left\{\|\boldsymbol{X}^\star\|_{2,\infty},\|\boldsymbol{Y}^\star\|_{2,\infty}\right\}\\ &\overset{\text{(i)}}{\lesssim} C_\infty\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\max\left\{\|\boldsymbol{X}^\star\|_{2,\infty},\|\boldsymbol{Y}^\star\|_{2,\infty}\right\}\\ &\overset{\text{(ii)}}{\le} C_\infty\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\sqrt{\frac{\mu r}{n}}\|\boldsymbol{X}^\star\|\\ &\overset{\text{(iii)}}{\le} \frac{c}{\kappa\sqrt{n}}\|\boldsymbol{X}^\star\|.\end{aligned}$$

Here, (i) follows from the choice $\lambda \asymp \sigma\sqrt{np}$; (ii) relies on the incoherence assumption (34); and (iii) holds true under the noise condition $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^4\mu r\log n}}$. A similar bound holds for $\|\boldsymbol{Y}-\boldsymbol{Y}^\star\|_{2,\infty}$.

The next lemma shows that for all $(\boldsymbol{X},\boldsymbol{Y})$ close to $(\boldsymbol{X}^\star,\boldsymbol{Y}^\star)$, $\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}^\star)$ is uniformly close to its expectation $p(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}^\star)$. The proof can be found in Appendix C.3.2.

Lemma 8. *Suppose that $n^2p \gg \kappa^4\mu^2r^2n\log^2 n$ and $\sigma\sqrt{n(\log n)/p} \ll \sigma_{\min}/\kappa$. With probability exceeding $1-O(n^{-10})$, one has*

$$\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}^\star\right)-p\left(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}^\star\right)\right\| < \lambda/8$$

simultaneously for any $(\boldsymbol{X},\boldsymbol{Y})$ obeying (26), provided that $\lambda = C_\lambda\sigma\sqrt{np}$ for some constant $C_\lambda > 0$.

C.3.1 Proof of Lemma 7

By definition, any $\boldsymbol{H} \in T$ can be expressed as

$$\boldsymbol{H} = \boldsymbol{X}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^\top \tag{80}$$

for some $\boldsymbol{A},\boldsymbol{B} \in \mathbb{R}^{n\times r}$. Given that this is an underdetermined linear system of equations, there might be numerous $(\boldsymbol{A},\boldsymbol{B})$'s compatible with (80). We take a specific choice as follows

$$\begin{aligned}(\boldsymbol{A},\boldsymbol{B}) := \arg\min_{(\tilde{\boldsymbol{A}},\tilde{\boldsymbol{B}})} \quad & 0.5\|\tilde{\boldsymbol{A}}\|_{\mathrm{F}}^2 + 0.5\|\tilde{\boldsymbol{B}}\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \boldsymbol{H} = \boldsymbol{X}\tilde{\boldsymbol{A}}^\top + \tilde{\boldsymbol{B}}\boldsymbol{Y}^\top.\end{aligned} \tag{81}$$

which satisfies a property that plays an important role in the subsequent analysis:

$$\boldsymbol{X}^\top\boldsymbol{B} = \boldsymbol{A}^\top\boldsymbol{Y}. \tag{82}$$

To see this, consider the Lagrangian

$$\mathcal{L}(\tilde{\boldsymbol{A}},\tilde{\boldsymbol{B}},\boldsymbol{\Lambda}) := 0.5\|\tilde{\boldsymbol{A}}\|_{\mathrm{F}}^2 + 0.5\|\tilde{\boldsymbol{B}}\|_{\mathrm{F}}^2 + \langle\boldsymbol{\Lambda}, \boldsymbol{X}\tilde{\boldsymbol{A}}^\top + \tilde{\boldsymbol{B}}\boldsymbol{Y}^\top - \boldsymbol{H}\rangle.$$

Taking the derivatives w.r.t. $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{B}}$ and setting them to zero yield

$$\boldsymbol{A} = -\boldsymbol{\Lambda}^\top\boldsymbol{X} \qquad \text{and} \qquad \boldsymbol{B} = -\boldsymbol{\Lambda}\boldsymbol{Y}$$

for some Lagrangian multiplier matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times n}$. The claim (82) then follows immediately.

The remaining proof consists of two steps.

- First, we would like to show that

$$\|\boldsymbol{H}\|_{\mathrm{F}}^2 \le 8\sigma_{\max}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right). \tag{83}$$

- Second, we prove that

$$\frac{1}{2p}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{H}\right)\right\|_{\mathrm{F}}^{2}=\frac{1}{2p}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{A}^{\top}+\boldsymbol{B}\boldsymbol{Y}^{\top}\right)\right\|_{\mathrm{F}}^{2}\geq\frac{\sigma_{\min}}{8}\left(\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{B}\right\|_{\mathrm{F}}^{2}\right). \tag{84}$$

Taking (83) and (84) together immediately yields the claimed bounds in the lemma. In what follows, we shall establish these two bounds separately.

1. Regarding the upper bound (83), it follows from elementary inequalities that

$$\begin{aligned}\left\|\boldsymbol{H}\right\|_{\mathrm{F}}^{2}&=\left\|\boldsymbol{X}\boldsymbol{A}^{\top}+\boldsymbol{B}\boldsymbol{Y}^{\top}\right\|_{\mathrm{F}}^{2}\leq2\big(\left\|\boldsymbol{X}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{B}\boldsymbol{Y}^{\top}\right\|_{\mathrm{F}}^{2}\big)\\&\leq2\big(\left\|\boldsymbol{X}\right\|^{2}\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{Y}\right\|^{2}\left\|\boldsymbol{B}\right\|_{\mathrm{F}}^{2}\big)\\&\leq2\max\left\{\left\|\boldsymbol{X}\right\|^{2},\left\|\boldsymbol{Y}\right\|^{2}\right\}\left(\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{B}\right\|_{\mathrm{F}}^{2}\right).\end{aligned} \tag{85}$$

It then suffices to control $\max\{\|\boldsymbol{X}\|,\|\boldsymbol{Y}\|\}$. In view of the assumption (79), one has

$$\left\|\boldsymbol{X}-\boldsymbol{X}^{\star}\right\|\leq\left\|\boldsymbol{X}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}\leq\sqrt{n}\left\|\boldsymbol{X}-\boldsymbol{X}^{\star}\right\|_{2,\infty}\leq\frac{c}{\kappa}\left\|\boldsymbol{X}^{\star}\right\|\leq\left\|\boldsymbol{X}^{\star}\right\|, \tag{86}$$

as long as $c<1$. This together with the triangle inequality reveals that

$$\left\|\boldsymbol{X}\right\|\leq\left\|\boldsymbol{X}^{\star}\right\|+\left\|\boldsymbol{X}-\boldsymbol{X}^{\star}\right\|\leq2\left\|\boldsymbol{X}^{\star}\right\|\leq2\sqrt{\sigma_{\max}}.$$

Similarly, one has $\|\boldsymbol{Y}\|\leq2\sqrt{\sigma_{\max}}$. Substitution into (85) yields the desired upper bound (83).

2. We now move on to the lower bound (84). To this end, one first decomposes

$$\frac{1}{2p}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{A}^{\top}+\boldsymbol{B}\boldsymbol{Y}^{\top}\right)\right\|_{\mathrm{F}}^{2}=\underbrace{\frac{1}{2p}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{A}^{\top}+\boldsymbol{B}\boldsymbol{Y}^{\top}\right)\right\|_{\mathrm{F}}^{2}-\frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{A}^{\top}+\boldsymbol{B}\boldsymbol{Y}^{\top}\right\|_{\mathrm{F}}^{2}}_{:=\alpha_1}+\underbrace{\frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{A}^{\top}+\boldsymbol{B}\boldsymbol{Y}^{\top}\right\|_{\mathrm{F}}^{2}}_{:=\alpha_2}.$$

The basic idea is to demonstrate that (1) α_2 is bounded from below, and (2) α_1 is sufficiently small compared to α_2.

(a) We start by controlling α_2, towards which we can expand

$$\alpha_{2}=\frac{1}{2}\left(\left\|\boldsymbol{X}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{B}\boldsymbol{Y}^{\top}\right\|_{\mathrm{F}}^{2}\right)+\mathrm{Tr}\left(\boldsymbol{X}^{\top}\boldsymbol{B}\boldsymbol{Y}^{\top}\boldsymbol{A}\right).$$

The property $\boldsymbol{X}^{\top}\boldsymbol{B}=\boldsymbol{A}^{\top}\boldsymbol{Y}$ (see (82)) implies that

$$\mathrm{Tr}\left(\boldsymbol{X}^{\top}\boldsymbol{B}\boldsymbol{Y}^{\top}\boldsymbol{A}\right)=\left\|\boldsymbol{X}^{\top}\boldsymbol{B}\right\|_{\mathrm{F}}^{2}\geq0\qquad\Longrightarrow\qquad\alpha_{2}\geq\frac{1}{2}\left(\left\|\boldsymbol{X}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{B}\boldsymbol{Y}^{\top}\right\|_{\mathrm{F}}^{2}\right).$$

Write $\boldsymbol{\Delta}_{\boldsymbol{X}}=\boldsymbol{X}-\boldsymbol{X}^{\star}$ and $\boldsymbol{\Delta}_{\boldsymbol{Y}}=\boldsymbol{Y}-\boldsymbol{Y}^{\star}$. We have

$$\begin{aligned}\left\|\boldsymbol{X}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}&=\left\|\left(\boldsymbol{X}^{\star}+\boldsymbol{\Delta}_{\boldsymbol{X}}\right)\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}=\left\|\boldsymbol{X}^{\star}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}+2\left\langle\boldsymbol{X}^{\star}\boldsymbol{A}^{\top},\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^{\top}\right\rangle\\&\geq\left\|\boldsymbol{X}^{\star}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}-2\left\|\boldsymbol{X}^{\star}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}\\&\geq\left\|\boldsymbol{X}^{\star}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}-2\left\|\boldsymbol{X}^{\star}\right\|\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^{2},\end{aligned}$$

where the second line arises from the Cauchy-Schwarz inequality. Recalling from (86) that $\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|\leq c\|\boldsymbol{X}^{\star}\|/\kappa$, we arrive at

$$\left\|\boldsymbol{X}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}\geq\left\|\boldsymbol{X}^{\star}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}-2c\sigma_{\min}\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^{2}\geq\left\|\boldsymbol{X}^{\star}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}-\sigma_{\min}\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^{2}/100,$$

provided that $c\leq1/200$. A similar bound holds for $\|\boldsymbol{B}\boldsymbol{Y}^{\top}\|_{\mathrm{F}}^{2}$, thus leading to

$$\alpha_{2}\geq\frac{1}{2}\left(\left\|\boldsymbol{X}^{\star}\boldsymbol{A}^{\top}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{B}\boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}^{2}\right)-\frac{1}{100}\sigma_{\min}\left(\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{B}\right\|_{\mathrm{F}}^{2}\right).$$

(b) Next, we control α_1. First, it is seen that

$$\begin{aligned}\boldsymbol{X}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^\top &= \left(\boldsymbol{X}^\star + \boldsymbol{\Delta}_{\boldsymbol{X}}\right)\boldsymbol{A}^\top + \boldsymbol{B}\left(\boldsymbol{Y}^\star + \boldsymbol{\Delta}_{\boldsymbol{Y}}\right)^\top \\ &= \boldsymbol{X}^\star\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^{\star\top} + \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top.\end{aligned}$$

As a result, we can expand α_1 as

$$\begin{aligned}\alpha_1 &= \frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^{\star\top} + \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right)\right\|_{\mathrm{F}}^2 - \frac{1}{2}\left\|\boldsymbol{X}^\star\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^{\star\top} + \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\|_{\mathrm{F}}^2 \\ &= \underbrace{\frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^{\star\top}\right)\right\|_{\mathrm{F}}^2 - \frac{1}{2}\left\|\boldsymbol{X}^\star\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}^2}_{:=\gamma_1} \\ &\quad + \underbrace{\frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right)\right\|_{\mathrm{F}}^2 - \frac{1}{2}\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right\|_{\mathrm{F}}^2}_{:=\gamma_2} + \underbrace{\frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right)\right\|_{\mathrm{F}}^2 - \frac{1}{2}\left\|\boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\|_{\mathrm{F}}^2}_{:=\gamma_3} \\ &\quad + \underbrace{\frac{1}{p}\left\langle\mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right)\right\rangle - \left\langle\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top, \boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\rangle}_{:=\gamma_4} \\ &\quad + \underbrace{\frac{1}{p}\left\langle\mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^{\star\top}\right), \mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right)\right\rangle - \left\langle\boldsymbol{X}^\star\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^{\star\top}, \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\rangle}_{:=\gamma_5}.\end{aligned}$$

i. Regarding γ_1, it follows from the bounds in [CR09, Section 4.2] that

$$|\gamma_1| \le \frac{1}{64}\left\|\boldsymbol{X}^\star\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}^2 \le \frac{1}{32}\left(\left\|\boldsymbol{X}^\star\boldsymbol{A}^\top\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{B}\boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}^2\right),$$

as long as $np \gg \mu r \log n$.

ii. Invoke Lemma 19 to show that

$$\begin{aligned}|\gamma_2| &\le \frac{3n}{2}\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|_{2,\infty}^2\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^2 \le \frac{3c^2}{2\kappa}\sigma_{\min}\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^2 \le \frac{1}{100}\sigma_{\min}\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^2, \\ |\gamma_3| &\le \frac{3n}{2}\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\right\|_{2,\infty}^2\left\|\boldsymbol{B}\right\|_{\mathrm{F}}^2 \le \frac{3c^2}{2\kappa}\sigma_{\min}\left\|\boldsymbol{B}\right\|_{\mathrm{F}}^2 \le \frac{1}{100}\sigma_{\min}\left\|\boldsymbol{B}\right\|_{\mathrm{F}}^2,\end{aligned}$$

as long as $n^2 p \gg n\log n$ and $c > 0$ is sufficiently small. Here we have utilized the assumption that $\max\{\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{2,\infty}, \|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{2,\infty}\} \le c\|\boldsymbol{X}^\star\|/(\kappa\sqrt{n})$.

iii. The term γ_4 can be controlled via Lemma 21:

$$\begin{aligned}|\gamma_4| &\le \left\|\frac{1}{p}\mathcal{P}_\Omega\left(\mathbf{1}\mathbf{1}^\top\right) - \mathbf{1}\mathbf{1}^\top\right\|\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|_{2,\infty}\left\|\boldsymbol{A}\right\|_{\mathrm{F}}\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\right\|_{2,\infty}\left\|\boldsymbol{B}\right\|_{\mathrm{F}} \\ &\lesssim \sqrt{\frac{n}{p}}\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|_{2,\infty}\left\|\boldsymbol{A}\right\|_{\mathrm{F}}\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\right\|_{2,\infty}\left\|\boldsymbol{B}\right\|_{\mathrm{F}},\end{aligned}$$

where the second line uses the bound $\|p^{-1}\mathcal{P}_\Omega\left(\mathbf{1}\mathbf{1}^\top\right) - \mathbf{1}\mathbf{1}^\top\| \lesssim \sqrt{n/p}$ guaranteed by [KMO10a, Lemma 3.2]. Continue the upper bound to get

$$|\gamma_4| \overset{\text{(i)}}{\lesssim} n\frac{c^2}{\kappa^2 n}\sigma_{\max}\left\|\boldsymbol{A}\right\|_{\mathrm{F}}\left\|\boldsymbol{B}\right\|_{\mathrm{F}} \overset{\text{(ii)}}{\le} \frac{c^2}{2\kappa}\sigma_{\min}\left(\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{B}\right\|_{\mathrm{F}}^2\right) \overset{\text{(iii)}}{\le} \frac{1}{100}\sigma_{\min}\left(\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{B}\right\|_{\mathrm{F}}^2\right).$$

Here the first relation (i) arises from the assumption that $np \gg 1$. The second inequality (ii) applies the elementary inequality $ab \le (a^2 + b^2)/2$ and the last one (iii) holds with the proviso that $c > 0$ is small enough.

iv. The last term γ_5 can be further decomposed into the sum of four terms. For brevity, we take one out as an example, namely the term

$$\frac{1}{p}\left\langle \mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{A}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right)\right\rangle - \left\langle \boldsymbol{X}^\star\boldsymbol{A}^\top, \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right\rangle.$$

Apply the triangle inequality to obtain

$$\begin{aligned}
&\left|\frac{1}{p}\left\langle \mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{A}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right)\right\rangle - \left\langle \boldsymbol{X}^\star\boldsymbol{A}^\top, \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right\rangle\right| \\
&\quad\le \left|\frac{1}{p}\left\langle \mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{A}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right)\right\rangle\right| + \left|\left\langle \boldsymbol{X}^\star\boldsymbol{A}^\top, \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right\rangle\right| \\
&\quad\le \frac{1}{\sqrt{p}}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{A}^\top\right)\right\|_{\mathrm{F}} \frac{1}{\sqrt{p}}\left\|\mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right)\right\|_{\mathrm{F}} + \left\|\boldsymbol{X}^\star\boldsymbol{A}^\top\right\|_{\mathrm{F}}\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right\|_{\mathrm{F}}.
\end{aligned}$$

In light of [CR09, Section 4.2] and [ZL16, Lemma 9], we have

$$\begin{aligned}
&\frac{1}{\sqrt{p}}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{A}^\top\right)\right\|_{\mathrm{F}} \le 1.1\left\|\boldsymbol{X}^\star\boldsymbol{A}^\top\right\|_{\mathrm{F}}; \\
&\frac{1}{\sqrt{p}}\left\|\mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right)\right\|_{\mathrm{F}} \le \sqrt{2n}\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|_{2,\infty}\|\boldsymbol{A}\|_{\mathrm{F}}.
\end{aligned}$$

Taking the above three bounds collectively yields

$$\begin{aligned}
&\left|\frac{1}{p}\left\langle \mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{A}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right)\right\rangle - \left\langle \boldsymbol{X}^\star\boldsymbol{A}^\top, \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right\rangle\right| \\
&\quad\le 5\left\|\boldsymbol{X}^\star\boldsymbol{A}^\top\right\|_{\mathrm{F}}\sqrt{n}\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|_{2,\infty}\|\boldsymbol{A}\|_{\mathrm{F}} + \left\|\boldsymbol{X}^\star\boldsymbol{A}^\top\right\|_{\mathrm{F}}\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right\|_{\mathrm{F}} \\
&\quad\le 5\sqrt{n}\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \sqrt{n}\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|\|\boldsymbol{A}\|_{\mathrm{F}}^2 \\
&\quad= 6\sqrt{n}\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|\|\boldsymbol{A}\|_{\mathrm{F}}^2.
\end{aligned}$$

Using the assumption that $\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{2,\infty} \le c\|\boldsymbol{X}^\star\|/(\kappa\sqrt{n})$, one has

$$\left|\frac{1}{p}\left\langle \mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{A}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right)\right\rangle - \left\langle \boldsymbol{X}^\star\boldsymbol{A}^\top, \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{A}^\top\right\rangle\right| \lesssim \sqrt{n}\frac{c}{\kappa\sqrt{n}}\sigma_{\max}\|\boldsymbol{A}\|_{\mathrm{F}}^2 \le \frac{1}{100}\sigma_{\min}\|\boldsymbol{A}\|_{\mathrm{F}}^2$$

for $c > 0$ small enough. The same argument applies to the remaining three terms, resulting in

$$|\gamma_5| \le \frac{1}{50}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right).$$

v. Combining the previous bounds on γ_1 through γ_5, we arrive at

$$\begin{aligned}
|\alpha_1| &\le |\gamma_1| + |\gamma_2| + |\gamma_3| + |\gamma_4| + |\gamma_5| \\
&\le \frac{1}{32}\left(\left\|\boldsymbol{X}^\star\boldsymbol{A}^\top\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{B}\boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}^2\right) + \frac{1}{25}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right).
\end{aligned}$$

(c) Taking the preceding bounds on α_1 and α_2 collectively yields

$$\begin{aligned}
\frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^\top\right)\right\|_{\mathrm{F}}^2 &\ge \alpha_2 - |\alpha_1| \\
&\ge \frac{15}{32}\left(\left\|\boldsymbol{X}^\star\boldsymbol{A}^\top\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{B}\boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}^2\right) - \frac{1}{5}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right) \\
&\ge \frac{15}{32}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right) - \frac{1}{5}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right) \\
&\ge \frac{1}{8}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right).
\end{aligned}$$

The proof is then complete.

C.3.2 Proof of Lemma 8

To start with, we have

$$\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star = (\boldsymbol{X}-\boldsymbol{X}^\star)\,\boldsymbol{Y}^\top + \boldsymbol{X}^\star\,(\boldsymbol{Y}-\boldsymbol{Y}^\star)^\top,$$

which together with the triangle inequality implies

$$\big\|\mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\big\| \le \big\|\mathcal{P}_\Omega^{\mathsf{debias}}\left[(\boldsymbol{X}-\boldsymbol{X}^\star)\,\boldsymbol{Y}^\top\right]\big\| + \big\|\mathcal{P}_\Omega^{\mathsf{debias}}\left[\boldsymbol{X}^\star\,(\boldsymbol{Y}-\boldsymbol{Y}^\star)^\top\right]\big\|.$$

Apply [CL17, Lemma 4.5] to obtain

$$\begin{aligned}\big\|\mathcal{P}_\Omega^{\mathsf{debias}}\left[(\boldsymbol{X}-\boldsymbol{X}^\star)\,\boldsymbol{Y}^\top\right]\big\| &\le \big\|\mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{1}\boldsymbol{1}^\top\right)\big\|\,\|\boldsymbol{X}-\boldsymbol{X}^\star\|_{2,\infty}\,\|\boldsymbol{Y}\|_{2,\infty}\\ &\lesssim \sqrt{np}\,\|\boldsymbol{X}-\boldsymbol{X}^\star\|_{2,\infty}\,\|\boldsymbol{Y}\|_{2,\infty},\end{aligned}$$

where the second line is due to $\|\mathcal{P}_\Omega^{\mathsf{debias}}(\boldsymbol{1}\boldsymbol{1}^\top)\| \lesssim \sqrt{np}$ (cf. [KMO10a, Lemma 3.2]). Similarly,

$$\big\|\mathcal{P}_\Omega^{\mathsf{debias}}\left[\boldsymbol{X}^\star\,(\boldsymbol{Y}-\boldsymbol{Y}^\star)^\top\right]\big\| \lesssim \sqrt{np}\,\|\boldsymbol{Y}-\boldsymbol{Y}^\star\|_{2,\infty}\,\|\boldsymbol{X}^\star\|_{2,\infty}.$$

In addition, the assumption (26) yields

$$\begin{aligned}\|\boldsymbol{Y}\|_{2,\infty} &\le \|\boldsymbol{Y}-\boldsymbol{Y}^\star\|_{2,\infty} + \|\boldsymbol{Y}^\star\|_{2,\infty}\\ &\le C_\infty\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{Y}^\star\|_{2,\infty} + \|\boldsymbol{Y}^\star\|_{2,\infty}\\ &\le 2\,\|\boldsymbol{Y}^\star\|_{2,\infty},\end{aligned}$$

as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} \ll 1/\kappa$ (recall that $\lambda = C_\lambda\sigma\sqrt{np}$ for some constant $C_\lambda > 0$). As a consequence, one obtains

$$\begin{aligned}\big\|\mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\big\| &\lesssim \sqrt{np}\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{X}^\star\|_{2,\infty}\,\|\boldsymbol{Y}^\star\|_{2,\infty}\\ &\le \sqrt{np}\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\frac{\mu r\sigma_{\max}}{n},\end{aligned} \tag{87}$$

where the last inequality follows from the upper bound $\max\{\|\boldsymbol{X}^\star\|_{2,\infty}, \|\boldsymbol{Y}^\star\|_{2,\infty}\} \le \sqrt{\mu r\sigma_{\max}/n}$ (cf. (34)). Rearrange the right-hand side of (87) to reach

$$\big\|\mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\big\| \lesssim \sigma\sqrt{np}\cdot\sqrt{\frac{\kappa^4\mu^2r^2\log n}{np}} + \lambda\sqrt{\frac{\kappa^4\mu^2r^2}{np}} < \lambda/8,$$

where the last line holds because of the assumption $n^2p \gg \kappa^4\mu^2r^2n\log n$ as well as the choice of λ.

