Noisy Matrix Completion: Understanding the Stability of Convex Relaxation via Nonconvex Optimization

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

February 13, 2019

Abstract
This paper studies noisy low-rank matrix completion: given partial and corrupted 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 the empirical success.

We make progress towards demystifying the practical efficacy of convex relaxation vis-à-vis random noise. When the rank of the unknown matrix is 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, allowing us to transfer the desired stability guarantees of the nonconvex approach to its convex counterpart.

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

1 Introduction
Suppose that we have 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 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. Solutions that are provably efficient have been developed, among which one of the most notable paradigms is convex relaxation [FHB04, CR09, CT10]. See [DR16, CC18b] for an overview of recent developments on this topic.

This paper focuses on noisy 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 $M^* = [M^*_{ij}]_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$, and suppose that this needs to be performed on the basis of a subset of noisy entries

$$M_{ij} = M^*_{ij} + E_{ij}, \quad (i,j) \in \Omega,$$

1It 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 exhibiting excellent stability in practice, is much 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

$$
\minimize_{Z \in \mathbb{R}^{n \times n}} \frac{1}{2} \sum_{(i,j) \in \Omega} (Z_{ij} - M_{ij})^2 + \lambda \text{rank}(Z),
$$

where $\lambda > 0$ is some proper regularization parameter. In words, this approach optimizes certain trade-off between the goodness of fit (through the squared error 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:

$$
\minimize_{Z \in \mathbb{R}^{n \times n}} g(Z) \triangleq \frac{1}{2} \sum_{(i,j) \in \Omega} (Z_{ij} - M_{ij})^2 + \lambda \|Z\|_*,
$$

where $\|Z\|_*$ denotes the nuclear norm (i.e. the sum of singular values) of $Z$—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, Che15, CP10, CT10, Gro11, Rec11].

In contrast, the performance of convex relaxation remains largely unclear when it comes to more practically relevant noisy scenarios. To begin with, the stability of an equivalent variant of (3) was first studied by Candès and Plan [CP10].2 The estimation error $\|Z_{\text{cvx}} - M^*\|_F$ derived therein, of the solution $Z_{\text{cvx}}$ to (3), is significantly larger than the minimax limit. This does not explain well the effectiveness of (3) in practice. In order to improve the stability guarantees, several variants of (3) have been put forward, most notably by Negahban and Wainwright [NW12] and Koltchinskii et al. [KLT11]; see Section 1.4 for more details. Nevertheless, the stability results derived therein [NW12, KLT11] could often be suboptimal when the magnitude of the noise is not sufficiently large; in fact, their estimation error bounds do not vanish as the size of the noise approaches zero.

All of these give rise to the following natural questions: Where does the convex program (3) stand in terms of its stability vis-à-vis additive noise? Can we establish a theoretical statistical guarantee that matches its practical effectiveness? Is it capable of accommodating a wider (and more practical) range of noise levels?

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 $Z$ as $Z = XY^T$ via the low-rank factors $X, Y \in \mathbb{R}^{n \times r}$, and proceeds by solving the following nonconvex (regularized) least-squares problem directly [KMO10a]

$$
\minimize_{X,Y \in \mathbb{R}^{n \times r}} \frac{1}{2} \sum_{(i,j) \in \Omega} [(XY^T)_{ij} - M_{ij}]^2 + \text{reg}(X, Y),
$$

Here, $\text{reg}(\cdot, \cdot)$ denotes a certain regularization term that promotes additional structural properties. To see its intimate connection [MHT10, SSS11] with the convex program (3), we make the following observation: if

2Technically, [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.
the solution to (3) has rank $r$, then it must coincide with the solution to

$$\minimize_{X,Y} \frac{1}{2} \sum_{(i,j) \in \Omega} [(XY^T)_{ij} - M_{ij}]^2 + \frac{\lambda}{2} \|X\|_F^2 + \frac{\lambda}{2} \|Y\|_F^2. \quad (5)$$

This can be easily verified by recognizing the elementary fact that

$$\|Z\|_* = \inf_{X,Y \in \mathbb{R}^{n \times r} : XY^T = Z} \{\frac{1}{2} \|X\|_F^2 + \frac{1}{2} \|Y\|_F^2\} \quad (6)$$

for any rank-$r$ matrix $Z$ [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 (3) — can possibly hold.

Despite the nonconvexity of (4), simple first-order optimization methods — in conjunction with proper initialization — are often found to be 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 noisy 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 [CLC18, CC18b] for a summary of recent results. Such intriguing stability 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 $M^* = X^*Y^T$, where $X^*, Y^* \in \mathbb{R}^{n \times r}$ are random orthonormal matrices. Each entry $M'_{ij}$ of $M^*$ is revealed with probability $p = 0.2$ independently, and then corrupted by an independent Gaussian noise $E_{ij} \sim \mathcal{N}(0, \sigma^2)$. Throughout the experiment, we set $\lambda = 5\sigma \sqrt{n/p}$. 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 [CLC18] for details). Let $Z_{\text{cvx}}$ (resp. $Z_{\text{ncvx}} = X_{\text{ncvx}}Y_{\text{ncvx}}^T$) be the solution returned by the convex program (3) (resp. the nonconvex program (5)). Figure 1 displays the relative estimation errors of both methods ($\|Z_{\text{cvx}} - M^*\|_F / \|M^*\|_F$ and $\|Z_{\text{ncvx}} - M^*\|_F / \|M^*\|_F$) as well as the relative distance $\|Z_{\text{cvx}} - Z_{\text{ncvx}}\|_F / \|M^*\|_F$ 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. $\|Z_{\text{cvx}} - Z_{\text{ncvx}}\|_F / \|M^*\|_F$ is typically below $10^{-7}$); in comparison, the relative estimation errors of both $Z_{\text{cvx}}$ and $Z_{\text{ncvx}}$ 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 minimax optimal errors vis-à-vis random noise [MWCC17], this suggests that the convex program is equally stable — a phenomenon that was not captured by the prior theory [CP10]. Can we leverage existing theory for the nonconvex scheme to improve the stability analysis of the convex relaxation approach?

### 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 stability guarantees for the convex program (3), which we formally present in this subsection.

Before proceeding, we introduce a few model assumptions that play a crucial role in our theory.
Estimation error: convex
Estimation error: nonconvex
Distance between solutions

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

**Assumption 1.**

(a) (Random sampling) Each index $(i, j)$ belongs to the index set $\Omega$ independently with probability $p$.

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

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

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

$$\|U^*\|_{2,\infty} \leq \sqrt{\frac{\mu}{n}} \|U^*\|_F = \sqrt{\frac{\mu r}{n}} \quad \text{and} \quad \|V^*\|_{2,\infty} \leq \sqrt{\frac{\mu}{n}} \|V^*\|_F = \sqrt{\frac{\mu r}{n}}.$$  

Here, $\|U\|_{2,\infty}$ denotes the largest $\ell_2$ norm of all rows of a matrix $U$.

With these in place, we are ready to present our improved stability guarantees for convex relaxation.

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

1. Any minimizer $Z_{cvx}$ of (3) obeys

$$\|Z_{cvx} - M^*\|_F \lesssim \kappa \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\mu}{p}} \|M^*\|_F,$$  

$$\|Z_{cvx} - M^*\|_\infty \lesssim \sqrt{\kappa^3\mu r} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} \|M^*\|_\infty,$$  

$$\|Z_{cvx} - M^*\| \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\mu}{p}} \|M^*\|.$$  

(7a) (7b) (7c)
2. Letting $Z_{\text{cvx},r} \triangleq \arg \min_{\text{rank}(Z) \leq r} \|Z - Z_{\text{cvx}}\|_F$ be the best rank-$r$ approximation of $Z_{\text{cvx}}$, we have

$$\|Z_{\text{cvx},r} - Z_{\text{cvx}}\|_F \leq \frac{1}{2} \cdot \frac{\sigma}{\sigma_{\min}} \sqrt{n} \|M^*\|,$$

and the error bounds in (7) continue to hold if $Z_{\text{cvx}}$ is replaced by $Z_{\text{cvx},r}$.

Remark 1. Here and throughout, $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ means $\lim_{n \to \infty} |f(n)|/|g(n)| \leq C$ for some constant $C > 0$; $f(n) \gtrsim g(n)$ means $\lim_{n \to \infty} |f(n)|/|g(n)| \geq C$ for some constant $C > 0$; and $f(n) \sim g(n)$ means $C_1 \leq \lim_{n \to \infty} |f(n)|/|g(n)| \leq C_2$ for some constants $C_1, C_2 > 0$. In addition, $\| \cdot \|_\infty$ denotes the entrywise $\ell_\infty$ norm, whereas $\| \cdot \|$ is the spectral norm.

Remark 2. The factor $1/n^4$ in (8) can be replaced by $1/n^c$ for any fixed constant $c > 0$.

Several implications of Theorem 1 follow immediately. The discussions below concentrate on the case where $r, \mu$ and $\kappa$ are all $O(1)$, under our random noise assumption.

- **Improved stability.** Our results reveal that the Euclidean error of any convex optimizer $Z_{\text{cvx}}$ of (3) obeys

$$\|Z_{\text{cvx}} - M^*\|_F \lesssim \sigma \sqrt{n/p},$$

implying that the performance of convex relaxation degrades gracefully as the signal-to-noise ratio decreases. This result matches the minimax lower bound derived in [KLT11, NW12] for the range of noise obeying $\sigma \lesssim \|M^*\|_\infty$, which improves upon prior art in the following aspects:

- Candès and Plan [CP10] provided a stability guarantee in the presence of arbitrary bounded noise. When applied to the random noise model described here, their results yield $\|Z_{\text{cvx}} - M^*\|_F \lesssim \sigma n^{1/3}$, which could be $O(\sqrt{n/p})$ times more conservative than our bound (9).

- Koltchinskii et al. [KLT11] proposed to replace $\sum_{i,j} |(Z_{ij} - M_{ij})^2|$ in (3) with $\sum_{i,j} (Z_{ij} - \frac{1}{p} M_{ij})^2$, where $M_{ij}$ is set to 0 for any unobserved entry (i.e. those with $(i,j) \notin \Omega$). This variant effectively performs singular value thresholding on a rescaled zero-padded data matrix. Under our conditions, their results read (up to some logarithmic factor)

$$\|\hat{Z} - M^*\|_F \lesssim \max \{\sigma, \|M^*\|_\infty\} \sqrt{n/p},$$

where $\hat{Z}$ is the estimate returned by their algorithm. This becomes suboptimal when $\sigma \ll \|M^*\|_\infty$ — a highly relevant regime covered by our analysis.

- Negahban and Wainwright [NW12] proposed to enforce an extra constraint $\|Z\|_\infty \leq \alpha$ when solving (3), in order to explicitly control the spikiness of the estimate. When applied to our model, their error bound is the same as (10) (modulo some log factor), which also becomes increasingly loose as $\sigma$ decreases.

- **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 $Z_{\text{cvx}}$. Our theorem formally justifies that there is no loss of statistical accuracy — measured in terms of either $\| \cdot \|_F$, $\| \cdot \|_1$, 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. Fig. 2 displays the relative estimation errors in both the $\ell_\infty$ norm and the spectral norm, under the same setting as in Fig. 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 $\|Z\|_\infty \leq \alpha$ 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.
Figure 2: The relative estimation error of $Z_{cvx}$ measured by both $\| \cdot \|_\infty$ (i.e. $\| Z_{cvx} - M^* \|_\infty / \| M^* \|_\infty$) and $\| \cdot \|$ (i.e. $\| Z_{cvx} - M^* \| / \| M^* \|)$ vs. the standard deviation $\sigma$ 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.

- **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 $Z$ with $g(Z) \leq g(Z_{cvx}) + \varepsilon$ enjoys the same error bounds as in (7) (with $Z_{cvx}$ replaced by $Z$ 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 reliable estimate.

Finally, we remark that our results are likely suboptimal when $r$ and $\kappa$ are allowed to scale with $n$. Interested readers are referred to Section 4 for more detailed discussions.

## 2 Strategy and novelty

In this section, we introduce the strategy for proving our main theorem. 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 sharp contrast to the noiseless setting, where one clearly anticipates the ground truth $M^*$ 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 notations. Let $P_\Omega(\cdot) : \mathbb{R}^{n \times n} \mapsto \mathbb{R}^{n \times n}$ represent the projection onto the subspace of matrices supported on $\Omega$, namely,

$$[P_\Omega(Z)]_{ij} = \begin{cases} Z_{ij}, & \text{for } (i,j) \in \Omega, \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

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

$$T = \{UA^T + BV^T \mid A, B \in \mathbb{R}^{n \times r} \} \quad (12)$$

Correspondingly, let $P_T(\cdot)$ be the orthogonal projection onto the subspace $T$, that is,

$$P_T(Z) = UU^T Z + ZV V^T - UU^T ZV V^T \quad (13)$$

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

$$X^* = U^*(\Sigma^*)^{1/2} \quad \text{and} \quad Y^* = V^*(\Sigma^*)^{1/2}. \quad (14)$$
The nonconvex problem (5) is equivalent to
\[
\min_{X,Y \in \mathbb{R}^{n \times r}} f(X, Y) \triangleq \frac{1}{2p} \| \Omega (XY^T - M) \|^2_F + \frac{\lambda}{2p} \| X \|^2_F + \frac{\lambda}{2p} \| Y \|^2_F,
\]
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 $Z \in \mathbb{R}^{n \times n}$ is a (primal) solution to (3) with SVD $Z = UV^\top$.\(^3\) As before, let $T$ be the tangent space of $Z$, and let $T^\bot$ be the orthogonal complement of $T$. Then the first-order optimality condition for (3) reads: there exists a matrix $W \in T^\bot$ (called a dual certificate) such that
\[
\frac{1}{\lambda} \mathcal{P}_\Omega (M - Z) = UV^\top + W;
\]
\[
\| W \| \leq 1.
\]
This condition is not only necessary to certify the optimality of $Z$, but also “almost sufficient” in guaranteeing the uniqueness of the solution $Z$; see Appendix B for details.

The challenge then boils down to identifying such a primal-dual pair $(Z, W)$ satisfying the optimality condition (16). For the noise-free case, the primal solution is clearly $Z = M^\star$ 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 $Z$, as it depends on the random noise in a complicated way. In fact, the lack of a suitable guess of $Z$ was the major hurdle that prior works faced when carrying out the duality analysis.

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 $X, Y \in \mathbb{R}^{n \times r}$ is a critical point of (5) with $\text{rank}(X) = \text{rank}(Y) = r$.\(^4\) Then the first-order condition reads
\[
\frac{1}{\lambda} \mathcal{P}_\Omega (M - X Y^\top) Y = X;
\]
\[
\frac{1}{\lambda} \mathcal{P}_\Omega (M - X Y^\top)^\top X = Y.
\]
To develop some intuition about the connection between (16) and (17), let us take a look at the case with $r = 1$. Denote $X = x$ and $Y = y$ and assume that the two factors are “balanced”, namely, $\| x \|_2 = \| y \|_2 \neq 0$. It then follows from (17) that $\lambda^{-1} \mathcal{P}_\Omega (M - xy^\top)$ has a singular value 1, whose corresponding left and right singular vectors are $x/\| x \|_2$ and $y/\| y \|_2$, respectively. In other words, one can express
\[
\frac{1}{\lambda} \mathcal{P}_\Omega (M - xy^\top) = \frac{1}{\| x \|_2 \| y \|_2} xy^\top + W,
\]
where $W$ is orthogonal to the tangent space of $xy^\top$; this is precisely the condition (16a). It remains to argue that (16b) is valid as well. Towards this end, the first-order condition (17) 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 $xy^\top$ is not far away from the ground truth $M^\star$, then one can guarantee $\| W \| < 1$ as required in (16b).

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 $\lambda$.

---

\(^3\)Here and below, we use $Z$ (rather than $Z_{\text{cvx}}$) for notational simplicity, whenever it is clear from the context.

\(^4\)Once again, we abuse the notation $(X,Y)$ (instead of using $(X_{\text{cvx}}, Y_{\text{cvx}})$) for notational simplicity, whenever it is clear from the context.
**Condition 1 (Regularization parameter).** The regularization parameter $\lambda$ satisfies

(a) (Relative to noise) $\|P_\Omega(E)\| < \lambda/8$.

(b) (Relative to nonconvex solution) $\|P_\Omega(XY^T - M^*) - p(XY^T - M^*)\| < \lambda/8$.

**Remark 3.** Condition 1 requires that the regularization parameter $\lambda$ should dominate certain norms of the noise, as well as of the deviation $XY^T - M^*$; as will be seen shortly, the latter condition can be met when $(X,Y)$ is sufficiently close to $(X^*,Y^*)$.

With the above condition in place, the following result demonstrates that a critical point $(X,Y)$ 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 $(X,Y)$ is a critical point of (5) satisfying $\text{rank}(X) = \text{rank}(Y) = r$, and the sampling operator $P_\Omega$ is injective when restricted to the elements of the tangent space $T$ of $XY^T$, namely,

$$P_\Omega(H) = 0 \iff H = 0, \text{ for all } H \in T.$$  \hspace{1cm} (19)

Under Condition 1, the point $Z \triangleq XY^T$ 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.

### 2.3 Approximate nonconvex optimizers

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

$$\nabla_X f(X,Y) \approx 0 \quad \text{and} \quad \nabla_Y f(X,Y) \approx 0.$$ 

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 (19), which is stated below.

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

$$p^{-1} \|P_\Omega(H)\|_F^2 \geq c_{\text{inj}} \|H\|_F^2, \quad \text{for all } H \in T.$$  \hspace{1cm} (20)

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 $(X,Y)$ obeys

$$\|\nabla f(X,Y)\|_F \leq c \frac{\sqrt{c_{\text{inj}}} p}{\kappa} \cdot \frac{\lambda}{\sqrt{\sigma_{\text{min}}}}.$$  \hspace{1cm} (21)

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

$$\|XY^T - Z_{\text{cvx}}\|_F \lesssim \frac{\kappa}{c_{\text{inj}}} \frac{1}{\sqrt{\sigma_{\text{min}}}} \|\nabla f(X,Y)\|_F.$$  \hspace{1cm} (22)
Here, $\epsilon > 0$ is the objective function of (3). 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 $(X, Y)$ results in an approximate optimizer of (3), namely, $g(XY^\top) = g(Z_{cva}) + \epsilon$ for some sufficiently small $\epsilon > 0$. Then for any $Z$ obeying $g(Z) \leq g(XY^\top) = g(Z_{cva}) + \epsilon$, one has

$$\|XY^\top - Z\|_F \leq \frac{\kappa}{c_{mj} \sqrt{\sigma_{\min}}} \|\nabla f (X, Y)\|_F.$$  \hspace{1cm} \text{(23)}$$

As a result, as long as the abovementioned algorithms converge in terms of the objective value, they must return a solution obeying (23), which is exceedingly close to $XY^\top$ if $\nabla f (X, Y)$ is small.

Remark 4. In fact, this lemma continues to hold if $Z_{cva}$ is replaced by any $Z$ obeying $g(Z) \leq g(XY^\top)$, where $g(\cdot)$ is the objective function of (3). 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 $(X, Y)$ results in an approximate optimizer of (3), namely, $g(XY^\top) = g(Z_{cva}) + \epsilon$ for some sufficiently small $\epsilon > 0$. Then for any $Z$ obeying $g(Z) \leq g(XY^\top) = g(Z_{cva}) + \epsilon$, one has

$$\|XY^\top - Z\|_F \leq \frac{\kappa}{c_{mj} \sqrt{\sigma_{\min}}} \|\nabla f (X, Y)\|_F.$$  \hspace{1cm} \text{(23)}$$

As a result, as long as the abovementioned algorithms converge in terms of the objective value, they must return a solution obeying (23), which is exceedingly close to $XY^\top$ if $\nabla f (X, Y)$ is small.

It is clear from Lemma 2 that, as the size of the gradient $\nabla f (X, Y)$ gets smaller, the nonconvex estimate $XY^\top$ 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 $XY^\top$; see (22).

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 $n^2p \geq Cn \log^2 n$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - O(n^{-10})$, one has $\|P_{\Omega} (E)\| \leq \sigma \sqrt{np}$. As a result, Condition 1 holds, i.e. $\|P_{\Omega} (E)\| < \lambda/8$ 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. \hfill \Box

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. Instantiate the notations and assumptions of Theorem 1. Then with probability exceeding $1 - O(n^{-10})$, Condition 1(b), i.e.

$$\|P_{\Omega} (XY^\top - M^\star) - p(XY^\top - M^\star)\| < \lambda/8,$$

and Condition 2 with $c_{mj} = (32\kappa)^{-1}$, i.e.

$$\frac{1}{p} \|P_{\Omega} (H)\|_F^2 \geq \frac{1}{32\kappa} \|H\|_F^2, \quad \forall H \in T$$

hold simultaneously for all $(X, Y)$ obeying

$$\max \left\{ \|X - X^\star\|_{2, \infty}, \|Y - Y^\star\|_{2, \infty} \right\}$$

$$\leq C_{\infty} \kappa \left[ \frac{\sigma}{\sigma_{\min}} \sqrt{n \log n} + \frac{\lambda}{p\sigma_{\min}} \right] \max \left\{ \|X^\star\|_{2, \infty}, \|Y^\star\|_{2, \infty} \right\}.$$  \hspace{1cm} \text{(24)}$$

Here, $T$ denotes the tangent space of $XY^\top$, and $C_{\infty} > 0$ is an absolute constant.

This lemma is a uniform result, namely, the bounds hold irrespective of the statistical dependency between $(X, Y)$ and $\Omega$. As a consequence, to demonstrate the proximity between the convex and the nonconvex solutions (cf. (22)), it remains to identify a point $(\hat{X}, \hat{Y})$ with vanishingly small gradient (cf. (21)) that is sufficiently close to the truth (cf. (24)).

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 (21) and (24). We shall note that Algorithm 1 is not practical since it starts from the ground truth $(X^\star, Y^\star)$; this is adopted to simplify the theoretical analysis.
Algorithm 1 Construction of an approximate primal solution.

**Initialization:** \( X^0 = X^*; Y^0 = Y^* \).

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

\[
X^{t+1} = X^t - \eta \nabla_X f(X^t, Y^t) = X^t - \frac{\eta}{p} \left( \mathcal{P}_\Omega \left( X^t Y^{tT} - M \right) Y^t + \lambda X^t \right); \tag{25a}
\]

\[
Y^{t+1} = Y^t - \eta \nabla_Y f(X^t, Y^t) = Y^t - \frac{\eta}{p} \left( \mathcal{P}_\Omega \left( X^t Y^{tT} - M \right) \right)^T X^t + \lambda Y^t. \tag{25b}
\]

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 (24). 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 \( X^* \) and \( Y^* \) is invariant under global orthonormal transformation, namely, for any orthonormal matrix \( R \in \mathbb{R}^{r \times r} \) one has \( X^* R (Y^* R)^T = X^* Y^* \). 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

\[
H^t \triangleq \arg \min_{R \in \mathcal{O}^{r \times r}} \left( \| X^t R - X^* \|_F^2 + \| Y^t R - Y^* \|_F^2 \right)^{1/2}, \tag{26}
\]

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 1. With probability at least \( 1 - O \left( n^{-3} \right) \), the iterates \( \{(X^t, Y^t)\}_{0 \leq t \leq t_0} \) of Algorithm 1 satisfy

\[
\max \left\{ \| X^t H^t - X^* \|_F, \| Y^t H^t - Y^* \|_F \right\} \leq C_F \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \| X^* \|_F, \tag{27a}
\]

\[
\max \left\{ \| X^t H^t - X^* \|, \| Y^t H^t - Y^* \| \right\} \leq C_{op} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \| X^* \|, \tag{27b}
\]

\[
\max \left\{ \| X^t H^t - X^* \|_2, \| Y^t H^t - Y^* \|_2 \right\} \leq C_{\infty} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \max \left\{ \| X^* \|_2, \| Y^* \|_2 \right\}, \tag{27c}
\]

\[
\min_{0 \leq t < t_0} \| \nabla f(X^t, Y^t) \|_F \leq \frac{1}{\eta} \left( \frac{\lambda}{p \sigma_{\min}} \sqrt{\frac{n \log n}{p}} \right) \max \left\{ \| X^* \|_2, \| Y^* \|_2 \right\}, \tag{28}
\]

where \( C_F, C_{\infty}, C_{op} > 0 \) are some absolute constants, provided that \( \eta \approx 1/(n \kappa^3 \sigma_{\max}) \) 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 Euclidean 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. (28)) guarantees the existence of a small-gradient point within this sequence \( \{(X^t, Y^t)\}_{0 \leq t \leq t_0} \), a somewhat straightforward property of GD tailored to smooth problems [Nes12].

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

\[
\| X^t Y^T - M^* \|_F \leq 3 \kappa C_F \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \| M^* \|_F, \tag{29a}
\]
\[ \|X'Y'' - M^*\|_\infty \leq 3C_\infty \sqrt{\kappa^3 \mu r} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \|M^*\|_\infty \]  
(29b)

\[ \|X'Y'' - M^*\| \leq 3C_{op} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \|M^*\| \]  
(29c)

for all \(0 \leq t \leq t_0\). The proof is deferred to Appendix D.12.

