Bridging Convex and Nonconvex Optimization in Robust PCA: Noise, Outliers, and Missing Data

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

February 28, 2021

Abstract

This paper delivers improved theoretical guarantees for the convex programming approach in low-rank matrix estimation, in the presence of (1) random noise, (2) gross sparse outliers, and (3) missing data. This problem, often dubbed as *robust principal component analysis (robust PCA)*, finds applications in various domains. Despite the wide applicability of convex relaxation, the available statistical support (particularly the stability analysis vis-à-vis random noise) remains highly suboptimal, which we strengthen in this paper. When the unknown matrix is well-conditioned, incoherent, and of constant rank, we demonstrate that a principled convex program achieves near-optimal statistical accuracy, in terms of both the Euclidean loss and the ℓ_∞ loss. All of this happens even when nearly a constant fraction of observations are corrupted by outliers with arbitrary magnitudes. The key analysis idea lies in bridging the convex program in use and an auxiliary nonconvex optimization algorithm, and hence the title of this paper.

Keywords: robust principal component analysis, nonconvex optimization, convex relaxation, ℓ_∞ guarantees, leave-one-out analysis

Contents

Author names are sorted alphabetically. Corresponding author: Yuxin Chen (Email: `yuxin.chen@princeton.edu`).

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

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

‡Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, CA 94720, USA; Email: `congm@berkeley.edu`.

1 Introduction

A diverse array of science and engineering applications (e.g. video surveillance, joint shape matching, graph clustering, covariance modeling, graphical models) involves estimation of low-rank matrices [CLC19, CLMW11, CGH14, JCSX11, CPW12, FLM13, DR16]. The imperfectness of data acquisition processes, however, presents several common yet critical challenges: (1) random noise: which reflects the uncertainty of the environment and/or the measurement processes; (2) outliers: which represent a sort of corruption that exhibits abnormal behavior; and (3) incomplete data, namely, we might only get to observe a fraction of the matrix entries. Low-rank matrix estimation algorithms aimed at addressing these challenges have been extensively studied under the umbrella of *robust principal component analysis (Robust PCA)* [CSPW11,CLMW11], a terminology popularized by the seminal work [CLMW11].

To formulate the above-mentioned problem in a more precise manner, imagine that we seek to estimate an unknown low-rank matrix $\boldsymbol{L}^\star \in \mathbb{R}^{n_1\times n_2}$. What we can obtain is a collection of partially observed and corrupted entries as follows

$$M_{ij} = L^\star_{ij} + S^\star_{ij} + E_{ij}, \qquad (i,j)\in\Omega_{\mathsf{obs}}, \tag{1.1}$$

where $\boldsymbol{S}^\star = [S^\star_{ij}]$ is a matrix consisting of outliers, $\boldsymbol{E} = [E_{ij}]$ represents the random noise, and we only observe entries over an index subset $\Omega_{\mathsf{obs}} \subseteq [n_1]\times[n_2]$ with $[n] := \{1,2,\cdots,n\}$. The current paper assumes that $\boldsymbol{S}^\star$ is a relatively sparse matrix whose non-zero entries might have arbitrary magnitudes. This assumption has been commonly adopted in prior work to model gross outliers, while enabling reliable disentanglement of the outlier component and the low-rank component [CSPW11,CLMW11,CJSC13,Li13]. In addition, we suppose that the entries $\{E_{ij}\}$ are independent zero-mean sub-Gaussian random variables, as commonly assumed in the statistics literature to model a large family of random noise. The aim is to reliably estimate $\boldsymbol{L}^\star$ given the grossly corrupted and possibly incomplete data (1.1). Ideally, this task should be accomplished without knowing the locations and magnitudes of the outliers $\boldsymbol{S}^\star$.

1.1 A principled convex programming approach

Focusing on the noiseless case with $\boldsymbol{E} = \boldsymbol{0}$, the papers by [CSPW11, CLMW11] delivered a positive and somewhat surprising message: both the low-rank component $\boldsymbol{L}^\star$ and the sparse component $\boldsymbol{S}^\star$ can be efficiently recovered with absolutely no error by means of a principled convex program

$$\underset{\boldsymbol{L},\boldsymbol{S}\in\mathbb{R}^{n_1\times n_2}}{\text{minimize}} \quad \|\boldsymbol{L}\|_* + \tau\|\boldsymbol{S}\|_1 \qquad \text{subject to} \quad \mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}+\boldsymbol{S}-\boldsymbol{M}) = \boldsymbol{0}, \tag{1.2}$$

provided that certain "separation" and "incoherence" conditions on $(\boldsymbol{L}^\star, \boldsymbol{S}^\star, \Omega_{\mathsf{obs}})$ hold[1] and that the regularization parameter τ is properly chosen. Here, $\|\boldsymbol{L}\|_*$ denotes the nuclear norm (i.e. the sum of the singular values) of $\boldsymbol{L}$, $\|\boldsymbol{S}\|_1 = \sum_{i,j}|S_{ij}|$ denotes the usual entrywise ℓ_1 norm, and $\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{M})$ denotes the Euclidean projection of a matrix $\boldsymbol{M}$ onto the subspace of matrices supported on Ω_{obs}. Given that the nuclear norm $\|\cdot\|_*$ (resp. the ℓ_1 norm $\|\cdot\|_1$) is the convex relaxation of the rank function $\mathsf{rank}(\cdot)$ (resp. the ℓ_0 counting norm $\|\cdot\|_0$), the rationale behind (1.2) is rather clear: it seeks a decomposition $(\boldsymbol{L},\boldsymbol{S})$ of $\boldsymbol{M}$ by promoting the low-rank structure of $\boldsymbol{L}$ as well as the sparsity structure of $\boldsymbol{S}$.

Moving on to the more realistic noisy setting, a natural strategy is to solve the following regularized least-squares problem

$$\underset{\boldsymbol{L},\boldsymbol{S}\in\mathbb{R}^{n_1\times n_2}}{\text{minimize}} \quad \frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}+\boldsymbol{S}-\boldsymbol{M})\right\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{L}\|_* + \tau\|\boldsymbol{S}\|_1. \tag{1.3}$$

With the regularization parameters $\lambda, \tau > 0$ properly chosen, one hopes to strike a balance between enhancing the goodness of fit (by enforcing $\boldsymbol{L}+\boldsymbol{S}-\boldsymbol{M}$ to be small) and promoting the desired low-complexity structures (by regularizing both the nuclear norm of $\boldsymbol{L}$ and the ℓ_1 norm of $\boldsymbol{S}$). A natural and important question comes into our mind:

> *Where does the algorithm (1.3) stand in terms of its statistical performance vis-à-vis random noise, sparse outliers and missing data?*

Unfortunately, however simple this program (1.3) might seem, the existing theoretical support remains far from satisfactory, as we shall discuss momentarily.

1.2 Theory-practice gaps under random noise

To assess the tightness of prior statistical guarantees for (1.3), we find it convenient to first look at a simple setting where (i) $n_1 = n_2 = n$, (ii) $\boldsymbol{E}$ consists of independent Gaussian components, namely, $E_{ij} \sim \mathcal{N}(0,\sigma^2)$, and (iii) there is no missing data. This simple scenario is sufficient to illustrate the sub-optimality of prior theory.

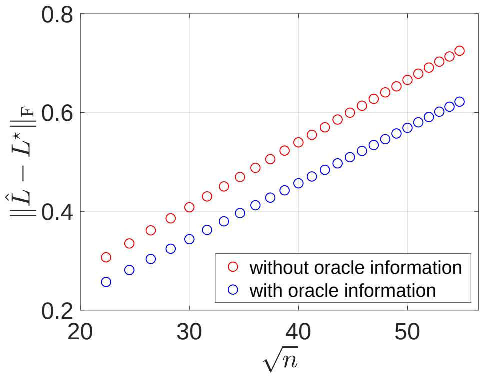
Prior statistical guarantees The paper [ZLW+10] was the first to derive a sort of statistical performance guarantees for the above convex program. Under mild conditions, [ZLW+10] demonstrated that any minimizer $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{S}})$ of (1.3) achieves[2]

$$\left\|\widehat{\boldsymbol{L}} - \boldsymbol{L}^\star\right\|_{\mathrm{F}} = O\left(n\left\|\boldsymbol{E}\right\|_{\mathrm{F}}\right) = O(\sigma n^2) \tag{1.4}$$

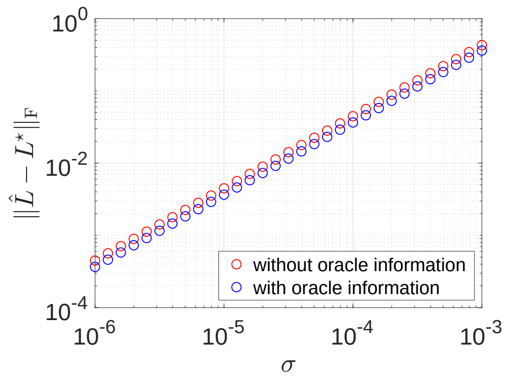
with high probability, where we have substituted in the well-known high-probability bound $\|\boldsymbol{E}\|_{\mathrm{F}} = O(\sigma n)$ under i.i.d. Gaussian noise. While this theory corroborates the potential stability of convex relaxation against both additive noise and sparse outliers, it remains unclear whether the estimation error bound (1.4) reflects the true performance of the convex program in use. In what follows, we shall compare it with an oracle error bound and collect some numerical evidence.

[1] Clearly, if the low-rank matrix $\boldsymbol{L}^\star$ is also sparse, one cannot possibly separate $\boldsymbol{S}^\star$ from $\boldsymbol{L}^\star$. The same holds true if the matrix $\boldsymbol{S}^\star$ is simultaneously sparse and low-rank.

[2] Mathematically, [ZLW+10] investigated an equivalent constrained form of (1.3) and developed an upper bound on the corresponding estimation error.



(a)



(b)

Figure 1: (a) Euclidean estimation errors of (1.3) and (1.5) vs. the problem size $\sqrt{n}$, where we fix $r = 5, \sigma = 10^{-3}$; (b) Euclidean estimation errors of (1.3) and (1.5) vs. the noise level σ in a log-log plot, where we fix $n = 1000, r = 5$. For both plots, we take $\lambda = 5\sigma\sqrt{n}$ and $\tau = 2\sigma\sqrt{\log n}$. The results are averaged over 50 independent trials.

Comparisons with an oracle bound Consider an idealistic scenario where an oracle informs us of the outlier matrix $\boldsymbol{S}^\star$. With the assistance of this oracle, the task of estimating $\boldsymbol{L}^\star$ reduces to a low-rank matrix denoising problem [DG14]. By fixing $\boldsymbol{S}$ to be $\boldsymbol{S}^\star$ in (1.3), we arrive at a simplified convex program

$$\underset{\boldsymbol{L}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \frac{1}{2}\left\|\boldsymbol{L} - (\boldsymbol{L}^\star + \boldsymbol{E})\right\|_{\mathrm{F}}^2 + \lambda \left\|\boldsymbol{L}\right\|_* . \tag{1.5}$$

It is known that (e.g. [DG14, CCF+20]): under mild conditions and with a properly chosen λ, the estimation error of (1.5) satisfies

$$\left\|\widehat{\boldsymbol{L}} - \boldsymbol{L}^\star\right\|_{\mathrm{F}} = O\left(\sigma\sqrt{nr}\right), \tag{1.6}$$

where we abuse the notation and denote by $\widehat{\boldsymbol{L}}$ the minimizer of (1.5). The large gap between the above two bounds (1.4) and (1.6) is self-evident; in particular, if $r = O(1)$, the gap between these two bounds can be as large as an order of $n^{1.5}$.

A numerical example without oracles One might naturally wonder whether the discrepancy between the two bounds (1.4) and (1.6) stems from the magical oracle information (i.e. $\boldsymbol{S}^\star$) which (1.3) does not have the luxury to know. To demonstrate that this is not the case, we conduct some numerical experiments to assess the importance of such oracle information. Generate $\boldsymbol{L}^\star = \boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}$, where $\boldsymbol{X}^\star, \boldsymbol{Y}^\star \in \mathbb{R}^{n\times r}$ are random orthonormal matrices. Each entry of $\boldsymbol{S}^\star$ is generated independently from a mixed distribution: with probability $1/10$, the entry is drawn from $\mathcal{N}(0, 10)$; otherwise, it is set to be zero. In other words, approximately 10% of the entries in $\boldsymbol{L}^\star$ are corrupted by large outliers. Throughout the experiments, we set $\lambda = 5\sigma\sqrt{n}$ and $\tau = 2\sigma\sqrt{\log n}$ with σ the standard deviation of each noise entry $\{E_{ij}\}$. Figure 1(a) fixes $r = 5, \sigma = 10^{-3}$ and examines the dependency of the Euclidean error $\|\widehat{\boldsymbol{L}} - \boldsymbol{L}^\star\|_{\mathrm{F}}$ on the size $\sqrt{n}$. Similarly, Figure 1(b) fixes $r = 5, n = 1000$ and displays the estimation error $\|\widehat{\boldsymbol{L}} - \boldsymbol{L}^\star\|_{\mathrm{F}}$ as the noise size σ varies in a log-log plot. As can be seen from Figure 1, the performance of the oracle-aided estimator (1.5) matches the theoretical prediction (1.6), namely, the numerical estimation error $\|\widehat{\boldsymbol{L}} - \boldsymbol{L}^\star\|_{\mathrm{F}}$ is proportional to both $\sqrt{n}$ and σ. Perhaps more intriguingly, even without the help of the oracle, the original estimator (1.3) performs quite well and behaves qualitatively similarly. In comparison with the bound (1.4) derived in the prior work [ZLW+10], our numerical experiments suggest that the convex estimator (1.3) might perform much better than previously predicted.

All in all, there seems to be a large gap between the practical performance of (1.3) and the existing theoretical support. This calls for a new theory that better explains practice, which we pursue in the current paper. We remark in passing that statistical guarantees have been developed in [ANW12, KLT17] for other convex estimators (i.e. the ones that are different from the convex estimator (1.3) considered herein). We shall compare our results with theirs later in Section 1.4.

1.3 Models, assumptions and notation

As it turns out, the appealing empirical performance of the convex program (1.3) in the presence of both sparse outliers and zero-mean random noise can be justified in theory. Towards this end, we need to introduce several notations and model assumptions that will be used throughout. Let $\boldsymbol{U}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}$ be the singular value decomposition (SVD) of the unknown rank-r matrix $\boldsymbol{L}^\star \in \mathbb{R}^{n_1\times n_2}$, where $\boldsymbol{U}^\star \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{V}^\star \in \mathbb{R}^{n_2\times r}$ consist of orthonormal columns and $\boldsymbol{\Sigma}^\star = \mathsf{diag}\{\sigma_1^\star,\dots,\sigma_r^\star\}$ is a diagonal matrix. Here, we let

$$\sigma_{\max} \coloneqq \sigma_1^\star \ge \sigma_2^\star \ge \cdots \ge \sigma_r^\star \eqqcolon \sigma_{\min} \qquad \text{and} \qquad \kappa \coloneqq \sigma_{\max}/\sigma_{\min}$$

represent the singular values and the condition number of $\boldsymbol{L}^\star$, respectively. We denote by $\Omega^\star$ the support set of $\boldsymbol{S}^\star$, that is,

$$\Omega^\star \coloneqq \{(i,j)\in\Omega_{\mathsf{obs}} : S_{ij}^\star \neq 0\}. \tag{1.7}$$

With this set of notation in place, we list below our key model assumptions.

Assumption 1 (**Incoherence**). *The low-rank matrix $\boldsymbol{L}^\star$ with SVD $\boldsymbol{L}^\star = \boldsymbol{U}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}$ is assumed to be μ-incoherent in the sense that*

$$\|\boldsymbol{U}^\star\|_{2,\infty} \le \sqrt{\frac{\mu}{n_1}}\|\boldsymbol{U}^\star\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{n_1}} \quad \textit{and} \quad \|\boldsymbol{V}^\star\|_{2,\infty} \le \sqrt{\frac{\mu}{n_2}}\|\boldsymbol{V}^\star\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{n_2}}. \tag{1.8}$$

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

Assumption 2 (**Random sampling**). *Each entry is observed independently with probability p, namely,*

$$\mathbb{P}\{(i,j)\in\Omega_{\mathsf{obs}}\} = p. \tag{1.9}$$

Assumption 3 (**Random locations of outliers**). *Each observed entry is independently corrupted by an outlier with probability ρ_{s}, namely,*

$$\mathbb{P}\{(i,j)\in\Omega^\star \mid (i,j)\in\Omega_{\mathsf{obs}}\} = \rho_{\mathsf{s}}, \tag{1.10}$$

where $\Omega^\star \subseteq \Omega_{\mathsf{obs}}$ is the support of the outlier matrix $\boldsymbol{S}^\star$.

Assumption 4 (**Random signs of outliers**). *The signs of the nonzero entries of $\boldsymbol{S}^\star$ are i.i.d. symmetric Bernoulli random variables (independent from the locations), namely,*

$$\mathsf{sign}(S_{ij}^\star) \overset{\text{ind.}}{=} \begin{cases} 1, & \textit{with probability } 1/2, \\ -1, & \textit{else}, \end{cases} \qquad \textit{for all } (i,j)\in\Omega^\star. \tag{1.11}$$

Assumption 5 (**Random noise**). *The noise matrix $\boldsymbol{E} = [E_{ij}]$ is composed of independent symmetric*[3] *zero-mean sub-Gaussian random variables with sub-Gaussian norm at most $\sigma > 0$, i.e. $\|E_{ij}\|_{\psi_2} \le \sigma$ (see [Ver12, Definition 5.7] for precise definitions).*

We take a moment to expand on our model assumptions. Assumption 1 is standard in the low-rank matrix recovery literature [CR09, CLMW11, Che15, CLC19]. If μ is small, then this assumption specifies that the singular spaces of $\boldsymbol{L}^\star$ is not sparse in the standard basis, thus ensuring that $\boldsymbol{L}^\star$ is not simultaneously low-rank and sparse. Assumption 3 requires the sparsity pattern of the outliers $\boldsymbol{S}^\star$ to be random, which precludes it from being simultaneously sparse and low-rank. In essence, Assumptions 1 and 3 are identifiability conditions, taken together as a sort of separation condition on $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$, which plays a crucial role in guaranteeing exact recovery in the noiseless case (i.e. $\boldsymbol{E} = \boldsymbol{0}$); see [CLMW11] for more discussions on these conditions.

[3]In fact, we only require E_{ij} to be symmetric for all $(i,j)\in\Omega^\star$.

Assumption 4 requires the signs of the outliers to be random, which has also been made in [ZLW+10, WL17].[4] We shall discuss in detail the crucial role of this random sign assumption (as opposed to deterministic sign patterns) in Section 1.6.

Finally, we introduce several notation to be used throughout. Denote by $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ the condition $|f(n)| \le Cg(n)$ for some constant $C > 0$ when n is sufficiently large; we use $f(n) \gtrsim g(n)$ to denote $f(n) \ge C|g(n)|$ for some constant $C > 0$ when n is sufficiently large; we also use $f(n) \asymp g(n)$ to indicate that $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$ hold simultaneously. The notation $f(n) \gg g(n)$ (resp. $f(n) \ll g(n)$) means that there exists a sufficiently large (resp. small) constant $c_1 > 0$ (resp. $c_2 > 0$) such that $f(n) \ge c_1 g(n)$ (resp. $f(n) \le c_2 g(n)$). For any subspace T, we denote by $\mathcal{P}_T(\boldsymbol{M})$ the Euclidean projection of a matrix $\boldsymbol{M}$ onto the subspace T, and let $\mathcal{P}_{T^\perp}(\boldsymbol{M}) \coloneqq \boldsymbol{M} - \mathcal{P}_T(\boldsymbol{M})$. For any index set Ω, we denote by $\mathcal{P}_\Omega(\boldsymbol{M})$ the Euclidean projection of a matrix $\boldsymbol{M}$ onto the subspace of matrices supported on Ω, and define $\mathcal{P}_{\Omega^c}(\boldsymbol{M}) \coloneqq \boldsymbol{M} - \mathcal{P}_\Omega(\boldsymbol{M})$. For any matrix $\boldsymbol{M}$, we let $\|\boldsymbol{M}\|$, $\|\boldsymbol{M}\|_{\mathrm{F}}$, $\|\boldsymbol{M}\|_*$, $\|\boldsymbol{M}\|_1$ and $\|\boldsymbol{M}\|_\infty$ denote its spectral norm, Frobenius norm, nuclear norm, entrywise ℓ_1 norm, and entrywise ℓ_∞ norm, respectively.

1.4 Main results

Armed with the above model assumptions, we are positioned to present our improved statistical guarantees for convex relaxation (1.3) in the random noise setting. Without loss of generality, assume that

$$n_1 \ge n_2.$$

As we shall elucidate in Section 1.5 and Section 3, our theory is established by exploiting an intriguing and intimate connection between convex relaxation and nonconvex optimization, and hence the title of this paper.

For the sake of simplicity, we shall start by presenting our statistical guarantees when the rank r, the condition number κ and the incoherence parameter μ of $\boldsymbol{L}^\star$ are all bounded by some constants. Despite its simplicity, this setting subsumes as special cases a wide array of fundamentally important applications, including angular and phase synchronization [Sin11] in computational biology, joint shape mapping problem [HG13, CGH14] in computer vision, and so on. All of these problems involve estimating a very well-conditioned matrix $\boldsymbol{L}^\star$ with a small rank.

Theorem 1. *Suppose that Assumptions 1-5 hold, and that $r, \kappa, \mu = O(1)$. Take $\lambda = C_\lambda \sigma \sqrt{n_1 p}$ and $\tau = C_\tau \sigma \sqrt{\log n_2}$ in (1.3) for some large enough constants $C_\lambda, C_\tau > 0$. Assume that*

$$n_1 n_2 p \ge C_{\mathsf{sample}} n_1 \log^6 n_1, \qquad \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n_1}{p}} \le \frac{c_{\mathsf{noise}}}{\sqrt{\log n_1}} \quad \text{and} \quad \rho_s \le \frac{c_{\mathsf{outlier}}}{\log n_1} \tag{1.12}$$

for some sufficiently large constant $C_{\mathsf{sample}} > 0$ and some sufficiently small constants $c_{\mathsf{noise}}, c_{\mathsf{outlier}} > 0$. Then with probability exceeding $1 - O(n_2^{-3})$, the following holds:

1. *Any minimizer $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ of the convex program (1.3) obeys*

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \le C_{\mathsf{err}} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n_1}{p}} \|\boldsymbol{L}^\star\|_{\mathrm{F}} \tag{1.13a}$$

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_\infty \le C_{\mathsf{err}} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n_1 \log n_1}{p}} \|\boldsymbol{L}^\star\|_\infty \tag{1.13b}$$

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\| \le C_{\mathsf{err}} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n_1}{p}} \|\boldsymbol{L}^\star\| \tag{1.13c}$$

   for some constant $C_{\mathsf{err}} > 0$.

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

$$\|\boldsymbol{L}_{\mathsf{cvx},r} - \boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}} \le \frac{1}{n_2^5} \cdot \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n_1}{p}} \|\boldsymbol{L}^\star\|_{\mathrm{F}}, \tag{1.14}$$

[4] Note that while the theorems in [ZLW+10, WL17] do not make explicit this random sign assumption, the proofs therein do rely on this assumption to guarantee the existence of certain approximate dual certificates.

and the statistical guarantees (1.13) hold unchanged if $\boldsymbol{L}_{\mathsf{cvx}}$ is replaced by $\boldsymbol{L}_{\mathsf{cvx},r}$.

Before we embark on interpreting our statistical guarantees, let us first parse the required conditions (1.12) in Theorem 1. For simplicity we assume that $n_1 = n_2 = n$.

- *Missing data.* Theorem 1 accommodates the case where a dominant fraction of entries are unobserved (more precisely, the sample size can be as low as an order of $n \operatorname{poly}\log n$). This is an appealing result since, even when there is no noise and no outlier (i.e. $\boldsymbol{E} = \boldsymbol{0}$ and $\rho_s = 0$), the minimal sample size required for exact matrix completion is at least on the order of $n \log n$ [CT10]. In comparison, prior theory on robust PCA with both sparse outliers and dense additive noise is either based on full observations [ZLW+10, ANW12], or assumes the sampling rate p exceeds some universal constant [WL17]. In other words, these prior results require the number of observed entries to exceed the order of n^2. The only exception is [KLT17], which also allows a significant amount of missing data (i.e. $p \gtrsim (\operatorname{poly}\log n)/n$).
- *Noise levels.* The noise condition, namely $\sigma\sqrt{n \log n / p} \lesssim \sigma_{\min}$, accommodates a wide range of noise levels. To see this, it is straightforward to check that this noise condition is equivalent to

$$\sigma \lesssim \sqrt{\frac{np}{\log n}}\, \|\boldsymbol{L}^\star\|_\infty$$

as long as $r, \mu, \kappa \asymp 1$. In other words, the entrywise noise level σ is allowed to be significantly larger than the maximum magnitude of the entries in the low-rank matrix $\boldsymbol{L}^\star$, as long as $p \gg (\log n)/n$.
- *Tolerable fraction of outliers.* The above theorem assumes that no more than a fraction $\rho_s \lesssim 1/\log n$ of observations are corrupted by outliers. In words, our theory allows *nearly* a constant proportion (up to a logarithmic order) of the entries of $\boldsymbol{L}^\star$ to be corrupted with arbitrary magnitudes.

Next, we move on to the interpretation of our statistical guarantees. Note that we still assume that $n_1 = n_2 = n$ for ease of presentation.

- *Near-optimal statistical guarantees.* Our first result (1.13a) gives an Euclidean estimation error bound of (1.3)

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}. \tag{1.15}$$

This cannot be improved even when an oracle has informed us of the outliers $\boldsymbol{S}^\star$ and the tangent space of $\boldsymbol{L}^\star$; see [CP10, Section III.B]. We remark that under similar model assumptions, the paper [WL17] derived an estimation error bound for a constrained version of the convex program (1.3), which asserts that this convex estimator $\widetilde{\boldsymbol{L}}_{\mathsf{cvx}}$ satisfies [5]

$$\left\|\widetilde{\boldsymbol{L}}_{\mathsf{cvx}} - \boldsymbol{L}^\star\right\|_{\mathrm{F}} \lesssim \sigma n^{1.5}, \tag{1.16}$$

with the proviso that p is at least on the constant order. The restriction on p arises from the dual certificate constructed in [CLMW11], which is also used in the Proof of Theorem 4 in [WL17]. While this is sub-optimal compared to our results in the setting considered herein, it is worth pointing out that the bound therein accommodates arbitrary noise matrix $\boldsymbol{E}$ (e.g. deterministic, adversary), and here in (1.16) we specialize their result to the random noise setting, namely the noise $\boldsymbol{E}$ obeys Assumption 5. In addition, under the full observation (i.e. $p = 1$) setting, the paper [ANW12] derived an estimation error bound for a convex program similar to (1.3), but with an additional constraint regularizing the spikiness of the low-rank component. Note that instead of imposing the incoherence condition as in Assumption 1, the prior work [ANW12] assumes a milder spikiness condition on $\boldsymbol{L}^\star$, which only constrains the maximum entry in the matrix $\boldsymbol{L}^\star$ is not too large. When $\{E_{ij}\}$ are i.i.d. drawn from $\mathcal{N}(0, \sigma^2)$ and when there is

[5] More specifically, [WL17, Theorem 4] studies the following convex program $\operatorname{minimize}_{\boldsymbol{L},\boldsymbol{S}\in\mathbb{R}^{n\times n}} \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1$ s.t. $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}+\boldsymbol{S}-\boldsymbol{M})\|_{\mathrm{F}} \le \delta$. Here, the quantity δ needs to be larger than $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}+\boldsymbol{S}-\boldsymbol{M})\|_{\mathrm{F}}$. Under our setting, the minimum level of δ should be a high-probability upper bound on $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}$, which is on the order of $\sigma n\sqrt{p}$. With this choice of δ, [WL17, Theorem 4] yields $\|\widetilde{\boldsymbol{L}}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \le [2 + 8\sqrt{n}(1+\sqrt{8/p})]\delta \lesssim \sigma n^{1.5}$.

Table 1: Comparison of our statistical guarantee and prior theory.

	Euclidean estimation error	Accounting for missing data
[ZLW+10]	σn^2	no
[ANW12]	$\sigma\sqrt{n}\max\{\sqrt{r},\sqrt{n\rho_{\mathsf{s}}\log n}\}+\|\boldsymbol{L}^\star\|_\infty n\sqrt{\rho_{\mathsf{s}}}$	no
[WL17]	$\sigma n^{1.5}$	yes ($p\gtrsim 1$)
[KLT17]	$\max\{\sigma,\|\boldsymbol{L}^\star\|_\infty,\|\boldsymbol{S}^\star\|_\infty\}\sqrt{(n\log n)/p}\max\{1,\sqrt{np\rho_{\mathsf{s}}}\}$	yes ($p\gtrsim(\mathrm{poly}\log n)/n$)
This paper	$\sigma\sqrt{nr/p}$	yes ($p\gtrsim\kappa^4\mu^2r^2(\mathrm{poly}\log n)/n$)

no missing data (i.e. $p=1$), the Euclidean estimation error bound achievable by their estimator $\boldsymbol{L}_{\mathsf{cvx}}^{\mathsf{ANW}}$ reads

$$\left\|\boldsymbol{L}_{\mathsf{cvx}}^{\mathsf{ANW}}-\boldsymbol{L}^\star\right\|_{\mathrm{F}}\lesssim\sigma\sqrt{n}\max\left\{1,\sqrt{n\rho_{\mathsf{s}}\log n}\right\}+\|\boldsymbol{L}^\star\|_\infty n\sqrt{\rho_{\mathsf{s}}},\tag{1.17}$$

which is sub-optimal compared to our results. In particular, (i) the bound (1.17) does not vanish even as the noise level decreases to zero, and (ii) it becomes looser as ρ_{s} grows (e.g. if $\rho_{\mathsf{s}}\asymp 1/\log n$, the bound (1.17) is $O(\sqrt{n})$ larger than our bound). Moreover, the work [ANW12] did not account for missing data. Similar to [ANW12] (but with an additional spikiness condition on $\boldsymbol{S}^\star$), the paper [KLT17] derived an estimation error bound for a constrained convex program, with a new constraint regularizing the spikiness of the sparse outliers. Their Euclidean estimation error bound reads

$$\left\|\boldsymbol{L}_{\mathsf{cvx}}^{\mathsf{KLT}}-\boldsymbol{L}^\star\right\|_{\mathrm{F}}\lesssim\max\left\{\sigma,\|\boldsymbol{L}^\star\|_\infty,\|\boldsymbol{S}^\star\|_\infty\right\}\sqrt{\frac{n\log n}{p}}\max\left\{1,\sqrt{np\rho_{\mathsf{s}}}\right\},\tag{1.18}$$

which is also sub-optimal compared to our results. In particular, (1) their error bound degrades as the magnitude $\|\boldsymbol{S}^\star\|_\infty$ of the outlier increases; (2) when there is no missing data (i.e. $p=1$), their bound might be off by a factor as large as $O(\sqrt{n})$. It is worth emphasizing that the theory developed in these prior works is developed to accommodate a broader range of matrices. For example, both [ANW12] and [KLT17] study the set of entrywise bounded low-rank matrices (without assuming the incoherence condition); [ANW12] even allows $\boldsymbol{L}^\star$ to be approximately low rank. To ease comparison, Table 1 displays a summary of our results vs. prior statistical guarantees when specialized to the settings considered herein.

- *Entrywise and spectral norm error control.* Moving beyond Euclidean estimation errors, our theory also provides statistical guarantees measured by two other important metrics: the entrywise ℓ_∞ norm (cf. (1.13b)) and the spectral norm (cf. (1.13c)). In particular, our entrywise error bound (1.13b) in reads

$$\|\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^\star\|_\infty\lesssim\sigma\sqrt{\frac{\log n}{np}}\tag{1.19}$$

as long as $r,\kappa,\mu\asymp 1$, which is about $O(n)$ times small than the Euclidean loss (1.15) modulo some logarithmic factor. This uncovers an appealing "delocalization" behavior of the estimation errors, namely, the estimation errors of $\boldsymbol{L}^\star$ are fairly spread out across all entries. This can also be viewed as an "implicit regularization" phenomenon: the convex program automatically controls the spikiness of the low-rank solution, without the need of explicitly regularizing it (e.g. adding a constraint $\|\boldsymbol{L}\|_\infty\leq\alpha$ as adopted in the prior work [ANW12, KLT17]). See Figure 2 for the numerical evidence for the relative entrywise and spectral norm error of $\boldsymbol{L}_{\mathsf{cvx}}$.

- *Approximate low-rank structure of the convex estimator $\boldsymbol{L}_{\mathsf{cvx}}$.* Last but not least, Theorem 1 ensures that the convex estimate $\boldsymbol{L}_{\mathsf{cvx}}$ is nearly rank-r, so that a rank-r approximation of $\boldsymbol{L}_{\mathsf{cvx}}$ is extremely accurate. In other words, the convex program automatically adapts to the true rank of $\boldsymbol{L}^\star$ without having any prior knowledge about r. As we shall see shortly, this is a crucial observation underlying the intimate connection between convex relaxation and a certain nonconvex approach.

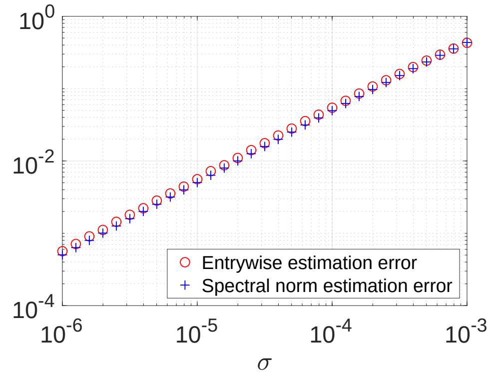


Figure 2: The relative estimation error of $\boldsymbol{L}_{\mathsf{cvx}}$ measured by both $\|\cdot\|_\infty$ (i.e. $\|\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^\star\|_\infty/\|\boldsymbol{L}^\star\|_\infty$) and $\|\cdot\|$ (i.e. $\|\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^\star\|/\|\boldsymbol{L}^\star\|$) vs. the standard deviation σ of the noise in a log-log plot. The results are reported for $n=1000$, $r=5$, $p=0.2$, $\rho_s=0.1$, $\lambda=5\sigma\sqrt{np}$, $\tau=2\sigma\sqrt{\log n}$, and are averaged over 50 independent trials. In addition, the data generating process is similar to that in Figure 1.

Moving beyond the setting with $r,\kappa,\mu\asymp 1$, we have developed theoretical guarantees that allow r,κ,μ to grow with the problem dimension n_1,n_2. The result is this.