D Analysis of the nonconvex gradient descent algorithm

Lemma 5 shares similar spirit as [MWCC17, Theorem 2] and [CLL19, Lemma 3.5] with one difference: the nonconvex loss function (17) has an additional term $\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2$ to balance the scale of $\boldsymbol{X}$ and $\boldsymbol{Y}$. To simplify the presentation, we find it convenient to introduce a few notations. Denote

$$\boldsymbol{F}^t \triangleq \begin{bmatrix}\boldsymbol{X}^t\\ \boldsymbol{Y}^t\end{bmatrix} \in \mathbb{R}^{2n\times r} \qquad \text{and} \qquad \boldsymbol{F}^\star \triangleq \begin{bmatrix}\boldsymbol{X}^\star\\ \boldsymbol{Y}^\star\end{bmatrix} \in \mathbb{R}^{2n\times r}. \tag{88}$$

It is easily seen from (28) that

$$\boldsymbol{H}^t = \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{F}^t\boldsymbol{R} - \boldsymbol{F}^\star\right\|_{\mathrm{F}}. \tag{89}$$

Algorithm 2 Construction of the lth leave-one-out sequence.

Initialization: $\boldsymbol{X}^{0,(l)} = \boldsymbol{X}^\star$; $\boldsymbol{Y}^{0,(l)} = \boldsymbol{Y}^\star$; Set $\boldsymbol{F}^{0,(l)} \triangleq \begin{bmatrix} \boldsymbol{X}^{0,(l)} \\ \boldsymbol{Y}^{0,(l)} \end{bmatrix}$.

Gradient updates: **for** $t = 0, 1, \ldots, t_0 - 1$ **do**

$$\boldsymbol{F}^{t+1,(l)} \triangleq \begin{bmatrix} \boldsymbol{X}^{t+1,(l)} \\ \boldsymbol{Y}^{t+1,(l)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^{t,(l)} - \eta \nabla_{\boldsymbol{X}} f^{(l)}(\boldsymbol{X}^{t,(l)}, \boldsymbol{Y}^{t,(l)}) \\ \boldsymbol{Y}^{t,(l)} - \eta \nabla_{\boldsymbol{Y}} f^{(l)}(\boldsymbol{X}^{t,(l)}, \boldsymbol{Y}^{t,(l)}) \end{bmatrix},$$

where $\eta > 0$ is the step size.

Similar to [MWCC17,CLL19], we resort to the leave-one-out sequences to control the ℓ_2/ℓ_∞ error. Specifically, for each $1 \le l \le n$ (corresponding to row indices), we construct $\{\boldsymbol{F}^{t,(l)}\}_{t\ge 0}$ to be the gradient descent iterates (see Algorithm 2) w.r.t. the following auxiliary loss function

$$f^{(l)}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{2p}\left\|\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \frac{1}{2}\left\|\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{Y}\|_{\mathrm{F}}^2. \tag{91}$$

Here $\mathcal{P}_{\Omega_{-l,\cdot}}(\cdot)$ (resp. $\mathcal{P}_{l,\cdot}(\cdot)$) denotes the orthogonal projection onto the space of matrices which are supported on the index set $\Omega_{-l,\cdot} = \{(i,j) \in \Omega | i \ne l\}$ (resp. $\{(i,j)|i = l\}$). Mathematically, we have for any matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$

$$\left[\mathcal{P}_{\Omega_{-l,\cdot}}(\boldsymbol{B})\right]_{ij} = \begin{cases} B_{ij}, & \text{if } (i,j) \in \Omega \text{ and } i \ne l, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \left[\mathcal{P}_{l,\cdot}(\boldsymbol{B})\right]_{ij} = \begin{cases} B_{ij}, & \text{if } i = l, \\ 0, & \text{otherwise.} \end{cases} \tag{92}$$

Similarly, for each $n+1 \le l \le 2n$ (with $l-n$ corresponding to the column index), we define $\{\boldsymbol{F}^{t,(l)}\}_{t\ge 0}$ to be the GD iterates (see Algorithm 2) operating on

$$f^{(l)}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{2p}\left\|\mathcal{P}_{\Omega_{\cdot,-(l-n)}}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \frac{1}{2}\left\|\mathcal{P}_{\cdot,(l-n)}\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{Y}\|_{\mathrm{F}}^2,$$

where $\mathcal{P}_{\Omega_{\cdot,-(l-n)}}(\cdot)$ and $\mathcal{P}_{\cdot,(l-n)}(\cdot)$ are defined as

$$\left[\mathcal{P}_{\Omega_{\cdot,-(l-n)}}(\boldsymbol{B})\right]_{ij} = \begin{cases} B_{ij}, & \text{if } (i,j) \in \Omega \text{ and } j \ne l-n, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \left[\mathcal{P}_{\cdot,(l-n)}(\boldsymbol{B})\right]_{ij} = \begin{cases} B_{ij}, & \text{if } j = l-n, \\ 0, & \text{otherwise,} \end{cases}$$

for any matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$. The key ideas are: (1) the iterates are not perturbed by much when one drops a small number of samples (and hence $\boldsymbol{F}^t$ and $\boldsymbol{F}^{t,(l)}$ remain sufficiently close); (2) the auxiliary iterates $\boldsymbol{F}^{t,(l)}$ are independent of the samples directly related to the lth row of $\boldsymbol{M}$, which in turn allows to exploit certain statistical independence to control the lth row of $\boldsymbol{F}^{t,(l)}$ (and hence $\boldsymbol{F}^t$). See [MWCC17, Section 5] for a detailed explanation. Last but not least, the step size is set to be η, and we take $\boldsymbol{F}^{0,(l)} = \boldsymbol{F}^\star$ for all $1 \le l \le 2n$ (the same initialization as in Algorithm 1).

With the help of the leave-one-out sequences, we are ready to establish Lemma 5 in an inductive manner. Concretely we aim at proving that

$$\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\right\|_{\mathrm{F}} \le C_{\mathrm{F}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{X}^\star\|_{\mathrm{F}}, \tag{93a}$$

$$\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\right\| \le C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{X}^\star\|, \tag{93b}$$

$$\max_{1\le l\le 2n}\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} \le C_3\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}, \tag{93c}$$

$$\max_{1\le l\le 2n}\left\|\left(\boldsymbol{F}^{t,(l)}\boldsymbol{H}^{t,(l)} - \boldsymbol{F}^\star\right)_{l,\cdot}\right\|_2 \le C_4\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}, \tag{93d}$$

$$\left\|\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right\|_{2,\infty}\le C_\infty\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{F}^\star\right\|_{2,\infty}, \tag{93e}$$

$$\left\|\boldsymbol{X}^{t\top}\boldsymbol{X}^t-\boldsymbol{Y}^{t\top}\boldsymbol{Y}^t\right\|_{\mathrm{F}}\le C_{\mathrm{B}}\kappa\eta\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2 \tag{93f}$$

hold for all $0\le t\le t_0=n^{18}$ and for some constants $C_{\mathrm{F}},C_{\mathrm{op}},C_3,C_4,C_\infty,C_{\mathrm{B}}>0$, provided that $\eta\asymp 1/(n\kappa^3\sigma_{\max})$. In addition, we also intend to establish that

$$f\left(\boldsymbol{X}^t,\boldsymbol{Y}^t\right)\le f\left(\boldsymbol{X}^{t-1},\boldsymbol{Y}^{t-1}\right)-\frac{\eta}{2}\left\|\nabla f\left(\boldsymbol{X}^{t-1},\boldsymbol{Y}^{t-1}\right)\right\|_{\mathrm{F}}^2 \tag{94}$$

holds for all $1\le t\le t_0=n^{18}$. Here, $\boldsymbol{H}^{t,(l)}$ and $\boldsymbol{R}^{t,(l)}$ are rotation matrices defined as

$$\boldsymbol{H}^{t,(l)}\triangleq\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}-\boldsymbol{F}^\star\right\|_{\mathrm{F}}; \tag{95a}$$

$$\boldsymbol{R}^{t,(l)}\triangleq\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}-\boldsymbol{F}^t\boldsymbol{H}^t\right\|_{\mathrm{F}}. \tag{95b}$$

Note that the induction hypotheses (93a), (93b) and (93e) readily imply the statements (29a), (29b) and (29c) in Lemma 5, respectively, whereas the last bound on the size of the gradient (30) follows from (94). We summarize the last connection in the following lemma, whose proof is in Appendix D.2.

Lemma 9 (**Small gradient** (30))**.** *Set $\lambda=C_\lambda\sigma\sqrt{np}$ for some large constant $C_\lambda>0$. Suppose that the sample size obeys $n^2p\gg\kappa\mu rn\log^2 n$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\ll\frac{1}{\sqrt{\kappa^4\mu r\log n}}$. If the induction hypotheses (93) hold for all $0\le t\le t_0$ and that (94) holds for all $1\le t\le t_0$, then*

$$\min_{0\le t<t_0}\left\|\nabla f\left(\boldsymbol{X}^t,\boldsymbol{Y}^t\right)\right\|_{\mathrm{F}}\le\frac{1}{n^5}\frac{\lambda}{p}\sqrt{\sigma_{\min}},$$

as long as $\eta\asymp 1/(n\kappa^3\sigma_{\max})$.

The rest of this section is devoted to proving the hypotheses (93) and (94) via induction. We start with the base case, i.e. $t=0$. All the induction hypotheses (93) are easily verified by noting that

$$\boldsymbol{F}^0=\boldsymbol{F}^{0,(l)}=\boldsymbol{F}^\star,\qquad\text{for all } 1\le l\le 2n.$$

We now proceed to the induction step, which are demonstrated via the following lemmas. All the proofs are in subsequent subsections.

Lemma 10 (**Frobenius norm error** (93a))**.** *Set $\lambda=C_\lambda\sigma\sqrt{np}$ for some large constant $C_\lambda>0$. Suppose that the sample size obeys $n^2p\gg\kappa\mu rn\log^2 n$ and the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\ll\frac{1}{\sqrt{\kappa^4\mu r\log n}}$. If the iterates satisfy (93) at the tth iteration, then with probability at least $1-O(n^{-100})$,*

$$\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1}-\boldsymbol{F}^\star\right\|_{\mathrm{F}}\le C_{\mathrm{F}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}},$$

holds as long as $0<\eta\ll 1/(\kappa^{5/2}\sigma_{\max})$ and $C_{\mathrm{F}}>0$ is large enough.

Lemma 11 (**Spectral norm error** (93b))**.** *Set $\lambda=C_\lambda\sigma\sqrt{np}$ for some large constant $C_\lambda>0$. Suppose the sample size obeys $n^2p\gg\kappa^4\mu^2r^2n\log^2 n$ and the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\ll\frac{1}{\sqrt{\kappa^4\log n}}$. If the iterates satisfy (93) at the tth iteration, then with probability at least $1-O(n^{-100})$,*

$$\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1}-\boldsymbol{F}^\star\right\|\le C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{X}^\star\right\|$$

holds with the proviso that $0<\eta\ll 1/(\kappa^3\sigma_{\max}\sqrt{r})$ and that $C_{\mathrm{op}}\gg 1$.

Lemma 12 (**Leave-one-out perturbation** (93c)). *Set $\lambda = C_\lambda \sigma\sqrt{np}$ for some large constant $C_\lambda > 0$. Suppose that the sample size satisfies $n^2 p \gg \kappa^4 \mu^2 r^2 n \log^3 n$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^4 \mu r \log n}}$. If the iterates satisfy (93) at the tth iteration, then with probability at least $1 - O(n^{-99})$,*

$$\max_{1\le l\le 2n} \left\| \boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1} - \boldsymbol{F}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)} \right\|_{\mathrm{F}} \le C_3 \left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}} \right) \|\boldsymbol{F}^\star\|_{2,\infty}$$

holds, provided that $0 < \eta \ll 1/(\kappa^2 \sigma_{\max} n)$ and that $C_3 > 0$ is some sufficiently large constant.

Lemma 13 (ℓ_2/ℓ_∞ **norm error of leave-one-out sequences** (93d)). *Set $\lambda = C_\lambda \sigma\sqrt{np}$ for some large constant $C_\lambda > 0$. Suppose that the sample size obeys $n^2 p \gg \kappa^2 \mu^2 r^2 n \log^3 n$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2 \log n}}$. If the iterates satisfy (93) at the tth iteration, then with probability at least $1 - O(n^{-99})$,*

$$\max_{1\le l\le 2n} \left\| \left(\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t+1,(l)} - \boldsymbol{F}^\star\right)_{l,\cdot} \right\|_2 \le C_4 \kappa \left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}} \right) \|\boldsymbol{F}^\star\|_{2,\infty}$$

holds, provided that $0 < \eta \ll 1/(\kappa^2 \sqrt{r}\sigma_{\max})$, $C_{\mathrm{op}} \gg 1$ and $C_4 \gg C_{\mathrm{op}}$.

Lemma 14 (ℓ_2/ℓ_∞ **norm error** (93e)). *Set $\lambda = C_\lambda \sigma\sqrt{np}$ for some large constant $C_\lambda > 0$. Suppose that $n \ge \mu r$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2 \log n}}$. If the iterates satisfy (93) at the tth iteration, then with probability at least $1 - O(n^{-99})$,*

$$\left\| \boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1} - \boldsymbol{F}^\star \right\|_{2,\infty} \le C_\infty \kappa \left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}} \right) \|\boldsymbol{F}^\star\|_{2,\infty},$$

holds provided that $C_\infty \ge 5C_3 + C_4$.

Lemma 15 (**Approximate balancedness** (93f)). *Set $\lambda = C_\lambda \sigma\sqrt{np}$ for some large constant $C_\lambda > 0$. Suppose that the sample size satisfies $n^2 p \gg \kappa^2 \mu^2 r^2 n \log n$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2 \log n}}$. If the iterates satisfy (93) at the tth iteration, then with probability at least $1 - O(n^{-100})$,*

$$\left\| \boldsymbol{X}^{t+1\top}\boldsymbol{X}^{t+1} - \boldsymbol{Y}^{t+1\top}\boldsymbol{Y}^{t+1} \right\|_{\mathrm{F}} \le C_{\mathrm{B}} \kappa\eta \left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}} \right) \sqrt{r}\sigma_{\max}^2,$$

$$\max_{1\le l\le 2n} \left\| \boldsymbol{X}^{t+1,(l)\top}\boldsymbol{X}^{t+1,(l)} - \boldsymbol{Y}^{t+1,(l)\top}\boldsymbol{Y}^{t+1,(l)} \right\|_{\mathrm{F}} \le C_{\mathrm{B}} \kappa\eta \left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}} \right) \sqrt{r}\sigma_{\max}^2,$$

holds for some sufficiently large constant $C_{\mathrm{B}} \gg C_{\mathrm{op}}^2$, provided that $0 < \eta < 1/\sigma_{\min}$.

Lemma 16 (**Decreasing of function values** (94)). *Set $\lambda = C_\lambda \sigma\sqrt{np}$ for some large constant $C_\lambda > 0$. Suppose that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1/\sqrt{r}$. If the iterates satisfy (93) at the tth iteration, then with probability at least $1 - O(n^{-99})$,*

$$f\left(\boldsymbol{X}^{t+1}, \boldsymbol{Y}^{t+1}\right) \le f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t\right) - \frac{\eta}{2}\left\|\nabla f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t\right)\right\|_{\mathrm{F}}^2,$$

as long as $\eta \ll 1/(\kappa n \sigma_{\max})$.

D.1 Preliminaries and notations

Before proceeding to the proofs, we collect a few useful facts and notations. To begin with, for any matrix $\boldsymbol{A}$, we denote by $\boldsymbol{A}_{l,\cdot}$ (resp. $\boldsymbol{A}_{\cdot,l}$) the lth row (reps. column) of $\boldsymbol{A}$.

Define an augmented loss function $f_{\mathsf{aug}}(\boldsymbol{X}, \boldsymbol{Y})$ to be

$$f_{\mathsf{aug}}\left(\boldsymbol{X},\boldsymbol{Y}\right) \triangleq \frac{1}{2p}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{M}\right)\right\|_{\mathrm{F}}^{2}+\frac{\lambda}{2p}\left\|\boldsymbol{X}\right\|_{\mathrm{F}}^{2}+\frac{\lambda}{2p}\left\|\boldsymbol{Y}\right\|_{\mathrm{F}}^{2}+\frac{1}{8}\left\|\boldsymbol{X}^{\top}\boldsymbol{X}-\boldsymbol{Y}^{\top}\boldsymbol{Y}\right\|_{\mathrm{F}}^{2}. \tag{96}$$

As the name suggests, this new function augments the original loss function (cf. (17)) with an additional term $\|\boldsymbol{X}^{\top}\boldsymbol{X}-\boldsymbol{Y}^{\top}\boldsymbol{Y}\|_{\mathrm{F}}^{2}/8$, which is commonly used in the literature of asymmetric low-rank matrix factorization to balance the scale of $\boldsymbol{X}$ and $\boldsymbol{Y}$ [TBS+16, YPCC16, CLL19]. We emphasize that, in contrast to aforementioned works, here our gradient descent algorithm (cf. Algorithm 1) operates on $f(\cdot,\cdot)$ instead of $f_{\mathsf{aug}}(\cdot,\cdot)$. The introduction of $f_{\mathsf{aug}}(\cdot,\cdot)$ is mainly to simplify the proof.

It is easily seen that the gradients of $f_{\mathsf{aug}}(\cdot,\cdot)$ are given by

$$\nabla_{\boldsymbol{X}} f_{\mathsf{aug}}(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{M}\right)\boldsymbol{Y}+\frac{\lambda}{p}\boldsymbol{X}+\frac{1}{2}\boldsymbol{X}\left(\boldsymbol{X}^{\top}\boldsymbol{X}-\boldsymbol{Y}^{\top}\boldsymbol{Y}\right); \tag{97a}$$

$$\nabla_{\boldsymbol{Y}} f_{\mathsf{aug}}(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{M}\right)^{\top}\boldsymbol{X}+\frac{\lambda}{p}\boldsymbol{Y}+\frac{1}{2}\boldsymbol{Y}\left(\boldsymbol{Y}^{\top}\boldsymbol{Y}-\boldsymbol{X}^{\top}\boldsymbol{X}\right). \tag{97b}$$

Correspondingly, define the difference between gradients of $\nabla f(\boldsymbol{X},\boldsymbol{Y})$ and $\nabla f_{\mathsf{aug}}(\boldsymbol{X},\boldsymbol{Y})$ as follows

$$\nabla_{\boldsymbol{X}} f_{\mathsf{diff}}(\boldsymbol{X},\boldsymbol{Y}) = -\boldsymbol{X}\left(\boldsymbol{X}^{\top}\boldsymbol{X}-\boldsymbol{Y}^{\top}\boldsymbol{Y}\right)/2; \tag{98a}$$

$$\nabla_{\boldsymbol{Y}} f_{\mathsf{diff}}(\boldsymbol{X},\boldsymbol{Y}) = -\boldsymbol{Y}\left(\boldsymbol{Y}^{\top}\boldsymbol{Y}-\boldsymbol{X}^{\top}\boldsymbol{X}\right)/2, \tag{98b}$$

such that

$$\nabla_{\boldsymbol{X}} f\left(\boldsymbol{X},\boldsymbol{Y}\right) = \nabla_{\boldsymbol{X}} f_{\mathsf{aug}}\left(\boldsymbol{X},\boldsymbol{Y}\right)+\nabla_{\boldsymbol{X}} f_{\mathsf{diff}}\left(\boldsymbol{X},\boldsymbol{Y}\right); \tag{99a}$$

$$\nabla_{\boldsymbol{Y}} f\left(\boldsymbol{X},\boldsymbol{Y}\right) = \nabla_{\boldsymbol{Y}} f_{\mathsf{aug}}\left(\boldsymbol{X},\boldsymbol{Y}\right)+\nabla_{\boldsymbol{Y}} f_{\mathsf{diff}}\left(\boldsymbol{X},\boldsymbol{Y}\right). \tag{99b}$$

Regarding $\boldsymbol{F}^{\star}$, simple algebra reveals that

$$\sigma_1\left(\boldsymbol{F}^{\star}\right) = \left\|\boldsymbol{F}^{\star}\right\| = \sqrt{2\sigma_{\max}}, \qquad \sigma_r\left(\boldsymbol{F}^{\star}\right) = \sqrt{2\sigma_{\min}}, \tag{100a}$$

$$\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty} = \max\left\{\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty},\left\|\boldsymbol{Y}^{\star}\right\|_{2,\infty}\right\} \le \sqrt{\mu r\sigma_{\max}/n}, \tag{100b}$$

where the last one follows from the incoherence assumption (34).

We start with a lemma that characterizes the local geometry of the nonconvex loss function, whose proof is given in Appendix D.10.

Lemma 17. *Set $\lambda = C_{\lambda}\sigma\sqrt{np}$ for some constant $C_{\lambda} > 0$. Suppose that the sample size obeys $n^2 p \ge C\kappa\mu r n\log^2 n$ for some sufficiently large constant $C > 0$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1$. Recall the function $f_{\mathsf{aug}}(\cdot,\cdot)$ defined in (96). Then with probability at least $1-O(n^{-10})$,*

$$\mathsf{vec}\left(\boldsymbol{\Delta}\right)^{\top}\nabla^2 f_{\mathsf{aug}}\left(\boldsymbol{X},\boldsymbol{Y}\right)\mathsf{vec}\left(\boldsymbol{\Delta}\right) \ge \tfrac{1}{10}\sigma_{\min}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}^{2},$$

$$\max\left\{\left\|\nabla^2 f_{\mathsf{aug}}\left(\boldsymbol{X},\boldsymbol{Y}\right)\right\|,\left\|\nabla^2 f\left(\boldsymbol{X},\boldsymbol{Y}\right)\right\|\right\} \le 10\sigma_{\max}$$

hold uniformly over all $\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}$ obeying

$$\left\|\begin{bmatrix}\boldsymbol{X}-\boldsymbol{X}^{\star}\\ \boldsymbol{Y}-\boldsymbol{Y}^{\star}\end{bmatrix}\right\|_{2,\infty} \le \frac{1}{1000\kappa\sqrt{n}}\left\|\boldsymbol{X}^{\star}\right\|$$

and all $\boldsymbol{\Delta} = \begin{bmatrix}\boldsymbol{\Delta}_{\boldsymbol{X}}\\ \boldsymbol{\Delta}_{\boldsymbol{Y}}\end{bmatrix}\in\mathbb{R}^{2n\times r}$ lying in the set

$$\left\{\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{Y}_1\end{bmatrix}\hat{\boldsymbol{H}}-\begin{bmatrix}\boldsymbol{X}_2\\ \boldsymbol{Y}_2\end{bmatrix}\;\middle|\;\left\|\begin{bmatrix}\boldsymbol{X}_2-\boldsymbol{X}^{\star}\\ \boldsymbol{Y}_2-\boldsymbol{Y}^{\star}\end{bmatrix}\right\| \le \frac{1}{500\kappa}\left\|\boldsymbol{X}^{\star}\right\|,\hat{\boldsymbol{H}} \triangleq \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{Y}_1\end{bmatrix}\boldsymbol{R}-\begin{bmatrix}\boldsymbol{X}_2\\ \boldsymbol{Y}_2\end{bmatrix}\right\|_{\mathrm{F}}\right\}.$$

Last but not least, a few immediate consequences of (93) are gathered in the following lemma, whose proof is given in Appendix D.11.