### 2.6 Proof of Theorem 1

Let \(t_* \triangleq \arg \min_{0 \leq t < t_0} \| \nabla f(X', Y') \|_F\), and take \((X_{ncvx}, Y_{ncvx}) = (X', H'^*, Y', H'^*)\) (cf. (26)). It is straightforward to verify that \((X_{ncvx}, Y_{ncvx})\) obeys (i) the small-gradient condition (21), and (ii) the proximity condition (24). We are now positioned to invoke Lemma 2: for any optimizer \(Z_{ncvx}\) of (3), one has

\[ \|Z_{ncvx} - X_{ncvx}Y_{ncvx}^\top\|_F \lesssim \kappa \frac{1}{\sqrt{\sigma_{\text{inj}}}} \| \nabla f(X_{ncvx}, Y_{ncvx}) \|_F \lesssim \frac{\kappa^2 \lambda}{n^5/p} \]  
(30)

The last line arises since \(n \gg \kappa\) — a consequence of the sample complexity condition \(np \gtrsim \kappa^4 \mu^2 r^2 \log^2 n\) (and hence \(n \geq np \gtrsim \kappa^4 \mu^2 r^2 \log n \gg \kappa^4\)). This taken collectively with the property (29) implies that

\[ \|Z_{ncvx} - M^*\|_F \leq \|Z_{ncvx} - X_{ncvx}Y_{ncvx}^\top\|_F + \|X_{ncvx}Y_{ncvx}^\top - M^*\|_F \lesssim \frac{1}{n^4 \mu \sigma_{\text{inj}}} \|M^*\| + \kappa \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \|M^*\|_F \]  
(31)

In other words, since \(X_{ncvx}Y_{ncvx}^\top\) and \(Z_{ncvx}\) are exceedingly close, the error \(Z_{ncvx} - M^*\) is mainly accredited to \(X_{ncvx}Y_{ncvx}^\top - M^*\). Similar arguments also lead to

\[ \|Z_{ncvx} - M^*\| \lesssim \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \|M^*\|, \]  

\[ \|Z_{ncvx} - M^*\|_\infty \lesssim \sqrt{\kappa^3 \mu r} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \|M^*\|_\infty. \]

We are left with proving the properties of \(Z_{ncvx,r}\). Since \(Z_{ncvx,r}\) is defined to be the best rank-\(r\) approximation of \(Z_{ncvx}\), one can invoke (30) to derive

\[ \|Z_{ncvx} - Z_{ncvx,r}\|_F \leq \|Z_{ncvx} - X_{ncvx}Y_{ncvx}^\top\|_F \lesssim \frac{1}{n^4 \mu \sigma_{\text{inj}}} \|M^*\|, \]  
from which (8) follows. Repeating the above calculations implies that (7) holds if \(Z_{ncvx}\) is replaced by \(Z_{ncvx,r}\), 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. \(E = 0\), it amounts to solving the following constrained convex program

\[ \minimize_{Z \in \mathbb{R}^{n \times n}} \|Z\|_* \quad \text{subject to} \quad P_\Omega(Z) = P_\Omega(M^*), \]  
(31)
which enjoys great theoretical success. Informally, this approach enables exact recovery of a rank-$r$ matrix $\mathbf{M}^* \in \mathbb{R}^{n \times n}$ 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]. 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 works provide optimal stability guarantees under our model when $r = O(1)$. Other related papers such as [Klo14,CZ16] include similar stability 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,CFMW17,AFWZ17,LMCC18,DC18,CCFM18]. In the absence of noise and omitting logarithmic factors, [MWCC17] showed that $O(n^3)$ samples are sufficient for vanilla GD to yield $\varepsilon$ accuracy in $O(\log \frac{1}{\varepsilon})$ iterations (without the need of extra regularization procedures); the sample complexity was further improved to $O(n^2)$ by [CLL19]. Apart from gradient descent, other nonconvex methods (e.g. [RS05,JMD10,WYZ12,JNS13,FRW11,Van13,LYX13,Har14,JKN16,RT+11,WCCL16,DC18,GAGG13,ZWL15]) and landscape/geometry properties have been investigated [GLM16,CL17,PKCS17,GJJZ17,SXZ18]; 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 (31) for instance: these results revealed the existence of a solution of rank at most $O(\sqrt{nr})$ — the intrinsic degrees of freedom of a rank-$r$ matrix — which certainly outperforms our current results. It is worth noting that the existing theory for nonconvex matrix factorization typically falls short of providing optimal scaling in $r$ and $\kappa$ [KMO10a,SL16,CW15,MWCC17,CLL19]. Thus, tightening the dependency of sample complexity on $r$ and $\kappa$ might call for new analysis tools.

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 stability 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 $\kappa$.** While our theory is optimal when $r$ and $\kappa$ are both constants, it becomes increasingly looser as either $r$ or $\kappa$ 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 $\kappa$ — which certainly outperforms our current results. It is worth noting that the existing theory for nonconvex matrix factorization typically falls short of providing optimal scaling in $r$ and $\kappa$ [KMO10a,SL16,CW15,MWCC17,CLL19]. Thus, tightening the dependency of sample complexity on $r$ and $\kappa$ might call for new analysis tools.

- **Approximate low-rank structure.** So far our theory is built upon the assumption that the ground-truth matrix $\mathbf{M}^*$ is exactly low-rank, which falls short of accommodating the more realistic scenario where $\mathbf{M}^*$ 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]. 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 stability guarantees for such problems.

**Acknowledgements**

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**


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

\[ \|X^*\|_{2,\infty} \leq \sqrt{\mu r/n} \|X^*\| \quad \text{and} \quad \|Y^*\|_{2,\infty} \leq \sqrt{\mu r/n} \|Y^*\|. \]  

(32)

This follows from

\[ \|X^*\|_{2,\infty} = \|U^* (\Sigma^*)^{1/2}\|_{2,\infty} \leq \|U^*\|_{2,\infty} \| (\Sigma^*)^{1/2}\| \leq \sqrt{\mu r/n} \|X^*\|. \]

The bound for \( Y^* \) follows from the same argument.

Finally, for notational convenience, we shall often denote

\[ \mathcal{P}_{\Omega}^{\text{debias}}(B) \triangleq \mathcal{P}_{\Omega}(B) - pB, \quad \text{for all } B \in \mathbb{R}^{n \times n}. \]  

(33)

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 \( Z = U \Sigma V^T \) be the SVD of \( Z \in \mathbb{R}^{n \times n} \). Denote by \( T \) be the tangent space of \( Z \) and by \( T^\perp \) its orthogonal complement. Suppose that there exists \( W \in T^\perp \) such that

\[ \frac{1}{\lambda} \mathcal{P}_{\Omega}(M - Z) = UV^T + W. \]  

(34)

Then \( Z \) is the unique minimizer of (3) if

1. \( \|W\| < 1 \);

2. The operator \( \mathcal{P}_{\Omega}(\cdot) \) restricted to elements in \( T \) is injective, i.e. \( \mathcal{P}_{\Omega}(H) = 0 \) implies \( H = 0 \) for any \( H \in T \).

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

\[ UV^T + W \in \partial \| Z \|_*, \]

where \( \partial \| Z \|_* \) denotes the subdifferential of \( \| \cdot \|_* \) at \( Z \). This combined with (34) reveals that

\[ \frac{1}{\lambda} \mathcal{P}_{\Omega}(M - Z) \in \partial \| Z \|_*, \]  

(35)

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

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

Claim 1. Suppose that \( Z_1 \) and \( Z_2 \) are both minimizers of (3). Then one has \( \mathcal{P}_{\Omega}(Z_1) = \mathcal{P}_{\Omega}(Z_2) \).

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

\[ g(Z + H) = \frac{1}{2} \| \mathcal{P}_{\Omega}(Z + H - M) \|_F^2 + \lambda \| Z + H \|_*, \]

(36)

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

\[ \mathcal{P}_{T}(S) = UV^T, \quad \| \mathcal{P}_{T^\perp}(S) \| \leq 1 \quad \text{and} \quad \langle \mathcal{P}_{T^\perp}(S), \mathcal{P}_{T^\perp}(H) \rangle = \| \mathcal{P}_{T^\perp}(H) \|_*. \]  

(37)

Using the convexity of \( \| \cdot \|_* \), one can further lower bound (36) by

\[ g(Z + H) \geq \frac{1}{2} \| \mathcal{P}_{\Omega}(Z - M) \|_F^2 + \lambda (\| Z \|_* + \langle S, H \rangle) \]
\[ g(Z) + \lambda \langle S, H \rangle \]
\[ = g(Z) + \lambda \langle UV^\top + W, H \rangle + \lambda \langle S - UV^\top - W, H \rangle \]
\[ = g(Z) + \lambda \langle S - UV^\top - W, H \rangle \quad \text{(i)} \]
\[ = g(Z) + \lambda \langle \mathcal{P}_{T^\perp}(S) - W, H \rangle . \quad \text{(ii)} \]

Here, (i) follows from our assumption that \( UV^\top + W \) is supported on \( \Omega \) (cf. (34)) and the fact that \( \mathcal{P}_\Omega(H) = 0 \), and (ii) holds since \( \mathcal{P}_\mathcal{T}(S) = UV^\top \) (cf. (37)). We can now expand the above expression as
\[ g(Z + H) \geq g(Z) + \lambda \langle \mathcal{P}_{T^\perp}(S), \mathcal{P}_{T^\perp}(H) \rangle - \lambda \langle W, \mathcal{P}_{T^\perp}(H) \rangle \]
\[ \geq g(Z) + \lambda (1 - \|W\|) \|\mathcal{P}_{T^\perp}(H)\|_*, \quad (38) \]
where the last inequality holds by using the last property of (37) and invoking the elementary inequality
\[ \langle W, \mathcal{P}_{T^\perp}(H) \rangle \leq \|W\| \|\mathcal{P}_{T^\perp}(H)\|_* . \]

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

**Proof of Claim 1.** Consider any minimizers \( Z_1 \neq Z_2 \), and suppose instead that \( \mathcal{P}_\Omega(Z_1 - Z_2) \neq 0 \). For any \( 0 < \alpha < 1 \), define
\[ Z_\alpha \triangleq \alpha Z_1 + (1 - \alpha) Z_2 . \]

Since \( \| \cdot \|_* \) is convex, we have
\[ g(Z_\alpha) = \frac{1}{2} \| \mathcal{P}_\Omega(\alpha Z_1 + (1 - \alpha) Z_2 - M) \|_F^2 + \lambda \alpha \| Z_1 \|_* + (1 - \alpha) \| Z_2 \|_* . \]

Furthermore, by the strong convexity of \( \| \cdot \|_F^2 \) we have
\[ g(Z_\alpha) < \frac{1}{2} \| \mathcal{P}_\Omega(Z_1 - M) \|_F^2 + \frac{1}{2} \| \mathcal{P}_\Omega(Z_2 - M) \|_F^2 + \alpha \lambda \| Z_1 \|_* + (1 - \alpha) \lambda \| Z_2 \|_* . \]
This contradicts the fact that \( Z_1 \) is a minimizer of (3), thus completing the proof. \( \square \)

**C Connections between convex and nonconvex solutions**

**C.1 Proof of Lemma 1**

First of all, since \( (X, Y) \) is a stationary point of (5), we have the first-order optimality conditions
\[ \mathcal{P}_\Omega(M - XY^\top) Y = \lambda X; \quad (40a) \]
\[ [\mathcal{P}_\Omega(M - XY^\top)]^\top X = \lambda Y. \quad (40b) \]

As an immediate consequence, one has
\[ X^\top X = \lambda^{-1} X^\top \mathcal{P}_\Omega(M - XY^\top) Y = Y^\top Y . \quad (41) \]

In words, any stationary point \( (X, Y) \) has “balanced” scale.

Let \( U \Sigma V^\top \) be the singular value decomposition of \( XY^\top \) with \( U, V \in \mathbb{R}^{n \times r} \) orthonormal and \( \Sigma \in \mathbb{R}^{r \times r} \) diagonal. In view of the balanced scale of \( (X, Y) \) (namely, (41)) and Lemma 20, we can write
\[ X = U \Sigma^{1/2} R \quad \text{and} \quad Y = V \Sigma^{1/2} R \quad (42) \]
for some orthonormal matrix \( R \in \mathbb{R}^{r \times r} \). Substitution into (40) results in

\[
\mathcal{P}_\Omega (M - XY^\top) V = \lambda U; 
\]

\[
[\mathcal{P}_\Omega (M - XY^\top)]^T U = \lambda V, 
\]

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

\[
\frac{1}{\lambda} \mathcal{P}_\Omega (M - XY^\top) = UV^\top + W, 
\]

where \( W \in T^\perp \); recall that \( T \) is the tangent space of \( XY^\top \) and also \( UV^\top \). In view of Lemma 6, it suffices to show that \( \|W\| < 1 \), which is the content of the rest of the proof.

One can rewrite \( \mathcal{P}_\Omega (M - XY^\top) \) as

\[
\mathcal{P}_\Omega (M - XY^\top) = p (M^* - XY^\top) + \mathcal{P}_{\Omega}^{\text{debias}} (M^* - XY^\top) + \mathcal{P}_\Omega (E).
\]

Substitute this identity into (43) and rearrange terms to obtain

\[
[pM^* + \mathcal{P}_{\Omega}^{\text{debias}} (M^* - XY^\top) + \mathcal{P}_\Omega (E)] V = U (p\Sigma + \lambda I_r); 
\]

\[
[pM^* + \mathcal{P}_{\Omega}^{\text{debias}} (M^* - XY^\top) + \mathcal{P}_\Omega (E)]^T U = V (p\Sigma + \lambda I_r).
\]

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

\[
pM^* + \mathcal{P}_{\Omega}^{\text{debias}} (M^* - XY^\top) + \mathcal{P}_\Omega (E),
\]

which is equivalent to saying that

\[
pM^* + \mathcal{P}_{\Omega}^{\text{debias}} (M^* - XY^\top) + \mathcal{P}_\Omega (E) = U (p\Sigma + \lambda I_r) V^\top + \lambda W_2, 
\]

for some \( W_2 \in T^\perp \). One can then derive from (44) that

\[
\|\mathcal{P}_\Omega (M - XY^\top)\| = \frac{1}{\lambda} \left\|pM^* - pXY^\top + \mathcal{P}_{\Omega}^{\text{debias}} (M^* - XY^\top) + \mathcal{P}_\Omega (E)\right\| 
\]

\[
= \frac{1}{\lambda} \left\|pM^* - pXY^\top + \mathcal{P}_{\Omega}^{\text{debias}} (M^* - XY^\top) + \mathcal{P}_\Omega (E)\right\| 
\]

\[
= \frac{1}{\lambda} \left\|U (p\Sigma + \lambda I_r) V^\top + \lambda W_2\right\| 
\]

where (i), (ii) and (iv) arise from the facts that \( UV^\top \in T, XY^\top \in T \) and \( U(p\Sigma + \lambda I_r)V^\top \in T \), respectively, and (iii) relies on the identity (45).

It then suffices to control \( \|W_2\| \). To this end, apply Weyl’s inequality to (45) to obtain that: for \( r + 1 \leq i \leq n \), the \( i \)th largest singular value of \( U(p\Sigma + \lambda I_r)V^\top + \lambda W_2 \) obeys

\[
\sigma_i (U (p\Sigma + \lambda I_r) V^\top + \lambda W_2) \leq p \sigma_i (M^*) + \left\|\mathcal{P}_{\Omega}^{\text{debias}} (M^* - XY^\top) + \mathcal{P}_\Omega (E)\right\| 
\]

\[
\leq \left\|\mathcal{P}_{\Omega}^{\text{debias}} (M^* - XY^\top)\right\| + \left\|\mathcal{P}_\Omega (E)\right\| < \lambda,
\]

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

\[
\|\mathcal{P}_\Omega (M - XY^\top)\| = \|W_2\| = \frac{1}{\lambda} \max_{r + 1 \leq i \leq n} \sigma_i (U (p\Sigma + \lambda I_r) V^\top + \lambda W_2) < 1.
\]

This together with Lemma 6 completes the proof.

\(^5\)Here, the pre-factor \( \lambda \) 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 (15) is given by

\[
\nabla f(X, Y) = \frac{1}{p} \left[ \nabla \Omega (XY^T - M) Y + \lambda X \right] = \frac{1}{p} \left[ \nabla \Omega (XY^T - M) Y + \lambda X \right],
\]

which together with the small-gradient assumption \( \| \nabla f(X, Y) \|_F \leq c \lambda \sqrt{c_{\text{inj}}} \sigma_{\min}/\kappa^2/p \) implies that

\[
\begin{align*}
\| \nabla \Omega (XY^T - M) Y + \lambda X \|_F &\leq p \| \nabla f(X, Y) \|_F \\
&\leq c \lambda \sqrt{c_{\text{inj}}} \sigma_{\min}/\kappa^2; \\
\| (\nabla \Omega (XY^T - M))^T X + \lambda Y \|_F &\leq p \| \nabla f(X, Y) \|_F \\
&\leq c \lambda \sqrt{c_{\text{inj}}} \sigma_{\min}/\kappa^2.
\end{align*}
\]

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

\[
\sigma_{\min}/2 \leq \sigma_{\min}(\Sigma) \leq \sigma_{\max}(\Sigma) \leq 2\sigma_{\max}.
\]

This can be easily seen from the following two inequalities

\[
\begin{align*}
\sigma_{\max}(\Sigma) &= \| XY^T \| \leq \| X \| \| Y \| \leq 2\sigma_{\max}; \\
\sigma_{\min}(\Sigma) &= \sigma_{\min}(XY^T) \geq \sigma_{\min}(X) \sigma_{\min}(Y) \geq \sigma_{\min}/2.
\end{align*}
\]

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

Claim 2. Instate the notations and assumptions of Lemma 2. One has

\[
P_\Omega (XY^T - M) = -\lambda UV^T + R,
\]

where \( R \) is some residual matrix satisfying

\[
\| P_T(R) \|_F \leq \frac{72 \kappa}{\sqrt{\sigma_{\min}}} p \| \nabla f(X, Y) \|_F \quad \text{and} \quad \| P_{T^\perp}(R) \| < \lambda/2.
\]

With Claim 2 in place, we are ready to prove Lemma 2. Let \( Z_{\text{cvx}} \) be any minimizer of (3) and denote \( \Delta \triangleq Z_{\text{cvx}} - XY^T \). The proof can be divided into the following steps.

- First, show that the difference \( \Delta \) primarily lies in the tangent space of \( XY^T \); see (56).
- Next, utilize this information to connect \( \| P_\Omega(\Delta) \|_F^2 \) with the size of the gradient \( \nabla f(X, Y) \); see (58).
- In the end, obtain a lower bound on \( \| \Delta \|_F^2 \) in terms of \( \| \Delta \|_F \) using the injectivity property; see (59).

The desired upper bound on \( \| \Delta \|_F \) stated 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 \( Z_{\text{cvx}} = XY^T + \Delta \) reveals that

\[
\frac{1}{2} \| P_\Omega (XY^T + \Delta - M) \|_F^2 + \lambda \| XY^T + \Delta \|_* \leq \frac{1}{2} \| P_\Omega (XY^T - M) \|_F^2 + \lambda \| XY^T \|_*.
\]

A little algebra allows us to rearrange terms as follows

\[
\frac{1}{2} \| P_\Omega(\Delta) \|_F^2 \leq -\langle P_\Omega (XY^T - M), \Delta \rangle + \lambda \| XY^T \|_* - \lambda \| XY^T + \Delta \|_*.
\]

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

\[
\| XY^T + \Delta \|_* \geq \| XY^T \|_* + \langle UV^T + W, \Delta \rangle,
\]

21
for any $W \in T^\perp$ obeying $\|W\| \leq 1$, where $UV^T + W$ serves as a subgradient of $\|\cdot\|_*$ at $XY^T$. In what follows, we shall pick $W$ such that $(W, \Delta) = \|P_{T^\perp}(\Delta)\|_*$. Combining this with (50) and (51), we reach
\begin{align*}
\frac{1}{2} \|P_\Omega(\Delta)\|_F^2 & \leq - \langle P_\Omega(XX^T - M), \Delta \rangle - \lambda \langle UV^T, \Delta \rangle - \lambda \langle W, \Delta \rangle \\
& = - \langle P_T(R), \Delta \rangle - \langle P_{T^\perp}(R), \Delta \rangle - \lambda \langle UV^T, \Delta \rangle - \lambda \|P_{T^\perp}(\Delta)\|_* .
\end{align*}
(52)

This together with the decomposition (48) leads to
\begin{align*}
0 \leq \frac{1}{2} \|P_\Omega(\Delta)\|_F^2 & \leq - \langle R, \Delta \rangle - \lambda \|P_{T^\perp}(\Delta)\|_* \\
& = - \langle P_T(R), \Delta \rangle - \langle P_{T^\perp}(R), \Delta \rangle - \lambda \|P_{T^\perp}(\Delta)\|_* ,
\end{align*}
(53)

and therefore
\begin{align*}
\langle P_T(R), \Delta \rangle + \langle P_{T^\perp}(R), \Delta \rangle + \lambda \|P_{T^\perp}(\Delta)\|_* & \leq 0.
\end{align*}
(54)

In addition, elementary inequalities give
\begin{align*}
- \|P_T(R)\|_F \|P_T(\Delta)\|_F - \|P_{T^\perp}(R)\| \|P_{T^\perp}(\Delta)\|_* + \lambda \|P_{T^\perp}(\Delta)\|_* \\
\leq \langle P_T(R), \Delta \rangle + \langle P_{T^\perp}(R), \Delta \rangle + \lambda \|P_{T^\perp}(\Delta)\|_* & \leq 0.
\end{align*}

From the condition (49) we have $\|P_{T^\perp}(R)\| \leq \lambda/2$ and hence
\begin{align*}
\|P_T(R)\|_F \|P_T(\Delta)\|_F \geq - \|P_{T^\perp}(R)\| \|P_{T^\perp}(\Delta)\|_* + \lambda \|P_{T^\perp}(\Delta)\|_* \geq \frac{1}{2} \|P_{T^\perp}(\Delta)\|_* ,
\end{align*}
(55)

which together with the condition (49) on $\|P_T(R)\|_F$ and the small gradient assumption (21) yields
\begin{align*}
\|P_{T^\perp}(\Delta)\|_* \leq 144 \frac{p}{\lambda \sqrt{\sigma_{\text{min}}}} \|\nabla f(X, Y)\|_F \|P_T(\Delta)\|_F \leq 144c \sqrt{e_{\text{min}}p} \|P_T(\Delta)\|_F .
\end{align*}
(56)

This essentially means that $\Delta$ lies primarily in the tangent space of $XY^T$ for $c$ sufficiently small. As an immediate consequence,
\begin{align*}
\|P_{T^\perp}(\Delta)\|_F \leq \|P_{T^\perp}(\Delta)\|_* \leq 144c \sqrt{e_{\text{min}}p} \|P_T(\Delta)\|_F \leq \|P_T(R)\|_F ,
\end{align*}
(57)

as long as $c$ is sufficiently small. Note that we also use the elementary fact that $e_{\text{min}} \leq 1/p$ (otherwise we will have the contradictory inequality $p^{-1}\|P_\Omega(H)\|_F^2 \geq c_{\text{min}} \|H\|_F^2 > p^{-1}\|H\|_F^2$).

2. Continue the upper bound in (53) to get
\begin{align*}
\frac{1}{2} \|P_\Omega(\Delta)\|_F^2 & \leq - \langle P_T(R), \Delta \rangle - \langle P_{T^\perp}(R), \Delta \rangle - \lambda \|P_{T^\perp}(\Delta)\|_* \\
& \leq \langle P_T(R)\|_F \|P_T(\Delta)\|_F - \frac{1}{2} \|P_{T^\perp}(\Delta)\|_* .
\end{align*}

Here, the last line uses the fact that $- \langle P_{T^\perp}(R), \Delta \rangle \leq \|P_{T^\perp}(R)\| \|P_{T^\perp}(\Delta)\|_* \leq \frac{1}{2} \|P_{T^\perp}(\Delta)\|_*$, which follows from (49). Therefore, using the condition (49) we reach
\begin{align*}
\frac{1}{2} \|P_\Omega(\Delta)\|_F^2 & \leq \|P_T(R)\|_F \|P_T(\Delta)\|_F \leq 72 \frac{p}{\sqrt{\sigma_{\text{min}}}} \|\nabla f(X, Y)\|_F \|\Delta\|_F .
\end{align*}
(58)

3. We are left with lower bounding $\|P_\Omega(\Delta)\|_F^2$. Using the decomposition $\Delta = P_T(\Delta) + P_{T^\perp}(\Delta)$, we get
\begin{align*}
\frac{1}{\sqrt{p}} \|P_\Omega(\Delta)\|_F = \frac{1}{\sqrt{p}} \|P_T(\Omega) + P_{T^\perp}(\Omega)\|_F \geq \frac{1}{\sqrt{p}} \|P_T(\Delta)\|_F - \frac{1}{\sqrt{p}} \|P_{T^\perp}(\Delta)\|_F \\
\geq \sqrt{e_{\text{min}}} \|P_T(\Delta)\|_F - \frac{1}{\sqrt{p}} \|P_{T^\perp}(\Delta)\|_F ,
\end{align*}

where the last inequality follows from the injectivity assumption (20). In addition, (56) implies
\begin{align*}
\frac{1}{\sqrt{p}} \|P_{T^\perp}(\Delta)\|_F \leq \frac{1}{\sqrt{p}} \|P_{T^\perp}(\Delta)\|_* \leq \frac{1}{\sqrt{p}} 144c \sqrt{e_{\text{min}}p} \|P_T(\Delta)\|_F \leq \frac{\sqrt{e_{\text{min}}}}{2} \|P_T(\Delta)\|_F .
\end{align*}
as long as $c$ is sufficiently small. As a result,
\[
\frac{1}{\sqrt{p}} \| \mathcal{P}_\Omega(\Delta) \|_F \geq \frac{\sqrt{\inj}}{2} \| \mathcal{P}_T(\Delta) \|_F.
\]
In addition, by (57) we have
\[
\| \Delta \|_F \leq \| \mathcal{P}_T(\Delta) \|_F + \| \mathcal{P}_{T^\perp}(\Delta) \|_F \leq 2 \| \mathcal{P}_T(\Delta) \|_F,
\]
and therefore
\[
\frac{1}{\sqrt{p}} \| \mathcal{P}_\Omega(\Delta) \|_F \geq \frac{\sqrt{\inj}}{2} \| \mathcal{P}_T(\Delta) \|_F \geq \frac{\sqrt{\inj}}{4} \| \Delta \|_F.
\]
Taking (58) and (59) collectively yields
\[
\frac{\inj}{12} \| \Delta \|_F^2 \leq \frac{1}{2p} \| \mathcal{P}_\Omega(\Delta) \|_F^2 \leq 72 \kappa \sqrt{\frac{1}{\sigma_{\min}}} \| \nabla f(X,Y) \|_F \| \Delta \|_F,
\]
thus indicating that
\[
\| \Delta \|_F \leq \frac{\kappa}{\inj} \frac{1}{\sqrt{\sigma_{\min}}} \| \nabla f(X,Y) \|_F.
\]

### 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 $U\Sigma V^T$ be the SVD of $XY^T$. There exists an invertible matrix $Q \in \mathbb{R}^{r \times r}$ such that $X = U\Sigma^{1/2}Q$, $Y = V\Sigma^{1/2}Q^{-T}$ and
\[
\| \Sigma Q - \Sigma Q^{-1} \|_F \leq 8\sqrt{n} \frac{p}{\lambda \sqrt{\sigma_{\min}}} \| \nabla f(X,Y) \|_F \leq 8c\sqrt{\inj p / \kappa},
\]
where $U_Q\Sigma_QV_Q^T$ is the SVD of $Q$.

