Optimal Tuning-Free Convex Relaxation for Noisy Matrix Completion

Yuepeng Yang and Cong Ma

Abstract—This paper is concerned with noisy matrix completion—the problem of recovering a low-rank matrix from partial and noisy entries. Under uniform sampling and incoherence assumptions, we prove that a tuning-free square-root matrix completion estimator (square-root MC) achieves optimal statistical performance for solving the noisy matrix completion problem. Similar to the square-root Lasso estimator in high-dimensional linear regression, square-root MC does not rely on the knowledge of the size of the noise. While solving square-root MC is a convex program, our statistical analysis of square-root MC hinges on its intimate connections to a nonconvex rank-constrained estimator.

Index Terms—Matrix completion, tuning-free, convex relaxation.

I. INTRODUCTION

LOW-RANK matrix completion [1], [2] aims to reconstruct a low-rank data matrix from its partially observed entries. This problem finds numerous applications in collaborative filtering [3], causal inference [4], sensor network localization [5], etc.

In this paper, we focus on the noisy matrix completion problem, in which the revealed entries are further corrupted by random noise. Mathematically, let $L^* \in \mathbb{R}^{n \times n}$ be a rank-$r$ matrix of interest, and $E \in \mathbb{R}^{n \times n}$ denotes the noise matrix. We observe a subset of entries

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

where $\Omega \subseteq \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}$ represents the index set of the observations. The goal of noisy matrix completion is to recover the underlying low-rank matrix $L^*$ given the observation $M = [M_{ij}]$.

Arguably, one of the most natural approaches to solving noisy matrix completion is the following nuclear norm regularized least-squares estimator [6], [7]:

$$\min_{L \in \mathbb{R}^{n \times n}} \sum_{(i,j) \in \Omega} (L_{ij} - M_{ij})^2 + \lambda \|L\|_*,$$  \hspace{1cm} (2)

Here, the least-squares loss $\sum_{(i,j) \in \Omega} (L_{ij} - M_{ij})^2$ measures the fidelity of the estimate $L$ to the observation $M$, while the nuclear norm penalty $\lambda \|L\|_*$ encourages the low-rank property of the solution. In a recent work [7], it has been shown that with properly chosen regularization parameter $\lambda$, the nuclear norm regularized least-squares estimator (2) achieves optimal statistical performance in terms of estimating the low-rank matrix $L^*$. However, this optimal choice depends on the noise size, which is often unknown in practice. This begs the question:

Can we develop an estimator for noisy matrix completion that does not rely on the unknown noise size (a.k.a., tuning-free), and at the same time achieves optimal statistical performance?

Motivated by the success of the square-root Lasso estimator [8] for sparse recovery problems, we consider in this paper the following square-root matrix completion estimator (dubbed square-root MC):

$$\min_{L \in \mathbb{R}^{n \times n}} \sqrt{\sum_{(i,j) \in \Omega} (L_{ij} - M_{ij})^2 + \lambda \|L\|_*}.$$  \hspace{1cm} (3)

A notable difference from the vanilla least-squares estimator (2) is that square-root MC (3) aims at minimizing the regularized $\ell_2$ error instead of the regularized squared $\ell_2$ error.

Our Contributions: The main result of this paper (cf. Theorem 1) shows that square-root MC (3) with a noise-size-oblivious choice $\lambda = 1/\sqrt{n}$ (e.g., $\lambda = 32/\sqrt{n}$) achieves the optimal error guarantees for recovering the low-rank matrix $L^*$ over a wide range of noise sizes. Such guarantees are on par with those established for the vanilla least-squares estimator (2) with a choice of $\lambda$ depending on the noise size [7]. Clearly, the tuning-free property and statistical optimality of square-root MC together answer our motivating question in the affirmative.

To put our contributions into context, we would like to immediately point out two relevant pieces of prior work, while deferring other related ones to Section V. First and foremost, a variant of the square-root MC estimator has been proposed and studied by Klopp [9], in which an extra element-wise max norm constraint is added to the problem (3). In the same paper, it was shown that square-root MC achieves optimal statistical performance when the size of the noise is sufficiently large compared to the entries of the low-rank matrix. However, when the noise size is relatively small, the upper bound proved therein fails to uncover the optimal performance of the square-root MC estimator. In particular,
it falls short of uncovering the exact recovery property when there is no noise, i.e., when \( E = 0 \). More recently, Zhang et al. \[10\] focuses on a closely related noisy robust PCA problem \[11\], \[12\] and studies a similar tuning-free estimator. Their results, however, even in the full observation setting (i.e., \( \Omega = \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\} \)), has a poor dependence on the problem dimension \( n \), which is far from optimality. Detailed comparisons between our results and those in the papers \[9\], \[10\] can be found in Section II.