Lemma 18. *We have the following four sets of consequences of the induction hypotheses (29).*

1. *Suppose that the sample size obeys $n \gg \mu r \log n$. If the tth iterates obey (93), then one has*

$$\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^{\star}\right\|_{2,\infty} \le (C_{\infty}\kappa + C_3)\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^{\star}\|_{2,\infty}, \tag{101a}$$

$$\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^{\star}\right\| \le 2C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{X}^{\star}\|. \tag{101b}$$

2. *Suppose that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2\log n}}$. If the tth iterates obey (93), then one has*

$$\left\|\boldsymbol{F}^{t}\boldsymbol{H}^{t}-\boldsymbol{F}^{\star}\right\| \le \|\boldsymbol{X}^{\star}\|, \quad \left\|\boldsymbol{F}^{t}\boldsymbol{H}^{t}-\boldsymbol{F}^{\star}\right\|_{\mathrm{F}} \le \|\boldsymbol{X}^{\star}\|_{\mathrm{F}}, \quad \left\|\boldsymbol{F}^{t}\boldsymbol{H}^{t}-\boldsymbol{F}^{\star}\right\|_{2,\infty} \le \|\boldsymbol{F}^{\star}\|_{2,\infty}, \tag{102a}$$

$$\left\|\boldsymbol{F}^{t}\right\| \le 2\|\boldsymbol{X}^{\star}\|, \quad \left\|\boldsymbol{F}^{t}\right\|_{\mathrm{F}} \le 2\|\boldsymbol{X}^{\star}\|_{\mathrm{F}}, \quad \left\|\boldsymbol{F}^{t}\right\|_{2,\infty} \le 2\|\boldsymbol{F}^{\star}\|_{2,\infty}. \tag{102b}$$

3. *Suppose that $n \gg \kappa^2\mu r\log n$ and that $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2\log n}}$. If the tth iterates obey (93), then we have*

$$\left\|\boldsymbol{F}^{t}\boldsymbol{H}^{t}-\boldsymbol{F}^{t,(l)}\boldsymbol{H}^{t,(l)}\right\|_{\mathrm{F}} \le 5\kappa\left\|\boldsymbol{F}^{t}\boldsymbol{H}^{t}-\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}.$$

4. *Suppose that $n \ge \kappa\mu$ and that $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2\log n}}$. If the tth iterates obey (93), then (102) also holds for $\boldsymbol{F}^{t,(l)}\boldsymbol{H}^{t,(l)}$. In addition, one has*

$$\sigma_{\min}/2 \le \sigma_{\min}\left((\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)})^{\top}\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)}\right) \le \sigma_{\max}\left((\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)})^{\top}\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)}\right) \le 2\sigma_{\max}.$$

D.2 Proof of Lemma 9

Summing (94) from $t=1$ to $t=t_0$ leads to a telescopic sum

$$f\left(\boldsymbol{X}^{t_0},\boldsymbol{Y}^{t_0}\right) \le f\left(\boldsymbol{X}^{0},\boldsymbol{Y}^{0}\right) - \frac{\eta}{2}\sum_{t=0}^{t_0-1}\left\|\nabla f\left(\boldsymbol{X}^{t},\boldsymbol{Y}^{t}\right)\right\|_{\mathrm{F}}^2.$$

This further implies that

$$\min_{0\le t<t_0}\left\|\nabla f\left(\boldsymbol{X}^{t},\boldsymbol{Y}^{t}\right)\right\|_{\mathrm{F}} \le \left\{\frac{1}{t_0}\sum_{t=0}^{t_0-1}\left\|\nabla f\left(\boldsymbol{X}^{t},\boldsymbol{Y}^{t}\right)\right\|_{\mathrm{F}}^2\right\}^{1/2} \le \left\{\frac{2}{\eta t_0}\left[f\left(\boldsymbol{X}^{\star},\boldsymbol{Y}^{\star}\right)-f\left(\boldsymbol{X}^{t_0},\boldsymbol{Y}^{t_0}\right)\right]\right\}^{1/2}, \tag{103}$$

where we have used the assumption that $(\boldsymbol{X}^0,\boldsymbol{Y}^0)=(\boldsymbol{X}^{\star},\boldsymbol{Y}^{\star})$.

It remains to control $f(\boldsymbol{X}^{\star},\boldsymbol{Y}^{\star})-f(\boldsymbol{X}^{t_0},\boldsymbol{Y}^{t_0})$. Towards this end, we can use the fact that $f(\boldsymbol{X},\boldsymbol{Y}) = f(\boldsymbol{X}\boldsymbol{R},\boldsymbol{Y}\boldsymbol{R})$ for any $\boldsymbol{R}\in\mathcal{O}^{r\times r}$ to obtain

$$f\left(\boldsymbol{F}^{t_0}\right) = f\left(\boldsymbol{F}^{t_0}\boldsymbol{H}^{t_0}\right) = f\left(\boldsymbol{F}^{\star}\right)+\left\langle\nabla f\left(\boldsymbol{F}^{\star}\right),\boldsymbol{F}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{F}^{\star}\right\rangle+\frac{1}{2}\mathsf{vec}\left(\boldsymbol{F}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{F}^{\star}\right)^{\top}\nabla^2 f\left(\tilde{\boldsymbol{F}}\right)\mathsf{vec}\left(\boldsymbol{F}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{F}^{\star}\right),$$

where $\tilde{\boldsymbol{F}}$ lies in the line segment connecting $\boldsymbol{F}^{t_0}\boldsymbol{H}^{t_0}$ and $\boldsymbol{F}^{\star}$. Apply the triangle inequality to see

$$\begin{aligned} f\left(\boldsymbol{F}^{\star}\right)-f\left(\boldsymbol{F}^{t_0}\right) &\le \left\|\nabla f\left(\boldsymbol{F}^{\star}\right)\right\|_{\mathrm{F}}\left\|\boldsymbol{F}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{F}^{\star}\right\|_{\mathrm{F}}-\frac{1}{2}\mathsf{vec}\left(\boldsymbol{F}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{F}^{\star}\right)^{\top}\nabla^2 f\left(\tilde{\boldsymbol{F}}\right)\mathsf{vec}\left(\boldsymbol{F}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{F}^{\star}\right)\\ &\le \left\|\nabla f\left(\boldsymbol{F}^{\star}\right)\right\|_{\mathrm{F}}\left\|\boldsymbol{F}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{F}^{\star}\right\|_{\mathrm{F}}+5\sigma_{\max}\left\|\boldsymbol{F}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{F}^{\star}\right\|_{\mathrm{F}}^2. \end{aligned}$$

Here the second line follows from the fact that $\|\nabla^2 f(\tilde{\boldsymbol{F}})\| \le 10\sigma_{\max}$. To see this, use (93e) to obtain that

$$\left\|\tilde{\boldsymbol{F}}-\boldsymbol{F}^{\star}\right\|_{2,\infty} \le \left\|\boldsymbol{F}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{F}^{\star}\right\|_{2,\infty} \le C_{\infty}\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^{\star}\|_{2,\infty}$$

$$\le C_\infty \kappa \left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \sqrt{\frac{\mu r}{n}} \sqrt{\sigma_{\max}}$$
$$\le \frac{1}{2000 \kappa \sqrt{n}} \sqrt{\sigma_{\max}}, \tag{104}$$

where the second line arises from the incoherence assumption (100b) and the last inequality holds as long as $\lambda \asymp \sigma\sqrt{np}$ and $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^4 \mu r \log n}}$. Apply Lemma 17 to conclude that $\|\nabla^2 f(\tilde{\boldsymbol{F}})\| \le 10\sigma_{\max}$. Recognize that

$$\begin{aligned}
\|\nabla f(\boldsymbol{F}^\star)\|_{\mathrm{F}} &\le \|\nabla_{\boldsymbol{X}} f(\boldsymbol{F}^\star)\|_{\mathrm{F}} + \|\nabla_{\boldsymbol{Y}} f(\boldsymbol{F}^\star)\|_{\mathrm{F}} \\
&\le \frac{1}{p}\|\mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{Y}^\star\|_{\mathrm{F}} + \frac{\lambda}{p}\|\boldsymbol{X}^\star\|_{\mathrm{F}} + \frac{1}{p}\left\|\mathcal{P}_\Omega(\boldsymbol{E})^\top \boldsymbol{X}^\star\right\|_{\mathrm{F}} + \frac{\lambda}{p}\|\boldsymbol{Y}^\star\|_{\mathrm{F}} \\
&\le \left(\frac{1}{p}\|\mathcal{P}_\Omega(\boldsymbol{E})\| + \frac{\lambda}{p}\right)\left(\|\boldsymbol{X}^\star\|_{\mathrm{F}} + \|\boldsymbol{Y}^\star\|_{\mathrm{F}}\right),
\end{aligned} \tag{105}$$

where we have used the fact that $\nabla_{\boldsymbol{X}} f(\boldsymbol{F}^\star) = \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top} - \boldsymbol{M}^\star - \boldsymbol{E})\boldsymbol{Y}^\star + \frac{\lambda}{p}\boldsymbol{X}^\star = -\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{Y}^\star + \frac{\lambda}{p}\boldsymbol{X}^\star$ (similar expression holds true for $\nabla_{\boldsymbol{Y}} f(\boldsymbol{F}^\star)$). This together with Lemma 3 and the assumption that $\lambda \asymp \sigma\sqrt{np}$ yields

$$\|\nabla f(\boldsymbol{F}^\star)\|_{\mathrm{F}} \lesssim \left(\sigma\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\sqrt{r\sigma_{\max}} \asymp \frac{\lambda}{p}\sqrt{r\sigma_{\max}}. \tag{106}$$

The above bounds together with the induction hypothesis (93a) for $t = t_0$ give

$$\begin{aligned}
f(\boldsymbol{F}^\star) - f\left(\boldsymbol{F}^{t_0}\right) &\lesssim \frac{\lambda}{p}\sqrt{r\sigma_{\max}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\|\boldsymbol{X}^\star\|_{\mathrm{F}} + \sigma_{\max}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)^2\|\boldsymbol{X}^\star\|_{\mathrm{F}}^2 \\
&\lesssim r\kappa^2\left(\frac{\lambda}{p}\right)^2,
\end{aligned}$$

where the last relation arises from $\sigma\sqrt{np} \asymp \lambda$. Substitution into (103) results in

$$\min_{0\le t<t_0}\left\|\nabla f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t\right)\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{1}{\eta t_0} r\kappa^2\left(\frac{\lambda}{p}\right)^2} \le \frac{1}{n^5}\frac{\lambda}{p}\sqrt{\sigma_{\min}},$$

provided that $\eta \asymp 1/(n\kappa^3\sigma_{\max})$, $t_0 = n^{18}$ and that $n \ge \kappa$, which is a consequence of our sample complexity $n \ge np \gg \kappa\mu r \log^2 n$.

D.3 Proof of Lemma 10

From the definitions of $\boldsymbol{H}^{t+1}$ (cf. (89)), ∇f_{aug} (cf. (97)) and ∇f_{diff} (cf. (98)), we have

$$\begin{aligned}
\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1} - \boldsymbol{F}^\star\right\|_{\mathrm{F}} &\le \left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^t - \boldsymbol{F}^\star\right\|_{\mathrm{F}} = \left\|\left[\boldsymbol{F}^t - \eta\nabla f\left(\boldsymbol{F}^t\right)\right]\boldsymbol{H}^t - \boldsymbol{F}^\star\right\|_{\mathrm{F}} \\
&\overset{\text{(i)}}{=} \left\|\boldsymbol{F}^t\boldsymbol{H}^t - \eta\nabla f\left(\boldsymbol{F}^t\boldsymbol{H}^t\right) - \boldsymbol{F}^\star\right\|_{\mathrm{F}} \\
&\overset{\text{(ii)}}{\le} \underbrace{\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \eta\nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^t\boldsymbol{H}^t\right) - \left[\boldsymbol{F}^\star - \eta\nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^\star\right)\right]\right\|_{\mathrm{F}}}_{:=\alpha_1} + \underbrace{\eta\left\|\nabla f_{\mathsf{diff}}\left(\boldsymbol{F}^t\boldsymbol{H}^t\right)\right\|_{\mathrm{F}}}_{:=\alpha_2} + \underbrace{\eta\left\|\nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^\star\right)\right\|_{\mathrm{F}}}_{:=\alpha_3}.
\end{aligned}$$

Here (i) uses the fact that $\nabla f(\boldsymbol{F}\boldsymbol{R}) = \nabla f(\boldsymbol{F})\boldsymbol{R}$ for all $\boldsymbol{R} \in \mathcal{O}^{r\times r}$; the last relation (ii) uses the decomposition (99) and the triangle inequality. In the following, we bound α_1, α_2 and α_3 in the reverse order.

1. First, regarding α_3, since $\boldsymbol{X}^{\star\top}\boldsymbol{X}^\star = \boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star$, one has $\eta\|\nabla f(\boldsymbol{F}^\star)\|_{\mathrm{F}} = \eta\|\nabla f_{\mathsf{aug}}(\boldsymbol{F}^\star)\|_{\mathrm{F}}$. Repeating our arguments for (105) and (106) gives

$$\alpha_3 = \eta\left\|\nabla f\left(\boldsymbol{F}^\star\right)\right\|_{\mathrm{F}} \le 4\eta\frac{\lambda}{p}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}$$

as long as $\lambda \asymp \sigma\sqrt{np}$. Here the last inequality also relies on the fact that $\|\boldsymbol{X}^\star\|_{\mathrm{F}} = \|\boldsymbol{Y}^\star\|_{\mathrm{F}}$.

2. We now move on to α_2, for which one has

$$\begin{aligned}\alpha_2 &\le \frac{\eta}{2}\left(\left\|\boldsymbol{X}^t\left(\boldsymbol{X}^{t\top}\boldsymbol{X}^t-\boldsymbol{Y}^{t\top}\boldsymbol{Y}^t\right)\boldsymbol{H}^t\right\|_{\mathrm{F}}+\left\|\boldsymbol{Y}^t\left(\boldsymbol{Y}^{t\top}\boldsymbol{Y}^t-\boldsymbol{X}^{t\top}\boldsymbol{X}^t\right)\boldsymbol{H}^t\right\|_{\mathrm{F}}\right)\\ &\le \frac{\eta}{2}\left(\left\|\boldsymbol{X}^t\right\|+\left\|\boldsymbol{Y}^t\right\|\right)\left\|\boldsymbol{X}^{t\top}\boldsymbol{X}^t-\boldsymbol{Y}^{t\top}\boldsymbol{Y}^t\right\|_{\mathrm{F}}.\end{aligned}$$

Utilize the fact that $\max\{\|\boldsymbol{X}^t\|,\|\boldsymbol{Y}^t\|\}\le\|\boldsymbol{F}^\star\|\le 2\|\boldsymbol{X}^\star\|$ (see Lemma 18) and the induction hypothesis (93f) to obtain

$$\begin{aligned}\alpha_2 &\le 2\eta\sqrt{\sigma_{\max}}\cdot C_{\mathrm{B}}\kappa\eta\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2\\ &\le \left(2C_{\mathrm{B}}\kappa^{5/2}\eta\sigma_{\max}\right)\sigma_{\min}\eta\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{X}^\star\|_{\mathrm{F}}\\ &\le \sigma_{\min}\eta\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{X}^\star\|_{\mathrm{F}},\end{aligned}$$

where the second inequality uses $\|\boldsymbol{X}^\star\|_{\mathrm{F}}\ge\sqrt{r\sigma_{\min}}$ and the last one holds as long as $2C_{\mathrm{B}}\kappa^{5/2}\sigma_{\max}\eta\le 1$.

3. In the end, for α_1, the fundamental theorem of calculus [Lan93, Chapter XIII, Theorem 4.2] reveals that

$$\begin{aligned}&\mathsf{vec}\left[\boldsymbol{F}^t\boldsymbol{H}^t-\eta\nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^t\boldsymbol{H}^t\right)-\left[\boldsymbol{F}^\star-\eta\nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^\star\right)\right]\right]\\ &\quad=\mathsf{vec}\left[\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right]-\eta\cdot\mathsf{vec}\left[\nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^t\boldsymbol{H}^t\right)-\nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^\star\right)\right]\\ &\quad=\Bigg(\boldsymbol{I}_{2nr}-\eta\underbrace{\int_0^1\nabla^2 f_{\mathsf{aug}}\left(\boldsymbol{F}(\tau)\right)\mathrm{d}\tau}_{:=\boldsymbol{A}}\Bigg)\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right),\end{aligned}\tag{107}$$

where we denote $\boldsymbol{F}(\tau)\triangleq\boldsymbol{F}^\star+\tau(\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star)$ for all $0\le\tau\le 1$. Taking the squared Euclidean norm of both sides of the equality (107) leads to

$$\begin{aligned}\alpha_1^2&=\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right)^\top\left(\boldsymbol{I}_{2nr}-\eta\boldsymbol{A}\right)^2\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right)\\ &=\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right)^\top\left(\boldsymbol{I}_{2nr}-2\eta\boldsymbol{A}+\eta^2\boldsymbol{A}^2\right)\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right)\\ &\le\left\|\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right\|_{\mathrm{F}}^2+\eta^2\|\boldsymbol{A}\|^2\left\|\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right\|_{\mathrm{F}}^2-2\eta\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right)^\top\boldsymbol{A}\,\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right),\end{aligned}\tag{108}$$

where (108) results from the fact that

$$\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right)^\top\boldsymbol{A}^2\,\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right)\le\|\boldsymbol{A}\|^2\left\|\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right\|_{\mathrm{F}}^2.$$

Applying the same argument as in (104), one gets for all $0\le\tau\le 1$, $\|\boldsymbol{F}(\tau)-\boldsymbol{F}^\star\|_{2,\infty}\le\frac{1}{2000\kappa\sqrt{n}}\|\boldsymbol{X}^\star\|$. Invoke Lemma 17 with $\boldsymbol{X}=\boldsymbol{X}^\star+\tau(\boldsymbol{X}^t\boldsymbol{H}^t-\boldsymbol{X}^\star)$, $\boldsymbol{Y}=\boldsymbol{Y}^\star+\tau(\boldsymbol{Y}^t\boldsymbol{H}^t-\boldsymbol{Y}^\star)$, $(\boldsymbol{X}_1,\boldsymbol{Y}_1)=(\boldsymbol{X}^t,\boldsymbol{Y}^t)$ and $(\boldsymbol{X}_2,\boldsymbol{Y}_2)=(\boldsymbol{X}^\star,\boldsymbol{Y}^\star)$ to obtain $\|\boldsymbol{A}\|\le 10\sigma_{\max}$ and

$$\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right)^\top\boldsymbol{A}\,\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right)\ge\frac{1}{10}\sigma_{\min}\left\|\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right\|_{\mathrm{F}}^2.$$

Putting these two bounds back to (108) yields

$$\alpha_1^2\le\left(1+100\eta^2\sigma_{\max}^2-\frac{1}{5}\eta\sigma_{\min}\right)\left\|\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right\|_{\mathrm{F}}^2\le\left(1-\frac{\sigma_{\min}}{10}\eta\right)\left\|\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right\|_{\mathrm{F}}^2.$$

Here the last relation holds as long as $0\le\eta\le 1/(1000\kappa\sigma_{\max})$. As a result, we have

$$\alpha_1\le\left(1-\frac{\sigma_{\min}}{20}\eta\right)\left\|\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right\|_{\mathrm{F}}.$$

Combine the above bounds on α_1, α_2 and α_3 to conclude that

$$\begin{aligned}
&\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1}-\boldsymbol{F}^{\star}\right\|_{\mathrm{F}} \le \left(1-\frac{\sigma_{\min}}{20}\eta\right)\left\|\boldsymbol{F}^{t}\boldsymbol{H}^{t}-\boldsymbol{F}^{\star}\right\|_{\mathrm{F}} + 4\eta\frac{\lambda}{p}\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} + \eta\sigma_{\min}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} \\
&\le \left(1-\frac{\sigma_{\min}}{20}\eta\right)C_{\mathrm{F}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} + 4\eta\sigma_{\min}\frac{\lambda}{p\,\sigma_{\min}}\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} + \eta\sigma_{\min}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} \\
&\le C_{\mathrm{F}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}},
\end{aligned}$$

provided that $C_{\mathrm{F}} > 0$ is large enough.

D.4 Proof of Lemma 11

To facilitate analysis, we define an auxiliary point $\tilde{\boldsymbol{F}}^{t+1} \triangleq \begin{bmatrix} \tilde{\boldsymbol{X}}^{t+1} \\ \tilde{\boldsymbol{Y}}^{t+1} \end{bmatrix}$ as

$$\tilde{\boldsymbol{X}}^{t+1} = \boldsymbol{X}^{t}\boldsymbol{H}^{t} - \eta\left[\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{t}\boldsymbol{Y}^{t\top}-\boldsymbol{M}^{\star}-\boldsymbol{E}\right)\boldsymbol{Y}^{\star} + \frac{\lambda}{p}\boldsymbol{X}^{\star} + \frac{1}{2}\boldsymbol{X}^{\star}\boldsymbol{H}^{t\top}\left(\boldsymbol{X}^{t\top}\boldsymbol{X}^{t}-\boldsymbol{Y}^{t\top}\boldsymbol{Y}^{t}\right)\boldsymbol{H}^{t}\right]; \quad (109a)$$

$$\tilde{\boldsymbol{Y}}^{t+1} = \boldsymbol{Y}^{t}\boldsymbol{H}^{t} - \eta\left[\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{t}\boldsymbol{Y}^{t\top}-\boldsymbol{M}^{\star}-\boldsymbol{E}\right)^{\top}\boldsymbol{X}^{\star} + \frac{\lambda}{p}\boldsymbol{Y}^{\star} + \frac{1}{2}\boldsymbol{Y}^{\star}\boldsymbol{H}^{t\top}\left(\boldsymbol{Y}^{t\top}\boldsymbol{Y}^{t}-\boldsymbol{X}^{t\top}\boldsymbol{X}^{t}\right)\boldsymbol{H}^{t}\right]. \quad (109b)$$

Then the triangle inequality tells us that

$$\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1}-\boldsymbol{F}^{\star}\right\| \le \underbrace{\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1}-\tilde{\boldsymbol{F}}^{t+1}\right\|}_{:=\alpha_1} + \underbrace{\left\|\tilde{\boldsymbol{F}}^{t+1}-\boldsymbol{F}^{\star}\right\|}_{:=\alpha_2}. \quad (110)$$

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

1. We start with α_2. By the triangle inequality again we have

$$\begin{aligned}
\alpha_2 \le &\underbrace{\left\|\begin{bmatrix} \boldsymbol{X}^{t}\boldsymbol{H}^{t}-\eta\left[\left(\boldsymbol{X}^{t}\boldsymbol{Y}^{t\top}-\boldsymbol{M}^{\star}\right)\boldsymbol{Y}^{\star}+\frac{1}{2}\boldsymbol{X}^{\star}\boldsymbol{H}^{t\top}\left(\boldsymbol{X}^{t\top}\boldsymbol{X}^{t}-\boldsymbol{Y}^{t\top}\boldsymbol{Y}^{t}\right)\boldsymbol{H}^{t}\right]-\boldsymbol{X}^{\star} \\ \boldsymbol{Y}^{t}\boldsymbol{H}^{t}-\eta\left[\left(\boldsymbol{X}^{t}\boldsymbol{Y}^{t\top}-\boldsymbol{M}^{\star}\right)^{\top}\boldsymbol{X}^{\star}+\frac{1}{2}\boldsymbol{Y}^{\star}\boldsymbol{H}^{t\top}\left(\boldsymbol{Y}^{t\top}\boldsymbol{Y}^{t}-\boldsymbol{X}^{t\top}\boldsymbol{X}^{t}\right)\boldsymbol{H}^{t}\right]-\boldsymbol{Y}^{\star}\end{bmatrix}\right\|}_{:=\beta_1} \\
&+ \underbrace{\frac{\eta}{p}\left\|\begin{bmatrix}\mathcal{P}_{\Omega}(\boldsymbol{E})\boldsymbol{Y}^{\star} \\ \mathcal{P}_{\Omega}(\boldsymbol{E})^{\top}\boldsymbol{X}^{\star}\end{bmatrix}\right\| + \eta\frac{\lambda}{p}\left\|\begin{bmatrix}\boldsymbol{X}^{\star} \\ \boldsymbol{Y}^{\star}\end{bmatrix}\right\|}_{:=\beta_2} + \underbrace{\eta\left\|\begin{bmatrix}\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{t}\boldsymbol{Y}^{t\top}-\boldsymbol{M}^{\star}\right)\boldsymbol{Y}^{\star}-\left(\boldsymbol{X}^{t}\boldsymbol{Y}^{t\top}-\boldsymbol{M}^{\star}\right)\boldsymbol{Y}^{\star} \\ \frac{1}{p}\left[\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{t}\boldsymbol{Y}^{t\top}-\boldsymbol{M}^{\star}\right)\right]^{\top}\boldsymbol{X}^{\star}-\left(\boldsymbol{X}^{t}\boldsymbol{Y}^{t\top}-\boldsymbol{M}^{\star}\right)^{\top}\boldsymbol{X}^{\star}\end{bmatrix}\right\|}_{:=\beta_3}
\end{aligned}$$