Theorem 2. *Suppose that Assumptions 1-5 hold and that $n_1\geq n_2$. Take $\lambda=C_\lambda\sigma\sqrt{n_1p}$ and $\tau=C_\tau\sigma\sqrt{\log n_2}$ in (1.3) for some large enough constants $C_\lambda,C_\tau>0$. Assume that*

$$n_1n_2p\geq C_{\mathsf{sample}}\kappa^4\mu^2r^2n_1\log^6 n_1,\qquad \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n_1}{p}}\leq\frac{c_{\mathsf{noise}}}{\sqrt{\kappa^4\mu r\log n_1}},\quad \text{and}\quad \rho_s\leq\frac{c_{\mathsf{outlier}}}{\kappa^3\mu r\log n_1} \tag{1.20}$$

for some sufficiently large constant $C_{\mathsf{sample}}>0$ and some sufficiently small constants $c_{\mathsf{noise}},c_{\mathsf{outlier}}>0$. Then with probability exceeding $1-O(n_2^{-3})$, the following holds:

1. *Any minimizer $(\boldsymbol{L}_{\mathsf{cvx}},\boldsymbol{S}_{\mathsf{cvx}})$ of the convex program (1.3) obeys*

$$\|\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^\star\|_{\mathrm{F}}\leq C_{\mathsf{err}}\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n_1}{p}}\,\|\boldsymbol{L}^\star\|_{\mathrm{F}} \tag{1.21a}$$

$$\|\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^\star\|_\infty\leq C_{\mathsf{err}}\sqrt{\kappa^3\mu r}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n_1\log n_1}{p}}\,\|\boldsymbol{L}^\star\|_\infty \tag{1.21b}$$

$$\|\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^\star\|\leq C_{\mathsf{err}}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n_1}{p}}\,\|\boldsymbol{L}^\star\| \tag{1.21c}$$

   for some constant $C_{\mathsf{err}}>0$.

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

$$\|\boldsymbol{L}_{\mathsf{cvx},r}-\boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}}\leq\frac{1}{n_2^5}\cdot\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n_1}{p}}\,\|\boldsymbol{L}^\star\|_{\mathrm{F}}\,, \tag{1.22}$$

   and the statistical guarantees (1.21) hold unchanged if $\boldsymbol{L}_{\mathsf{cvx}}$ is replaced by $\boldsymbol{L}_{\mathsf{cvx},r}$.

Similar to Theorem 1, our general theory (i.e. Theorem 2) provides the estimation error of the convex estimator $\boldsymbol{L}_{\mathsf{cvx}}$ in three different norms (i.e. the Euclidean, entrywise and operator norms), and reveals the near low-rankness of the convex estimator (cf. (1.22)) as well as the implicit regularization phenomenon (cf. (1.21b)).

Finally, we make note of several aspects of our general theory that call for further improvement. For instance, when there is no missing data and $n_1=n_2=n$, the rank r of the unknown matrix $\boldsymbol{L}^\star$ needs to satisfy $r\lesssim\sqrt{n}$. On the positive side, our result allows r to grow with the problem dimension n. However,

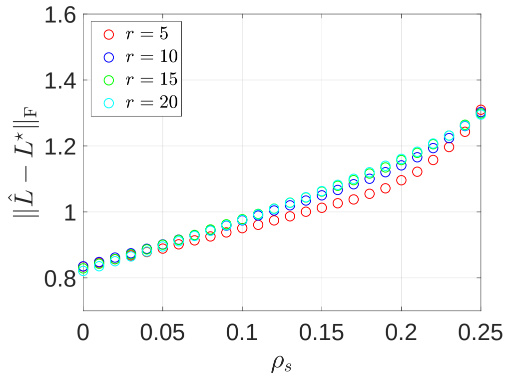


Figure 3: Euclidean estimation errors of (1.3) vs. ρ_{s} under four different ranks $r = 5, 10, 15, 20$. The results are reported for $n = 1000$, $p = 0.04r$, $\sigma = 10^{-3}$, $\lambda = 5\sigma\sqrt{np}$, $\tau = 2\sigma\sqrt{\log n}$, and are averaged over 50 independent trials. In addition, the data generating process is similar to that in Figure 1.

prior results in the noiseless case [CLMW11, Li13] allow r to grow almost linearly with n. This unsatisfactory aspect arises from the suboptimal analysis (in terms of the dependency on r) of a tightly related nonconvex estimation algorithm (to be elaborated on later), which, to the best of our knowledge, has not been resolved in the nonconvex low-rank matrix recovery literature [MWCC20, CLL20]. See Section 2 for more discussions about this point. Moreover, when $\boldsymbol{E} = \boldsymbol{0}$, it is known that ρ_s can be as large as a constant even when the rank r is allowed to grow with the dimension n [Li13, CJSC13]. Our current theory, however, fails to cover the case with $\rho_s \asymp 1$ in the presence of noise. We demonstrate through numerical experiments that the dependence of ρ_{s} on r might indeed by suboptimal in our current theory. More specifically, Figure 3 depicts the numerical Euclidean estimation errors w.r.t. the corruption probability ρ_{s} as we vary the rank while fixing the sampling ratio. It can be seen that the estimation error curves corresponding to different ranks align very well with each other, thus suggesting the capability of convex relaxation in tolerating a constant fraction ρ_{s} of outliers.

1.5 A peek at our technical approach

Before delving into the proof details, we immediately highlight our key technical ideas and novelties. For simplicity we assume $n_1 = n_2 = n$ throughout this section.

Connections between convex and nonconvex optimization. Instead of directly analyzing the convex program (1.3), we turn attention to a seemingly different, but in fact closely related, nonconvex program

$$\operatorname*{minimize}_{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r},\boldsymbol{S}\in\mathbb{R}^{n\times n}} \quad \frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{S} - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\left(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2\right) + \tau\left\|\boldsymbol{S}\right\|_1. \tag{1.23}$$

This idea is inspired by an interesting numerical finding (cf. Figure 4) that the solution to the convex program (1.3), and an estimate obtained by attempting to solve the nonconvex formulation (1.23), are exceedingly close in our experiments. If such an intimate connection can be formalized, then it suffices to analyze the statistical performance of the nonconvex approach instead.[6] Fortunately, recent advances in nonconvex low-rank factorization (see [CLC19] for an overview) provide powerful tools for analyzing nonconvex low-rank estimation, allowing us to derive the desired statistical guarantees that can then be transferred to the convex approach. Of course, this is merely a high-level picture of our proof strategy, and we defer the details to Section 3.

It is worth emphasizing that our key idea — that is, bridging convex and nonconvex optimization — is drastically different from previous technical approaches for analyzing convex estimators (e.g. (1.3)). As it

[6]On the surface, the convex program (1.3) and the nonconvex one (1.23) are closely related: the convex solution $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ coincides with that of the nonconvex program (1.23) if $\boldsymbol{L}_{\mathsf{cvx}}$ is rank-r. This is an immediate consequence of the algebraic identity $\|\boldsymbol{Z}\|_* = \inf_{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}:\boldsymbol{X}\boldsymbol{Y}^\top=\boldsymbol{Z}}(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2)$ [SS05, MHT10]. However, it is difficult to know *a priori* the rank of the convex solution. Hence such a connection does not prove useful in establishing the statistical properties of the convex estimator.

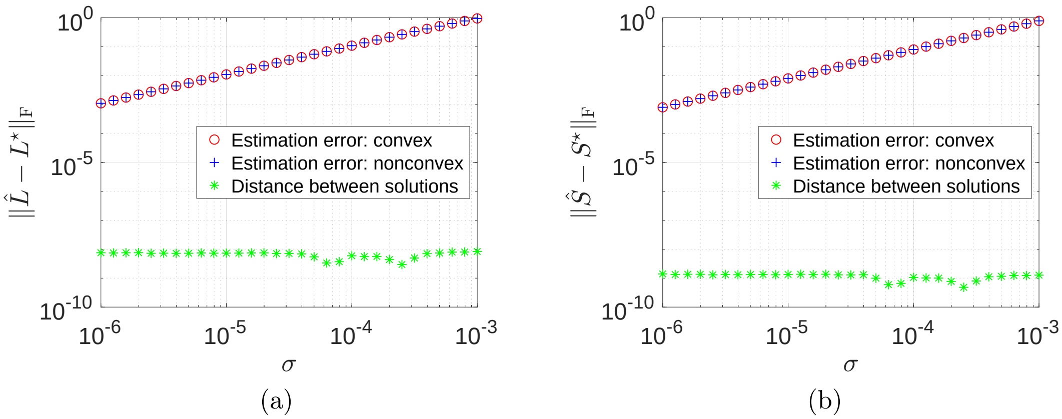


Figure 4: (a) The relative estimation errors of both $\boldsymbol{L}_{\mathsf{cvx}}$ (the convex estimator (1.3)) and $\boldsymbol{L}_{\mathsf{ncvx}}$ (the estimate returned by the nonconvex approach tailored to (1.23)) and the relative distance between them vs. the standard deviation σ of the noise. (b) The relative estimation errors of both $\boldsymbol{S}_{\mathsf{cvx}}$ (the convex estimator in (1.3)) and $\boldsymbol{S}_{\mathsf{ncvx}}$ (the estimate returned by the nonconvex approach tailored to (1.23)) and the relative distance between them vs. the standard deviation σ of the noise. The results are reported for $n = 1000$, $r = 5$, $p = 0.2$, $\rho_s = 0.1$, $\lambda = 5\sigma\sqrt{np}$, $\tau = 2\sigma\sqrt{\log n}$ and are averaged over 50 independent trials.

turns out, these prior approaches, which include constructing dual certificates and/or exploiting restricted strong convexity, have their own deficiencies in analyzing (1.3) and fall short of explaining the effectiveness of (1.3) in the random noise setting. For instance, constructing dual certificates in the noisy case is notoriously challenging given that we do not have closed-form expressions for the primal solutions (so that it is difficult to invoke the powerful dual construction strategies like the golfing scheme [Gro11] developed for the noiseless case). If we directly utilize the dual certificates constructed for the noiseless case, we would end up with an overly conservative bound like (1.4), which is exactly why the results in [ZLW+10, WL17] are sub-optimal. On the other hand, while it is viable to show certain strong convexity of (1.3) when restricted to some highly local sets and directions, it is unclear how (1.3) forces its solution to stay within the desired set and follow the desired directions, without adding further (and often unnecessary) constraints to (1.3).

Nonconvex low-rank estimation with nonsmooth loss functions. It is worth noting that a similar connection between convex and nonconvex optimization has been pointed out by [CCF+20] towards understanding the power of convex relaxation for noisy matrix completion. Due to the absence of sparse outliers in the noisy matrix completion problem, the nonconvex loss function considered therein is smooth in nature, which greatly simplifies both the algorithmic and theoretical development. By contrast, the nonsmoothness inherent in (1.23) makes it particularly challenging to achieve the two desiderata mentioned above, namely, connecting the convex and nonconvex solutions and establishing the optimality of the nonconvex solution. In fact, to establish the connection between convex and nonconvex solutions, we put forward a novel two-step analysis strategy. Specifically, we first develop a crude upper bound on the Euclidean estimation error leveraging the idea of approximate dual certificates; see Theorem 3. While this crude upper bound is far from optimal, it serves as an important starting point towards formalizing the intimate relation between the convex solution $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ and the nonconvex solution $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S})$, since it is challenging to establish $\boldsymbol{X}\boldsymbol{Y} \approx \boldsymbol{L}_{\mathsf{cvx}}$ and $\boldsymbol{S} \approx \boldsymbol{S}_{\mathsf{cvx}}$ simultaneously without the aid of a crude bound. Second, in establishing the optimality of the nonconvex solution, the nonsmoothness nature of the nonconvex loss prevents us from applying the vanilla gradient descent scheme (as has been done in [CCF+20]). To address this issue, we develop an alternating minimization scheme — which alternates between gradient updates on $(\boldsymbol{X}, \boldsymbol{Y})$ and minimization of $\boldsymbol{S}$ — aimed at minimizing the nonsmooth nonconvex loss function (1.23); see Algorithm 1 for details. As it turns out, such a simple algorithm allows us to track the proximity of the convex and nonconvex solutions and establish the optimality of the nonconvex solution all at once.

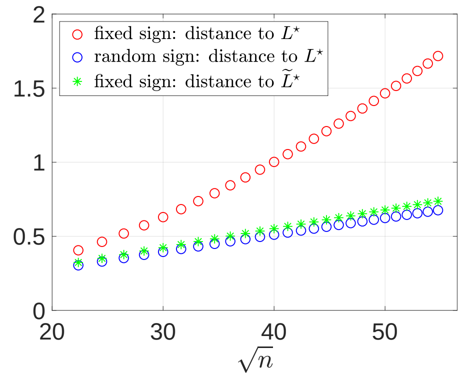


Figure 5: The red (resp. blue) line displays the Euclidean estimation error of (1.3) vs. $\sqrt{n}$ under fixed (resp. random) sign patterns of $\boldsymbol{S}^\star$. The green line displays the Euclidean distance between $\boldsymbol{L}_{\mathsf{cvx}}$ and $\widetilde{\boldsymbol{L}}^\star$ under fixed sign patterns of $\boldsymbol{S}^\star$. The results are reported for $r = 5$, $p = 1$, $\sigma = 10^{-3}$, and $\rho_{\mathsf{s}} = 1/\log n$, with $\lambda = 5\sigma\sqrt{np}$ and $\tau = 2\sigma\sqrt{\log n}$ and are averaged over 50 independent trials. For the random sign setting, the nonzero entries of $\boldsymbol{S}^\star$ are independently generated as $z \cdot 5\sigma$, where z follows a Rademacher distribution. For the fixed sign setting, each nonzero entry of $\boldsymbol{S}^\star$ equals to 5σ.

1.6 Random signs of outliers

The careful reader might wonder whether it is possible to remove the random sign assumption on $\boldsymbol{S}^\star$ (namely, Assumption 4) without compromising our statistical guarantees. After all, the results of [CSPW11, CLMW11, Li13] derived for the noise-free case do not rely on such a random sign assumption at all.[7] Unfortunately, removal of such a condition might be problematic in general, as illustrated by the following example.

An example with non-random signs Suppose that (i) $n_1 = n_2 = n$, (ii) each non-zero entry of $\boldsymbol{S}^\star$ obeys $S_{ij}^\star = c_0\sigma$, (iii) $\rho_{\mathsf{s}} = c_1/\log n$ for some sufficiently small constant $c_1 > 0$, and (iv) there is no missing data (i.e. $p = 1$). In such a scenario, the data matrix can be decomposed as

$$\boldsymbol{M} = \boldsymbol{L}^\star + \boldsymbol{S}^\star + \boldsymbol{E} = \underbrace{\boldsymbol{L}^\star + \mathbb{E}[\boldsymbol{S}^\star]}_{=:\widetilde{\boldsymbol{L}}^\star} + \underbrace{\boldsymbol{S}^\star - \mathbb{E}[\boldsymbol{S}^\star] + \boldsymbol{E}}_{=:\widetilde{\boldsymbol{E}}}.$$

Two observations are worth noting: (1) given that $\mathbb{E}[\boldsymbol{S}^\star] = c_0\rho_{\mathsf{s}}\sigma\boldsymbol{1}\boldsymbol{1}^\top$ with $\boldsymbol{1}$ the all-one vector, the rank of the matrix $\widetilde{\boldsymbol{L}}^\star = \boldsymbol{L}^\star + \mathbb{E}[\boldsymbol{S}^\star]$ is at most $r+1$; (2) $\widetilde{\boldsymbol{E}}$ is a zero-mean random matrix consisting of independent entries with sub-Gaussian norm $O(\sigma)$. In other words, the decomposition $\boldsymbol{M} = \widetilde{\boldsymbol{L}}^\star + \widetilde{\boldsymbol{E}}$ corresponds to a case with random noise but no outliers. Consequently, we can invoke Theorem 1 to conclude that (assuming $r = O(1)$ and $\widetilde{\boldsymbol{L}}^\star$ is incoherent with condition number $O(1)$): any minimizer $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ of (1.3) obeys

$$\begin{aligned}\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star - \rho_{\mathsf{s}}\sigma\boldsymbol{1}\boldsymbol{1}^\top\|_{\mathrm{F}} = \|\boldsymbol{L}_{\mathsf{cvx}} - \widetilde{\boldsymbol{L}}^\star\|_{\mathrm{F}} &\lesssim \frac{\sigma}{\sigma_{\min}(\widetilde{\boldsymbol{L}}^\star)}\sqrt{n}\|\widetilde{\boldsymbol{L}}^\star\|_{\mathrm{F}} \\ &\lesssim \sigma\sqrt{nr}\frac{\sigma_{\max}(\widetilde{\boldsymbol{L}}^\star)}{\sigma_{\min}(\widetilde{\boldsymbol{L}}^\star)} \lesssim \sigma\sqrt{n}\end{aligned}$$

with high probability. Here the last step follows since $\widetilde{\boldsymbol{L}}^\star$ is of constant rank and condition number. This, however, leads to a lower bound on the estimation error

$$\begin{aligned}\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \geq \|c_0\rho_{\mathsf{s}}\sigma\boldsymbol{1}\boldsymbol{1}^\top\|_{\mathrm{F}} - \|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star - \rho_{\mathsf{s}}\sigma\boldsymbol{1}\boldsymbol{1}^\top\|_{\mathrm{F}} &= \sigma\big(c_0\rho_{\mathsf{s}}n - O(\sqrt{n})\big) \\ &= (1 - o(1))\frac{c_0c_1\sigma n}{\log n},\end{aligned}$$

[7] Notably, in the noisy setting, prior theory [ZLW+10, WL17] also implicitly assumes this random sign condition, while [ANW12, KLT17] do not require this condition.

which can be $O(\sqrt{n}/\log n)$ times larger than the desired estimation error $O(\sigma\sqrt{n})$. Numerical experiments under the above setting (with $c_0 = 5$ and $c_1 = 1$) also suggest that (i) the estimation error under the fixed sign setting might be orderwise larger than that under the random sign setting; and (ii) under the fixed sign setting, the estimator (1.3) approximately recovers $\widetilde{\boldsymbol{L}}^\star$ instead of $\boldsymbol{L}^\star$; see Figure 5.

The take-away message is this: when the entries of $\boldsymbol{S}^\star$ are of non-random signs, it might sometimes be possible to decompose $\boldsymbol{S}^\star$ into (1) a low-rank bias component with a large Euclidean norm, and (2) a random fluctuation component whose typical size does not exceed that of $\boldsymbol{E}$. If this is the case, then the convex program (1.3) might mistakenly treat the bias component as a part of the low-rank matrix $\boldsymbol{L}^\star$, thus dramatically hampering its estimation accuracy.

2 Prior art

Principal component analysis (PCA) [Pea01, Jol11, FSZZ18] is one of the most widely used statistical methods for dimension reduction in data analysis. However, PCA is known to be quite sensitive to adversarial outliers — even a single corrupted data point can make PCA completely off. This motivated the investigation of robust PCA, which aims at making PCA robust to gross adversarial outliers. As formulated in [CLMW11, CSPW11], this is closely related to the problem of disentangling a low-rank matrix $\boldsymbol{L}^\star$ and a sparse outlier matrix $\boldsymbol{S}^\star$ (with unknown locations and magnitudes) from a superposition of them. Consequently, robust PCA can be viewed as an outlier-robust extension of the low-rank matrix estimation/completion tasks [CR09, KMO10, CLC19]. In a similar vein, robust PCA has also been extensively studied in the context of structured covariance estimation under approximate factor models [FFL08, FLM13, FWZ18, FWZ19], where the population covariance of certain random sample vectors is a mixture of a low-rank matrix and a sparse matrix, corresponding to the factor component and the idiosyncratic component, respectively.

Focusing on the convex relaxation approach, [CSPW11, CLMW11] started by considering the noiseless case with no missing data (i.e. $\boldsymbol{E} = \boldsymbol{0}$ and $p = 1$) and demonstrated that, under mild conditions, convex relaxation succeeds in exactly decomposing both $\boldsymbol{L}^\star$ and $\boldsymbol{S}^\star$ from the data matrix $\boldsymbol{L}^\star + \boldsymbol{S}^\star$. More specifically, [CSPW11] adopted a deterministic model without assuming any probabilistic structure on the outlier matrix $\boldsymbol{S}^\star$. As shown in [CSPW11] and several subsequent work [CJSC13, HKZ11], convex relaxation is guaranteed to work as long as the fraction of outliers in each row/column does not exceed $O(1/r)$. In contrast, [CLMW11] proposed a random model by assuming that $\boldsymbol{S}^\star$ has random support (cf. Assumption 3); under this model, exact recovery is guaranteed even if a constant fraction of the entries of $\boldsymbol{S}^\star$ are nonzero with arbitrary magnitudes. Following the random location model proposed in [CLMW11], the paper [GWL+10] showed that, in the absence of noise, convex programming can provably tolerate a dominant fraction of outliers, provided that the signs of the nonzero entries of $\boldsymbol{S}^\star$ are randomly generated (cf. Assumption 4). Later, the papers [CJSC13, Li13] extended these results to the case when most entries of the matrix are unseen; even in the presence of highly incomplete data, convex relaxation still succeeds when a constant proportion of the observed entries are arbitrarily corrupted. It is worth noting that the results of [CJSC13] accommodated both models proposed in [CSPW11] and [CLMW11], while the results of [Li13] focused on the latter model.

The literature on robust PCA with not only sparse outliers but also dense noise — namely, when the measurements take the form $\boldsymbol{M} = \mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}^\star + \boldsymbol{S}^\star + \boldsymbol{E})$ — is relatively scarce. [ZLW+10, ANW12] were among the first to present a general theory for robust PCA with dense noise, which was further extended in [WL17, KLT17]. As we mentioned before, the first three [ZLW+10, ANW12, WL17] accommodated arbitrary noise with the last one [KLT17] focusing on the random noise. As we have discussed in Section 1.4, the statistical guarantees provided in these papers are highly suboptimal when it comes to the random noise setting considered herein. The paper [CC14] extended the robust PCA results to the case where the truth is not only low-rank but also of Hankel structure. The results therein, however, suffered from the same sub-optimality issue.

Moving beyond convex relaxation methods, another line of work proposed nonconvex approaches for robust PCA [NNS+14, GWL16, YPCC16, CGJ17, ZWG18, CCD+19, LMCC19, CCW19], largely motivated by the recent success of nonconvex methods in low-rank matrix factorization [CLC19, KMO10, CLS15, SL16, CC17, CW15, ZCL16, CC18, JNS13, NJS13, MWCC20, CCFM19, WGE17, WCCL16, CW18, ZL16, CDDD19]. Following the deterministic model of [CSPW11], the paper [NNS+14] proposed an alternating projection / minimization scheme to seek a low-rank and sparse decomposition of the observed data matrix. In the noiseless setting, i.e. $\boldsymbol{E} = \boldsymbol{0}$, this alternating minimization scheme provably disentangles the low-rank and sparse matrix from their superposition under mild conditions. In addition, [NNS+14] extended their result to the arbitrary noise case where the size of the noise is extremely small, namely, $\|\boldsymbol{E}\|_\infty \ll \sigma_{\min}/n$. When the noise $\{E_{ij}\} \sim \mathcal{N}(0, \sigma^2)$, this is equivalent to the condition $\sigma \ll \sigma_{\min}/(n\sqrt{\log n})$. Comparing this with our noise condition $\sigma \ll \sigma_{\min}/(\sqrt{n \log n})$ (cf. (1.12)) when $r, \mu, \kappa \asymp 1$, one sees that our theoretical guarantees cover a wider range of noise levels. Similarly, [YPCC16] applied regularized gradient descent on a smooth nonconvex loss function which enjoys provable convergence guarantees to $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ under the noiseless and partial observation setting. A recent paper [CCD+19] considered the nonsmooth nonconvex formulation for robust PCA and established rigorously the convergence of subgradient-type methods in the rank-1 setting, i.e. $r = 1$. However, the extension to more general rank remains out of reach.

It is worth noting that noisy matrix completion problem [CP10, CCF+20] is subsumed as a special case by the model studied in this paper (namely, it is a special case with $\boldsymbol{S}^\star = \boldsymbol{0}$). Statistical optimality under the random noise setting (cf. Assumption 5) — including the convex relaxation approach [CCF+20, NW12, KLT11, Klo14] and the nonconvex approach [MWCC20, CLL20] — has been extensively studied. Focusing on arbitrary deterministic noise, [CP10] established the stability of the convex approach, whose resulting estimation error bound is similar to the one established for robust PCA with noise in [ZLW+10]) (see (1.4)). The paper [KS20] later confirmed that the estimation error bound established in [CP10] is the best one can hope for in the arbitrary noise setting for matrix completion, although it might be highly suboptimal if we restrict attention to random noise.

Finally, there is also a large literature considering robust PCA under different settings and/or from different perspectives. For instance, the computational efficiency in solving the convex optimization problem (1.3) and its variants has been studied in the optimization literature (e.g. [TY11,GMS13,SWZ14,MA18]). The problem has also been investigated under a streaming / online setting [GQV14, QV10, FXY13, ZLGV16, QVLH14, VN18]. These are beyond the scope of the current paper.

3 Architecture of the proof

In this section, we give an outline for proving Theorem 2. The proof of Theorem 1 follows immediately as it is a special case of Theorem 2. For simplicity of presentation, our proof sets $n_1 = n_2 = n$. It is straightforward to obtain the proof for the general rectangular case via minor modification.

The main ingredient of the proof lies in establishing an intimate link between convex and nonconvex optimization. Unless otherwise noted, we shall set the regularization parameters as

$$\lambda = C_\lambda \sigma \sqrt{np} \qquad \text{and} \qquad \tau = C_\tau \sigma \sqrt{\log n} \tag{3.1}$$

throughout. In addition, the soft thresholding operator at level τ is defined such that

$$\mathcal{S}_\tau(x) := \mathsf{sign}(x) \max(|x| - \tau, 0) \tag{3.2}$$

For any matrix $\boldsymbol{X}$, the matrix $\mathcal{S}_\tau(\boldsymbol{X})$ is obtained by applying the soft thresholding operator $\mathcal{S}_\tau(\cdot)$ to each entry of $\boldsymbol{X}$ separately. Additionally, we define the true low-rank factors as follows

$$\boldsymbol{X}^\star := \boldsymbol{U}^\star (\boldsymbol{\Sigma}^\star)^{1/2} \qquad \text{and} \qquad \boldsymbol{Y}^\star := \boldsymbol{V}^\star (\boldsymbol{\Sigma}^\star)^{1/2}, \tag{3.3}$$

where $\boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$ is the SVD of the true low-rank matrix $\boldsymbol{L}^\star$.

3.1 Crude estimation error bounds for convex relaxation

We start by delivering a crude upper bound on the Euclidean estimation error, built upon the (approximate) duality certificate previously constructed in [CJSC13]. The proof is postponed to Appendix D.

Theorem 3. *Consider any given $\lambda > 0$ and set $\tau \asymp \lambda\sqrt{(\log n)/np}$. Suppose that Assumptions 1-4 hold, and that*

$$n^2 p \geq C\mu^2 r^2 n \log^6 n \qquad \textit{and} \qquad \rho_{\mathsf{s}} \leq c$$

hold for some sufficiently large (resp. small) constant $C > 0$ (resp. $c > 0$). Then with probability at least $1 - O(n^{-10})$, any minimizer $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ of the convex program (1.3) satisfies

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}}^2 + \|\boldsymbol{S}_{\mathsf{cvx}} - \boldsymbol{S}^\star\|_{\mathrm{F}}^2 \lesssim \lambda^2 n^5 \log^3 n + \frac{n}{\lambda^2} \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\right\|_{\mathrm{F}}^4. \tag{3.4}$$

It is worth noting that the above theorem holds true for an arbitrary noise matrix $\boldsymbol{E}$. When specialized to the case with independent sub-Gaussian noise, this crude bound admits a simpler expression as follows.

Corollary 1. *Take $\lambda = C_\lambda \sigma\sqrt{np}$ and $\tau = C_\tau \sigma \sqrt{\log n}$ for some universal constant $C_\lambda, C_\tau > 0$. Under the assumptions of Theorem 3 and Assumption 5, we have — with probability exceeding $1 - O(n^{-10})$ — that*

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_{\mathrm{F}} \lesssim \sigma n^3 \log^{3/2} n \quad \text{and} \quad \|\boldsymbol{S}_{\mathsf{cvx}} - \boldsymbol{S}^\star\|_{\mathrm{F}} \lesssim \sigma n^3 \log^{3/2} n. \tag{3.5}$$

Proof. This corollary follows immediately by combining Theorem 3 and Lemma 1 below. □

Lemma 1. *Suppose that Assumption 5 holds and that $n^2 p > C_1 n \log^2 n$ for some sufficiently large constant $C_1 > 0$. Then with probability exceeding $1 - O(n^{-10})$, one has*

$$\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\right\| \lesssim \sigma\sqrt{np} \qquad \text{and} \qquad \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\right\|_{\mathrm{F}} \lesssim \sigma n \sqrt{p}.$$

While the above results often lose a polynomial factor in n vis-à-vis the optimal error bound, it serves as an important starting point that paves the way for subsequent analytical refinement.

3.2 Approximate stationary points of the nonconvex formulation

Instead of analyzing the convex estimator directly, we take a detour by considering the following nonconvex optimization problem

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r},\, \boldsymbol{S}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad F(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S}) := \underbrace{\frac{1}{2p}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{S} - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{Y}\|_{\mathrm{F}}^2}_{=:\, f(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{S})} + \frac{\tau}{p}\|\boldsymbol{S}\|_1. \tag{3.6}$$

Here, $f(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{S})$ is a function of $\boldsymbol{X}$ and $\boldsymbol{Y}$ with $\boldsymbol{S}$ frozen, which contains the smooth component of the loss function $F(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S})$. As it turns out, the solution to convex relaxation (1.3) is exceedingly close to an estimate $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S})$ obtained by a nonconvex algorithm aimed at solving (3.6) — to be detailed in Section 3.3. This fundamental connection between the two algorithmic paradigms provides a powerful framework that allows us to understand convex relaxation by studying nonconvex optimization.

In what follows, we set out to develop the afforementioned intimate connection. Before proceeding, we first state the following conditions concerned with the interplay between the noise size, the estimation accuracy of the nonconvex estimate $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S})$, and the regularization parameters.

Condition 1. *The regularization parameters λ and $\tau \asymp \lambda\sqrt{(\log n)/np}$ satisfy*

- $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\| < \lambda/16$ *and* $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_\infty \leq \tau/4$;
- $\|\boldsymbol{S} - \boldsymbol{S}^\star\| < \lambda/16$ *and* $\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{L}^\star\|_\infty \leq \tau/4$;
- $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{L}^\star) - p(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{L}^\star)\| < \lambda/8$.

As an interpretation, the above condition says that: (1) the regularization parameters are not too small compared to the size of the noise, so as to ensure that we enforce a sufficiently large degree of regularization; (2) the estimate represented by the point $(\boldsymbol{X}\boldsymbol{Y}^\top, \boldsymbol{S})$ is sufficiently close to the truth. At this point, whether this condition is meaningful or not remains far from clear; we shall return to justify its feasibility shortly.

In addition, we need another condition concerning the injectivity of $\mathcal{P}_{\Omega^\star}$ w.r.t. a certain tangent space. For a rank-r matrix $\boldsymbol{L}$ with singular value decomposition $\boldsymbol{U\Sigma V}^\top$ where $\boldsymbol{U},\boldsymbol{V} \in \mathbb{R}^{n\times r}$, the tangent space of the set of rank-r matrices at the point $\boldsymbol{L}$ is given by

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

Again, the validity of this condition will be discussed momentarily.

Condition 2 (**Injectivity**)**.** *Let T be the tangent space of the set of rank-r matrices at the point $\boldsymbol{X}\boldsymbol{Y}^{\top}$. Assume that there exist a constants $c_{\mathrm{inj}} > 0$ such that for all $\boldsymbol{H} \in T$, one has*

$$p^{-1} \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{H})\right\|_{\mathrm{F}}^{2} \geq \frac{c_{\mathrm{inj}}}{\kappa} \left\|\boldsymbol{H}\right\|_{\mathrm{F}}^{2} \qquad \text{and} \qquad p^{-1} \left\|\mathcal{P}_{\Omega^{\star}}(\boldsymbol{H})\right\|_{\mathrm{F}}^{2} \leq \frac{c_{\mathrm{inj}}}{4\kappa} \left\|\boldsymbol{H}\right\|_{\mathrm{F}}^{2}.$$

With the above conditions in place, we are ready to make precise the intimate link between convex relaxation and a candidate nonconvex solution. The proof is deferred to Appendix E.

Theorem 4. *Suppose that $n \geq \kappa$ and $\rho_{\mathsf{s}} \leq c/\kappa$ for some sufficiently small constant $c > 0$. Assume that there exists a triple $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S})$ such that*

$$\left\|\nabla f\left(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{S}\right)\right\|_{\mathrm{F}} \leq \frac{1}{n^{20}} \frac{\lambda}{p} \sqrt{\sigma_{\min}}, \quad \text{and} \quad \boldsymbol{S} = \mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\mathcal{S}_{\tau}\left(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^{\top}\right)\right). \tag{3.7}$$

Further, assume that any singular value of $\boldsymbol{X}$ and $\boldsymbol{Y}$ lies in $[\sqrt{\sigma_{\min}/2}, \sqrt{2\sigma_{\max}}]$. If the solution $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ to the convex program (1.3) admits the following crude error bound

$$\left\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^{\star}\right\|_{\mathrm{F}} \lesssim \sigma n^{4}, \tag{3.8}$$

then under Conditions 1-2 we have

$$\left\|\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{L}_{\mathsf{cvx}}\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^{5}} \qquad \text{and} \qquad \left\|\boldsymbol{S} - \boldsymbol{S}_{\mathsf{cvx}}\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^{5}}.$$

This theorem is a deterministic result, focusing on some sort of "approximate stationary points" of $F(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S})$. To interpret this, observe that in view of (3.7), one has $\nabla f\left(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{S}\right) \approx \boldsymbol{0}$, and $\boldsymbol{S}$ minimizes $F(\boldsymbol{X}, \boldsymbol{Y}, \cdot)$ for any fixed $\boldsymbol{X}$ and $\boldsymbol{Y}$. If one can identify such an approximate stationary point that is sufficiently close to the truth (so that it satisfies Condition 1), then under mild conditions our theory asserts that

$$\boldsymbol{X}\boldsymbol{Y}^{\top} \approx \boldsymbol{L}_{\mathsf{cvx}} \qquad \text{and} \qquad \boldsymbol{S} \approx \boldsymbol{S}_{\mathsf{cvx}}.$$

This would in turn formalize the intimate relation between the solution to convex relaxation and an approximate stationary point of the nonconvex formulation. The existence of such approximate stationary points will be verified shortly in Section 3.3.

The careful reader might immediately remark that this theorem does not say anything explicit about the minimizer of the nonconvex optimization problem (3.6); rather, it only pays attention to a special class of approximate stationary points of the nonconvex formulation. This arises mainly due to a technical consideration: it seems more difficult to analyze the nonconvex optimizer directly than to study certain approximate stationary points. Fortunately, our theorem indicates that any approximate stationary point obeying the above conditions serves as an extremely tight approximation of the convex estimate, and, therefore, it suffices to identify and analyze any such points.

3.3 Constructing an approximate stationary point via nonconvex algorithms

By virtue of Theorem 4, the key to understanding convex relaxation is to construct an approximate stationary point of the nonconvex problem (3.6) that enjoys desired statistical properties. For this purpose, we resort to the following iterative algorithm (Algorithm 1) to solve the nonconvex program (3.6).

In a nutshell, Algorithm 1 alternates between one iteration of gradient updates (w.r.t. the decision matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$) and optimization of the non-smooth problem w.r.t. $\boldsymbol{S}$ (with $\boldsymbol{X}$ and $\boldsymbol{Y}$ frozen).[8] For the sake of simplicity, we initialize this algorithm from the ground truth $(\boldsymbol{X}^{\star}, \boldsymbol{Y}^{\star}, \boldsymbol{S}^{\star})$, but our analysis framework might be extended to accommodate other more practical initialization (e.g. the one obtained by a spectral method [CCFM20]).