In light of the assumptions (46), one has
\[
\mathcal{P}_\Omega(XY^T - M)Y = -\lambda X + B_1 \quad \text{and} \quad [\mathcal{P}_\Omega(XY^T - M)]^T X = -\lambda Y + B_2
\]
for some $B_1 \in \mathbb{R}^{n \times r}$ and $B_2 \in \mathbb{R}^{n \times r}$, where $\max \{ \| B_1 \|_F, \| B_2 \|_F \} \leq p \| \nabla f(X,Y) \|_F$. Recall that
\[
\mathcal{P}_\Omega(XY^T - M) = -\lambda UV^T + R.
\]
In the sequel, we shall prove the upper bounds on both $\| \mathcal{P}_T(R) \|_F$ and $\| \mathcal{P}_{T^\perp}(R) \|_F$ separately.

1. From the definition of $\mathcal{P}_T(\cdot)$ (see (13)), we have
\[
\| \mathcal{P}_T(R) \|_F = \| UU^T R(I - VV^T) + RVV^T \|_F \\
\leq \| U^T R(I - VV^T) \|_F + \| R \|_F \\
\leq \| U^T R \|_F + \| R \|_F.
\]
In addition, invoke Claim 3 to obtain
\[
X = U\Sigma^{1/2}Q \quad \text{and} \quad Y = V\Sigma^{1/2}Q^{-T}
\]
for some invertible matrix $Q \in \mathbb{R}^{r \times r}$, whose SVD $U_Q\Sigma_QV_Q^T$ obeys (60). Combine (61) and (62) to see
\[
-\lambda UV^TY + RY = -\lambda X + B_1,
\]
which together with (64) yields
\[
RV = \lambda U\Sigma^{1/2}(I_r - QQ^T)\Sigma^{-1/2} + B_1Q^T\Sigma^{-1/2}.
\]
Apply the triangle inequality to get
\[
\|RV\|_F \leq \|\lambda U\Sigma^{1/2} (I_r - QQ^\top) \Sigma^{-1/2}\|_F + \|B, Q^\top \Sigma^{-1/2}\|_F
\leq \lambda \|\Sigma^{1/2}\| \|\Sigma^{-1/2}\| \|IQQ^\top - I_r\|_F + \|Q\| \|\Sigma^{-1/2}\| \|B_1\|_F.
\] (65)

In order to further upper bound (65), we first recognize that (47) implies
\[
\|\Sigma^{1/2}\| \leq \sqrt{2\sigma_{\max}}, \quad \text{and} \quad \|\Sigma^{-1/2}\| = 1/\sqrt{\sigma_{\min}}(\Sigma) \leq \sqrt{2/\sigma_{\min}}.
\]

Second, Claim 3 yields
\[
\|\Sigma_Q - \Sigma_Q^{-1}\|_F \leq 8\sqrt{\kappa} \frac{p}{\lambda\sqrt{\sigma_{\min}}} \|\nabla f(X, Y)\|_F \leq 8\epsilon c\sqrt{c_{nij}/p} \kappa \ll 1,
\]
with the proviso that \( c \) is sufficiently small. Here we have used the facts that \( c_{nij} \leq 1/p \) and that \( \kappa \geq 1 \).

This in turn implies that \( \|Q\| = \|\Sigma_Q\| \leq 2 \). Putting the above bounds together yields
\[
\|RV\|_F \leq \lambda \sqrt{2\sigma_{\max}} \sqrt{\frac{2}{\lambda\sqrt{\sigma_{\min}}} \|\Sigma_Q - I_r\|_F} + 2\sqrt{\frac{2}{\lambda\sqrt{\sigma_{\min}}} \|\nabla f(X, Y)\|_F}
\leq 2\lambda \sqrt{2\sigma_{\max}} \sqrt{\frac{2}{\lambda\sqrt{\sigma_{\min}}} \|\Sigma_Q - \Sigma_Q^{-1}\|_F} + 2\sqrt{\frac{2}{\lambda\sqrt{\sigma_{\min}}} \|\nabla f(X, Y)\|_F}
\leq 2\lambda \sqrt{2\sigma_{\max}} \sqrt{\frac{2}{\lambda\sqrt{\sigma_{\min}}} \|\nabla f(X, Y)\|_F} + 2\sqrt{\frac{2}{\lambda\sqrt{\sigma_{\min}}} \|\nabla f(X, Y)\|_F}
\leq 36\kappa \frac{p}{\sqrt{\sigma_{\min}}} \|\nabla f(X, Y)\|_F.
\]

Similarly we can show that \( \|U^\top R\|_F \leq 36\kappa p \|\nabla f(X, Y)\|_F/\sqrt{\sigma_{\min}} \). These bounds together with (63) result in
\[
\|P_T(R)\|_F \leq 72\kappa \frac{p}{\sqrt{\sigma_{\min}}} \|\nabla f(X, Y)\|_F.
\] (66)

2. We now move on to bounding \( \|P_{T^\bot}(R)\| \). By the definition of \( P_{\Omega}^\text{debias}(\cdot) \) in (33), we can rearrange (61) to get
\[
[pM^* + P_\Omega(E) - P_{\Omega}^\text{debias}(XY^\top - M^*)] Y = pXY^\top Y + \lambda X - B_1,
[pM^* + P_\Omega(E) - P_{\Omega}^\text{debias}(XY^\top - M^*)]^\top X = pXY^\top X + \lambda Y - B_2.
\]

In view of the representation \( X = U\Sigma^{1/2} Q \) and \( Y = V\Sigma^{1/2} Q^{-\top} \), the above identities are equivalent to
\[
[pM^* + P_\Omega(E) - P_{\Omega}^\text{debias}(XY^\top - M^*)] V = pU\Sigma + \lambda U\Sigma^{1/2} QQ^\top \Sigma^{-1/2} - B_1 Q^\top \Sigma^{-1/2},
[pM^* + P_\Omega(E) - P_{\Omega}^\text{debias}(XY^\top - M^*)]^\top U = pV\Sigma + \lambda V\Sigma^{1/2} Q^{-\top} Q^{-1} \Sigma^{-1/2} - B_2 Q^{-1} \Sigma^{-1/2}.
\]

Letting
\[
pM^* + P_\Omega(E) - P_{\Omega}^\text{debias}(XY^\top - M^*) = pU\Sigma V^\top + \lambda U\Sigma^{1/2} QQ^\top \Sigma^{-1/2} V^\top + \tilde{R}
\] (67)
for some residual matrix \( \tilde{R} \in \mathbb{R}^{n \times n} \), we have
\[
\|P_{T^\bot}(R)\| \overset{(i)}{=} \|P_{T^\bot} [P_\Omega (XY^\top - M^* - E)]\|
\overset{(ii)}{=} \|P_{T^\bot} [p (XY^\top - M^*) + P_{\Omega}^\text{debias}(XY^\top - M^*) - P_\Omega (E)]\|
\overset{(iii)}{=} \|P_{T^\bot} [pM^* + P_\Omega(E) - P_{\Omega}^\text{debias}(XY^\top - M^*)]\|
\overset{(iv)}{=} \|P_{T^\bot}(\tilde{R})\|,
\] (68)
where (i) follows from the definition of $R$ and the fact that $UV^T \in T$, (ii) uses the definition of $P_{\Omega \text{ debias}}(\cdot)$, (iii) relies on the fact that $XY^T \in T$, and (iv) applies (67) and the facts that $U\Sigma V^T \in T$ and that $U\Sigma^{1/2}QQ^T \Sigma^{-1/2}V^T \in T$. Therefore, it suffices to bound $\|P_{\perp T}(\hat{R})\|$. Rewrite (67) as

$$pM^* + P_{\Omega}(E) - P_{\Omega \text{ debias}}(XY^T - M^*) - P_T(\hat{R}) = U(p\Sigma + \lambda \Sigma^{1/2}QQ^T \Sigma^{-1/2})V^T + P_{\perp T}(\hat{R}).$$

(69)

Suppose for the moment that

$$\|P_T(\hat{R})\| \leq \lambda/4.$$  

(70)

This together with the assumptions that $\|P_{\Omega \text{ debias}}(XY^T - M^*)\| < \lambda/8$ and $\|P_{\Omega}(E)\| < \lambda/8$ reveals that

$$\|P_{\Omega}(E) - P_{\Omega \text{ debias}}(XY^T - M^*) - P_T(\hat{R})\| < \lambda/2.$$  

By Weyl’s inequality, one has

$$\sigma_i \left[ U(p\Sigma + \lambda \Sigma^{1/2}QQ^T \Sigma^{-1/2})V^T + P_{T \perp}(\hat{R}) \right] < \sigma_i (pM^*) + \lambda/2 < \lambda/2,$$

(71)

for $r + 1 \leq i \leq n$, where $\sigma_i(A)$ denotes the $i$th largest singular value of a matrix $A$. Here, we have used the fact that $M^*$ has rank $r$. In addition, one sees that

$$\left\| \Sigma^{1/2}QQ^T \Sigma^{-1/2} - I_r \right\| = \left\| \Sigma^{1/2}(QQ^T - I_r) \Sigma^{-1/2} \right\|$$

$$\leq \left\| \Sigma^{1/2} \right\| \left\| \Sigma^{-1/2} \right\| \left\| QQ^T - I_r \right\|_{F}.$$  

Note that in (65), we have obtained the bound

$$\left\| \Sigma^{1/2} \right\| \left\| \Sigma^{-1/2} \right\| \left\| QQ^T - I_r \right\|_{F} \leq 2\sqrt{2\sigma_{\max}} \sqrt{\frac{2}{\sigma_{\min}} 8c \sqrt{c_{\text{inj}} p / \kappa}} \leq \frac{1}{10}$$

as long as $c$ is sufficiently small. Therefore we know

$$\sigma_i \left[ U(p\Sigma + \lambda \Sigma^{1/2}QQ^T \Sigma^{-1/2})V^T \right] \geq \sigma_r \left[ p\Sigma + \lambda I_r + \lambda (\Sigma^{1/2}QQ^T \Sigma^{-1/2} - I_r) \right]$$

$$\geq \sigma_r (p\Sigma + \lambda I_r) - \lambda \left\| \Sigma^{1/2}QQ^T \Sigma^{-1/2} - I_r \right\|$$

$$\geq 0 + \sigma_r (\lambda I_r) - \lambda \left\| \Sigma^{1/2}QQ^T \Sigma^{-1/2} - I_r \right\|$$

$$\geq \lambda - \lambda/10 > \lambda/2,$$

for any $1 \leq i \leq r$, where the second inequality uses Weyl’s inequality again. This combined with (68) and (71) yields

$$\|P_{T \perp}(\hat{R})\| = \|P_{T \perp}(\hat{R})\| < \lambda/2;$$

this happens because there are at least $n - r$ singular values of $U(p\Sigma + \lambda \Sigma^{1/2}QQ^T \Sigma^{-1/2}) V^T + P_{T \perp}(\hat{R})$ are no larger than $\lambda/2$ and they cannot correspond to directions simultaneously in the column space spanned by $U$ and the row space spanned by $V^T$.

The proof is then complete by verifying (70). To this end, observe that

$$\hat{R}V = -B_1Q^T \Sigma^{-1/2}, \quad \hat{R}^T U = \lambda V \Sigma^{1/2}Q^T \Sigma^{-1/2} - \lambda V \Sigma^{-1/2}QQ^T \Sigma^{1/2} - B_2Q^{-1} \Sigma^{-1/2}.$$  

Then following similar technique used to bound $\|P_T(\hat{R})\|$, we have

$$\|P_T(\hat{R})\| \leq \|P_T(\hat{R})\|_F \leq \|U^T \hat{R}\|_F + \|\hat{R}V\|_F \lesssim c\sqrt{c_{\text{inj}}} \lambda < \lambda/4$$

(72)

as long as $c$ is small enough.
C.2.2 Proof of Claim 3

Let
\[ P_\Omega (XY^T - M) Y + \lambda X = B_1 \quad \text{and} \quad [P_\Omega (XY^T - M)]^T X + \lambda Y = B_2 \] (73)
for some \( B_1, B_2 \in \mathbb{R}^{n \times r} \). Clearly, it is seen from the assumption (46) that
\[ \max\{\|B_1\|_F, \|B_2\|_F\} \leq p\|\nabla f(X, Y)\|_F. \] (74)

In addition, the identities (73) allow us to obtain
\[ \|X^T X - Y^T Y\|_F = \frac{1}{\lambda} \|X^T (B_1 - P_\Omega (XY^T - M) Y) - (B_2 - [P_\Omega (XY^T - M)]^T X)^T Y\|_F \]
\[ = \frac{1}{\lambda} \|X^T B_1 - B_2^T Y\|_F \]
\[ \leq \frac{1}{\lambda} \|X\|_F \|B_1\|_F + \frac{1}{\lambda} \|B_2\|_F \|Y\| \]
\[ \leq 2\frac{p}{\lambda} \sqrt{2\sigma_{\max}} \|\nabla f(X, Y)\|_F. \] (75)

Here, the last line makes use of (74) and the assumption that \( \|X\|_F, \|Y\| \leq \sqrt{2\sigma_{\max}} \). In view of Lemma 20, one can find an invertible \( Q \) such that \( X = U \Sigma^{1/2} Q, Y = V \Sigma^{1/2} Q^{-T} \) and
\[ \|\Sigma Q - \Sigma Q^{-1}\|_F \leq \frac{1}{\sigma_{\min}(\Sigma)} \|X^T X - Y^T Y\|_F \]
\[ \leq \frac{2}{\sigma_{\min}} \cdot 2 \frac{p}{\lambda} \sqrt{2\sigma_{\max}} \|\nabla f(X, Y)\|_F \]
\[ \leq 8\sqrt{c} \frac{p}{\lambda \sigma_{\min}} \|\nabla f(X, Y)\|_F \]
\[ \leq 8\sqrt{c} \frac{c_{\min} p}{\lambda \sigma_{\min} n}, \] (75)

where \( \Sigma_Q \) is a diagonal matrix consisting of all singular values of \( Q \). Here, (i) follows from (47) as well as the bound (75), and the last inequality (ii) uses the assumption (21). 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 \((X, Y)\) is sufficiently close to \((X^*, Y^*)\), the operator \(P_\Omega(\cdot)\) restricted to the tangent space \(T\) of \(XY^T\) is injective. The proof is deferred to Appendix C.3.1.

**Lemma 7.** Suppose that the sample complexity obeys \(n^2p \geq C_{\mu r n} \log n\) for some sufficiently large constant \(C > 0\). Then with probability exceeding \(1 - O(n^{-10})\),
\[ \frac{1}{p} \|P_\Omega(H)\|_F^2 \geq \frac{1}{32\kappa} \|H\|_F^2, \quad \forall H \in T \]
holds simultaneously for all \((X, Y)\) obeying
\[ \max \left\{ \|X - X^*\|_{2, \infty}, \|Y - Y^*\|_{2, \infty} \right\} \leq \frac{c}{\kappa \sqrt{n}} \|X^*\|. \] (76)

Here, \(c > 0\) is some sufficiently small constant, and \(T\) denotes the tangent space of \(XY^T\).

**Remark** 5. In the prior literature, the injectivity of \(P_\Omega(\cdot)\) has been mostly studied when restricted to a fixed tangent space independent of \(\Omega\) (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 \(\Omega\).
First, we would like to show that here, (i) follows from the choice

\[ \sigma_{\text{true}} \leq \frac{\lambda}{p \sigma_{\text{min}}} \]

under the noise condition

\[ (\sigma_{\text{true}})^2 \lesssim \frac{\lambda^2}{\sigma_{\text{min}}^2} \]

Remark 6. Note that the condition (76) on \( (X, Y) \) is weaker than (24) under the assumptions of Lemma 4. To see this, if (24) holds, then one necessarily has

\[ \| X - X^\ast \|_{2,\infty} \leq C_n \kappa \left( \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\text{min}}} \right) \max \left\{ \| X^\ast \|_{2,\infty}, \| Y^\ast \|_{2,\infty} \right\} \]

\[ \lesssim C_n \kappa \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n \log n}{p}} \max \left\{ \| X^\ast \|_{2,\infty}, \| Y^\ast \|_{2,\infty} \right\} \]

\[ \leq C_n \kappa \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n \log n}{p}} \sqrt{\frac{\mu \rho}{n}} \| X^\ast \| \]

\[ \leq \frac{c}{\kappa \sqrt{n}} \| X^\ast \|. \]

Here, (i) follows from the choice \( \lambda \approx \sigma \sqrt{\frac{n \rho}{p}} \); (ii) relies on the incoherence assumption (32); and (iii) holds true under the noise condition \( \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{\mu \rho \log n}{p}} \lesssim \frac{1}{\sqrt{\kappa \sigma_{\text{min}} \log n}} \). A similar bound holds for \( \| Y - Y^\ast \|_{2,\infty} \).

The next lemma shows that for all \( (X, Y) \) close to \( (X^\ast, Y^\ast) \), \( \mathbb{P}_0 (XY^\top - M^\ast) \) is uniformly close to its expectation \( p(XY^\top - M^\ast) \). The proof can be found in Appendix C.3.2.

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

\[ \| p \Omega (XY^\top - M^\ast) - p (XY^\top - M^\ast) \| < \lambda / 8 \]

simultaneously for any \((X, Y)\) obeying (24), provided that \( \lambda = C_\lambda \sigma \sqrt{\rho p} \) for some constant \( C_\lambda > 0 \).

C.3.1 Proof of Lemma 7

By definition, any \( H \in \mathcal{T} \) can be expressed as

\[ H = XA^\top + BY^\top \]

for some \( A, B \in \mathbb{R}^{n \times r} \). Given that this is an underdetermined linear system of equations, there might be numerous \((A, B)\)'s compatible with (77). We take a specific choice as follows

\[ (A, B) := \arg \min_{(\tilde{A}, \tilde{B})} 0.5 \| \tilde{A} \|_F^2 + 0.5 \| \tilde{B} \|_F^2 \]

subject to \( H = X\tilde{A}^\top + \tilde{B}Y^\top \).

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

\[ X^\top B = A^\top Y. \]

To see this, consider the Lagrangian

\[ \mathcal{L}(A, \tilde{B}, \Lambda) := 0.5 \| \tilde{A} \|_F^2 + 0.5 \| \tilde{B} \|_F^2 + \langle \Lambda, X\tilde{A}^\top + \tilde{B}Y^\top - H \rangle. \]

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

\[ A = -\Lambda^\top X \quad \text{and} \quad B = -\Lambda Y \]

for some Lagrangian multiplier matrix \( \Lambda \in \mathbb{R}^{n \times n} \). The claim (79) then follows immediately.

The remaining proof consists of two steps.

- First, we would like to show that

\[ \| H \|_F^2 \leq 8 \sigma_{\text{max}} (\| A \|_F^2 + \| B \|_F^2). \]

(80)
• Second, we prove that
\[
\frac{1}{2p} \| \mathcal{P}_\Omega (H) \|_F^2 = \frac{1}{2p} \| \mathcal{P}_\Omega (X A^T + B Y^T) \|_F^2 \geq \frac{\sigma_{\min}}{8} (\| A \|_F^2 + \| B \|_F^2).
\]

(81)

Taking (80) and (81) together immediately yields the claimed bounds in the lemma. In what follows, we shall establish these two bounds separately.

1. Regarding the upper bound (80), it follows from elementary inequalities that
\[
\| H \|_F^2 = \| X A^T + B Y^T \|_F^2 \leq 2 (\| X A^T \|_F^2 + \| B Y^T \|_F^2 )
\]
\[
\leq 2 (\| X \|_F^2 \| A \|_F^2 + \| Y \|_F^2 \| B \|_F^2 )
\]
\[
\leq 2 \max \left \{ \| X \|_F^2, \| Y \|_F^2 \right \} (\| A \|_F^2 + \| B \|_F^2 ).
\]

(82)

It then suffices to control \( \max \{ \| X \|, \| Y \| \} \). In view of the assumption (76), one has
\[
\| X - X^* \| \leq \| X - X^* \|_F \leq \sqrt{n} \| X - X^* \|_{2,\infty} \leq \frac{c}{\kappa} \| X^* \| \leq \| X^* \|,
\]
\[
\text{as long as } c < 1. \text{ This together with the triangle inequality reveals that}
\]
\[
\| X \| \leq \| X^* \| + \| X - X^* \| \leq 2 \| X^* \| \leq 2 \sqrt{\sigma_{\text{max}}}. \]

Similarly, one has \( \| Y \| \leq 2 \sqrt{\sigma_{\text{max}}} \). Substitution into (82) yields the desired upper bound (80).

2. We now move on to the lower bound (81). To this end, one first decomposes
\[
\frac{1}{2p} \| \mathcal{P}_\Omega (X A^T + B Y^T) \|_F^2 = \frac{1}{2p} \| \mathcal{P}_\Omega (X A^T + B Y^T) \|_F^2 - \frac{1}{2} \| X A^T + B Y^T \|_F^2 + \frac{1}{2} \| X A^T + B Y^T \|_F^2.
\]
\[
= \alpha_1 - \alpha_2.
\]

The basic idea is to demonstrate that (1) \( \alpha_2 \) is bounded from below, and (2) \( \alpha_1 \) is sufficiently small compared to \( \alpha_2 \).

(a) We start by controlling \( \alpha_2 \), towards which we can expand
\[
\alpha_2 = \frac{1}{2} \left( \| X A^T \|_F^2 + \| B Y^T \|_F^2 \right) + \text{Tr} \left( X^T B Y^T A \right).
\]

The property \( X^T B = A^T Y \) (see (79)) implies that
\[
\text{Tr} \left( X^T B Y^T A \right) = \| X^T B \|_F^2 \geq 0 \implies \alpha_2 \geq \frac{1}{2} \left( \| X A^T \|_F^2 + \| B Y^T \|_F^2 \right).
\]

Write \( \Delta X = X - X^* \) and \( \Delta Y = Y - Y^* \). We have
\[
\| X A^T \|_F^2 = \| (X^* + \Delta X) A \|_F^2 \geq \| X^* A \|_F^2 + \| \Delta X A \|_F^2 + 2 \langle X^* A, \Delta X A \rangle
\]
\[
\geq \| X^* A \|_F^2 - 2 \| X^* A \|_F \| \Delta X A \|_F + 2 \langle X^* A, \Delta X A \rangle
\]
\[
\geq \| X^* A \|_F^2 - 2 \| X^* \| \| \Delta X \| \| A \|_F^2,
\]

where the second line arises from the Cauchy-Schwarz inequality. Recalling from (83) that \( \| \Delta X \| \leq c \| X^* \| / \kappa \), we arrive at
\[
\| X A^T \|_F^2 \geq \| X^* A \|_F^2 - 2c \| A \|_F^2 \geq \| X^* A \|_F^2 - \sigma_{\min} \| A \|_F^2 / 100,
\]

provided that \( c \leq 1/200. \) A similar bound holds for \( \| B Y^T \|_F^2 \), thus leading to
\[
\alpha_2 \geq \frac{1}{2} \left( \| X^* A \|_F^2 + \| B Y^* \|_F^2 \right) - \frac{1}{100} \sigma_{\min} \left( \| A \|_F^2 + \| B \|_F^2 \right).
\]

28
(b) Next, we control $\alpha$. First, it is seen that

\[
XA^T + BY^T = (X^* + \Delta X)A^T + B(Y^* + \Delta Y)^T
\]

\[
= X^*A^T + BY^* + \Delta XA^T + B\Delta Y^T.
\]

As a result, we can expand $\alpha$ as

\[
\alpha = \frac{1}{2p}\|P_\Omega (X^*A^T + BY^*)\|_F^2 - \frac{1}{2}\|X^*A^T + BY^* + \Delta XA^T + B\Delta Y^T\|_F^2
\]

\[
= \frac{1}{2p}\|P_\Omega (X^*A^T + BY^*)\|_F^2 - \frac{1}{2}\|\Delta XA^T\|_F^2 + \frac{1}{2p}\|P_\Omega (B\Delta Y^T)\|_F^2 - \frac{1}{2}\|B\Delta Y^T\|_F^2
\]

\[
+ \frac{1}{p}\langle P_\Omega (\Delta XA^T), P_\Omega (B\Delta Y^T) \rangle - \langle \Delta XA^T, B\Delta Y^T \rangle
\]

\[
+ \frac{1}{p}\langle P_\Omega (X^*A^T + BY^*), P_\Omega (\Delta XA^T + B\Delta Y^T) \rangle - \langle X^*A^T + BY^*, \Delta XA^T + B\Delta Y^T \rangle.
\]

i. Regarding $\gamma_1$, it follows from the bounds in [CR09, Section 4.2] that

\[
|\gamma_1| \leq \frac{1}{64} \|X^*A^T + BY^*\|_F^2 \leq \frac{1}{32} \left( \|X^*A^T\|_F^2 + \|BY^*\|_F^2 \right),
\]

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

ii. Invoke Lemma 19 to show that

\[
|\gamma_2| \leq \frac{3n}{2} \|\Delta X\|_{2,\infty}^2 \|A\|_F^2 \leq \frac{3c^2}{2k} \sigma_{\min} \|A\|_F^2 \leq \frac{1}{100} \sigma_{\min} \|A\|_F^2,
\]

\[
|\gamma_3| \leq \frac{3n}{2} \|\Delta Y\|_{2,\infty}^2 \|B\|_F^2 \leq \frac{3c^2}{2k} \sigma_{\min} \|B\|_F^2 \leq \frac{1}{100} \sigma_{\min} \|B\|_F^2,
\]