   Denote $\boldsymbol{\Delta}^{t} \triangleq \boldsymbol{F}^{t}\boldsymbol{H}^{t}-\boldsymbol{F}^{\star} = \begin{bmatrix}\boldsymbol{\Delta}_{\boldsymbol{X}}^{t} \\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^{t}\end{bmatrix}$. The term β_1 is the same as the term α_2 in [CLL19, Section 4.2]. Therefore we can adopt the bound therein to obtain

$$\beta_1 \le (1-\eta\sigma_{\min})\left\|\boldsymbol{\Delta}^{t}\right\| + 4\eta\left\|\boldsymbol{\Delta}^{t}\right\|^{2}\left\|\boldsymbol{X}^{\star}\right\|.$$

   Moving to β_2, one has

$$\beta_2 = \eta\left\|\begin{bmatrix}\frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{E}) & \boldsymbol{0} \\ \boldsymbol{0} & \frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{E})^{\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{Y}^{\star} \\ \boldsymbol{X}^{\star}\end{bmatrix}\right\| + \eta\frac{\lambda}{p}\left\|\boldsymbol{F}^{\star}\right\| \le \frac{\eta}{p}\left\|\mathcal{P}_{\Omega}(\boldsymbol{E})\right\|\left\|\boldsymbol{F}^{\star}\right\| + \eta\frac{\lambda}{p}\left\|\boldsymbol{F}^{\star}\right\| \le C\eta\left(\sigma\sqrt{\frac{n}{p}}+\frac{\lambda}{p}\right)\left\|\boldsymbol{X}^{\star}\right\|$$

   for some constant $C > 0$. Here the last inequality arises from Lemma 3 and the fact that $\|\boldsymbol{F}^{\star}\| = \sqrt{2}\|\boldsymbol{X}^{\star}\|$ (cf. (100a)). We are now left with the term β_3, which is exactly the term α_1 in [CLL19, Section 4.2]. Reusing their results, we have

$$\begin{aligned}
\beta_3 &\le \frac{2\eta}{p}\left\|\boldsymbol{X}^{\star}\right\|\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{1}\boldsymbol{1}^{\top}\right)-p\boldsymbol{1}\boldsymbol{1}^{\top}\right\|\left(\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}^{t}\right\|_{2,\infty}\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t}\right\|_{2,\infty} + \left\|\boldsymbol{\Delta}_{\boldsymbol{X}}^{t}\right\|_{2,\infty}\left\|\boldsymbol{Y}^{\star}\right\|_{2,\infty} + \left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t}\right\|_{2,\infty}\right) \\
&\lesssim \eta\sqrt{\frac{n}{p}}\left\|\boldsymbol{\Delta}^{t}\right\|_{2,\infty}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{X}^{\star}\right\|.
\end{aligned}$$

The last line follows from the facts that $\|\mathcal{P}_\Omega(\mathbf{1}\mathbf{1}^\top) - p\mathbf{1}\mathbf{1}^\top\| \lesssim \sqrt{np}$ (see [KMO10a, Lemma 3.2]) and that $\max\{\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|_{2,\infty}, \|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\|_{2,\infty}\} \le \|\boldsymbol{F}^\star\|_{2,\infty}$, provided that $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2 \log n}}$ (see Lemma 18). Combining the above three bounds gives

$$\begin{aligned}\alpha_2 &\le (1-\eta\sigma_{\min})\left\|\boldsymbol{\Delta}^t\right\| + 4\eta\left\|\boldsymbol{\Delta}^t\right\|^2\|\boldsymbol{X}^\star\| + \tilde{C}\eta\left(\sigma\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\|\boldsymbol{X}^\star\| + \tilde{C}\eta\sqrt{\frac{n}{p}}\left\|\boldsymbol{\Delta}^t\right\|_{2,\infty}\|\boldsymbol{F}^\star\|_{2,\infty}\|\boldsymbol{X}^\star\| \\ &\le \left(1-\frac{\eta}{2}\sigma_{\min}\right)\left\|\boldsymbol{\Delta}^t\right\| + \tilde{C}\eta\left(\sigma\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\|\boldsymbol{X}^\star\| + \tilde{C}\eta\sqrt{\frac{n}{p}}\left\|\boldsymbol{\Delta}^t\right\|_{2,\infty}\|\boldsymbol{F}^\star\|_{2,\infty}\|\boldsymbol{X}^\star\| \end{aligned} \tag{111}$$

for some sufficiently large constant $\tilde{C} > 0$. Here the second inequality arises from the condition

$$4\left\|\boldsymbol{\Delta}^t\right\|\|\boldsymbol{X}^\star\| \le \sigma_{\min}/2,$$

which would hold if $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\kappa}$. An immediate consequence of (111) is that

$$\alpha_2 = \left\|\tilde{\boldsymbol{F}}^{t+1} - \boldsymbol{F}^\star\right\| \le (\sqrt{2}\kappa)^{-1}\|\boldsymbol{X}^\star\|. \tag{112}$$

To see this, apply the induction hypotheses (93b) and (93e) to get

$$\begin{aligned}\alpha_2 &\le \left(1-\frac{\eta\sigma_{\min}}{2}\right)C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\|\boldsymbol{X}^\star\| + \tilde{C}\eta\left(\sigma\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\|\boldsymbol{X}^\star\| \\ &\quad + \tilde{C}\eta\sqrt{\frac{n}{p}}C_\infty\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}^2\|\boldsymbol{X}^\star\| \\ &\overset{\text{(i)}}{\le} C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\|\boldsymbol{X}^\star\| \\ &\overset{\text{(ii)}}{\le} (\sqrt{2}\kappa)^{-1}\|\boldsymbol{X}^\star\|. \end{aligned} \tag{113}$$

Here (i) holds under the assumptions that $C_{\mathrm{op}} \gg \tilde{C} + \tilde{C}C_\infty\kappa^2\sqrt{\frac{\mu^2 r^2\log n}{np}}$ and that $\|\boldsymbol{F}^\star\|_{2,\infty} \le \sqrt{\frac{\mu r}{n}}\|\boldsymbol{X}^\star\|$ (cf. (100b)); (ii) arises since $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1/\kappa$ and $\lambda \asymp \sigma\sqrt{np}$. Under the sample complexity $n^2 p \gg \kappa^4\mu^2 r^2 n\log n$, the first condition can be simplified to $C_{\mathrm{op}} \gg 2\tilde{C} \gg 1$.

2. Next we bound α_1, towards which we first observe that

$$\alpha_1 = \left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1} - \tilde{\boldsymbol{F}}^{t+1}\right\| = \left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^t\boldsymbol{H}^{t\top}\boldsymbol{H}^{t+1} - \tilde{\boldsymbol{F}}^{t+1}\right\|.$$

It is straightforward to verify that $\boldsymbol{H}^{t\top}\boldsymbol{H}^{t+1}$ is the best rotation matrix to align $\boldsymbol{F}^{t+1}\boldsymbol{H}^t$ and $\boldsymbol{F}^\star$ (in the sense of (89)). Regarding $\tilde{\boldsymbol{F}}^{t+1}$, we obtain the following claim, which demonstrates that it is already aligned with $\boldsymbol{F}^\star$, i.e. $\boldsymbol{I}_r$ is the best rotation matrix to align $\tilde{\boldsymbol{F}}^{t+1}$ and $\boldsymbol{F}^\star$.

Claim 4. *Suppose (113) holds true, one has*

$$\boldsymbol{I}_r = \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\tilde{\boldsymbol{F}}^{t+1}\boldsymbol{R} - \boldsymbol{F}^\star\right\|_{\mathrm{F}}.$$

Now we intend to apply Lemma 22 with

$$\boldsymbol{F}_0 = \boldsymbol{F}^\star, \quad \boldsymbol{F}_1 = \tilde{\boldsymbol{F}}^{t+1}, \quad \boldsymbol{F}_2 = \boldsymbol{F}^{t+1}\boldsymbol{H}^t,$$

for which we need to check the two conditions therein. First, in view of (112), one has

$$\|\boldsymbol{F}_1 - \boldsymbol{F}_0\|\|\boldsymbol{F}_0\| = \left\|\tilde{\boldsymbol{F}}^{t+1} - \boldsymbol{F}^\star\right\|\|\boldsymbol{F}^\star\| \le \frac{1}{\sqrt{2}\kappa}\|\boldsymbol{X}^\star\|\|\boldsymbol{F}^\star\| = \sigma_{\min} = \frac{1}{2}\sigma_r^2(\boldsymbol{F}_0).$$

Second, making use of the gradient update rules (27) and the decomposition (99), we obtain

$$\begin{aligned}\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^t-\tilde{\boldsymbol{F}}^{t+1}\right\| &= \left\|\left(\boldsymbol{F}^t-\eta\nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^t\right)-\eta\nabla f_{\mathsf{diff}}\left(\boldsymbol{F}^t\right)\right)\boldsymbol{H}^t-\tilde{\boldsymbol{F}}^{t+1}\right\| \\ &\le \underbrace{\left\|\left(\boldsymbol{F}^t-\eta\nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^t\right)\right)\boldsymbol{H}^t-\tilde{\boldsymbol{F}}^{t+1}\right\|}_{:=\theta_1}+\underbrace{\eta\left\|\nabla f_{\mathsf{diff}}\left(\boldsymbol{F}^t\right)\right\|}_{:=\theta_2}.\end{aligned}$$

The term θ_2 has been controlled as α_2 in the proof of Lemma 10, where we obtained

$$\theta_2 \le 2C_{\mathrm{B}}\kappa\eta^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2\left\|\boldsymbol{X}^\star\right\|.$$

We now move on to θ_1, for which we have

$$\begin{aligned}\theta_1 \le \underbrace{\left\|\left(\boldsymbol{F}^t-\eta\left\{\nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^t\right)-\eta\begin{bmatrix}\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}^t\\ \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)^\top\boldsymbol{X}^t\end{bmatrix}-\eta\frac{\lambda}{p}\begin{bmatrix}\boldsymbol{X}^t\\ \boldsymbol{Y}^t\end{bmatrix}\right\}\right)\boldsymbol{H}^t \right.}_{:=\xi_1} \\ \underbrace{\left.-\tilde{\boldsymbol{F}}^{t+1}-\eta\begin{bmatrix}\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}^\star\\ \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)^\top\boldsymbol{X}^\star\end{bmatrix}-\eta\frac{\lambda}{p}\begin{bmatrix}\boldsymbol{X}^\star\\ \boldsymbol{Y}^\star\end{bmatrix}\right\|}_{:=\xi_1} \\ +\underbrace{\eta\left\|\begin{bmatrix}\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}^t\\ \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)^\top\boldsymbol{X}^t\end{bmatrix}\boldsymbol{H}^t+\frac{\lambda}{p}\begin{bmatrix}\boldsymbol{X}^t\\ \boldsymbol{Y}^t\end{bmatrix}\boldsymbol{H}^t-\begin{bmatrix}\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\boldsymbol{Y}^\star\\ \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)^\top\boldsymbol{X}^\star\end{bmatrix}-\frac{\lambda}{p}\begin{bmatrix}\boldsymbol{X}^\star\\ \boldsymbol{Y}^\star\end{bmatrix}\right\|}_{:=\xi_2}.\end{aligned}$$

Combining [CLL19, Equation (4.13)] and [KMO10a, Lemma 3.2] yields

$$\begin{aligned}\xi_1 &\lesssim \eta\sqrt{\frac{n}{p}}\left(\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\right\|_{2,\infty}\left\|\boldsymbol{Y}^\star\right\|_{2,\infty}+\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}+\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\right\|_{2,\infty}\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\right\|_{2,\infty}\right)\left\|\boldsymbol{\Delta}^t\right\| \\ &\quad+\eta\left(\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\right\|\left\|\boldsymbol{Y}^\star\right\|+\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\right\|\left\|\boldsymbol{X}^\star\right\|+\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\right\|\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\right\|+2\left\|\boldsymbol{X}^\star\right\|\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\right\|+2\left\|\boldsymbol{Y}^\star\right\|\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\right\|+\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\right\|^2+\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^t\right\|^2\right)\left\|\boldsymbol{\Delta}_t\right\| \\ &\lesssim \eta\sqrt{\frac{n}{p}}\left\|\boldsymbol{\Delta}^t\right\|_{2,\infty}\left\|\boldsymbol{F}^\star\right\|_{2,\infty}\left\|\boldsymbol{\Delta}^t\right\|+\eta\left\|\boldsymbol{\Delta}^t\right\|^2\left\|\boldsymbol{X}^\star\right\| \\ &\le \frac{1}{15\kappa}\frac{\sigma_{\min}}{4}\eta\left\|\boldsymbol{\Delta}^t\right\|.\end{aligned}$$

Here the penultimate inequality arises from the facts that $\max\{\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|_{2,\infty},\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|_{2,\infty}\} \le \|\boldsymbol{\Delta}^t\|_{2,\infty} \le \|\boldsymbol{F}^\star\|_{2,\infty}$ and similarly $\max\{\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|,\|\boldsymbol{\Delta}_{\boldsymbol{X}}^t\|\} \le \|\boldsymbol{\Delta}^t\| \le \|\boldsymbol{X}^\star\|$; see Lemma 18. In addition, the last line holds because of the induction hypotheses (93b) and (93e), provided that

$$C_\infty\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\sqrt{\frac{\mu^2r^2\log n}{np}} \ll \frac{1}{\kappa^2} \qquad \text{and} \qquad C_{\mathrm{op}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\kappa^2}.$$

Again, the first condition would be guaranteed by the sample size condition $n^2p \gg \kappa^4\mu^2r^2n\log n$ and the noise condition $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1/\kappa$. Next, the term ξ_2 can be easily controlled as follows

$$\begin{aligned}\xi_2 &\le \eta\left\|\begin{bmatrix}\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{E})\left(\boldsymbol{Y}^t\boldsymbol{H}^t-\boldsymbol{Y}^\star\right)\\ \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{E})^\top\left(\boldsymbol{X}^t\boldsymbol{H}^t-\boldsymbol{X}^\star\right)\end{bmatrix}\right\|+\eta\frac{\lambda}{p}\left\|\boldsymbol{F}^t\boldsymbol{H}^t-\boldsymbol{F}^\star\right\| \\ &\le \tilde{C}\eta\left(\sigma\sqrt{\frac{n}{p}}+\frac{\lambda}{p}\right)\left\|\boldsymbol{\Delta}^t\right\|,\end{aligned}$$

where the last line follows from the same argument for bounding β_2 above. Taking the bounds on θ_1 and θ_2 collectively yields

$$\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^t-\tilde{\boldsymbol{F}}^{t+1}\right\| \le \frac{1}{15\kappa}\frac{\sigma_{\min}}{4}\eta\left\|\boldsymbol{\Delta}^t\right\|+\tilde{C}\eta\left(\sigma\sqrt{\frac{n}{p}}+\frac{\lambda}{p}\right)\left\|\boldsymbol{\Delta}^t\right\|+2C_{\mathrm{B}}\kappa\eta^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2\left\|\boldsymbol{X}^\star\right\|$$

$$\leq \frac{1}{5\kappa}\frac{\sigma_{\min}}{4}\eta\left\|\boldsymbol{\Delta}^t\right\| + 2C_{\mathrm{B}}\kappa\eta^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2\left\|\boldsymbol{X}^\star\right\|. \tag{114}$$

The final inequality is true as long as $\lambda \asymp \sigma\sqrt{np}$ and $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\kappa}$. An immediate consequence of (114) is that

$$\left\|\tilde{\boldsymbol{F}}^{t+1} - \boldsymbol{F}^{t+1}\boldsymbol{H}^t\right\| \leq (2\sqrt{2}\kappa)^{-1}\left\|\boldsymbol{X}^\star\right\|, \tag{115}$$

as long as $\eta \ll 1/(C_{\mathrm{B}}\kappa^2\sigma_{\max}\sqrt{r})$, $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\kappa}$ and $\lambda \asymp \sigma\sqrt{np}$. As a result, one obtains

$$\left\|\boldsymbol{F}_1 - \boldsymbol{F}_2\right\|\left\|\boldsymbol{F}_0\right\| = \left\|\tilde{\boldsymbol{F}}^{t+1} - \boldsymbol{F}^{t+1}\boldsymbol{H}^t\right\|\left\|\boldsymbol{F}^\star\right\| \leq (2\sqrt{2}\kappa)^{-1}\left\|\boldsymbol{X}^\star\right\|\left\|\boldsymbol{F}^\star\right\| = \sigma_{\min}/2 = \sigma_{\min}^2(\boldsymbol{F}_0)/4.$$

Armed with these two conditions, we can invoke Lemma 22 to obtain

$$\begin{aligned}
\alpha_1 = \left\|\tilde{\boldsymbol{F}}^{t+1} - \boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1}\right\| &\leq 5\kappa\left\|\tilde{\boldsymbol{F}}^{t+1} - \boldsymbol{F}^{t+1}\boldsymbol{H}^t\right\| \\
&\leq \frac{1}{4}\sigma_{\min}\eta\left\|\boldsymbol{\Delta}^t\right\| + 10C_{\mathrm{B}}\kappa^2\eta^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2\left\|\boldsymbol{X}^\star\right\| \\
&\leq \frac{1}{4}\sigma_{\min}\eta\left\|\boldsymbol{\Delta}^t\right\| + \eta\left(\sigma\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\left\|\boldsymbol{X}^\star\right\|,
\end{aligned}$$

provided that $\eta \ll 1/(C_{\mathrm{B}}\kappa^3\sigma_{\max}\sqrt{r})$.

Combine the bounds on α_1 and α_2 to reach

$$\begin{aligned}
&\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1} - \boldsymbol{F}^\star\right\| \\
&\quad\leq \left(1 - \frac{\eta}{2}\sigma_{\min}\right)\left\|\boldsymbol{\Delta}^t\right\| + \left(\tilde{C}+1\right)\eta\left(\sigma\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\left\|\boldsymbol{X}^\star\right\| + \tilde{C}\eta\sqrt{\frac{n}{p}}\left\|\boldsymbol{\Delta}^t\right\|_{2,\infty}\left\|\boldsymbol{F}^\star\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\| + \frac{\sigma_{\min}}{4}\eta\left\|\boldsymbol{\Delta}^t\right\| \\
&\quad\leq \left(1 - \frac{\eta}{4}\sigma_{\min}\right)C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{X}^\star\right\| + \left(\tilde{C}+1\right)\eta\left(\sigma\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\left\|\boldsymbol{X}^\star\right\| \\
&\qquad + \tilde{C}\eta\sqrt{\frac{n}{p}}C_\infty\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{F}^\star\right\|_{2,\infty}^2\left\|\boldsymbol{X}^\star\right\| \\
&\quad\leq C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{X}^\star\right\|,
\end{aligned}$$

with the proviso that $C_{\mathrm{op}} \gg 1$ and $n^2 p \gg \kappa^4\mu^2 r^2 n\log n$. Here the last line follows from the same argument as in bounding (113). This completes the proof.

Proof of Claim 4. In view of [MWCC17, Lemma 35], it suffices to show that $\boldsymbol{F}^{\star\top}\tilde{\boldsymbol{F}}^{t+1}$ is symmetric and positive semidefinite. Recognizing that $\boldsymbol{F}^{\star\top}\boldsymbol{F}^t\boldsymbol{H}^t = (\boldsymbol{X}^{\star\top}\boldsymbol{X}^t + \boldsymbol{Y}^{\star\top}\boldsymbol{Y}^t)\boldsymbol{H}^t$ is symmetric (see [MWCC17, Lemma 35]), it is straightforward to verify that $\boldsymbol{F}^{\star\top}\tilde{\boldsymbol{F}}^{t+1}$ is also symmetric (which we omit here for brevity). In addition, by (112) we have

$$\left\|\boldsymbol{F}^{\star\top}\tilde{\boldsymbol{F}}^{t+1} - \boldsymbol{F}^{\star\top}\boldsymbol{F}^\star\right\| \leq \left\|\boldsymbol{F}^\star\right\|\left\|\tilde{\boldsymbol{F}}^{t+1} - \boldsymbol{F}^\star\right\| = \alpha_2\left\|\boldsymbol{F}^\star\right\| \leq \frac{1}{\sqrt{2}\kappa}\left\|\boldsymbol{X}^\star\right\|\left\|\boldsymbol{F}^\star\right\| = \sigma_{\min}.$$

Since $\boldsymbol{F}^{\star\top}\boldsymbol{F}^\star = \boldsymbol{X}^{\star\top}\boldsymbol{X}^\star + \boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star = 2\boldsymbol{\Sigma}^\star$, Weyl's inequality gives

$$\lambda_{\min}(\boldsymbol{F}^{\star\top}\tilde{\boldsymbol{F}}^{t+1}) \geq 2\sigma_{\min} - \left\|\boldsymbol{F}^{\star\top}\tilde{\boldsymbol{F}}^{t+1} - \boldsymbol{F}^{\star\top}\boldsymbol{F}^\star\right\| \geq \sigma_{\min} > 0,$$

where $\lambda_{\min}(\boldsymbol{A})$ stands for the minimum eigenvalue of a matrix $\boldsymbol{A}$. To conclude, $\boldsymbol{F}^{\star\top}\tilde{\boldsymbol{F}}^{t+1}$ is both symmetric and positive semidefinite, thus establishing the claim. □

D.5 Proof of Lemma 12

Without loss of generality, we consider the case when $1 \le l \le n$; the case with $n+1 \le l \le 2n$ can be derived in a similar way. From the definition of $\boldsymbol{R}^{t+1,(l)}$ (cf. (95b)), we have

$$\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1} - \boldsymbol{F}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)}\right\|_{\mathrm{F}} \le \left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t} - \boldsymbol{F}^{t+1,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}.$$

The gradient update rules (27) and (90) give

$$\begin{aligned}
&\boldsymbol{F}^{t+1}\boldsymbol{H}^{t} - \boldsymbol{F}^{t+1,(l)}\boldsymbol{R}^{t,(l)} \\
&= \left[\boldsymbol{F}^{t} - \eta\nabla f\left(\boldsymbol{F}^{t}\right)\right]\boldsymbol{H}^{t} - \left[\boldsymbol{F}^{t,(l)} - \eta\nabla f^{(l)}\left(\boldsymbol{F}^{t,(l)}\right)\right]\boldsymbol{R}^{t,(l)} \\
&= \boldsymbol{F}^{t}\boldsymbol{H}^{t} - \eta\nabla f\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right) - \left[\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)} - \eta\nabla f^{(l)}\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right)\right] \\
&= \underbrace{\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t} - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right) - \eta\left[\nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right) - \nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right)\right]}_{:=\boldsymbol{A}_1} - \underbrace{\eta\left[\nabla f_{\mathsf{diff}}\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right) - \nabla f_{\mathsf{diff}}\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right)\right]}_{:=\boldsymbol{A}_2} \\
&\quad + \underbrace{\eta\left[\nabla f^{(l)}\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right) - \nabla f\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right)\right]}_{:=\boldsymbol{A}_3},
\end{aligned}$$

where we have used the facts that $\nabla f(\boldsymbol{F})\boldsymbol{R} = \nabla f(\boldsymbol{F}\boldsymbol{R})$ and $\nabla f^{(l)}(\boldsymbol{F})\boldsymbol{R} = \nabla f^{(l)}(\boldsymbol{F}\boldsymbol{R})$ for any orthonormal matrix $\boldsymbol{R} \in \mathcal{O}^{r\times r}$.

In what follows, we shall bound $\boldsymbol{A}_1, \boldsymbol{A}_2$ and $\boldsymbol{A}_3$ sequentially.

1. The first term $\boldsymbol{A}_1$ is similar to α_1 in the proof of Lemma 10. Going through the same derivations therein, we obtain
$$\|\boldsymbol{A}_1\|_{\mathrm{F}} \le \left(1 - \frac{\sigma_{\min}}{20}\eta\right)\left\|\boldsymbol{F}^{t}\boldsymbol{H}^{t} - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}, \tag{116}$$
provided that $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^4\mu r\log n}}$ and that $0 \le \eta \le 1/(1000\kappa\sigma_{\max})$.