The following theorem makes precise the statistical guarantees of the above nonconvex optimization algorithm; the proof is deferred to Appendix F. Here and throughout, we define

$$\boldsymbol{H}^{t} \coloneqq \underset{\boldsymbol{R} \in \mathcal{O}^{r \times r}}{\arg\min} \ \left(\left\|\boldsymbol{X}^{t}\boldsymbol{R} - \boldsymbol{X}^{\star}\right\|_{\mathrm{F}}^{2} + \left\|\boldsymbol{Y}^{t}\boldsymbol{R} - \boldsymbol{Y}^{\star}\right\|_{\mathrm{F}}^{2}\right)^{1/2}, \tag{3.10}$$

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

[8]Note that for any given $\boldsymbol{X}$ and $\boldsymbol{Y}$, the solution to minimize$_{\boldsymbol{S}}$ $F(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S})$ is given precisely by $\mathcal{S}_{\tau}(\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^{\top}))$.

Algorithm 1 Alternating minimization method for solving the nonconvex problem (3.6).

Suitable initialization: $\boldsymbol{X}^0=\boldsymbol{X}^\star$, $\boldsymbol{Y}^0=\boldsymbol{Y}^\star$, $\boldsymbol{S}^0=\boldsymbol{S}^\star$.

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

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t-\eta\nabla_{\boldsymbol{X}} f\left(\boldsymbol{X}^t,\boldsymbol{Y}^t;\boldsymbol{S}^t\right)=\boldsymbol{X}^t-\frac{\eta}{p}\left[\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}^t\boldsymbol{Y}^{t\top}+\boldsymbol{S}^t-\boldsymbol{M}\right)\boldsymbol{Y}^t+\lambda\boldsymbol{X}^t\right]; \tag{3.9a}$$

$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t-\eta\nabla_{\boldsymbol{Y}} f\left(\boldsymbol{X}^t,\boldsymbol{Y}^t;\boldsymbol{S}^t\right)=\boldsymbol{Y}^t-\frac{\eta}{p}\left\{\left[\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}^t\boldsymbol{Y}^{t\top}+\boldsymbol{S}^t-\boldsymbol{M}\right)\right]^\top\boldsymbol{X}^t+\lambda\boldsymbol{Y}^t\right\}; \tag{3.9b}$$

$$\boldsymbol{S}^{t+1} = \mathcal{S}_\tau\left[\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{M}-\boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top}\right)\right]. \tag{3.9c}$$

Theorem 5. *Instate the assumptions of Theorem 2 and define*

$$\delta_n := \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}.$$

Take $t_0=n^{47}$ and $\eta\asymp 1/(n\kappa^3\sigma_{\max})$ in Algorithm 1. With probability at least $1-O(n^{-3})$, the iterates $\{(\boldsymbol{X}^t,\boldsymbol{Y}^t,\boldsymbol{S}^t)\}_{0\le t\le t_0}$ of Algorithm 1 satisfy

$$\max\left\{\left\|\boldsymbol{X}^t\boldsymbol{H}^t-\boldsymbol{X}^\star\right\|_{\mathrm{F}},\left\|\boldsymbol{Y}^t\boldsymbol{H}^t-\boldsymbol{Y}^\star\right\|_{\mathrm{F}}\right\}\lesssim\delta_n\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{3.11a}$$

$$\max\left\{\left\|\boldsymbol{X}^t\boldsymbol{H}^t-\boldsymbol{X}^\star\right\|,\left\|\boldsymbol{Y}^t\boldsymbol{H}^t-\boldsymbol{Y}^\star\right\|\right\}\lesssim\delta_n\left\|\boldsymbol{X}^\star\right\|, \tag{3.11b}$$

$$\max\left\{\left\|\boldsymbol{X}^t\boldsymbol{H}^t-\boldsymbol{X}^\star\right\|_{2,\infty},\left\|\boldsymbol{Y}^t\boldsymbol{H}^t-\boldsymbol{Y}^\star\right\|_{2,\infty}\right\}\lesssim\kappa\sqrt{\log n}\,\delta_n\max\left\{\left\|\boldsymbol{X}^\star\right\|_{2,\infty},\left\|\boldsymbol{Y}^\star\right\|_{2,\infty}\right\}, \tag{3.11c}$$

$$\left\|\boldsymbol{S}^t-\boldsymbol{S}^\star\right\|\lesssim\sigma\sqrt{np}. \tag{3.11d}$$

In addition, with probability at least $1-O(n^{-3})$, one has

$$\min_{0\le t<t_0}\left\|\nabla f\left(\boldsymbol{X}^t,\boldsymbol{Y}^t;\boldsymbol{S}^t\right)\right\|_{\mathrm{F}}\le\frac{1}{n^{20}}\frac{\lambda}{p}\sqrt{\sigma_{\min}}. \tag{3.12}$$

In short, the bounds (3.11a)-(3.11c) reveal that the entire sequence $\{\boldsymbol{X}^t,\boldsymbol{Y}^t\}_{t=0}^{t_0}$ stays sufficiently close to the truth (measured by $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|$, and more importantly, $\|\cdot\|_{2,\infty}$), the inequality (3.11d) demonstrates the goodness of fit of $\{\boldsymbol{S}^t\}_{0\le t\le t_0}$ in terms of the spectral norm accuracy, whereas the last bound (3.12) indicates that there is at least one point in the sequence $\{\boldsymbol{X}^t,\boldsymbol{Y}^t,\boldsymbol{S}^t\}_{0\le t\le t_0}$ that can serve as an approximate stationary point of the nonconvex formulation.

We shall also gather a few immediate consequences of Theorem 5 as follows, which contain basic properties that will be useful throughout.

Corollary 2. *Instate the assumptions of Theorem 5. Suppose that the sample size obeys $n^2p\gg\kappa^4\mu^2r^2n\log^4 n$, the noise satisfies $\delta_n\ll 1/\sqrt{\kappa^4\mu r\log n}$, the outlier fraction satisfies $\rho_{\mathsf{s}}\ll 1/(\kappa^3\mu r\log n)$. With probability at least $1-O(n^{-3})$, the iterates of Algorithm 1 satisfy*

$$\left\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top}-\boldsymbol{L}^\star\right\|_{\mathrm{F}}\lesssim\kappa\delta_n\left\|\boldsymbol{L}^\star\right\|_{\mathrm{F}}, \tag{3.13a}$$

$$\left\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top}-\boldsymbol{L}^\star\right\|_{\infty}\lesssim\sqrt{\kappa^3\mu r\log n}\,\delta_n\left\|\boldsymbol{L}^\star\right\|_{\infty}, \tag{3.13b}$$

$$\left\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top}-\boldsymbol{L}^\star\right\|\lesssim\delta_n\left\|\boldsymbol{L}^\star\right\| \tag{3.13c}$$

simultaneously for all $t\le t_0$.

Proof. See [CCF+20, Appendix D.12]. □

3.4 Proof of Theorem 2

Define

$$t_*:=\arg\min_{0\le t<t_0}\|\nabla f(\boldsymbol{X}^t,\boldsymbol{Y}^t;\boldsymbol{S}^t)\|_{\mathrm{F}}; \tag{3.14}$$

$$\left(\boldsymbol{X}_{\mathsf{ncvx}}, \boldsymbol{Y}_{\mathsf{ncvx}}, \boldsymbol{S}_{\mathsf{ncvx}}\right) := \left(\boldsymbol{X}^{t_*}\boldsymbol{H}^{t_*}, \boldsymbol{Y}^{t_*}\boldsymbol{H}^{t_*}, \boldsymbol{S}^{t_*}\right). \tag{3.15}$$

Theorem 5 and Corollary 2 have established appealing statistical performance of the nonconvex solution $(\boldsymbol{X}_{\mathsf{ncvx}}, \boldsymbol{Y}_{\mathsf{ncvx}}, \boldsymbol{S}_{\mathsf{ncvx}})$. To transfer this desired statistical property to that of $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$, it remains to show that the nonconvex estimator $\left(\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^{\top}, \boldsymbol{S}_{\mathsf{ncvx}}\right)$ is extremely close to the convex estimator $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$. Towards this end, we intend to invoke Theorem 4; therefore, it boils down to verifying the conditions therein.

1. The small gradient condition (cf. (3.7)) holds automatically under (3.12).
2. By virtue of the spectral norm bound (3.11b), one has

$$\|\boldsymbol{X}_{\mathsf{ncvx}} - \boldsymbol{X}^{\star}\| = \left\|\boldsymbol{X}^{t_*}\boldsymbol{H}^{t_*} - \boldsymbol{X}^{\star}\right\| \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\,\|\boldsymbol{L}^{\star}\| \leq \frac{\sqrt{\sigma_{\min}}}{10},$$

   as long as $\sigma\sqrt{\kappa n/p} \ll \sigma_{\min}$. This together with the Weyl inequality verifies the constraints on the singular values of $(\boldsymbol{X}_{\mathsf{ncvx}}, \boldsymbol{Y}_{\mathsf{ncvx}})$.
3. The crude error bounds are valid in view of Theorem 3.
4. Regarding Condition 1 and Condition 2, Lemma 1 and standard inequalities about sub-Gaussian random variables imply that $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\| < \lambda/16$ and $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\infty} \leq \tau/4$. In addition, the bounds (3.11d) and (3.13b) ensure the second assumption $\|\boldsymbol{S}_{\mathsf{ncvx}} - \boldsymbol{S}^{\star}\| \leq \lambda/16$ and $\|\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{L}^{\star}\|_{\infty} \leq \tau/4$ in Condition 1. We are left with the last assumption in Condition 1 and Condition 2, which are guaranteed to hold in view of the following lemma (see Appendix C for the proof).

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

$$\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M}^{\star}\right) - p\left(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M}^{\star}\right)\right\| < \lambda/8, \tag{3.16a}$$

$$\frac{1}{p}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{H})\right\|_{\mathrm{F}}^{2} \geq \frac{1}{32\kappa}\|\boldsymbol{H}\|_{\mathrm{F}}^{2}, \qquad \forall \boldsymbol{H} \in T, \tag{3.16b}$$

$$p^{-1}\left\|\mathcal{P}_{\Omega^{\star}}(\boldsymbol{H})\right\|_{\mathrm{F}}^{2} \leq \frac{1}{128\kappa}\|\boldsymbol{H}\|_{\mathrm{F}}^{2}, \qquad \forall \boldsymbol{H} \in T \tag{3.16c}$$

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

$$\|\boldsymbol{X} - \boldsymbol{X}^{\star}\|_{2,\infty} \leq C_{\infty}\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\max\left\{\|\boldsymbol{X}^{\star}\|_{2,\infty}, \|\boldsymbol{Y}^{\star}\|_{2,\infty}\right\}; \tag{3.17a}$$

$$\|\boldsymbol{Y} - \boldsymbol{Y}^{\star}\|_{2,\infty} \leq C_{\infty}\kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\max\left\{\|\boldsymbol{X}^{\star}\|_{2,\infty}, \|\boldsymbol{Y}^{\star}\|_{2,\infty}\right\}. \tag{3.17b}$$

   Here, T denotes the tangent space of the set of rank-r matrices at the point $\boldsymbol{X}\boldsymbol{Y}^{\top}$, and $C_{\infty} > 0$ is an absolute constant.

Armed with the above conditions, we can readily invoke Theorem 4 to reach

$$\left\|\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^{\top} - \boldsymbol{L}_{\mathsf{cvx}}\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^{5}} \qquad \text{and} \qquad \|\boldsymbol{S}_{\mathsf{ncvx}} - \boldsymbol{S}_{\mathsf{cvx}}\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^{5}}$$

with high probability. This taken collectively with Corollary 2 gives

$$\begin{aligned}
\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^{\star}\|_{\mathrm{F}} &\leq \left\|\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^{\top} - \boldsymbol{L}_{\mathsf{cvx}}\right\|_{\mathrm{F}} + \left\|\boldsymbol{X}_{\mathsf{ncvx}}\boldsymbol{Y}_{\mathsf{ncvx}}^{\top} - \boldsymbol{L}^{\star}\right\|_{\mathrm{F}} \\
&\lesssim \frac{\sigma}{n^{5}} + \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\,\|\boldsymbol{L}^{\star}\|_{\mathrm{F}} \\
&\asymp \kappa\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\,\|\boldsymbol{L}^{\star}\|_{\mathrm{F}}.
\end{aligned}$$

Similar arguments lead to the advertised high-probability bounds

$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\|_\infty \lesssim \sqrt{\kappa^3 \mu r} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} \|\boldsymbol{L}^\star\|_\infty ,$$
$$\|\boldsymbol{L}_{\mathsf{cvx}} - \boldsymbol{L}^\star\| \lesssim \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|\boldsymbol{L}^\star\| .$$

Finally, given that $\boldsymbol{X}_{\mathsf{ncvx}} \boldsymbol{Y}_{\mathsf{ncvx}}^\top$ is a rank-r matrix, the rank-r approximation $\boldsymbol{L}_{\mathsf{cvx},r} := \arg\min_{\boldsymbol{Z}:\mathrm{rank}(\boldsymbol{Z})\leq r} \|\boldsymbol{Z} - \boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}}$ of $\boldsymbol{L}_{\mathsf{cvx}}$ necessarily satisfies

$$\|\boldsymbol{L}_{\mathsf{cvx},r} - \boldsymbol{L}_{\mathsf{cvx}}\|_{\mathrm{F}} \leq \left\|\boldsymbol{X}_{\mathsf{ncvx}} \boldsymbol{Y}_{\mathsf{ncvx}}^\top - \boldsymbol{L}_{\mathsf{cvx}}\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^5} \leq \frac{1}{n^5} \cdot \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|\boldsymbol{L}^\star\| ,$$

which establishes (1.22). In view of the triangle inequality, the properties (1.21) hold unchanged if $\boldsymbol{L}_{\mathsf{cvx}}$ is replaced by $\boldsymbol{L}_{\mathsf{cvx},r}$.

4 Discussion

This paper investigates the unreasonable effectiveness of convex programming in estimating an unknown low-rank matrix from grossly corrupted data. We develop an improved theory that confirms the optimality of convex relaxation in the presence of random noise, gross sparse outliers, and missing data. In particular, our results significantly improve upon the prior statistical guarantees [ZLW$^+$10] under random noise, while further allowing for missing data. Our theoretical analysis is built upon an appealing connection between convex and nonconvex optimization, which has not been established previously.

Having said this, our current work leaves open several important issues that call for further investigation. To begin with, the conditions (1.20) stated in the main theorem are likely suboptimal in terms of the dependency on both the rank r and the condition number κ. For example, we shall keep in mind that in the noise-free setting, the sample size can be as low as $O(nr \mathrm{poly} \log n)$ and the tolerable outlier fraction can be as large as a constant [Li13, CJSC13], both of which exhibit more favorable scalings w.r.t. r and κ compared to our current condition (1.20). Moving forward, our analysis ideas suggest a possible route for analyzing convex relaxation for other structured estimation problems under both random noise and outliers, including but not limited to sparse PCA (the case with a simultaneously low-rank and sparse matrix) [CMW13], low-rank Hankel matrix estimation (the case involving a low-rank Hankel matrix) [CC14], and blind deconvolution[9] (the case that aims to recover a low-rank matrix from structured Fourier measurements) [ARR14]. Last but not least, we would like to point out that it is possible to design a similar debiasing procedure as in [CFMY19] for correcting the bias in the convex estimator, which further allows uncertainty quantification and statistical inference on the unknown low-rank matrix of interest.

Acknowledgements

Y. Chen is supported in part by the AFOSR YIP award FA9550-19-1-0030, by the ONR grant N00014-19-1-2120, by the ARO grants W911NF-20-1-0097 and W911NF-18-1-0303, by the NSF grants CCF-1907661, IIS-1900140 and DMS-2014279, and by the Princeton SEAS innovation award. J. Fan is supported in part by the NSF grants DMS-1662139 and DMS-1712591, the ONR grant N00014-19-1-2120, and the NIH grant 2R01-GM072611-14.

A An equivalent probabilistic model of $\Omega^\star$ used throughout the proof

Recall that $\Omega^\star$ is the support of the sparse component $\boldsymbol{S}^*$. In this section, we introduce an equivalent probabilistic model of $\Omega^\star$, which is more amenable to analysis and shall be assumed throughout the proof.

[9] Our ongoing work [CFWY20] is pursuing this direction.

- **The original model.** Recall from Assumption 3 the way we generate $\Omega^\star$: (1) sample Ω_{obs} from the i.i.d. Bernoulli model with parameter p; (2) for each $(i,j) \in \Omega_{\mathsf{obs}}$, let $(i,j) \in \Omega^\star$ independently with probability ρ_{s}.
- **An equivalent model.** The model involves over-sampling and rejection method: (1) sample Ω_{obs} from the i.i.d. Bernoulli model with parameter p; (2) generate an augmented index set $\Omega_{\mathsf{aug}} \subseteq \Omega_{\mathsf{obs}}$ such that: for each $(i,j) \in \Omega_{\mathsf{obs}}$, we generate $(i,j) \in \Omega_{\mathsf{aug}}$ independently with probability ρ_{aug}; (3) for any $(i,j) \in \Omega_{\mathsf{aug}}$, include (i,j) in $\Omega^\star$ independently with probability $\rho_{\mathsf{s}}/\rho_{\mathsf{aug}}$.

It is straightforward to verify that the two models for $\Omega^\star$ are equivalent as long as $\rho_{\mathsf{s}} \leq \rho_{\mathsf{aug}} \leq 1$. Two important remarks are in order. First, by construction, we have $\Omega^\star \subseteq \Omega_{\mathsf{aug}}$. Second, the choice of ρ_{aug} can vary as needed as long as $\rho_{\mathsf{s}} \leq \rho_{\mathsf{aug}} \leq 1$.

The introduction of this augmented index set Ω_{aug} comes in handy when we would like to control the size $\|\mathcal{P}_{\Omega^\star}(\boldsymbol{A})\|_{\mathrm{F}}$ for some matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. The first inclusion property $\Omega^\star \subseteq \Omega_{\mathsf{aug}}$ allows us to upper bound $\|\mathcal{P}_{\Omega^\star}(\boldsymbol{A})\|_{\mathrm{F}}$ by $\|\mathcal{P}_{\Omega_{\mathsf{aug}}}(\boldsymbol{A})\|_{\mathrm{F}}$, and the freedom to choose ρ_{aug} allows us to leverage stronger concentration results, which might not hold for the smaller ρ_{s}. See Corollary 3 in the next section for an example.

B Preliminaries

B.1 A few preliminary facts

This subsection collects several results that are useful throughout the proof. To begin with, the incoherence assumption (cf. Assumption 1) asserts that

$$\|\boldsymbol{X}^\star\|_{2,\infty} \leq \sqrt{\mu r/n}\,\|\boldsymbol{X}^\star\| \qquad \text{and} \qquad \|\boldsymbol{Y}^\star\|_{2,\infty} \leq \sqrt{\mu r/n}\,\|\boldsymbol{Y}^\star\|. \tag{B.1}$$

This is because

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

where the first inequality comes from the elementary inequality $\|\boldsymbol{A}\boldsymbol{B}\|_{2,\infty} \leq \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|$, and the last inequality is a consequence of the incoherence assumption as well as the fact that $\|(\boldsymbol{\Sigma}^\star)^{1/2}\| = \|\boldsymbol{X}^\star\|$.

The next lemma is extensively used in the low-rank matrix completion literature.

Lemma 3. *Suppose that each (i,j) is included in $\Omega_0 \subseteq [n]\times[n]$ independently with probability ρ_0. Then with probability exceeding $1 - O(n^{-10})$, one has*

$$\left\|\mathcal{P}_{T^\star} - \rho_0^{-1}\mathcal{P}_{T^\star}\mathcal{P}_{\Omega_0}\mathcal{P}_{T^\star}\right\| \leq \frac{1}{2}, \tag{B.2}$$

provided that $n^2\rho_0 \gg \mu r n \log n$. Here, $T^\star$ denotes the tangent space of the set of rank-r matrices at the point $\boldsymbol{L}^\star = \boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}$.

Proof. See [CR09, Theorem 4.1] □

In fact, the bound (B.2) uncovers certain near-isometry of the operator $\rho_0^{-1}\mathcal{P}_{\Omega_0}(\cdot)$ when restricted to the tangent space $T^\star$. This property is formalized in the following fact.

Fact 1. *Suppose that $\|\mathcal{P}_{T^\star} - \rho_0^{-1}\mathcal{P}_{T^\star}\mathcal{P}_{\Omega_0}\mathcal{P}_{T^\star}\| \leq 1/2$. Then one has*

$$\frac{1}{2}\|\boldsymbol{H}\|_{\mathrm{F}}^2 \leq \frac{1}{\rho_0}\|\mathcal{P}_{\Omega_0}(\boldsymbol{H})\|_{\mathrm{F}}^2 \leq \frac{3}{2}\|\boldsymbol{H}\|_{\mathrm{F}}^2, \qquad \textit{for all } \boldsymbol{H} \in T^\star.$$

Proof. The proof has actually been documented in the literature. For completeness, we present the proof for the lower bound here; the upper bound follows from a very similar argument. For any $\boldsymbol{H} \in \mathbb{R}^{n\times n}$, one has

$$\begin{aligned}\|\mathcal{P}_{\Omega_0}\mathcal{P}_{T^\star}(\boldsymbol{H})\|_{\mathrm{F}}^2 &= \langle\mathcal{P}_{\Omega_0}\mathcal{P}_{T^\star}(\boldsymbol{H}), \mathcal{P}_{\Omega_0}\mathcal{P}_{T^\star}(\boldsymbol{H})\rangle = \langle\mathcal{P}_{T^\star}(\boldsymbol{H}), \mathcal{P}_{T^\star}\mathcal{P}_{\Omega_0}\mathcal{P}_{T^\star}(\boldsymbol{H})\rangle \\ &= \rho_0\|\mathcal{P}_{T^\star}(\boldsymbol{H})\|_{\mathrm{F}}^2 - \rho_0\left\langle\mathcal{P}_{T^\star}(\boldsymbol{H}), \left(\mathcal{P}_{T^\star} - \rho_0^{-1}\mathcal{P}_{T^\star}\mathcal{P}_{\Omega_0}\mathcal{P}_{T^\star}\right)(\mathcal{P}_{T^\star}(\boldsymbol{H}))\right\rangle\end{aligned}$$

$$\geq \rho_0 \|\mathcal{P}_{T^\star}(\boldsymbol{H})\|_{\mathrm{F}}^2 - \rho_0 \left\|\mathcal{P}_{T^\star} - \rho_0^{-1}\mathcal{P}_{T^\star}\mathcal{P}_{\Omega_0}\mathcal{P}_{T^\star}\right\| \|\mathcal{P}_{T^\star}(\boldsymbol{H})\|_{\mathrm{F}}^2$$
$$\geq \frac{\rho_0}{2} \|\mathcal{P}_{T^\star}(\boldsymbol{H})\|_{\mathrm{F}}^2 .$$

Here, the penultimate inequality relies on the elementary fact that $\langle \boldsymbol{A}, \boldsymbol{B}\rangle \leq \|\boldsymbol{A}\|_{\mathrm{F}}\|\boldsymbol{B}\|_{\mathrm{F}}$, and the last step follows from the assumption $\|\mathcal{P}_{T^\star} - \rho_0^{-1}\mathcal{P}_{T^\star}\mathcal{P}_{\Omega_0}\mathcal{P}_{T^\star}\| \leq 1/2$. □

The following corollary is an immediate consequence of Lemma 3 and Fact 1.

Corollary 3. *Suppose that $\rho_{\mathsf{s}} \leq \rho_{\mathsf{aug}} \leq 1/12$ and that $n^2 p\rho_{\mathsf{aug}} \gg \mu r n \log n$. Then with probability at least $1 - O(n^{-10})$, we have*

$$\|\mathcal{P}_{\Omega^\star}\mathcal{P}_{T^\star}\|^2 \leq p/8.$$

Proof. Recall the auxiliary index set Ω_{aug} introduced in Appendix A. Since $\Omega^\star \subseteq \Omega_{\mathsf{aug}}$, we have for any $\boldsymbol{H} \in \mathbb{R}^{n\times n}$

$$\|\mathcal{P}_{\Omega^\star}\mathcal{P}_{T^\star}(\boldsymbol{H})\|_{\mathrm{F}}^2 \leq \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\mathcal{P}_{T^\star}(\boldsymbol{H})\right\|_{\mathrm{F}}^2 \leq \frac{3p\rho_{\mathsf{aug}}}{2} \|\mathcal{P}_{T^\star}(\boldsymbol{H})\|_{\mathrm{F}}^2 \leq \frac{3p\rho_{\mathsf{aug}}}{2} \|\boldsymbol{H}\|_{\mathrm{F}}^2 .$$

Here, the second inequality arises from Lemma 3 and Fact 1 (by taking $\Omega_0 = \Omega_{\mathsf{aug}}$ and $\rho_0 = p\rho_{\mathsf{aug}}$). The proof is complete by recognizing the assumption $\rho_{\mathsf{aug}} \leq 1/12$. □

As it turns out, the near-isometry property of $\rho_0^{-1}\mathcal{P}_{\Omega_0}(\cdot)$ can be strengthened to a uniform version (uniform over a large collection of tangent spaces), as shown in the lemma below.

Lemma 4. *Suppose that each (i, j) is included in $\Omega_0 \subseteq [n] \times [n]$ independently with probability ρ_0, and that $n^2\rho_0 \gg \mu r n \log n$. Then with probability at least $1 - O(n^{-10})$,*

$$\frac{1}{32\kappa} \|\boldsymbol{H}\|_{\mathrm{F}}^2 \leq \frac{1}{\rho_0} \|\mathcal{P}_{\Omega_0}(\boldsymbol{H})\|_{\mathrm{F}}^2 \leq 40\kappa \|\boldsymbol{H}\|_{\mathrm{F}}^2 , \qquad \text{for all } \boldsymbol{H} \in T$$

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

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

Here, $c > 0$ is some sufficiently small constant, and T denotes the tangent space of the set of rank-r matrices at the point $\boldsymbol{X}\boldsymbol{Y}^\top$.

Proof. See Appendix B.2. □

In the end, we recall a useful lemma which relates the operator norm to the $\ell_{2,\infty}$ norm of a matrix.

Lemma 5. *Suppose that each (i, j) is included in $\Omega_0 \subseteq [n] \times [n]$ independently with probability ρ_0, and that $n^2\rho_0 \gg n \log n$. Then there exists some absolute constant $C > 0$ such that with probability at least $1 - O(n^{-10})$,*

$$\left\|\mathcal{P}_{\Omega_0}\left(\boldsymbol{A}\boldsymbol{B}^\top\right) - \rho_0\boldsymbol{A}\boldsymbol{B}^\top\right\| \leq C\sqrt{n\rho_0} \|\boldsymbol{A}\|_{2,\infty} \|\boldsymbol{B}\|_{2,\infty}$$

holds simultaneously for all $\boldsymbol{A}$ and $\boldsymbol{B}$.

Proof. See [CLL20, Lemmas 4.2 and 4.3]. □

B.2 Proof of Lemma 4

The lower bound has been established in [CCF+20, Lemma 7], and hence we focus on the upper bound. We start by expressing $\boldsymbol{H} \in T$ as $\boldsymbol{H} = \boldsymbol{X}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^\top$, where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times r}$ are chosen to be

$$(\boldsymbol{A}, \boldsymbol{B}) := \mathop{\arg\min}_{(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}):\, \boldsymbol{H} = \boldsymbol{X}\tilde{\boldsymbol{A}}^\top + \tilde{\boldsymbol{B}}\boldsymbol{Y}^\top} \left\{\|\tilde{\boldsymbol{A}}\|_{\mathrm{F}}^2/2 + \|\tilde{\boldsymbol{B}}\|_{\mathrm{F}}^2/2\right\}.$$

The optimality condition of $(\boldsymbol{A}, \boldsymbol{B})$ requires

$$\boldsymbol{X}^\top\boldsymbol{B} = \boldsymbol{A}^\top\boldsymbol{Y}; \tag{B.3}$$

see [CCF+20, Section C.3.1] for the justification of this identity. The proof then consists of two steps:

1. Showing that $\|\boldsymbol{H}\|_{\mathrm{F}}^2$ is bounded from below, namely,

$$\|\boldsymbol{H}\|_{\mathrm{F}}^2 \geq \frac{49}{100}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right).$$

To see this, we can invoke the bound on α_2 stated in [CCF+20, Appendix C.3.1] to yield

$$\begin{aligned}\|\boldsymbol{H}\|_{\mathrm{F}}^2 = \left\|\boldsymbol{X}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^\top\right\|_{\mathrm{F}}^2 &\geq \frac{1}{2}\left(\left\|\boldsymbol{X}^\star\boldsymbol{A}^\top\right\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\boldsymbol{Y}^\star\|_{\mathrm{F}}^2\right) - \frac{1}{100}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right)\\ &\geq \frac{1}{2}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right) - \frac{1}{100}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right) \geq \frac{49}{100}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right).\end{aligned}$$

2. Showing that $\|\mathcal{P}_{\Omega^\star}(\boldsymbol{H})\|_{\mathrm{F}}^2$ is bounded from above, namely,

$$\frac{1}{2\rho_0}\left\|\mathcal{P}_{\Omega^\star}(\boldsymbol{H})\right\|_{\mathrm{F}}^2 \leq 9\sigma_{\max}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right).$$

To this end, one starts with the following decomposition

$$\frac{1}{2\rho_0}\left\|\mathcal{P}_{\Omega_0}(\boldsymbol{H})\right\|_{\mathrm{F}}^2 = \frac{1}{2}\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \frac{1}{2\rho_0}\left\|\mathcal{P}_{\Omega_0}(\boldsymbol{H})\right\|_{\mathrm{F}}^2 - \frac{1}{2}\|\boldsymbol{H}\|_{\mathrm{F}}^2. \tag{B.4}$$

Apply [CCF+20, Equation (83)] to obtain

$$\frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^\top\right\|_{\mathrm{F}}^2 \leq 8\sigma_{\max}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right).$$

In addition, the bound on α_1 stated in [CCF+20, Appendix C.3.1] tells us that

$$\begin{aligned}\frac{1}{2\rho_0}\left\|\mathcal{P}_{\Omega_0}(\boldsymbol{H})\right\|_{\mathrm{F}}^2 - \frac{1}{2}\|\boldsymbol{H}\|_{\mathrm{F}}^2 &= \frac{1}{2\rho_0}\left\|\mathcal{P}_{\Omega_0}\left(\boldsymbol{X}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^\top\right)\right\|_{\mathrm{F}}^2 - \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{Y}^\top\right\|_{\mathrm{F}}^2\\ &\leq \frac{1}{32}\left(\left\|\boldsymbol{X}^\star\boldsymbol{A}^\top\right\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\boldsymbol{Y}^\star\|_{\mathrm{F}}^2\right) + \frac{1}{25}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right)\\ &\leq \left(\frac{1}{32}\sigma_{\max} + \frac{1}{25}\sigma_{\min}\right)\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right).\end{aligned}$$

Substitution into (B.4) gives

$$\begin{aligned}\frac{1}{2\rho_0}\left\|\mathcal{P}_{\Omega_0}(\boldsymbol{H})\right\|_{\mathrm{F}}^2 &\leq 8\sigma_{\max}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right) + \left(\frac{1}{32}\sigma_{\max} + \frac{1}{25}\sigma_{\min}\right)\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right)\\ &\leq 9\sigma_{\max}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right).\end{aligned}$$

Putting the above two bounds together, we conclude that

$$\frac{1}{2\rho_0}\left\|\mathcal{P}_{\Omega_0}(\boldsymbol{H})\right\|_{\mathrm{F}}^2 \leq 9\sigma_{\max}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right) \leq \frac{900}{49}\kappa\cdot\frac{49}{100}\sigma_{\min}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}\|_{\mathrm{F}}^2\right) \leq 20\kappa\|\boldsymbol{H}\|_{\mathrm{F}}^2$$

as claimed.

C Proof of Lemma 2

With Lemma 4 in place, we can immediately justify Lemma 2.

To begin with, the first two parts (3.16a) and (3.16b) are the same as [CCF+20, Lemma 4]. Hence, it suffices to verify the last one (3.16c). Recall from Appendix A that $\Omega^\star \subseteq \Omega_{\mathsf{aug}}$, where Ω_{aug} is randomly sampled such that each (i,j) is included in Ω_{aug} independently with probability $p\rho_{\mathsf{aug}}$. Applying Lemma 4 on Ω_{aug} finishes the proof, with the proviso that $\rho_{\mathsf{aug}} \asymp 1/\kappa^2$ and $\rho_{\mathsf{s}} \leq \rho_{\mathsf{aug}}$.

D Crude error bounds (Proof of Theorem 3)

This section is devoted to establishing our crude statistical error bounds on $\|\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^{\star}\|_{\mathrm{F}}$ and $\|\boldsymbol{S}_{\mathsf{cvx}}-\boldsymbol{S}^{\star}\|_{\mathrm{F}}$. Without loss of generality, we only consider the case when $\tau=\lambda\sqrt{\frac{\log n}{np}}$. The proof works for general choices $\tau\asymp\lambda\sqrt{\frac{\log n}{np}}$ with slight modification. To simplify the notation hereafter, we denote

$$\boldsymbol{\Lambda}_{\boldsymbol{L}}\coloneqq\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^{\star},\qquad\text{and}\qquad\boldsymbol{\Lambda}_{\boldsymbol{S}}\coloneqq\boldsymbol{S}_{\mathsf{cvx}}-\boldsymbol{S}^{\star},$$
$$\boldsymbol{\Lambda}^{+}\coloneqq(\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}})+\boldsymbol{\Lambda}_{\boldsymbol{S}})/2,\qquad\text{and}\qquad\boldsymbol{\Lambda}^{-}\coloneqq(\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}})-\boldsymbol{\Lambda}_{\boldsymbol{S}})/2,$$

which immediately imply

$$\boldsymbol{\Lambda}_{\boldsymbol{L}}=\boldsymbol{\Lambda}^{+}+\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}}),\qquad\text{and}\qquad\boldsymbol{\Lambda}_{\boldsymbol{S}}=\boldsymbol{\Lambda}^{+}-\boldsymbol{\Lambda}^{-}.$$

These in turn allow us to decompose $\|\boldsymbol{\Lambda}_{\boldsymbol{L}}\|_{\mathrm{F}}^{2}+\|\boldsymbol{\Lambda}_{\boldsymbol{S}}\|_{\mathrm{F}}^{2}$ as follows

$$\begin{aligned}\|\boldsymbol{\Lambda}_{\boldsymbol{L}}\|_{\mathrm{F}}^{2}+\|\boldsymbol{\Lambda}_{\boldsymbol{S}}\|_{\mathrm{F}}^{2}&=\left\|\boldsymbol{\Lambda}^{+}+\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}})\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{\Lambda}^{+}-\boldsymbol{\Lambda}^{-}\right\|_{\mathrm{F}}^{2}\\&=\left\|\boldsymbol{\Lambda}^{+}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}})\right\|_{\mathrm{F}}^{2}+2\left\langle\boldsymbol{\Lambda}^{+},\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}})\right\rangle+\left\|\boldsymbol{\Lambda}^{+}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{\Lambda}^{-}\right\|_{\mathrm{F}}^{2}-2\left\langle\boldsymbol{\Lambda}^{+},\boldsymbol{\Lambda}^{-}\right\rangle\\&=2\left\|\boldsymbol{\Lambda}^{+}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}})\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{\Lambda}^{-}\right\|_{\mathrm{F}}^{2}+2\left\langle\boldsymbol{\Lambda}^{+},\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}})\right\rangle.\end{aligned}\tag{D.1}$$

Since $(\boldsymbol{L}_{\mathsf{cvx}},\boldsymbol{S}_{\mathsf{cvx}})$ is the minimizer of (1.3), it is self-evident that $\boldsymbol{S}_{\mathsf{cvx}}$ must be supported on Ω_{obs}. Then by construction, $\boldsymbol{\Lambda}_{\boldsymbol{S}},\boldsymbol{\Lambda}^{+}$ and $\boldsymbol{\Lambda}^{-}$ are all necessarily supported on Ω_{obs}, thus indicating that

$$\langle\boldsymbol{\Lambda}^{+},\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}})\rangle=0.$$

Making use of this relation, we can continue the derivation (D.1) above to obtain

$$\begin{aligned}&\|\boldsymbol{\Lambda}_{\boldsymbol{L}}\|_{\mathrm{F}}^{2}+\|\boldsymbol{\Lambda}_{\boldsymbol{S}}\|_{\mathrm{F}}^{2}=2\left\|\boldsymbol{\Lambda}^{+}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}})\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{\Lambda}^{-}\right\|_{\mathrm{F}}^{2}\\&\quad=\underbrace{2\left\|\boldsymbol{\Lambda}^{+}\right\|_{\mathrm{F}}^{2}}_{=:\alpha_{1}}+\underbrace{\left\|\mathcal{P}_{T^{\star}}\left(\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}})\right)\right\|_{\mathrm{F}}^{2}+\left\|\mathcal{P}_{\Omega^{\star}}(\boldsymbol{\Lambda}^{-})\right\|_{\mathrm{F}}^{2}}_{=:\alpha_{2}}+\underbrace{\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}})\right)\right\|_{\mathrm{F}}^{2}+\left\|\mathcal{P}_{\Omega^{\star\mathrm{c}}}(\boldsymbol{\Lambda}^{-})\right\|_{\mathrm{F}}^{2}}_{=:\alpha_{3}}.\end{aligned}$$

In the sequel, we shall control the three terms α_1,α_2 and α_3 separately.