In establishing the optimal performance of square-root MC, we make the following technical contributions. First, we introduce a new decision variable \( \theta \) to convert a non-smooth loss function to a smooth one to facilitate later analysis. We then establish a novel connection between the convex square-root MC estimator and a smooth nonconvex estimator. In the end, we manage to show that an iterative algorithm allows one to find a statistically optimal solution to the nonconvex program. While this general proof strategy has been laid out in \[7\], novel considerations need to be taken to handle the non-smooth loss function and the new decision variable \( \theta \). We defer detailed discussions to relevant places in later analysis.

Notation: For a vector \( v \), we use \( \|v\|_2 \) to denote its Euclidean norm. For a matrix \( M \), we use \( \|M\|, \|M\|_F \), and \( \|M\|_\infty \) to denote its spectral norm, Frobenius norm, and the elementwise \( \ell_\infty \) norm. In addition, \( \|M\|_{2, \infty} \) denotes the largest \( \ell_2 \) norm of the rows. We also use \( \sigma_j(M) \) to denote the \( j \)-th largest singular value of \( M \).

Additionally, the standard notation \( f(n) = O(g(n)) \) or \( f(n) \lesssim g(n) \) means that there exists a constant \( c > 0 \) such that \( |f(n)| \leq cg(n) \). \( f(n) \gtrsim g(n) \) means that there exists a constant \( c > 0 \) such that \( |f(n)| \geq cg(n) \). Also, \( f(n) \gg g(n) \) means that there exists some large enough constant \( c > 0 \) such that \( |f(n)| \geq cg(n) \). Similarly, \( f(n) \ll g(n) \) means that there exists some sufficiently small constant \( c > 0 \) such that \( |f(n)| \leq cg(n) \).

II. MAIN RESULTS

We start with introducing the model assumptions for noisy matrix completion. The first assumption is on the observation pattern.

Assumption 1: Each index \( (i, j) \) belongs to the set \( \Omega \) independently with probability \( p \).

The next assumption is concerned with the noise matrix.

Assumption 2: The noise matrix \( E = [E_{ij}] \) is composed of i.i.d. zero-mean sub-Gaussian random variables with variance \( \sigma^2 \) and sub-Gaussian norm \( O(\sigma) \), i.e., \( \|E_{ij}\|_{\psi_2} = O(\sigma) \); see Definition 5.7 in the article \[13\].

In the end, we turn to the assumptions on the groundtruth matrix \( L^* \). Let \( \sigma_{\min}, \sigma_{\max} \) be the smallest and largest singular values of \( L^* \), respectively, and let \( \kappa := \sigma_{\max}/\sigma_{\min} \) be its condition number. We require the matrix \( L^* \) to be \( \mu \)-incoherent defined in the following way.

Assumption 3: The rank-\( r \) matrix \( L^* \) with SVD \( L^* = U^* \Sigma^* V^*\top \) is \( \mu \)-incoherent in the sense that

\[
\|V^*\|_{2, \infty} \leq \sqrt{\frac{\mu}{n}} \|V^*\|_F = \sqrt{\frac{\mu}{n}}.
\]

Now we are in position to state our main results regarding the square-root MC estimator, with the proof deferred to Section III.