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

iii. The term $\gamma_4$ can be controlled via Lemma 21:

\[
|\gamma_4| \leq \frac{1}{p} \|p^{-1}P_\Omega (11^T) - 11^T\| \|\Delta X\|_{2,\infty} \|A\|_F \|\Delta Y\|_{2,\infty} \|B\|_F
\]

\[
\leq \sqrt{\frac{n}{p}} \|\Delta X\|_{2,\infty} \|A\|_F \|\Delta Y\|_{2,\infty} \|B\|_F,
\]

where the second line uses the bound $\|p^{-1}P_\Omega (11^T) - 11^T\| \lesssim \sqrt{n}/p$ guaranteed by [KMO10a, Lemma 3.2]. Continue the upper bound to get

\[
|\gamma_4| \overset{(i)}{\lesssim} \frac{c^2}{2k^2} \sigma_{\max} \|A\|_F \|B\|_F \leq \frac{c^2}{2k} \sigma_{\min} \left( \|A\|_F^2 + \|B\|_F^2 \right) \overset{(ii)}{\leq} \frac{1}{100} \sigma_{\min} \left( \|A\|_F^2 + \|B\|_F^2 \right).
\]

Here the first relation (i) arises from the assumption that $np \gg 1$. The second inequality (ii) applies the elementary inequality $ab \leq (a^2 + b^2)/2$ and the last one (iii) holds with the proviso that $c > 0$ is small enough.
iv. The last term \(\gamma_5\) can be further decomposed into the sum of four terms. For brevity, we take one out as an example, namely the term
\[
\frac{1}{p} \left\langle \mathcal{P}_\Omega \left( X^* A^T \right), \mathcal{P}_\Omega \left( \Delta X A^T \right) \right\rangle - \left\langle X^* A^T, \Delta X A^T \right\rangle.
\]
Apply the triangle inequality to obtain
\[
\left| \frac{1}{p} \left\langle \mathcal{P}_\Omega \left( X^* A^T \right), \mathcal{P}_\Omega \left( \Delta X A^T \right) \right\rangle - \left\langle X^* A^T, \Delta X A^T \right\rangle \right| \\
\leq \frac{1}{p} \left| \left\langle \mathcal{P}_\Omega \left( X^* A^T \right), \mathcal{P}_\Omega \left( \Delta X A^T \right) \right\rangle \right| + \left| \left\langle X^* A^T, \Delta X A^T \right\rangle \right| \\
\leq \frac{1}{\sqrt{p}} \|\mathcal{P}_\Omega \left( X^* A^T \right)\|_F \frac{1}{\sqrt{p}} \|\mathcal{P}_\Omega \left( \Delta X A^T \right)\|_F + \|X^* A^T\|_F \|\Delta X A^T\|_F.
\]
In light of [CR09, Section 4.2] and [ZL16, Lemma 9], we have
\[
\frac{1}{\sqrt{p}} \|\mathcal{P}_\Omega \left( X^* A^T \right)\|_F \leq 1.1 \|X^* A^T\|_F; \\
\frac{1}{\sqrt{p}} \|\mathcal{P}_\Omega \left( \Delta X A^T \right)\|_F \leq \sqrt{2\pi} \|\Delta X\|_{2,\infty} \|A\|_F.
\]
Taking the above three bounds collectively yields
\[
\left| \frac{1}{p} \left\langle \mathcal{P}_\Omega \left( X^* A^T \right), \mathcal{P}_\Omega \left( \Delta X A^T \right) \right\rangle - \left\langle X^* A^T, \Delta X A^T \right\rangle \right| \\
\leq 5 \|X^* A^T\|_F \sqrt{\pi} \|\Delta X\|_{2,\infty} \|A\|_F + \|X^* A^T\|_F \|\Delta X A^T\|_F \\
\leq 5\sqrt{\pi} \|\Delta X\|_{2,\infty} \|X^*\|_F \|A\|_F + \sqrt{\pi} \|\Delta X\|_{2,\infty} \|X^*\|_F \|A\|_F \\
= 6\sqrt{\pi} \|\Delta X\|_{2,\infty} \|X^*\|_F \|A\|_F^2.
\]
Using the assumption that \(\|\Delta X\|_{2,\infty} \leq c\|X^*\|/(\kappa \sqrt{\pi})\), one has
\[
\left| \frac{1}{p} \left\langle \mathcal{P}_\Omega \left( X^* A^T \right), \mathcal{P}_\Omega \left( \Delta X A^T \right) \right\rangle - \left\langle X^* A^T, \Delta X A^T \right\rangle \right| \leq \sqrt{n} \frac{c}{\kappa \sqrt{n}} \sigma_{\max} \|A\|_F^2 \leq \frac{1}{100} \sigma_{\min} \|A\|_F^2
\]
for \(c > 0\) small enough. The same argument applies to the remaining three terms, resulting in
\[
|\gamma_5| \leq \frac{1}{50} \sigma_{\min} \left( \|A\|_F^2 + \|B\|_F^2 \right).
\]
v. Combining the previous bounds on \(\gamma_1\) through \(\gamma_5\), we arrive at
\[
|\alpha_1| \leq |\gamma_1| + |\gamma_2| + |\gamma_3| + |\gamma_4| + |\gamma_5| \\
\leq \frac{1}{32} \left( \|X^* A^T\|_F^2 + \|BY^* T\|_F^2 \right) + \frac{1}{25} \sigma_{\min} \left( \|A\|_F^2 + \|B\|_F^2 \right).
\]
(c) Taking the preceding bounds on \(\alpha_1\) and \(\alpha_2\) collectively yields
\[
\frac{1}{2p} \left\|\mathcal{P}_\Omega \left( X A^T + B Y^T \right) \right\|_F^2 \geq \alpha_2 - |\alpha_1| \\
\geq \frac{15}{32} \left( \|X^* A^T\|_F^2 + \|BY^* T\|_F^2 \right) - \frac{1}{5} \sigma_{\min} \left( \|A\|_F^2 + \|B\|_F^2 \right) \\
\geq \frac{15}{32} \sigma_{\min} \left( \|A\|_F^2 + \|B\|_F^2 \right) - \frac{1}{5} \sigma_{\min} \left( \|A\|_F^2 + \|B\|_F^2 \right) \\
\geq \frac{1}{8} \sigma_{\min} \left( \|A\|_F^2 + \|B\|_F^2 \right).
\]
The proof is then complete.
C.3.2 Proof of Lemma 8

To start with, we have

\[ XY^\top - M^* = (X - X^*) Y^\top + X^* (Y - Y^*)^\top, \]

which together with the triangle inequality implies

\[ \| P^\text{debias}_\Omega (XY^\top - M^*) \| \leq \| P^\text{debias}_\Omega [(X - X^*) Y^\top] \| + \| P^\text{debias}_\Omega [X^* (Y - Y^*)^\top] \|. \]

Apply [CL17, Lemma 4.5] to obtain

\[ \| P^\text{debias}_\Omega [(X - X^*) Y^\top] \| \leq \| P^\text{debias}_\Omega (11^\top) \| \| X - X^* \|_{2,\infty} \| Y \|_{2,\infty} \]

\[ \lesssim \sqrt{np} \| X - X^* \|_{2,\infty} \| Y \|_{2,\infty}, \]

where the second line is due to \( \| P^\text{debias}_\Omega (11^\top) \| \lesssim \sqrt{np} \) (cf. [KMO10a, Lemma 3.2]). Similarly,

\[ \| P^\text{debias}_\Omega [X^* (Y - Y^*)^\top] \| \lesssim \sqrt{np} \| Y - Y^* \|_{2,\infty} \| X^* \|_{2,\infty}. \]

In addition, the assumption (24) yields

\[ \| Y \|_{2,\infty} \leq \| Y - Y^* \|_{2,\infty} + \| Y^* \|_{2,\infty} \]

\[ \leq C_{\infty} \kappa \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \| Y^* \|_{2,\infty} + \| Y^* \|_{2,\infty} \]

\[ \leq 2 \| Y^* \|_{2,\infty}, \]

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

\[ \| P^\text{debias}_\Omega (XY^\top - M^*) \| \lesssim \sqrt{np} \kappa \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \| X^* \|_{2,\infty} \| Y^* \|_{2,\infty} \]

\[ \lesssim \sqrt{np} \kappa \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \frac{\mu r \sigma_{\max}}{n}, \quad (84) \]

where the last inequality follows from the upper bound \( \max \{ \| X^* \|_{2,\infty}, \| Y^* \|_{2,\infty} \} \leq \sqrt{\mu r \sigma_{\max}/n} \) (cf. (32)). Rearrange the right-hand side of (84) to reach

\[ \| P^\text{debias}_\Omega (XY^\top - M^*) \| \lesssim \sigma \sqrt{np} \cdot \sqrt{\frac{\kappa^4 \mu^2 r^2 \log n}{np} + \lambda \sqrt{\frac{\kappa^4 \mu^2 r^2}{np}}} < \lambda/8, \]

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

D Analysis of the nonconvex gradient descent algorithm

Lemma 5 shares similar spirit as [MWCC17, CLL19] with one difference: the nonconvex loss function (15) has an additional term \( \| X \|_F^2 + \| Y \|_F^2 \) to balance the scale of \( X \) and \( Y \). To simplify the presentation, we find it convenient to introduce a few notations. Denote

\[ F^t \triangleq \begin{bmatrix} X^t \\ Y^t \end{bmatrix} \in \mathbb{R}^{2n \times r} \quad \text{and} \quad F^* \triangleq \begin{bmatrix} X^* \\ Y^* \end{bmatrix} \in \mathbb{R}^{2n \times r}. \]

It is easily seen from (26) that

\[ H^t = \arg \min_{R \in \mathbb{O}^{r \times r}} \| F^t R - F^* \|_F. \]

31
Concretely, we aim at proving that

\[ 1 \leq c \] for any matrix \( F \),

where \( c \) is the statistical independence to control the \( F \) for the following auxiliary loss function

\[ \min \{ \sum_{i,j} B_{ij} \} \] on the index set

\[ \Omega \].

Similarly, for each \( n \), we define \( \{ F_t \}_{t \geq 0} \) to be the GD iterates (see Algorithm 2) w.r.t.

the following auxiliary loss function

\[ f_t(X,Y) = \frac{1}{2p} \| P_{\Omega^{-t}} (XY^T - M) \|^2_F + \frac{1}{2} \| P_t (XY^T - M^*) \|^2_F + \lambda \| X \|^2_F + \frac{\lambda}{2p} \| Y \|^2_F. \] (88)

Here \( P_{\Omega^{-t}}(\cdot) \) denotes the orthogonal projection onto the space of matrices which are supported on the index set \( \Omega^{-t} = \{(i,j) \mid \Omega \cap t = \emptyset \} \) (resp. \( \{(i,j) | i = l \} \)). Mathematically, we have for any matrix \( B \in \mathbb{R}^{n \times n} \)

\[ [P_{\Omega^{-t}}(B)]_{ij} = \begin{cases} B_{ij}, & \text{if } (i,j) \in \Omega \text{ and } i \neq l, \\ 0, & \text{otherwise}, \end{cases} \quad \text{and} \quad [P_t(B)]_{ij} = \begin{cases} B_{ij}, & \text{if } i = l, \\ 0, & \text{otherwise}. \end{cases} \] (89)

Similarly, for each \( n + 1 \leq l \leq 2n \), we define \( \{ F_{t} \}_{t \geq 0} \) to be the GD iterates (see Algorithm 2) operating on

\[ f_t(X,Y) = \frac{1}{2p} \| P_{\Omega^{-l-n}} (XY^T - M) \|^2_F + \frac{1}{2} \| P_{l-(n-1)} (XY^T - M^*) \|^2_F + \lambda \| X \|^2_F + \frac{\lambda}{2p} \| Y \|^2_F, \]

where \( P_{\Omega^{-l-n}}(\cdot) \) and \( P_{l-(n-1)}(\cdot) \) are defined as

\[ [P_{\Omega^{-l-n}}(B)]_{ij} = \begin{cases} B_{ij}, & \text{if } (i,j) \in \Omega \text{ and } j \neq l-n, \\ 0, & \text{otherwise}, \end{cases} \quad \text{and} \quad [P_{l-(n-1)}(B)]_{ij} = \begin{cases} B_{ij}, & \text{if } j = l-n, \\ 0, & \text{otherwise}. \end{cases} \]

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

\[ \| F^t H^t - F^* \|_2 \leq C_F \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p\sigma_{\min}} \right) \| X^* \|_F, \] (90a)

\[ \| F^t H^t - F^* \|_2 \leq C_{op} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p\sigma_{\min}} \right) \| X^* \|_2 \] (90b)

\[ \max_{1 \leq l \leq 2n} \| F^t H^t - F_{l-}^t \|_2 \leq C_3 \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p\sigma_{\min}} \right) \| F^* \|_{2,\infty}, \] (90c)

\[ \max_{1 \leq l \leq 2n} \| (F^t H^t - F^*) \|_{l,2} \leq C_4 \kappa \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p\sigma_{\min}} \right) \| F^* \|_{2,\infty}, \] (90d)

\[ \| F^t H^t - F^* \|_{2,\infty} \leq C_\infty \kappa \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p\sigma_{\min}} \right) \| F^* \|_{2,\infty} \] (90e)
\begin{equation}
\|X^{tT}X^t - Y^{tT}Y^t\|_F \leq C_F n \eta \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \sqrt{r_0^2} \quad (90f)
\end{equation}

hold for all \(0 \leq t \leq t_0 = n^{18}\) and for some constants \(C_F, C_{op}, C_3, C_4, C_\infty, C_B > 0\), provided that \(\eta \ll 1/(n \kappa^3 \sigma_{\max})\). In addition, we also intend to establish that
\[
\frac{\eta}{2} \|\nabla f(X^{t-1}, Y^{t-1})\|_F^2
\]
holds for all \(1 \leq t \leq t_0 = n^{18}\). Here, \(H^{t,(l)}\) and \(R^{t,(l)}\) are rotation matrices defined as
\begin{align}
H^{t,(l)} &\triangleq \arg \min_{R \in O^{d \times r}} \|F^{t,(l)}R - F^*\|_F; \\
R^{t,(l)} &\triangleq \arg \min_{R \in O^{d \times r}} \|F^{t,(l)}R - F^tH^t\|_F. \quad (92a) \tag{92a}
\end{align}

Note that the induction hypotheses \((90a), (90b)\) and \((90c)\) readily imply the statements \((27a), (27b)\) and \((27c)\) in Lemma 5, respectively, whereas the last bound on the size of the gradient \((28)\) follows from \((91)\). We summarize the last connection in the following lemma, whose proof is in Appendix D.2.

**Lemma 9.** Set \(\lambda = C_\lambda \sigma \sqrt{np}\) for some large constant \(C_\lambda > 0\). Suppose that \(n \geq \kappa\) and that \((91)\) holds for all \(1 \leq t \leq t_0\), then
\[
\min_{0 \leq \ell < t_0} \|\nabla f(X^\ell, Y^\ell)\|_F \leq \frac{1}{n^3} \frac{\lambda}{p \sigma_{\min}},
\]
as long as \(\eta \ll 1/(n \kappa^3 \sigma_{\max})\).

The rest of this section is devoted to proving the hypotheses \((90)\) and \((91)\) via induction. We start with the base case, i.e., \(t = 0\). All the induction hypotheses \((90)\) are easily verified by noting that
\[
F^0 = F^{0,(l)} = F^*, \quad \text{for all } 1 \leq l \leq 2n.
\]

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 (90a)).** Set \(\lambda = C_\lambda \sigma \sqrt{np}\) for some large constant \(C_\lambda > 0\). Suppose that the sample size obeys \(n^2p \gg \kappa \mu \eta n \log n\) and the noise satisfies \(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^4 \mu \eta \log n}}\). If the iterates satisfy \((90)\) at the \(t\)th iteration, then with probability at least \(1 - O(n^{-100})\),
\[
\|F^{t+1}H^{t+1} - F^*\|_F \leq C_F \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \|X^*\|_F,
\]
holds as long as \(0 < \eta \ll 1/(\kappa^5 \sigma_{\max})\) and \(C_F > 0\) is large enough.

**Lemma 11 (Spectral norm error (90b)).** Set \(\lambda = C_\lambda \sigma \sqrt{np}\) for some large constant \(C_\lambda > 0\). Suppose that the sample size obeys \(n^2p \gg \kappa^4 \mu^2 \eta^2 n \log n\) and the noise satisfies \(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^4 \eta \log n}}\). If the iterates satisfy \((90)\) at the \(t\)th iteration, then with probability at least \(1 - O(n^{-100})\),
\[
\|F^{t+1}H^{t+1} - F^*\| \leq C_{op} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \|X^*\|
\]
holds with the proviso that \(0 < \eta \ll 1/(\kappa^3 \sigma_{\max} \sqrt{\kappa})\) and that \(C_{op} \gg 1\).

**Lemma 12 (Leave-one-out perturbation (90c)).** Set \(\lambda = C_\lambda \sigma \sqrt{np}\) for some large constant \(C_\lambda > 0\). Suppose that the sample size satisfies \(n^2p \gg \kappa^4 \mu^2 \eta^2 n \log^3 n\) and that the noise satisfies \(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^4 \mu \eta \log n}}\). If the iterates satisfy \((90)\) at the \(t\)th iteration, then with probability at least \(1 - O(n^{-99})\),
\[
\max_{1 \leq t < 2n} \|F^{t+1}H^{t+1} - F^{t+1,l}R^{t+1,l}\|_F \leq C_3 \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \|F^*\|_{2,\infty}
\]
holds, provided that \(0 < \eta \ll 1/(\kappa^2 \sigma_{\max} n)\) and that \(C_3 > 0\) is some sufficiently large constant.
Lemma 13 ($\ell_2/\ell_\infty$ norm error of leave-one-out sequences (90d)). Set $\lambda = C_\lambda \sigma \sqrt{np}$ for some large constant $C_\lambda > 0$. Suppose that the sample size obeys $n^2 p \gg \kappa^2 \mu^2 r^2 n \log n$ 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 (90) at the $t$th iteration, then with probability at least $1 - O(n^{-100})$,
\[ \max_{1 \leq l \leq 2n} \| (F^{t+1,l} H^{t+1,l} - F^*)_{1:1} \|_2 \leq C_4 \kappa \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \| F^* \|_{2,\infty}, \]
holds, provided that $0 < \eta \ll 1/(\kappa^2 \sqrt{\sigma_{\max}})$, $C_{\text{opt}} \gg 1$ and $C_4 \gg C_{\text{opt}}$.

Lemma 14 ($\ell_2/\ell_\infty$ norm error (90c)). Set $\lambda = C_\lambda \sigma \sqrt{np}$ for some large constant $C_\lambda > 0$. Suppose that $n \geq \mu r$ 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 (90) at the $t$th iteration, then with probability at least $1 - O(n^{-99})$,
\[ \| F^{t+1} H^{t+1} - F^* \|_{2,\infty} \leq C_\infty \kappa \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \| F^* \|_{2,\infty}, \]
holds provided that $C_\infty \geq 5C_3 + C_4$.

Lemma 15 (Approximate balancedness (90f)). 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 (90) at the $t$th iteration, then with probability at least $1 - O(n^{-100})$,
\[ \| X^{t+1} X^{t+1} - Y^{t+1} Y^{t+1} \|_F \leq C_{B} \kappa \eta \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \sqrt{\sigma_{\max}^2}, \]
\[ \max_{1 \leq l \leq 2n} \| X^{t+1,l} X^{t+1,l} - Y^{t+1,l} Y^{t+1,l} \|_F \leq C_{B} \kappa \eta \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \sqrt{\sigma_{\max}^2}, \]
holds for some sufficiently large constant $C_{B} \gg C_{\text{opt}}^2$, provided that $0 < \eta \ll 1/\sigma_{\min}$.

Lemma 16 (Decreasing of function values (91)). 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 (90) at the $t$th iteration, then with probability at least $1 - O(n^{-99})$,
\[ f(X^{t+1}, Y^{t+1}) \leq f(X^t, Y^t) - \eta \| \nabla f(X^t, Y^t) \|_F^2, \]
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 $A$, we denote by $A_{l,\cdot}$ (resp. $A_{\cdot,l}$) the $l$th row (resp. column) of $A$.

Define an augmented loss function $f_{\text{aug}}(X, Y)$ to be
\[ f_{\text{aug}}(X, Y) \triangleq \frac{1}{2p} \| \mathcal{P}_\Omega (XY^T - M) \|_F^2 + \frac{\lambda}{2p} \| X \|_F^2 + \frac{\lambda}{2p} \| Y \|_F^2 + \frac{1}{8} \| X^T X - Y^T Y \|_F^2. \] (93)
As the name suggests, this new function augments the original loss function (cf. (15)) with an additional term $\| X^T X - Y^T Y \|_F^2/8$, which is commonly used in the literature of asymmetric low-rank matrix factorization to balance the scale of $X$ and $Y$ [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 $f_{\text{aug}}(\cdot, \cdot)$. The introduction of $f_{\text{aug}}(\cdot, \cdot)$ is mainly to simplify the proof.

It is easily seen that the gradients of $f_{\text{aug}}(\cdot, \cdot)$ are given by
\[ \nabla X f_{\text{aug}}(X, Y) = \frac{1}{p} \mathcal{P}_\Omega (XY^T - M) Y + \frac{\lambda}{p} X + \frac{1}{2} X (X^T X - Y^T Y); \]
(94a)
\[ \nabla_Y f_{\text{aug}}(X, Y) = \frac{1}{p} \mathcal{P}_\mathcal{O} \left( XY^T - M \right)^T X + \frac{\lambda}{p} Y + \frac{1}{2} Y (Y^T Y - X^T X) . \] (94b)

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

\[ \nabla_X f_{\text{diff}}(X, Y) = -X (X^T X - Y^T Y) / 2; \] (95a)
\[ \nabla_Y f_{\text{diff}}(X, Y) = -Y (Y^T Y - X^T X) / 2, \] (95b)

such that

\[ \nabla f(X, Y) = \nabla f_{\text{aug}}(X, Y) + \nabla f_{\text{diff}}(X, Y), \] (96a)
\[ \nabla f(X, Y) = \nabla f_{\text{aug}}(X, Y) + \nabla f_{\text{diff}}(X, Y). \] (96b)

Regarding \( F^* \), simple algebra reveals that

\[ \sigma_1(F^*) = \| F^* \| = \sqrt{2\sigma_{\text{max}}}, \quad \sigma_\infty(F^*) = \sqrt{2\sigma_{\text{min}}}, \] (97a)
\[ \| F^* \|_{2, \infty} = \max \{ \| X^* \|_{2, \infty}, \| Y^* \|_{2, \infty} \} \leq \sqrt{\mu \sigma_{\text{max}} / n}, \] (97b)

where the last one follows from the incoherence assumption (32).

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 \( \lambda = C\sigma \sqrt{\mu} \) for some constant \( C > 0 \). Suppose that the sample size obeys \( n^2 \mu \geq C \kappa \mu \log n \) for some sufficiently large constant \( C > 0 \) and that the noise satisfies \( \sigma / \sigma_{\text{min}} \sqrt{\frac{n}{p}} \ll 1 \). Recall the function \( f_{\text{aug}} \) defined in (93). Then with probability at least \( 1 - O(n^{-10}) \),

\[ \text{vec} (\Delta)^T \nabla^2 f_{\text{aug}}(X, Y) \text{vec} (\Delta) \geq \frac{1}{10} \sigma_{\text{min}} \| \Delta \|_F^2 \]
\[ \max \{ \| \nabla^2 f_{\text{aug}}(X, Y) \|, \| \nabla^2 f(X, Y) \| \} \leq 10 \sigma_{\text{max}} \]

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

\[ \left\| \begin{bmatrix} X - X^* \\ Y - Y^* \end{bmatrix} \right\|_{2, \infty} \leq \frac{1}{1000 \kappa \sqrt{n}} \| X^* \| \]

and all \( \Delta = \begin{bmatrix} \Delta_X \\ \Delta_Y \end{bmatrix} \in \mathbb{R}^{2n \times r} \) lying in the set

\[ \left\{ \begin{bmatrix} X_1 \\ Y_1 \end{bmatrix} \right\} \left\| \begin{bmatrix} X_2 - X^* \\ Y_2 - Y^* \end{bmatrix} \right\| \leq \frac{1}{500 \kappa} \| X^* \| , \quad \hat{H} = \arg \min_{R \in \mathbb{O}^{r \times r}} \left\{ \left\| \begin{bmatrix} X_1 \\ Y_1 \end{bmatrix} - \left[ \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix} \right] \right\|_F \right\} . \]

Last but not least, a few immediate consequences of (90) 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 (27).

1. Suppose that the sample size obeys \( n \gg \mu \log n \). If the \( t \)-th iterates obey (90), then one has

\[ \left\| F_t^{(t)} R_t^{(t)} - F^* \right\|_{2, \infty} \leq \left( C_{\infty} \kappa + C_3 \right) \left( \sigma / \sigma_{\text{min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{\mu \sigma_{\text{min}}} \right) \| F^* \|_{2, \infty} , \] (98a)
\[ \left\| F_t^{(t)} R_t^{(t)} - F^* \right\| \leq 2C_{\text{op}} \left( \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n}{p}} + \frac{\lambda}{\mu \sigma_{\text{min}}} \right) \| X^* \|. \] (98b)

2. Suppose that the noise satisfies \( \sigma / \sigma_{\text{min}} \sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\mu} \log n} \). If the \( t \)-th iterates obey (90), then one has

\[ \| F^t H^t - F^* \| \leq \| X^* \| , \quad \| F^t H^t - F^* \|_F \leq \| X^* \|_F , \quad \| F^t H^t - F^* \|_{2, \infty} \leq \| F^* \|_{2, \infty} , \] (99a)
\[ \| F^t \| \leq 2 \| X^* \| , \quad \| F^t \|_F \leq 2 \| X^* \|_F , \quad \| F^t \|_{2, \infty} \leq 2 \| F^* \|_{2, \infty} . \] (99b)
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 $t$th iterates obey (90), then we have

$$\|F^t H^t - F^{t-1}(l) H^{t-1}(l)\|_F \leq 5\kappa \|F^t H^t - F^{t-1}(l) R^{t-1}(l)\|_F.$$ 

4. Suppose that $n \geq \kappa \mu$ and that $\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2 \log n}}$. If the $t$th iterates obey (90), then (99) also holds for $F^{t-1}(l) H^{t-1}(l)$. In addition, one has

$$\frac{\sigma_{\min}}{2} \leq \sigma_{\min} \left( (Y^{t-1}(l) H^{t-1}(l))^\top Y^{t-1}(l) H^{t-1}(l) \right) \leq \sigma_{\max} \left( (Y^{t-1}(l) H^{t-1}(l))^\top Y^{t-1}(l) H^{t-1}(l) \right) \leq 2\sigma_{\max}.$$

D.2 Proof of Lemma 9

Summing (91) over $t = 1$ to $t = t_0$ leads to a telescopic sum

$$f(X^{t_0}, Y^{t_0}) \leq f(X^0, Y^0) - \frac{\eta}{2} \sum_{t=0}^{t_0-1} \|\nabla f(X^t, Y^t)\|^2_F.$$

This further implies that

$$\min_{0 \leq t < t_0} \|\nabla f(X^t, Y^t)\|_F \leq \left\{ \frac{1}{t_0} \sum_{t=0}^{t_0-1} \|\nabla f(X^t, Y^t)\|^2_F \right\}^{1/2} \leq \left\{ \frac{2}{\eta t_0} \left[ f(X^*, Y^*) - f(X^{t_0}, Y^{t_0}) \right] \right\}^{1/2}, \quad (100)$$

where we have used the assumption that $(X^0, Y^0) = (X^*, Y^*)$.