Step 1: bounding α_1. By definition, we have

$$\begin{aligned}\alpha_{1}=2\left\|\boldsymbol{\Lambda}^{+}\right\|_{\mathrm{F}}^{2}&=\frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}})+\boldsymbol{\Lambda}_{\boldsymbol{S}}\right\|_{\mathrm{F}}^{2}=\frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}}+\boldsymbol{\Lambda}_{\boldsymbol{S}})\right\|_{\mathrm{F}}^{2}\\&=\frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}_{\mathsf{cvx}}+\boldsymbol{S}_{\mathsf{cvx}}-\boldsymbol{M}+\boldsymbol{M}-\boldsymbol{L}^{\star}-\boldsymbol{S}^{\star})\right\|_{\mathrm{F}}^{2}\\&\leq\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}_{\mathsf{cvx}}+\boldsymbol{S}_{\mathsf{cvx}}-\boldsymbol{M})\right\|_{\mathrm{F}}^{2}+\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}^{\star}+\boldsymbol{S}^{\star}-\boldsymbol{M})\right\|_{\mathrm{F}}^{2}\\&=\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}_{\mathsf{cvx}}+\boldsymbol{S}_{\mathsf{cvx}}-\boldsymbol{M})\right\|_{\mathrm{F}}^{2}+\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\right\|_{\mathrm{F}}^{2},\end{aligned}\tag{D.2}$$

where the third identity holds true since $\boldsymbol{\Lambda}_{\boldsymbol{S}}=\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Lambda}_{\boldsymbol{S}})$, the penultimate relation is due to the elementary inequality $\|\boldsymbol{A}+\boldsymbol{B}\|_{\mathrm{F}}^{2}\leq2\|\boldsymbol{A}\|_{\mathrm{F}}^{2}+2\|\boldsymbol{B}\|_{\mathrm{F}}^{2}$, and the last line follows since $\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}^{\star}+\boldsymbol{S}^{\star}-\boldsymbol{M})=\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})$. To upper bound $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}_{\mathsf{cvx}}+\boldsymbol{S}_{\mathsf{cvx}}-\boldsymbol{M})\|_{\mathrm{F}}^{2}$, we leverage the optimality of $(\boldsymbol{L}_{\mathsf{cvx}},\boldsymbol{S}_{\mathsf{cvx}})$ w.r.t. the convex program (1.3) to obtain

$$\begin{aligned}&\frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}_{\mathsf{cvx}}+\boldsymbol{S}_{\mathsf{cvx}}-\boldsymbol{M})\right\|_{\mathrm{F}}^{2}+\lambda\left\|\boldsymbol{L}_{\mathsf{cvx}}\right\|_{*}+\tau\left\|\boldsymbol{S}_{\mathsf{cvx}}\right\|_{1}\\&\qquad\leq\frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}^{\star}+\boldsymbol{S}^{\star}-\boldsymbol{M})\right\|_{\mathrm{F}}^{2}+\lambda\left\|\boldsymbol{L}^{\star}\right\|_{*}+\tau\left\|\boldsymbol{S}^{\star}\right\|_{1}.\end{aligned}\tag{D.3}$$

Recognizing again that $\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}^{\star}+\boldsymbol{S}^{\star}-\boldsymbol{M})=\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})$, we can rearrange terms in (D.3) to derive

$$\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}_{\mathsf{cvx}}+\boldsymbol{S}_{\mathsf{cvx}}-\boldsymbol{M})\right\|_{\mathrm{F}}^{2}\leq\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\right\|_{\mathrm{F}}^{2}+2\lambda\left\|\boldsymbol{L}^{\star}\right\|_{*}+2\tau\left\|\boldsymbol{S}^{\star}\right\|_{1}-2\lambda\left\|\boldsymbol{L}_{\mathsf{cvx}}\right\|_{*}-2\tau\left\|\boldsymbol{S}_{\mathsf{cvx}}\right\|_{1}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{E}\right)\right\|_{\mathrm{F}}^{2}+2\lambda\left\|\boldsymbol{\Lambda}_{\boldsymbol{L}}\right\|_{*}+2\tau\left\|\boldsymbol{\Lambda}_{\boldsymbol{S}}\right\|_{1}\\
&\overset{\text{(ii)}}{\leq} \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{E}\right)\right\|_{\mathrm{F}}^{2}+2\lambda\sqrt{n}\left\|\boldsymbol{\Lambda}_{\boldsymbol{L}}\right\|_{\mathrm{F}}+2\tau\sqrt{\left|\Omega_{\mathsf{obs}}\right|}\left\|\boldsymbol{\Lambda}_{\boldsymbol{S}}\right\|_{\mathrm{F}}\\
&\overset{\text{(iii)}}{\leq} \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{E}\right)\right\|_{\mathrm{F}}^{2}+2\sqrt{2}\lambda\sqrt{n\log n}\left(\left\|\boldsymbol{\Lambda}_{\boldsymbol{L}}\right\|_{\mathrm{F}}+\left\|\boldsymbol{\Lambda}_{\boldsymbol{S}}\right\|_{\mathrm{F}}\right),
\end{aligned}
\tag{D.4}
$$

where $|\Omega_{\mathsf{obs}}|$ denotes the cardinality of Ω_{obs}. Here, the relation (i) results from the triangle inequality, the inequality (ii) holds true since $\|\boldsymbol{A}\|_* \leq \sqrt{n}\|\boldsymbol{A}\|_{\mathrm{F}}$ for any $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\|\boldsymbol{\Lambda}_{\boldsymbol{S}}\|_1 = \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Lambda}_{\boldsymbol{S}})\|_1 \leq \sqrt{|\Omega_{\mathsf{obs}}|}\|\boldsymbol{\Lambda}_{\boldsymbol{S}}\|_{\mathrm{F}}$, and the last line (iii) arises from the fact that $|\Omega_{\mathsf{obs}}| \leq 2n^2p$ with high probability as well as the choice $\tau = \lambda\sqrt{\frac{\log n}{np}}$. Combine (D.2) and (D.4) to reach

$$
\begin{aligned}
\alpha_1 &\leq 2\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{E}\right)\|_{\mathrm{F}}^{2}+2\sqrt{2}\lambda\sqrt{n\log n}\left(\left\|\boldsymbol{\Lambda}_{\boldsymbol{L}}\right\|_{\mathrm{F}}+\left\|\boldsymbol{\Lambda}_{\boldsymbol{S}}\right\|_{\mathrm{F}}\right)\\
&\leq 2\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{E}\right)\right\|_{\mathrm{F}}^{2}+4\lambda\sqrt{n\log n}\sqrt{\left\|\boldsymbol{\Lambda}_{\boldsymbol{L}}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{\Lambda}_{\boldsymbol{S}}\right\|_{\mathrm{F}}^{2}},
\end{aligned}
\tag{D.5}
$$

where we use the elementary inequality $a+b \leq \sqrt{2}\cdot\sqrt{a^2+b^2}$.

Step 2: bounding α_2 via α_3. To relate α_2 to α_3, the following lemma plays a crucial role, whose proof is deferred to Appendix D.1.

Lemma 6. *Suppose that $\|\mathcal{P}_{\Omega^\star}\mathcal{P}_{T^\star}\|^2 \leq p/8$ and that $\|\mathcal{P}_{T^\star} - p^{-1}\mathcal{P}_{T^\star}\mathcal{P}_{\Omega_{\mathsf{obs}}}\mathcal{P}_{T^\star}\| \leq 1/2$. Then for any pair $(\boldsymbol{A},\boldsymbol{B})$ of matrices, we have*

$$
\left\|\mathcal{P}_{T^\star}\left(\boldsymbol{A}\right)\right\|_{\mathrm{F}}^{2}+\left\|\mathcal{P}_{\Omega^\star}\left(\boldsymbol{B}\right)\right\|_{\mathrm{F}}^{2} \leq \frac{4}{p}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left[\mathcal{P}_{T^\star}\left(\boldsymbol{A}\right)+\mathcal{P}_{\Omega^\star}\left(\boldsymbol{B}\right)\right]\right\|_{\mathrm{F}}^{2}. \tag{D.6}
$$

Suppose for the moment that the assumptions of Lemma 6 hold. Taking $(\boldsymbol{A},\boldsymbol{B})$ as $(\boldsymbol{\Lambda}^- + \mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}(\boldsymbol{\Lambda}_{\boldsymbol{L}}), -\boldsymbol{\Lambda}^-)$ in Lemma 6 yields

$$
\alpha_2 = \left\|\mathcal{P}_{T^\star}\left(\boldsymbol{\Lambda}^- + \mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}\left(\boldsymbol{\Lambda}_{\boldsymbol{L}}\right)\right)\right\|_{\mathrm{F}}^{2}+\left\|\mathcal{P}_{\Omega^\star}\left(\boldsymbol{\Lambda}^-\right)\right\|_{\mathrm{F}}^{2} \leq \frac{4}{p}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left[\mathcal{P}_{T^\star}\left(\boldsymbol{\Lambda}^- + \mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}\left(\boldsymbol{\Lambda}_{\boldsymbol{L}}\right)\right)-\mathcal{P}_{\Omega^\star}\left(\boldsymbol{\Lambda}^-\right)\right]\right\|_{\mathrm{F}}^{2}.
$$

By virtue of the identity

$$
\begin{aligned}
\mathcal{P}_{T^\star}\left(\boldsymbol{\Lambda}^- + \mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}\left(\boldsymbol{\Lambda}_{\boldsymbol{L}}\right)\right)-\mathcal{P}_{\Omega^\star}\left(\boldsymbol{\Lambda}^-\right) &= \boldsymbol{\Lambda}^- + \mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}\left(\boldsymbol{\Lambda}_{\boldsymbol{L}}\right) - \mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^- + \mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}\left(\boldsymbol{\Lambda}_{\boldsymbol{L}}\right)\right) - \boldsymbol{\Lambda}^- + \mathcal{P}_{(\Omega^\star)^{\mathsf{c}}}(\boldsymbol{\Lambda}^-)\\
&= \mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}\left(\boldsymbol{\Lambda}_{\boldsymbol{L}}\right) - \mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^- + \mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}\left(\boldsymbol{\Lambda}_{\boldsymbol{L}}\right)\right) + \mathcal{P}_{(\Omega^\star)^{\mathsf{c}}}(\boldsymbol{\Lambda}^-),
\end{aligned}
$$

we further obtain

$$
\begin{aligned}
\alpha_2 &\leq \frac{4}{p}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left[\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}\left(\boldsymbol{\Lambda}_{\boldsymbol{L}}\right) - \mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^- + \mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}\left(\boldsymbol{\Lambda}_{\boldsymbol{L}}\right)\right) + \mathcal{P}_{(\Omega^\star)^{\mathsf{c}}}(\boldsymbol{\Lambda}^-)\right]\right\|_{\mathrm{F}}^{2}\\
&= \frac{4}{p}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left[\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^- + \mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}\left(\boldsymbol{\Lambda}_{\boldsymbol{L}}\right)\right) - \mathcal{P}_{(\Omega^\star)^{\mathsf{c}}}(\boldsymbol{\Lambda}^-)\right]\right\|_{\mathrm{F}}^{2}\\
&\leq \frac{4}{p}\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^- + \mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}\left(\boldsymbol{\Lambda}_{\boldsymbol{L}}\right)\right) - \mathcal{P}_{(\Omega^\star)^{\mathsf{c}}}\left(\boldsymbol{\Lambda}^-\right)\right\|_{\mathrm{F}}^{2}\\
&\leq \frac{8}{p}\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^- + \mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathsf{c}}}\left(\boldsymbol{\Lambda}_{\boldsymbol{L}}\right)\right)\right\|_{\mathrm{F}}^{2} + \frac{8}{p}\left\|\mathcal{P}_{(\Omega^\star)^{\mathsf{c}}}\left(\boldsymbol{\Lambda}^-\right)\right\|_{\mathrm{F}}^{2} = \frac{8}{p}\alpha_3.
\end{aligned}
\tag{D.7}
$$

Once again, the derivation has made use of the elementary inequality $\|\boldsymbol{A}+\boldsymbol{B}\|_{\mathrm{F}}^2 \leq 2\|\boldsymbol{A}\|_{\mathrm{F}}^2 + 2\|\boldsymbol{B}\|_{\mathrm{F}}^2$.

Step 3: bounding α_3 via α_1 and $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}$. The following lemma proves useful in linking α_3 with α_1, and we postpone the proof to Appendix D.2.

Lemma 7. *Assume that $n^2p \gg n\log n$, $\rho_{\mathsf{s}} \ll 1$ and $\|\mathcal{P}_{T^\star} - p^{-1}(1-\rho_{\mathsf{s}})^{-1}\mathcal{P}_{T^\star}\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}\mathcal{P}_{T^\star}\| \leq 1/2$. Further assume that there exists a dual certificate $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ such that*

$$
\left\|\mathcal{P}_{T^\star}\left[\lambda\boldsymbol{W} + \tau\mathsf{sgn}\left(\boldsymbol{S}^\star\right) - \lambda\boldsymbol{U}^\star\boldsymbol{V}^{\star\top}\right]\right\|_{\mathrm{F}} \leq \tau/\sqrt{n}, \tag{D.8a}
$$

$$\left\|\mathcal{P}_{T^{\star\perp}}\left[\lambda \boldsymbol{W}+\tau \mathsf{sign}\left(\boldsymbol{S}^{\star}\right)\right]\right\|<\lambda/2, \tag{D.8b}$$

$$\mathcal{P}_{(\Omega_{\mathsf{obs}}\setminus\Omega^{\star})^{\mathrm{c}}}\left(\boldsymbol{W}\right)=\boldsymbol{0}, \tag{D.8c}$$

$$\left\|\lambda \boldsymbol{W}\right\|_{\infty}<\tau/2, \tag{D.8d}$$

where $\mathsf{sign}\left(\boldsymbol{S}^{\star}\right):=[\mathsf{sign}(S^{\star}_{ij})]_{1\le i,j\le n}$. *Then for any* $\boldsymbol{H_L},\boldsymbol{H_S}\in\mathbb{R}^{n\times n}$ *satisfying* $\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{H_L})+\boldsymbol{H_S}=\boldsymbol{0}$*, one has*

$$\lambda\left\|\boldsymbol{L}^{\star}+\boldsymbol{H_L}\right\|_{*}+\tau\left\|\boldsymbol{S}^{\star}+\boldsymbol{H_S}\right\|_{1}\ge\lambda\left\|\boldsymbol{L}^{\star}\right\|_{*}+\tau\left\|\boldsymbol{S}^{\star}\right\|_{1}+\frac{\lambda}{4}\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{H_L}\right)\right\|_{*}+\frac{\tau}{4}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\left(\boldsymbol{H_S}\right)\right\|_{1}.$$

Again, we assume for the moment that the assumptions in Lemma 7 hold. Setting $\boldsymbol{H}_L=\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}(\boldsymbol{\Lambda_L})$ and $\boldsymbol{H_S}=-\boldsymbol{\Lambda}^{-}$ in Lemma 7 gives

$$\begin{aligned}&\lambda\left\|\boldsymbol{L}^{\star}+\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}\left(\boldsymbol{\Lambda_L}\right)\right\|_{*}+\tau\left\|\boldsymbol{S}^{\star}-\boldsymbol{\Lambda}^{-}\right\|_{1}\\&\quad\ge\lambda\left\|\boldsymbol{L}^{\star}\right\|_{*}+\tau\left\|\boldsymbol{S}^{\star}\right\|_{1}+\frac{\lambda}{4}\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}\left(\boldsymbol{\Lambda_L}\right)\right)\right\|_{*}+\frac{\tau}{4}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\left(\boldsymbol{\Lambda}^{-}\right)\right\|_{1}.\end{aligned}$$

In addition, recalling the identities $\boldsymbol{L}_{\mathsf{cvx}}=\boldsymbol{L}^{\star}+\boldsymbol{\Lambda}^{+}+\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}(\boldsymbol{\Lambda_L})$ and $\boldsymbol{S}_{\mathsf{cvx}}=\boldsymbol{S}^{\star}+\boldsymbol{\Lambda}^{+}-\boldsymbol{\Lambda}^{-}$, we can invoke the triangle inequality to obtain

$$\begin{aligned}\lambda\left\|\boldsymbol{L}_{\mathsf{cvx}}\right\|_{*}+\tau\left\|\boldsymbol{S}_{\mathsf{cvx}}\right\|_{1}&=\lambda\left\|\boldsymbol{L}^{\star}+\boldsymbol{\Lambda}^{-}+\boldsymbol{\Lambda}^{+}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}\left(\boldsymbol{\Lambda_L}\right)\right\|_{*}+\tau\left\|\boldsymbol{S}^{\star}-\boldsymbol{\Lambda}^{-}+\boldsymbol{\Lambda}^{+}\right\|_{1}\\&\ge\lambda\left\|\boldsymbol{L}^{\star}+\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}\left(\boldsymbol{\Lambda_L}\right)\right\|_{*}+\tau\left\|\boldsymbol{S}^{\star}-\boldsymbol{\Lambda}^{-}\right\|_{1}-\lambda\left\|\boldsymbol{\Lambda}^{+}\right\|_{*}-\tau\left\|\boldsymbol{\Lambda}^{+}\right\|_{1}.\end{aligned}$$

Adding the above two inequalities and using the fact $\mathsf{support}(\boldsymbol{\Lambda}^{-})\subseteq\Omega_{\mathrm{obs}}$ lead to

$$\begin{aligned}&\frac{\lambda}{4}\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}\left(\boldsymbol{\Lambda_L}\right)\right)\right\|_{*}+\frac{\tau}{4}\left\|\mathcal{P}_{(\Omega^{\star})^{\mathrm{c}}}\left(\boldsymbol{\Lambda}^{-}\right)\right\|_{1}=\frac{\lambda}{4}\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}\left(\boldsymbol{\Lambda_L}\right)\right)\right\|_{*}+\frac{\tau}{4}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\left(\boldsymbol{\Lambda}^{-}\right)\right\|_{1}\\&\quad\le\lambda\left\|\boldsymbol{L}_{\mathsf{cvx}}\right\|_{*}+\tau\left\|\boldsymbol{S}_{\mathsf{cvx}}\right\|_{1}+\lambda\left\|\boldsymbol{\Lambda}^{+}\right\|_{*}+\tau\left\|\boldsymbol{\Lambda}^{+}\right\|_{1}-\lambda\left\|\boldsymbol{L}^{\star}\right\|_{*}-\tau\left\|\boldsymbol{S}^{\star}\right\|_{1}\\&\quad\le\frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{L}^{\star}+\boldsymbol{S}^{\star}-\boldsymbol{M}\right)\right\|_{\mathrm{F}}^{2}-\frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{L}_{\mathsf{cvx}}+\boldsymbol{S}_{\mathsf{cvx}}-\boldsymbol{M}\right)\right\|_{\mathrm{F}}^{2}+\lambda\left\|\boldsymbol{\Lambda}^{+}\right\|_{*}+\tau\left\|\boldsymbol{\Lambda}^{+}\right\|_{1}\\&\quad\le\frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{E}\right)\right\|_{\mathrm{F}}^{2}+4\lambda\sqrt{n\log n}\left\|\boldsymbol{\Lambda}^{+}\right\|_{\mathrm{F}}.\end{aligned}\tag{D.9}$$

Here, the penultimate line results from the inequality (D.3) and last line follows from the same argument in obtaining (D.4).

We are now ready to establish the upper bound on α_3. Invoke the elementary inequalities $\|\boldsymbol{A}\|_{\mathrm{F}}\le\|\boldsymbol{A}\|_{*}$ and $\|\boldsymbol{A}\|_{\mathrm{F}}\le\|\boldsymbol{A}\|_{1}$ for any $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ to show that

$$\begin{aligned}\alpha_3&=\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}\left(\boldsymbol{\Lambda_L}\right)\right)\right\|_{\mathrm{F}}^{2}+\left\|\mathcal{P}_{(\Omega^{\star})^{\mathrm{c}}}\left(\boldsymbol{\Lambda}^{-}\right)\right\|_{\mathrm{F}}^{2}\le\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}\left(\boldsymbol{\Lambda_L}\right)\right)\right\|_{*}^{2}+\left\|\mathcal{P}_{(\Omega^{\star})^{\mathrm{c}}}\left(\boldsymbol{\Lambda}^{-}\right)\right\|_{1}^{2}\\&\le\left(\frac{16}{\lambda^{2}}+\frac{16}{\tau^{2}}\right)\left(\frac{\lambda}{4}\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{\Lambda}^{-}+\mathcal{P}_{\Omega_{\mathsf{obs}}^{\mathrm{c}}}\left(\boldsymbol{\Lambda_L}\right)\right)\right\|_{*}+\frac{\tau}{4}\left\|\mathcal{P}_{(\Omega^{\star})^{\mathrm{c}}}\left(\boldsymbol{\Lambda}^{-}\right)\right\|_{1}\right)^{2}.\end{aligned}$$

This combined with (D.9) allows us to obtain

$$\alpha_3\le\left(\frac{16}{\lambda^{2}}+\frac{16}{\tau^{2}}\right)\left(\frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{E}\right)\right\|_{\mathrm{F}}^{2}+4\lambda\sqrt{n}\left\|\boldsymbol{\Lambda}^{+}\right\|_{\mathrm{F}}\right)^{2}\le\left(\frac{32}{\lambda^{2}}+\frac{32}{\tau^{2}}\right)\left(\frac{1}{4}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{E}\right)\right\|_{\mathrm{F}}^{4}+16\lambda^{2}n\log n\left\|\boldsymbol{\Lambda}^{+}\right\|_{\mathrm{F}}^{2}\right),$$

where we have used the elementary inequality $(a+b)^2\le 2a^2+2b^2$. Recalling that $\tau=\lambda/\sqrt{np/\log n}$ and that $np\ge 1$, we arrive at

$$\alpha_3\le\frac{64np}{\lambda^{2}}\left(\frac{1}{4}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{E}\right)\right\|_{\mathrm{F}}^{4}+16\lambda^{2}n\log n\left\|\boldsymbol{\Lambda}^{+}\right\|_{\mathrm{F}}^{2}\right)=\frac{16np}{\lambda^{2}}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{E}\right)\right\|_{\mathrm{F}}^{4}+2^{9}n^{2}p\alpha_{1}\log n,\tag{D.10}$$

where we have identified $2\|\boldsymbol{\Lambda}^{+}\|_{\mathrm{F}}^{2}$ with α_1.

Step 4: putting the above bounds on $\alpha_1, \alpha_2, \alpha_3$ together. Taking the preceding bounds on α_1, α_2 and α_3 collectively yields

$$\begin{aligned}
\|\mathbf{\Lambda}_{\boldsymbol{L}}\|_{\mathrm{F}}^2 + \|\mathbf{\Lambda}_{\boldsymbol{S}}\|_{\mathrm{F}}^2 = \alpha_1 + \alpha_2 + \alpha_3 &\overset{\text{(i)}}{\leq} \alpha_1 + \left(1 + \frac{8}{p}\right)\alpha_3 \overset{\text{(ii)}}{\leq} \alpha_1 + \frac{16}{p}\alpha_3 \\
&\overset{\text{(iii)}}{\leq} \left(2^{13} n^2 \log n + 1\right)\alpha_1 + \frac{2^8 n}{\lambda^2}\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}^4 \\
&\overset{\text{(iv)}}{\leq} \left(2^{13} n^2 \log n + 1\right)\left[2\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}^2 + 4\lambda\sqrt{n\log n}\sqrt{\|\mathbf{\Lambda}_{\boldsymbol{L}}\|_{\mathrm{F}}^2 + \|\mathbf{\Lambda}_{\boldsymbol{S}}\|_{\mathrm{F}}^2}\right] + \frac{2^8 n}{\lambda^2}\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}^4 .
\end{aligned}$$

Here, the first inequality (i) comes from (D.7), the second inequality (ii) follows from the fact $1 \leq 8/p$, the third relation (iii) is a consequence of (D.10), and the last line (iv) results from (D.5). Note that this forms a quadratic inequality in $\sqrt{\|\mathbf{\Lambda}_{\boldsymbol{L}}\|_{\mathrm{F}}^2 + \|\mathbf{\Lambda}_{\boldsymbol{S}}\|_{\mathrm{F}}^2}$. Solving the inequality yields the claimed bound

$$\|\mathbf{\Lambda}_{\boldsymbol{L}}\|_{\mathrm{F}}^2 + \|\mathbf{\Lambda}_{\boldsymbol{S}}\|_{\mathrm{F}}^2 \lesssim \lambda^2 n^5 \log^3 n + n^2 \log n \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}^2 + \frac{n}{\lambda^2}\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}^4 .$$

Further, the elementary inequality $a^2 + b^2 \geq 2ab$ yields

$$\lambda^2 n^5 \log^3 n + \frac{n}{\lambda^2}\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}^4 \geq 2n^3 \log^{3/2} n \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}^2 \geq n^2 \log n \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}^2 ,$$

leading to the simplified bound

$$\|\mathbf{\Lambda}_{\boldsymbol{L}}\|_{\mathrm{F}}^2 + \|\mathbf{\Lambda}_{\boldsymbol{S}}\|_{\mathrm{F}}^2 \lesssim \lambda^2 n^5 \log^3 n + \frac{n}{\lambda^2}\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\|_{\mathrm{F}}^4 .$$

Step 5: checking the conditions in Lemmas 6 and 7. We are left with proving that the conditions in Lemmas 6 and 7 hold with high probability. In view of Lemma 3 and Corollary 3, the conditions $\|\mathcal{P}_{T^\star} - p^{-1}\mathcal{P}_{T^\star}\mathcal{P}_{\Omega_{\mathsf{obs}}}\mathcal{P}_{T^\star}\| \leq 1/2$ and $\|\mathcal{P}_{\Omega^\star}\mathcal{P}_{T^\star}\|^2 \leq p/8$ hold with high probability, provided that $n^2 p \gg \mu r n \log n$ and $\rho_{\mathsf{s}} \leq 1/12$. In addition, Lemma 3 ensures that $\|\mathcal{P}_{T^\star} - p^{-1}(1-\rho_{\mathsf{s}})^{-1}\mathcal{P}_{T^\star}\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}\mathcal{P}_{T^\star}\| \leq 1/2$ holds with high probability, with the proviso that $n^2 p(1-\rho_{\mathsf{s}}) \gg \mu r n \log n$, which holds true under the assumptions $\rho_{\mathsf{s}} \leq 1/12$ and $n^2 p \gg \mu r n \log n$. Last but not least, the existence of the dual certificate $\boldsymbol{W}$ obeying (D.8) is guaranteed with high probability according to [CJSC13, Section III.D], under the conditions $\rho_{\mathsf{s}} \ll 1$ and $n^2 p \gg \mu^2 r^2 n \log^6 n$.[10]

D.1 Proof of Lemma 6

Expand $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\mathcal{P}_{T^\star}(\boldsymbol{A}) + \mathcal{P}_{\Omega^\star}(\boldsymbol{B}))\|_{\mathrm{F}}^2$ to obtain

$$\begin{aligned}
\|\mathcal{P}_{\Omega_{\mathsf{obs}}}[\mathcal{P}_{T^\star}(\boldsymbol{A}) + \mathcal{P}_{\Omega^\star}(\boldsymbol{B})]\|_{\mathrm{F}}^2 &= \|\mathcal{P}_{\Omega_{\mathsf{obs}}}\mathcal{P}_{T^\star}(\boldsymbol{A})\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\Omega^\star}(\boldsymbol{B})\|_{\mathrm{F}}^2 + 2\langle\mathcal{P}_{\Omega_{\mathsf{obs}}}\mathcal{P}_{T^\star}(\boldsymbol{A}), \mathcal{P}_{\Omega^\star}(\boldsymbol{B})\rangle \\
&\geq \frac{p}{2}\|\mathcal{P}_{T^\star}(\boldsymbol{A})\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\Omega^\star}(\boldsymbol{B})\|_{\mathrm{F}}^2 + 2\langle\mathcal{P}_{\Omega_{\mathsf{obs}}}\mathcal{P}_{T^\star}(\boldsymbol{A}), \mathcal{P}_{\Omega^\star}(\boldsymbol{B})\rangle .
\end{aligned}$$

Here, the equality uses the fact $\Omega^\star \subseteq \Omega_{\mathsf{obs}}$, and the inequality holds because of the assumption $\|\mathcal{P}_{T^\star} - p^{-1}\mathcal{P}_{T^\star}\mathcal{P}_{\Omega_{\mathsf{obs}}}\mathcal{P}_{T^\star}\| \leq 1/2$ and Fact 1. Use $\Omega^\star \subseteq \Omega_{\mathsf{obs}}$ once again to obtain

$$\begin{aligned}
2\langle\mathcal{P}_{\Omega_{\mathsf{obs}}}\mathcal{P}_{T^\star}(\boldsymbol{A}), \mathcal{P}_{\Omega^\star}(\boldsymbol{B})\rangle &= 2\langle\mathcal{P}_{\Omega^\star}\mathcal{P}_{T^\star}\mathcal{P}_{T^\star}(\boldsymbol{A}), \mathcal{P}_{\Omega^\star}(\boldsymbol{B})\rangle \\
&\geq -2\|\mathcal{P}_{\Omega^\star}\mathcal{P}_{T^\star}\|\|\mathcal{P}_{T^\star}(\boldsymbol{A})\|_{\mathrm{F}}\|\mathcal{P}_{\Omega^\star}(\boldsymbol{B})\|_{\mathrm{F}} \\
&\geq -2\|\mathcal{P}_{\Omega^\star}\mathcal{P}_{T^\star}\|^2\|\mathcal{P}_{T^\star}(\boldsymbol{A})\|_{\mathrm{F}}^2 - \frac{1}{2}\|\mathcal{P}_{\Omega^\star}(\boldsymbol{B})\|_{\mathrm{F}}^2 .
\end{aligned}$$

Here, the last relation arises from the elementary inequality $ab \leq (a^2+b^2)/2$ and the fact that $\|\mathcal{P}_{\Omega^\star}\mathcal{P}_{T^\star}\| \leq 1$. Combine the above two inequalities to obtain

$$\|\mathcal{P}_{\Omega_{\mathsf{obs}}}[\mathcal{P}_{T^\star}(\boldsymbol{A}) + \mathcal{P}_{\Omega^\star}(\boldsymbol{B})]\|_{\mathrm{F}}^2 \geq \left(\frac{p}{2} - 2\|\mathcal{P}_{\Omega^\star}\mathcal{P}_{T^\star}\|^2\right)\|\mathcal{P}_{T^\star}(\boldsymbol{A})\|_{\mathrm{F}}^2 + \frac{1}{2}\|\mathcal{P}_{\Omega^\star}(\boldsymbol{B})\|_{\mathrm{F}}^2$$

[10] Note that [CJSC13, Section III.D] requires $n^2 p \gg \max\{\mu, \mu_2\} r n \log^6 n$ under an additional incoherence condition $\|\boldsymbol{U}^\star\boldsymbol{V}^{\star\top}\|_\infty \leq \sqrt{\mu_2 r/n^2}$. While we do not impose this extra condition, it is easily seen that $\|\boldsymbol{U}^\star\boldsymbol{V}^{\star\top}\|_\infty \leq \|\boldsymbol{U}^\star\|_{2,\infty}\|\boldsymbol{V}^\star\|_{2,\infty} \leq \mu r/n$ and hence $\mu_2 \leq \mu^2 r$.

$$\geq \frac{p}{4}\left\|\mathcal{P}_{T^\star}(\boldsymbol{A})\right\|_{\mathrm{F}}^2+\frac{1}{2}\left\|\mathcal{P}_{\Omega^\star}(\boldsymbol{B})\right\|_{\mathrm{F}}^2$$
$$\geq \frac{p}{4}\left(\left\|\mathcal{P}_{T^\star}(\boldsymbol{A})\right\|_{\mathrm{F}}^2+\left\|\mathcal{P}_{\Omega^\star}(\boldsymbol{B})\right\|_{\mathrm{F}}^2\right)$$

as claimed, where we have used the assumption $\|\mathcal{P}_{\Omega^\star}\mathcal{P}_{T^\star}\|^2 \leq p/8$ in the middle line and the fact $1/2 \geq p/4$ in the last inequality.