Theorem 1: Suppose that Assumptions 1-3 hold. In addition, assume that the sample size and the noise level satisfy

\[
n^2p \geq C_{\text{sample}} \kappa^4 \mu^2 r^2 n \log^3 n,
\]

for some sufficient large (resp. small) constant \( C_{\text{sample}} > 0 \) (resp. \( C_{\text{noise}} > 0 \)). Set \( \lambda = C_{\Lambda}/\sqrt{n} \) for the square-root MC estimator (3), where \( C_{\Lambda} \) is some large absolute constant (e.g., 32). With probability at least \( 1 - O(n^{-3}) \), any solution \( L_{\text{cvx}} \) to the square-root MC problem (3) obeys

\[
\|L_{\text{cvx}} - L^*\|_F \leq C_F \kappa \frac{\sigma}{\sigma_{\min}} \sqrt{n_p} \|L^*\|_F; \tag{4a}
\]

\[
\|L_{\text{cvx}} - L^*\|_{\infty} \leq C_{\infty} \sqrt{\kappa^3 \mu r} \frac{\sigma}{\sigma_{\min}} \sqrt{n \log n} \|L^*\|_{\infty}; \tag{4b}
\]

\[
\|L_{\text{cvx}} - L^*\|_2 \leq C_{\text{op}} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|L^*\|_2. \tag{4c}
\]

Here \( C_F, C_{\infty}, C_{\text{op}} > 0 \) are three universal constants. Several remarks on Theorem 1 are in order.

A. Minimax-Optimal \( \ell_F \) Estimation Error

When the condition number \( \kappa \) is of a constant order, the square-root MC estimator enjoys minimax-optimal \( \ell_F \) estimation error \[7\], \[14\]. In contrast, the upper bound in the paper \[9\] reads

\[
\|L_{\text{cvx}} - L^*\|_F \lesssim \max \{\sigma, \|L^*\|_{\infty}\} \sqrt{n \log n / p},
\]

which is only statistically optimal when \( \sigma \gtrsim \|L^*\|_{\infty} \). In addition, translating the bound in the paper \[10\] from robust PCA to the matrix completion setting, one obtains \( \|L_{\text{cvx}} - L^*\|_F \lesssim \sigma n^2 \), which has a much worse (and hence sub-optimal) dependence on the problem dimension \( n \).

B. Noise, Sample Complexity, and Dependency on \( \kappa, r \)

Our assumption on noise level and sample complexity is consistent with \[7\]. Furthermore, these assumptions are necessary for a non-trivial guarantee as otherwise, a naive zero estimator would achieve the optimal rate. Regarding \( \kappa \) and \( r \), while we mostly focus on the case where they are of constant size, their dependency on the error rate can be of interest. In particular, the dependency on \( r \) is the best rate known and is consistent with some other nonconvex methods \[15\], \[16\]. However, the exact sharp dependency of both \( r \) and \( \kappa \) remains an open problem. Reference \[7\] also discusses these in their marks in Section I.
C. Tuning-Free Property

More importantly, the optimal performance of square-root MC is achieved in a completely tuning-free fashion. The regularization parameter $\lambda$ can be set to be $32/\sqrt{n}$, that does not depend on the noise variance $\sigma^2$, the observation probability $p$, nor the true rank $r$ of the matrix $L^*$. This is in stark contrast to the vanilla nuclear norm regularized least-squares estimator (2) in which $\lambda$ is set to be on the order of $\sigma \sqrt{\frac{n}{p}}$ (cf. [7]).

D. Entrywise Error Guarantees

Also, our main results provide upper bounds on the entrywise estimation error (cf. bound (4b)). Compared to the $\ell_p$ estimation error (4a), it can be seen that the square-root MC estimator is uniformly good in the sense that there is no spiky entry estimate with large estimation error.

To corroborate our main results, we perform numerical experiments on noisy matrix completion with simulated data. We fix the rank $r$ to be 5 throughout the experiment. For each problem dimension $n$, we generate two $n \times r$ random orthonormal matrices as $X^*$ and $Y^*$ and take $L^* := X^* Y^* \top$ as the rank-$r$ $n \times n$ groundtruth matrix. The entrywise noise is taken to be Gaussian with variance $\sigma^2$. For all the experiments, we set $\lambda = 4/\sqrt{n}$ in square-root MC, and report the average results over 20 Monte-Carlo simulations. Figure 1 reports the relative error of the square-root MC estimator in Frobenius, spectral, and infinity norms. More specifically, Figure 1(a) fixes $n = 500$, $p = 0.5$, and varies $\sigma$; Figure 1(b) fixes $\sigma = 10^{-4}$, $p = 0.5$, and varies $n$; Figure 1(c) fixes $\sigma = 10^{-4}$, $n = 2000$, and varies $p$. Overall, the plots showcase a linear relationship between the performance and the noise size $\sigma$, the problem dimension $\sqrt{n}$, and the observation probability $p$. This is consistent with the $O(\sigma \sqrt{n}/p)$ scaling proved in Theorem 1.