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

$$f(F^{t_0}) = f(F^{t_0} H^{t_0}) = f(F^*) + \langle \nabla f(F^*), F^{t_0} H^{t_0} - F^* \rangle + \frac{1}{2} \text{vec} (F^{t_0} H^{t_0} - F^*)^\top \nabla^2 f(\tilde{F}) \text{vec} (F^{t_0} H^{t_0} - F^*),$$

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

$$f(F^*) - f(F^{t_0} H^{t_0}) \leq \|\nabla f(F^*)\|_F \|F^{t_0} H^{t_0} - F^*\|_F + \frac{1}{2} \text{vec} (F^{t_0} H^{t_0} - F^*)^\top \nabla^2 f(\tilde{F}) \text{vec} (F^{t_0} H^{t_0} - F^*) \leq \|\nabla f(F^*)\|_F \|F^{t_0} H^{t_0} - F^*\|_F + 5\sigma_{\max} \|F^{t_0} H^{t_0} - F^*\|_F^2.$$

Here the second line follows from the fact that $\|\nabla^2 f(\tilde{F})\| \leq 10\sigma_{\max}$; see Lemma 17. Recognize that

$$\|\nabla f(F^*)\|_F \leq \|\nabla_X f(F^*)\|_F + \|\nabla_Y f(F^*)\|_F \leq \frac{1}{p} \|P_{\Omega}(E) Y^*\|_F + \frac{\lambda}{p} \|X^*\|_F + \frac{1}{p} \|P_{\Omega}(E)^\top X^*\|_F + \frac{\lambda}{p} \|Y^*\|_F \leq \left( \frac{1}{p} \|P_{\Omega}(E)\| + \frac{\lambda}{p} \right) (\|X^*\|_F + \|Y^*\|_F), \quad (101)$$

where we have used the fact that $\nabla_X f(F^*) = \frac{1}{p} P_{\Omega}(X Y^*^\top - M^* - E) Y^* + \frac{1}{p} X^* = -\frac{1}{p} P_{\Omega}(E) Y^* + \frac{1}{p} X^*$. This together with Lemma 3 and the assumption that $\lambda \asymp \sigma / \sqrt{mp}$ yields

$$\|\nabla f(F^*)\|_F \lesssim \left( \sqrt{\frac{n}{p}} + \frac{\lambda}{p} \right) \sqrt{r \sigma_{\max}} \asymp \frac{\lambda}{p} \sqrt{r \sigma_{\max}}. \quad (102)$$

The above bounds together with the induction hypothesis (90a) for $t = t_0$ give

$$f(F^*) - f(F^{t_0} H^{t_0}) \lesssim \frac{\lambda}{p} \sqrt{r \sigma_{\max}} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p} \sigma_{\min} \right) \|X^*\|_F + \sigma_{\max} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p} \sigma_{\min} \right)^2 \|X^*\|_F^2.$$

36
\[
\leq r\kappa^2 \left( \frac{\lambda}{p} \right)^2 ,
\]
where the last relation arises from \( \sigma \sqrt{mp} \asymp \lambda \). Substitution into (100) results in
\[
\min_{0 \leq t < t_0} \| \nabla f (X^t, Y^t) \|_F \leq \sqrt{\frac{1}{\eta t_0} r \kappa^2 \left( \frac{\lambda}{p} \right)^2} \leq \frac{1}{n^3} \lambda \sqrt{\sigma_{\min}},
\]
provided that \( \eta \asymp 1/(\kappa^3 \sigma_{\max}) \), \( t_0 = n^{18} \) and that \( n \geq \kappa \).

D.3 Proof of Lemma 10

From the definitions of \( H^{t+1} \) (cf. (86)), \( \nabla f_{\text{aug}} \) (cf. (94)) and \( \nabla f_{\text{diff}} \) (cf. (95)), we have
\[
\begin{align*}
\| F^{t+1} H^{t+1} - F^* \|_F & \leq \| F^{t+1} H^t - F^* \|_F + \| \left[ F^t - \eta \nabla f \left( F^t \right) \right] H^t - F^* \|_F \\
& \leq \| F^{t+1} H^t - \eta \nabla f \left( F^t \right) H^t \|_F \\
& \leq \| F^{t+1} H^t - \eta \nabla f_{\text{aug}} \left( F^t \right) \|_F + \| \nabla f_{\text{diff}} \left( F^t \right) \|_F + \eta \| \nabla f_{\text{aug}} \left( F^* \right) \|_F .
\end{align*}
\]
Here (i) uses the fact that \( \nabla f (FR) = \nabla f (F) R \) for all \( R \in \mathcal{O}^{r \times r} \); the last relation (ii) uses the decomposition (96) and the triangle inequality. In the following, we bound \( \alpha_1, \alpha_2 \) and \( \alpha_3 \) in the reverse order.

1. First, regarding \( \alpha_3 \), since \( X^*^T X^* = Y^*^T Y^* \), one has \( \eta \| \nabla f (F^* \|_F = \eta \| \nabla f_{\text{aug}} (F^* \|_F \). Repeating our arguments for (101) and (102) gives
\[
\alpha_3 = \eta \| \nabla f (F^* \|_F \leq 4n^3 \frac{\lambda}{p} \| X^* \|_F
\]
as long as \( \lambda \asymp \sigma \sqrt{mp} \). Here the last inequality relies on the fact that \( \| X^* \|_F = \| Y^* \|_F \).

2. We now move on to \( \alpha_2 \), for which one has
\[
\begin{align*}
\alpha_2 & \leq \eta \left( \frac{1}{2} \left( \| X^t (X^T X^t T - Y^T Y^t \right) H^t \|_F + \| Y^t (Y^T Y^t - X^T X^t) H^t \|_F ) \\
& \leq \frac{\eta}{2} \left( \| X^t \| + \| Y^t \| \right) \| X^T X^t T - Y^T Y^t \|_F .
\end{align*}
\]
Utilize the fact that \( \max \{ \| X^t \|, \| Y^t \| \} \leq \| F^* \| \leq 2 \| X^* \| \) (see Lemma 18) and the induction hypothesis (90f) to obtain
\[
\begin{align*}
\alpha_2 & \leq 2n \sigma_{\max} \cdot C_B \kappa \eta \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\eta}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \sqrt{\sigma_{\max}^2} \\
& \leq (2C_B \kappa^{3/2} \eta \sigma_{\max}) \sigma_{\min} \eta \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\eta}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \| X^* \|_F \\
& \leq \sigma_{\min} \eta \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\eta}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \| X^* \|_F ,
\end{align*}
\]
where the second inequality uses \( \| X^* \|_F \geq \sqrt{\sigma_{\min}} \) and the last one holds as long as \( 2C_B \kappa^{3/2} \sigma_{\max} \eta \leq 1 \).

3. In the end, for \( \alpha_1 \), the fundamental theorem of calculus [Lan93, Chapter XIII, Theorem 4.2] reveals that
\[
\begin{align*}
\vec \left[ F^t H^t - \eta \nabla f_{\text{aug}} \left( F^t H^t \right) \right] - \left[ F^t - \eta \nabla f_{\text{aug}} \left( F^* \right) \right] \\
= \vec \left[ F^t H^t - F^* \right] - \eta \cdot \vec \left[ \nabla f_{\text{aug}} \left( F^t H^t \right) - \nabla f_{\text{aug}} \left( F^* \right) \right] \\
= \left( I_{2nr} - \eta \int_0^1 \nabla^2 f_{\text{aug}} \left( F(\tau) \right) d\tau \right) \vec \left( F^t H^t - F^* \right),
\end{align*}
\]
(103)
where we denote \( \mathbf{F}(\tau) \triangleq \mathbf{F}^* + \tau(\mathbf{F}^t \mathbf{H}^t - \mathbf{F}^*) \) for all \( 0 \leq \tau \leq 1 \). Taking the squared Euclidean norm of both sides of the equality (103) leads to
\[
\alpha_1^2 = \text{vec} \left( \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \right) \top (\mathbf{I}_{2n_r - \eta \mathbf{A}})^2 \text{vec} \left( \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \right)
= \text{vec} \left( \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \right) \top (\mathbf{I}_{2n_r - 2\eta \mathbf{A}}^2 + \eta^2 \mathbf{A}^2) \text{vec} \left( \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \right)
\leq \| \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \|^2_F + \eta^2 \mathbf{A}^2 \| \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \|^2_F - 2\eta \text{vec} \left( \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \right) \top \mathbf{A} \text{vec} \left( \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \right),
\]
where (104) results from the fact that
\[
\text{vec} \left( \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \right) \top \mathbf{A}^2 \text{vec} \left( \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \right) \leq \mathbf{A} \| \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \|^2_F.
\]
Based on the (90e), it is easily seen that \( \forall \tau \in [0, 1], \)
\[
\| \mathbf{F}(\tau) - \mathbf{F}^* \|_{2, \infty} \leq \| \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \|_{2, \infty} \leq C_{\infty} \kappa \left\| \sigma \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right\| \mathbf{F}^* \|_{2, \infty}
\leq \frac{1}{1000 \kappa \sqrt{n}} \sqrt{\sigma_{\max}},
\]
where the second line arises from the incoherence assumption (97b) and the last inequality holds as long as \( \frac{\sigma_{\min}}{\sigma_{\max}} \sqrt{\frac{n \log n}{p}} \ll \frac{1}{\kappa \mu \log n}. \) Invoke Lemma 17 with \( \mathbf{X} = \mathbf{X}^* + \tau(\mathbf{X}^t \mathbf{H}^t - \mathbf{X}^*), \mathbf{Y} = \mathbf{Y}^* + \tau(\mathbf{Y}^t \mathbf{H}^t - \mathbf{Y}^*), \)
\( (\mathbf{X}_1, \mathbf{Y}_1) = (\mathbf{X}^t, \mathbf{Y}^t) \) and \( (\mathbf{X}_2, \mathbf{Y}_2) = (\mathbf{X}^*, \mathbf{Y}^*) \) to obtain \( \| \mathbf{A} \| \leq 10 \sigma_{\max} \) and
\[
\text{vec} \left( \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \right) \top \mathbf{A} \text{vec} \left( \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \right) \geq \frac{1}{10} \sigma_{\min} \| \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \|^2_F,
\]
Putting these two bounds back to (104) yields
\[
\alpha_1^2 \leq \left( 1 + 100 \eta^2 \sigma_{\max}^2 - \frac{1}{5} \eta \sigma_{\min} \right) \| \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \|^2_F \leq \left( 1 - \frac{\sigma_{\min}}{10} \eta \right) \| \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \|^2_F.
\]
Here the last relation holds as long as \( 0 \leq \eta \leq 1/(1000 \kappa \sigma_{\max}). \) As a result, we have
\[
\alpha_1 \leq \left( 1 - \frac{\sigma_{\min}}{20} \eta \right) \| \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \|_F.
\]
Combine the above bounds on \( \alpha_1, \alpha_2 \) and \( \alpha_3 \) to conclude that
\[
\| \mathbf{F}^{t+1} \mathbf{H}^{t+1} - \mathbf{F}^* \|_F \leq \left( 1 - \frac{\sigma_{\min}}{20} \eta \right) \| \mathbf{F}^t \mathbf{H}^t - \mathbf{F}^* \|_F + 4 \eta^2 p \lambda \sigma_{\min} \left( \frac{\sigma_{\min}}{\sqrt{\frac{n \log n}{p}}} + \frac{\lambda}{p \sigma_{\min}} \right) \| \mathbf{X}^* \|_F
\leq \left( 1 - \frac{\sigma_{\min}}{20} \eta \right) C_F \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \| \mathbf{X}^* \|_F + 4 \eta \sigma_{\min} \lambda \sigma_{\min} \left( \frac{\sigma_{\min}}{\sqrt{\frac{n \log n}{p}}} + \frac{\lambda}{p \sigma_{\min}} \right) \| \mathbf{X}^* \|_F
\leq C_F \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \| \mathbf{X}^* \|_F,
\]
provided that \( C_F > 0 \) is large enough.

**D.4 Proof of Lemma 11**

To facilitate analysis, we define an auxiliary point \( \mathbf{\hat{F}}^{t+1} \triangleq \begin{bmatrix} \mathbf{\hat{X}}^{t+1} \\ \mathbf{\hat{Y}}^{t+1} \end{bmatrix} \) given by
\[
\mathbf{\hat{X}}^{t+1} = \mathbf{X}^t \mathbf{H}^t - \eta \left[ \frac{1}{p} \mathbf{P}_\Omega (\mathbf{X}^t \mathbf{Y}^t \mathbf{Y}^t - \mathbf{M}^* \mathbf{E}) \mathbf{Y}^* + \frac{\lambda}{p} \mathbf{X}^* + \frac{1}{2} \mathbf{X}^* \mathbf{H}^t \mathbf{H}^t (\mathbf{X}^t \mathbf{X}^t - \mathbf{Y}^t \mathbf{Y}^t) \mathbf{H}^t \right].
\]

(105a)
\[ Y^{t+1} = Y^t H^t - \eta \left[ \frac{1}{p} \mathcal{P}_\Omega (X^t Y^{t\top} - M^* - E)^\top X^* + \frac{\lambda}{p} Y^* + \frac{1}{2} Y^* H^{t\top} (Y^{t\top} Y^t - X^{t\top} X^t) H^t \right]. \] (105b)

Then the triangle inequality tells us that
\[ \| F^{t+1} H^{t+1} - F^* \| \leq \| F^{t+1} H^{t+1} - \tilde{F}^{t+1} \| + \| \tilde{F}^{t+1} - F^* \|. \] (106)

In what follows, we shall control \( \alpha_1 \) and \( \alpha_2 \) separately.

1. We start with \( \alpha_2 \). By the triangle inequality again we have
\[
\alpha_2 \leq \left\| \frac{1}{p} \mathcal{P}_\Omega (E) Y^* \right\| + \eta \lambda \left\| \frac{1}{p} \mathcal{P}_\Omega (E)^\top X^* \right\| + \eta \left\| \frac{1}{p} \mathcal{P}_\Omega (X^t Y^{t\top} - M^*) Y^* - (X^t Y^{t\top} - M^*) Y^* \right\|
\]

Denote \( \Delta^t \triangleq F^t H^t - F^* = \left[ \frac{\Delta^t_X}{\Delta^t_Y} \right] \). The term \( \beta_1 \) is the same as the term \( \alpha_2 \) in [CLL19, Section 4.2]. Therefore we can adopt the bound therein to obtain
\[
\beta_1 \leq (1 - \eta \sigma_{\min}) \| \Delta^t \| + 4 \eta \| \Delta^t \|^2 \| X^* \|.
\]

Moving to \( \beta_2 \), one has
\[
\beta_2 = \eta \left\| \frac{1}{p} \mathcal{P}_\Omega (E) \right\| + \eta \lambda \left\| \frac{1}{p} \mathcal{P}_\Omega (E)^\top \right\| \leq \eta \| \mathcal{P}_\Omega (E) \| + \eta \lambda \leq C \eta \left( \sigma \sqrt{\frac{n}{p}} + \frac{\lambda}{p} \right) \| X^* \|
\]

for some constant \( C > 0 \). Here the last inequality arises from Lemma 3, and the fact that \( \| F^* \| = \sqrt{2} \| X^* \| \) (cf. (97a)). We are now left with the term \( \beta_3 \), which is exactly the same as the \( \alpha_1 \) in [CLL19, Section 4.2]. Reusing their results we have
\[
\beta_3 \leq \eta \| \Delta^t \| \| F^* \| \| X^* \|
\]

The last line follows from the facts that \( \| \mathcal{P}_\Omega (11^\top) - p11^\top \| \leq \sqrt{n p} \) (see [KMO10a, Lemma 3.2]) and that \( \max \{ \| \Delta^t_X \|_{2,\infty}, \| \Delta^t_Y \|_{2,\infty} \} \leq \| F^* \|_{2,\infty} \), 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
\[
\alpha_2 \leq \left( 1 - \eta \sigma_{\min} \right) \| \Delta^t \| + 4 \eta \| \Delta^t \|^2 \| X^* \| + \tilde{C} \eta \left( \frac{\sqrt{n}}{p} + \frac{\lambda}{p} \right) \| X^* \| + \tilde{C} \eta \left( \frac{\sqrt{n}}{p} \right) \| \Delta^t \|_{2,\infty} \| F^* \|_{2,\infty} \| X^* \|
\]

for some sufficiently large constant \( \tilde{C} > 0 \). Here the second inequality arises from the condition
\[ 4 \| \Delta^t \| \| X^* \| \leq \sigma_{\min}/2, \]

which would hold if \( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll \frac{1}{\kappa} \). An immediate consequence of (107) is that
\[ \alpha_2 = \| \tilde{F}^{t+1} - F^* \| \leq (\sqrt{2\kappa})^{-1} \| X^* \|. \] (108)
To see this, apply the induction hypotheses (90b) and (90c) to get

\[
\alpha_2 \leq \left(1 - \frac{\eta \sigma_{\min}}{2}\right) C_{\text{op}} \left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p} + \frac{\lambda}{p \sigma_{\min}}} \|X^*\| + \tilde{C} \eta \left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p} + \frac{\lambda}{p \sigma_{\min}}} \|X^*\|\right) + \tilde{C} \eta \sqrt{\frac{n}{p} C_{\text{op}} K} \left(\frac{\sigma}{\sigma_{\min}} \frac{n \log n}{p} \|F^*\|_{2, \infty} \|X^*\|\right)\right)
\]

(109)

Here (i) holds under the assumptions that \(C_{\text{op}} \gg \tilde{C} + \tilde{C} C_{\text{op}} K^2 \sqrt{\frac{\mu r^2 \log n}{np}}\) and that \(\|F^*\|_{2, \infty} \leq \sqrt{\frac{m}{n}} \|X^*\|\) (cf. 97b); (ii) arises since \(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p} \ll 1/\kappa \text{ and } \lambda \gg \sigma \sqrt{mp}}\). Under the sample complexity \(n^2 p \gg \kappa^4 \mu^2 r^2 n \log n\), the first condition can be simplified to \(C_{\text{op}} \gg 2 \tilde{C} \gg 1\).

2. Next we bound \(\alpha_1\), towards which we first observe that

\[
\alpha_1 = \|F^{t+1} H^{t+1} - \tilde{F}^{t+1}\| = \|F^{t+1} H^t H^t H^{t+1} - \tilde{F}^{t+1}\|.
\]

It is straightforward to verify that \(H^{t+1} H^t H^{t+1}\) is the best rotation matrix to align \(F^{t+1} H^t\) and \(F^*\) (in the sense of (86)). Regarding \(\tilde{F}^{t+1}\), we obtain the following claim, which demonstrates that it is already aligned with \(F^*\), i.e. \(I_r\) is the best rotation matrix to align \(\tilde{F}^{t+1}\) and \(F^*\).

**Claim 4.** Suppose (109) holds true, one has

\[
I_r = \arg \min_{R \in O_{r \times r}} \|\tilde{F}^{t+1} R - F^*\|_F.
\]

Now we intend to apply Lemma 22 with \(F_0 = F^*\), \(F_1 = \tilde{F}^{t+1}\), \(F_2 = F^{t+1} H^t\), for which we need to check the two conditions therein. First, in view of (108), one has

\[
\|F_1 - F_0\| \|F_0\| = \|\tilde{F}^{t+1} - F^*\| \|F^*\| \leq \frac{1}{\sqrt{2 \kappa}} \|X^*\| \|F^*\| = \sigma_{\min} = \frac{1}{2} \sigma_r^2(F_0).
\]

Second, making use of the gradient update rules (25) and the decomposition (96), we obtain

\[
\|F^{t+1} H^t - \tilde{F}^{t+1}\| = \left\| \left(F^t - \eta \nabla f_{\text{aug}}(F^t) - \eta \nabla f_{\text{diff}}(F^t) \right) H^t - \tilde{F}^{t+1} \right\|
\]

\[
\leq \left\| \left(F^t - \eta \nabla f_{\text{aug}}(F^t) \right) H^t - \tilde{F}^{t+1} \right\| + \eta \|\nabla f_{\text{diff}}(F^t)\|.
\]

The term \(\theta_2\) has been controlled as \(\alpha_2\) in the proof of Lemma 10, where we obtained

\[
\theta_2 \leq 2 C_{\text{B}} \eta^2 \left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p} + \frac{\lambda}{p \sigma_{\min}}} \right) \sqrt{r} \sigma_{\max}^2 \|X^*\|.
\]

We now move on to \(\theta_1\), for which we have

\[
\theta_1 \leq \left\| F^t - \eta \left\{ \nabla f_{\text{aug}}(F^t) - \eta \left[ \frac{1}{p} \mathcal{P}_\Omega(E) Y^t \right] - \eta \left[ \frac{1}{p} \mathcal{P}_\Omega(E) X^t \right] \right\} H^t \right\|
\]

\[
- \tilde{F}^{t+1} - \eta \left[ \frac{1}{p} \mathcal{P}_\Omega(E) Y^* \right] - \eta \left[ \frac{1}{p} \mathcal{P}_\Omega(E) X^* \right].
\]

\[
:= \xi_1
\]

40
\[
\begin{align*}
+ \eta \left[ \| \frac{1}{p} \mathcal{P}_\Omega(E) Y^t \| H^t \right] + \frac{\lambda}{p} \left[ X^t - \frac{1}{p} \mathcal{P}_\Omega(E) Y^* \right] \right] H^t - \left[ \frac{1}{p} \mathcal{P}_\Omega(E) X^* \right] \right].
\end{align*}
\]

\[\xi_2 := \xi_2\]

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

\[
\xi_1 \lesssim \eta \sqrt{n} \left( \| \Delta_{\mathbf{x}} \|_{2, \infty} \| Y^* \|_{2, \infty} + \| \Delta_{\mathbf{y}} \|_{2, \infty} \| X^* \|_{2, \infty} + \| \Delta_{\mathbf{x}} \|_{2, \infty} \| \Delta_{\mathbf{y}} \|_{2, \infty} \right) \| \Delta^t \|
\]

\[
+ \eta \left( \| \Delta_{\mathbf{x}} \| \| Y^* \| + \| \Delta_{\mathbf{y}} \| \| X^* \| + \| \Delta_{\mathbf{x}} \| \| \Delta_{\mathbf{y}} \| + 2 \| X^* \| \| \Delta_{\mathbf{x}} \| + 2 \| Y^* \| \| \Delta_{\mathbf{y}} \| + \| \Delta_{\mathbf{x}} \|^2 + \| \Delta_{\mathbf{y}} \|^2 \right) \| \Delta^t \|
\]

\[
\lesssim \eta \sqrt{n} \| \Delta^t \|_{2, \infty} \left( \| F^* \|_{2, \infty} \| \Delta^t \| + \eta \| \Delta^t \|^2 \right)
\]

\[
\leq \frac{1}{15 \kappa} \frac{\sigma_{\min}}{4} \eta \| \Delta^t \|.
\]

Here the penultimate inequality arises from the facts that max{\| \Delta_{\mathbf{x}} \|_{2, \infty}, \| \Delta_{\mathbf{y}} \|_{2, \infty}} \leq \| \Delta^t \|_{2, \infty} \leq \| F^* \|_{2, \infty} and similarly max{\| \Delta_{\mathbf{x}} \|, \| \Delta_{\mathbf{y}} \|} \leq \| \Delta^t \| \leq \| X^* \|; see Lemma 18. In addition, the last line holds because of the induction hypotheses (90b) and (90e), provided that

\[
C_{\infty} \kappa \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \sqrt{rac{\mu^2 r^2 \log n}{np}} \ll \frac{1}{\kappa^2} \quad \text{and} \quad C_{\text{op}} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll \frac{1}{\kappa^2}.
\]

Again, the first condition would be guaranteed by the sample size condition \(n^2 p \gg \kappa^4 \mu^2 r^2 n \log n\) and the noise condition \(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll 1/\kappa\). Next, the term \(\xi_2\) can be easily controlled as follows

\[
\xi_2 \leq \eta \left( \| \frac{1}{p} \mathcal{P}_\Omega(E) (Y^t H^t - Y^*) \| + \frac{\lambda}{p} \| F^t H^t - F^* \| \right)
\]

\[
\leq \tilde{C} \eta \left( \sigma \sqrt{\frac{n}{p}} + \frac{\lambda}{p} \right) \| \Delta^t \|,
\]

where the last line follows from the same argument for bounding \(\beta_2\) above. Taking the bounds on \(\theta_1\) and \(\theta_2\) collectively yields

\[
\| F^{t+1} H^t - \hat{F}^{t+1} \| \leq \frac{1}{15 \kappa} \frac{\sigma_{\min}}{4} \eta \| \Delta^t \| + \tilde{C} \eta \left( \sigma \sqrt{\frac{n}{p}} + \frac{\lambda}{p} \right) \| \Delta^t \| + 2 C_B \kappa \eta^2 \sigma_{\min} \sqrt{\frac{n}{p}} + \frac{\lambda}{p} \frac{\lambda}{p} \sigma_{\min} \sqrt{\frac{n}{p}} \| X^* \|
\]

\[
\leq \frac{1}{5 \kappa} \frac{\sigma_{\min}}{4} \eta \| \Delta^t \| + 2 C_B \kappa \eta^2 \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p} \sigma_{\min} \sqrt{\frac{n}{p}} \right) \sqrt{\sigma_{\max}^2} \| X^* \|.
\]

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 (110) is that

\[
\| \hat{F}^{t+1} - F^{t+1} \| \leq (2 \sqrt{2} \kappa)^{-1} \| X^* \|,
\]

as long as \(\eta \ll 1/(C_B \kappa^2 \sigma_{\max} \sqrt{\kappa})\) and \(\lambda \asymp \sqrt{np}\). As a result, one obtains

\[
\| F_1 - F_2 \| \| F_0 \| = \| \hat{F}^{t+1} - F^{t+1} H^t \| \| F^* \| \leq (2 \sqrt{2} \kappa)^{-1} \| X^* \| \| F^* \| = \sigma_{\min}/2 = \sigma_{\min}(F^0)/4.
\]

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

\[
\alpha_1 = \| \hat{F}^{t+1} - F^{t+1} H^t+1 \| \leq 5 \kappa \| \hat{F}^{t+1} - F^{t+1} H^t \|
\]

\[
\leq \frac{1}{4} \sigma_{\min} \| \Delta^t \| + 10 C_B \kappa^2 \eta^2 \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p} \sigma_{\min} \sqrt{\frac{n}{p}} \right) \sqrt{\sigma_{\max}^2} \| X^* \|
\]

\[
\leq \frac{1}{4} \sigma_{\min} \| \Delta^t \| + \eta \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p} \right) \| X^* \|,
\]

provided that \(\eta \ll 1/(C_B \kappa^3 \sigma_{\max} \sqrt{\kappa})\).
Combine the bounds on $\alpha_1$ and $\alpha_2$ to reach

$$
\|F^{t+1}H^{t+1} - F^*\| \\
\leq \left(1 - \frac{\eta}{2}\sigma_{\min}\right)\|\Delta^t\| + \left(\tilde{C} + 1\right)\eta\left(\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\|X^*\| + \tilde{C}\eta\sqrt{\frac{n}{p}}\|\Delta^t\|_2,\infty\|F^*\|_2,\infty\|X^*\| + \frac{\sigma_{\min}}{4}\eta\|\Delta^t\|_2,\infty\|X^*\|,
$$

where $\lambda = \max\left\{\frac{\|\nabla f\|_2}{\delta}\right\}$.