D.2 Proof of Lemma 7

In view of the convexity of the nuclear norm $\|\cdot\|_*$, one has

$$\|\boldsymbol{L}^\star+\boldsymbol{H}_{\boldsymbol{L}}\|_* \geq \|\boldsymbol{L}^\star\|_* + \left\langle \boldsymbol{U}^\star\boldsymbol{V}^{\star\top}+\boldsymbol{G}_1, \boldsymbol{H}_{\boldsymbol{L}}\right\rangle = \|\boldsymbol{L}^\star\|_* + \left\langle \boldsymbol{U}^\star\boldsymbol{V}^{\star\top}, \boldsymbol{H}_{\boldsymbol{L}}\right\rangle + \|\mathcal{P}_{T^{\star\perp}}(\boldsymbol{H}_{\boldsymbol{L}})\|_*.$$

Here, $\boldsymbol{U}^\star\boldsymbol{V}^{\star\top}+\boldsymbol{G}_1$ is a sub-gradient of $\|\cdot\|_*$ at $\boldsymbol{L}^\star$. The last identity holds by choosing $\boldsymbol{G}_1$ such that $\langle \boldsymbol{G}_1, \boldsymbol{H}_{\boldsymbol{L}}\rangle = \|\mathcal{P}_{T^{\star\perp}}(\boldsymbol{H}_{\boldsymbol{L}})\|_*$. Similarly, using the assumption $\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{H}_{\boldsymbol{L}})+\boldsymbol{H}_{\boldsymbol{S}}=\boldsymbol{0}$ and the convexity of the ℓ_1 norm $\|\cdot\|_1$, we can obtain

$$\begin{aligned}
\|\boldsymbol{S}^\star+\boldsymbol{H}_{\boldsymbol{S}}\|_1 &= \|\boldsymbol{S}^\star-\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{H}_{\boldsymbol{L}})\|_1 \\
&\geq \|\boldsymbol{S}^\star\|_1 - \langle \mathsf{sign}(\boldsymbol{S}^\star)+\boldsymbol{G}_2, \mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{H}_{\boldsymbol{L}})\rangle \\
&\overset{\text{(i)}}{=} \|\boldsymbol{S}^\star\|_1 - \langle \mathsf{sign}(\boldsymbol{S}^\star), \mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{H}_{\boldsymbol{L}})\rangle + \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}(\boldsymbol{H}_{\boldsymbol{L}})\right\|_1 \\
&\overset{\text{(ii)}}{=} \|\boldsymbol{S}^\star\|_1 - \langle \mathsf{sign}(\boldsymbol{S}^\star), \boldsymbol{H}_{\boldsymbol{L}}\rangle + \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}(\boldsymbol{H}_{\boldsymbol{L}})\right\|_1,
\end{aligned}$$

where $\mathsf{sign}(\boldsymbol{S}^\star) := [\mathsf{sign}(S^\star_{ij})]_{1\leq i,j\leq n}$, and $\mathsf{sign}(\boldsymbol{S}^\star)+\boldsymbol{G}_2$ is a sub-gradient of $\|\cdot\|_1$ at $\boldsymbol{S}^\star$. The first equality (i) holds by choosing $\boldsymbol{G}_2$ such that $-\langle \boldsymbol{G}_2, \mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{H}_{\boldsymbol{L}})\rangle = \|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}(\boldsymbol{H}_{\boldsymbol{L}})\|_1$, and the last relation (ii) arises since $\mathsf{sign}(\boldsymbol{S}^\star)$ is supported on $\Omega^\star \subseteq \Omega_{\mathsf{obs}}$. Combine the above two bounds to deduce that

$$\begin{aligned}
\Delta &:= \lambda\|\boldsymbol{L}^\star+\boldsymbol{H}_{\boldsymbol{L}}\|_* + \tau\|\boldsymbol{S}^\star+\boldsymbol{H}_{\boldsymbol{S}}\|_1 - \lambda\|\boldsymbol{L}^\star\|_* - \tau\|\boldsymbol{S}^\star\|_1 \\
&\geq \lambda\left\langle \boldsymbol{U}^\star\boldsymbol{V}^{\star\top}, \boldsymbol{H}_{\boldsymbol{L}}\right\rangle + \lambda\|\mathcal{P}_{T^{\star\perp}}(\boldsymbol{H}_{\boldsymbol{L}})\|_* - \tau\langle \mathsf{sgn}(\boldsymbol{S}^\star), \boldsymbol{H}_{\boldsymbol{L}}\rangle + \tau\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}(\boldsymbol{H}_{\boldsymbol{L}})\right\|_1 \\
&= \left\langle \lambda\boldsymbol{U}^\star\boldsymbol{V}^{\star\top} - \tau\mathsf{sgn}(\boldsymbol{S}^\star), \boldsymbol{H}_{\boldsymbol{L}}\right\rangle + \lambda\|\mathcal{P}_{T^{\star\perp}}(\boldsymbol{H}_{\boldsymbol{L}})\|_* + \tau\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}(\boldsymbol{H}_{\boldsymbol{L}})\right\|_1 \\
&= \underbrace{\left\langle \lambda\boldsymbol{U}^\star\boldsymbol{V}^{\star\top} - \tau\mathsf{sgn}(\boldsymbol{S}^\star) - \lambda\boldsymbol{W}, \boldsymbol{H}_{\boldsymbol{L}}\right\rangle}_{=:\theta_1} + \underbrace{\left\langle \lambda\boldsymbol{W}, \boldsymbol{H}_{\boldsymbol{L}}\right\rangle}_{=:\theta_2} + \lambda\|\mathcal{P}_{T^{\star\perp}}(\boldsymbol{H}_{\boldsymbol{L}})\|_* + \tau\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}(\boldsymbol{H}_{\boldsymbol{L}})\right\|_1, \qquad \text{(D.11)}
\end{aligned}$$

where $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ is the dual certificate stated in Lemma 7.

In what follows, we shall lower bound the right-hand side of (D.11). To begin with, for θ_1 we have

$$\begin{aligned}
\theta_1 &= \left\langle \mathcal{P}_{T^\star}\left[\lambda\boldsymbol{U}^\star\boldsymbol{V}^{\star\top} - \tau\mathsf{sign}(\boldsymbol{S}^\star) - \lambda\boldsymbol{W}\right], \mathcal{P}_{T^\star}(\boldsymbol{H}_{\boldsymbol{L}})\right\rangle + \left\langle \mathcal{P}_{T^{\star\perp}}\left[\lambda\boldsymbol{U}^\star\boldsymbol{V}^{\star\top} - \tau\mathsf{sign}(\boldsymbol{S}^\star) - \lambda\boldsymbol{W}\right], \mathcal{P}_{T^{\star\perp}}(\boldsymbol{H}_{\boldsymbol{L}})\right\rangle \\
&= \left\langle \mathcal{P}_{T^\star}\left[\lambda\boldsymbol{U}^\star\boldsymbol{V}^{\star\top} - \tau\mathsf{sign}(\boldsymbol{S}^\star) - \lambda\boldsymbol{W}\right], \mathcal{P}_{T^\star}(\boldsymbol{H}_{\boldsymbol{L}})\right\rangle - \left\langle \mathcal{P}_{T^{\star\perp}}\left[\tau\mathsf{sign}(\boldsymbol{S}^\star) + \lambda\boldsymbol{W}\right], \mathcal{P}_{T^{\star\perp}}(\boldsymbol{H}_{\boldsymbol{L}})\right\rangle \\
&\geq -\left\|\mathcal{P}_{T^\star}\left[\lambda\boldsymbol{U}^\star\boldsymbol{V}^{\star\top} - \tau\mathsf{sign}(\boldsymbol{S}^\star) - \lambda\boldsymbol{W}\right]\right\|_{\mathrm{F}}\|\mathcal{P}_{T^\star}(\boldsymbol{H}_{\boldsymbol{L}})\|_{\mathrm{F}} - \left\|\mathcal{P}_{T^{\star\perp}}\left[\tau\mathsf{sign}(\boldsymbol{S}^\star) + \lambda\boldsymbol{W}\right]\right\|\|\mathcal{P}_{T^{\star\perp}}(\boldsymbol{H}_{\boldsymbol{L}})\|_* \\
&\geq -\frac{\tau}{\sqrt{n}}\|\mathcal{P}_{T^\star}(\boldsymbol{H}_{\boldsymbol{L}})\|_{\mathrm{F}} - \frac{\lambda}{2}\|\mathcal{P}_{T^{\star\perp}}(\boldsymbol{H}_{\boldsymbol{L}})\|_*.
\end{aligned}$$

Here, the penultimate line uses the fact $\boldsymbol{U}^\star\boldsymbol{V}^{\star\top} \in T^\star$ and the elementary inequalities $|\langle \boldsymbol{A}, \boldsymbol{B}\rangle| \leq \|\boldsymbol{A}\|_{\mathrm{F}}\|\boldsymbol{B}\|_{\mathrm{F}}$ and $|\langle \boldsymbol{A}, \boldsymbol{B}\rangle| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|_*$, whereas the last inequality relies on the properties of the dual certificate $\boldsymbol{W}$, namely, (D.8a) and (D.8b). Moving on to θ_2, one has

$$\begin{aligned}
\theta_2 &= \left\langle \lambda\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}(\boldsymbol{W}), \boldsymbol{H}_{\boldsymbol{L}}\right\rangle + \left\langle \lambda\mathcal{P}_{(\Omega_{\mathsf{obs}}\setminus\Omega^\star)^c}(\boldsymbol{W}), \boldsymbol{H}_{\boldsymbol{L}}\right\rangle \\
&= \left\langle \lambda\boldsymbol{W}, \mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}(\boldsymbol{H}_{\boldsymbol{L}})\right\rangle \overset{\text{(i)}}{\geq} -\|\lambda\boldsymbol{W}\|_\infty\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}(\boldsymbol{H}_{\boldsymbol{L}})\right\|_1 \overset{\text{(ii)}}{\geq} -\frac{\tau}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}(\boldsymbol{H}_{\boldsymbol{L}})\right\|_1.
\end{aligned}$$

Here, the second identity uses the assumption (D.8c), the first inequality (i) uses the elementary inequality $|\langle \boldsymbol{A}, \boldsymbol{B}\rangle| \leq \|\boldsymbol{A}\|_\infty\|\boldsymbol{B}\|_1$, and the last relation (ii) holds because of the assumption (D.8d). Substituting the above two bounds back into (D.11) gives

$$\Delta \geq -\frac{\tau}{\sqrt{n}}\|\mathcal{P}_{T^\star}(\boldsymbol{H}_{\boldsymbol{L}})\|_{\mathrm{F}} + \frac{\lambda}{2}\|\mathcal{P}_{T^{\star\perp}}(\boldsymbol{H}_{\boldsymbol{L}})\|_* + \frac{\tau}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^\star}(\boldsymbol{H}_{\boldsymbol{L}})\right\|_1. \qquad \text{(D.12)}$$

Continuing the lower bound, we have

$$\begin{aligned}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{1} &\overset{\text{(i)}}{\geq}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}}=\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\mathcal{P}_{T^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)+\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}}\\ &\overset{\text{(ii)}}{\geq}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\mathcal{P}_{T^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}}-\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}}\\ &\geq\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\mathcal{P}_{T^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}}-\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}},\end{aligned}$$

where (i) holds because $\|\boldsymbol{A}\|_1 \geq \|\boldsymbol{A}\|_{\mathrm{F}}$ for any matrix $\boldsymbol{A}$, and (ii) arises from the triangle inequality. Putting the above relation and (D.12) together results in

$$\begin{aligned}\Delta &\geq -\frac{\tau}{\sqrt{n}}\left\|\mathcal{P}_{T^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}}+\left(\frac{\lambda}{2}-\frac{\tau}{4}\right)\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{*}+\frac{\tau}{4}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\mathcal{P}_{T^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}}+\frac{\tau}{4}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{1}\\ &\geq\frac{\tau}{4}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\mathcal{P}_{T^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}}-\frac{\tau}{\sqrt{n}}\left\|\mathcal{P}_{T^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}}+\frac{\lambda}{4}\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{*}+\frac{\tau}{4}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{1},\end{aligned}\tag{D.13}$$

where the last line holds since $\lambda=\tau\sqrt{np/\log n}\geq\tau$ (as long as $np\geq\log n$). Everything then boils down to lower bounding $\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\mathcal{P}_{T^{\star}}(\boldsymbol{H}_{\boldsymbol{L}})\|_{\mathrm{F}}$. To this end, one can use the assumption $\|\mathcal{P}_{T^{\star}}-p^{-1}(1-\rho_{\mathsf{s}})^{-1}\mathcal{P}_{T^{\star}}\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\mathcal{P}_{T^{\star}}\|\leq 1/2$ and Fact 1 to obtain

$$\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\mathcal{P}_{T^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}}^{2}\geq\frac{1}{2}p\left(1-\rho_{\mathsf{s}}\right)\left\|\mathcal{P}_{T^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}}^{2}.\tag{D.14}$$

Take (D.13) and (D.14) collectively to yield

$$\begin{aligned}\Delta &\geq\left(\frac{\tau}{4}\sqrt{\frac{1}{2}p\left(1-\rho_{\mathsf{s}}\right)}-\frac{\tau}{\sqrt{n}}\right)\left\|\mathcal{P}_{T^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{\mathrm{F}}+\frac{\lambda}{4}\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{*}+\frac{\tau}{4}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{1}\\ &\geq\frac{\lambda}{4}\left\|\mathcal{P}_{T^{\star\perp}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{*}+\frac{\tau}{4}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}\left(\boldsymbol{H}_{\boldsymbol{L}}\right)\right\|_{1},\end{aligned}$$

where the last relation is guaranteed by $np\gg 1$ and $\rho_{\mathsf{s}}\ll 1$. Recognizing that $\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}(\boldsymbol{H}_{\boldsymbol{L}})=-\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega^{\star}}(\boldsymbol{H}_{\boldsymbol{S}})$ finishes the proof.

E Equivalence between convex and nonconvex solutions (Proof of Theorem 4)

The goal of this section is to establish the intimate connection between the convex and nonconvex solutions (cf. Theorem 4). Before continuing, we remind the readers of the following notations:

- $\boldsymbol{X}\boldsymbol{Y}^{\top}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$: the rank-$r$ singular value decomposition of $\boldsymbol{X}\boldsymbol{Y}^{\top}$;
- T: the tangent space of the set of rank-r matrices at the estimate $\boldsymbol{X}\boldsymbol{Y}^{\top}$.

In addition, we define

$$\boldsymbol{\Delta}_{\boldsymbol{L}}\coloneqq\boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{X}\boldsymbol{Y}^{\top},\qquad\boldsymbol{\Delta}_{\boldsymbol{S}}\coloneqq\boldsymbol{S}_{\mathsf{cvx}}-\boldsymbol{S},\tag{E.1}$$

and denote the support of $\boldsymbol{S}$ by

$$\Omega\coloneqq\{(i,j)\,|\,S_{ij}\neq 0\}.\tag{E.2}$$

E.1 Preliminary facts

We begin with two useful lemmas which demonstrate that the point $(\boldsymbol{X}\boldsymbol{Y}^{\top},\boldsymbol{S})$ described in Theorem 4 satisfies approximate optimality conditions w.r.t. the convex program (1.3).

Lemma 8. *Instate the assumptions in Theorem 4. The triple $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S})$ as stated in Theorem 4 satisfies*

$$\frac{1}{\lambda}\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}+\boldsymbol{S}-\boldsymbol{M}\right)=-\boldsymbol{U}\boldsymbol{V}^{\top}+\boldsymbol{R}_{1} \tag{E.3}$$

for some matrix $\boldsymbol{R}_1 \in \mathbb{R}^{n\times n}$ obeying

$$\left\Vert \mathcal{P}_{T}\left(\boldsymbol{R}_{1}\right)\right\Vert _{\mathrm{F}}\lesssim\frac{\kappa p\left\Vert \nabla f\left(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{S}\right)\right\Vert _{\mathrm{F}}}{\lambda\sqrt{\sigma_{\min}}}\lesssim\frac{1}{n^{19}}\qquad\text{and}\qquad\left\Vert \mathcal{P}_{T^{\perp}}\left(\boldsymbol{R}_{1}\right)\right\Vert \leq\frac{1}{2}. \tag{E.4}$$

Proof. The proof can be straightforwardly adapted from [CCF+20, Claim 2] by replacing $\boldsymbol{E}$ therein with $\boldsymbol{E}+\boldsymbol{S}^{\star}-\boldsymbol{S}$. We omit it for the sake of brevity. $\square$

Lemma 9. *The point $(\boldsymbol{X}\boldsymbol{Y}^{\top}, \boldsymbol{S})$ as stated in Theorem 4 obeys*

$$\frac{1}{\tau}\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}+\boldsymbol{S}-\boldsymbol{M}\right)=-\mathsf{sign}\left(\boldsymbol{S}\right)+\boldsymbol{R}_{2} \tag{E.5}$$

for some matrix $\boldsymbol{R}_2 \in \mathbb{R}^{n\times n}$, where $\boldsymbol{R}_2$ satisfies

$$\mathcal{P}_{\Omega}\left(\boldsymbol{R}_{2}\right)=\boldsymbol{0}\qquad\text{and}\qquad\left\Vert \mathcal{P}_{\Omega^{\mathrm{c}}}(\boldsymbol{R}_{2})\right\Vert _{\infty}\leq1 \tag{E.6}$$

with Ω defined in (E.2).

Proof. By definition, one has $\boldsymbol{S}=\mathcal{P}_{\Omega_{\mathsf{obs}}}[\mathcal{S}_{\tau}(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^{\top})]$. Clearly, this is equivalent to saying that $\boldsymbol{S}$ is the unique minimizer of the following convex program

$$\boldsymbol{S}=\underset{\widehat{\boldsymbol{S}}\in\mathbb{R}^{n\times n}}{\arg\min}\ \frac{1}{2}\left\Vert \mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}+\widehat{\boldsymbol{S}}-\boldsymbol{M}\right)\right\Vert _{\mathrm{F}}^{2}+\frac{\lambda}{2}\left\Vert \boldsymbol{X}\right\Vert _{\mathrm{F}}^{2}+\frac{\lambda}{2}\left\Vert \boldsymbol{Y}\right\Vert _{\mathrm{F}}^{2}+\tau\big\Vert \widehat{\boldsymbol{S}}\big\Vert _{1}. \tag{E.7}$$

The claim of this lemma then follows from the optimality condition of this convex program (E.7). $\square$

Additionally, in view of the crude error bound (3.8) and Condition 1, the matrix $\boldsymbol{\Delta}_{\boldsymbol{L}}$ (cf. (E.1)) obeys

$$\begin{aligned}\left\Vert \boldsymbol{\Delta}_{\boldsymbol{L}}\right\Vert _{\mathrm{F}} & =\left\Vert \boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{X}\boldsymbol{Y}^{\top}\right\Vert _{\mathrm{F}}\leq\left\Vert \boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^{\star}\right\Vert _{\mathrm{F}}+\left\Vert \boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{L}^{\star}\right\Vert _{\mathrm{F}}\leq\left\Vert \boldsymbol{L}_{\mathsf{cvx}}-\boldsymbol{L}^{\star}\right\Vert _{\mathrm{F}}+n\left\Vert \boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{L}^{\star}\right\Vert _{\infty}\\ & \lesssim\sigma n^{4}+n\tau\asymp\sigma n^{4},\end{aligned} \tag{E.8}$$

where we use the the elementary inequality $\|\boldsymbol{A}\|_{\mathrm{F}}\leq n\|\boldsymbol{A}\|_{\infty}$ and the fact that $\tau\asymp\sigma\sqrt{\log n}$.

E.2 Proof of Theorem 4

We now present the proof of Theorem 4, which consists of three main steps:

1. Showing that $(\boldsymbol{X}\boldsymbol{Y}^{\top}, \boldsymbol{S})$ is not far from $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}})$ over Ω_{obs}, in the sense that $\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta}_{\boldsymbol{L}}+\boldsymbol{\Delta}_{\boldsymbol{S}})\approx\boldsymbol{0}$;
2. Showing that $\boldsymbol{\Delta}_{\boldsymbol{L}}$ (resp. $\boldsymbol{\Delta}_{\boldsymbol{S}}$) is extremely small outside the tangent space T (resp. the support $\Omega^{\star}$), and hence most of the energy of $\boldsymbol{\Delta}_{\boldsymbol{L}}$ (resp. $\boldsymbol{\Delta}_{\boldsymbol{S}}$) — if it is not vanishingly small — has to reside within T (resp. $\Omega^{\star}$);
3. Showing that $\boldsymbol{\Delta}_{\boldsymbol{S}}\approx\boldsymbol{0}$ and $\boldsymbol{\Delta}_{\boldsymbol{L}}\approx\boldsymbol{0}$, with the assistance of the preceding two steps.

In what follows, we shall detail each of these steps.

E.2.1 Step 1: showing that $\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S})\approx\boldsymbol{0}$

Since $(\boldsymbol{L}_{\mathsf{cvx}},\boldsymbol{S}_{\mathsf{cvx}})$ is the minimizer of the convex program (1.3), we have

$$\begin{aligned}&\tfrac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{L}_{\mathsf{cvx}}+\boldsymbol{S}_{\mathsf{cvx}}-\boldsymbol{M}\right)\right\|_{\mathrm{F}}^{2}+\lambda\left\|\boldsymbol{L}_{\mathsf{cvx}}\right\|_{*}+\tau\left\|\boldsymbol{S}_{\mathsf{cvx}}\right\|_{1}\\&\quad=\tfrac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}+\boldsymbol{\Delta_L}+\boldsymbol{S}+\boldsymbol{\Delta_S}-\boldsymbol{M}\right)\right\|_{\mathrm{F}}^{2}+\lambda\left\|\boldsymbol{X}\boldsymbol{Y}^{\top}+\boldsymbol{\Delta_L}\right\|_{*}+\tau\left\|\boldsymbol{S}+\boldsymbol{\Delta_S}\right\|_{1}\\&\quad\leq\tfrac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}+\boldsymbol{S}-\boldsymbol{M}\right)\right\|_{\mathrm{F}}^{2}+\lambda\left\|\boldsymbol{X}\boldsymbol{Y}^{\top}\right\|_{*}+\tau\left\|\boldsymbol{S}\right\|_{1}.\end{aligned}$$

Here, the equality arises from the relations $\boldsymbol{L}_{\mathsf{cvx}}=\boldsymbol{X}\boldsymbol{Y}^{\top}+\boldsymbol{\Delta_L}$ and $\boldsymbol{S}_{\mathsf{cvx}}=\boldsymbol{S}+\boldsymbol{\Delta_S}$. Expanding the squares and rearranging terms, we arrive at

$$\begin{aligned}\tfrac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}}^{2}\leq&-\left\langle\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}+\boldsymbol{S}-\boldsymbol{M}\right),\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S}\right\rangle+\lambda\left\|\boldsymbol{X}\boldsymbol{Y}^{\top}\right\|_{*}-\lambda\left\|\boldsymbol{X}\boldsymbol{Y}^{\top}+\boldsymbol{\Delta_L}\right\|_{*}\\&+\tau\left\|\boldsymbol{S}\right\|_{1}-\tau\left\|\boldsymbol{S}+\boldsymbol{\Delta_S}\right\|_{1}.\end{aligned}\tag{E.9}$$

In view of the convexity of the nuclear norm $\|\cdot\|_*$ and the ℓ_1 norm $\|\cdot\|_1$, one has

$$\left\|\boldsymbol{X}\boldsymbol{Y}^{\top}+\boldsymbol{\Delta_L}\right\|_{*}\geq\left\|\boldsymbol{X}\boldsymbol{Y}^{\top}\right\|_{*}+\left\langle\boldsymbol{U}\boldsymbol{V}^{\top}+\boldsymbol{W},\boldsymbol{\Delta_L}\right\rangle\overset{\text{(i)}}{=}\left\|\boldsymbol{X}\boldsymbol{Y}^{\top}\right\|_{*}+\left\langle\boldsymbol{U}\boldsymbol{V}^{\top},\boldsymbol{\Delta_L}\right\rangle+\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta_L})\right\|_{*};\tag{E.10a}$$

$$\left\|\boldsymbol{S}+\boldsymbol{\Delta_S}\right\|_{1}\geq\left\|\boldsymbol{S}\right\|_{1}+\left\langle\mathsf{sign}\left(\boldsymbol{S}\right)+\boldsymbol{G},\boldsymbol{\Delta_S}\right\rangle\overset{\text{(ii)}}{=}\left\|\boldsymbol{S}\right\|_{1}+\left\langle\mathsf{sign}\left(\boldsymbol{S}\right),\boldsymbol{\Delta_S}\right\rangle+\left\|\mathcal{P}_{\Omega^{c}}(\boldsymbol{\Delta_S})\right\|_{1}.\tag{E.10b}$$

Here, $\boldsymbol{U}\boldsymbol{V}^{\top}+\boldsymbol{W}$ is a sub-gradient of $\|\cdot\|_*$ at $\boldsymbol{X}\boldsymbol{Y}^{\top}$. The identity (i) holds by choosing $\boldsymbol{W}$ such that $\langle\boldsymbol{W},\boldsymbol{\Delta_L}\rangle=\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta_L})\|_*$. Similarly, $\mathsf{sign}(\boldsymbol{S})+\boldsymbol{G}$ is a sub-gradient of $\|\cdot\|_1$ at $\boldsymbol{S}$ and one can choose $\boldsymbol{G}$ obeying $\langle\boldsymbol{G},\boldsymbol{\Delta_S}\rangle=\|\mathcal{P}_{\Omega^c}(\boldsymbol{\Delta_S})\|_1$ to make (ii) valid. These taken together with (E.9) lead to

$$\begin{aligned}\tfrac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}}^{2}\leq&-\left\langle\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}+\boldsymbol{S}-\boldsymbol{M}\right),\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S}\right\rangle-\lambda\left\langle\boldsymbol{U}\boldsymbol{V}^{\top},\boldsymbol{\Delta_L}\right\rangle-\lambda\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{\Delta_L}\right)\right\|_{*}\\&-\tau\left\langle\mathsf{sign}\left(\boldsymbol{S}\right),\boldsymbol{\Delta_S}\right\rangle-\tau\left\|\mathcal{P}_{\Omega^{c}}\left(\boldsymbol{\Delta_S}\right)\right\|_{1}.\end{aligned}$$

Recall the definitions of $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ from Lemmas 8 and 9. We can then simplify the above inequality as

$$\tfrac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}}^{2}\leq\underbrace{-\lambda\left\langle\boldsymbol{R}_1,\boldsymbol{\Delta_L}\right\rangle-\lambda\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta_L})\right\|_{*}}_{=:\theta_1}\underbrace{-\tau\left\langle\boldsymbol{R}_2,\boldsymbol{\Delta_S}\right\rangle-\tau\left\|\mathcal{P}_{\Omega^{c}}(\boldsymbol{\Delta_S})\right\|_{1}}_{=:\theta_2}.\tag{E.11}$$

In the sequel, we develop bounds on θ_1 and θ_2.

1. With regards to θ_1, one can further decompose it into

$$\begin{aligned}\theta_1&=-\lambda\left\langle\boldsymbol{R}_1,\mathcal{P}_T\left(\boldsymbol{\Delta_L}\right)\right\rangle-\lambda\left\langle\boldsymbol{R}_1,\mathcal{P}_{T^{\perp}}\left(\boldsymbol{\Delta_L}\right)\right\rangle-\lambda\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{\Delta_L}\right)\right\|_{*}\\&\leq\lambda\left\|\mathcal{P}_T\left(\boldsymbol{R}_1\right)\right\|_{\mathrm{F}}\left\|\mathcal{P}_T\left(\boldsymbol{\Delta_L}\right)\right\|_{\mathrm{F}}-\lambda\left(1-\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{R}_1\right)\right\|\right)\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{\Delta_L}\right)\right\|_{*}\\&\leq\lambda\left\|\mathcal{P}_T\left(\boldsymbol{R}_1\right)\right\|_{\mathrm{F}}\left\|\mathcal{P}_T\left(\boldsymbol{\Delta_L}\right)\right\|_{\mathrm{F}}-\frac{\lambda}{2}\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{\Delta_L}\right)\right\|_{*},\end{aligned}\tag{E.12}$$

   where the middle line arises from the elementary inequalities $|\langle\boldsymbol{A},\boldsymbol{B}\rangle|\leq\|\boldsymbol{A}\|_{\mathrm{F}}\|\boldsymbol{B}\|_{\mathrm{F}}$ and $|\langle\boldsymbol{A},\boldsymbol{B}\rangle|\leq\|\boldsymbol{A}\|\|\boldsymbol{B}\|_*$, and the last inequality holds since $\|\mathcal{P}_{T^{\perp}}(\boldsymbol{R}_1)\|\leq 1/2$ (see Lemma 8).

2. Similarly, one can decompose θ_2 into

$$\begin{aligned}\theta_2&=-\tau\left\langle\boldsymbol{R}_2,\mathcal{P}_{\Omega}\left(\boldsymbol{\Delta_S}\right)\right\rangle-\tau\left\langle\boldsymbol{R}_2,\mathcal{P}_{\Omega^c}\left(\boldsymbol{\Delta_S}\right)\right\rangle-\tau\left\|\mathcal{P}_{\Omega^c}\left(\boldsymbol{\Delta_S}\right)\right\|_{1}\\&\leq\tau\left\langle\mathcal{P}_{\Omega}\left(\boldsymbol{R}_2\right),\mathcal{P}_{\Omega}\left(\boldsymbol{\Delta_S}\right)\right\rangle-\tau\left(1-\left\|\mathcal{P}_{\Omega^c}\left(\boldsymbol{R}_2\right)\right\|_{\infty}\right)\left\|\mathcal{P}_{\Omega^c}\left(\boldsymbol{\Delta_S}\right)\right\|_{1}\leq 0.\end{aligned}\tag{E.13}$$

   Here, the first inequality comes from the facts that $|\langle\boldsymbol{A},\boldsymbol{B}\rangle|\leq\|\boldsymbol{A}\|_{\infty}\|\boldsymbol{B}\|_1$ and $|\langle\boldsymbol{A},\boldsymbol{B}\rangle|\leq\|\boldsymbol{A}\|_{\mathrm{F}}\|\boldsymbol{B}\|_{\mathrm{F}}$, and the second one utilizes the facts that $\mathcal{P}_{\Omega}(\boldsymbol{R}_2)=\boldsymbol{0}$ and $\|\mathcal{P}_{\Omega^c}(\boldsymbol{R}_2)\|_{\infty}\leq 1$ (cf. Lemma 9).

Combining (E.11), (E.12) and (E.13) yields

$$\tfrac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}}^{2}\leq\lambda\left\|\mathcal{P}_T(\boldsymbol{R}_1)\right\|_{\mathrm{F}}\left\|\mathcal{P}_T(\boldsymbol{\Delta_L})\right\|_{\mathrm{F}}-\frac{\lambda}{2}\left\|\mathcal{P}_{T^{\perp}}(\boldsymbol{\Delta_L})\right\|_{*}\tag{E.14}$$

$$\lesssim \frac{\lambda}{n^{19}} \left\|\boldsymbol{\Delta_L}\right\|_{\mathrm{F}} \lesssim \frac{\sigma\sqrt{np}}{n^{19}}\sigma n^4 \lesssim \frac{\sigma^2}{n^{14.5}},$$

where we make use of the upper bound $\|\mathcal{P}_T(\boldsymbol{R}_1)\|_{\mathrm{F}} \lesssim n^{-19}$ (cf. Lemma 8), the choice $\lambda \asymp \sigma\sqrt{np}$ as well as the crude error bound $\|\boldsymbol{\Delta_L}\|_{\mathrm{F}} \lesssim \sigma n^4$ (cf. (E.8)). Consequently, we have demonstrated that $\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S}) \approx \mathbf{0}$ in the sense that

$$\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^{7.25}} \leq \frac{\sigma}{n^7}. \tag{E.15}$$

E.2.2 Step 2: showing that $\mathcal{P}_{T^\perp}(\boldsymbol{\Delta_L}) \approx \mathbf{0}$ and $\mathcal{P}_{(\Omega^\star)^{\mathrm{c}}}(\boldsymbol{\Delta_S}) \approx \mathbf{0}$

We begin by demonstrating that $\mathcal{P}_{T^\perp}(\boldsymbol{\Delta_L}) \approx \mathbf{0}$. From the inequality (E.14), we have

$$\begin{aligned}
\frac{1}{2}\left\|\mathcal{P}_{T^\perp}(\boldsymbol{\Delta_L})\right\|_* &\leq \|\mathcal{P}_T(\boldsymbol{R}_1)\|_{\mathrm{F}}\,\|\mathcal{P}_T(\boldsymbol{\Delta_L})\|_{\mathrm{F}} - \frac{1}{2\lambda}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}}^2 \\
&\leq \|\mathcal{P}_T(\boldsymbol{R}_1)\|_{\mathrm{F}}\,\|\mathcal{P}_T(\boldsymbol{\Delta_L})\|_{\mathrm{F}} \lesssim \frac{1}{n^{19}}\left\|\boldsymbol{\Delta_L}\right\|_{\mathrm{F}},
\end{aligned}$$

where the last inequality again results from the estimate $\|\mathcal{P}_T(\boldsymbol{R}_1)\|_{\mathrm{F}} \lesssim n^{-19}$ given in Lemma 8. Invoking the condition $\|\boldsymbol{\Delta_L}\|_{\mathrm{F}} \lesssim \sigma n^4$ (cf. (E.8)) yields

$$\left\|\mathcal{P}_{T^\perp}(\boldsymbol{\Delta_L})\right\|_{\mathrm{F}} \leq \left\|\mathcal{P}_{T^\perp}(\boldsymbol{\Delta_L})\right\|_* \lesssim \frac{1}{n^{19}}\sigma n^4 \lesssim \frac{\sigma}{n^{15}} \leq \frac{\sigma}{n^{14}}, \tag{E.16}$$

which demonstrates that the energy of $\boldsymbol{\Delta_L}$ outside T is extremely small.

We now move on to $\mathcal{P}_{(\Omega^\star)^{\mathrm{c}}}(\boldsymbol{\Delta_S})$. This term obeys

$$\mathcal{P}_{(\Omega^\star)^{\mathrm{c}}}(\boldsymbol{\Delta_S}) = \mathcal{P}_{\Omega_{\mathsf{obs}}\backslash\Omega^\star}(\boldsymbol{\Delta_S}),$$

where the relation holds since $\boldsymbol{\Delta_S}$ is supported on Ω_{obs}. To facilitate the analysis of $\Omega_{\mathsf{obs}}\backslash\Omega^\star$, we introduce another index subset

$$\Omega_1 \coloneqq \left\{(i,j) \in \Omega_{\mathsf{obs}} \,:\, |(\boldsymbol{\Delta_S})_{ij}| \leq \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S})\|_\infty\right\}. \tag{E.17}$$

The usefulness of Ω_1 can be seen through the following claim, whose claim is postponed to the end of this section.

Claim 1. Under Condition 1, we have

$$\Omega_{\mathsf{obs}}\backslash\Omega^\star \subseteq \Omega_1.$$

An immediate consequence of Claim 1 is that

$$\begin{aligned}
\left\|\mathcal{P}_{(\Omega^\star)^{\mathrm{c}}}\left(\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}} &= \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\backslash\Omega^\star}\left(\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}} \leq \left\|\mathcal{P}_{\Omega_1}\left(\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}} \\
&\leq n\left\|\mathcal{P}_{\Omega_1}\left(\boldsymbol{\Delta_S}\right)\right\|_\infty \leq n\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S}\right)\right\|_\infty \\
&\leq n\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}} \leq \frac{\sigma}{n^6},
\end{aligned} \tag{E.18}$$

which justifies our assertion that the energy of $\boldsymbol{\Delta_S}$ outside $\Omega^\star$ is extremely small. Here, the last inequality arises from (E.15).