III. OUTLINE OF THE PROOF

In this section, we provide the key steps for proving our main result, i.e., Theorem 1. The proof follows the general strategy of bridging convex and nonconvex solutions, first appeared in the paper [7], with several important modifications to handle the non-smooth $\ell_p$ norm (as opposed to the smooth squared $\ell_p$ norm).

A central object in our analysis is the following nonconvex optimization problem

$$
\min_{X,Y,\theta \geq 0} f(X,Y,\theta) = \min_{X,Y,\theta \geq 0} \frac{1}{2\theta} \|P_\Omega (XY^\top - M)\|_F^2 + \frac{\theta}{2} + \frac{\lambda}{2} \left( \|X\|_F^2 + \|Y\|_F^2 \right),
$$

which is closely related to the original convex square-root MC formulation (3). To see this, first, for any rank-$r$ matrix $Z$, one has

$$
\|Z\|_* = \min_{X,Y \in \mathbb{R}^{n \times r}, XY^\top = Z} \frac{1}{2} \left( \|X\|_F^2 + \|Y\|_F^2 \right).
$$

Second and more importantly, we have that for any matrix $Z = XY^\top$,

$$
\|P_\Omega (Z - M)\|_F = \inf_{\theta > 0} \frac{1}{2\theta} \|P_\Omega (XY^\top - M)\|_F^2 + \frac{\theta}{2}.
$$

It turns out that the (approximate) solution to the nonconvex optimization problem (5) serves as an extremely tight approximation to the square-root MC estimator, which facilitates the statistical analysis of the latter.

FIG. 1. (a) Relative estimation error of $L_{\text{cvx}}$ vs. noise size $\sigma$ on a log-log scale, where we fix $n = 500$, $r = 5$, $p = 0.5$; (b) Relative estimation error of $L_{\text{cvx}}$ vs. problem size $\sqrt{n}$, where we fix $r = 5$, $\sigma = 10^{-4}$, $p = 0.5$; (c) Relative estimation error of $L_{\text{cvx}}$ vs. observation probability $p$ on a log-log scale, where we fix $n = 2000$, $r = 5$, $\sigma = 10^{-4}$. For all three plots, $\lambda = 4/\sqrt{n}$ and each point represents the average of 20 independent trials.
Algorithm 1 Gradient Descent on the Nonconvex Formulation of Square Root Matrix Completion

\textbf{Input:} initialization

\begin{align*}
X_0 = X^*, Y_0 = Y^*, \theta_0 = \|P_\Omega(X^*Y^*^T - M)\|_F,
\end{align*}

step size $\eta \preccurlyeq \sigma / (\sqrt{p}c_3\sigma_{\text{max}})$, and total number of iterations $t_0 = n^{18}$.

\textbf{Gradient updates:} for $t = 0, 1, \ldots, t_0 - 1$ do

\begin{align*}
X_{t+1} &= X_t - \eta \nabla_X f(X_t, Y_t, \theta_t) \\
&= X_t - \left( \frac{1}{\sigma^2} P_\Omega(X_tY_t^T - M)Y_t + \lambda X_t \right);
\end{align*}

\begin{align*}
Y_{t+1} &= Y_t - \eta \nabla_Y f(X_t, Y_t, \theta_t) \\
&= Y_t - \left( \frac{1}{\sigma^2} P_\Omega(X_tY_t^T - M)^T X_t + \lambda Y_t \right);
\end{align*}

\begin{align*}
\theta_{t+1} &= \|P_\Omega(X_{t+1}Y_{t+1}^T - M)\|_F.
\end{align*}

\textbf{Define}

\begin{align*}
t^* &:= \arg \min_{0 \leq t \leq t_0} \|\nabla_X X \cdot Y(X_t, Y_t, \theta_t)\|_F,
\end{align*}

where

\begin{align*}
\nabla_X X \cdot Y(X_t, Y_t, \theta_t) &= \left[ \frac{1}{\sigma^2} P_\Omega(X_tY_t^T - M)Y_t + \lambda X_t \right]^T X_t + \lambda Y_t.
\end{align*}

\textbf{Output:}

\begin{align*}
L_{\text{ncvx}} := X^*, Y^*, L_{\text{cvx}} := X^*, Y_{\text{cvx}} := Y^*.
\end{align*}

In sum, our proof involves two main steps:

1) We first show—via an explicit construction—that an approximate stationary point $L_{\text{ncvx}}$ of the nonconvex problem (5) exists and is also close to the groundtruth matrix $L^*$.