2. Next, we turn attention to $\boldsymbol{A}_2$, which clearly obeys
$$\|\boldsymbol{A}_2\|_{\mathrm{F}} \le \eta\left\|\nabla f_{\mathsf{diff}}\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right)\right\|_{\mathrm{F}} + \eta\left\|\nabla f_{\mathsf{diff}}\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right)\right\|_{\mathrm{F}}.$$
Recall from the term α_2 in the proof of Lemma 10 that
$$\eta\left\|\nabla f_{\mathsf{diff}}\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right)\right\|_{\mathrm{F}} \le 2C_{\mathrm{B}}\kappa\eta^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2\left\|\boldsymbol{X}^{\star}\right\|.$$
Applying Lemma 15 and going through the same derivation as in bounding α_2 in the proof of Lemma 10, one gets
$$\eta\left\|\nabla f_{\mathsf{diff}}\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right)\right\|_{\mathrm{F}} \le 2C_{\mathrm{B}}\kappa\eta^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2\left\|\boldsymbol{X}^{\star}\right\|.$$
Combine the above three inequalities to obtain
$$\begin{aligned}
\|\boldsymbol{A}_2\|_{\mathrm{F}} &\le 4C_{\mathrm{B}}\kappa\eta^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2\left\|\boldsymbol{X}^{\star}\right\| \\
&\le 4\sqrt{n}C_{\mathrm{B}}\kappa\eta^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty} \\
&\le \eta\left(\sigma\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}.
\end{aligned}$$
Here the second inequality arises from the elementary inequality $\|\boldsymbol{X}^{\star}\| \le \sqrt{n}\|\boldsymbol{X}^{\star}\|_{2,\infty}$, whereas the last one holds true because of $\|\boldsymbol{X}^{\star}\|_{2,\infty} \le \|\boldsymbol{F}^{\star}\|_{2,\infty}$ and the condition that $\eta \ll \frac{1}{n\kappa^2\sigma_{\max}}$.

3. We are now left with $\boldsymbol{A}_3$. To this end, we first observe that

$$\boldsymbol{A}_3 = \eta \begin{bmatrix} \underbrace{\left[\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\right) - p^{-1}\mathcal{P}_{\Omega_{l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\right)\right]\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}}_{:=\boldsymbol{B}_1} + \underbrace{p^{-1}\mathcal{P}_{\Omega_{l,\cdot}}(\boldsymbol{E})\,\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}}_{:=\boldsymbol{C}_1} \\ \underbrace{\left[\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\right) - p^{-1}\mathcal{P}_{\Omega_{l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\right)\right]^\top\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}}_{:=\boldsymbol{B}_2} + \underbrace{p^{-1}\mathcal{P}_{\Omega_{l,\cdot}}(\boldsymbol{E})^\top\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}}_{:=\boldsymbol{C}_2} \end{bmatrix}.$$

The following claims allow one to bound $\boldsymbol{B}_1, \boldsymbol{B}_2$ and $\boldsymbol{C}_1, \boldsymbol{C}_2$; the proofs are deferred to the end of this subsection.

Claim 5. *Suppose that* $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2\log n}}$ *and that* $np \gg \log^2 n$. *With probability at least* $1 - O(n^{-100})$,

$$\|\boldsymbol{B}_1\|_{\mathrm{F}} \lesssim \sqrt{\frac{\mu^2 r^2 \log n}{np}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{F}^\star\right\|_{2,\infty}\sigma_{\max}. \tag{117}$$

Claim 6. *Suppose that* $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2\log n}}$ *and that* $np \gg \log n$. *With probability at least* $1 - O(n^{-100})$,

$$\|\boldsymbol{B}_2\|_{\mathrm{F}} \lesssim \sqrt{\frac{\mu^2 r^2 \log n}{np}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{F}^\star\right\|_{2,\infty}\sigma_{\max}. \tag{118}$$

Claim 7. *Suppose that* $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2\log n}}$ *and that* $np \gg \log^3 n$. *With probability at least* $1 - O(n^{-100})$,

$$\max\left\{\|\boldsymbol{C}_1\|_{\mathrm{F}}, \|\boldsymbol{C}_2\|_{\mathrm{F}}\right\} \lesssim \sigma\sqrt{\frac{n\log n}{p}}\,\|\boldsymbol{F}^\star\|_{2,\infty}. \tag{119}$$

With these claims in place, one can readily obtain that

$$\begin{aligned}
\|\boldsymbol{A}_3\|_{\mathrm{F}} &\le \eta\left(\|\boldsymbol{B}_1\|_{\mathrm{F}} + \|\boldsymbol{B}_2\|_{\mathrm{F}} + \|\boldsymbol{C}_1\|_{\mathrm{F}} + \|\boldsymbol{C}_2\|_{\mathrm{F}}\right) \\
&\lesssim \eta\sigma\sqrt{\frac{n\log n}{p}}\,\|\boldsymbol{F}^\star\|_{2,\infty} + \eta\sqrt{\frac{\mu^2 r^2\log n}{np}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{F}^\star\right\|_{2,\infty}\sigma_{\max} \\
&\le \eta\sigma\sqrt{\frac{n\log n}{p}}\,\|\boldsymbol{F}^\star\|_{2,\infty} + \eta\sqrt{\frac{\mu^2 r^2\log n}{np}}\left(C_\infty\kappa + C_3\right)\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}\,\sigma_{\max},
\end{aligned}$$

where the last line follows from the induction hypotheses (93c) and (93e).

This together with the bounds on $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ gives: for some constant $\tilde{C} > 0$,

$$\begin{aligned}
&\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1} - \boldsymbol{F}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)}\right\|_{\mathrm{F}} \le \|\boldsymbol{A}_1\|_{\mathrm{F}} + \|\boldsymbol{A}_2\|_{\mathrm{F}} + \|\boldsymbol{A}_3\|_{\mathrm{F}} \\
&\le \left(1 - \frac{\sigma_{\min}}{20}\eta\right)\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} + \eta\left(\sigma\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\|\boldsymbol{F}^\star\|_{2,\infty} \\
&\quad + \tilde{C}\eta\sigma\sqrt{\frac{n\log n}{p}}\,\|\boldsymbol{F}^\star\|_{2,\infty} + \tilde{C}\eta\sqrt{\frac{\mu^2 r^2\log n}{np}}\left(C_\infty\kappa + C_3\right)\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}\,\sigma_{\max} \\
&\overset{\text{(i)}}{\le} \left(1 - \frac{\sigma_{\min}}{20}\eta\right)C_3\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty} + \eta\left(\sigma\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\|\boldsymbol{F}^\star\|_{2,\infty} \\
&\quad + \tilde{C}\eta\sigma\sqrt{\frac{n\log n}{p}}\,\|\boldsymbol{F}^\star\|_{2,\infty} + \tilde{C}\eta\sqrt{\frac{\mu^2 r^2\log n}{np}}\left(C_\infty\kappa + C_3\right)\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}\,\sigma_{\max}
\end{aligned}$$

$$\overset{\text{(ii)}}{\leq} C_3\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}$$

as claimed. Here, (i) invokes the induction hypothesis (93c), whereas (ii) holds as long as C_3 is large enough and the sample size satisfies $n^2p\gg\kappa^4\mu^2r^2n\log n$.

Proof of Claim 5. For notational simplicity, we denote

$$\boldsymbol{C}\triangleq\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top}-\boldsymbol{M}^\star=\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top}-\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}. \tag{120}$$

Since the Frobenius norm is unitarily invariant, we have

$$\|\boldsymbol{B}_1\|_{\mathrm{F}}=\Big\|\underbrace{\left[p^{-1}\mathcal{P}_{\Omega_{l,\cdot}}(\boldsymbol{C})-\mathcal{P}_{l,\cdot}(\boldsymbol{C})\right]}_{:=\boldsymbol{W}}\boldsymbol{Y}^{t,(l)}\Big\|_{\mathrm{F}}.$$

All nonzero entries of the matrix $\boldsymbol{W}$ reside in its lth row and therefore

$$p\,\|\boldsymbol{B}_1\|_{\mathrm{F}}=\Big\|\sum\nolimits_{j=1}^{n}(\delta_{lj}-p)C_{lj}\boldsymbol{Y}_{j,\cdot}^{t,(l)}\Big\|_2,$$

where $\delta_{lj}\triangleq\mathbb{1}_{\{(l,j)\in\Omega\}}$. Notice that conditional on $\boldsymbol{X}^{t,(l)}$ and $\boldsymbol{Y}^{t,(l)}$, the right-hand side is composed of a sum of independent random vectors, where the randomness comes from $\{\delta_{lj}\}_{1\leq j\leq n}$. It then follows that

$$\begin{aligned}
L&\triangleq\max_{1\leq j\leq n}\Big\|(\delta_{l,j}-p)\,C_{l,j}\boldsymbol{Y}_{j,\cdot}^{t,(l)}\Big\|_2\leq\|\boldsymbol{C}\|_\infty\left\|\boldsymbol{Y}^{t,(l)}\right\|_{2,\infty}\overset{\text{(i)}}{\leq}2\,\|\boldsymbol{C}\|_\infty\|\boldsymbol{Y}^\star\|_{2,\infty},\\
V&\triangleq\Big\|\sum_{j=1}^{n}\mathbb{E}\big[(\delta_{l,j}-p)^2\big]C_{l,j}^2\boldsymbol{Y}_{j,\cdot}^{t,(l)}\boldsymbol{Y}_{j,\cdot}^{t,(l)\top}\Big\|\leq p\|\boldsymbol{C}\|_\infty^2\Big\|\sum_{j=1}^{n}\boldsymbol{Y}_{j,\cdot}^{t,(l)}\boldsymbol{Y}_{j,\cdot}^{t,(l)\top}\Big\|\\
&=p\left\|\boldsymbol{C}\right\|_\infty^2\left\|\boldsymbol{Y}^{t,(l)}\right\|_{\mathrm{F}}^2\overset{\text{(ii)}}{\leq}4p\left\|\boldsymbol{C}\right\|_\infty^2\left\|\boldsymbol{Y}^\star\right\|_{\mathrm{F}}^2.
\end{aligned}$$

Here, both (i) and (ii) arise from Lemma 18, as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\ll\frac{1}{\sqrt{\kappa^2\log n}}$. The matrix Bernstein inequality [Tro15, Theorem 6.1.1] reveals that

$$\Big\|\sum_{j=1}^{n}(\delta_{l,j}-p)\,C_{l,j}\boldsymbol{Y}_{j,\cdot}^{t,(l)}\Big\|_2\lesssim\sqrt{V\log n}+L\log n\lesssim\sqrt{p\,\|\boldsymbol{C}\|_\infty^2\,\|\boldsymbol{Y}^\star\|_{\mathrm{F}}^2\log n}+\|\boldsymbol{C}\|_\infty\|\boldsymbol{Y}^\star\|_{2,\infty}\log n$$

with probability exceeding $1-O(n^{-100})$. As a result, we arrive at

$$p\,\|\boldsymbol{B}_1\|_{\mathrm{F}}\lesssim\sqrt{p\log n}\,\|\boldsymbol{C}\|_\infty\|\boldsymbol{Y}^\star\|_{\mathrm{F}}+\sqrt{np}\,\|\boldsymbol{C}\|_\infty\|\boldsymbol{Y}^\star\|_{2,\infty} \tag{121}$$

as soon as $np\gg\log^2 n$.

To finish up, we make the observation that

$$\begin{aligned}
\|\boldsymbol{C}\|_\infty&=\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\left(\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\right)^\top-\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}\right\|_\infty\\
&\leq\left\|\left(\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^\star\right)\left(\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\right)^\top\right\|_\infty+\left\|\boldsymbol{X}^\star\left(\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{Y}^\star\right)^\top\right\|_\infty\\
&\leq\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{2,\infty}+\|\boldsymbol{X}^\star\|_{2,\infty}\left\|\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{Y}^\star\right\|_{2,\infty}\\
&\leq3\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^\star\right\|_{2,\infty}\|\boldsymbol{F}^\star\|_{2,\infty},
\end{aligned} \tag{122}$$

where the last line arises from Lemma 18. This combined with (121) gives

$$\|\boldsymbol{B}_1\|_{\mathrm{F}}\lesssim\sqrt{\frac{\log n}{p}}\,\|\boldsymbol{C}\|_\infty\|\boldsymbol{Y}^\star\|_{\mathrm{F}}+\sqrt{\frac{n}{p}}\,\|\boldsymbol{C}\|_\infty\|\boldsymbol{Y}^\star\|_{2,\infty}$$

$$\overset{\text{(i)}}{\lesssim} \sqrt{\frac{\log n}{p}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{Y}^{\star}\right\|_{\mathrm{F}}+\sqrt{\frac{n}{p}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}^{2}$$
$$\overset{\text{(ii)}}{\lesssim} \sqrt{\frac{\log n}{p}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^{\star}\right\|_{2,\infty}\sqrt{\frac{\mu r^{2}}{n}}\sigma_{\max}+\sqrt{\frac{n}{p}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^{\star}\right\|_{2,\infty}\frac{\mu r}{n}\sigma_{\max}$$
$$\lesssim \sqrt{\frac{\mu^{2}r^{2}\log n}{np}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^{\star}\right\|_{2,\infty}\sigma_{\max},$$

where (i) comes from (122), and (ii) makes use of the incoherence condition $\|\boldsymbol{F}^{\star}\|_{2,\infty}\le\sqrt{\mu r\sigma_{\max}/n}$ and the fact that $\|\boldsymbol{Y}^{\star}\|_{\mathrm{F}}\le\sqrt{r\sigma_{\max}}$. □

Proof of Claim 6. Instate the notation in proof of Claim 5. By the unitary invariance of Frobenius norm and the fact that all nonzero entries of the matrix $\boldsymbol{W}$ reside in its lth row, we have

$$p\left\|\boldsymbol{B}_{2}\right\|_{\mathrm{F}}=\left\|p\boldsymbol{W}^{\top}\boldsymbol{X}^{t,(l)}\right\|_{\mathrm{F}}=\Bigg\|\underbrace{\begin{bmatrix}(\delta_{l1}-p)\,C_{l1}\\ \vdots\\ (\delta_{ln}-p)\,C_{ln}\end{bmatrix}}_{:=\boldsymbol{b}}\boldsymbol{X}_{l,\cdot}^{t,(l)}\Bigg\|_{\mathrm{F}}=\left\|\boldsymbol{b}\right\|_{2}\left\|\boldsymbol{X}_{l,\cdot}^{t,(l)}\right\|_{2}.$$

We can write $\boldsymbol{b}$ as

$$\boldsymbol{b}=\sum\nolimits_{j=1}^{n}\underbrace{\boldsymbol{e}_{j}\left(\delta_{lj}-p\right)C_{lj}}_{:=\boldsymbol{u}_{j}}=\sum\nolimits_{j=1}^{n}\boldsymbol{u}_{j}.$$

Note that for all j, one has

$$L\triangleq\max_{1\le j\le n}\left\|\boldsymbol{u}_{j}\right\|_{2}\le\left\|\boldsymbol{C}\right\|_{\infty},$$
$$V\triangleq\left\|\sum\nolimits_{j=1}^{n}\mathbb{E}\big[\left(\delta_{lj}-p\right)^{2}\big]C_{lj}^{2}\boldsymbol{e}_{j}^{\top}\boldsymbol{e}_{j}\right\|\le p\|\boldsymbol{C}\|_{\infty}^{2}\left\|\sum\nolimits_{j=1}^{n}\boldsymbol{e}_{j}^{\top}\boldsymbol{e}_{j}\right\|=np\left\|\boldsymbol{C}\right\|_{\infty}^{2}.$$

Then the matrix Bernstein inequality [Tro15, Theorem 6.1.1] reveals that

$$\begin{aligned}\left\|\boldsymbol{b}\right\|_{2}&\lesssim\sqrt{V\log n}+L\log n\lesssim\sqrt{np\log n}\left\|\boldsymbol{C}\right\|_{\infty}+\left\|\boldsymbol{C}\right\|_{\infty}\log n\\&\lesssim\sqrt{np\log n}\left\|\boldsymbol{C}\right\|_{\infty}\\&\lesssim\sqrt{np\log n}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}\end{aligned}$$

with probability exceeding $1-O\left(n^{-100}\right)$ as long as $np\gg\log n$. Here the last relation uses (122). Observe that $\|\boldsymbol{X}^{t,(l)}\|_{2,\infty}\le 2\|\boldsymbol{F}^{\star}\|_{2,\infty}$ as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\ll\frac{1}{\sqrt{\kappa^{2}\log n}}$; see Lemma 18. Making use of the incoherence condition (100a) to get

$$\left\|\boldsymbol{B}_{2}\right\|_{\mathrm{F}}\lesssim\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}^{2}\lesssim\sqrt{\frac{\mu^{2}r^{2}\log n}{np}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^{\star}\right\|_{2,\infty}\sigma_{\max}.$$

We can then conclude the proof. □

Proof of Claim 7. By the unitary invariance of the Frobenius norm, one has

$$\left\|\boldsymbol{C}_{1}\right\|_{\mathrm{F}}=p^{-1}\left\|\mathcal{P}_{\Omega_{l,\cdot}}\left(\boldsymbol{E}\right)\boldsymbol{Y}^{t,(l)}\right\|_{\mathrm{F}}.$$

Since the entries of $\mathcal{P}_{\Omega_{l,\cdot}}(\boldsymbol{E})$ are all zero except those on the lth row, we have

$$p\left\|\boldsymbol{C}_{1}\right\|_{\mathrm{F}}=\Big\|\sum\nolimits_{j=1}^{n}\underbrace{\delta_{lj}E_{lj}\boldsymbol{Y}_{j,\cdot}^{t,(l)}}_{:=\boldsymbol{u}_{j}}\Big\|_{2},$$

where we denote $\delta_{lj} \triangleq \mathbb{1}_{(l,j)\in\Omega}$. Since $\boldsymbol{Y}^{t,(l)}$ is independent of $\{\delta_{lj}\}_{1\le j\le n}$ and $\{E_{lj}\}_{1\le j\le n}$, the vectors $\{\boldsymbol{u}_j\}_{1\le j\le n}$ are independent conditioning on $\boldsymbol{Y}^{t,(l)}$. Therefore, from now on we shall condition on a fixed $\boldsymbol{Y}^{t,(l)}$. It is easy to verify that

$$\big\|\|\boldsymbol{u}_j\|_2\big\|_{\psi_1} \le \big\|\boldsymbol{Y}^{t,(l)}\big\|_{2,\infty}\,\|\delta_{lj}E_{lj}\|_{\psi_1} \lesssim \sigma\big\|\boldsymbol{Y}^{t,(l)}\big\|_{2,\infty},$$

where $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm [KLT11, Section 6]. Further, one can calculate

$$V := \Big\|\mathbb{E}\Big[\sum\nolimits_{j=1}^{n}(\delta_{lj}E_{lj})^2\,\boldsymbol{Y}_{j,\cdot}^{t,(l)}\boldsymbol{Y}_{j,\cdot}^{t,(l)\top}\Big]\Big\| \lesssim p\sigma^2\Big\|\mathbb{E}\Big[\sum\nolimits_{j=1}^{n}\boldsymbol{Y}_{j,\cdot}^{t,(l)}\boldsymbol{Y}_{j,\cdot}^{t,(l)\top}\Big]\Big\| = p\sigma^2\big\|\boldsymbol{Y}^{t,(l)}\big\|_{\mathrm{F}}^2.$$

Invoke the matrix Bernstein inequality [KLT11, Proposition 2] to discover that with probability at least $1-O\left(n^{-100}\right)$,

$$\begin{aligned}\Big\|\sum\nolimits_{j=1}^{n}\boldsymbol{u}_j\Big\|_2 &\lesssim \sqrt{V\log n} + \Big\|\|\boldsymbol{u}_j\|_2\Big\|_{\psi_1}\log^2 n\\ &\lesssim \sqrt{p\sigma^2\big\|\boldsymbol{Y}^{t,(l)}\big\|_{\mathrm{F}}^2\log n} + \sigma\big\|\boldsymbol{Y}^{t,(l)}\big\|_{2,\infty}\log^2 n\\ &\lesssim \sigma\sqrt{np\log n}\big\|\boldsymbol{Y}^{t,(l)}\big\|_{2,\infty} + \sigma\big\|\boldsymbol{Y}^{t,(l)}\big\|_{2,\infty}\log^2 n\\ &\lesssim \sigma\sqrt{np\log n}\big\|\boldsymbol{Y}^{t,(l)}\big\|_{2,\infty},\end{aligned}$$

where the third inequality follows from $\|\boldsymbol{Y}^{t,(l)}\|_{\mathrm{F}}^2 \le n\|\boldsymbol{Y}^{t,(l)}\|_{2,\infty}^2$, and the last inequality holds if $np \gg \log^3 n$. We then complete the proof by observing that $\|\boldsymbol{Y}^{t,(l)}\|_{2,\infty} \le 2\|\boldsymbol{F}^\star\|_{2,\infty}$ as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2\log n}}$; see Lemma 18. The bound on $\boldsymbol{C}_2$ follows from similar arguments to that used to bound $\boldsymbol{B}_2$. □

D.6 Proof of Lemma 13

Without loss of generality, we assume $1 \le l \le n$. One can then decompose $(\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t+1,(l)} - \boldsymbol{F}^\star)_{l,\cdot}$ as

$$\begin{aligned}&(\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t+1,(l)} - \boldsymbol{F}^\star)_{l,\cdot} = \boldsymbol{X}_{l,\cdot}^{t+1,(l)}\boldsymbol{H}^{t+1,(l)} - \boldsymbol{X}_{l,\cdot}^\star\\ &= \Big\{\boldsymbol{X}_{l,\cdot}^{t,(l)} - \eta\big[(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star)_{l,\cdot}\boldsymbol{Y}^{t,(l)} + \tfrac{\lambda}{p}\boldsymbol{X}_{l,\cdot}^{t,(l)}\big]\Big\}\boldsymbol{H}^{t+1,(l)} - \boldsymbol{X}_{l,\cdot}^\star\\ &= \boldsymbol{X}_{l,\cdot}^{t,(l)}\boldsymbol{H}^{t+1,(l)} - \boldsymbol{X}_{l,\cdot}^\star - \eta\big[\big(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\big)_{l,\cdot}\boldsymbol{Y}^{t,(l)} + \tfrac{\lambda}{p}\boldsymbol{X}_{l,\cdot}^{t,(l)}\big]\boldsymbol{H}^{t+1,(l)}\\ &= \underbrace{\boldsymbol{X}_{l,\cdot}^{t,(l)}\boldsymbol{H}^{t,(l)} - \boldsymbol{X}_{l,\cdot}^\star - \eta\big[\big(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\big)_{l,\cdot}\boldsymbol{Y}^{t,(l)} + \tfrac{\lambda}{p}\boldsymbol{X}_{l,\cdot}^{t,(l)}\big]\boldsymbol{H}^{t,(l)}}_{:=\boldsymbol{a}_1}\\ &\quad + \underbrace{\Big\{\boldsymbol{X}_{l,\cdot}^{t,(l)}\boldsymbol{H}^{t,(l)} - \eta\big[\big(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\big)_{l,\cdot}\boldsymbol{Y}^{t,(l)} + \tfrac{\lambda}{p}\boldsymbol{X}_{l,\cdot}^{t,(l)}\big]\boldsymbol{H}^{t,(l)}\Big\}\Big[\big(\boldsymbol{H}^{t,(l)}\big)^{-1}\boldsymbol{H}^{t+1,(l)} - \boldsymbol{I}_r\Big]}_{:=\boldsymbol{a}_2}.\end{aligned}$$

Note that here $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are r-dimensional row vectors. In the sequel, let us control $\|\boldsymbol{a}_1\|_2$ and $\|\boldsymbol{a}_2\|_2$ separately.