E.2.3 Step 3: controlling the size of $\boldsymbol{\Delta_S}$ (and hence that of $\boldsymbol{\Delta_L}$)

In view of (E.15) and the triangle inequality, we have

$$\begin{aligned}
\frac{\sigma}{n^7} &\geq \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}} \\
&\geq \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\mathcal{P}_T\left(\boldsymbol{\Delta_L}\right)\right\|_{\mathrm{F}} - \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\mathcal{P}_{T^\perp}\left(\boldsymbol{\Delta_L}\right)\right\|_{\mathrm{F}} - \left\|\mathcal{P}_{\Omega^\star}\left(\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}} - \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}\backslash\Omega^\star}\left(\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}} \\
&\geq \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\mathcal{P}_T\left(\boldsymbol{\Delta_L}\right)\right\|_{\mathrm{F}} - \left\|\mathcal{P}_{\Omega^\star}\left(\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}} - \left\|\mathcal{P}_{T^\perp}\left(\boldsymbol{\Delta_L}\right)\right\|_{\mathrm{F}} - \left\|\mathcal{P}_{(\Omega^\star)^{\mathrm{c}}}\left(\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}} \\
&\geq \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\mathcal{P}_T(\boldsymbol{\Delta_L})\right\|_{\mathrm{F}} - \left\|\mathcal{P}_{\Omega^\star}\left(\boldsymbol{\Delta_S}\right)\right\|_{\mathrm{F}} - \frac{\sigma}{n^6} - \frac{\sigma}{n^{14}},
\end{aligned} \tag{E.19}$$

where the last step follows from (E.16) and (E.18). By Condition 2, we have

$$\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\mathcal{P}_T(\boldsymbol{\Delta_L})\|_{\mathrm{F}} \geq \sqrt{\frac{c_{\mathrm{inj}}}{\kappa}p}\,\|\mathcal{P}_T(\boldsymbol{\Delta_L})\|_{\mathrm{F}} \quad \text{and} \quad \|\mathcal{P}_{\Omega^\star}\mathcal{P}_T(\boldsymbol{\Delta_L})\|_{\mathrm{F}} \leq \frac{1}{2}\sqrt{\frac{c_{\mathrm{inj}}}{\kappa}p}\,\|\mathcal{P}_T(\boldsymbol{\Delta_L})\|_{\mathrm{F}},$$

given that $\mathcal{P}_T(\boldsymbol{\Delta_L}) \in T$. The latter inequality combined with (E.15) and (E.16) further gives

$$\begin{aligned}
\|\mathcal{P}_{\Omega^\star}(\boldsymbol{\Delta_S})\|_{\mathrm{F}} &\leq \|\mathcal{P}_{\Omega^\star}(\boldsymbol{\Delta_S}+\boldsymbol{\Delta_L})\|_{\mathrm{F}} + \|\mathcal{P}_{\Omega^\star}(\boldsymbol{\Delta_L})\|_{\mathrm{F}} \\
&\leq \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta_S}+\boldsymbol{\Delta_L})\|_{\mathrm{F}} + \|\mathcal{P}_{\Omega^\star}\mathcal{P}_T(\boldsymbol{\Delta_L})\|_{\mathrm{F}} + \|\mathcal{P}_{\Omega^\star}\mathcal{P}_{T^\perp}(\boldsymbol{\Delta_L})\|_{\mathrm{F}} \\
&\leq \frac{\sigma}{n^7} + \frac{1}{2}\sqrt{\frac{c_{\mathrm{inj}}}{\kappa}p}\,\|\mathcal{P}_T\boldsymbol{\Delta_L}\|_{\mathrm{F}} + \frac{\sigma}{n^{14}}.
\end{aligned}$$

Substituting the above bounds into (E.19) gives

$$\begin{aligned}
\frac{\sigma}{n^7} &\geq \sqrt{\frac{c_{\mathrm{inj}}}{\kappa}p}\,\|\mathcal{P}_T(\boldsymbol{\Delta_L})\|_{\mathrm{F}} - \frac{1}{2}\sqrt{\frac{c_{\mathrm{inj}}}{\kappa}p}\,\|\mathcal{P}_T(\boldsymbol{\Delta_L})\|_{\mathrm{F}} - \frac{\sigma}{n^7} - \frac{\sigma}{n^6} - \frac{2\sigma}{n^{14}} \\
&\geq \frac{1}{2}\sqrt{\frac{c_{\mathrm{inj}}}{\kappa}p}\,\|\mathcal{P}_T(\boldsymbol{\Delta_L})\|_{\mathrm{F}} - \frac{2\sigma}{n^6},
\end{aligned}$$

which further yields

$$\|\mathcal{P}_T(\boldsymbol{\Delta_L})\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^6}\sqrt{\frac{\kappa}{p}} \leq \frac{\sigma}{n^5},$$

provided that $n^2 p \gg \kappa$. This combined with (E.16) allows one to control the size of $\boldsymbol{\Delta_L}$:

$$\|\boldsymbol{\Delta_L}\|_{\mathrm{F}} \leq \|\mathcal{P}_T(\boldsymbol{\Delta_L})\|_{\mathrm{F}} + \|\mathcal{P}_{T^\perp}(\boldsymbol{\Delta_L})\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^5}.$$

In view of (E.15) and the fact that $\boldsymbol{\Delta_S}$ is supported on Ω_{obs}, we have

$$\begin{aligned}
\|\boldsymbol{\Delta_S}\|_{\mathrm{F}} = \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta_S})\|_{\mathrm{F}} &\leq \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S})\|_{\mathrm{F}} + \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta_L})\|_{\mathrm{F}} \\
&\leq \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S})\|_{\mathrm{F}} + \|\boldsymbol{\Delta_L}\|_{\mathrm{F}} \lesssim \frac{\sigma}{n^5},
\end{aligned}$$

thus concluding the proof.

E.2.4 Proof of Claim 1

We first recall the facts that

$$\boldsymbol{S} = \mathcal{P}_{\Omega_{\mathsf{obs}}}\left[\mathcal{S}_\tau\left(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^\top\right)\right] \qquad \text{and} \qquad \boldsymbol{S}+\boldsymbol{\Delta_S} = \mathcal{P}_{\Omega_{\mathsf{obs}}}\left[\mathcal{S}_\tau\left(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{\Delta_L}\right)\right],$$

where the second identity follows since $(\boldsymbol{L}_{\mathsf{cvx}}, \boldsymbol{S}_{\mathsf{cvx}}) = (\boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{\Delta_L}, \boldsymbol{S}+\boldsymbol{\Delta_S})$ is the optimizer of the convex program (1.3). These allow us to write

$$\begin{aligned}
\boldsymbol{\Delta_S} &= \mathcal{P}_{\Omega_{\mathsf{obs}}}\left[\mathcal{S}_\tau\left(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{\Delta_L}\right) - \mathcal{S}_\tau\left(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^\top\right)\right] \\
&= \mathcal{P}_{\Omega_{\mathsf{obs}}}\left[\mathcal{S}_\tau\left[\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^\top+\boldsymbol{\Delta_S}-(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S})\right] - \mathcal{S}_\tau\left(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^\top\right)\right]. \qquad \text{(E.20)}
\end{aligned}$$

This characterization of $\boldsymbol{\Delta_S}$ turns out to be crucial when establishing the inclusion $\Omega_{\mathsf{obs}}\backslash\Omega^\star \subseteq \Omega_1$. Towards this end, we need to introduce another index subset

$$\Omega_2 \coloneqq \left\{(i,j) \in \Omega_{\mathsf{obs}} \,:\, \tau - \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta_L}+\boldsymbol{\Delta_S})\|_\infty \leq \left|\left(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{Y}^\top\right)_{ij}\right| \leq \tau\right\}.$$

As it turns out, the sets Ω, Ω_1 and Ω_2 obey the following three conditions

$$\Omega_2 \cap \Omega = \varnothing, \qquad \Omega_{\mathsf{obs}}\backslash\Omega \subseteq \Omega_1 \cup \Omega_2, \qquad \text{and} \qquad \Omega \cup \Omega_2 \subseteq \Omega^\star,$$

which immediately lead to

$$\Omega_{\mathsf{obs}}\backslash\Omega^\star \overset{\text{(i)}}{\subseteq} \Omega_{\mathsf{obs}}\backslash(\Omega\cup\Omega_2) \overset{\text{(ii)}}{=} (\Omega_{\mathsf{obs}}\backslash\Omega)\backslash\Omega_2 \overset{\text{(iii)}}{\subseteq} (\Omega_1\cup\Omega_2)\backslash\Omega_2 \subseteq \Omega_1.$$

Here, (i) follows since $\Omega\cup\Omega_2 \subseteq \Omega^\star$, (ii) holds true since $\Omega_2\cap\Omega = \varnothing$, and (iii) results from the condition $\Omega_{\mathsf{obs}}\backslash\Omega \subseteq \Omega_1\cup\Omega_2$. It then boils down to proving each of the above three conditions.

1. The first one $\Omega_2 \cap \Omega = \varnothing$ is straightforward to establish. Note that for any $(i,j) \in \Omega_2$, one must have $\left|\left(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top\right)_{ij}\right| \le \tau$ and hence $\left[\mathcal{S}_\tau(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top)\right]_{ij} = 0$, which means that $(i,j) \notin \Omega$. This proves the relation $\Omega_2 \cap \Omega = \varnothing$.

2. Moving on to the second one $\Omega_{\mathsf{obs}} \backslash \Omega \subseteq \Omega_1 \cup \Omega_2$, we prove this via contradiction. Suppose that this inclusion is false, i.e. there exits an index $(i,j) \in \Omega_{\mathsf{obs}} \backslash \Omega$ such that
$$|(\boldsymbol{\Delta}_{\boldsymbol{S}})_{ij}| > \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta}_{\boldsymbol{L}} + \boldsymbol{\Delta}_{\boldsymbol{S}})\|_\infty \quad \text{and} \quad |(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top)_{ij}| < \tau - \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta}_{\boldsymbol{L}} + \boldsymbol{\Delta}_{\boldsymbol{S}})\|_\infty. \tag{E.21}$$
Here, we have taken into account the fact that
$$|(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top)_{ij}| \le \tau, \qquad \text{for any } (i,j) \in \Omega_{\mathsf{obs}} \backslash \Omega.$$
To reach contradiction, we find it convenient to state the following simple fact.

Fact 2. *Suppose that $|a| \le \tau$ and that $\mathcal{S}_\tau(a+b) \neq 0$. Then*
$$\left|\mathcal{S}_\tau(a+b)\right| \le |b| + |a| - \tau.$$

Proof. Given that $\mathcal{S}_\tau(a+b) \neq 0$, one necessarily has $|a+b| > \tau$. Without loss of generality, assume that $a+b > 0$, which gives
$$\mathcal{S}_\tau(a+b) = a + b - \tau > 0.$$
This together with the fact $\tau \ge |a|$ yields $|\mathcal{S}_\tau(a+b)| = a + b - \tau \le |b| + |a| - \tau$. ☐

With this fact in mind, we can deduce that
$$\begin{aligned}
\left|\left(\boldsymbol{\Delta}_{\boldsymbol{S}}\right)_{ij}\right| &= \left|\left\{\mathcal{S}_\tau\left[\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{\Delta}_{\boldsymbol{S}} - (\boldsymbol{\Delta}_{\boldsymbol{L}} + \boldsymbol{\Delta}_{\boldsymbol{S}})\right] - \mathcal{S}_\tau\left(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top\right)\right\}_{ij}\right| \\
&\overset{\text{(i)}}{=} \left|\left\{\mathcal{S}_\tau\left[\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{\Delta}_{\boldsymbol{S}} - (\boldsymbol{\Delta}_{\boldsymbol{L}} + \boldsymbol{\Delta}_{\boldsymbol{S}})\right]\right\}_{ij}\right| \\
&\overset{\text{(ii)}}{\le} \left|\left[\boldsymbol{\Delta}_{\boldsymbol{S}} - (\boldsymbol{\Delta}_{\boldsymbol{L}} + \boldsymbol{\Delta}_{\boldsymbol{S}})\right]_{ij}\right| + \left|\left(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top\right)_{ij}\right| - \tau \\
&\overset{\text{(iii)}}{<} \left|\left(\boldsymbol{\Delta}_{\boldsymbol{S}}\right)_{ij}\right| + \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta}_{\boldsymbol{L}} + \boldsymbol{\Delta}_{\boldsymbol{S}})\|_\infty - \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta}_{\boldsymbol{L}} + \boldsymbol{\Delta}_{\boldsymbol{S}})\|_\infty \\
&= \left|\left(\boldsymbol{\Delta}_{\boldsymbol{S}}\right)_{ij}\right|,
\end{aligned} \tag{E.22}$$
where (i) holds true since $\left[\mathcal{S}_\tau(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top)\right]_{ij} = 0$ for any $(i,j) \in \Omega_{\mathsf{obs}} \backslash \Omega$, (ii) follows from Fact 2 (by taking $a = (\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top)_{ij}$ and $b = [\boldsymbol{\Delta}_{\boldsymbol{S}} - (\boldsymbol{\Delta}_{\boldsymbol{L}} + \boldsymbol{\Delta}_{\boldsymbol{S}})]_{ij}$), and (iii) is a consequence of (E.21) as well as the triangle inequality. The inequality (E.22), however, is clearly impossible. This establishes that $\Omega_{\mathsf{obs}} \backslash \Omega \subseteq \Omega_1 \cup \Omega_2$.

3. We are left with the last one $\Omega \cup \Omega_2 \subseteq \Omega^\star$, which is equivalent to saying $\Omega \subseteq \Omega^\star$ and $\Omega_2 \subseteq \Omega^\star$. First, for any $(i,j) \in \Omega$, one has
$$\begin{aligned}
\left|S_{ij}\right| > 0 \quad &\Longrightarrow \quad (i,j) \in \Omega_{\mathsf{obs}} \quad \text{and} \quad \left|\left(\boldsymbol{L}^\star + \boldsymbol{S}^\star + \boldsymbol{E} - \boldsymbol{X}\boldsymbol{Y}^\top\right)_{ij}\right| > \tau \\
&\Longrightarrow \quad (i,j) \in \Omega_{\mathsf{obs}} \quad \text{and} \quad \left|S^\star_{ij}\right| > \tau - \left\|\boldsymbol{L}^\star - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_\infty - \|\boldsymbol{E}\|_\infty > 0.
\end{aligned}$$
Here, the last step comes from the triangle inequality and Condition 1. This reveals that $\Omega \subseteq \Omega^\star$. Similarly, for any $(i,j) \in \Omega_2$ we have
$$\begin{aligned}
& \tau - \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta}_{\boldsymbol{L}} + \boldsymbol{\Delta}_{\boldsymbol{S}})\|_\infty \le \left|\left(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top\right)_{ij}\right| \\
\Longleftrightarrow \quad & \tau - \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta}_{\boldsymbol{L}} + \boldsymbol{\Delta}_{\boldsymbol{S}})\|_\infty \le \left|\left(\boldsymbol{S}^\star + \boldsymbol{L}^\star - \boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{E}\right)_{ij}\right| \\
\Longrightarrow \quad & \left|S^\star_{ij}\right| \ge \tau - \left\|\boldsymbol{L}^\star - \boldsymbol{X}\boldsymbol{Y}^\top\right\|_\infty - \|\boldsymbol{E}\|_\infty - \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{\Delta}_{\boldsymbol{L}} + \boldsymbol{\Delta}_{\boldsymbol{S}})\|_{\mathrm{F}} \ge \frac{\tau}{2} - \frac{\sigma}{n^7} > 0,
\end{aligned}$$
where we have used Condition 1, the bound (E.15), and the fact that $\tau \gg \sigma$. This demonstrates that $\Omega_2 \subseteq \Omega^\star$. We have therefore justified that $\Omega \cup \Omega_2 \subseteq \Omega^\star$.

F Analysis of the nonconvex procedure (Proof of Theorem 5)

This section is devoted to establishing Theorem 5. For notational convenience, we introduce

$$\boldsymbol{F}^t := \left[\boldsymbol{X}^{t\top}, \boldsymbol{Y}^{t\top}\right]^\top \in \mathbb{R}^{2n\times r} \qquad \text{and} \qquad \boldsymbol{F}^\star := \left[\boldsymbol{X}^{\star\top}, \boldsymbol{Y}^{\star\top}\right]^\top \in \mathbb{R}^{2n\times r}. \tag{F.1}$$

These allow us to express succinctly the rotation matrix $\boldsymbol{H}^t$ defined in (3.10) as

$$\boldsymbol{H}^t = \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{F}^t\boldsymbol{R} - \boldsymbol{F}^\star\right\|_{\mathrm{F}}. \tag{F.2}$$

With the definitions of $\boldsymbol{F}^t$ and $\boldsymbol{H}^t$ in mind, it suffices to justify that: for all $0 \le t \le t_0 = n^{47}$, the following hypotheses

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

$$\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\right\| \le C_{\mathrm{op}}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\left\|\boldsymbol{X}^\star\right\|, \tag{F.3b}$$

$$\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\right\|_{2,\infty} \le C_{\infty}\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\left\|\boldsymbol{F}^\star\right\|_{2,\infty}, \tag{F.3c}$$

$$\left\|\boldsymbol{X}^{t\top}\boldsymbol{X}^t - \boldsymbol{Y}^{t\top}\boldsymbol{Y}^t\right\|_{\mathrm{F}} \le C_{\mathrm{B}}\kappa\eta\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\sqrt{r}\sigma_{\max}^2, \tag{F.3d}$$

$$\left\|\boldsymbol{S}^t - \boldsymbol{S}^\star\right\| \le C_{\mathrm{S}}\sigma\sqrt{np} \tag{F.3e}$$

hold for some universal constants C_{F}, C_{op}, C_{∞}, C_{B}, $C_{\mathrm{S}} > 0$, and, in addition,

$$F\left(\boldsymbol{X}^t, \boldsymbol{Y}^t; \boldsymbol{S}^t\right) \le F\left(\boldsymbol{X}^{t-1}, \boldsymbol{Y}^{t-1}; \boldsymbol{S}^{t-1}\right) - \frac{\eta}{2}\left\|\nabla f\left(\boldsymbol{X}^{t-1}, \boldsymbol{Y}^{t-1}; \boldsymbol{S}^{t-1}\right)\right\|_{\mathrm{F}}^2 \tag{F.4}$$

holds for all $1 \le t \le t_0 = n^{47}$.

Clearly, the bounds (3.11a), (3.11b), (3.11c), and (3.11d) in Theorem 5 follow immediately from (F.3a), (F.3b), (F.3c), and (F.3e), respectively. It remains to justify the small gradient bound (3.12) on the basis of (F.3) and (F.4), which is exactly the content of the following lemma.

Lemma 10 (**Small gradient**). *Set $\lambda = C_\lambda\sigma\sqrt{np\log n}$ for some large constant $C_\lambda > 0$. Suppose that $n^2p \gg \kappa^3\mu rn\log^2 n$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^4\mu r\log n}}$. Take $\eta \asymp 1/(n\kappa^3\sigma_{\max})$. If the iterates satisfy (F.3) for all $0 \le t \le t_0$ and (F.4) for all $1 \le t \le t_0$, then with probability at least $1 - O(n^{-50})$, one has*

$$\min_{0\le t<t_0}\left\|\nabla f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t; \boldsymbol{S}^t\right)\right\|_{\mathrm{F}} \le \frac{1}{n^{20}}\frac{\lambda}{p}\sqrt{\sigma_{\min}}.$$

Proof. See Appendix F.3. $\square$

The remainder of this section is thus dedicated to showing that (F.3) and (F.4) hold for $\{(\boldsymbol{F}^t, \boldsymbol{S}^t)\}_{0\le t\le t_0}$, which we accomplish via mathematical induction. Throughout this section, we let $\boldsymbol{X}_{l,\cdot}$ denote the lth row of a matrix $\boldsymbol{X}$.

F.1 Leave-one-out analysis

The above hypotheses (F.3) require, among other things, sharp control of the $\ell_{2,\infty}$ estimation errors, which calls for fine-grained statistical analyses. In order to decouple complicated statistical dependency, we resort to the following leave-one-out analysis framework that has been successfully applied to analyze other nonconvex algorithms [ZB18, MWCC20, CLL20, CCFM19, CCF+20, CLPC20, LWC+20, DC20].

Leave-one-out loss functions. For each $1 \leq l \leq n$, we define the following auxiliary loss functions

$$F^{(l)}\left(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S}\right) \\ := \underbrace{\frac{1}{2p}\left\|\mathcal{P}_{(\Omega_{\mathsf{obs}})_{-l,\cdot}}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}+\boldsymbol{S}-\boldsymbol{M}\right)\right\|_{\mathrm{F}}^{2}+\frac{1}{2}\left\|\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{L}^{\star}\right)\right\|_{\mathrm{F}}^{2}+\frac{\lambda}{2p}\left\|\boldsymbol{X}\right\|_{\mathrm{F}}^{2}+\frac{\lambda}{2p}\left\|\boldsymbol{Y}\right\|_{\mathrm{F}}^{2}}_{=:f^{(l)}(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{S})}+\frac{\tau}{p}\left\|\boldsymbol{S}\right\|_{1}.$$

Here, $\mathcal{P}_{(\Omega_{\mathsf{obs}})_{-l,\cdot}}(\cdot)$ (resp. $\mathcal{P}_{l,\cdot}(\cdot)$) denotes orthogonal projection onto the space of matrices supported on the index set $\{(i,j) \in \Omega_{\mathsf{obs}} \mid i \neq l\}$ (resp. $\{(i,j) \mid i = l\}$), namely, for any matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ one has

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

The above auxiliary loss function is obtained by dropping the randomness coming from the lth row of $\boldsymbol{M}$, which, as we shall see shortly, facilitates analysis in establishing the incoherence properties (F.3c). Similarly, we define for each $n+1 \leq l \leq 2n$ that

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

where the projection operators $\mathcal{P}_{(\Omega_{\mathsf{obs}})_{\cdot,-(l-n)}}(\cdot)$ and $\mathcal{P}_{\cdot,(l-n)}(\cdot)$ are defined such that for any matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$,

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

Again, this auxiliary loss function is produced in a way that is independent from the $(l-n)$-th column of $\boldsymbol{M}$. In the above notation, $f^{(l)}\left(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{S}\right)$ is a function of $\boldsymbol{X}$ and $\boldsymbol{Y}$ with $\boldsymbol{S}$ frozen.

Leave-one-out auxiliary sequences. For each $1 \leq l \leq 2n$, we construct a sequence of leave-one-out iterates $\{\boldsymbol{F}^{t,(l)}, \boldsymbol{S}^{t,(l)}\}_{t\geq 0}$ via Algorithm 2.

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

Initialization: $\boldsymbol{X}^{0,(l)} = \boldsymbol{X}^{\star}$, $\boldsymbol{Y}^{0,(l)} = \boldsymbol{Y}^{\star}$, $\boldsymbol{S}^{0,(l)} = \boldsymbol{S}^{\star}$, $\boldsymbol{F}^{0,(l)} := \begin{bmatrix}\boldsymbol{X}^{0,(l)}\\ \boldsymbol{Y}^{0,(l)}\end{bmatrix}$, and the step size $\eta > 0$.

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

$$\boldsymbol{F}^{t+1,(l)} := \begin{bmatrix}\boldsymbol{X}^{t+1,(l)}\\ \boldsymbol{Y}^{t+1,(l)}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}^{t,(l)}-\eta\nabla_{\boldsymbol{X}}f^{(l)}(\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)};\boldsymbol{S}^{t,(l)})\\ \boldsymbol{Y}^{t,(l)}-\eta\nabla_{\boldsymbol{Y}}f^{(l)}(\boldsymbol{X}^{t,(l)},\boldsymbol{Y}^{t,(l)};\boldsymbol{S}^{t,(l)})\end{bmatrix}; \tag{F.5a}$$

$$\boldsymbol{S}^{t+1,(l)} := \begin{cases}\mathcal{S}_{\tau}\left[\mathcal{P}_{(\Omega_{\mathsf{obs}})_{-l,\cdot}}\left(\boldsymbol{M}-\boldsymbol{X}^{t+1,(l)}\boldsymbol{Y}^{t+1,(l)\top}\right)\right]+\mathcal{P}_{l,\cdot}\left(\boldsymbol{S}^{\star}\right), & \text{if } 1\leq l\leq n,\\ \mathcal{S}_{\tau}\left[\mathcal{P}_{(\Omega_{\mathsf{obs}})_{\cdot,-(l-n)}}\left(\boldsymbol{M}-\boldsymbol{X}^{t+1,(l)}\boldsymbol{Y}^{t+1,(l)\top}\right)\right]+\mathcal{P}_{\cdot,(l-n)}\left(\boldsymbol{S}^{\star}\right), & \text{if } n+1\leq l\leq 2n.\end{cases} \tag{F.5b}$$

Properties of leave-one-out sequences. There are several features of the leave-one-out sequences that prove useful for our statistical analysis: (1) for the lth leave-one-out sequence, one can exploit the statistical independence to control the estimation error of $\boldsymbol{F}^{t,(l)}$ in the lth row; (2) the leave-one-out sequences and the original sequence $(\boldsymbol{F}^t, \boldsymbol{S}^t)$ are exceedingly close (since we have only discarded a small amount of information). These properties taken collectively allow us to control the estimation error of $\boldsymbol{F}^t$ in each row. To formalize these features, we make an additional set of induction hypotheses

$$\left\|\boldsymbol{F}^{t}\boldsymbol{H}^{t}-\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}\leq C_{1}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\sigma_{\min}}\right)\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}, \tag{F.6a}$$

$$\max_{1\le l\le 2n}\left\|\left(\boldsymbol{F}^{t,(l)}\boldsymbol{H}^{t,(l)}-\boldsymbol{F}^{\star}\right)_{l,\cdot}\right\|_{2}\le C_{2}\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\sigma_{\min}}\right)\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}, \tag{F.6b}$$

$$\max_{1\le l\le n}\left\|\mathcal{P}_{-l,\cdot}\left(\boldsymbol{S}^{t}-\boldsymbol{S}^{t,(l)}\right)\right\|_{\mathrm{F}}\le C_{3}\frac{\sigma}{\sigma_{\min}}\sqrt{n\log n}\left\|\boldsymbol{F}^{\star}\right\|\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}, \tag{F.6c}$$

$$\max_{n<l\le 2n}\left\|\mathcal{P}_{\cdot,-(l-n)}\left(\boldsymbol{S}^{t}-\boldsymbol{S}^{t,(l)}\right)\right\|_{\mathrm{F}}\le C_{3}\frac{\sigma}{\sigma_{\min}}\sqrt{n\log n}\left\|\boldsymbol{F}^{\star}\right\|\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}. \tag{F.6d}$$

Here, the rotation matrices $\boldsymbol{H}^{t,(l)}$ and $\boldsymbol{R}^{t,(l)}$ are defined respectively by

$$\boldsymbol{H}^{t,(l)}\coloneqq\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}-\boldsymbol{F}^{\star}\right\|_{\mathrm{F}},\qquad\text{and}\qquad\boldsymbol{R}^{t,(l)}\coloneqq\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}-\boldsymbol{F}^{t}\right\|_{\mathrm{F}}.$$

F.2 Key lemmas for establishing the induction hypotheses

This subsection establishes the induction hypotheses made in Appendix F.1, namely (F.3), (F.4) and (F.6). Before continuing, we find it convenient to introduce another function of $\boldsymbol{X}$ and $\boldsymbol{Y}$ (with $\boldsymbol{S}$ frozen) as follows

$$f_{\mathsf{aug}}\left(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{S}\right)\coloneqq\frac{1}{2p}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}\boldsymbol{Y}^{\top}+\boldsymbol{S}-\boldsymbol{M}\right)\right\|_{\mathrm{F}}^{2}+\frac{\lambda}{2p}\left\|\boldsymbol{X}\right\|_{\mathrm{F}}^{2}+\frac{\lambda}{2p}\left\|\boldsymbol{Y}\right\|_{\mathrm{F}}^{2}+\frac{1}{8}\left\|\boldsymbol{X}^{\top}\boldsymbol{X}-\boldsymbol{Y}^{\top}\boldsymbol{Y}\right\|_{\mathrm{F}}^{2}. \tag{F.7}$$

The difference between f_{aug} and f lies in the following balancing term

$$f_{\mathsf{diff}}\left(\boldsymbol{X},\boldsymbol{Y}\right)\coloneqq-\frac{1}{8}\left\|\boldsymbol{X}^{\top}\boldsymbol{X}-\boldsymbol{Y}^{\top}\boldsymbol{Y}\right\|_{\mathrm{F}}^{2},$$

that is, $f=f_{\mathsf{aug}}+f_{\mathsf{diff}}$.

The following four lemmas, which are inherited from [CCF+20] with little modification, are concerned with local strong convexity as well as the hypotheses (F.3a), (F.3b), (F.3d), (F.6b) and (F.3c).

Lemma 11 (**Restricted strong convexity**). *Set $\lambda=C_{\lambda}\sigma\sqrt{np}$ for some large enough constant $C_{\lambda}>0$. Suppose that the sample size obeys $n^{2}p\gg\kappa\mu rn\log n$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\ll1$. Let the function f_{aug} be defined in (F.7). Then with probability at least $1-O(n^{-100})$,*

$$\mathsf{vec}\left(\boldsymbol{\Delta}\right)^{\top}\nabla^{2}f_{\mathsf{aug}}\left(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{S}\right)\mathsf{vec}\left(\boldsymbol{\Delta}\right)\ge\tfrac{1}{10}\sigma_{\min}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}^{2}$$
$$\max\left\{\left\|\nabla^{2}f_{\mathsf{aug}}\left(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{S}\right)\right\|,\left\|\nabla^{2}f\left(\boldsymbol{X},\boldsymbol{Y}\right)\right\|\right\}\le10\sigma_{\max}$$

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

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

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

$$\left\{\begin{bmatrix}\boldsymbol{X}_{1}\\ \boldsymbol{Y}_{1}\end{bmatrix}\hat{\boldsymbol{H}}-\begin{bmatrix}\boldsymbol{X}_{2}\\ \boldsymbol{Y}_{2}\end{bmatrix}\,\middle|\,\left\|\begin{bmatrix}\boldsymbol{X}_{2}-\boldsymbol{X}^{\star}\\ \boldsymbol{Y}_{2}-\boldsymbol{Y}^{\star}\end{bmatrix}\right\|\le\frac{1}{500\kappa}\left\|\boldsymbol{X}^{\star}\right\|,\hat{\boldsymbol{H}}\coloneqq\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\begin{bmatrix}\boldsymbol{X}_{1}\\ \boldsymbol{Y}_{1}\end{bmatrix}\boldsymbol{R}-\begin{bmatrix}\boldsymbol{X}_{2}\\ \boldsymbol{Y}_{2}\end{bmatrix}\right\|_{\mathrm{F}}\right\}.$$

Lemma 12 (**Frobenius norm error w.r.t. $\boldsymbol{F}$**). *Set $\lambda=C_{\lambda}\sigma\sqrt{np}$ for some large enough constant $C_{\lambda}>0$. Suppose that the sample size obeys $n^{2}p\gg\kappa\mu rn\log^{2}n$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\ll\frac{1}{\sqrt{\kappa^{4}\mu r\log n}}$. If the iterates satisfy (F.3) in the tth iteration, then with probability at least $1-O(n^{-100})$,*

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

holds as long as $0<\eta\ll1/(\kappa^{5/2}\sigma_{\max})$.

Lemma 13 (**Spectral norm error w.r.t. $\boldsymbol{F}$**). *Set $\lambda = C_\lambda \sigma\sqrt{np}$ for some large enough constant $C_\lambda > 0$. Suppose that the sample size obeys $n^2 p \gg \kappa^4\mu^2 r^2 n \log^2 n$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^4 \log n}}$. If the iterates satisfy (F.3) in the tth iteration, then with probability at least $1 - O(n^{-100})$, one has*

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

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

Lemma 14 (**Approximate balancedness**). *Set $\lambda = C_\lambda \sigma\sqrt{np}$ for some large enough 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 (F.3) in the tth iteration, then with probability at least $1 - O(n^{-100})$,*

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

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

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

Lemma 15 (**$\ell_{2,\infty}$ norm error of leave-one-out sequences**). *Set $\lambda = C_\lambda \sigma\sqrt{np}$ for some large enough constant $C_\lambda > 0$. Suppose that the sample size obeys $n^2 p \gg \kappa^4\mu^2 r^2 n \log^3 n$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2 \log n}}$. If the iterates satisfy (F.3) in the tth iteration, then with probability at least $1 - O(n^{-100})$,*

$$\max_{1\le l\le 2n}\left\|\left(\boldsymbol{F}^{t+1,(l)}\boldsymbol{H}^{t+1,(l)} - \boldsymbol{F}^\star\right)_{l,\cdot}\right\|_2 \le C_2\kappa\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}$$

holds, provided that $0 < \eta \ll 1/(\kappa^2\sqrt{r}\sigma_{\max})$, $C_{\mathrm{op}} \gg 1$ and $C_2 \gg C_{\mathrm{op}}$.

Lemma 16 (**$\ell_{2,\infty}$ norm error of the true sequence**). *Set $\lambda = C_\lambda \sigma\sqrt{np}$ for some large enough constant $C_\lambda > 0$. Suppose that $n \ge \mu r$ and that the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2 \log n}}$. If the iterates satisfy (F.3) and (F.6) in the tth iteration, then with probability at least $1 - O(n^{-99})$, one has*

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

provided that $C_\infty \ge 5C_1 + C_2$.

Proof of Lemmas 11, 12, 13, 14, 15 and 16. As it turns out, Lemmas 11, 12, 13 and 14 follow immediately from the proofs of [CCF+20, Lemmas 17, 10, 11, 15] respectively. More specifically, the proofs can be accomplished by replacing $\boldsymbol{E}$ in the proofs therein with $\tilde{\boldsymbol{E}} \coloneqq \boldsymbol{E} + \boldsymbol{S}^\star - \boldsymbol{S}^t$. To see this, we remark that the only property of the perturbation matrix $\boldsymbol{E}$ utilized in the proofs therein is that $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\| \lesssim \sigma\sqrt{np}$ with probability at least $1 - O(n^{-10})$; under our hypotheses, the new matrix $\tilde{\boldsymbol{E}}$ clearly satisfies this property since

$$\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\tilde{\boldsymbol{E}})\| \le \|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{E})\| + \left\|\boldsymbol{S}^\star - \boldsymbol{S}^t\right\| \lesssim \sigma\sqrt{np}.$$

Regarding Lemma 15, we note that $\boldsymbol{S}_{l,\cdot}^{t,(l)} \equiv \boldsymbol{S}_{l,\cdot}^\star$ by construction. Therefore, the update rule regarding the lth row of $\{\boldsymbol{X}_{l,\cdot}^{t,(l)}\}_{t\ge 0}$ and $\{\boldsymbol{Y}_{l,\cdot}^{t,(l)}\}_{t\ge 0}$ is exactly the same as that in the leave-one-out sequence introduced in [CCF+20]. Thus, Lemma 15 follows immediately from the proof of [CCF+20, Lemma 13].

Finally, the proof of Lemma 16 is exactly the same as the proof of [CCF+20, Lemma 14]. □

Next, we justify the hypotheses (F.3e), (F.6a) and (F.6c) in the following three lemmas, which require more careful analysis of the properties about $\{\boldsymbol{S}^t\}$.

Lemma 17 (**Spectral norm error w.r.t. $\boldsymbol{S}$**). *Set $\tau = C_\tau \sigma \sqrt{\log n}$ for some large enough constant $C_\tau > 0$. Suppose that the sample size obeys $n^2 p \gg \kappa^4 \mu r n \log n$, the noise satisfies $\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll 1/\sqrt{\kappa^2 \log n}$, the outlier fraction satisfies $\rho_{\mathsf{s}} \le \rho_{\mathsf{aug}} \ll 1/\sqrt{\kappa^5 \mu r \log^2 n}$ and $n^2 p \rho_{\mathsf{aug}} \gg \mu n r \log^2 n$. If the iterates satisfy (F.3b) and (F.3c) in the $(t+1)$-th iteration, then with probability at least $1 - O(n^{-100})$,*

$$\left\| \boldsymbol{S}^{t+1} - \boldsymbol{S}^\star \right\| \le C_{\mathrm{S}} \sigma \sqrt{np}$$

holds for some constant $C_{\mathrm{S}} > 0$ that does not rely on the choice of other constants.