With the proviso that $C_{op} \gg 1$ and $n^2p \gg \kappa^4\mu^2r^2n\log n$. Here the last line follows from the same argument as in bounding (109). This completes the proof.

**Proof of Claim 4.** In view of [MWCC17, Lemma 35], it suffices to show that $F^*\tilde{F}^{t+1}$ is symmetric and positive semidefinite. Recognizing that $F^*\tilde{F}^{t+1} = (X^*X^t + Y^*Y^t)H^t$ is symmetric (see [MWCC17, Lemma 35]), it is straightforward to verify that $F^*\tilde{F}^{t+1}$ is also symmetric (which we omit here for brevity). In addition, by (108) we have

$$
\|F^*\tilde{F}^{t+1} - F^*F^*\| \leq \|F^*\| \|\tilde{F}^{t+1} - F^*\| = \alpha_2\|F^*\| \leq \frac{1}{\sqrt{2\kappa}}\|X^*\|\|F^*\| = \sigma_{\min}.
$$

Since $F^*F^* = X^*X^t + Y^*Y^t = 2\Sigma^*$, Weyl’s inequality gives

$$
\lambda_{\min}(F^*\tilde{F}^{t+1}) \geq 2\sigma_{\min} - \|F^*F^*\tilde{F}^{t+1} - F^*F^*\| \geq \sigma_{\min} > 0,
$$

where $\lambda_{\min}(A)$ stands for the minimum eigenvalue of a matrix $A$. To conclude, $F^*\tilde{F}^{t+1}$ 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 $1 \leq l \leq n$; the case with $n + 1 \leq l \leq 2n$ can be derived in a similar way. From the definition of $R^{t+1,l}$ (cf. (92b)), we have

$$
\|F^{t+1}H^{t+1} - F^{t+1,l}(R^{t+1,l})\|_F \leq \|F^{t+1}H^t - F^{t+1,l}(R^{t,l})\|_F.
$$

The gradient update rules (25) and (87) give

$$
F^{t+1}H^t - F^{t+1,l}(R^{t,l}) \\
= \left[F^t - \eta\nabla f(F^t)\right]H^t - \left[F^{t,l} - \eta\nabla f^{(t,l)}(F^{t,l})\right]R^{t,l} \\
= \left[F^tH^t - \eta\nabla f(F^tH^t) - \left[F^{t,l}R^{t,l} - \eta\nabla f^{(t,l)}(F^{t,l}R^{t,l})\right]\right] \\
= \left[F^tH^t - F^{t,l}(R^{t,l})\right] - \eta\left[\nabla f_{\text{aug}}(F^tH^t) - \nabla f_{\text{aug}}(F^{t,l}R^{t,l})\right] - \eta\left[\nabla f_{\text{diff}}(F^tH^t) - \nabla f_{\text{diff}}(F^{t,l}R^{t,l})\right] \\
+ \eta\left[\nabla f^{(t,l)}(F^{t,l}R^{t,l})\right]
$$

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

In what follows, we shall bound $A_1, A_2$ and $A_3$ sequentially.
1. The first term $A_1$ is similar to $\alpha_1$ in the proof of Lemma \ref{lem:alpha1}. Going through the same derivations therein, we obtain

$$\|A_1\|_F \leq \left(1 - \frac{\sigma_{\min}}{20}\right) \|F^t H^t - F^{t,(l)} R^{t,(l)}\|_F,$$

provided that $\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2 \mu_{\text{r}} \log n}}$ and that $0 \leq \eta \leq 1/(1000\kappa \sigma_{\text{max}})$.

2. Next, we turn attention to $A_2$, which clearly obeys

$$\|A_2\|_F \leq \eta \|\nabla f_{\text{diff}}(F^t H^t)\|_F + \eta \|\nabla f_{\text{diff}}(F^{t,(l)} R^{t,(l)})\|_F.$$

Recall from the term $\theta_2$ in the proof of Lemma \ref{lem:theta1} that

$$\eta \|\nabla f_{\text{diff}}(F^t H^t)\|_F = \eta \|\nabla f_{\text{diff}}(F^t)\|_F \leq 2C_B \kappa^2 \left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{\sigma_{\min}}\right) \sqrt{n} \sigma_{\text{max}}^2 \|X^*\|.$$

Applying Lemma \ref{lem:Jensen} and going through the same derivation as in bounding $\theta_2$ in the proof of Lemma \ref{lem:theta1}, one gets

$$\eta \|\nabla f_{\text{diff}}(F^{t,(l)} R^{t,(l)})\|_F \leq 2C_B \kappa^2 \left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{\sigma_{\min}}\right) \sqrt{n} \sigma_{\text{max}}^2 \|X^*\|_2,\infty.$$

Combine the above three inequalities to obtain

$$\|A_2\|_F \leq 4C_B \kappa^2 \left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{\sigma_{\min}}\right) \sqrt{n} \sigma_{\text{max}}^2 \|X^*\|$$

$$\leq 4\sqrt{n}C_B \kappa^2 \left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{\sigma_{\min}}\right) \sqrt{n} \sigma_{\text{max}}^2 \|X^*\|_2,\infty$$

$$\leq \eta \left(\frac{n \sigma}{\sqrt{p}} + \frac{\lambda}{p}\right) \|F^*\|_2,\infty.$$

Here the second inequality arises from the elementary inequality $\|X^*\| \leq \sqrt{n}\|X^*\|_2,\infty$, whereas the last one holds true because of $\|X^*\|_2,\infty \leq \|F^*\|_2,\infty$ and the condition that $\eta \ll \frac{1}{\kappa^2 \sigma_{\text{max}}}$.

3. We are now left with $A_3$. To this end, we first observe that

$$A_3 = \eta \left[ \begin{array}{cc} \mathcal{P}_{\mathcal{B}_1} \left( X^{t,(l)} Y^{t,(l)\top} - M^* \right) - p^{-1} \mathcal{P}_{\mathcal{O}_{\mathcal{B}_1}} \left( X^{t,(l)} Y^{t,(l)\top} - M^* \right) & \left( Y^{t,(l)} R^{t,(l)} \right)^\top \right] \right. \left[ \begin{array}{cc} \mathcal{P}_{\mathcal{B}_2} \left( X^{t,(l)} Y^{t,(l)\top} - M^* \right) - p^{-1} \mathcal{P}_{\mathcal{O}_{\mathcal{B}_2}} \left( X^{t,(l)} Y^{t,(l)\top} - M^* \right) & \left( Y^{t,(l)} R^{t,(l)} \right)^\top \right]$$

The following claims allow one to bound $C_1, C_2$ and $B_1, B_2$; 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 $np \gg \log^3 n$. With probability at least $1 - O(n^{-100})$,

$$\max \left\{ \|C_1\|_F, \|C_2\|_F \right\} \leq \sigma \sqrt{\frac{n \log n}{p}} \|F^*\|_2,\infty.$$  \hfill (113)

**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 $1 - O(n^{-100})$,

$$\|B_1\|_F \leq \sqrt{\frac{\mu \eta^2 \log n}{np}} \|F^{t,(l)} R^{t,(l)} - F^*\|_2,\infty \sigma_{\text{max}}.$$  \hfill (114)

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

$$\|B_2\|_F \leq \sqrt{\frac{\mu \eta^2 \log n}{np}} \|F^{t,(l)} R^{t,(l)} - F^*\|_2,\infty \sigma_{\text{max}}.$$  \hfill (115)
With these claims in place, one can readily obtain that
\[
\| A_3 \|_F \leq \eta \left( \| B_1 \|_F + \| B_2 \|_F + \| C_1 \|_F + \| C_2 \|_F \right) \\
\leq \eta \sqrt{\frac{n \log n}{p}} \left( \| F^* \|_{2, \infty} + \eta \sqrt{\frac{\mu^2 r^2 \log n}{np}} \right) \| F^{t, (i)} R^{t, (i)} - F^* \|_{2, \infty} \sigma_{\text{max}} \\
\leq \eta \sqrt{\frac{n \log n}{p}} \| F^* \|_{2, \infty} + \eta \sqrt{\frac{\mu^2 r^2 \log n}{np}} \left( C_{\infty} \kappa + C_3 \right) \left( \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\text{min}}} \right) \| F^* \|_{2, \infty} \sigma_{\text{max}},
\]
where the last line follows from the induction hypotheses (90c) and (90e).

This together with the bounds on \( A_1 \) and \( A_2 \) gives: for some constant \( \tilde{C} > 0 \),
\[
\| F^{t+1} H^{t+1} - F^{t+1, (i)} R^{t+1, (i)} \|_F \leq \| A_1 \|_F + \| A_2 \|_F + \| A_3 \|_F \\
\leq \left( 1 - \frac{\sigma_{\text{min}}}{20} \eta \right) \| F^t H^t - F^{t, (i)} R^{t, (i)} \|_F + \eta \left( \sigma \sqrt{\frac{n}{p}} + \frac{\lambda}{p} \right) \| F^* \|_{2, \infty} \\
+ \tilde{C} \eta \sqrt{\frac{n \log n}{p}} \| F^* \|_{2, \infty} + \tilde{C} \eta \sqrt{\frac{n \log n}{p}} \left( C_{\infty} \kappa + C_3 \right) \left( \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\text{min}}} \right) \| F^* \|_{2, \infty} \sigma_{\text{max}} \\
\leq \left( 1 - \frac{\sigma_{\text{min}}}{20} \eta \right) C_3 \left( \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\text{min}}} \right) \| F^* \|_{2, \infty} + \eta \left( \sigma \sqrt{\frac{n}{p}} + \frac{\lambda}{p} \right) \| F^* \|_{2, \infty} \\
+ \tilde{C} \eta \sqrt{\frac{n \log n}{p}} \| F^* \|_{2, \infty} + \tilde{C} \eta \sqrt{\frac{n \log n}{p}} \left( C_{\infty} \kappa + C_3 \right) \left( \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\text{min}}} \right) \| F^* \|_{2, \infty} \sigma_{\text{max}} \\
\leq C_3 \left( \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\text{min}}} \right) \| F^* \|_{2, \infty}
\]
as claimed. Here, (i) invokes the induction hypothesis (90c), whereas (ii) holds as long as \( C_3 \) is large enough and the sample size satisfies \( n^2 p \gg \kappa^4 \mu^2 r^2 n \log n \).

**Proof of Claim 5.** By the unitary invariance of the Frobenius norm, one has
\[
\| C_1 \|_F = p^{-1} \| \mathcal{P}_{\Omega_t} (E) Y^{t, (i)} \|_F.
\]
Since the entries of \( \mathcal{P}_{\Omega_t} (E) \) are all zero except those on the \( l \)th row, we have
\[
p \| C_1 \|_F = \left\| \sum_{j=1}^n \delta_{l j} E_{l j} Y_{j, l}^{t, (i)} \right\|_2,
\]
where we denote \( \delta_{l j} \triangleq I_{(l,j) \in \Omega_t} \). Since \( Y^{t, (i)} \) is independent of \( \{ \delta_{l j} \}_{1 \leq j \leq n} \) and \( \{ E_{l j} \}_{1 \leq j \leq n} \), the vectors \( \{ u_j \}_{1 \leq j \leq n} \) are independent conditioning on \( Y^{t, (i)} \). Therefore, from now on we shall condition on a fixed \( Y^{t, (i)} \). It is easy to verify that
\[
\left\| u_j \right\|_2 \leq \left\| Y^{t, (i)} \right\|_{2, \infty} \left\| \delta_{l j} E_{l j} \right\|_{\psi_1} \leq \sigma \left\| Y^{t, (i)} \right\|_{2, \infty},
\]
where \( \| \cdot \|_{\psi_1} \) denotes the sub-exponential norm [KLT11, Section 6]. Further, one can calculate
\[
V := \left\| \mathbb{E} \left[ \sum_{j=1}^n (\delta_{l j} E_{l j})^2 Y_{j, l}^{t, (i)} Y_{j, l}^{t, (i) \top} \right] \right\| \leq p \sigma^2 \left\| \mathbb{E} \left[ \sum_{j=1}^n Y_{j, l}^{t, (i)} Y_{j, l}^{t, (i) \top} \right] \right\| = p \sigma^2 \left\| Y^{t, (i)} \right\|_F^2.
\]
Invoke the matrix Bernstein inequality [KLT11, Proposition 2] to discover that with probability at least \( 1 - O \left( n^{-100} \right) \),
\[
\left\| \sum_{j=1}^n u_j \right\|_2 \leq \sqrt{V \log n} + \left\| u_j \right\|_{\psi_1} \log^2 n.
\]
Here, both (i) and (ii) arise from Lemma 18, as long as $\sigma = \sqrt{\frac{\log n}{np}}$, and the last inequality holds if $np \gg \log^3 n$.

We then complete the proof by observing that $\|Y^{t, l}\|_{2, \infty} \leq 2\|F^*\|_{2, \infty}$ as long as $\frac{\sigma}{\min \sqrt{\frac{\log n}{np}} \ll \frac{1}{\sqrt{n^2 \log n}}}$; see Lemma 18. The bound on $C_2$ follows from similar arguments.

**Proof of Claim 6.** For notational simplicity, we denote

$$C := X^{t, l} Y^{t, l} - M^* = X^{t, l} Y^{t, l} - X^* Y^{* \top}.$$  \hfill (116)

Since the Frobenius norm is unitarily invariant, we have

$$\|B_1\|_F = \left\| \left(p^{-1} p_{l, i} (C) - p_{l, j} (C)\right) Y^{t, l} \right\|_F$$

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

$$L \triangleq \max_{1 \leq j \leq n} \left\| \left(\delta_{i, j} - p\right) C_{l, j} Y^{t, l}_{j, i} \right\|_2 \leq \|C\|_\infty \|Y^{t, l}\|_{2, \infty} \leq \|C\|_\infty \|Y^*\|_{2, \infty} \leq \|C\|_\infty \|Y^{* \top}\|_{2, \infty},$$

$$V \triangleq \left\| \sum_{j=1}^n \mathbb{E}\left[\left(\delta_{i, j} - p\right)^2\right] C_{l, j}^2 Y^{t, l}_{j, i} Y^{t, l}_{j, i} \right\| \leq p\|C\|_2^2 \left\| \sum_{j=1}^n Y^{t, l}_{j, i} Y^{t, l}_{j, i} \right\|$$

$$= p\|C\|_2^2 \left\| Y^{t, l} \right\|_F^2 \leq 4p\|C\|_2^2 \|Y^*\|_F^2.$$  \hfill (ii)

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

$$\left\| \sum_{j=1}^n (\delta_{i, j} - p) C_{l, j} Y^{t, l}_{j, i} \right\|_2 \leq \sqrt{V \log n} + L \log n \leq \sqrt{p \|C\|_2^2 \|Y^*\|_F^2 \log n + \|C\|_\infty \|Y^{* \top}\|_{2, \infty} \log n}$$

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

$$p\|B_1\|_F \leq \sqrt{p \log n} \|C\|_\infty \|Y^*\|_F + \sqrt{np} \|C\|_\infty \|Y^*\|_{2, \infty} \leq \sqrt{p \log n} \|C\|_\infty \|Y^*\|_F + \sqrt{np} \|C\|_\infty \|Y^{* \top}\|_{2, \infty} \leq \sqrt{p \log n} \|C\|_\infty \|Y^*\|_F + \sqrt{np} \|C\|_\infty \|Y^{* \top}\|_{2, \infty}$$

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

To finish up, we make the observation that

$$\|C\|_\infty = \left\| X^{t, l} R^{t, l} (Y^{t, l} R^{t, l}) - X^* Y^{* \top} \right\|_{\infty}$$

$$\leq \left\| X^{t, l} R^{t, l} - X^* \right\|_{2, \infty} \left\| (Y^{t, l} R^{t, l}) Y^{* \top} \right\|_{\infty} + \left\| X^* (Y^{t, l} R^{t, l} - Y^*) \right\|_{\infty}$$

$$\leq \|X^{t, l} R^{t, l} - X^*\|_{2, \infty} \|Y^{t, l} R^{t, l}\|_{2, \infty} + \|X^*\|_{2, \infty} \|Y^{t, l} R^{t, l} - Y^*\|_{2, \infty}$$

$$\leq 3\|F^{t, l} R^{t, l} - F^*\|_{2, \infty} \|F^*\|_{2, \infty},$$

(118)
where the last line arises from Lemma 18. This combined with (117) gives
\[ \| B_1 \|_F \lesssim \sqrt{\frac{\log n}{p}} \| C \|_\infty \| Y^* \|_F + \sqrt{\frac{\eta}{p}} \| C \|_\infty \| Y^* \|_{2,\infty} \]
\[ \overset{(i)}{\lesssim} \sqrt{\frac{\log n}{p}} \| F^{t,(l)} R^{t,(l)} - F^* \|_{2,\infty} \| Y^* \|_F + \sqrt{\frac{\eta}{p}} \| F^{t,(l)} R^{t,(l)} - F^* \|_{2,\infty} \| F^* \|_{2,\infty} \]
\[ \overset{(ii)}{\lesssim} \sqrt{\frac{\log n}{p}} \| F^{t,(l)} R^{t,(l)} - F^* \|_{2,\infty} \sqrt{\frac{\mu^2 r}{n} \sigma_{\text{max}}} + \sqrt{\frac{\eta}{p}} \| F^{t,(l)} R^{t,(l)} - F^* \|_{2,\infty} \frac{\mu r}{n} \sigma_{\text{max}} \]
\[ \lesssim \sqrt{\frac{\mu^2 r^2 \log n}{np}} \| F^{t,(l)} R^{t,(l)} - F^* \|_{2,\infty} \sigma_{\text{max}}, \]
where (i) comes from (118), and (ii) makes use of the incoherence condition \( \| F^* \|_{2,\infty} \lesssim \sqrt{\mu r \sigma_{\text{max}}/n} \) and the fact that \( \| Y^* \|_F \leq \sqrt{\sigma_{\text{max}}} \).

**Proof of Claim 7.** Instate the notation in proof of Claim 6. By the unitary invariance of Frobenius norm and the fact that all nonzero entries of the matrix \( W \) reside in its \( l \)th row, we have
\[ p \| B_2 \|_F = \left\| p W^T X^{t,(l)} \right\|_F = \left\| \begin{pmatrix} (\delta_{l1} - p) C_{l1} \\ \vdots \\ (\delta_{ln} - p) C_{ln} \end{pmatrix} X_{l\cdot}^{t,(l)} \right\|_F = \| b \|_2 \| X_{l\cdot}^{t,(l)} \|_2. \]

We can write \( b \) as
\[ b = \sum_{j=1}^n e_j (\delta_{lj} - p) C_{lj} = \sum_{j=1}^n u_j. \]

Note that for all \( j \), one has
\[ L \triangleq \max_{1 \leq j \leq n} \| u_j \|_2 \leq \| C \|_\infty, \]
\[ V \triangleq \left\| \sum_{j=1}^n \mathbb{E}[(\delta_{lj} - p)^2] C_{lj}^2 e_j^T e_j \right\| \leq p \| C \|_\infty^2 \left\| \sum_{j=1}^n e_j^T e_j \right\| = np \| C \|_\infty^2. \]

Then the matrix Bernstein inequality [Tro15, Theorem 6.1.1] reveals that
\[ \| b \|_2 \lesssim \sqrt{V \log n} + L \log n \lesssim \sqrt{np \log n} \| C \|_\infty + \| C \|_\infty \log n \]
\[ \lesssim \sqrt{np \log n} \| C \|_\infty \]
\[ \lesssim \sqrt{np \log n} \| F^{t,(l)} R^{t,(l)} - F^* \|_{2,\infty} \| F^* \|_{2,\infty} \]
with probability exceeding \( 1 - O(n^{-100}) \) as long as \( np \gg \log n \). Here the last relation uses (118). Observe that \( \| X^{t,(l)} \|_{2,\infty} \leq 2 \| F^* \|_{2,\infty} \) as long as \( \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n}{p}} \leq \frac{1}{\sqrt{\kappa^2 \log n}} \); see Lemma 18. Making use of the incoherence condition (97a) to get
\[ \| B_2 \|_F \lesssim \sqrt{\frac{n \log n}{p}} \| F^{t,(l)} R^{t,(l)} - F^* \|_{2,\infty} \| F^* \|_{2,\infty}^2 \lesssim \sqrt{\frac{\mu^2 r^2 \log n}{np}} \| F^{t,(l)} R^{t,(l)} - F^* \|_{2,\infty} \sigma_{\text{max}}. \]

We can then conclude the proof. \( \square \)

**D.6 Proof of Lemma 13**

Without loss of generality, we assume \( 1 \leq l \leq n \). One can then decompose \( (F^{t+1,(l)} H^{t+1,(l)} - F^*)_t \) as
\[ (F^{t+1,(l)} H^{t+1,(l)} - F^*)_t = X_{t\cdot}^{t+1,(l)} H^{t+1,(l)} - X_t^*. \]
Next, we turn attention to which together with the triangle inequality yields
\[ \Delta_{\sigma} = \sqrt{\|X_{l,i}^t H^{t+1,i} - X_{l,i}^t\|^2} \leq \eta \left( \|\Delta_{\Delta} X_{l,i}^t\|_{\infty} + \frac{\lambda}{\sigma_{\min}} \right) \|Y_{l,i}^t H^{t+1,i}\| \]

\[ \Rightarrow a_1 = X_{l,i}^t H^{t+1,i} - X_{l,i}^t - \eta \left[ \left( X_{l,i}^t Y_{l,i}^t H^{t+1,i} - M^* \right)_{l,i} Y_{l,i}^t + \frac{\lambda}{p} X_{l,i}^t \right] H^{t+1,i} \]
\[ := a_2 \]

Note that here \( a_1 \) and \( a_2 \) are \( r \)-dimensional row vectors. In the sequel, let us control \( \|a_1\|_2 \) and \( \|a_2\|_2 \) separately.