2) We then establish that such an approximate stationary point $L_{\text{ncvx}}$ is extremely close to the solution $L_{\text{cvx}}$ to the convex problem (3).

Combining the two key steps via triangle inequality finishes the proof.

\textit{Step 1: Nonconvex Optimization:} The nonconvex optimization problem (5) has two groups of decision variables, i.e., $(X, Y)$ and $\theta$. Also note that given a fixed pair $(X, Y)$, the optimal choice of $\theta$ is simply given by $\theta = \|P_\Omega(XY^T - M)\|_F$. Therefore it is natural to consider an alternating minimization method to construct an approximate stationary point of the nonconvex program (5); see Algorithm 1. Given a current iterate $(X_t, Y_t, \theta_t)$, the algorithm first runs one step of gradient descent on $(X, Y)$ while fixing $\theta_t$. Then updates $\theta_{t+1} = \|P_\Omega(X_{t+1}Y_{t+1}^T - M)\|_F$ to be the optimal choice given the new iterate $(X_{t+1}, Y_{t+1})$. In the end, Algorithm 1 returns the point $L_{\text{ncvx}}$ with the smallest gradient among the iterates as an approximate stationary point.

The following lemma ensures that $L_{\text{ncvx}}$ is an approximate stationary point of the nonconvex problem and more importantly is close to the groundtruth matrix $L^*$. The proof is deferred to Section III-A.

\textit{Lemma 1:} Instate the assumptions of Theorem 1. With probability at least $1 - O(n^{-3})$, one has

\begin{align*}
&\|L_{\text{ncvx}} - L^*\|_F \leq 3\sigma C_\text{C}\frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n}{p}} \|L^*\|_F, \\
&\|L_{\text{ncvx}} - L^*\|_\infty \leq 3\sqrt{k3} r^{C_\infty} \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n \log{n}}{p}} \|L^*\|_\infty.
\end{align*}

\textit{Step 2: Bridging Convex and Nonconvex Solutions:} It remains to show that $L_{\text{ncvx}}$ is extremely close to the convex solution $L_{\text{cvx}}$, which is provided in the following lemma.

\textit{Lemma 2:} Instate the assumptions of Theorem 1. With probability exceeding $1 - O(n^{-3})$, one has

\begin{align*}
&\|L_{\text{ncvx}} - L_{\text{cvx}}\|_F \leq \frac{1}{n^{5/2}} \frac{\lambda \sigma}{\sigma_{\text{min}}} \|L^*\|_F.
\end{align*}

See Section III-B for the proof of this lemma.

We remark in passing that the polynomial factor $n^{-5}$ in Lemma 2 is arbitrarily chosen, and the exponent 5 can be replaced with any large constant. The essence is that the difference between $L_{\text{ncvx}}$ and $L_{\text{cvx}}$ is orderwise much smaller compared to the estimation error of $L_{\text{ncvx}}$ itself. Such proximity between $L_{\text{ncvx}}$ and $L_{\text{cvx}}$ is verified empirically in Figure 2.

Now we are ready to combine the previous two steps and finish the proof of Theorem 1.

\textit{Proof of Theorem 1:} Combine Lemmas 1-2 with the triangle inequality to arrive at

\begin{align*}
&\|L_{\text{cvx}} - L^*\|_F \\
&\leq \|L_{\text{ncvx}} - L_{\text{cvx}}\|_F + \|L_{\text{ncvx}} - L^*\|_F \\
&\leq \left( \frac{1}{n^{5/2}} \frac{\lambda \sigma}{\sigma_{\text{min}}} + 3\sigma C_\text{C} \left( \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n}{p}} \right) \right) \|L^*\|_F
\end{align*}
where the last relation uses the facts that \( \lambda \approx 1/\sqrt{n} \) and that \( p \gtrsim 1/\sqrt{n} \). Redefine \( 4C_F \) to be \( C_F \) to complete the proof of the bound (4a). The other two bounds on the operator norm and the \( \ell_\infty \) norm follow from similar arguments. We omit here for brevity.