Proof. See Appendix F.4. □

Lemma 18 (**Leave-one-out perturbation w.r.t. $\boldsymbol{F}$**). *Set $\lambda = C_\lambda \sigma \sqrt{np}$ for some large enough constant $C_\lambda > 0$. Suppose that the sample size obeys $n^2 p \gg \kappa^4 \mu^2 r^2 n \log^4 n$, the noise satisfies $\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll 1/\sqrt{\kappa^4 \mu r \log n}$, the outlier fraction satisfies $\rho_{\mathsf{s}} \le \rho_{\mathsf{aug}} \ll 1/(\kappa^3 \mu r \log n)$ and $n^2 p \rho_{\mathsf{aug}} \gg \mu r n \log n$. If the iterates satisfy (F.3) and (F.6) in the tth iteration, then with probability at least $1 - O(n^{-100})$,*

$$\max_{1 \le l \le 2n} \left\| \boldsymbol{F}^{t+1} \boldsymbol{H}^{t+1} - \boldsymbol{F}^{t+1,(l)} \boldsymbol{R}^{t+1,(l)} \right\|_{\mathrm{F}} \le C_1 \left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p \sigma_{\min}} \right) \left\| \boldsymbol{F}^\star \right\|_{2,\infty}$$

holds for some constant $C_1 > 0$, provided that $\eta \ll 1/(n \kappa^2 \sigma_{\max})$ and $C_1 \gg C_3$.

Proof. See Appendix F.5. □

Lemma 19 (**Leave-one-out perturbation w.r.t. $\boldsymbol{S}$**). *Set $\tau = C_\tau \sigma \sqrt{\log n}$ for some large enough constant $C_\tau > 0$. Suppose that the sample size satisfies $n^2 p \gg \kappa^4 \mu^2 r^2 n \log n$, the noise obeys $\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll 1/\sqrt{\kappa^2 \log n}$ and the outlier fraction satisfies $\rho_{\mathsf{s}} \le \rho_{\mathsf{aug}} \ll 1/\kappa$. If the iterates satisfy (F.3b), (F.3c) and (F.6a) in the $(t+1)$-th iteration, then with probability at least $1 - O(n^{-100})$,*

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

$$\max_{n < l \le 2n} \left\| \mathcal{P}_{\cdot,-(l-n)} \left(\boldsymbol{S}^{t+1} - \boldsymbol{S}^{t+1,(l)} \right) \right\|_{\mathrm{F}} \le C_3 \frac{\sigma}{\sigma_{\min}} \sqrt{n \log n} \left\| \boldsymbol{F}^\star \right\| \left\| \boldsymbol{F}^\star \right\|_{2,\infty}$$

hold for some constant C_3 that does not rely on the choice of other constants.

Proof. See Appendix F.6. □

Finally, it remains to justify (F.4), which is a straightforward consequence from standard gradient descent theory and implies the existence of a point with nearly zero gradient.

Lemma 20 (**Monotonicity of the function values**). *Set $\lambda = C_\lambda \sigma \sqrt{np}$ for some large enough constant $C_\lambda > 0$. Suppose that the noise satisfies $\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \ll 1$. If the iterates satisfy (F.3) in the tth iteration, then with probability at least $1 - O(n^{-100})$,*

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

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

Proof. See Appendix F.7. □

F.3 Proof of Lemma 10

Summing (F.4) over $t = 1, \dots, t_0$ gives

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

which further implies

$$\begin{aligned}\min_{0\leq t<t_0} \left\|\nabla f\left(\boldsymbol{X}^{t}, \boldsymbol{Y}^{t}; \boldsymbol{S}^{t}\right)\right\|_{\mathrm{F}}^2 &\leq \frac{1}{t_0} \sum_{t=0}^{t_0-1} \left\|\nabla f\left(\boldsymbol{X}^{t}, \boldsymbol{Y}^{t}; \boldsymbol{S}^{t}\right)\right\|_{\mathrm{F}}^2 \\ &\leq \frac{2}{\eta t_0} \left[F\left(\boldsymbol{X}^{\star}, \boldsymbol{Y}^{\star}, \boldsymbol{S}^{\star}\right) - F\left(\boldsymbol{X}^{t_0}, \boldsymbol{Y}^{t_0}, \boldsymbol{S}^{t_0}\right)\right]. \qquad \text{(F.8)}\end{aligned}$$

Here, the last inequality results from our choice $(\boldsymbol{X}^0, \boldsymbol{Y}^0, \boldsymbol{S}^0) = (\boldsymbol{X}^\star, \boldsymbol{Y}^\star, \boldsymbol{S}^\star)$. Therefore, it suffices to control $F\left(\boldsymbol{X}^{\star}, \boldsymbol{Y}^{\star}, \boldsymbol{S}^{\star}\right) - F\left(\boldsymbol{X}^{t_0}, \boldsymbol{Y}^{t_0}, \boldsymbol{S}^{t_0}\right)$.

We first decompose the difference into

$$\begin{aligned}&F\left(\boldsymbol{X}^{\star}, \boldsymbol{Y}^{\star}, \boldsymbol{S}^{\star}\right) - F\left(\boldsymbol{X}^{t_0}, \boldsymbol{Y}^{t_0}, \boldsymbol{S}^{t_0}\right) = f\left(\boldsymbol{X}^{\star}, \boldsymbol{Y}^{\star}; \boldsymbol{S}^{\star}\right) + \frac{\tau}{p}\left\|\boldsymbol{S}^\star\right\|_1 - f\left(\boldsymbol{X}^{t_0}, \boldsymbol{Y}^{t_0}; \boldsymbol{S}^{t_0}\right) - \frac{\tau}{p}\left\|\boldsymbol{S}^{t_0}\right\|_1 \\ &\quad = \underbrace{f\left(\boldsymbol{X}^{\star}, \boldsymbol{Y}^{\star}; \boldsymbol{S}^{\star}\right) - f\left(\boldsymbol{X}^{t_0}, \boldsymbol{Y}^{t_0}; \boldsymbol{S}^{\star}\right)}_{=:\Delta_1} + \underbrace{f\left(\boldsymbol{X}^{t_0}, \boldsymbol{Y}^{t_0}; \boldsymbol{S}^{\star}\right) - f\left(\boldsymbol{X}^{t_0}, \boldsymbol{Y}^{t_0}; \boldsymbol{S}^{t_0}\right)}_{=:\Delta_2} + \underbrace{\frac{\tau}{p}\left\|\boldsymbol{S}^\star\right\|_1 - \frac{\tau}{p}\left\|\boldsymbol{S}^{t_0}\right\|_1}_{=:\Delta_3}.\end{aligned}$$

In what follows, we shall bound Δ_1, Δ_2 and Δ_3 separately.

1. In view of the proof of [CCF+20, Lemma 9], we have

$$|\Delta_1| \lesssim r\kappa^2 \left(\frac{\lambda}{p}\right)^2,$$

provided that $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^4 \mu r \log n}}$.

2. When it comes to Δ_2, we deduce that

$$\begin{aligned}|\Delta_2| &= \left|\frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}^{t_0}\boldsymbol{Y}^{t_0\top} + \boldsymbol{S}^\star - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2p}\|\boldsymbol{F}^{t_0}\|_{\mathrm{F}}^2 - \frac{1}{2}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}^{t_0}\boldsymbol{Y}^{t_0\top} + \boldsymbol{S}^{t_0} - \boldsymbol{M}\right)\right\|_{\mathrm{F}}^2 - \frac{\lambda}{2p}\|\boldsymbol{F}^{t_0}\|_{\mathrm{F}}^2\right| \\ &= \left|\left\langle\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}^{t_0}\boldsymbol{Y}^{t_0\top} + \boldsymbol{S}^{t_0} - \boldsymbol{M}\right), \boldsymbol{S}^\star - \boldsymbol{S}^{t_0}\right\rangle + \frac{1}{2}\left\|\boldsymbol{S}^\star - \boldsymbol{S}^{t_0}\right\|_{\mathrm{F}}^2\right| \\ &\leq \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}^{t_0}\boldsymbol{Y}^{t_0\top} + \boldsymbol{S}^{t_0} - \boldsymbol{M}\right)\right\|_{\mathrm{F}}\left\|\boldsymbol{S}^\star - \boldsymbol{S}^{t_0}\right\|_{\mathrm{F}} + \frac{1}{2}\left\|\boldsymbol{S}^\star - \boldsymbol{S}^{t_0}\right\|_{\mathrm{F}}^2, \qquad \text{(F.9)}\end{aligned}$$

where the last step arises from the elementary inequality $\langle \boldsymbol{A}, \boldsymbol{B}\rangle \leq \|\boldsymbol{A}\|_{\mathrm{F}}\|\boldsymbol{B}\|_{\mathrm{F}}$ and the triangle inequality. It is straightforward to derive from (F.3e) that

$$\|\boldsymbol{S}^\star - \boldsymbol{S}^{t_0}\|_{\mathrm{F}} \leq \sqrt{n}\|\boldsymbol{S}^\star - \boldsymbol{S}^{t_0}\| \lesssim \sigma n\sqrt{p}.$$

Moving on to the first term in (F.9), one has by the triangle inequality

$$\begin{aligned}\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}^{t_0}\boldsymbol{Y}^{t_0\top} + \boldsymbol{S}^{t_0} - \boldsymbol{M}\right)\right\|_{\mathrm{F}} &\leq \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}^{t_0}\boldsymbol{Y}^{t_0\top} - \boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top}\right)\right\|_{\mathrm{F}} + \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{S}^{t_0} - \boldsymbol{S}^\star\right)\right\|_{\mathrm{F}} + \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{E}\right)\right\|_{\mathrm{F}} \\ &\lesssim \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}^{t_0}\boldsymbol{Y}^{t_0\top} - \boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top}\right)\right\|_{\mathrm{F}} + \sigma n\sqrt{p},\end{aligned}$$

where the last bound follows from Lemma 1 and the bound above $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{S}^{t_0} - \boldsymbol{S}^\star)\|_{\mathrm{F}} \leq \|\boldsymbol{S}^{t_0} - \boldsymbol{S}^\star\|_{\mathrm{F}} \lesssim \sigma n\sqrt{p}$. We can further decompose $\|\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{X}^{t_0}\boldsymbol{Y}^{t_0\top} - \boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top})\|_{\mathrm{F}}$ into

$$\left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{X}^{t_0}\boldsymbol{Y}^{t_0\top} - \boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top}\right)\right\|_{\mathrm{F}} \leq \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left[\left(\boldsymbol{X}^{t_0}\boldsymbol{H}^{t_0} - \boldsymbol{X}^\star\right)\boldsymbol{Y}^{\star\top}\right]\right\|_{\mathrm{F}} + \left\|\mathcal{P}_{\Omega_{\mathsf{obs}}}\left[\boldsymbol{X}^{t_0}\boldsymbol{H}^{t_0}\left(\boldsymbol{Y}^{t_0}\boldsymbol{H}^{t_0} - \boldsymbol{Y}^\star\right)^\top\right]\right\|_{\mathrm{F}}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\lesssim} \sqrt{\kappa p}\left\|\left(\boldsymbol{X}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{X}^{\star}\right)\boldsymbol{Y}^{\star\top}\right\|_{\mathrm{F}}+\sqrt{\kappa p}\left\|\boldsymbol{X}^{t_0}\boldsymbol{H}^{t_0}\left(\boldsymbol{Y}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{Y}^{\star}\right)^{\top}\right\|_{\mathrm{F}} \\
&\lesssim \sqrt{\kappa p}\left\|\boldsymbol{X}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}\left\|\boldsymbol{Y}^{\star}\right\|+\sqrt{\kappa p}\left\|\boldsymbol{X}^{t_0}\boldsymbol{H}^{t_0}\right\|\left\|\boldsymbol{Y}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{Y}^{\star}\right\|_{\mathrm{F}} \\
&\overset{\text{(ii)}}{\lesssim} \sqrt{\kappa p}\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}+\frac{\lambda}{p\sigma_{\min}}\right)\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}\left\|\boldsymbol{X}^{\star}\right\| \lesssim \kappa^{3/2}\frac{\lambda}{\sqrt{p}}\sqrt{r}. \qquad \text{(F.10)}
\end{aligned}
$$

Here, the relation (i) utilizes Lemma 4, and the facts that $(\boldsymbol{X}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{X}^{\star})\boldsymbol{Y}^{\star\top} \in T^{\star}$ and that $\boldsymbol{X}^{t_0}\boldsymbol{H}^{t_0}(\boldsymbol{Y}^{t_0}\boldsymbol{H}^{t_0}-\boldsymbol{Y}^{\star})^{\top} \in T^{t_0}$, where T^{t_0} denotes the tangent space at $\boldsymbol{X}^{t_0}\boldsymbol{Y}^{t_0\top}$. In addition, the last line (ii) holds because of the hypothesis (F.3a) and the simple fact $\|\boldsymbol{X}^{t_0}\boldsymbol{H}^{t_0}\| \le 2\|\boldsymbol{X}^{\star}\|$, which is an immediate consequence of the hypothesis (F.3b) provided that $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1$. Collecting the bounds together, we arrive at

$$
|\Delta_2| \lesssim \left(\kappa^{3/2}\frac{\lambda}{\sqrt{p}}\sqrt{r}+\sigma n\sqrt{p}\right)\cdot\sigma n\sqrt{p}+\sigma^2 n^2 p+\frac{\lambda}{p}\cdot\kappa\frac{\lambda}{p}r \lesssim \sigma^2 n^2 p,
$$

with the proviso that $np \gg \kappa^3 r$.

3. In the end, we have the following upper bound on Δ_3:

$$
|\Delta_3| \le \frac{\tau}{p}\left\|\boldsymbol{S}^{t_0}-\boldsymbol{S}^{\star}\right\|_1 \le \frac{\tau}{p}n\left\|\boldsymbol{S}^{t_0}-\boldsymbol{S}^{\star}\right\|_{\mathrm{F}} \lesssim \frac{1}{p}\frac{\lambda}{\sqrt{np/\log n}}\sigma n^2\sqrt{p} \asymp \frac{\lambda}{p}\sigma n^{3/2}\sqrt{\log n},
$$

where we have made use of the elementary fact that $\|\boldsymbol{A}\|_1 \le n\|\boldsymbol{A}\|_{\mathrm{F}}$ for all $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

Putting the above bounds together, one can reach

$$
\left|F\left(\boldsymbol{X}^{\star},\boldsymbol{Y}^{\star},\boldsymbol{S}^{\star}\right)-F\left(\boldsymbol{X}^{t_0},\boldsymbol{Y}^{t_0},\boldsymbol{S}^{t_0}\right)\right| \lesssim r\kappa^2\left(\frac{\lambda}{p}\right)^2+\sigma^2 n^2 p+\frac{\lambda}{p}\sigma n^{3/2}\sqrt{\log n} \lesssim n\left(\frac{\lambda}{p}\right)^2\sqrt{\log n}
$$

as long as $n \gg \kappa^2 r$ and $\lambda \asymp \sigma\sqrt{np}$. Substitution into (F.8) allows us to conclude that

$$
\min_{0\le t\le t_0}\left\|\nabla f\left(\boldsymbol{X}^{t},\boldsymbol{Y}^{t};\boldsymbol{S}^{t}\right)\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{1}{\eta t_0}n\left(\frac{\lambda}{p}\right)^2\sqrt{\log n}} \le \frac{1}{n^{20}}\frac{\lambda}{p}\sqrt{\sigma_{\min}},
$$

provided that $\eta \asymp 1/(n\kappa^3\sigma_{\max})$, $t_0 \ge n^{47}$ and $n \ge \kappa$.

F.4 Proof of Lemma 17

In view of the definitions $\Omega^{\star} = \{(i,j) : S^{\star}_{ij} \ne 0\} \subseteq \Omega_{\mathsf{aug}} \subseteq \Omega_{\mathsf{obs}}$ and $\boldsymbol{S}^{t+1} = \mathcal{S}_{\tau}[\mathcal{P}_{\Omega_{\mathsf{obs}}}(\boldsymbol{L}^{\star}+\boldsymbol{S}^{\star}+\boldsymbol{E}-\boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top})]$, we have the decomposition

$$
\begin{aligned}
\boldsymbol{S}^{t+1}-\boldsymbol{S}^{\star} &= \mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{S}^{t+1}\right)-\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{S}^{\star}\right)+\mathcal{P}_{\Omega_{\mathsf{aug}}^{\mathrm{c}}}\left(\boldsymbol{S}^{t+1}\right) \\
&= \underbrace{\mathcal{S}_{\tau}\left[\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top}+\boldsymbol{S}^{\star}+\boldsymbol{E}-\boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top}\right)\right]-\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{S}^{\star}\right)}_{=:\boldsymbol{A}^{t+1}}+\underbrace{\mathcal{S}_{\tau}\left[\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega_{\mathsf{aug}}}\left(\boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top}+\boldsymbol{E}-\boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top}\right)\right]}_{=:\boldsymbol{B}^{t+1}}.
\end{aligned}
\qquad \text{(F.11)}
$$

We shall control $\|\boldsymbol{A}^{t+1}\|$ and $\|\boldsymbol{B}^{t+1}\|$ separately.

1. We begin by controlling the size of $\boldsymbol{A}^{t+1}$, which can be further decomposed into

$$
\begin{aligned}
\boldsymbol{A}^{t+1} = \mathcal{P}_{\Omega_{\mathsf{aug}}}(\boldsymbol{E})&+\underbrace{\mathcal{S}_{\tau}\left[\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{S}^{\star}+\boldsymbol{E}\right)\right]-\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{S}^{\star}+\boldsymbol{E}\right)}_{=:\boldsymbol{A}_1} \\
&+\underbrace{\mathcal{S}_{\tau}\left[\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top}-\boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top}+\boldsymbol{S}^{\star}+\boldsymbol{E}\right)\right]-\mathcal{S}_{\tau}\left[\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{S}^{\star}+\boldsymbol{E}\right)\right]}_{=:\boldsymbol{A}_2^{t+1}}. \qquad \text{(F.12)}
\end{aligned}
$$

First of all, we know that $\|\mathcal{P}_{\Omega_{\mathsf{aug}}}(\boldsymbol{E})\| \lesssim \sigma\sqrt{np\rho_{\mathsf{aug}}} \leq \sigma\sqrt{np}$, as long as $n^2 p\rho_{\mathsf{aug}} \gg n\log^2 n$. This arises from standard concentration results for the spectral norm of sub-Gaussian random matrices (cf. Lemma 1). Regarding $\boldsymbol{A}_1$, we know from the definition of $\mathcal{S}_\tau(\cdot)$ that $\|\boldsymbol{A}_1\|_\infty \leq \tau$. More precisely, we have

$$(\boldsymbol{A}_1)_{ij} = \begin{cases} -\tau & \text{if } S^\star_{ij} + E_{ij} \geq \tau, \\ -S^\star_{ij} - E_{ij} & \text{if } -\tau < S^\star_{ij} + E_{ij} < \tau, \\ \tau & \text{if } S^\star_{ij} + E_{ij} \leq -\tau. \end{cases}$$

Recall from Assumption 4 that $\boldsymbol{S}^\star$ has random signs on its support $\Omega^\star \subseteq \Omega_{\mathsf{aug}}$ and E_{ij} is symmetric around zero. It then follows from standard concentration results for the spectral norm of matrices with i.i.d. entries that

$$\|\boldsymbol{A}_1\| \lesssim \tau\sqrt{np\rho_{\mathsf{aug}}} = C_\tau \sigma\sqrt{np\rho_{\mathsf{aug}}\log n},$$

provided that $n^2 p\rho_{\mathsf{aug}} \gg n\log^2 n$. Moving on to $\boldsymbol{A}_2^{t+1}$, since it is supported on Ω_{aug}, we can further decompose $\|\boldsymbol{A}_2^{t+1}\|$ into

$$\left\|\boldsymbol{A}_2^{t+1}\right\| = \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{A}_2^{t+1}\right)\right\| \leq p\rho_{\mathsf{aug}}\left\|\boldsymbol{A}_2^{t+1}\right\| + \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{A}_2^{t+1}\right) - p\rho_{\mathsf{aug}}\boldsymbol{A}_2^{t+1}\right\|. \tag{F.13}$$

Invoking Lemma 5 with $\boldsymbol{A} = \boldsymbol{A}_2^{t+1}$, $\boldsymbol{B} = \boldsymbol{I}_n$ and $\rho_0 = p\rho_{\mathsf{aug}}$, we have

$$\left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{A}_2^{t+1}\right) - p\rho_{\mathsf{aug}}\boldsymbol{A}_2^{t+1}\right\| \leq C\sqrt{np\rho_{\mathsf{aug}}}\left\|\boldsymbol{A}_2^{t+1}\right\|_{2,\infty}, \tag{F.14}$$

with the proviso that $n^2 p\rho_{\mathsf{aug}} \gg n\log n$. Combine the above bounds to reach

$$\left\|\boldsymbol{A}_2^{t+1}\right\| \leq p\rho_{\mathsf{aug}}\left\|\boldsymbol{A}_2^{t+1}\right\| + C\sqrt{np\rho_{\mathsf{aug}}}\left\|\boldsymbol{A}_2^{t+1}\right\|_{2,\infty} \leq \frac{1}{2}\left\|\boldsymbol{A}_2^{t+1}\right\| + C\sqrt{np\rho_{\mathsf{aug}}}\left\|\boldsymbol{A}_2^{t+1}\right\|_{2,\infty},$$

as soon as $\rho_{\mathsf{s}} \leq \rho_{\mathsf{aug}} \leq 1/2$. We are then in need of an upper bound on $\|\boldsymbol{A}_2^{t+1}\|_{2,\infty}$, which is supplied in the following fact.

Fact 3. *Suppose that $n^2 p\rho_{\mathsf{aug}} \gg \mu r n\log n$ and $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} \ll 1/\kappa$. Then with probability exceeding $1 - O(n^{-100})$, one has*

$$\left\|\boldsymbol{A}_2^{t+1}\right\|_{2,\infty} \leq \sqrt{40\kappa p\rho_{\mathsf{aug}}}\left(C_\infty\kappa + 2C_{\mathrm{op}}\right)\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}\|\boldsymbol{X}^\star\|.$$

With the help of Fact 3, we can continue the upper bound as follows

$$\left\|\boldsymbol{A}_2^{t+1}\right\| \leq 2C\sqrt{np\rho_{\mathsf{aug}}}\left\|\boldsymbol{A}_2^{t+1}\right\|_{2,\infty} \lesssim \left(C_\infty + C_{\mathrm{op}}\right)\sqrt{np}\kappa^{3/2}\rho_{\mathsf{aug}}\frac{\sigma}{\sigma_{\min}}\sqrt{n\log n}\,\|\boldsymbol{F}^\star\|_{2,\infty}\|\boldsymbol{X}^\star\|.$$

All in all, we obtain the following bound on $\boldsymbol{A}^{t+1}$:

$$\begin{aligned} \|\boldsymbol{A}^{t+1}\| &\leq \|\mathcal{P}_{\Omega_{\mathsf{aug}}}(\boldsymbol{E})\| + \|\boldsymbol{A}_1\| + \left\|\boldsymbol{A}_2^{t+1}\right\| \\ &\lesssim \sigma\sqrt{np} + C_\tau\sigma\sqrt{np\rho_{\mathsf{aug}}\log n} + \left(C_\infty + C_{\mathrm{op}}\right)\sqrt{np}\kappa^{3/2}\rho_{\mathsf{aug}}\frac{\sigma}{\sigma_{\min}}\sqrt{n\log n}\,\|\boldsymbol{F}^\star\|_{2,\infty}\|\boldsymbol{X}^\star\| \\ &\leq C_S\sigma\sqrt{np}, \end{aligned}$$

with the proviso that $\rho_{\mathsf{s}} \leq \rho_{\mathsf{aug}} \ll 1/\sqrt{\kappa^5\mu r\log^2 n}$. Here, the last line uses the incoherence assumption $\|\boldsymbol{F}^\star\|_{2,\infty} \leq \sqrt{\mu r/n}\|\boldsymbol{X}^\star\|$ (cf. (B.1)).

2. When it comes to $\boldsymbol{B}^{t+1}$, we first note that

$$\begin{aligned} \left\|\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top} - \boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top}\right\|_\infty &= \left\|\left(\boldsymbol{X}^\star - \boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\right)\boldsymbol{Y}^{\star\top} + \boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\left(\boldsymbol{Y}^\star - \boldsymbol{Y}^{t+1}\boldsymbol{H}^{t+1}\right)^\top\right\|_\infty \\ &\leq \left\|\boldsymbol{X}^\star - \boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\right\|_{2,\infty}\|\boldsymbol{Y}^\star\|_{2,\infty} + \left\|\boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\right\|_{2,\infty}\left\|\boldsymbol{Y}^\star - \boldsymbol{Y}^{t+1}\boldsymbol{H}^{t+1}\right\|_{2,\infty} \end{aligned}$$

$$\leq 3C_\infty \kappa \left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p\sigma_{\min}} \right) \|\boldsymbol{F}^\star\|_{2,\infty}^2 \leq 3C_\infty \kappa \left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p\sigma_{\min}} \right) \frac{\mu r}{n} \sigma_{\max}. \tag{F.15}$$

Here, we have plugged in (F.3c) for the $(t+1)$-th iteration and its immediate consequence $\|\boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\|_{2,\infty} \leq \|\boldsymbol{F}^{t+1}\|_{2,\infty} \leq 2\|\boldsymbol{F}^\star\|_{2,\infty}$, as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} \ll 1/\kappa$. As a result, for all (i,j) we have

$$\begin{aligned}
\left|\left(\boldsymbol{M} - \boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top}\right)_{ij}\right| &= \left|\left(\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top} + \boldsymbol{E} - \boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top}\right)_{ij}\right| \leq |E_{ij}| + \left\|\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top} - \boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top}\right\|_\infty \\
&\overset{\text{(i)}}{\leq} |E_{ij}| + 3C_\infty\kappa \left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p\sigma_{\min}} \right) \frac{\mu r}{n}\sigma_{\max} \\
&\overset{\text{(ii)}}{\leq} C_\lambda \sigma \sqrt{\log n} = \tau.
\end{aligned}$$

Here, the inequality (i) comes from (F.15), and the last line (ii) relies on the property of sub-Gaussian random variables (namely, $|E_{ij}| \leq \tau/2$ with probability exceeding $1 - O(n^{-102})$) and the sample size condition $n^2 p \gg \kappa^4 \mu^2 r^2 n \log n$. An immediate consequence is that with probability at least $1 - O(n^{-100})$,

$$\boldsymbol{B}^{t+1} = \mathcal{S}_\tau\left[\mathcal{P}_{\Omega_{\mathsf{obs}}\setminus\Omega_{\mathsf{aug}}}\left(\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top} + \boldsymbol{E} - \boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top}\right)\right] = \boldsymbol{0}. \tag{F.16}$$

Substituting the above two bounds into (F.11), we conclude that $\left\|\boldsymbol{S}^{t+1} - \boldsymbol{S}^\star\right\| \leq C_S \sigma\sqrt{np}$ as claimed.

Proof of Fact 3. In view of the definition of $\boldsymbol{A}^{t+1}$ in (F.12), we have

$$\left\|\boldsymbol{A}_2^{t+1}\right\|_{2,\infty} \leq \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top} - \boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}\right)\right\|_{2,\infty},$$

where we use the non-expansiveness of the proximal operator $\mathcal{S}_\tau(\cdot)$. Apply a similar argument as in bounding (F.10) to obtain

$$\begin{aligned}
&\left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top} - \boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}\right)\right\|_{2,\infty} \\
&\quad\leq \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left[\left(\boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1} - \boldsymbol{X}^\star\right)\boldsymbol{Y}^{\star\top}\right]\right\|_{2,\infty} + \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left[\boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\left(\boldsymbol{Y}^{t+1}\boldsymbol{H}^{t+1} - \boldsymbol{Y}^\star\right)^\top\right]\right\|_{2,\infty} \\
&\quad\leq \sqrt{40\kappa p\rho_{\mathsf{aug}}}\left\|\left(\boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1} - \boldsymbol{X}^\star\right)\boldsymbol{Y}^{\star\top}\right\|_{2,\infty} + \sqrt{40\kappa p\rho_{\mathsf{aug}}}\left\|\boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\left(\boldsymbol{Y}^{t+1}\boldsymbol{H}^{t+1} - \boldsymbol{Y}^\star\right)^\top\right\|_{2,\infty} \\
&\quad\leq \sqrt{40\kappa p\rho_{\mathsf{aug}}}\left(C_\infty\kappa + 2C_{\mathrm{op}}\right)\left(\frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} + \frac{\lambda}{p\sigma_{\min}} \right)\|\boldsymbol{F}^\star\|_{2,\infty}\|\boldsymbol{X}^\star\|,
\end{aligned}$$

as long as $n^2 p\rho_{\mathsf{aug}} \gg \mu r n \log n$. Here, the last line uses the induction hypotheses (F.3b) and (F.3c) for the $(t+1)$-th iteration and their immediate consequence $\|\boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\|_{2,\infty} \leq 2\|\boldsymbol{F}^\star\|_{2,\infty}$, as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} \ll 1/\kappa$. Taking the preceding two bounds together concludes the proof. □

F.5 Proof of Lemma 18

Without loss of generality, we only consider the case when $1 \leq l \leq n$. The case with $n+1 \leq l \leq 2n$ can be derived similarly with very minor modification, and hence we omit it for the sake of brevity.

To begin with, since $(\boldsymbol{H}^{t+1}, \boldsymbol{R}^{t+1,(l)})$ is the choice of the rotation matrix that best aligns $\boldsymbol{F}^{t+1}$ and $\boldsymbol{F}^{t+1,(l)}$, we have

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

In view of the gradient update rule, one has

$$\begin{aligned}
&\boldsymbol{F}^{t+1}\boldsymbol{H}^t - \boldsymbol{F}^{t+1,(l)}\boldsymbol{R}^{t,(l)} \\
&\quad= \left[\boldsymbol{F}^t - \eta\nabla f\left(\boldsymbol{F}^t;\boldsymbol{S}^t\right)\right]\boldsymbol{H}^t - \left[\boldsymbol{F}^{t,(l)} - \eta\nabla f^{(l)}\left(\boldsymbol{F}^{t,(l)};\boldsymbol{S}^{t,(l)}\right)\right]\boldsymbol{R}^{t,(l)}
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{F}^t\boldsymbol{H}^t - \eta\nabla f\left(\boldsymbol{F}^t\boldsymbol{H}^t;\boldsymbol{S}^t\right) - \left[\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)} - \eta\nabla f^{(l)}(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)};\boldsymbol{S}^{t,(l)})\right]\\
&= \underbrace{\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)} - \eta\left[\nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^t\boldsymbol{H}^t;\boldsymbol{S}^t\right) - \nabla f_{\mathsf{aug}}\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)};\boldsymbol{S}^t\right)\right]}_{=:\boldsymbol{C}_1} \underbrace{- \eta\left[\nabla f_{\mathsf{diff}}\left(\boldsymbol{F}^t\boldsymbol{H}^t\right) - \nabla f_{\mathsf{diff}}\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right)\right]}_{=:\boldsymbol{C}_2}\\
&\quad + \underbrace{\eta\left[\nabla f^{(l)}\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)};\boldsymbol{S}^{t,(l)}\right) - \nabla f\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)};\boldsymbol{S}^{t,(l)}\right)\right]}_{=:\boldsymbol{C}_3} + \underbrace{\eta\left[\nabla f\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)};\boldsymbol{S}^{t,(l)}\right) - \nabla f\left(\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)};\boldsymbol{S}^{t}\right)\right]}_{=:\boldsymbol{C}_4}.
\end{aligned}
$$

Here, the second identity relies on the facts that $\nabla f(\boldsymbol{F};\boldsymbol{S})\boldsymbol{R} = \nabla f(\boldsymbol{F}\boldsymbol{R};\boldsymbol{S})$ and $\nabla f^{(l)}(\boldsymbol{F};\boldsymbol{S})\boldsymbol{R} = \nabla f^{(l)}(\boldsymbol{F}\boldsymbol{R};\boldsymbol{S})$ for any orthonormal matrix $\boldsymbol{R}\in\mathcal{O}^{r\times r}$. We shall then control $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, $\boldsymbol{C}_3$ and $\boldsymbol{C}_4$ separately.

Employing the same strategy used to bound $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ in the proof of [CCF+20, Lemma 12], we can demonstrate that

$$
\|\boldsymbol{C}_1\|_{\mathrm{F}} \le \left(1-\frac{\sigma_{\min}}{20}\eta\right)\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} \quad\text{and}\quad \|\boldsymbol{C}_2\|_{\mathrm{F}} \le \eta\left(\sigma\sqrt{\frac{n}{p}}+\frac{\lambda}{p}\right)\|\boldsymbol{F}^\star\|_{2,\infty},
$$

provided that $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll 1/\sqrt{\kappa^4\mu r\log n}$ and $\eta \ll 1/(n\kappa^2\sigma_{\max})$. With regards to $\boldsymbol{C}_3$, it is seen from the definitions of ∇f and $\nabla f^{(l)}$ that

$$
\boldsymbol{C}_3 = \eta\begin{bmatrix} \left[\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)}-\boldsymbol{L}^\star\right) - p^{-1}\mathcal{P}_{(\Omega_{\mathsf{obs}})_{l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)}-\boldsymbol{L}^\star\right)\right]\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} + p^{-1}\mathcal{P}_{(\Omega_{\mathsf{obs}})_{l,\cdot}}(\boldsymbol{E})\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} \\ \left[\mathcal{P}_{l,\cdot}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)}-\boldsymbol{L}^\star\right) - p^{-1}\mathcal{P}_{(\Omega_{\mathsf{obs}})_{l,\cdot}}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{Y}^{t,(l)}-\boldsymbol{L}^\star\right)\right]^\top\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} + p^{-1}\mathcal{P}_{(\Omega_{\mathsf{obs}})_{l,\cdot}}(\boldsymbol{E})^\top\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} \end{bmatrix},
$$

which has the same form as $\boldsymbol{A}_3$ in the proof of [CCF+20, Lemma 12]. It thus follows from [CCF+20, Claim 5, 6 and 7] that

$$
\|\boldsymbol{C}_3\|_{\mathrm{F}} \lesssim \eta\sigma\sqrt{\frac{n\log n}{p}}\|\boldsymbol{F}^\star\|_{2,\infty} + \eta\sqrt{\frac{\mu^2r^2\log n}{np}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{F}^\star\right\|_{2,\infty}\sigma_{\max},
$$

provided that $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}} \ll \frac{1}{\sqrt{\kappa^2\log n}}$ and that $n^2p \gg n\log^3 n$.

We are then left with controlling the term $\boldsymbol{C}_4$. Towards this, we invoke the definition of f to decompose

$$
\begin{aligned}
\boldsymbol{C}_4 &= \eta\begin{bmatrix} p^{-1}\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} \\ p^{-1}\mathcal{P}_{\Omega_{\mathsf{obs}}}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)^\top\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} \end{bmatrix}\\
&= \underbrace{\frac{\eta}{p}\begin{bmatrix} \mathcal{P}_{-l,\cdot}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} \\ \left[\mathcal{P}_{-l,\cdot}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)\right]^\top\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} \end{bmatrix}}_{=:\boldsymbol{D}_1} + \underbrace{\frac{\eta}{p}\begin{bmatrix} \mathcal{P}_{l,\cdot}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} \\ \left[\mathcal{P}_{l,\cdot}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)\right]^\top\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} \end{bmatrix}}_{=:\boldsymbol{D}_2}.
\end{aligned}
$$

Here, we have used the fact that both $\boldsymbol{S}^{t,(l)}$ and $\boldsymbol{S}^t$ are supported on $\Omega^\star\subseteq\Omega_{\mathsf{obs}}$. Regarding the first matrix $\boldsymbol{D}_1$, we have the following fact.