1. We begin with $\boldsymbol{a}_1$. For notational convenience, define $\boldsymbol{\Delta}^{t,(l)} \triangleq \begin{bmatrix}\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}\\ \boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\end{bmatrix}$, where $\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)} \triangleq \boldsymbol{X}^{t,(l)}\boldsymbol{H}^{t,(l)} - \boldsymbol{X}^\star$ and $\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} \triangleq \boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)} - \boldsymbol{Y}^\star$. Then $\boldsymbol{a}_1$ can be rewritten as

$$\begin{aligned}\boldsymbol{a}_1 &= (\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot} - \eta\Big[(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}(\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)})^\top + \boldsymbol{X}_{l,\cdot}^\star\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)\top}\Big]\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)} - \eta\frac{\lambda}{p}\boldsymbol{X}_{l,\cdot}^{t,(l)}\boldsymbol{H}^{t,(l)}\\ &= (\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}\Big[\boldsymbol{I}_r - \eta(\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)}\Big] - \eta\boldsymbol{X}_{l,\cdot}^\star\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)\top}\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)} - \eta\frac{\lambda}{p}\boldsymbol{X}_{l,\cdot}^{t,(l)}\boldsymbol{H}^{t,(l)},\end{aligned}$$

which together with the triangle inequality yields

$$\|\boldsymbol{a}_1\|_2 \le \big\|\boldsymbol{I}_r - \eta(\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)}\big\|\cdot\big\|\big(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}\big)_{l,\cdot}\big\|_2$$

$$+\eta\|\boldsymbol{X}^\star_{l,\cdot}\|_2\cdot\left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\right\|\cdot\left\|\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)}\right\|+\eta\frac{\lambda}{p}\|\boldsymbol{X}^{t,(l)}\boldsymbol{H}^{t,(l)}\|_{2,\infty}.$$

In view of Lemma 18, we have

$$\begin{aligned}
\sigma_{\min}/2\le\sigma_{\min}\left[(\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)}\right]&\le\sigma_{\max}\left[(\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)})^\top\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)}\right]\le 2\sigma_{\max},\\
\|\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)}\|\le\|\boldsymbol{\Delta}^{t,(l)}\|&\le 2C_{\rm op}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{X}^\star\|,\qquad(123)\\
\|\boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)}\|\le\|\boldsymbol{F}^{t,(l)}\|&\le 2\|\boldsymbol{X}^\star\|\qquad\text{and}\\
\|\boldsymbol{X}^{t,(l)}\boldsymbol{H}^{t,(l)}\|_{2,\infty}\le\|\boldsymbol{F}^{t,(l)}\|_{2,\infty}&\le 2\|\boldsymbol{F}^\star\|_{2,\infty},
\end{aligned}$$

provided that the sample size obeys $n\gg\kappa\mu$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\ll\frac{1}{\sqrt{\kappa^2\log n}}$. These allow us to further upper bound $\|\boldsymbol{a}_1\|_2$ by

$$\|\boldsymbol{a}_1\|_2\le\left(1-\frac{\eta\sigma_{\min}}{2}\right)\left\|(\boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)})_{l,\cdot}\right\|_2+4\eta C_{\rm op}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sigma_{\max}\|\boldsymbol{F}^\star\|_{2,\infty}+2\eta\frac{\lambda}{p}\|\boldsymbol{F}^\star\|_{2,\infty},$$

as long as $\eta\le 1/(2\sigma_{\max})$. As an immediate consequence,

$$\begin{aligned}
\|\boldsymbol{a}_1\|_2&\le C_4\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}+4\eta C_{\rm op}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sigma_{\max}\|\boldsymbol{F}^\star\|_{2,\infty}+\frac{2\eta\lambda}{p}\|\boldsymbol{F}^\star\|_{2,\infty}\\
&\le\|\boldsymbol{F}^\star\|_{2,\infty},\qquad(124)
\end{aligned}$$

where the first inequality follows from the induction hypothesis (93d) and the last one holds as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\ll\frac{1}{\sqrt{\kappa^2\log n}}$ and $\eta\ll 1/\sigma_{\max}$.

2. Next, we turn attention to $\|\boldsymbol{a}_2\|_2$, which satisfies

$$\|\boldsymbol{a}_2\|_2=\left\|(\boldsymbol{a}_1+\boldsymbol{X}^\star_{l,\cdot})\left[(\boldsymbol{H}^{t,(l)})^{-1}\boldsymbol{H}^{t+1,(l)}-\boldsymbol{I}_r\right]\right\|_2\le\left\|(\boldsymbol{H}^{t,(l)})^{-1}\boldsymbol{H}^{t+1,(l)}-\boldsymbol{I}_r\right\|\left\|\boldsymbol{a}_1+\boldsymbol{X}^\star_{l,\cdot}\right\|_2.$$

From (124), it is easily seen that

$$\left\|\boldsymbol{a}_1+\boldsymbol{X}^\star_{l,\cdot}\right\|_2\le\|\boldsymbol{a}_1\|_2+\|\boldsymbol{F}^\star\|_{2,\infty}\le 2\left\|\boldsymbol{F}^\star\right\|_{2,\infty}.$$

Regarding the term $\|(\boldsymbol{H}^{t,(l)})^{-1}\boldsymbol{H}^{t+1,(l)}-\boldsymbol{I}_r\|$, we find the following claim useful.

Claim 8. *With probability at least $1-O(n^{-100})$, we have*

$$\left\|(\boldsymbol{H}^{t,(l)})^{-1}\boldsymbol{H}^{t+1,(l)}-\boldsymbol{I}_r\right\|\lesssim\eta\kappa C_{\rm op}^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)^2\sigma_{\max}+\eta^2C_{\rm B}\kappa^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2,$$

provided that $C_{\rm op}\gg 1$.

Finally, taking the bounds on $\|\boldsymbol{a}_1\|_2$ and $\|\boldsymbol{a}_2\|_2$ collectively yields that: for some absolute constant $\tilde{C}>0$,

$$\begin{aligned}
&\left\|\left(\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t+1,(l)}-\boldsymbol{F}^\star\right)_{l,\cdot}\right\|_2\le\|\boldsymbol{a}_1\|_2+\|\boldsymbol{a}_2\|_2\\
&\le\left(1-\frac{\eta}{2}\sigma_{\min}\right)C_4\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}+4C_{\rm op}\eta\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sigma_{\max}\|\boldsymbol{F}^\star\|_{2,\infty}+\frac{2\eta\lambda}{p}\|\boldsymbol{F}^\star\|_{2,\infty}\\
&\quad+\tilde{C}\eta\kappa C_{\rm op}^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)^2\sigma_{\max}\|\boldsymbol{F}^\star\|_{2,\infty}+\tilde{C}\eta^2C_{\rm B}\kappa^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2\|\boldsymbol{F}^\star\|_{2,\infty}\\
&\le C_4\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty},
\end{aligned}$$

provided that $C_4\gg C_{\rm op}$, $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\ll 1/\kappa$ and $\eta\ll 1/(\kappa^2\sqrt{r}\sigma_{\max})$. This finishes the proof of the lemma. It remains to establish Claim 8.

Proof of Claim 8. To facilitate analysis, we introduce an auxiliary point $\tilde{\boldsymbol{F}}^{t+1} \triangleq \begin{bmatrix} \tilde{\boldsymbol{X}}^{t+1,(l)} \\ \tilde{\boldsymbol{Y}}^{t+1,(l)} \end{bmatrix}$ where

$$\begin{aligned}
\tilde{\boldsymbol{X}}^{t+1,(l)} &= \boldsymbol{X}^{t,(l)}\boldsymbol{H}^{t,(l)} - \eta\left[\tfrac{1}{p}\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star - \boldsymbol{E}\right) + \mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\right)\right]\boldsymbol{Y}^\star \\
&\quad - \eta\frac{\lambda}{p}\boldsymbol{X}^\star - \frac{\eta}{2}\boldsymbol{X}^\star\boldsymbol{H}^{t,(l)\top}\left(\boldsymbol{X}^{t,(l)\top}\boldsymbol{X}^{t,(l)} - \boldsymbol{Y}^{t,(l)\top}\boldsymbol{Y}^{t,(l)}\right)\boldsymbol{H}^{t,(l)}, \\
\tilde{\boldsymbol{Y}}^{t+1,(l)} &= \boldsymbol{Y}^{t,(l)}\boldsymbol{H}^{t,(l)} - \eta\left[\tfrac{1}{p}\mathcal{P}_{\Omega_{-l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star - \boldsymbol{E}\right) + \mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\right)\right]^\top\boldsymbol{X}^\star \\
&\quad - \eta\frac{\lambda}{p}\boldsymbol{Y}^\star - \frac{\eta}{2}\boldsymbol{Y}^\star\boldsymbol{H}^{t,(l)\top}\left(\boldsymbol{Y}^{t,(l)\top}\boldsymbol{Y}^{t,(l)} - \boldsymbol{X}^{t,(l)\top}\boldsymbol{X}^{t,(l)}\right)\boldsymbol{H}^{t,(l)}.
\end{aligned}$$

We first claim that $\boldsymbol{I}_r$ is the best rotation matrix to align $\tilde{\boldsymbol{F}}^{t+1,(l)}$ and $\boldsymbol{F}^\star$; its proof is similar to that of Claim 4 and is hence omitted for brevity.

Claim 9. *One has*

$$\boldsymbol{I}_r = \arg\min_{\boldsymbol{R}\in\mathcal{O}^r}\left\|\tilde{\boldsymbol{F}}^{t+1,(l)}\boldsymbol{R} - \boldsymbol{F}^\star\right\|_{\mathrm{F}} \qquad \text{and} \qquad \sigma_{\min}\big(\tilde{\boldsymbol{F}}^{t+1,(l)\top}\boldsymbol{F}^\star\big) \geq \sigma_{\min}/2.$$

With this claim at hand, we intend to invoke Lemma 23 with

$$\boldsymbol{S} = \tilde{\boldsymbol{F}}^{t+1,(l)\top}\boldsymbol{F}^\star, \quad \boldsymbol{K} = \big(\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t,(l)} - \tilde{\boldsymbol{F}}^{t+1,(l)}\big)^\top\boldsymbol{F}^\star$$

to get

$$\begin{aligned}
\left\|(\boldsymbol{H}^{t,(l)})^{-1}\boldsymbol{H}^{t+1,(l)} - \boldsymbol{I}_r\right\| &= \|\mathsf{sgn}(\boldsymbol{S}+\boldsymbol{K}) - \mathsf{sgn}(\boldsymbol{S})\| \leq \frac{1}{\sigma_{\min}(\boldsymbol{S})}\|\boldsymbol{K}\| \\
&= \frac{1}{\sigma_{\min}(\tilde{\boldsymbol{F}}^{t+1,(l)\top}\boldsymbol{F}^\star)}\left\|\big(\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t,(l)} - \tilde{\boldsymbol{F}}^{t+1,(l)}\big)^\top\boldsymbol{F}^\star\right\|. \\
&\leq \frac{2}{\sigma_{\min}}\left\|\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t,(l)} - \tilde{\boldsymbol{F}}^{t+1,(l)}\right\|\|\boldsymbol{F}^\star\|, \qquad (125)
\end{aligned}$$

where the last line uses Claim 9. Here $\mathsf{sgn}(\boldsymbol{A}) = \boldsymbol{U}\boldsymbol{V}^\top$ for any matrix $\boldsymbol{A}$ with SVD $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. It then boils down to controlling $\|\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t,(l)} - \tilde{\boldsymbol{F}}^{t+1,(l)}\|$, for which we have

$$\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t,(l)} - \tilde{\boldsymbol{F}}^{t+1,(l)} = \eta\begin{bmatrix}\boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\Delta}_{\boldsymbol{Y}}^{t,(l)} \\ \boldsymbol{\Delta}_{\boldsymbol{X}}^{t,(l)}\end{bmatrix} + \frac{\eta}{2}\begin{bmatrix}\boldsymbol{X}^\star \\ -\boldsymbol{Y}^\star\end{bmatrix}\boldsymbol{H}^{t,(l)\top}\boldsymbol{C}\boldsymbol{H}^{t,(l)} - \eta\frac{\lambda}{p}\boldsymbol{\Delta}^{t,(l)},$$

where we denote

$$\begin{aligned}
&\boldsymbol{B} \triangleq -p^{-1}\mathcal{P}_{\Omega_{-l,\cdot}}\big(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star - \boldsymbol{E}\big) - \mathcal{P}_{l,\cdot}\big(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\big); \\
&\boldsymbol{C} \triangleq \boldsymbol{X}^{t,(l)\top}\boldsymbol{X}^{t,(l)} - \boldsymbol{Y}^{t,(l)\top}\boldsymbol{Y}^{t,(l)}.
\end{aligned}$$

This enables us to obtain

$$\left\|\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t,(l)} - \tilde{\boldsymbol{F}}^{t+1,(l)}\right\| \leq \eta\|\boldsymbol{B}\|\left\|\boldsymbol{\Delta}^{t,(l)}\right\| + \frac{\eta}{2}\|\boldsymbol{F}^\star\|\|\boldsymbol{C}\|_{\mathrm{F}} + \frac{\eta\lambda}{p}\|\boldsymbol{\Delta}^{t,(l)}\|. \qquad (126)$$

In view of Lemma 15, one has

$$\|\boldsymbol{C}\|_{\mathrm{F}} \leq C_{\mathrm{B}}\kappa\eta\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2. \qquad (127)$$

We are left with bounding $\|\boldsymbol{B}\|$. Decompose $\boldsymbol{B}$ into

$$\boldsymbol{B} = \underbrace{-\tfrac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\right)}_{:=\boldsymbol{B}_1} + \underbrace{\tfrac{1}{p}\mathcal{P}_{\Omega_{l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\right) - \mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top} - \boldsymbol{M}^\star\right)}_{:=\boldsymbol{B}_2} + \underbrace{\tfrac{1}{p}\mathcal{P}_{\Omega_{-l,\cdot}}(\boldsymbol{E})}_{:=\boldsymbol{B}_3}.$$

To control $\boldsymbol{B}_1$, following the same argument in Lemma 8, we see that

$$\begin{aligned}\|\boldsymbol{B}_1\| &\le \left\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top}-\boldsymbol{M}^\star)-(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top}-\boldsymbol{M}^\star)\right\|+\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)\top}-\boldsymbol{M}^\star\right\| \\ &\lesssim \sqrt{n/p}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^\star\right\|_{2,\infty}\|\boldsymbol{F}^\star\|_{2,\infty}+\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^\star\right\|\|\boldsymbol{F}^\star\|.\end{aligned}$$

We now move on to $\|\boldsymbol{B}_2\|$, which is equal to $\|\boldsymbol{b}\|_2/p$ defined in the proof of Claim 6, namely,

$$\|\boldsymbol{B}_2\| = \|\boldsymbol{b}\|_2/p \lesssim \sqrt{n\log n/p}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^\star\right\|_{2,\infty}\|\boldsymbol{F}^\star\|_{2,\infty}.$$

The last term $\boldsymbol{B}_3$ can be easily bound via Lemma 3, that is,

$$\|\boldsymbol{B}_3\| \le p^{-1}\|\mathcal{P}_\Omega(\boldsymbol{E})\| \lesssim \sigma\sqrt{n/p}.$$

Combining the above three bounds with Lemma 18, we arrive at

$$\begin{aligned}\|\boldsymbol{B}\| &\le \|\boldsymbol{B}_1\|+\|\boldsymbol{B}_2\|+\|\boldsymbol{B}_3\| \\ &\lesssim \sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^\star\right\|_{2,\infty}\|\boldsymbol{F}^\star\|_{2,\infty}+\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^\star\right\|\|\boldsymbol{F}^\star\|+\sigma\sqrt{\frac{n}{p}} \\ &\lesssim \sqrt{\frac{n\log n}{p}}\left(C_\infty\kappa+C_3\right)\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\frac{\mu r}{n}\sigma_{\max}+2C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sigma_{\max}+\sigma\sqrt{\frac{n}{p}} \\ &\lesssim C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sigma_{\max}, \qquad (128)\end{aligned}$$

provided that $n^2p \gg \kappa^2\mu^2r^2n\log^2 n$ and $C_{\mathrm{op}}>0$ is large enough. Taking (126), (127) and (128) collectively, we arrive at

$$\begin{aligned}&\left\|\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t,(l)}-\tilde{\boldsymbol{F}}^{t+1,(l)}\right\| \\ &\lesssim \eta C_{\mathrm{op}}^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)^2\sigma_{\max}\|\boldsymbol{X}^\star\|+\eta^2C_{\mathrm{B}}\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2\|\boldsymbol{X}^\star\| \\ &\quad+\eta\frac{\lambda}{p}C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{X}^\star\| \\ &\lesssim \eta C_{\mathrm{op}}^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)^2\sigma_{\max}\|\boldsymbol{X}^\star\|+\eta^2C_{\mathrm{B}}\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2\|\boldsymbol{X}^\star\|,\end{aligned}$$

provided that C_{op} is large enough. Here the last relation uses (123). Substitution into (125) yields

$$\|(\boldsymbol{H}^{t,(l)})^{-1}\boldsymbol{H}^{t+1,(l)}-\boldsymbol{I}_r\| \lesssim \eta\kappa C_{\mathrm{op}}^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)^2\sigma_{\max}+\eta^2C_{\mathrm{B}}\kappa^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2,$$

which concludes the proof. □

D.7 Proof of Lemma 14

Fix any $1\le l\le 2n$. Apply the triangle inequality to see that

$$\begin{aligned}\left\|\left(\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1}-\boldsymbol{F}^\star\right)_{l,\cdot}\right\|_2 &\le \left\|\left(\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1}-\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t+1,(l)}\right)_{l,\cdot}\right\|_2+\left\|\left(\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t+1,(l)}-\boldsymbol{F}^\star\right)_{l,\cdot}\right\|_2 \\ &\le \left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1}-\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t+1,(l)}\right\|_{\mathrm{F}}+C_4\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}, \qquad (129)\end{aligned}$$

where the second line follows from Lemma 13. Apply Lemma 18 to the $(t+1)$th iterates to see that

$$\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1}-\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t+1,(l)}\right\|_{\mathrm{F}} \le 5\kappa\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1}-\boldsymbol{F}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)}\right\|_{\mathrm{F}}$$

$$\leq 5\kappa C_3 \left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}. \tag{130}$$

Here the second line follows from Lemma 12. Combine (129) and (130) to reach

$$\begin{aligned}\left\|\left(\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1}-\boldsymbol{F}^\star\right)_{l,\cdot}\right\|_2 &\leq 5\kappa C_3\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty} + C_4\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}\\ &\leq C_\infty\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}\end{aligned}$$

as long as $C_\infty \geq 5C_3 + C_4$. The proof is then complete since this holds for all $1 \leq l \leq 2n$.

D.8 Proof of Lemma 15

To simplify the notation hereafter, we denote

$$\boldsymbol{A}^t \triangleq \boldsymbol{X}^{t\top}\boldsymbol{X}^t - \boldsymbol{Y}^{t\top}\boldsymbol{Y}^t \qquad \text{and} \qquad \boldsymbol{A}^{t+1} \triangleq \boldsymbol{X}^{t+1\top}\boldsymbol{X}^{t+1} - \boldsymbol{Y}^{t+1\top}\boldsymbol{Y}^{t+1}.$$

In view of the gradient descent update rules (27), we have

$$\begin{aligned}\boldsymbol{X}^{t+1\top}\boldsymbol{X}^{t+1} &= \boldsymbol{X}^{t\top}\boldsymbol{X}^t - \eta\left[\boldsymbol{X}^{t\top}\nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t) + \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t)^\top\boldsymbol{X}^t\right] + \eta^2\nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t)^\top\nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t),\\ \boldsymbol{Y}^{t+1\top}\boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^{t\top}\boldsymbol{Y}^t - \eta\left[\boldsymbol{Y}^{t\top}\nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t) + \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t)^\top\boldsymbol{Y}^t\right] + \eta^2\nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t)^\top\nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t).\end{aligned}$$

This gives rise to the following identity

$$\boldsymbol{A}^{t+1} = \boldsymbol{A}^t - \eta\boldsymbol{B}^t + \eta^2\boldsymbol{C}^t, \tag{131}$$

where we denote

$$\begin{aligned}\boldsymbol{B}^t &\triangleq \boldsymbol{X}^{t\top}\nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t) + \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t)^\top\boldsymbol{X}^t - \boldsymbol{Y}^{t\top}\nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t) - \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t)^\top\boldsymbol{Y}^t,\\ \boldsymbol{C}^t &\triangleq \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t)^\top\nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t) - \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t)^\top\nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t,\boldsymbol{Y}^t).\end{aligned}$$

Denoting

$$\boldsymbol{D}^t \triangleq p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}), \tag{132}$$

we have

$$\nabla_{\boldsymbol{X}} f\left(\boldsymbol{X}^t,\boldsymbol{Y}^t\right) = \boldsymbol{D}^t\boldsymbol{Y}^t + \tfrac{\lambda}{p}\boldsymbol{X}^t \qquad \text{and} \qquad \nabla_{\boldsymbol{Y}} f\left(\boldsymbol{X}^t,\boldsymbol{Y}^t\right) = \boldsymbol{D}^{t\top}\boldsymbol{X}^t + \tfrac{\lambda}{p}\boldsymbol{Y}^t.$$

With these in mind, a little calculation reveals that

$$\boldsymbol{B}^t = \boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t + \boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{X}^t + \tfrac{2\lambda}{p}\boldsymbol{X}^{t\top}\boldsymbol{X}^t - \boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{X}^t - \boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t - \tfrac{2\lambda}{p}\boldsymbol{Y}^{t\top}\boldsymbol{Y}^t = \tfrac{2\lambda}{p}\boldsymbol{A}^t$$

as well as

$$\begin{aligned}\boldsymbol{C}^t &= \left(\boldsymbol{D}^t\boldsymbol{Y}^t + \tfrac{\lambda}{p}\boldsymbol{X}^t\right)^\top\left(\boldsymbol{D}^t\boldsymbol{Y}^t + \tfrac{\lambda}{p}\boldsymbol{X}^t\right) - \left(\boldsymbol{D}^{t\top}\boldsymbol{X}^t + \tfrac{\lambda}{p}\boldsymbol{Y}^t\right)^\top\left(\boldsymbol{D}^{t\top}\boldsymbol{X}^t + \tfrac{\lambda}{p}\boldsymbol{Y}^t\right)\\ &= \boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t + \tfrac{\lambda}{p}\boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{X}^t + \tfrac{\lambda}{p}\boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t + \left(\tfrac{\lambda}{p}\right)^2\boldsymbol{X}^{t\top}\boldsymbol{X}^t\\ &\quad - \boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{D}^{t\top}\boldsymbol{X}^t - \tfrac{\lambda}{p}\boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t - \tfrac{\lambda}{p}\boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{X}^t - \left(\tfrac{\lambda}{p}\right)^2\boldsymbol{Y}^{t\top}\boldsymbol{Y}^t\\ &= \left(\boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t - \boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{D}^{t\top}\boldsymbol{X}^t\right) + \left(\tfrac{\lambda}{p}\right)^2\boldsymbol{A}^t.\end{aligned}$$

Substituting the identities for $\boldsymbol{B}^t$ and $\boldsymbol{C}^t$ into (131) yields

$$\boldsymbol{A}^{t+1} = \boldsymbol{A}^t - 2\eta\tfrac{\lambda}{p}\boldsymbol{A}^t + \eta^2\left(\boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t - \boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{D}^{t\top}\boldsymbol{X}^t\right) + \eta^2\left(\tfrac{\lambda}{p}\right)^2\boldsymbol{A}^t$$

$$= (1-\lambda\eta/p)^2 \boldsymbol{A}^t + \eta^2 \left(\boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t - \boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{D}^{t\top}\boldsymbol{X}^t\right),$$

which together with the triangle inequality gives

$$\begin{aligned}\left\|\boldsymbol{A}^{t+1}\right\|_{\mathrm{F}} &\le (1-\lambda\eta/p)^2 \left\|\boldsymbol{A}^t\right\|_{\mathrm{F}} + \eta^2 \left\|\boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t - \boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{D}^{t\top}\boldsymbol{X}^t\right\|_{\mathrm{F}} \\ &\le (1-\lambda\eta/p) \left\|\boldsymbol{A}^t\right\|_{\mathrm{F}} + \eta^2 \left\|\boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t - \boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{D}^{t\top}\boldsymbol{X}^t\right\|_{\mathrm{F}},\end{aligned}$$

as long as $\lambda\eta/p < 1$ — a condition that is guaranteed by our assumptions on λ and η. It then boils down to controlling $\|\boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t - \boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{D}^{t\top}\boldsymbol{X}^t\|_{\mathrm{F}}$, which is supplied in the following claim.