A. Proof of Lemma 1

Since Algorithm 1 operates in the space of low-rank factors, we start with establishing guarantees for the stacked low-rank factor

\[
F_t := [X_t, Y_t] \in \mathbb{R}^{2n \times r},
\]

and then translate the guarantees to the matrix space \( L_t = X_tY_t^\top \). Special care is needed as the decomposition \( L = XY^\top \) is not unique in \( (X, Y) \), and hence we need to account for the rotational ambiguity in \( (X, Y) \). To this end, for each \( t \geq 0 \), we define the optimal rotation matrix to be

\[
H_t := \arg\min_{R \in O(r \times r)} \|X_tR - X^\star\|_F^2 + \|Y_tR - Y^\star\|_F^2.
\]

1) Introducing Leave-One-Out Sequences: In order to control the \( \ell_2, \infty \) error of \( F_t \) (and hence \( \ell_\infty \) error of \( L_t \)), we construct 2\( n \) leave-one-out auxiliary sequences \( \{F_t^{(l)}\}_{1 \leq l \leq 2n, t \geq 0} \). The hope is that \( \{F_t^{(l)}\}_{1 \leq l \leq 2n, t \geq 0} \) serves as a good approximation to the original sequence \( \{F_t\}_{t \geq 0} \), while at the same time is more amenable to statistical analysis.

To formally construct such leave-one-out sequences, we first define 2\( n \) auxiliary loss functions. For each \( 1 \leq l \leq n \), define

\[
f^{(l)}(X, Y, \theta) = \frac{1}{2\theta} \left( \|P_{\Omega_{l-1}}(L - M)\|_F^2 + p\|P_{\Omega_{l}}(L - M)\|_F^2 \right) + \frac{\theta}{2} + \frac{\lambda}{2} (\|X\|_F^2 + \|Y\|_F^2),
\]

where

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

\[
[P_{\Omega_{l}}, (B)]_{ij} = \begin{cases} B_{ij}, & \text{if } i = l, \\ 0, & \text{otherwise} \end{cases}
\]

Similarly, for each \( n + 1 \leq l \leq 2n \), we define

\[
f^{(l)}(X, Y, \theta) = \frac{1}{2\theta} \left( \|P_{\Omega_{(l-n)}}(L - M)\|_F^2 + p\|P_{\Omega_{l}}(L - M)\|_F^2 \right) + \frac{\theta}{2} + \frac{\lambda}{2} (\|X\|_F^2 + \|Y\|_F^2),
\]

where

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

\[
[P_{\Omega_{l}}, (B)]_{ij} = \begin{cases} B_{ij}, & \text{if } j = l - n, \\ 0, & \text{otherwise} \end{cases}
\]

With these notations in place, Algorithm 2 details the way we construct the leave-one-out sequences.

Similar constructions have been deployed in the papers [7] and [12]. However, it is worth pointing out that the sequence \( \{\theta_t\} \) is produced according to the original sequence, instead of the leave-one-out sequence. This change is tailored to the analysis of the square-root MLE estimator as it aligns better with the original loss function \( f \), while allowing us to reuse several keys results in the paper [7].

2) Properties of the Iterates: As planned, we aim to show that the leave-one-out iterates \( \{F_t^{(l)}\}_{1 \leq l \leq 2n, t \geq 0} \) stay extremely close to the original iterates \( \{F_t\}_{t \geq 0} \), and that \( \{F_t\}_{t \geq 0} \) is close to the groundtruth factor \( F^\star \). Such properties are collected in the following lemma.