1. We begin with \( a_1 \). For notational convenience, define \( \Delta_{\Delta} \triangleq \begin{bmatrix} \Delta_{\Delta X} \\Delta_{\Delta Y} \end{bmatrix} \), where \( \Delta_{\Delta X} \triangleq X_{l,i}^t H^{t+1,i} - X^* \) and \( \Delta_{\Delta Y} \triangleq Y_{l,i}^t H^{t+1,i} - Y^* \). Then \( a_1 \) can be rewritten as

\[ a_1 = (\Delta_{\Delta X})_{l,i} - \eta \left[ (\Delta_{\Delta X})_{l,i} (Y_{l,i}^t H^{t+1,i})^\top + X_{l,i}^* \Delta_{\Delta Y}^\top \right] Y_{l,i}^t H^{t+1,i} - \eta \frac{\lambda}{p} X_{l,i}^t H^{t+1,i} \]
\[ := \begin{bmatrix} I - \eta (Y_{l,i}^t H^{t+1,i})^\top Y_{l,i}^t H^{t+1,i} \\ - \eta X_{l,i}^* \Delta_{\Delta Y}^\top H^{t+1,i} - \eta \frac{\lambda}{p} X_{l,i}^t H^{t+1,i} \end{bmatrix} \]

which together with the triangle inequality yields

\[ \|a_1\|_2 \leq \|I - \eta (Y_{l,i}^t H^{t+1,i})^\top Y_{l,i}^t H^{t+1,i}\| \cdot \|\Delta_{\Delta X}\|_{l,i} + \eta \|X_{l,i}^*\|_2 \cdot \|\Delta_{\Delta Y}\| + \eta \frac{\lambda}{p} \|X_{l,i}^t H^{t+1,i}\|_{2,\infty} \]

In view of Lemma 18, we have
\[ \sigma_{\min}/2 \leq \sigma_{\min} \left[ (Y_{l,i}^t H^{t+1,i})^\top Y_{l,i}^t H^{t+1,i} \right] \leq \sigma_{\max} \left[ (Y_{l,i}^t H^{t+1,i})^\top Y_{l,i}^t H^{t+1,i} \right] \leq 2\sigma_{\max} \]

\[ \|\Delta_{\Delta} \| \leq \|\Delta_{\Delta} \|_{l,i} \leq 2C_{\Delta} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \|X^*\|_{l,i}, \] (19)

\[ \|Y_{l,i}^t H^{t+1,i}\| \leq \|F_{l,i}^t\|_{l,i} \leq 2\|X^*\|_{l,i} \] and
\[ \|X_{l,i}^t H^{t+1,i}\|_{2,\infty} \leq \|F_{l,i}^t\|_{2,\infty} \leq 2\|F^*\|_{2,\infty} \]

provided that the sample size obeys \( n \gg \kappa \mu \) and that the noise satisfies \( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll \frac{1}{\kappa^2 \log n} \). These allow us to further upper bound \( \|a_1\|_2 \) by

\[ \|a_1\|_2 \leq \left( 1 - \frac{\eta \sigma_{\min}}{2} \right) \|\Delta_{\Delta} \|_{l,i} + 4\eta C_{\Delta} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \sigma_{\max} \|F^*\|_{2,\infty} + 2\eta \frac{\lambda}{p} \|F^*\|_{2,\infty} \]

as long as \( \eta \leq 1/(2\sigma_{\max}) \). As an immediate consequence,

\[ \|a_1\|_2 \leq C_{\Delta} \kappa \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \|F^*\|_{2,\infty} + 4\eta C_{\Delta} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \sigma_{\max} \|F^*\|_{2,\infty} + 2\eta \frac{\lambda}{p} \|F^*\|_{2,\infty} \]

\[ \leq \|F^*\|_{2,\infty} \] (120)

where the first inequality follows from the induction hypothesis (90d) and the last one holds as long as \( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll \frac{1}{\kappa^2 \log n} \) and \( \eta \ll 1/\sigma_{\max} \).

2. Next, we turn attention to \( \|a_2\|_2 \), which satisfies

\[ \|a_2\|_2 = \left\| (a_1 + X_{l,i}^t) \left[ (H^{t+1,i})^{-1} H^{t+1,i} - I \right] \right\|_2 \leq \left\| (H^{t+1,i})^{-1} H^{t+1,i} - I \right\| \|a_1 + X_{l,i}^t\|_2 \]
From (120), it is easily seen that
\[ \|a_1 + X_t^i\|_2 \leq \|a_1\|_2 + \|F^*\|_{2,\infty} \leq 2\|F^*\|_{2,\infty}. \]

Regarding the term \(\|(H^{t,(i)})^{-1}H^{t+1,(i)} - I_r\|\), we find the following claim useful.

**Claim 8.** With probability at least 1 \(- \Theta(n^{-100})\), we have
\[
\left\| (H^{t,(i)})^{-1}H^{t+1,(i)} - I_r \right\| \lesssim \eta \kappa C_{op}^2 \left( \frac{\sigma}{\sigma_{min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{min}} \right)^2 \sigma_{max} + \eta^2 C_B \kappa^2 \left( \frac{\sigma}{\sigma_{min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{min}} \right) \sqrt{\tau \sigma_{max}^2}.
\]

Finally, taking the bounds on \(\|a_1\|_2\) and \(\|a_2\|_2\) collectively yields that: for some absolute constant \(\tilde{C} > 0\),
\[
\left\| (F^{t+1,(i)}H^{t+1,(i)} - F^*) \right\|_{1,\infty} \leq \|a_1\|_2 + \|a_2\|_2
\leq \left( 1 - \frac{\eta}{2\sigma_{min}} \right) C_4 \kappa \left( \frac{\sigma}{\sigma_{min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{min}} \right) \|F^*\|_{2,\infty} + 4C_{op}\eta \left( \frac{\sigma}{\sigma_{min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{min}} \right) \sigma_{max} \|F^*\|_{2,\infty} + \frac{2\eta \lambda}{p} \|F^*\|_{2,\infty}
\]
\[+ \tilde{C} \eta \kappa C_{op}^2 \left( \frac{\sigma}{\sigma_{min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{min}} \right)^2 \sigma_{max} \|F^*\|_{2,\infty} + \tilde{C} \eta^2 C_B \kappa^2 \left( \frac{\sigma}{\sigma_{min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{min}} \right) \sqrt{\tau \sigma_{max}^2} \|F^*\|_{2,\infty}
\]
\[\leq C_4 \kappa \left( \frac{\sigma}{\sigma_{min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{min}} \right) \|F^*\|_{2,\infty},
\]
provided that \(C_4 \gg C_{op}, \sigma_{min} \sqrt{\frac{n}{p}} \ll 1/\kappa\) and \(\eta \ll 1/(\kappa^2 \sqrt{\tau \sigma_{max}})\). 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 \(\tilde{F}^{t+1} \triangleq \begin{bmatrix} \tilde{X}^{t+1,(i)} \\ \tilde{Y}^{t+1,(i)} \end{bmatrix}\) where
\[
\tilde{X}^{t+1,(i)} = X^{t,(i)}H^{t,(i)} - \eta \frac{1}{2} P_{\Omega_{1,i}} \left( X^{t,(i)}Y^{t,(i),\top} - \bar{M}^* - \bar{E} \right) + P_{\Omega_{1,i}} \left( X^{t,(i)}Y^{t,(i),\top} - \bar{M}^* \right) Y^*,
\]
\[- \frac{\eta}{2} \bar{X}^* - \frac{\eta}{2} \bar{X}^* H^{t,(i),\top} \left( X^{t,(i),\top} Y^{t,(i),\top} - \bar{M}^* - \bar{E} \right) + P_{\Omega_{1,i}} \left( X^{t,(i),\top} Y^{t,(i),\top} - \bar{M}^* \right) H^{t,(i)},
\]
\[
\tilde{Y}^{t+1,(i)} = Y^{t,(i)}H^{t,(i)} - \eta \frac{1}{2} P_{\Omega_{1,i}} \left( X^{t,(i)}Y^{t,(i),\top} - \bar{M}^* - \bar{E} \right) + P_{\Omega_{1,i}} \left( X^{t,(i)}Y^{t,(i),\top} - \bar{M}^* \right) Y^*,
\]
\[- \frac{\eta}{2} \bar{Y}^* - \frac{\eta}{2} \bar{Y}^* H^{t,(i),\top} \left( Y^{t,(i),\top} Y^{t,(i),\top} - \bar{M}^* - \bar{E} \right) + P_{\Omega_{1,i}} \left( X^{t,(i),\top} Y^{t,(i),\top} - \bar{M}^* \right) H^{t,(i)}.
\]
We first claim that \(I_r\) is the best rotation matrix to align \(\tilde{F}^{t+1,(i)}\) and \(F^*\); its proof is similar to that of Claim 4 and is hence omitted for brevity.

**Claim 9.** One has
\[
I_r = \arg \min_{R \in GL} \|\tilde{F}^{t+1,(i)} R - F^*\|_F \quad \text{and} \quad \sigma_{min}(\tilde{F}^{t+1,(i),\top} F^*) \geq \sigma_{min}/2.
\]

With this claim at hand, we intend to invoke Lemma 23 with
\[
S = \tilde{F}^{t+1,(i),\top} F^*, \quad K = (F^{t+1,(i)}H^{t,(i)} - \tilde{F}^{t+1,(i)})^\top F^*
\]
to get
\[
\left\| (H^{t,(i)})^{-1}H^{t+1,(i)} - I_r \right\| = \|\text{sgn}(S + K) - \text{sgn}(S)\| \leq \frac{1}{\sigma_{min}(S)} \|K\|
\]
\[= \frac{1}{\sigma_{min}(\tilde{F}^{t+1,(i),\top} F^*))} \| (F^{t+1,(i)}H^{t,(i)} - \tilde{F}^{t+1,(i)})^\top F^* \|.
\]
\[
\frac{2}{\sigma_{\min}} \| F^{t+1, (l)} H^{t, (l)} - \tilde{F}^{t+1, (l)} \|_F, \quad (121)
\]

where the last line uses Claim 9. Here \( \text{sgn}(A) = UV^T \) for any matrix \( A \) with SVD \( U \Sigma V^T \). It then boils down to controlling \( \| F^{t+1, (l)} H^{t, (l)} - \tilde{F}^{t+1, (l)} \|_F \), for which we have

\[
F^{t+1, (l)} H^{t, (l)} - \tilde{F}^{t+1, (l)} = \eta \begin{bmatrix} B & 0 \\ 0 & B^T \end{bmatrix} \begin{bmatrix} \Delta Y^{t, (l)} \\ \Delta X^{t, (l)} \end{bmatrix} + \frac{\eta}{2} \begin{bmatrix} X^* \\ -Y^* \end{bmatrix} H^{t, (l)}^T C H^{t, (l)} - \eta \frac{\lambda}{p} \Delta^{t, (l)},
\]

where we denote

\[
B \triangleq -p^{-1} P_{\Delta t, l} (X^{t, (l)} Y^{t, (l)} - M^* - E) - P_{t, l} (X^{t, (l)} Y^{t, (l)} - M^*);
\]
\[
C \triangleq X^{t, (l)} H^{t, (l)} - Y^{t, (l)} H^{t, (l)}.
\]

This enables us to obtain

\[
\| F^{t+1, (l)} H^{t, (l)} - \tilde{F}^{t+1, (l)} \| \leq \eta \| B \| \| \Delta^{t, (l)} \| + \frac{\eta}{2} \| F^* \| \| C \| + \frac{\eta \lambda}{p} \| \Delta^{t, (l)} \|. \quad (122)
\]

In view of Lemma 15, one has

\[
\| C \|_F \leq C_{B \kappa \eta} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \sqrt{\tau} \sigma_\max^2. \quad (123)
\]

We are left with bounding \( \| B \| \). Decompose \( B \) into

\[
B = -\frac{1}{p} P_{\Delta t, l} (X^{t, (l)} Y^{t, (l)} T - M^*) + \frac{1}{p} P_{t, l} (X^{t, (l)} Y^{t, (l)} T - M^*) - \frac{1}{p} P_{t, l} (X^{t, (l)} Y^{t, (l)} T - M^*) \triangleq B_1 + B_2 + B_3.
\]

To control \( B_1 \), following the same argument in Lemma 8 we see that

\[
\| B_1 \| \leq \| p^{-1} P_{\Delta t, l} (X^{t, (l)} Y^{t, (l)} T - M^*) - (X^{t, (l)} Y^{t, (l)} T - M^*) \| + \| X^{t, (l)} Y^{t, (l)} T - M^* \|
\lesssim \sqrt{n/p} \| F^{t, (l)} R^{t, (l)} - F^* \|_{2, \infty} \| F^* \|_{2, \infty} + \| F^{t, (l)} R^{t, (l)} - F^* \| F^* \|.
\]

We now move on to \( \| B_2 \| \), which is equal to \( \| b \|_2 / p \) defined in the proof of Claim 7, namely,

\[
\| B_2 \| = \| b \|_2 / p \lesssim \sqrt{n \log n / p} \| F^{t, (l)} R^{t, (l)} - F^* \|_{2, \infty} \| F^* \|_{2, \infty}.
\]

The last term \( B_3 \) can be easily bound via Lemma 3, that is,

\[
\| B_3 \| \leq p^{-1} \| P_{t, l} (E) \| \lesssim \sigma \sqrt{n / p}.
\]

Combining the above three bounds with Lemma 18, we arrive at

\[
\| B \| \leq \| B_1 \| + \| B_2 \| + \| B_3 \|
\lesssim \sqrt{\frac{n \log n}{p}} \| F^{t, (l)} R^{t, (l)} - F^* \|_{2, \infty} \| F^* \|_{2, \infty} + \| F^{t, (l)} R^{t, (l)} - F^* \| F^* \| + \sigma \sqrt{n / p}
\lesssim \sqrt{\frac{n \log n}{p}} (C_{\infty} \kappa + C_0) \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \frac{\mu}{n} \sigma_\max + 2 C_0 \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \sigma_\max + \sigma \sqrt{n / p}
\lesssim C_0 \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{\log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \sigma_\max, \quad (124)
\]

provided that \( n^2 p \gg \kappa^2 \mu^2 \tau^2 n \log n \) and \( C_0 > 0 \) is large enough. Taking (122), (123) and (124) collectively, we arrive at

\[
\| F^{t+1, (l)} H^{t, (l)} - \tilde{F}^{t+1, (l)} \|
\]
provided that $C_{\text{op}}$ is large enough. Here the last relation uses (119). Substitution into (121) yields

$$
\|(H^{t,(l)})^{-1} H^{t+1,(l)} - I_r\| \lesssim \eta C_{\text{op}}^2 \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p} + \frac{\lambda}{p \sigma_{\min}}} \right)^2 \|X^*\| + \eta^2 C_{B \kappa} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p} + \frac{\lambda}{p \sigma_{\min}}} \right) \sqrt{\sigma_{\max}^2},
$$

which concludes the proof.

\[\Box\]

### D.7 Proof of Lemma 14

Fix any $1 \leq l \leq 2n$. Apply the triangle inequality to see that

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

where the second line follows from Lemma 13. Apply Lemma 18 to the $(t + 1)$th iterates to see that

$$
\|F^{t+1} H^{t+1} - F^{t+1,(l)} H^{t+1,(l)}\|_F \leq 5 \kappa \|F^{t+1} H^{t+1} - F^{t+1,(l)} R^{t+1,(l)}\|_F
\leq 5 \eta C_{\text{op}} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p} + \frac{\lambda}{p \sigma_{\min}}} \right) \|F^*\|_{2,\infty}.
$$

Here the second line follows from Lemma 12. Combine (125) and (126) to reach

$$
\|((F^{t+1} H^{t+1} - F^*)_{l,} \|_2 \leq 5 \eta C_{\text{op}} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p} + \frac{\lambda}{p \sigma_{\min}}} \right) \|F^*\|_{2,\infty} + C_4 \kappa \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p} + \frac{\lambda}{p \sigma_{\min}}} \right) \|F^*\|_{2,\infty}
\leq C_{\infty} \kappa \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p} + \frac{\lambda}{p \sigma_{\min}}} \right) \|F^*\|_{2,\infty}
\text{as long as } C_{\infty} \geq 5 C_{3} + C_{4}. \text{ 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

$$
A^t \triangleq X^{t^T} X^t - Y^{t^T} Y^t \quad \text{and} \quad A^{t+1} \triangleq X^{t+1^T} X^{t+1} - Y^{t+1^T} Y^{t+1}.
$$

In view of the gradient descent update rules (25), we have

$$
X^{t+1^T} X^{t+1} = X^{t^T} X^t - \eta \left[ X^{t^T} \nabla X f(X^t, Y^t) + \nabla X f(X^t, Y^t)^T X^t \right] + \eta^2 \nabla X f(X^t, Y^t)^T \nabla X f(X^t, Y^t),
$$

$$
Y^{t+1^T} Y^{t+1} = Y^{t^T} Y^t - \eta \left[ Y^{t^T} \nabla Y f(X^t, Y^t) + \nabla Y f(X^t, Y^t)^T Y^t \right] + \eta^2 \nabla Y f(X^t, Y^t)^T \nabla Y f(X^t, Y^t).
$$

This gives rise to the following identity

$$
A^{t+1} = A^t - \eta B^t + \eta^2 C^t,
$$

(127)
where we denote
\[
B^t \triangleq X^t \nabla_x f(X^t, Y^t) + \nabla_x f(X^t, Y^t)^\top X^t - Y^t \nabla_y f(X^t, Y^t) - \nabla_y f(X^t, Y^t)^\top Y^t,
\]
\[
C^t \triangleq \nabla_x f(X^t, Y^t)^\top \nabla_x f(X^t, Y^t) - \nabla_y f(X^t, Y^t)^\top \nabla_y f(X^t, Y^t).
\]

Denoting
\[
D^t \triangleq p^{-1} \mathcal{P}(X^t Y^t \top - M),
\]
we have
\[
\nabla_x f(X^t, Y^t) = D^t Y^t + \frac{\lambda}{p} X^t \quad \text{and} \quad \nabla_y f(X^t, Y^t) = D^t X^t + \frac{\lambda}{p} Y^t.
\]

With these in mind, a little calculation reveals that
\[
B^t = X^t D^t Y^t + Y^t D^t X^t + \frac{2\lambda}{p} X^t Y^t X^t - Y^t D^t Y^t - X^t D^t Y^t - \frac{2\lambda}{p} Y^t Y^t Y^t = \frac{2\lambda}{p} A^t
\]
as well as
\[
C^t = (D^t Y^t + \frac{\lambda}{p} X^t)^\top (D^t Y^t + \frac{\lambda}{p} X^t) - \left( (D^t X^t + \frac{\lambda}{p} Y^t)^\top (D^t X^t + \frac{\lambda}{p} Y^t) \right)
\]
\[
= Y^t D^t Y^t + \frac{\lambda}{p} Y^t D^t Y^t + \frac{\lambda}{p} X^t D^t Y^t + \left( \frac{\lambda}{p} \right)^2 X^t X^t
\]
\[
- X^t D^t X^t - \frac{\lambda}{p} X^t D^t Y^t - \frac{\lambda}{p} Y^t D^t X^t - \left( \frac{\lambda}{p} \right)^2 Y^t Y^t
\]
\[
= (Y^t D^t Y^t - X^t D^t D^t X^t) + \left( \frac{\lambda}{p} \right)^2 A^t.
\]

Substituting the identities for \( B^t \) and \( C^t \) into (127) yields
\[
A^{t+1} = A^t - 2\eta \frac{\lambda}{p} A^t + \eta^2 \left( Y^t D^t Y^t - X^t D^t X^t \right) + \eta^2 \left( \frac{\lambda}{p} \right)^2 A^t
\]
\[
= (1 - \frac{\lambda \eta}{p}) A^t + \eta^2 \left( Y^t D^t Y^t - X^t D^t X^t \right),
\]
which together with the triangle inequality gives
\[
\|A^{t+1}\|_F \leq (1 - \frac{\lambda \eta}{p}) \|A^t\|_F + \eta^2 \|Y^t D^t Y^t - X^t D^t X^t\|_F
\]
\[
\leq (1 - \frac{\lambda \eta}{p}) \|A^t\|_F + \eta^2 \| Y^t D^t Y^t - X^t D^t X^t \|_F,
\]
as long as \( \lambda \eta / p < 1 \) — a condition that is guaranteed by our assumptions on \( \lambda \) and \( \eta \). It then boils down to controlling \( \| Y^t D^t Y^t - X^t D^t X^t \|_F \), which is supplied in the following claim.

**Claim 10.** Suppose that the sample complexity satisfies \( \eta^2 p \gg \kappa^2 \mu^2 r^2 n \log n \) and that the noise satisfies \( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2 \log n}} \), then one has
\[
\|Y^t D^t D^t Y^t - X^t D^t D^t X^t\|_F \lesssim C_{op}^2 \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right)^2 \sqrt{\sigma_{\max}}.
\]

Taking the above bounds together, we arrive at for some constant \( \tilde{C} > 0 \),
\[
\|A^{t+1}\|_F \leq \left( 1 - \frac{\lambda \eta}{p} \right) \|A^t\|_F + \eta^2 \tilde{C} C_{op}^2 \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right)^2 \sqrt{\sigma_{\max}}
\]
\[
\leq \left( 1 - \frac{\lambda \eta}{p} \right) C_{B \kappa \eta} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \sqrt{\sigma_{\max}} + \eta^2 \tilde{C} C_{op}^2 \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right)^2 \sqrt{\sigma_{\max}}
\]
\[
\leq C_{B \kappa \eta} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \sqrt{\sigma_{\max}},
\]
as long as \( \lambda \geq \sigma \sqrt{n} p \) and \( C_B \gg C_{op}^2 \).
Proof of Claim 10. The triangle inequality yields
\[
\| Y^t D^t D' Y^t - X^t D^t D' X^t \|_F \leq \| Y^t D^t D' Y^t \|_F + \| X^t D^t D' X^t \|_F
\]
It is easy to see from Lemma 18 that
\[
\| Y^t \| \leq 2 \| Y^* \|, \quad \| Y^t \|_F \leq 2 \| Y^* \|_F, \quad \| X^t \| \leq 2 \| X^* \| \quad \text{and} \quad \| X^t \|_F \leq 2 \| X^* \|_F
\]
provided that \( \frac{1}{\sigma \min} \sqrt{\frac{n}{p}} \leq \frac{1}{\sqrt{\kappa^2 \log n}} \). These allow us to further upper bound (129) as
\[
\| Y^t D^t D' Y^t - X^t D^t D' X^t \|_F \leq 4 \| D^t \|_2 \| Y^* \|_F \| Y^* \|_p + 4 \| D^t \|_2 \| X^* \|_F \| X^* \|_p
\]
It remains to bound \( \| D^t \| \). To this end, recall from (128) that
\[
\| D^t \| \leq \frac{1}{p} \| \mathcal{P}_\Omega (E) \| + \frac{1}{p} \| \mathcal{P}^\text{debias} (X^t Y^\top - M^*) \| + \| X^t Y^\top - M^* \|.
\]
Next, repeating the arguments in the proof of Lemma 8 gives
\[
\| \mathcal{P}^\text{debias} (X^t Y^\top - M^*) \| = \| \mathcal{P}^\text{debias} [X^t H^t (Y^t H^t)^\top - M^*] \|
\]
which together with the induction hypothesis (90e) yields
\[
\frac{1}{p} \| \mathcal{P}^\text{debias} (X^t Y^\top - M^*) \| \leq \sqrt{\frac{n}{p}} C_{\infty \kappa} \left( \frac{\sigma}{\sigma \min} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma \min} \right) \| F^* \|_{2, \infty} \| F^* \|_{2, \infty}
\]
Here the last relation uses the incoherence assumption \( \| F^* \|_{2, \infty} \leq \sqrt{\mu \sigma \max / \eta} \) (cf. (97a)). Regarding \( \| X^t Y^\top - M^* \| \), the triangle inequality reveals that
\[
\| X^t Y^\top - M^* \| = \| X^t H^t (Y^t H^t)^\top - M^* \|
\]
\[
\leq \| X^t H^t (Y^t H^t)^\top - X^t H^t Y^*^\top \| + \| X^t H^t Y^*^\top - X^t Y^*^\top \|
\]
Combine the induction hypothesis (90b) and the fact that \( \| X^t H^t \| = \| X^t \| \leq 2 \| X^* \| \) to reach
\[
\| X^t Y^\top - M^* \| \leq C_{\op} \left( \frac{\sigma}{\sigma \min} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma \min} \right) \sigma \max.
\]
Putting together the previous three bounds, we arrive at
\[
\| D^t \| \leq \sigma \left( \frac{\sigma}{\sigma \min} \sqrt{\frac{n}{p}} \right) + C_{\op} \left( \frac{\sigma}{\sigma \min} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma \min} \right) \sqrt{\frac{\mu^2 r^2}{np} \sigma \max} + C_{\op} \left( \frac{\sigma}{\sigma \min} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma \min} \right) \sigma \max
\]
\[
\leq C_{\op} \left( \frac{\sigma}{\sigma \min} \sqrt{\frac{n}{p}} + \frac{\lambda}{p \sigma \min} \right) \sigma \max
\]
since \( np \gg \kappa^2 \mu^2 r^2 \log n \). Putting (131) back to (130) leads to the claimed upper bound.