Claim 10. *Suppose that the sample complexity satisfies $n^2 p \gg \kappa^2\mu^2 r^2 n\log n$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2\log n}}$, then one has*

$$\left\|\boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t - \boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{D}^{t\top}\boldsymbol{X}^t\right\|_{\mathrm{F}} \lesssim C_{\mathrm{op}}^2 \left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)^2 \sqrt{r}\sigma_{\max}^3.$$

Taking the above bounds together, we arrive at for some constant $\tilde{C} > 0$,

$$\begin{aligned}\left\|\boldsymbol{A}^{t+1}\right\|_{\mathrm{F}} &\le \left(1-\frac{\lambda}{p}\eta\right)\left\|\boldsymbol{A}^t\right\|_{\mathrm{F}} + \eta^2\tilde{C}C_{\mathrm{op}}^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)^2\sqrt{r}\sigma_{\max}^3 \\ &\le \left(1-\frac{\lambda}{p}\eta\right)C_{\mathrm{B}}\kappa\eta\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2 + \eta^2\tilde{C}C_{\mathrm{op}}^2\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)^2\sqrt{r}\sigma_{\max}^3 \\ &\le C_{\mathrm{B}}\kappa\eta\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2,\end{aligned}$$

as long as $\lambda \ge \sigma\sqrt{np}$ and $C_{\mathrm{B}} \gg C_{\mathrm{op}}^2$.

Proof of Claim 10. The triangle inequality yields

$$\begin{aligned}\left\|\boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t - \boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{D}^{t\top}\boldsymbol{X}^t\right\|_{\mathrm{F}} &\le \left\|\boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t\right\|_{\mathrm{F}} + \left\|\boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{D}^{t\top}\boldsymbol{X}^t\right\|_{\mathrm{F}} \\ &\le \left\|\boldsymbol{Y}^t\right\|\left\|\boldsymbol{D}^t\right\|^2\left\|\boldsymbol{Y}^t\right\|_{\mathrm{F}} + \left\|\boldsymbol{X}^t\right\|\left\|\boldsymbol{D}^t\right\|^2\left\|\boldsymbol{X}^t\right\|_{\mathrm{F}}. \qquad (133)\end{aligned}$$

It is easy to see from Lemma 18 that

$$\left\|\boldsymbol{Y}^t\right\| \le 2\left\|\boldsymbol{Y}^\star\right\|, \quad \left\|\boldsymbol{Y}^t\right\|_{\mathrm{F}} \le 2\left\|\boldsymbol{Y}^\star\right\|_{\mathrm{F}}, \quad \left\|\boldsymbol{X}^t\right\| \le 2\left\|\boldsymbol{X}^\star\right\| \quad \text{and} \quad \left\|\boldsymbol{X}^t\right\|_{\mathrm{F}} \le 2\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}$$

provided that $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2\log n}}$. These allow us to further upper bound (133) as

$$\begin{aligned}\left\|\boldsymbol{Y}^{t\top}\boldsymbol{D}^{t\top}\boldsymbol{D}^t\boldsymbol{Y}^t - \boldsymbol{X}^{t\top}\boldsymbol{D}^t\boldsymbol{D}^{t\top}\boldsymbol{X}^t\right\|_{\mathrm{F}} &\le 4\left\|\boldsymbol{D}^t\right\|^2\left\|\boldsymbol{Y}^\star\right\|\left\|\boldsymbol{Y}^\star\right\|_{\mathrm{F}} + 4\left\|\boldsymbol{D}^t\right\|^2\left\|\boldsymbol{X}^\star\right\|\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}} \\ &\le 8\left\|\boldsymbol{D}^t\right\|^2\sqrt{r}\sigma_{\max}. \qquad (134)\end{aligned}$$

It remains to bound $\|\boldsymbol{D}^t\|$. To this end, recall from (132) that

$$\left\|\boldsymbol{D}^t\right\| \le p^{-1}\left\|\mathcal{P}_\Omega(\boldsymbol{E})\right\| + p^{-1}\left\|\mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}^\star\right)\right\| + \left\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}^\star\right\|.$$

In the sequel we shall bound these three terms sequentially. First, Lemma 3 tells us that $\frac{1}{p}\left\|\mathcal{P}_\Omega(\boldsymbol{E})\right\| \lesssim \sigma\sqrt{\frac{n}{p}}$. Next, repeating the arguments in the proof of Lemma 8 gives

$$\begin{aligned}\left\|\mathcal{P}_\Omega^{\mathsf{debias}}\left(\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}^\star\right)\right\| &= \left\|\mathcal{P}_\Omega^{\mathsf{debias}}\left[\boldsymbol{X}^t\boldsymbol{H}^t\left(\boldsymbol{Y}^t\boldsymbol{H}^t\right)^\top - \boldsymbol{M}^\star\right]\right\| \\ &\lesssim \sqrt{np}\left(\left\|\boldsymbol{X}^t\boldsymbol{H}^t - \boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{Y}^\star\right\|_{2,\infty} + \left\|\boldsymbol{Y}^t\boldsymbol{H}^t - \boldsymbol{Y}^\star\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\right),\end{aligned}$$

which together with the induction hypothesis (93e) yields

$$\begin{aligned}\frac{1}{p}\left\|\mathcal{P}_{\Omega}^{\mathsf{debias}}\left(\boldsymbol{X}^{t}\boldsymbol{Y}^{t\top}-\boldsymbol{M}^{\star}\right)\right\| &\lesssim \sqrt{\frac{n}{p}}C_{\infty}\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}\\ &\lesssim C_{\infty}\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{\frac{\mu^{2}r^{2}}{np}}\sigma_{\max}.\end{aligned}$$

Here the last relation uses the incoherence assumption $\|\boldsymbol{F}^{\star}\|_{2,\infty}\le\sqrt{\mu r\sigma_{\max}/n}$ (cf. (100a)). Regarding $\|\boldsymbol{X}^{t}\boldsymbol{Y}^{t\top}-\boldsymbol{M}^{\star}\|$, the triangle inequality reveals that

$$\begin{aligned}\left\|\boldsymbol{X}^{t}\boldsymbol{Y}^{t\top}-\boldsymbol{M}^{\star}\right\| &= \left\|\boldsymbol{X}^{t}\boldsymbol{H}^{t}\left(\boldsymbol{Y}^{t}\boldsymbol{H}^{t}\right)^{\top}-\boldsymbol{M}^{\star}\right\|\\ &\le \left\|\boldsymbol{X}^{t}\boldsymbol{H}^{t}\left(\boldsymbol{Y}^{t}\boldsymbol{H}^{t}\right)^{\top}-\boldsymbol{X}^{t}\boldsymbol{H}^{t}\boldsymbol{Y}^{\star\top}\right\|+\left\|\boldsymbol{X}^{t}\boldsymbol{H}^{t}\boldsymbol{Y}^{\star\top}-\boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top}\right\|\\ &\le \left\|\boldsymbol{X}^{t}\boldsymbol{H}^{t}\right\|\left\|\boldsymbol{Y}^{t}\boldsymbol{H}^{t}-\boldsymbol{Y}^{\star}\right\|+\left\|\boldsymbol{X}^{t}\boldsymbol{H}^{t}-\boldsymbol{X}^{\star}\right\|\left\|\boldsymbol{Y}^{\star}\right\|.\end{aligned}$$

Combine the induction hypothesis (93b) and the fact that $\|\boldsymbol{X}^{t}\boldsymbol{H}^{t}\|=\|\boldsymbol{X}^{t}\|\le 2\|\boldsymbol{X}^{\star}\|$ to reach

$$\left\|\boldsymbol{X}^{t}\boldsymbol{Y}^{t\top}-\boldsymbol{M}^{\star}\right\|\lesssim C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sigma_{\max}.$$

Putting together the previous three bounds, we arrive at

$$\begin{aligned}\left\|\boldsymbol{D}^{t}\right\| &\lesssim \sigma\sqrt{\frac{n}{p}}+C_{\infty}\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{\frac{\mu^{2}r^{2}}{np}}\sigma_{\max}+C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sigma_{\max}\\ &\lesssim C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\,\sigma_{\min}}\right)\sigma_{\max}\end{aligned}\tag{135}$$

since $np\gg\kappa^{2}\mu^{2}r^{2}\log n$. Putting (135) back to (134) leads to the claimed upper bound.

The upper bound on the leave-one-out sequences can be derived similarly. For brevity, we omit it. □

D.9 Proof of Lemma 16

In light of the facts that $f(\boldsymbol{F}\boldsymbol{R})=f(\boldsymbol{F})$ and $\nabla f(\boldsymbol{F}\boldsymbol{R})=\nabla f(\boldsymbol{F})\boldsymbol{R}$ for any $\boldsymbol{R}\in\mathcal{O}^{r\times r}$, one has

$$\begin{aligned}f\left(\boldsymbol{F}^{t+1}\right) &= f\left(\boldsymbol{F}^{t+1}\boldsymbol{H}^{t}\right)=f\left(\left[\boldsymbol{F}^{t}-\eta\nabla f\left(\boldsymbol{F}^{t}\right)\right]\boldsymbol{H}^{t}\right)\\ &= f\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}-\eta\nabla f\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right)\right)\\ &= f\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right)-\eta\left\langle\nabla f\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right),\nabla f\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right)\right\rangle+\frac{\eta^{2}}{2}\mathsf{vec}\left(\nabla f\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right)\right)^{\top}\nabla^{2}f(\tilde{\boldsymbol{F}})\mathsf{vec}\left(\nabla f\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right)\right)\end{aligned}$$

for some $\tilde{\boldsymbol{F}}$ which lies between $\boldsymbol{F}^{t}\boldsymbol{H}^{t}$ and $\boldsymbol{F}^{t}\boldsymbol{H}^{t}-\eta\nabla f(\boldsymbol{F}^{t}\boldsymbol{H}^{t})$. Suppose for the moment that

$$\left\|\boldsymbol{F}^{t}\boldsymbol{H}^{t}-\boldsymbol{F}^{\star}\right\|_{2,\infty}\le\frac{1}{2000\kappa\sqrt{n}}\|\boldsymbol{X}^{\star}\|,\tag{136a}$$

$$\left\|\boldsymbol{F}^{t}\boldsymbol{H}^{t}-\eta\nabla f(\boldsymbol{F}^{t}\boldsymbol{H}^{t})-\boldsymbol{F}^{\star}\right\|_{2,\infty}\le\frac{1}{1000\kappa\sqrt{n}}\|\boldsymbol{X}^{\star}\|.\tag{136b}$$

One can invoke Lemma 17 to obtain $\|\nabla^{2}f(\tilde{\boldsymbol{F}})\|\le 10\sigma_{\max}$ and hence

$$\begin{aligned}f\left(\boldsymbol{F}^{t+1}\right) &\le f\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right)-\eta\left\|\nabla f\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right)\right\|_{\mathrm{F}}^{2}+5\eta^{2}\sigma_{\max}\left\|\nabla f\left(\boldsymbol{F}^{t}\boldsymbol{H}^{t}\right)\right\|_{\mathrm{F}}^{2}\\ &= f\left(\boldsymbol{F}^{t}\right)-\eta\left\|\nabla f\left(\boldsymbol{F}^{t}\right)\right\|_{\mathrm{F}}^{2}+5\eta^{2}\sigma_{\max}\left\|\nabla f\left(\boldsymbol{F}^{t}\right)\right\|_{\mathrm{F}}^{2}\\ &\le f\left(\boldsymbol{F}^{t}\right)-\tfrac{\eta}{2}\left\|\nabla f\left(\boldsymbol{F}^{t}\right)\right\|_{\mathrm{F}}^{2}.\end{aligned}$$

Here the equality uses again the facts that $f(\boldsymbol{F}\boldsymbol{R}) = f(\boldsymbol{F})$ and $\nabla f(\boldsymbol{F}\boldsymbol{R}) = \nabla f(\boldsymbol{F})\boldsymbol{R}$ for any $\boldsymbol{R} \in \mathcal{O}^{r\times r}$ and the last inequality holds as long as $\eta \le \frac{1}{10\sigma_{\max}}$. We are left with proving the aforementioned conditions (136). The first condition has been established in the proof of Lemma 9 and hence we concentrate on the second one, namely (136b). Apply the triangle inequality and the fundamental theorem of calculus [Lan93, Chapter XIII, Theorem 4.2] to obtain

$$\begin{aligned}\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \eta\nabla f(\boldsymbol{F}^t\boldsymbol{H}^t) - \boldsymbol{F}^\star\right\|_{2,\infty} &\le \left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\right\|_{2,\infty} + \eta\left\|\nabla f(\boldsymbol{F}^t\boldsymbol{H}^t) - \nabla f(\boldsymbol{F}^\star)\right\|_{\mathrm{F}} + \eta\left\|\nabla f(\boldsymbol{F}^\star)\right\|_{\mathrm{F}} \\ &\le \left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\right\|_{2,\infty} + \eta\left\|\int_0^1 \nabla^2 f\left(\boldsymbol{F}\left(\tau\right)\right)\mathrm{d}\tau\,\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\right)\right\|_2 + \eta\left\|\nabla f(\boldsymbol{F}^\star)\right\|_{\mathrm{F}},\end{aligned}$$

where $\boldsymbol{F}(\tau) \triangleq \boldsymbol{F}^\star + \tau(\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star)$ for $0 \le \tau \le 1$. Following similar arguments in the proof of Lemma 10 and the proof of Lemma 9, one obtains

$$\begin{aligned}&\eta\left\|\int_0^1 \nabla^2 f\left(\boldsymbol{F}\left(\tau\right)\right)\mathrm{d}\tau\,\mathsf{vec}\left(\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\right)\right\|_{\mathrm{F}} + \eta\left\|\nabla f(\boldsymbol{F}^\star)\right\|_{\mathrm{F}} \\ &\le \eta\cdot 10\sigma_{\max}\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\right\|_{\mathrm{F}} + \eta\frac{\lambda}{p}\sqrt{r\sigma_{\max}} \\ &\lesssim \eta\sigma_{\max}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{r}\|\boldsymbol{X}^\star\| + \eta\sigma_{\min}\frac{\lambda}{p\,\sigma_{\min}}\sqrt{r}\|\boldsymbol{X}^\star\| \le \frac{1}{2000\kappa\sqrt{n}}\|\boldsymbol{X}^\star\|.\end{aligned}$$

Here the middle inequality uses the induction hypothesis (93b) and the last relation holds true provided that $\lambda \asymp \sigma\sqrt{np}$, $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1/\sqrt{r}$ and that $\eta \ll 1/(\kappa n\sigma_{\max})$. This proves the second condition and also the whole lemma.

D.10 Proof of Lemma 17

We start by defining a new loss function

$$f_{\mathsf{clean}}\left(\boldsymbol{X},\boldsymbol{Y}\right) \triangleq \frac{1}{2p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\right)\right\|_{\mathrm{F}}^2 + \frac{1}{8}\left\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}\right\|_{\mathrm{F}}^2;$$

compared with $f_{\mathsf{aug}}(\cdot,\cdot)$, this new function $f_{\mathsf{clean}}(\cdot,\cdot)$ sets $\lambda = 0$ and excludes the noise $\boldsymbol{E}$ from consideration. It is straightforward to check that for any $\boldsymbol{\Delta} \in \mathbb{R}^{2n\times r}$,

$$\mathsf{vec}\left(\boldsymbol{\Delta}\right)^\top\nabla^2 f_{\mathsf{aug}}\left(\boldsymbol{X},\boldsymbol{Y}\right)\mathsf{vec}\left(\boldsymbol{\Delta}\right) = \mathsf{vec}\left(\boldsymbol{\Delta}\right)^\top\nabla^2 f_{\mathsf{clean}}\left(\boldsymbol{X},\boldsymbol{Y}\right)\mathsf{vec}\left(\boldsymbol{\Delta}\right) - \frac{2}{p}\left\langle\mathcal{P}_\Omega\left(\boldsymbol{E}\right),\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\rangle + \frac{\lambda}{p}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}^2.$$

It has been proven in [CLL19, Lemma 3.2] that under the assumptions stated in Lemma 17, one has

$$\mathsf{vec}\left(\boldsymbol{\Delta}\right)^\top\nabla^2 f_{\mathsf{clean}}\left(\boldsymbol{X},\boldsymbol{Y}\right)\mathsf{vec}\left(\boldsymbol{\Delta}\right) \ge \tfrac{1}{5}\sigma_{\min}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}^2 \qquad \text{and} \qquad \left\|\nabla^2 f_{\mathsf{clean}}\left(\boldsymbol{X},\boldsymbol{Y}\right)\right\| \le 5\sigma_{\max}. \tag{137}$$

It then boils down to controlling $-\frac{2}{p}\left\langle\mathcal{P}_\Omega\left(\boldsymbol{E}\right),\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\rangle + \frac{\lambda}{p}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}^2$. To this end, one has

$$\left|\tfrac{1}{p}\left\langle\mathcal{P}_\Omega\left(\boldsymbol{E}\right),\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\rangle\right| \le \left\|\tfrac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right\|\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\|_* \lesssim \sigma\sqrt{\tfrac{n}{p}}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}^2, \tag{138}$$

where the last relation holds due to Lemma 3 and the elementary fact about the nuclear norm (6), i.e.

$$2\left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\|_* \le \left\|\boldsymbol{\Delta}_{\boldsymbol{X}}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\right\|_{\mathrm{F}}^2 = \left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}^2.$$

Regarding the term $\lambda\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2/p$, it is easy to see from the assumption $\lambda \asymp \sigma\sqrt{np}$ that $\frac{\lambda}{p}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}^2 \asymp \sigma\sqrt{\frac{n}{p}}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}^2$. Combine the above two bounds and use the triangle inequality to reach

$$\left|-\tfrac{2}{p}\left\langle\mathcal{P}_\Omega\left(\boldsymbol{E}\right),\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\right\rangle + \tfrac{\lambda}{p}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}^2\right| \lesssim \sigma\sqrt{\tfrac{n}{p}}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}^2 \le \tfrac{1}{10}\sigma_{\min}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}^2, \tag{139}$$

with the proviso that $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1$. Taking (137) and (139) together immediately establishes the claims on $\nabla^2 f_{\mathsf{aug}}(\cdot,\cdot)$.

Moving on to $\nabla^2 f(\boldsymbol{X},\boldsymbol{Y})$, one has

$$\begin{aligned}
&\mathsf{vec}(\boldsymbol{\Delta})^\top \nabla^2 f\left(\boldsymbol{X},\boldsymbol{Y}\right)\mathsf{vec}(\boldsymbol{\Delta}) = \tfrac{1}{p}\big\|\mathcal{P}_\Omega\big(\boldsymbol{X}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top + \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{Y}^\top\big)\big\|_{\mathrm{F}}^2 + \tfrac{2}{p}\big\langle \mathcal{P}_\Omega\big(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star - \boldsymbol{E}\big), \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\big\rangle + \tfrac{\lambda}{p}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \\
&= \underbrace{\big\|\boldsymbol{X}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top + \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{Y}^\top\big\|_{\mathrm{F}}^2}_{:=\alpha_1} + \underbrace{2\big\langle \boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star, \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\big\rangle}_{:=\alpha_2} + \underbrace{\big(-\tfrac{2}{p}\big\langle \mathcal{P}_\Omega(\boldsymbol{E}), \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\big\rangle + \tfrac{\lambda}{p}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2\big)}_{:=\alpha_3} \\
&\quad + \underbrace{\tfrac{1}{p}\big\|\mathcal{P}_\Omega\big(\boldsymbol{X}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top + \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{Y}^\top\big)\big\|_{\mathrm{F}}^2 + \tfrac{2}{p}\big\langle \mathcal{P}_\Omega\big(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\big), \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\big\rangle - \big\|\boldsymbol{X}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top + \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{Y}^\top\big\|_{\mathrm{F}}^2 - 2\big\langle \boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star, \boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\big\rangle}_{:=\alpha_4}.
\end{aligned}$$

The term α_4 can be bounded by [CLL19, Equation A.4]

$$|\alpha_4| \le \tfrac{1}{5}\sigma_{\min}\big(\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{\mathrm{F}}^2 + \|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{\mathrm{F}}^2\big) + \tfrac{1}{5}\big(\big\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\big\|_{\mathrm{F}}^2 + \big\|\boldsymbol{X}^\star\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\big\|_{\mathrm{F}}^2\big) \le \tfrac{2}{5}\sigma_{\max}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2.$$

The term α_3 has been bounded in (139) where $|\alpha_3| \le \sigma_{\max}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2$ provided that $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1$. The term α_2 can be written as

$$|\alpha_2| \le 2\big\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}^\star\big\|\big\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\big\|_* \le \left(\|\boldsymbol{X}-\boldsymbol{X}^\star\|\,\|\boldsymbol{Y}\| + \|\boldsymbol{X}^\star\|\,\|\boldsymbol{Y}-\boldsymbol{Y}^\star\|\right)\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2.$$

Since

$$\left\|\begin{bmatrix}\boldsymbol{X}-\boldsymbol{X}^\star\\ \boldsymbol{Y}-\boldsymbol{Y}^\star\end{bmatrix}\right\| \le \left\|\begin{bmatrix}\boldsymbol{X}-\boldsymbol{X}^\star\\ \boldsymbol{Y}-\boldsymbol{Y}^\star\end{bmatrix}\right\|_{\mathrm{F}} \le \sqrt{2n}\left\|\begin{bmatrix}\boldsymbol{X}-\boldsymbol{X}^\star\\ \boldsymbol{Y}-\boldsymbol{Y}^\star\end{bmatrix}\right\|_{2,\infty} \le \frac{1}{500\kappa}\|\boldsymbol{X}^\star\|,$$

we immediately have

$$|\alpha_2| \le \frac{3}{500\kappa}\sigma_{\max}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \le \frac{1}{2}\sigma_{\max}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2.$$

The term α_1 can be bounded by

$$\alpha_1 \le 2\left(\big\|\boldsymbol{X}^\star\boldsymbol{\Delta}_{\boldsymbol{Y}}^\top\big\|_{\mathrm{F}}^2 + \big\|\boldsymbol{\Delta}_{\boldsymbol{X}}\boldsymbol{Y}^{\star\top}\big\|_{\mathrm{F}}^2\right) \le 2\left(\|\boldsymbol{X}^\star\|^2\|\boldsymbol{\Delta}_{\boldsymbol{Y}}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^\star\|^2\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{\mathrm{F}}^2\right) = 2\sigma_{\max}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2.$$

Combining all these bounds yields

$$\mathsf{vec}(\boldsymbol{\Delta})^\top \nabla^2 f\left(\boldsymbol{X},\boldsymbol{Y}\right)\mathsf{vec}(\boldsymbol{\Delta}) \le 10\sigma_{\max}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2.$$

D.11 Proof of Lemma 18

The first set of consequences (101) follows straightforwardly from the triangle inequality. For instance, combine the induction hypotheses (93c) and (93e) to obtain

$$\begin{aligned}
\big\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{F}^\star\big\|_{2,\infty} &\le \big\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{F}^t\boldsymbol{H}^t\big\|_{2,\infty} + \big\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\big\|_{2,\infty} \\
&\le (C_\infty\kappa + C_3)\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}.
\end{aligned}$$

Similar bounds can be obtained for $\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{F}^\star\|$ provided that $n \gg \mu r\log n$.