Lemma 3: With probability at least \( 1 - O(n^{-3}) \), the following statements hold for all iterations \( 0 \leq t \leq t_0 \):

\[
\|F_tH_t - F^\star\|_F \leq C_F \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|X^\star\|_F, \tag{10a}
\]

\[
\|F_tH_t - F^\star\|_F \leq C_{\text{opt}} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|X^\star\|_F, \tag{10b}
\]

\[
\max_{1 \leq l \leq 2n} \|F_tH_t - F_t^{(l)} R_{t}^{(l)}\|_F \leq C_3 \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} \|F^\star\|_2, \tag{10c}
\]

\[
\max_{1 \leq l \leq 2n} \|(F_t^{(l)} H_t^{(l)} - F^\star)\|_2 \leq C_4 \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} \|F^\star\|_2, \tag{10d}
\]

\[
\|F_tH_t - F^\star\|_{2, \infty} \leq C_{\infty} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} \|F^\star\|_{2, \infty}, \tag{10e}
\]

for some positive constants \( C_F \), \( C_{\text{opt}} \), \( C_3 \), \( C_4 \), \( C_{\infty} \). Here \( H_t^{(l)} \) and \( R_{t}^{(l)} \) are defined as

\[
H_t^{(l)} := \arg\min_{R \in O(r \times r)} \|F_t^{(l)} R - F^\star\|_F,
\]

\[
R_{t}^{(l)} := \arg\min_{R \in O(r \times r)} \|F_t^{(l)} R - F_t H_t\|_F.
\]

Furthermore the output \( (X_t, Y_t) \) has small gradient:

\[
\|\nabla X, Y (X_t, Y_t, \theta_t)\|_F \leq C_{\text{grad}} \frac{1}{n^\theta} \sqrt{\frac{\sigma_{\max}}{p}}. \tag{11}
\]

See Section A for the proof of this lemma.
Now we are ready to prove Lemma 1 based on the results presented in Lemma 3.

3) Proof of Lemma 1: By the triangle inequality, one has
\[
\|X_tY_t^T - L^*\| \leq \|X_tY_t^T - X_tY_{t'} + X_tY_{t'} - L^*\| \\
+ \|X_tY_{t'} - L^*\| \\
\leq \|X_tY_{t'} - X_{t'}Y_{t'}\| + \|X_tY_{t'} - L^*\| + \|L^* - L\|. 
\]
Use relation (10b) to obtain
\[
\|X_tY_t^T - L^*\| \leq \|X_tY_t^T - X_tY_t^T\| + \|X_tY_t^T - L^*\| + \|L^* - L\|
\]
\[
\leq 3C_{\text{op}} \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n}{p}} \|X_t\| \|Y_t\|
\]
\[= 3C_{\text{op}} \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n}{p}} \|L^*\|.
\]
The first inequality uses \(\|X_t\| \leq 2\|X_t\|\), which is a direct consequence of (10b) and the last line uses \(\|L^*\| = \|X_t\|\). Similarly, we have
\[
\|X_tY_t^T - L^*\| \leq 3C_{\text{op}} \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n}{p}} \|L^*\|.
\]
and
\[
\|X_tY_t^T - L^*\| \leq 3C_{\text{op}} \frac{\sigma}{\sigma_{\text{min}}} \sqrt{\frac{n}{p}} \|L^*\|
\]
\[
\leq 3C_{\text{op}} \frac{\sigma}{\sigma_{\text{min}}} \sqrt{n \log n \frac{1}{p}} \|F^*\|_{2,\infty} \|X_t\|_{2,\infty}
\]
\[
\leq 3C_{\text{op}} \frac{\sigma}{\sigma_{\text{min}}} \sqrt{n \log n \frac{1}{p}} \|L^*\|_{\infty}.
\]

Lemma 5: In case the assumptions of Theorem 1. With probability exceeding \(1 - O(n^{-3})\), for all \(H \in T\)
\[
p^{-1/2} \|P_{\Omega}(H)\|_F \geq C_{\text{inj}} \|H\|_F
\]
where \(C_{\text{inj}} = (32\sigma)\)^{-1/2}.

Proof: This is an easy consequence of Lemma 4 in the paper [7] and the relation (10e).

Last but not least, the lemma collects several interesting properties of the nonconvex solution \(L_{\text{ncvx}}\), as well as its low-rank factors \(X_{\text{ncvx}}, Y_{\text{ncvx}}\).