Fact 4. *Suppose that the sample size obeys $n^2p \gg \kappa^4\mu^2r^2n\log n$, the noise satisfies $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} \ll 1/\kappa$, the outlier fraction satisfies $\rho_{\mathsf{s}} \le \rho_{\mathsf{aug}} \ll 1/\kappa^3$ and $n^2p\rho_{\mathsf{aug}} \gg \mu rn\log n$ hold. Then with probability at least $1-O(n^-100)$, we have*

$$
\left\|\boldsymbol{D}_1\right\|_{\mathrm{F}} \lesssim \eta\sigma\sqrt{\frac{n\log n}{p}}\|\boldsymbol{F}^\star\|_{2,\infty}.
$$

With regards to $\boldsymbol{D}_2$, recall that $\boldsymbol{S}^{t,(l)}_{l,\cdot} = \boldsymbol{S}^\star_{l,\cdot}$. Using the decomposition (F.11) in the proof of Lemma 17, and recalling that $\boldsymbol{B}^{t+1}=\boldsymbol{0}$ from the proof of Lemma 17, we obtain

$$
\mathcal{P}_{l,\cdot}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} = \mathcal{P}_{l,\cdot}\left(\boldsymbol{A}_1+\boldsymbol{E}\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)} + \mathcal{P}_{l,\cdot}\left(\boldsymbol{A}_2^t\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}, \tag{F.17}
$$

where

$$\boldsymbol{A}_1 = \mathcal{S}_\tau\left[\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{S}^\star + \boldsymbol{E}\right)\right] - \mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{S}^\star + \boldsymbol{E}\right);$$
$$\boldsymbol{A}_2^t \coloneqq \mathcal{S}_\tau\left[\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top} - \boldsymbol{X}^t\boldsymbol{Y}^{t\top} + \boldsymbol{S}^\star + \boldsymbol{E}\right)\right] - \mathcal{S}_\tau\left[\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{S}^\star + \boldsymbol{E}\right)\right].$$

For the first term $\mathcal{P}_{l,\cdot}(\boldsymbol{A}_1 + \boldsymbol{E})\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}$, the independence between $\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}$ and the l-th row of $\boldsymbol{A}_1 + \boldsymbol{E}$ allows us to obtain the following bound.

Fact 5. *Suppose that $\rho_{\mathsf{s}} \le \rho_{\mathsf{aug}} \ll 1/\log n$ and that $n^2 p \gg n\log^4 n$. Then with probability at least $1 - O(n^-100)$, we have*

$$\left\|\mathcal{P}_{l,\cdot}\left(\boldsymbol{A}_1\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} \lesssim \sigma\sqrt{np\log n}\left\|\boldsymbol{Y}^\star\right\|_{2,\infty}.$$

The term involving $\boldsymbol{A}_2^t$ is controlled in the following claim, which relies heavily on the small scale of the entries in $\boldsymbol{A}_2^t$.

Fact 6. *Suppose that $n \gg \kappa\mu r$, $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} \ll 1/\kappa$, $\rho_{\mathsf{s}} \le \rho_{\mathsf{aug}} \ll 1/(\kappa\mu r)$ and that $n^2 p\rho_{\mathsf{aug}} \gg n\log n$. Then with probability at least $1 - O(n^-100)$, we have*

$$\left\|\mathcal{P}_{l,\cdot}\left(\boldsymbol{A}_2^t\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} \lesssim \sigma\sqrt{np\log n}\left\|\boldsymbol{Y}^\star\right\|_{2,\infty}.$$

Combining the two bounds in Facts 5 and 6 gives

$$\left\|\mathcal{P}_{l,\cdot}(\boldsymbol{S}^{t,(l)} - \boldsymbol{S}^t)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} \lesssim \sigma\sqrt{np\log n}\left\|\boldsymbol{Y}^\star\right\|_{2,\infty}.$$

The same bound applies to $\|\mathcal{P}_{l,\cdot}(\boldsymbol{S}^{t,(l)} - \boldsymbol{S}^t)^\top\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\|_{\mathrm{F}}$ via the same technique. As a result, we have

$$\|\boldsymbol{D}_2\|_{\mathrm{F}} \lesssim \eta\sigma\sqrt{\frac{n\log n}{p}}\|\boldsymbol{F}^\star\|_{2,\infty}.$$

Putting the above bounds together yields

$$\begin{aligned}
&\left\|\boldsymbol{F}^{t+1}\boldsymbol{H}^{t+1} - \boldsymbol{F}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)}\right\|_{\mathrm{F}} \le \|\boldsymbol{C}_1\|_{\mathrm{F}} + \|\boldsymbol{C}_2\|_{\mathrm{F}} + \|\boldsymbol{C}_3\|_{\mathrm{F}} + \|\boldsymbol{D}_1\|_{\mathrm{F}} + \|\boldsymbol{D}_2\|_{\mathrm{F}} \\
&\quad\le \left(1 - \frac{\sigma_{\min}}{20}\eta\right)\left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} + \eta\left(\sigma\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\|\boldsymbol{F}^\star\|_{2,\infty} \\
&\qquad + \tilde{C}\left(\eta\sigma\sqrt{\frac{n\log n}{p}}\|\boldsymbol{F}^\star\|_{2,\infty} + \eta\sqrt{\frac{\mu^2 r^2\log n}{np}}\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{F}^\star\right\|_{2,\infty}\sigma_{\max}\right) + \tilde{C}\eta\sigma\sqrt{\frac{n\log n}{p}}\|\boldsymbol{F}^\star\|_{2,\infty} \\
&\quad\overset{\text{(i)}}{\le} \left(1 - \frac{\sigma_{\min}}{20}\eta\right)C_1\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty} + \eta\left(\sigma\sqrt{\frac{n}{p}} + \frac{\lambda}{p}\right)\|\boldsymbol{F}^\star\|_{2,\infty} + \tilde{C}\frac{\lambda}{p}\|\boldsymbol{F}^\star\|_{2,\infty} \\
&\qquad + \tilde{C}\eta\sqrt{\frac{\mu^2 r^2\log n}{np}}(C_\infty\kappa + C_1)\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}\sigma_{\max} + \tilde{C}\eta\sigma\sqrt{\frac{n\log n}{p}}\|\boldsymbol{F}^\star\|_{2,\infty} \\
&\quad\overset{\text{(ii)}}{\le} C_1\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty},
\end{aligned}$$

where (i) invokes (F.6a) and its immediate consequence that

$$\begin{aligned}
\left\|\boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{F}^\star\right\|_{2,\infty} &\le \left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} + \left\|\boldsymbol{F}^t\boldsymbol{H}^t - \boldsymbol{F}^\star\right\|_{2,\infty} && \text{(F.18)} \\
&\le (C_\infty\kappa + C_1)\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\|\boldsymbol{F}^\star\|_{2,\infty}. && \text{(F.19)}
\end{aligned}$$

The last line (ii) holds as long as $n^2 p \gg \kappa^4\mu^2 r^2 n\log n$ and C_1 is large enough.

Proof of Fact 4. First notice that $\boldsymbol{S}^t$ is supported on Ω_{aug}, which is a consequence of (F.11) and (F.16) as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} \ll 1/\kappa$ and $n^2p \gg \kappa^4\mu^2r^2n\log n$. By replacing $\boldsymbol{X}^{t+1}$ (resp. $\boldsymbol{Y}^{t+1}$) with $\boldsymbol{X}^{t+1,(l)}$ (resp. $\boldsymbol{Y}^{t+1,(l)}$)) and invoking (F.19) instead of (F.3c), the same arguments yield the fact that $\boldsymbol{S}^{t,(l)}$ is also supported on Ω_{aug}. Define $\omega_{ij} \coloneqq \mathbb{1}_{(i,j)\in\Omega_{\mathsf{aug}}}$. The Frobenius norm of the upper block of $\boldsymbol{D}_1$ can be bounded by

$$\begin{aligned}
&\left\|\mathcal{P}_{-l,\cdot}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}^2 = \sum_{i:i\neq l}\sum_{j=1}^{r}\left[\sum_{k=1}^{n}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)_{ik}Y_{kj}^{t,(l)}\right]^2 \\
&\quad= \sum_{i:i\neq l}\sum_{j=1}^{r}\left[\sum_{k=1}^{n}\omega_{ik}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)_{ik}Y_{kj}^{t,(l)}\right]^2 \le \sum_{i:i\neq l}\sum_{j=1}^{r}\left[\sum_{k=1}^{n}\omega_{ik}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)_{ik}^2\right]\left[\sum_{k=1}^{n}\omega_{ik}\big(Y_{kj}^{t,(l)}\big)^2\right],
\end{aligned}$$

where we use the Cauchy-Schwarz inequality in the last step. Converting to the matrix notation, we obtain

$$\sum_{k=1}^{n}\omega_{ik}\big(Y_{kj}^{t,(l)}\big)^2 = \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{Y}^{t,(l)\top}\right)\right\|_{\mathrm{F}}^2.$$

Applying a similar argument as in bounding (F.10), one can obtain from Lemma 4 that

$$\sum_{j=1}^{r}\left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{Y}^{t,(l)\top}\right)\right\|_{\mathrm{F}}^2 \lesssim \kappa p\rho_{\mathsf{aug}}\sum_{j=1}^{r}\|\boldsymbol{Y}_{\cdot,j}^{t,(l)}\|_{\mathrm{F}}^2 = \kappa p\rho_{\mathsf{aug}}\|\boldsymbol{Y}^{t,(l)}\|_{\mathrm{F}}^2,$$

provided that $n^2p\rho_{\mathsf{aug}} \gg \mu rn\log n$. This allows us to reach

$$\begin{aligned}
\left\|\mathcal{P}_{-l,\cdot}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} &\lesssim \sqrt{\sum_{i:i\neq l}\left[\sum_{k=1}^{n}\omega_{ik}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)_{ik}^2\right]}\cdot\sqrt{\kappa p\rho_{\mathsf{aug}}}\|\boldsymbol{Y}^{t,(l)}\|_{\mathrm{F}} \\
&= \sqrt{\kappa p\rho_{\mathsf{aug}}}\left\|\mathcal{P}_{-l,\cdot}\left(\boldsymbol{S}^{t,(l)}-\boldsymbol{S}^t\right)\right\|_{\mathrm{F}}\left\|\boldsymbol{Y}^{t,(l)}\right\|_{\mathrm{F}} \\
&\le C_3\sqrt{\kappa p\rho_{\mathsf{aug}}}\frac{\sigma}{\sigma_{\min}}\sqrt{n\log n}\left\|\boldsymbol{F}^\star\right\|^2\left\|\boldsymbol{F}^\star\right\|_{2,\infty} \\
&\lesssim \sigma\sqrt{np\log n}\left\|\boldsymbol{F}^\star\right\|_{2,\infty}.
\end{aligned}$$

Here, the penultimate step comes from the hypothesis (F.6c), whereas the last step holds as long as $\rho_{\mathsf{s}} \le \rho_{\mathsf{aug}} \ll 1/\kappa^3$. The Frobenius norm of the lower block of $\boldsymbol{D}_1$ admits the same bound. As a result, we obtain $\left\|\boldsymbol{D}_1\right\|_{\mathrm{F}} \lesssim \eta\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{F}^\star\right\|_{2,\infty}$ as claimed. $\square$

Proof of Fact 5. Regarding the first term on the right-hand side of (F.17), we can write

$$\left\|\mathcal{P}_{l,\cdot}\left(\boldsymbol{A}_1+\boldsymbol{E}\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} = \left\|\sum\nolimits_{j=1}^{n}\left(\boldsymbol{A}_1+\boldsymbol{E}\right)_{lj}\boldsymbol{Y}_{j,\cdot}^{t,(l)}\right\|_2 = \Big\|\sum\nolimits_{j=1}^{n}\underbrace{\omega_{lj}\left[\mathcal{S}_\tau(S_{lj}^\star+E_{ij})-S_{lj}^\star\right]\boldsymbol{Y}_{j,\cdot}^{t,(l)}}_{=:\boldsymbol{u}_j}\Big\|_2,$$

where $\omega_{lj} \coloneqq \mathbb{1}\{(l,j)\in\Omega_{\mathsf{aug}}\}$ is a Bernoulli random variable with mean $p\rho_{\mathsf{aug}}$. Since $\boldsymbol{Y}^{t,(l)}$ is independent of $\{\omega_{lj}\}_{1\le j\le n}$ and $\boldsymbol{S}_{l,\cdot}^\star$, the vectors $\{\boldsymbol{u}_j\}_{j=1}^n$ are statistically independent conditional on $\boldsymbol{Y}^{t,(l)}$. We can thus apply the matrix Bernstein inequality to control this term. Specifically, conditional on $\boldsymbol{Y}^{t,(l)}$, we have

$$\left\|\|\boldsymbol{u}_j\|_2\right\|_{\psi_1} \le \left\|\boldsymbol{Y}^{t,(l)}\right\|_{2,\infty}\left\|\omega_{lj}\left[\mathcal{S}_\tau(S_{lj}^\star+E_{ij})-S_{lj}^\star\right]\right\|_{\psi_1} \overset{\text{(i)}}{\lesssim} \tau\left\|\boldsymbol{Y}^{t,(l)}\right\|_{2,\infty},$$

$$V \coloneqq \left\|\mathbb{E}\left[\sum\nolimits_{j=1}^{n}\omega_{lj}^2\left(\mathcal{S}_\tau(S_{lj}^\star+E_{ij})-S_{lj}^\star\right)^2\boldsymbol{Y}_{j,\cdot}^{t,(l)}\boldsymbol{Y}_{j,\cdot}^{t,(l)\top}\right]\right\| \overset{\text{(ii)}}{\lesssim} p\rho_{\mathsf{aug}}\tau^2\left\|\boldsymbol{Y}^{t,(l)}\right\|_{\mathrm{F}}^2,$$

where $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm [Ver17]. Here, the relation (i) holds since

$$\left\|\omega_{lj}\left[\mathcal{S}_\tau(S_{lj}^\star+E_{ij})-S_{lj}^\star\right]\right\|_{\psi_1} \le \left\|\mathcal{S}_\tau(S_{lj}^\star+E_{ij})-S_{lj}^\star\right\|_{\psi_1} \le \left\|\mathcal{S}_\tau(S_{lj}^\star+E_{ij})-\left(S_{lj}^\star+E_{ij}\right)\right\|_{\psi_1} + \|E_{ij}\|_{\psi_1}$$

$$\leq \left|\mathcal{S}_\tau(S^\star_{lj}+E_{ij})-\left(S^\star_{lj}+E_{ij}\right)\right|+\|E_{ij}\|_{\psi_2}\leq 2\tau,$$

where we have used the fact that $|\mathcal{S}_\tau(x)-x|\leq\tau$ and $\|E_{ij}\|_{\psi_1}\leq\|E_{ij}\|_{\psi_2}\leq\sigma\leq\tau$. In addition, the second inequality (ii) comes from the identity $\mathbb{E}[\omega_{lj}^2]=p\rho_{\mathsf{aug}}$ and the fact that

$$\mathbb{E}\left[\left(\mathcal{S}_\tau(S^\star_{lj}+E_{ij})-S^\star_{lj}\right)^2\right]\leq 2\mathbb{E}\left[\left(\mathcal{S}_\tau(S^\star_{lj}+E_{ij})-S^\star_{lj}-E_{ij}\right)^2\right]+2\mathbb{E}\left[E_{ij}^2\right]\lesssim\tau^2.$$

With the aid of the above bounds, we can invoke the matrix Bernstein inequality [KLT11, Proposition 2] to reach

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

with probability at least $1-O(n^{-10})$. Here, the last inequality arises from $\|\boldsymbol{Y}^{t,(l)}\|_{\mathrm{F}}^2\leq n\|\boldsymbol{Y}^{t,(l)}\|_{2,\infty}^2$. Consequently, we conclude that, with high probability,

$$\left\|\mathcal{P}_{l,\cdot}\left(\boldsymbol{A}_1+\boldsymbol{E}\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}\lesssim\left(\tau\sqrt{np\rho_{\mathsf{aug}}\log n}+\tau\log^2 n\right)\left\|\boldsymbol{Y}^{t,(l)}\right\|_{2,\infty}\lesssim\sigma\sqrt{np\log n}\left\|\boldsymbol{Y}^\star\right\|_{2,\infty},$$

with the proviso that $\rho_{\mathsf{s}}\leq\rho_{\mathsf{aug}}\ll 1/\log n$ and $n^2p\gg n\log^4 n$. □

Proof of Fact 6. Regarding the second term on the right-hand side of (F.17), we have

$$\begin{aligned}\left\|\mathcal{P}_{l,\cdot}\left(\boldsymbol{A}_2^t\right)\boldsymbol{Y}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}&=\left\|\sum\nolimits_{j=1}^n\left(\boldsymbol{A}_2^t\right)_{lj}\boldsymbol{Y}_{j,\cdot}^{t,(l)}\right\|_2\\&\overset{\text{(i)}}{\leq}2np\rho_{\mathsf{aug}}\left\|\boldsymbol{A}_2^t\right\|_\infty\left\|\boldsymbol{Y}^{t,(l)}\right\|_{2,\infty}\\&\overset{\text{(ii)}}{\leq}12np\rho_{\mathsf{aug}}C_\infty\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\sigma_{\min}}\right)\frac{\mu r}{n}\sigma_{\max}\left\|\boldsymbol{Y}^\star\right\|_{2,\infty}\\&\overset{\text{(iii)}}{\lesssim}\sigma\sqrt{np\log n}\left\|\boldsymbol{Y}^\star\right\|_{2,\infty}.\end{aligned}$$

Here, the first upper bound (i) arises from the fact that $\{j\mid(\boldsymbol{A}_2^t)_{lj}\neq 0\}\subseteq\{j\mid(l,j)\in\Omega_{\mathsf{aug}}\}$, whose cardinality is upper bounded by $2np\rho_{\mathsf{aug}}$ with high probability as long as $np\rho_{\mathsf{aug}}\gg\log n$. The second inequality (ii) comes from the simple fact that $\|\boldsymbol{Y}^{t,(l)}\|_{2,\infty}\leq 2\|\boldsymbol{Y}^\star\|_{2,\infty}$ as well as the bound

$$\begin{aligned}\left\|\boldsymbol{A}_2^t\right\|_\infty&=\left\|\mathcal{S}_\tau\left[\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}-\boldsymbol{X}^t\boldsymbol{Y}^{t\top}+\boldsymbol{S}^\star+\boldsymbol{E}\right)\right]-\mathcal{S}_\tau\left[\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{S}^\star+\boldsymbol{E}\right)\right]\right\|_\infty\\&\leq\left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}-\boldsymbol{X}^t\boldsymbol{Y}^{t\top}+\boldsymbol{S}^\star+\boldsymbol{E}\right)-\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{S}^\star+\boldsymbol{E}\right)\right\|_\infty\\&\leq\left\|\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top}-\boldsymbol{X}^t\boldsymbol{Y}^{t\top}\right\|_\infty\\&\leq 3C_\infty\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}+\frac{\lambda}{p\sigma_{\min}}\right)\frac{\mu r}{n}\sigma_{\max},\end{aligned}$$

where we use the non-expansiveness of $\mathcal{S}_\tau(\cdot)$ and the established bound (F.15), which holds as long as $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\ll 1/\kappa$. Last but not least, the relation (iii) holds as long as $\rho_{\mathsf{s}}\leq\rho_{\mathsf{aug}}\ll 1/(\kappa\mu r)$ and $n\gg\kappa\mu r$. □

F.6 Proof of Lemma 19

Without loss of generality, we assume $1\leq l\leq n$. Following the definitions of $\boldsymbol{S}^{t+1,(l)}$ and $\boldsymbol{S}^{t+1}$, we have

$$\left\|\mathcal{P}_{-l,\cdot}\left(\boldsymbol{S}^{t+1,(l)}-\boldsymbol{S}^{t+1}\right)\right\|_{\mathrm{F}}=\left\|\mathcal{P}_{-l,\cdot}\left[\mathcal{S}_\tau\left(\boldsymbol{M}-\boldsymbol{X}^{t+1,(l)}\boldsymbol{Y}^{t+1,(l)\top}\right)-\mathcal{S}_\tau\left(\boldsymbol{M}-\boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top}\right)\right]\right\|_{\mathrm{F}}$$

$$\leq \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} + \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}^{\mathrm{c}}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}}, \tag{F.20}$$

where we denote $\boldsymbol{\Delta} \coloneqq \mathcal{S}_{\tau}(\boldsymbol{M} - \boldsymbol{X}^{t+1,(l)}\boldsymbol{Y}^{t+1,(l)\top}) - \mathcal{S}_{\tau}(\boldsymbol{M} - \boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top})$. Recall from Appendix A that each (i,j) is included in Ω_{aug} independently with probability $p\rho_{\mathsf{aug}}$, where $1 \geq \rho_{\mathsf{aug}} \geq \rho_{\mathsf{s}}$.

1. For the first term $\|\mathcal{P}_{\Omega_{\mathsf{aug}}}(\boldsymbol{\Delta})\|_{\mathrm{F}}$, the non-expansiveness of the proximal operator $\mathcal{S}_{\tau}(\cdot)$ yields

$$\left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} \leq \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{X}^{t+1,(l)}\boldsymbol{Y}^{t+1,(l)\top} - \boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top}\right)\right\|_{\mathrm{F}}.$$

Apply Lemma 4 and a similar argument in bounding (F.10) to obtain

$$\begin{aligned}
\left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} &\leq \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left[\boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\left(\boldsymbol{Y}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)} - \boldsymbol{Y}^{t+1}\boldsymbol{H}^{t+1}\right)^{\top}\right]\right\|_{\mathrm{F}} \\
&\quad + \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left[\left(\boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)} - \boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\right)\boldsymbol{R}^{t+1,(l)\top}\boldsymbol{Y}^{t+1,(l)\top}\right]\right\|_{\mathrm{F}} \\
&\lesssim \sqrt{\kappa p\rho_{\mathsf{aug}}}\left\|\boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\right\|\left\|\boldsymbol{Y}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)} - \boldsymbol{Y}^{t+1}\boldsymbol{H}^{t+1}\right\|_{\mathrm{F}} \\
&\quad + \sqrt{\kappa p\rho_{\mathsf{aug}}}\left\|\boldsymbol{Y}^{t+1,(l)}\boldsymbol{H}^{t+1,(l)}\right\|\left\|\boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)} - \boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\right\|_{\mathrm{F}},
\end{aligned}$$

with the proviso that $n^2 p\rho_{\mathsf{aug}} \gg \mu r n \log n$. In view of (F.6a) and the simple facts $\|\boldsymbol{X}^{t+1}\boldsymbol{H}^{t+1}\| \leq 2\|\boldsymbol{X}^{\star}\|$, $\|\boldsymbol{Y}^{t+1,(l)}\boldsymbol{H}^{t+1,(l)}\| \leq 2\|\boldsymbol{X}^{\star}\|$, one has

$$\begin{aligned}
\left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} &\lesssim \sqrt{\kappa p\rho_{\mathsf{aug}}}\left\|\boldsymbol{X}^{\star}\right\|\left(\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} + \frac{\lambda}{p\sigma_{\min}}\right)\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty} \\
&\leq C_3\frac{\sigma}{\sigma_{\min}}\sqrt{n\log n}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{F}^{\star}\right\|,
\end{aligned}$$

provided that $\rho_{\mathsf{aug}} \ll 1/\kappa$.

2. Regarding the second term $\|\mathcal{P}_{\Omega_{\mathsf{aug}}^{\mathrm{c}}}(\boldsymbol{\Delta})\|_{\mathrm{F}}$, we first recall from (F.16) that

$$\mathcal{S}_{\tau}\left[\mathcal{P}_{\Omega_{\mathsf{aug}}^{\mathrm{c}}}\left(\boldsymbol{M} - \boldsymbol{X}^{t+1}\boldsymbol{Y}^{t+1\top}\right)\right] = \boldsymbol{0}.$$

By replacing $\boldsymbol{X}^{t+1}$ (resp. $\boldsymbol{Y}^{t+1}$) with $\boldsymbol{X}^{t+1,(l)}$ (resp. $\boldsymbol{Y}^{t+1,(l)}$)) and invoking (F.19) instead of (F.3c), the same arguments that we used to prove (F.16) also allow us to demonstrate

$$\mathcal{S}_{\tau}\left[\mathcal{P}_{\Omega_{\mathsf{aug}}^{\mathrm{c}}}\left(\boldsymbol{M} - \boldsymbol{X}^{t+1,(l)}\boldsymbol{Y}^{t+1,(l)\top}\right)\right] = \boldsymbol{0}$$

provided that $n^2 p \gg \kappa^4\mu^2 r^2 n\log n$ and $\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}} \ll 1/\kappa$. Consequently, we have $\mathcal{P}_{\Omega_{\mathsf{aug}}^{\mathrm{c}}}(\boldsymbol{\Delta}) = \boldsymbol{0}$.

Substituting the above two bounds into (F.20), we conclude that

$$\left\|\mathcal{P}_{-l,\cdot}\left(\boldsymbol{S}^{t+1,(l)} - \boldsymbol{S}^{t+1}\right)\right\|_{\mathrm{F}} \leq \left\|\mathcal{P}_{\Omega_{\mathsf{aug}}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} \leq C_3\frac{\sigma}{\sigma_{\min}}\sqrt{n\log n}\left\|\boldsymbol{F}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{F}^{\star}\right\|.$$

F.7 Proof of Lemma 20

Following [CCF+20, Lemma 18], we already know that

$$f\left(\boldsymbol{X}^{t+1}, \boldsymbol{Y}^{t+1}; \boldsymbol{S}^{t}\right) \leq f\left(\boldsymbol{X}^{t}, \boldsymbol{Y}^{t}; \boldsymbol{S}^{t}\right) - \frac{\eta}{2}\left\|\nabla f\left(\boldsymbol{X}^{t}, \boldsymbol{Y}^{t}; \boldsymbol{S}^{t}\right)\right\|_{\mathrm{F}}^{2}. \tag{F.21}$$

As a result, one has

$$F\left(\boldsymbol{X}^{t+1}, \boldsymbol{Y}^{t+1}, \boldsymbol{S}^{t+1}\right) \overset{\text{(i)}}{\leq} F\left(\boldsymbol{X}^{t+1}, \boldsymbol{Y}^{t+1}, \boldsymbol{S}^{t}\right) = f\left(\boldsymbol{X}^{t+1}, \boldsymbol{Y}^{t+1}; \boldsymbol{S}^{t}\right) + \tau\left\|\boldsymbol{S}^{t}\right\|_{1}$$

$$
\begin{aligned}
&\overset{\text{(ii)}}{\leq} f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t; \boldsymbol{S}^t\right) - \frac{\eta}{2}\left\|\nabla f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t; \boldsymbol{S}^t\right)\right\|_{\mathrm{F}}^2 + \tau\left\|\boldsymbol{S}^t\right\|_1 \\
&= F\left(\boldsymbol{X}^t, \boldsymbol{Y}^t, \boldsymbol{S}^t\right) - \frac{\eta}{2}\left\|\nabla f\left(\boldsymbol{X}^t, \boldsymbol{Y}^t; \boldsymbol{S}^t\right)\right\|_{\mathrm{F}}^2,
\end{aligned}
$$

where (i) follows since, by construction, $\boldsymbol{S}^{t+1}$ is the minimizer of $F(\boldsymbol{X}^{t+1}, \boldsymbol{Y}^{t+1}, \boldsymbol{S})$ for any given $(\boldsymbol{X}^{t+1}, \boldsymbol{Y}^{t+1})$, and (ii) arises from (F.21).

References

[ANW12] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197, 2012.

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

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

[CC17] Yuxin Chen and Emmanuel J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Comm. Pure Appl. Math.*, 70(5):822–883, 2017.

[CC18] Yuxin Chen and Emmanuel Candès. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Communications on Pure and Applied Mathematics*, 71(8):1648–1714, 2018.

[CCD+19] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv preprint arXiv:1904.10020*, 2019.

[CCF+20] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020.

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

[CCFM20] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral methods for data science: A statistical perspective. *arXiv preprint arXiv:2012.08496*, 2020.

[CCW19] HanQin Cai, Jian-Feng Cai, and Ke Wei. Accelerated alternating projections for robust principal component analysis. *The Journal of Machine Learning Research*, 20(1):685–717, 2019.

[CDDD19] Vasileios Charisopoulos, Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Composite optimization for robust blind deconvolution. *arXiv preprint arXiv:1901.01624*, 2019.

[CFMY19] Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, 2019.

[CFWY20] Yuxin Chen, Jianqing Fan, Bingyan Wang, and Yuling Yan. Convex and nonconvex optimization are both minimax-optimal for noisy blind deconvolution. *arXiv preprint arXiv:2008.01724*, 2020.

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

[CGJ17] Yeshwanth Cherapanamjeri, Kartik Gupta, and Prateek Jain. Nearly optimal robust matrix completion. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 797–805. JMLR. org, 2017.

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

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

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

[CLL20] Ji Chen, Dekai Liu, and Xiaodong Li. Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *IEEE Transactions on Information Theory*, 2020.

[CLMW11] Emmanuel Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of ACM*, 58(3):11:1–11:37, Jun 2011.

[CLPC20] Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. Nonconvex low-rank tensor completion from noisy data. *Accepted to Operations Research*, 2020.

[CLS15] E. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, April 2015.

[CMW13] T Tony Cai, Zongming Ma, and Yihong Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

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

[CPW12] Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. *Annals of Statistics*, 40(4):1935–1967, 2012.

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

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

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

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

[CW18] Jian-Feng Cai and Ke Wei. Solving systems of phaseless equations via riemannian optimization with optimal sampling complexity. *arXiv preprint arXiv:1809.02773*, 2018.

[DC20] Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 2020.

[DG14] David Donoho and Matan Gavish. Minimax risk of matrix denoising by singular value thresholding. *The Annals of Statistics*, 42(6):2413–2440, 2014.

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

[FFL08] Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

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

[FSZZ18] Jianqing Fan, Qiang Sun, Wen-Xin Zhou, and Ziwei Zhu. Principal component analysis for big data. *arXiv preprint arXiv:1801.01602*, 2018.

[FWZ18] Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An ℓ_∞ eigenvector perturbation bound and its application. *Journal of Machine Learning Research*, 18(207):1–42, 2018.

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

[FXY13] Jiashi Feng, Huan Xu, and Shuicheng Yan. Online robust PCA via stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 404–412, 2013.

[GMS13] Donald Goldfarb, Shiqian Ma, and Katya Scheinberg. Fast alternating linearization methods for minimizing the sum of two convex functions. *Mathematical Programming*, 141(1-2):349–382, 2013.

[GQV14] Han Guo, Chenlu Qiu, and Namrata Vaswani. An online algorithm for separating sparse and low-dimensional signal sequences from their sum. *IEEE Transactions on Signal Processing*, 62(16):4284–4297, 2014.

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

[GWL+10] Arvind Ganesh, John Wright, Xiaodong Li, Emmanuel J Candes, and Yi Ma. Dense error correction for low-rank matrices via principal component pursuit. In *2010 IEEE international symposium on information theory*, pages 1513–1517. IEEE, 2010.

[GWL16] Quanquan Gu, Zhaoran Wang Wang, and Han Liu. Low-rank and sparse structure pursuit via alternating minimization. In *Artificial Intelligence and Statistics*, pages 600–609, 2016.

[HG13] Q. Huang and L. Guibas. Consistent shape maps via semidefinite programming. *Computer Graphics Forum*, 32(5):177–186, 2013.

[HKZ11] Daniel Hsu, Sham M Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, 2011.

[JCSX11] Ali Jalali, Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. In *International Conference on Machine Learning*, volume 11, pages 1001–1008, 2011.

[JNS13] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *ACM symposium on Theory of computing*, pages 665–674, 2013.

[Jol11] Ian Jolliffe. *Principal component analysis*. Springer, 2011.

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

[KLT11] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

[KLT17] Olga Klopp, Karim Lounici, and Alexandre B Tsybakov. Robust matrix completion. *Probability Theory and Related Fields*, 169(1-2):523–564, 2017.

[KMO10] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980 –2998, June 2010.

[KS20] Felix Krahmer and Dominik Stöger. On the convex geometry of blind deconvolution and matrix completion. *Communications on Pure and Applied Mathematics*, 2020.

[Li13] Xiaodong Li. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37:73–99, 2013.

[LMCC19] Yuanxin Li, Cong Ma, Yuxin Chen, and Yuejie Chi. Nonconvex matrix factorization from rank-one measurements. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1496–1505, 2019.

[LWC+20] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Neural Information Processing Systems*, 2020.

[MA18] Shiqian Ma and Necdet Serhat Aybat. Efficient optimization algorithms for robust principal component analysis and its variants. *Proceedings of the IEEE*, 106(8):1411–1426, 2018.

[MHT10] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

[MWCC20] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632, 2020.

[NJS13] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[NNS+14] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

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

[Pea01] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[QV10] Chenlu Qiu and Namrata Vaswani. Real-time robust principal components' pursuit. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 591–598. IEEE, 2010.

[QVLH14] Chenlu Qiu, Namrata Vaswani, Brian Lois, and Leslie Hogben. Recursive robust pca or recursive sparse recovery in large but structured noise. *IEEE Transactions on Information Theory*, 60(8):5007–5039, 2014.

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

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

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

[SWZ14] Yuan Shen, Zaiwen Wen, and Yin Zhang. Augmented lagrangian alternating direction method for matrix separation based on low-rank factorization. *Optimization Methods and Software*, 29(2):239–263, 2014.

[TY11] Min Tao and Xiaoming Yuan. Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM Journal on Optimization*, 21(1):57–81, 2011.

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

[Ver17] Roman Vershynin. High dimensional probability, 2017.

[VN18] Namrata Vaswani and Praneeth Narayanamurthy. Static and dynamic robust pca and matrix completion: A review. *Proceedings of the IEEE*, 106(8):1359–1379, 2018.

[WCCL16] K. Wei, J.F. Cai, T. Chan, and S. Leung. Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

[WGE17] Gang Wang, Georgios B Giannakis, and Yonina C Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 2017.

[WL17] Raymond KW Wong and Thomas Lee. Matrix completion with noisy entries and outliers. *The Journal of Machine Learning Research*, 18(1):5404–5428, 2017.

[YPCC16] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. In *NIPS*, pages 4152–4160, 2016.

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

[ZCL16] Huishuai Zhang, Yuejie Chi, and Yingbin Liang. Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow. In *International conference on machine learning*, pages 1022–1031, 2016.

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

[ZLGV16] Jinchun Zhan, Brian Lois, Han Guo, and Namrata Vaswani. Online (and offline) robust PCA: Novel algorithms and performance guarantees. In *Artificial intelligence and statistics*, pages 1488–1496, 2016.

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

[ZWG18] Xiao Zhang, Lingxiao Wang Wang, and Quanquan Gu. A unified framework for nonconvex low-rank plus sparse matrix recovery. In *International Conference on Artificial Intelligence and Statistics*, 2018.