We continue to establish the second set of consequences namely (102). Since $\|\cdot\|$ is unitarily invariant, one can apply the triangle inequality to get

$$\begin{aligned}
\big\|\boldsymbol{F}^t\big\| = \big\|\boldsymbol{F}^t\boldsymbol{H}^t\big\| &\le \big\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\big\| + \|\boldsymbol{F}^\star\| \\
&\overset{\text{(i)}}{\le} C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{X}^\star\| + \sqrt{2}\,\|\boldsymbol{X}^\star\| \overset{\text{(ii)}}{\le} 2\,\|\boldsymbol{X}^\star\|.
\end{aligned}$$

Here (i) uses the induction hypothesis (93b) and the fact that $\|\boldsymbol{F}^\star\| = \sqrt{2}\|\boldsymbol{X}^\star\|$, and (ii) holds as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1$. Similarly one can obtain $\|\boldsymbol{F}^t\|_{\mathrm{F}} \le 2\|\boldsymbol{X}^\star\|_{\mathrm{F}}$ provided that $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1$ and $\|\boldsymbol{F}^t\|_{2,\infty} \le 2\|\boldsymbol{F}^\star\|_{2,\infty}$ as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1/(\sqrt{\kappa^2\log n})$. Notice that along the way, we have also proven that

$$\big\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\big\| \le \|\boldsymbol{X}^\star\|, \quad \big\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\big\|_{\mathrm{F}} \le \|\boldsymbol{X}^\star\|_{\mathrm{F}} \quad \text{and} \quad \big\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\big\|_{2,\infty} \le \|\boldsymbol{F}^\star\|_{2,\infty}.$$

We now move on to $\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{H}^{t,(l)}\|_{\mathrm{F}}$, for which we intend to apply Lemma 22 to connect it with $\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\|_{\mathrm{F}}$. First, in view of the induction hypothesis (93b), one has

$$\begin{aligned}\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\right\|\|\boldsymbol{F}^\star\| &\le C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{X}^\star\|\,\|\boldsymbol{F}^\star\| \\ &= \sqrt{2}C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\sigma_{\max} \\ &\le \sigma_r^2\left(\boldsymbol{F}^\star\right)/2,\end{aligned}$$

where the equality arises since $\|\boldsymbol{F}^\star\| = \sqrt{2\sigma_{\max}}$ (see (100a)) and $\|\boldsymbol{X}^\star\| = \sqrt{\sigma_{\max}}$, and the last line holds as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1/\kappa$. In addition, it follows from the induction hypothesis (93c) that

$$\begin{aligned}\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|\|\boldsymbol{F}^\star\| &\le \left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}\|\boldsymbol{F}^\star\| \\ &\le C_3\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}\|\boldsymbol{F}^\star\| \\ &\le \sqrt{2}C_3\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\sqrt{\frac{\mu r}{n}}\sigma_{\max} \\ &\le \sigma_r^2\left(\boldsymbol{F}^\star\right)/4,\end{aligned}$$

where the penultimate inequality arises from the facts that $\|\boldsymbol{F}^\star\|_{2,\infty} \le \sqrt{\mu r\sigma_{\max}/n}$ and that $\|\boldsymbol{F}^\star\| = \sqrt{2\sigma_{\max}}$ (cf. (100)), and the last relation holds as long as $(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}})\sqrt{\frac{\mu r}{n}} \ll 1/\kappa$. Invoke Lemma 22 with

$$\boldsymbol{F}_0 = \boldsymbol{F}^\star, \qquad \boldsymbol{F}_1 = \boldsymbol{F}^t\boldsymbol{H}^t \qquad \text{and} \qquad \boldsymbol{F}_2 = \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}$$

to arrive at

$$\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{H}^{t,(l)}\right\|_{\mathrm{F}} \le 5\frac{\sigma_1^2\left(\boldsymbol{F}^\star\right)}{\sigma_r^2\left(\boldsymbol{F}^\star\right)}\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} = 5\kappa\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}.$$

The last set of consequences can be derived following similar arguments to that for establishing the first set. For brevity, we omit the proof.

D.12 Proof of the inequalities (31)

We single out the proof of $\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}^\star\|_\infty$, whereas the proofs of $\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}^\star\|_{\mathrm{F}}$ and $\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}^\star\|$ follow from the same argument. Recognize the following decomposition

$$\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}^\star = \left(\boldsymbol{X}^t\boldsymbol{H}^t - \boldsymbol{X}^\star\right)\left(\boldsymbol{Y}^t\boldsymbol{H}^t\right)^\top + \boldsymbol{X}^\star\left(\boldsymbol{Y}^t\boldsymbol{H}^t - \boldsymbol{Y}^\star\right)^\top,$$

which together with the triangle inequality gives

$$\begin{aligned}\left\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}^\star\right\|_\infty &\le \left\|\left(\boldsymbol{X}^t\boldsymbol{H}^t - \boldsymbol{X}^\star\right)\left(\boldsymbol{Y}^t\boldsymbol{H}^t\right)^\top\right\|_\infty + \left\|\boldsymbol{X}^\star\left(\boldsymbol{Y}^t\boldsymbol{H}^t - \boldsymbol{Y}^\star\right)^\top\right\|_\infty \\ &\le \left\|\boldsymbol{X}^t\boldsymbol{H}^t - \boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{Y}^t\boldsymbol{H}^t\right\|_{2,\infty} + \|\boldsymbol{X}^\star\|_{2,\infty}\left\|\boldsymbol{Y}^t\boldsymbol{H}^t - \boldsymbol{Y}^\star\right\|_{2,\infty}.\end{aligned}$$

In view of Lemma 18, one has $\|\boldsymbol{Y}^t\boldsymbol{H}^t\|_{2,\infty} \le 2\|\boldsymbol{F}^\star\|_{2,\infty}$ as long as the noise obeys $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2\log n}}$. This further implies that

$$\begin{aligned}\left\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{M}^\star\right\|_\infty &\le 3C_\infty\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}\|\boldsymbol{F}^\star\|_{2,\infty} \\ &\le 3C_\infty\sqrt{\kappa^3\mu r}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\,\sigma_{\min}}\right)\|\boldsymbol{M}^\star\|_\infty,\end{aligned}$$

where the last relation is $\|\boldsymbol{F}^\star\|_{2,\infty}\|\boldsymbol{F}^\star\|_{2,\infty} \le \sqrt{\kappa\mu r}\|\boldsymbol{M}^\star\|_\infty$. To see this, one has for any $1 \le i \le n$,

$$n\left\|\boldsymbol{M}^\star\right\|_\infty^2 \ge \sum_{j=1}^n (M_{ij}^\star)^2 = \boldsymbol{X}_{i,\cdot}^\star \boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\boldsymbol{X}_{i,\cdot}^{\star\top} \ge \left\|\boldsymbol{X}_{i,\cdot}^\star\right\|_2^2 \lambda_{\min}\left(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star\right) = \sigma_{\min}\left\|\boldsymbol{X}_{i,\cdot}^\star\right\|_2^2.$$

Here $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue. Since the inequality holds for all $1 \le i \le n$, we arrive at

$$\left\|\boldsymbol{X}^\star\right\|_{2,\infty} \le \sqrt{\frac{n}{\sigma_{\min}}}\left\|\boldsymbol{M}^\star\right\|_\infty.$$

Similarly one can obtain $\|\boldsymbol{Y}^\star\|_{2,\infty} \le \sqrt{n/\sigma_{\min}}\|\boldsymbol{M}^\star\|_\infty$, which further implies $\|\boldsymbol{F}^\star\|_{2,\infty} = \max\{\|\boldsymbol{X}^\star\|_{2,\infty}, \|\boldsymbol{Y}^\star\|_{2,\infty}\} \le \sqrt{n/\sigma_{\min}}\|\boldsymbol{M}^\star\|_\infty$. As a result, we arrive at

$$\left\|\boldsymbol{F}^\star\right\|_{2,\infty}\left\|\boldsymbol{F}^\star\right\|_{2,\infty} \le \sqrt{\frac{n}{\sigma_{\min}}}\left\|\boldsymbol{M}^\star\right\|_\infty \cdot \sqrt{\frac{\mu r}{n}}\sqrt{\sigma_{\max}} \le \sqrt{\kappa\mu r}\left\|\boldsymbol{M}^\star\right\|_\infty.$$

Here we used the incoherence assumption (100a).

E Technical lemmas

Lemma 19. *Suppose $n^2 p \ge Cn\log n$ for some sufficiently large constant $C > 0$. Then with probability exceeding $1 - O(n^{-10})$,*

$$\left|p^{-1}\left\|\mathcal{P}_\Omega\left(\boldsymbol{A}\boldsymbol{B}^\top\right)\right\|_{\mathrm{F}}^2 - \left\|\boldsymbol{A}\boldsymbol{B}^\top\right\|_{\mathrm{F}}^2\right| \le 3n\min\left\{\|\boldsymbol{A}\|_{2,\infty}^2\|\boldsymbol{B}\|_{\mathrm{F}}^2, \|\boldsymbol{B}\|_{2,\infty}^2\|\boldsymbol{A}\|_{\mathrm{F}}^2\right\}$$

holds uniformly for all matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times r}$.

Proof. In view of [ZL16, Lemma 9], one has

$$p^{-1}\left\|\mathcal{P}_\Omega\left(\boldsymbol{A}\boldsymbol{B}^\top\right)\right\|_{\mathrm{F}}^2 \le 2n\min\left\{\|\boldsymbol{A}\|_{2,\infty}^2\|\boldsymbol{B}\|_{\mathrm{F}}^2, \|\boldsymbol{B}\|_{2,\infty}^2\|\boldsymbol{A}\|_{\mathrm{F}}^2\right\}$$

with high probability. In addition, simple algebra reveals that

$$\left\|\boldsymbol{A}\boldsymbol{B}^\top\right\|_{\mathrm{F}}^2 \le \|\boldsymbol{A}\|_{\mathrm{F}}^2\|\boldsymbol{B}\|_{\mathrm{F}}^2 \le n\|\boldsymbol{A}\|_{2,\infty}^2\|\boldsymbol{B}\|_{\mathrm{F}}^2$$

and, similarly, $\|\boldsymbol{A}\boldsymbol{B}^\top\|_{\mathrm{F}}^2 \le n\|\boldsymbol{A}\|_{\mathrm{F}}^2\|\boldsymbol{B}\|_{2,\infty}^2$. Combining the previous bounds with the triangle inequality establishes the claim. □

Lemma 20. *Let $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the SVD of a rank-r matrix $\boldsymbol{X}\boldsymbol{Y}^\top$ with $\boldsymbol{X}, \boldsymbol{Y} \in \boldsymbol{R}^{n\times r}$. Then there exists an invertible matrix $\boldsymbol{Q} \in \mathbb{R}^{r\times r}$ such that $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}$ and $\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}^{-\top}$. In addition, one has*

$$\left\|\boldsymbol{\Sigma}_{\boldsymbol{Q}} - \boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\right\|_{\mathrm{F}} \le \frac{1}{\sigma_{\min}(\boldsymbol{\Sigma})}\left\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}\right\|_{\mathrm{F}}, \tag{140}$$

where $\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^\top$ is the SVD of $\boldsymbol{Q}$. In particular, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ have balanced scale, i.e. $\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y} = \boldsymbol{0}$, then $\boldsymbol{Q}$ must be a rotation matrix.

Proof. The existence of $\boldsymbol{Q}$ is trivial by setting

$$\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}^\top\boldsymbol{X}.$$

To see this, one has

$$\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}^\top\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{X} = \boldsymbol{X},$$

where the last equality follows from the fact that the columns of $\boldsymbol{U}$ are the left singular vectors of $\boldsymbol{X}$. The relation $\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}^{-\top}$ can also be verified by the identity

$$\boldsymbol{X}\boldsymbol{Y}^\top = \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Y}^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top.$$

We now move on to proving the perturbation bound (140). In view of the SVD of $\boldsymbol{Q}$, i.e. $\boldsymbol{Q}=\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}$, one can obtain

$$\begin{aligned}\boldsymbol{X}^{\top}\boldsymbol{X}-\boldsymbol{Y}^{\top}\boldsymbol{Y} &= \boldsymbol{Q}^{\top}\boldsymbol{\Sigma}\boldsymbol{Q}-\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}\boldsymbol{Q}^{-\top} \\ &= \boldsymbol{V}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{U}_{\boldsymbol{Q}}^{\top}\boldsymbol{\Sigma}\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}-\boldsymbol{V}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\boldsymbol{U}_{\boldsymbol{Q}}^{\top}\boldsymbol{\Sigma}\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}.\end{aligned}$$

Denote $\boldsymbol{B}:=\boldsymbol{U}_{\boldsymbol{Q}}^{\top}\boldsymbol{\Sigma}\boldsymbol{U}_{\boldsymbol{Q}}\succ 0$. Then we have

$$\left\|\boldsymbol{X}^{\top}\boldsymbol{X}-\boldsymbol{Y}^{\top}\boldsymbol{Y}\right\|_{\mathrm{F}}^{2}=\left\|\boldsymbol{V}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{B}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}-\boldsymbol{V}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\boldsymbol{B}\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}\right\|_{\mathrm{F}}^{2}=\left\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{B}\boldsymbol{\Sigma}_{\boldsymbol{Q}}-\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\boldsymbol{B}\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\right\|_{\mathrm{F}}^{2}.$$

Let $\boldsymbol{C}=\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{B}^{1/2}$ and $\boldsymbol{D}=\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\boldsymbol{B}^{1/2}$, and denote $\boldsymbol{\Delta}=\boldsymbol{C}-\boldsymbol{D}$. One then has

$$\begin{aligned}\left\|\boldsymbol{X}^{\top}\boldsymbol{X}-\boldsymbol{Y}^{\top}\boldsymbol{Y}\right\|_{\mathrm{F}}^{2} &= \left\|\boldsymbol{C}\boldsymbol{C}^{\top}-\boldsymbol{D}\boldsymbol{D}^{\top}\right\|_{\mathrm{F}}^{2}=\left\|\boldsymbol{C}\boldsymbol{\Delta}^{\top}+\boldsymbol{\Delta}\boldsymbol{C}^{\top}-\boldsymbol{\Delta}\boldsymbol{\Delta}^{\top}\right\|_{\mathrm{F}}^{2} \\ &= \mathrm{Tr}\left(2\boldsymbol{C}^{\top}\boldsymbol{C}\boldsymbol{\Delta}^{\top}\boldsymbol{\Delta}+\boldsymbol{\Delta}\boldsymbol{\Delta}^{\top}\boldsymbol{\Delta}\boldsymbol{\Delta}^{\top}+2\boldsymbol{C}^{\top}\boldsymbol{\Delta}\boldsymbol{C}^{\top}\boldsymbol{\Delta}-4\boldsymbol{C}^{\top}\boldsymbol{\Delta}\boldsymbol{\Delta}^{\top}\boldsymbol{\Delta}\right) \\ &= \mathrm{Tr}\Big[\left(\boldsymbol{\Delta}^{\top}\boldsymbol{\Delta}-\sqrt{2}\boldsymbol{C}^{\top}\boldsymbol{\Delta}\right)^{2}+(4-2\sqrt{2})\boldsymbol{C}^{\top}(\boldsymbol{C}-\boldsymbol{\Delta})\boldsymbol{\Delta}^{\top}\boldsymbol{\Delta}+(2\sqrt{2}-1)\boldsymbol{C}^{\top}\boldsymbol{C}\boldsymbol{\Delta}^{\top}\boldsymbol{\Delta}\Big].\end{aligned}$$

Note that $\boldsymbol{C}^{\top}\boldsymbol{D}=\boldsymbol{B}$ and that $\boldsymbol{C}^{\top}\boldsymbol{\Delta}=\boldsymbol{C}^{\top}\boldsymbol{C}-\boldsymbol{C}^{\top}\boldsymbol{D}=\boldsymbol{C}^{\top}\boldsymbol{C}-\boldsymbol{B}$ is symmetric. One can continue the bound as

$$\begin{aligned}\left\|\boldsymbol{X}^{\top}\boldsymbol{X}-\boldsymbol{Y}^{\top}\boldsymbol{Y}\right\|_{\mathrm{F}}^{2} &= \left\|\boldsymbol{\Delta}^{\top}\boldsymbol{\Delta}-\sqrt{2}\boldsymbol{C}^{\top}\boldsymbol{\Delta}\right\|_{\mathrm{F}}^{2}+(4-2\sqrt{2})\mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{\Delta}\boldsymbol{\Delta}^{\top}\right)+(2\sqrt{2}-1)\left\|\boldsymbol{C}\boldsymbol{\Delta}^{\top}\right\|_{\mathrm{F}}^{2} \\ &\geq \mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{\Delta}\boldsymbol{\Delta}^{\top}\right),\end{aligned}$$

where the inequality follows since $4-2\sqrt{2}\geq 1$. Write $\boldsymbol{B}=\boldsymbol{B}^{1/2}\cdot\boldsymbol{B}^{1/2}$ to see

$$\begin{aligned}\left\|\boldsymbol{X}^{\top}\boldsymbol{X}-\boldsymbol{Y}^{\top}\boldsymbol{Y}\right\|_{\mathrm{F}}^{2} &\geq \mathrm{Tr}(\boldsymbol{B}^{1/2}\boldsymbol{\Delta}\boldsymbol{\Delta}^{\top}\boldsymbol{B}^{1/2})=\left\|\boldsymbol{B}^{1/2}\boldsymbol{\Delta}\right\|_{\mathrm{F}}^{2} \\ &= \left\|\boldsymbol{B}^{1/2}(\boldsymbol{\Sigma}_{\boldsymbol{Q}}-\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1})\boldsymbol{B}^{1/2}\right\|_{\mathrm{F}}^{2} \\ &\geq \sigma_{\min}^{2}\left(\boldsymbol{B}\right)\left\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}-\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\right\|_{\mathrm{F}}^{2}.\end{aligned}$$

Recognizing that $\sigma_{\min}(\boldsymbol{B})=\sigma_{\min}(\boldsymbol{\Sigma})$ finishes the proof of (140).

Combining $\boldsymbol{X}^{\top}\boldsymbol{X}=\boldsymbol{Y}^{\top}\boldsymbol{Y}$ and (140) yields $\left\|\boldsymbol{\Sigma}_{\boldsymbol{Q}}-\boldsymbol{\Sigma}_{\boldsymbol{Q}}^{-1}\right\|_{\mathrm{F}}=0$, which implies $\boldsymbol{\Sigma}_{\boldsymbol{Q}}=\boldsymbol{I}$. Under this circumstance, $\boldsymbol{Q}=\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}=\boldsymbol{U}_{\boldsymbol{Q}}\boldsymbol{V}_{\boldsymbol{Q}}^{\top}$ is a rotation matrix. The proof is then complete. □

Lemma 21. *For all $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}\in\mathbb{R}^{n\times r}$, one has*

$$\left|\left\langle\mathcal{P}_{\Omega}\left(\boldsymbol{A}\boldsymbol{C}^{\top}\right),\mathcal{P}_{\Omega}\left(\boldsymbol{B}\boldsymbol{D}^{\top}\right)\right\rangle-p\left\langle\boldsymbol{A}\boldsymbol{C}^{\top},\boldsymbol{B}\boldsymbol{D}^{\top}\right\rangle\right|\leq\left\|\mathcal{P}_{\Omega}\left(\mathbf{1}\mathbf{1}^{\top}\right)-p\mathbf{1}\mathbf{1}^{\top}\right\|\left\|\boldsymbol{A}\right\|_{2,\infty}\left\|\boldsymbol{B}\right\|_{\mathrm{F}}\left\|\boldsymbol{C}\right\|_{2,\infty}\left\|\boldsymbol{D}\right\|_{\mathrm{F}}.$$

Proof. This is a simple consequence of [CL17, Lemma 4.4], where they have shown

$$\begin{aligned}&\left|\left\langle\mathcal{P}_{\Omega}\left(\boldsymbol{A}\boldsymbol{C}^{\top}\right),\mathcal{P}_{\Omega}\left(\boldsymbol{B}\boldsymbol{D}^{\top}\right)\right\rangle-p\left\langle\boldsymbol{A}\boldsymbol{C}^{\top},\boldsymbol{B}\boldsymbol{D}^{\top}\right\rangle\right| \\ &\leq\left\|\mathcal{P}_{\Omega}\left(\mathbf{1}\mathbf{1}^{\top}\right)-p\mathbf{1}\mathbf{1}^{\top}\right\|\sqrt{\sum_{k=1}^{n}\left\|\boldsymbol{A}_{k,\cdot}\right\|_{2}^{2}\left\|\boldsymbol{B}_{k,\cdot}\right\|_{2}^{2}}\sqrt{\sum_{k=1}^{n}\left\|\boldsymbol{C}_{k,\cdot}\right\|_{2}^{2}\left\|\boldsymbol{D}_{k,\cdot}\right\|_{2}^{2}}.\end{aligned}$$

Recognize that

$$\sum_{k=1}^{n}\left\|\boldsymbol{A}_{k,\cdot}\right\|_{2}^{2}\left\|\boldsymbol{B}_{k,\cdot}\right\|_{2}^{2}\leq\left\|\boldsymbol{A}\right\|_{2,\infty}^{2}\sum_{k=1}^{n}\left\|\boldsymbol{B}_{k,\cdot}\right\|_{2}^{2}=\left\|\boldsymbol{A}\right\|_{2,\infty}^{2}\left\|\boldsymbol{B}\right\|_{\mathrm{F}}^{2}$$

and, similarly, $\sum_{k}\left\|\boldsymbol{C}_{k,\cdot}\right\|_{2}^{2}\left\|\boldsymbol{D}_{k,\cdot}\right\|_{2}^{2}\leq\left\|\boldsymbol{C}\right\|_{2,\infty}^{2}\left\|\boldsymbol{D}\right\|_{\mathrm{F}}^{2}$. Putting these together concludes the proof. □

Lemma 22. *Suppose $\boldsymbol{F}_{1},\boldsymbol{F}_{2},\boldsymbol{F}_{0}\in\mathbb{R}^{2n\times r}$ are three matrices such that*

$$\left\|\boldsymbol{F}_{1}-\boldsymbol{F}_{0}\right\|\left\|\boldsymbol{F}_{0}\right\|\leq\sigma_{r}^{2}\left(\boldsymbol{F}_{0}\right)/2\qquad\text{and}\qquad\left\|\boldsymbol{F}_{1}-\boldsymbol{F}_{2}\right\|\left\|\boldsymbol{F}_{0}\right\|\leq\sigma_{r}^{2}\left(\boldsymbol{F}_{0}\right)/4,$$

where $\sigma_i(\boldsymbol{A})$ stands for the ith largest singular value of $\boldsymbol{A}$. Denote

$$\boldsymbol{R}_1 \triangleq \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{F}_1\boldsymbol{R}-\boldsymbol{F}_0\|_{\mathrm{F}} \qquad \text{and} \qquad \boldsymbol{R}_2 \triangleq \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{F}_2\boldsymbol{R}-\boldsymbol{F}_0\|_{\mathrm{F}}\,.$$

Then the following two inequalities hold true:

$$\|\boldsymbol{F}_1\boldsymbol{R}_1-\boldsymbol{F}_2\boldsymbol{R}_2\| \le 5\frac{\sigma_1^2(\boldsymbol{F}_0)}{\sigma_r^2(\boldsymbol{F}_0)}\|\boldsymbol{F}_1-\boldsymbol{F}_2\| \qquad \text{and} \qquad \|\boldsymbol{F}_1\boldsymbol{R}_1-\boldsymbol{F}_2\boldsymbol{R}_2\|_{\mathrm{F}} \le 5\frac{\sigma_1^2(\boldsymbol{F}_0)}{\sigma_r^2(\boldsymbol{F}_0)}\|\boldsymbol{F}_1-\boldsymbol{F}_2\|_{\mathrm{F}}\,.$$

Proof. This is the same as [MWCC17, Lemma 37]. □

Lemma 23. *Let $\boldsymbol{S}\in\mathbb{R}^{r\times r}$ be a nonsingular matrix. Then for any matrix $\boldsymbol{K}\in\mathbb{R}^{r\times r}$ with $\|\boldsymbol{K}\|\le\sigma_{\min}(\boldsymbol{S})$, one has*

$$\|\mathsf{sgn}(\boldsymbol{S}+\boldsymbol{K})-\mathsf{sgn}(\boldsymbol{S})\| \le \frac{2}{\sigma_{r-1}(\boldsymbol{S})+\sigma_r(\boldsymbol{S})}\|\boldsymbol{K}\|,$$

where $\mathsf{sgn}(\cdot)$ denotes the matrix sign function, i.e. $\mathsf{sgn}(\boldsymbol{A})=\boldsymbol{U}\boldsymbol{V}^\top$ for a matrix $\boldsymbol{A}$ with SVD $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$.

Proof. This is the same as [MWCC17, Lemma 36]. □