The upper bound on the leave-one-out sequences can be derived similarly. For brevity, we omit it. \( \square \)
D.9 Proof of Lemma 16

In light of the facts that \( f(\mathbf{FR}) = f(\mathbf{F}) \) and \( \nabla f(\mathbf{FR}) = \nabla f(\mathbf{F}) \mathbf{R} \) for any \( \mathbf{R} \in \mathcal{O}^{r \times r} \), one has

\[
\begin{align*}
    f(\mathbf{F}^{t+1}) &= f((\mathbf{F}^{t} - \eta \nabla f(\mathbf{F}^{t}))^{T} \mathbf{H}^{t}) \\
    &= f(\mathbf{F}^{t} \mathbf{H}^{t} - \eta \nabla f(\mathbf{F}^{t} \mathbf{H}^{t})) \\
    &= f(\mathbf{F}^{t} \mathbf{H}^{t} - \eta \nabla f(\mathbf{F}^{t} \mathbf{H}^{t})) + \frac{\eta^{2}}{2} \text{vec}(\nabla f(\mathbf{F}^{t} \mathbf{H}^{t}))^{T} \nabla^{2} f(\tilde{\mathbf{F}}) \text{vec}(\nabla f(\mathbf{F}^{t} \mathbf{H}^{t}))
\end{align*}
\]

for some \( \tilde{\mathbf{F}} \) which lies between \( \mathbf{F}^{t} \mathbf{H}^{t} \) and \( \mathbf{F}^{t} \mathbf{H}^{t} - \eta \nabla f(\mathbf{F}^{t} \mathbf{H}^{t}) \). Suppose for the moment that

\[
\begin{align*}
    \| \mathbf{F}^{t} \mathbf{H}^{t} - \mathbf{F}^{*} \|_{2,\infty} &\leq \frac{1}{2000 \sqrt{n}} \| \mathbf{X}^{*} \|, \quad (132a) \\
    \| \mathbf{F}^{t} \mathbf{H}^{t} - \eta \nabla f(\mathbf{F}^{t} \mathbf{H}^{t}) - \mathbf{F}^{*} \|_{2,\infty} &\leq \frac{1}{1000 \sqrt{n}} \| \mathbf{X}^{*} \|. \quad (132b)
\end{align*}
\]

One can invoke Lemma 17 to obtain \( \| \nabla^{2} f(\tilde{\mathbf{F}}) \| \leq 10\sigma_{\text{max}} \) and hence

\[
\begin{align*}
    f(\mathbf{F}^{t+1}) &\leq f(\mathbf{F}^{t} \mathbf{H}^{t}) - \eta \| \nabla f(\mathbf{F}^{t} \mathbf{H}^{t}) \|_{F}^{2} + 5\eta^{2} \sigma_{\text{max}} \| \nabla f(\mathbf{F}^{t} \mathbf{H}^{t}) \|_{F}^{2} \\
    &= f(\mathbf{F}^{t}) - \eta \| \nabla f(\mathbf{F}^{t}) \|_{F}^{2} + 5\eta^{2} \sigma_{\text{max}} \| \nabla f(\mathbf{F}^{t}) \|_{F}^{2} \\
    &\leq f(\mathbf{F}^{t}) - \frac{\eta}{2} \| \nabla f(\mathbf{F}^{t}) \|_{F}^{2}.
\end{align*}
\]

Here the equality uses again the facts that \( f(\mathbf{FR}) = f(\mathbf{F}) \) and \( \nabla f(\mathbf{FR}) = \nabla f(\mathbf{F}) \mathbf{R} \) for any \( \mathbf{R} \in \mathcal{O}^{r \times r} \) and the last inequality holds as long as \( \eta \leq \frac{1}{1000 \sqrt{n}} \). We are left with proving the aforementioned conditions (132). The first condition has been established in the proof of Lemma 10 and hence we concentrate on the second one, namely (132b). Apply the triangle inequality and the fundamental theorem of calculus [Lan93, Chapter XIII, Theorem 4.2] to obtain

\[
\begin{align*}
    \| \mathbf{F}^{t} \mathbf{H}^{t} - \eta \nabla f(\mathbf{F}^{t} \mathbf{H}^{t}) - \mathbf{F}^{*} \|_{2,\infty} &\leq \| \mathbf{F}^{t} \mathbf{H}^{t} - \mathbf{F}^{*} \|_{2,\infty} + \eta \| \nabla f(\mathbf{F}^{t} \mathbf{H}^{t}) - \nabla f(\mathbf{F}^{*}) \|_{F} + \eta \| \nabla f(\mathbf{F}^{*}) \|_{F} \\
    &\leq \| \mathbf{F}^{t} \mathbf{H}^{t} - \mathbf{F}^{*} \|_{2,\infty} + \eta \left( \int_{0}^{1} \| \nabla^{2} f(\mathbf{F}(\tau)) \|_{F} \, d\tau \right) \| \text{vec}(\mathbf{F}^{t} \mathbf{H}^{t} - \mathbf{F}^{*}) \|_{F} + \eta \| \nabla f(\mathbf{F}^{*}) \|_{F},
\end{align*}
\]

where \( \mathbf{F}(\tau) \triangleq \mathbf{F}^{*} + \tau(\mathbf{F}^{t} \mathbf{H}^{t} - \mathbf{F}^{*}) \) for \( 0 \leq \tau \leq 1 \). Following similar arguments in the proof of Lemma 10 and the proof of Lemma 9, one obtains

\[
\begin{align*}
    \eta \left( \int_{0}^{1} \| \nabla^{2} f(\mathbf{F}(\tau)) \|_{F} \, d\tau \right) \| \text{vec}(\mathbf{F}^{t} \mathbf{H}^{t} - \mathbf{F}^{*}) \|_{F} &\leq \eta \cdot 10\sigma_{\text{max}} \| \mathbf{F}^{t} \mathbf{H}^{t} - \mathbf{F}^{*} \|_{F} + \eta \frac{\lambda}{p} \sqrt{\sigma_{\text{max}}} \\
    &\lesssim \eta \sigma_{\text{max}} \left( \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n}{\eta}} + \frac{\lambda}{p \sigma_{\text{min}}} \right) \sqrt{\eta} \| \mathbf{X}^{*} \| + \eta \sigma_{\text{min}} \frac{\lambda}{p \sigma_{\text{min}}} \sqrt{\eta} \| \mathbf{X}^{*} \| \leq \frac{1}{2000 \sqrt{n}} \| \mathbf{X}^{*} \|.
\end{align*}
\]

Here the middle inequality uses the induction hypothesis (90b) and the last relation holds true provided that \( \lambda \simeq \sigma \sqrt{\frac{n}{p}} \), \( \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n}{p}} \ll 1/\sqrt{\eta} \) and that \( \eta \ll 1/(\kappa n \sigma_{\text{max}}) \). This proves the second condition and also the whole lemma.

D.10 Proof of Lemma 17

We start by defining a new loss function

\[
f_{\text{clean}}(\mathbf{X}, \mathbf{Y}) \triangleq \frac{1}{2p} \left( \left\| \mathbf{P}_{\Omega} \left( \mathbf{XY}^{T} - \mathbf{M}^{*} \right) \right\|_{F}^{2} + \frac{1}{8} \| \mathbf{X}^{T} \mathbf{X} - \mathbf{Y}^{T} \mathbf{Y} \|_{F}^{2} \right);
\]

compared with \( f_{\text{aug}} \), this new function \( f_{\text{clean}} \) sets \( \lambda = 0 \) and excludes the noise from consideration. It is therefore straightforward to check that for any \( \Delta \in \mathbb{R}^{2n \times r} \),

\[
\text{vec}(\Delta)^{T} \nabla^{2} f_{\text{aug}}(\mathbf{X}, \mathbf{Y}) \text{vec}(\Delta) = \text{vec}(\Delta)^{T} \nabla^{2} f_{\text{clean}}(\mathbf{X}, \mathbf{Y}) \text{vec}(\Delta) - \frac{2}{p} \langle \mathbf{P}_{\Omega}(\mathbf{E}), \Delta \mathbf{X} \Delta^{T} \mathbf{X} \rangle + \frac{\lambda}{p} \| \Delta \|_{F}^{2}.
\]
It has been proven in [CLL19, Lemma 3.2] that under the assumptions stated in Lemma 17, one has
\[
\text{vec}(\Delta)^T \nabla^2 f_{\text{clean}}(X, Y) \text{vec}(\Delta) \geq \frac{1}{2} \sigma_{\min} \|\Delta\|_F^2 \quad \text{and} \quad \|\nabla^2 f_{\text{clean}}(X, Y)\| \leq 5 \sigma_{\text{max}}. \tag{133}
\]
It then boils down to controlling \(-\frac{2}{p} \langle P_\Omega(E), \Delta X \Delta Y^T \rangle + \frac{1}{p} \|\Delta\|_F^2\). To this end, one has
\[
\left|\frac{1}{p} \langle P_\Omega(E), \Delta X \Delta Y^T \rangle \right| \leq \left\| \frac{1}{p} P_\Omega(E) \right\| \|\Delta X \Delta Y^T\|_* \lesssim \sigma \sqrt{\frac{2}{p}} \|\Delta\|_F^2, \tag{134}
\]
where the last relation holds due to Lemma 3 and the elementary fact about the nuclear norm (6), i.e.
\[
2 \|\Delta X \Delta Y^T\|_* \leq \|\Delta X\|_F^2 + \|\Delta Y\|_F^2 = \|\Delta\|_F^2.
\]
Regarding the term \(\lambda \|\Delta\|_F^2 / p\), it is easy to see from the assumption \(\lambda \approx \sigma \sqrt{n/p}\) that \(\frac{1}{p} \|\Delta\|_F^2 \approx \sigma \sqrt{\frac{2}{p}} \|\Delta\|_F^2\).

Combine the above two bounds and use the triangle inequality to reach
\[
\left|\frac{\alpha}{p} \langle P_\Omega(E), \Delta X \Delta Y^T \rangle + \frac{1}{p} \|\Delta\|_F^2 \right| \lesssim \sigma \sqrt{\frac{2}{p}} \|\Delta\|_F^2 \leq \frac{\alpha}{p} \sigma_{\min} \|\Delta\|_F^2, \tag{135}
\]
with the proviso that \(\frac{\sigma_{\min}}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll 1\). Taking (133) and (135) together immediately establishes the claim.

With regards to \(\nabla^2 f(X, Y)\), one has
\[
\text{vec}(\Delta)^T \nabla^2 f(X, Y) \text{vec}(\Delta) = \frac{1}{2} \|P_\Omega(X \Delta^T - M^* - E), \Delta X \Delta Y^T\|^2 + \frac{2}{p} \langle P_\Omega(YX^T - M^* - E), \Delta X \Delta Y^T \rangle + \frac{2}{p} \langle P_\Omega(XX^T - M^*), \Delta X \Delta Y^T \rangle + \frac{2}{p} \langle P_\Omega(YX^T - M^*), \Delta X \Delta Y^T \rangle + \frac{2}{p} \langle P_\Omega(XX^T - M^*), \Delta X \Delta Y^T \rangle - \|\Delta X \Delta Y^T\|_F^2 - 2 \langle XY^T - M^*, \Delta X \Delta Y^T \rangle.
\]

The term \(\alpha_4\) can be bounded by [CLL19, Equation A.4]
\[
|\alpha_4| \leq \frac{1}{2} \sigma_{\min} \left(\|\Delta X\|_F^2 + \|\Delta Y\|_F^2\right) + \frac{1}{2} \left(\|\Delta X Y^T\|_F^2 + \|X^* \Delta Y^T\|_F^2\right) \leq \frac{\sigma_{\max}}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|\Delta\|_F^2.
\]

The term \(\alpha_3\) has been bounded in (134) where \(|\alpha_3| \lesssim \sigma \sqrt{n/p} \|\Delta\|_F^2 \leq \sigma_{\max} \|\Delta\|_F^2\) provided that \(\frac{\sigma_{\min}}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll 1\).

The term \(\alpha_2\) can be written as
\[
|\alpha_2| \leq 2 \|XY^T - M^*\| \|\Delta X \Delta Y^T\|_* \leq (\|X - X^*\| \|Y\| + \|X^*\| \|Y - Y^*\|) \|\Delta\|_F^2.
\]
Since
\[
\left\| \begin{bmatrix} X - X^* \\ Y - Y^* \end{bmatrix} \right\|_F \leq \sqrt{2n} \left\| \begin{bmatrix} X - X^* \\ Y - Y^* \end{bmatrix} \right\|_{2,\infty} \leq \frac{1}{500\kappa} \|X^*\|,
\]
we immediately have
\[
|\alpha_2| \leq \frac{3}{500\kappa} \sigma_{\max} \|\Delta\|_F^2 \leq \frac{1}{2} \sigma_{\max} \|\Delta\|_F^2.
\]

The term \(\alpha_1\) can be bounded by
\[
\alpha_1 \leq 2 \left(\|X^* \Delta Y^T\|_F^2 + \|\Delta X Y^T\|_F^2\right) \leq 2 \left(\|X^*\|^2 \|\Delta Y\|_F^2 + \|Y^*\|^2 \|\Delta X\|_F^2\right) = 2 \sigma_{\max} \|\Delta\|_F^2.
\]
Combining all these bounds yields
\[
\text{vec}(\Delta)^T \nabla^2 f(X, Y) \text{vec}(\Delta) \leq 10 \sigma_{\max} \|\Delta\|_F^2.
\]
D.11 Proof of Lemma 18

The first set of consequences (98) follows straightforwardly from the triangle inequality. For instance, combine the induction hypotheses (90c) and (90e) to obtain

$$\|F^t(I)R^t(I) - F^*\|_{2,\infty} \leq \|F^t(I)R^t(I) - F^tH^t\|_{2,\infty} + \|F^tH^t - F^*\|_{2,\infty}$$

$$\leq (Ce\kappa + C_3) \left( \frac{n \log n}{p} \sqrt{\frac{n \log n}{\sigma_{\min}}} + \frac{\lambda}{p \sigma_{\min}} \right) \|F^*\|_{2,\infty}.$$ 

Similar bounds can be obtained for $\|F^t(I)R^t(I) - F^*\|$ provided that $n \gg \mu r \log n$.

We continue to establish the second set of consequences namely (99). Since $\|\cdot\|$ is unitarily invariant, one can apply the triangle inequality to get

$$\|F^t\| = \|F^tH^t - F^*\| + \|F^*\| \leq C_{\text{opt}} \left( \frac{n \log n}{p} + \frac{\lambda}{p \sigma_{\min}} \right) \|F^*\| + \sqrt{2} \|X^*\| \leq 2 \|X^*\|.$$ 

Here (i) uses the induction hypothesis (90b) and the fact that $\|F^*\| = \sqrt{2}\|X^*\|$, and (ii) holds as long as $\frac{n \log n}{p} \ll 1$. Similarly one can obtain $\|F^t\| \leq 2\|X^*\|$ provided that $\frac{n \log n}{p} \ll 1$ and $\|F^t\|_{2,\infty} \leq 2\|F^*\|_{2,\infty}$ as long as $\frac{n \log n}{p} \ll 1/(\sqrt{2}\log n)$. Notice that along the way, we have also proven that

$$\|F^tH^t - F^*\| \leq \|X^*\|, \quad \|F^tH^t - F^*\|_F \leq \|X^*\|_F \quad \text{and} \quad \|F^tH^t - F^*\|_{2,\infty} \leq \|F^*\|_{2,\infty}.$$ 

We now move on to $\|F^tH^t - F^t(I)H^t(I)\|_F$, for which we intend to apply Lemma 22 to connect it with $\|F^tH^t - F^t(I)R^t(I)\|_F$. First, in view of the induction hypothesis (90b), one has

$$\|F^tH^t - F^*\|_F \leq C_{\text{opt}} \left( \frac{n \log n}{p} + \frac{\lambda}{p \sigma_{\min}} \right) \|X^*\| \|F^*\| \leq \frac{\sigma_{\max}}{2},$$

where the equality arises since $\|F^*\| = \sqrt{2}\sigma_{\max}$ (see (97a)) and $\|X^*\| = \sqrt{\sigma_{\max}}$ and the last line holds as long as $\frac{n \log n}{p} \ll 1/\kappa$. In addition, it follows from the induction hypothesis (90c) that

$$\|F^tH^t - F^t(I)R^t(I)\|_F \leq \|F^tH^t - F^t(I)R^t(I)\|_{2,\infty} \leq C_3 \left( \frac{n \log n}{p} + \frac{\lambda}{p \sigma_{\min}} \right) \|F^*\| \leq \frac{\sigma_{\max}}{4},$$

where the penultimate inequality arises from the facts that $\|F^*\|_{2,\infty} \leq \sqrt{\mu r \sigma_{\max}/n}$ and that $\|F^*\| = \sqrt{2}\sigma_{\max}$ (cf. (97)), and the last relation holds as long as $\left( \frac{n \log n}{p} + \frac{\lambda}{p \sigma_{\min}} \right) \sqrt{\frac{\mu r \sigma_{\max}}{n}} \ll 1/\kappa$. Invoke Lemma 22 with

$$F_0 = F^*, \quad F_1 = F^tH^t \quad \text{and} \quad F_2 = F^t(I)R^t(I)$$

to arrive at

$$\|F^tH^t - F^t(I)H^t(I)\|_F \leq \frac{5\sigma_{\max}^2(F^*)^2}{\sigma_{\max}^2(F^*)} \|F^tH^t - F^t(I)R^t(I)\|_F = 5\kappa \|F^tH^t - F^t(I)R^t(I)\|_F.$$

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 (29)

We single out the proof of \( \|X^t Y^t - M^*\|_\infty \), whereas the proofs of \( \|X^t Y^t - M^*\|_F \) and \( \|X^t Y^t - M^*\|_F \) follow from the same argument. Recognize the following decomposition

\[
X^t Y^t - M^* = (X^t H^t - X^*) (Y^t H^t)^T + X^* (Y^t H^t - Y^*)^T,
\]

which together with the triangle inequality gives

\[
\|X^t Y^t - M^*\|_\infty \leq \left( \|X^t H^t - X^*\| \|Y^t H^t\|_\infty \right) + \|X^* \|_\infty \|Y^t H^t - Y^*\|_\infty.
\]

In view of Lemma 18, one has \( \|Y^t H^t\|_\infty \leq 2\|F^*\|_\infty \) as long as the noise obeys \( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \leq \frac{1}{\sqrt{k} \log n} \).

This further implies that

\[
\|X^t Y^t - M^*\|_\infty \leq 3C_n \kappa \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \|F^*\|_\infty \|F^*\|_\infty
\]

\[
\leq 3C_n \sqrt{\kappa \mu} \left( \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \|M^*\|_\infty,
\]

where the last relation is \( \|F^*\|_\infty \|F^*\|_\infty \leq \sqrt{\kappa \mu} \|M^*\|_\infty \). To see this, one has for any \( 1 \leq i \leq n \),

\[
n \|M^*\|_\infty \geq \sum_{j=1}^{n} (M^*_{ji})^2 = X^*_i, Y^* i^T Y^* i^T \geq \|X^*_i\|_2^2 \lambda_{\min} (Y^* Y^*) = \sigma_{\min} \|X^*_i\|_2^2.
\]

Here \( \lambda_{\min}(\cdot) \) denotes the minimum eigenvalue. Since the inequality holds for all \( 1 \leq i \leq n \), we arrive at

\[
\|X^*\|_\infty \leq \frac{n}{\sigma_{\min}} \|M^*\|_\infty.
\]

Similarly one can obtain \( \|Y^*\|_\infty \leq \sqrt{n/\sigma_{\min}} \|M^*\|_\infty \), which further implies \( \|F^*\|_\infty \leq \max\{\|X^*\|_\infty, \|Y^*\|_\infty\} \leq \sqrt{n/\sigma_{\min}} \|M^*\|_\infty \). As a result, we arrive at

\[
\|F^*\|_\infty \|F^*\|_\infty \leq \frac{n}{\sigma_{\min}} \|M^*\|_\infty \cdot \sqrt{\frac{\mu \kappa}{n} \sqrt{\sigma_{\max} \|M^*\|_\infty}}.
\]

Here we used the incoherence assumption (97a).

E Technical lemmas

Lemma 19. Suppose \( n^2 p \geq Cn \log n \) for some sufficiently large constant \( C > 0 \). Then with probability exceeding \( 1 - O(n^{-10}) \),

\[
\left| p^{-1} \|P_{\Omega} (AB^T)\|_F^2 - \|AB^T\|_F^2 \right| \leq 3n \min\{\|A\|_2^2, \|B\|_2^2, \|A\|_2^2 \|B\|_2^2 \}
\]

holds uniformly for all matrices \( A, B \in \mathbb{R}^{n \times r} \).

Proof. In view of [ZL16, Lemma 9], one has

\[
p^{-1} \|P_{\Omega} (AB^T)\|_F^2 \leq 2n \min\{\|A\|_2^2, \|B\|_2^2, \|A\|_2^2 \|B\|_2^2 \}
\]

with high probability. In addition, simple algebra reveals that

\[
\|AB^T\|_F^2 \leq \|A\|_F^2 \|B\|_F^2 \leq n \|\|A\|_F^2 \|B\|_2^2 \leq n \|A\|_2^2 \|B\|_2^2.
\]

and, similarly, \( \|AB^T\|_F^2 \leq n \|A\|_F^2 \|B\|_2^2 \|B\|_2^2 \). Combining the previous bounds with the triangle inequality establishes the claim.
Lemma 20. Let $U\Sigma V^T$ be the SVD of a rank-$r$ matrix $XY^T$ with $X, Y \in \mathbb{R}^{n \times r}$. Then there exists an invertible matrix $Q \in \mathbb{R}^{r \times r}$ such that $X = U\Sigma^{1/2}Q$ and $Y = V\Sigma^{1/2}Q^{-T}$. In addition, one has
\[
\|\Sigma Q - \Sigma Q^{-1}\|_F \leq \frac{1}{\sigma_{\min}(\Sigma)} \|X^T X - Y^T Y\|_F, \tag{136}
\]

where $U_Q\Sigma_Q V_Q^T$ is the SVD of $Q$. In particular, if $X$ and $Y$ have balanced scale, i.e. $X^T X - Y^T Y = 0$, then $Q$ must be a rotation matrix.

Proof. The existence of $Q$ is trivial by setting
\[
Q = \Sigma^{-1/2}U^T X.
\]

To see this, one has
\[
U\Sigma^{1/2}Q = U\Sigma^{1/2}\Sigma^{-1/2}U^T X = UU^T X = X,
\]

where the last equality follows from the fact that the columns of $U$ are the left singular vectors of $X$. The relation $Y = V\Sigma^{1/2}Q^{-T}$ can also be verified by the identity
\[
XY^T = U\Sigma^{1/2}QY^T = U\Sigma V^T.
\]

We now move on to proving the perturbation bound (136). In view of the SVD of $Q$, i.e. $Q = U_Q\Sigma_Q V_Q^T$, one can obtain
\[
X^T X - Y^T Y = Q^T \Sigma Q - Q^{-1} \Sigma Q^{-T}
\]
\[
= V_Q \Sigma_Q U_Q^T \Sigma U_Q \Sigma_Q V_Q^T - V_Q \Sigma_Q^{-1} U_Q^T \Sigma U_Q \Sigma_Q^{-1} V_Q^T.
\]

Denote $B := U_Q^T \Sigma U_Q > 0$. Then we have
\[
\|X^T X - Y^T Y\|_F^2 = \|V_Q \Sigma_Q B \Sigma Q V_Q^T - V_Q \Sigma_Q^{-1} B \Sigma_Q^{-1} V_Q^T\|_F^2 = \|\Sigma B \Sigma Q - \Sigma_Q^{-1} B \Sigma_Q^{-1}\|_F^2.
\]

Let $C = \Sigma_Q B^{1/2}$ and $D = \Sigma_Q^{-1} B^{1/2}$, and denote $\Delta = C - D$. One then has
\[
\|X^T X - Y^T Y\|_F^2 = \|CC^T - DD^T\|_F^2 = \|C\Delta^T + \Delta C^T - \Delta \Delta^T\|_F^2
\]
\[
= \text{Tr}\left[ (2C^T \Delta \Delta^T + \Delta \Delta^T C^T + 2C^T \Delta C^T \Delta - 4C^T \Delta \Delta^T \Delta) \right]
\]
\[
= \text{Tr}\left[ (\Delta^T - \sqrt{2} \Delta \Delta^T) + (4 - 2\sqrt{2})C^T (C - \Delta) \Delta^T \Delta + (2\sqrt{2} - 1)C^T C \Delta^T \Delta \right].
\]

Note that $C^T D = B$ and that $C^T \Delta = C^T C - C^T B = C^T C - B$ is symmetric. One can continue the bound as
\[
\|X^T X - Y^T Y\|_F^2 \geq \text{Tr}(B \Delta \Delta^T),
\]

where the inequality follows since $4 - 2\sqrt{2} \geq 1$. Write $B = B^{1/2} \cdot B^{1/2}$ to see
\[
\|X^T X - Y^T Y\|_F^2 \geq \text{Tr}(B^{1/2} \Delta \Delta^T B^{1/2}) = \|B^{1/2} \Delta\|_F^2
\]
\[
= \|B^{1/2}(\Sigma_Q - \Sigma_Q^{-1}) B^{1/2}\|_F^2
\]
\[
\geq \sigma_{\min}(B) \|\Sigma_Q - \Sigma_Q^{-1}\|_F^2.
\]

Recognizing that $\sigma_{\min}(B) = \sigma_{\min}(\Sigma)$ finishes the proof of (136).

Combining $X^T X = Y^T Y$ and (136) yields $\|\Sigma_Q - \Sigma_Q^{-1}\|_F = 0$, which implies $\Sigma_Q = I$. Under this circumstance, $Q = U_Q \Sigma_Q V_Q^T = U_Q V_Q^T$ is a rotation matrix. The proof is then complete. \qed
Lemma 21. For all $A, B, C, D \in \mathbb{R}^{n \times r}$, one has

$$\|\langle P_\Omega (AC^T), P_\Omega (BD^T) \rangle - p \langle AC^T, BD^T \rangle \| \leq \|P_\Omega (11^T) - p11^T\| \|A\|_{2,\infty} \|B\|_F \|C\|_{2,\infty} \|D\|_F.$$  

Proof. This is a simple consequence of [CL17, Lemma 4.4], where they have shown

$$\|\langle P_\Omega (AC^T), P_\Omega (BD^T) \rangle - p \langle AC^T, BD^T \rangle \| \leq \sqrt{\sum_{k=1}^n \|A_k\|_2^2 \|B_k\|_2^2} \sqrt{\sum_{k=1}^n \|C_k\|_2^2 \|D_k\|_2^2}.$$  

Recognize that

$$\sum_{k=1}^n \|A_k\|_2^2 \|B_k\|_2^2 \leq \|A\|_{2,\infty}^2 \|B\|_F^2$$  

and, similarly,

$$\sum_{k=1}^n \|B_k\|_2^2 \leq \|B\|_{2,\infty}^2 \|B\|_F^2.$$  

Putting these together concludes the proof.

Lemma 22. Suppose $F_1, F_2, F_0 \in \mathbb{R}^{2n \times r}$ are three matrices such that

$$\|F_1 - F_0\|_F \leq \frac{\sigma_r^2(F_0)}{2} \quad \text{and} \quad \|F_1 - F_2\|_F \leq \frac{\sigma_r^2(F_0)}{4},$$  

where $\sigma_i(A)$ stands for the $i$th largest singular value of $A$. Denote

$$R_1 \triangleq \arg \min_{R \in O^{r \times r}} \|F_1 R - F_0\|_F \quad \text{and} \quad R_2 \triangleq \arg \min_{R \in O^{r \times r}} \|F_2 R - F_0\|_F.$$  

Then the following two inequalities hold true:

$$\|F_1 R_1 - F_2 R_2\|_F \leq \frac{5 \sigma_r^2(F_0)}{\sigma_r^2(F_0)} \|F_1 - F_2\|_F \quad \text{and} \quad \|F_1 R_1 - F_2 R_2\|_F \leq \frac{5 \sigma_r^2(F_0)}{\sigma_r^2(F_0)} \|F_1 - F_2\|_F.$$  

Proof. This is the same as [MWCC17, Lemma 37].

Lemma 23. Let $S \in \mathbb{R}^{r \times r}$ be a nonsingular matrix. Then for any matrix $K \in \mathbb{R}^{r \times r}$ with $\|K\| \leq \sigma_{\min}(S)$, one has

$$\|\text{sgn}(S + K) - \text{sgn}(S)\| \leq \frac{2}{\sigma_{r-1}(S) + \sigma_r(S)} \|K\|,$$

where $\text{sgn}(\cdot)$ denotes the matrix sign function, i.e. $\text{sgn}(A) = UV^\top$ for a matrix $A$ with SVD $U \Sigma V^\top$.

Proof. This is the same as [MWCC17, Lemma 36].