Lemma 6: The approximate stationary point \(L_{\text{ncvx}}\) satisfies
\[
\sqrt{\frac{\sigma_{\text{min}}}{2}} \leq \sigma_{\text{min}}(X_{\text{ncvx}}) \leq \sigma_{\text{max}}(X_{\text{ncvx}}) \leq 2\sigma_{\text{max}};
\]
\[
\sqrt{\frac{\sigma_{\text{min}}}{2}} \leq \sigma_{\text{min}}(Y_{\text{ncvx}}) \leq \sigma_{\text{max}}(Y_{\text{ncvx}}) \leq 2\sigma_{\text{max}};
\]
\[
\frac{1}{2} np^{1/2} \sigma \leq \|P_{\Omega}(L_{\text{ncvx}} - M)\|_F \leq 2np^{1/2} \sigma;
\]
\[
\|P_{\Omega}(XY^T - L^*) - p(XY^T - L^*)\| \leq \frac{\lambda}{16} np^{1/2} \sigma.
\]
See Section E for the proof of this lemma.

For notational simplicity, we define
\[
g(X, Y) := f(X, Y, \|P_{\Omega}(XY^T - M)\|_F)
\]
\[
= \|P_{\Omega}(XY^T - M)\|_F
\]
\[
+ \frac{\lambda}{2} \left(\|X\|_F^2 + \|Y\|_F^2\right).
\]
In other words, \(g(X, Y)\) is the minimal value of \(f(X, Y, \theta)\) when \((X, Y)\) is fixed.

Now we are ready to present the key lemma of this section, which relates the difference between \(L_{\text{ncvx}}\) and \(L_{\text{cvx}}\) to the size of the gradient \(\nabla g(X_{\text{ncvx}}, Y_{\text{ncvx}})\). The proof is deferred to Section F.

Lemma 7: Suppose that \((X_{\text{ncvx}}, Y_{\text{ncvx}})\) has small gradient in the sense that
\[
\|\nabla g(X_{\text{ncvx}}, Y_{\text{ncvx}})\|_F \leq \frac{\sqrt{\sigma_{\text{min}}}}{280\kappa} \max \left\{C_{\text{inj}} \sqrt{p}, \frac{1}{2} \lambda^2 n \sigma\right\}.
\]
Then on the event that Lemmas 4-6 hold, any minimizer \(L_{\text{cvx}}\) of the convex program (3) satisfies
\[
\|L_{\text{ncvx}} - L_{\text{cvx}}\|_F \leq \frac{\lambda \kappa^2}{\sqrt{\sigma_{\text{min}}}} n \sigma \|\nabla g(X_{\text{ncvx}}, Y_{\text{ncvx}})\|_F.
\]

Remark 1: Observe that if \(\|\nabla g(X_{\text{ncvx}}, Y_{\text{ncvx}})\|_F = 0\), i.e., if \(L_{\text{ncvx}}\) is an exact stationary point of the nonconvex square-root MCD problem, \(L_{\text{ncvx}}\) is also a solution to the convex problem (3).

With the help of Lemma 7, we can prove Lemma 2 now.

Proof of Lemma 2: First, Lemma 3 tells us that the nonconvex solution \((X_{t'}, Y_{t'})\) satisfies the bound (15) on the size of the gradient. This together with Lemmas 4 to 6...
allows us to invoke Lemma 7 to obtain
\[
\|\mathbf{L}_{\text{ncvx}} - \mathbf{L}_{\text{cvx}}\|_F \\
\leq \frac{\lambda \kappa^2}{\sqrt{np}} \sigma \|\nabla g(\mathbf{X}_{\text{ncvx}}, \mathbf{Y}_{\text{ncvx}})\|_F \\
\leq \frac{1}{n^5} \lambda \sigma \|\mathbf{L}^*\|_F,
\]
where the last inequality uses the gradient upper bound (15),
\[
\|\mathbf{L}^*\|_F \geq \|\mathbf{L}^\star\| \geq \sigma_{\text{min}} = \kappa \sigma_{\text{min}},
\]
and the fact that the sample size assumption
\[
n^2p \geq C_{\text{sample}} \kappa^4 \mu^2 r^2 n \log^3 n
\]
implies \(np \gtrsim 1\) and \(\kappa \lesssim n\).

IV. SIMULATION

In this section, we further illustrate the performance of the tuning-free square root matrix completion through two sets of comparative simulation studies. First we compare the performance of square-root MC to the non-square-root estimator (2) with oracle and cross-validated parameters. This allows us to examine whether we sacrifice a significant amount of performance in achieving the tuning-free property. Second, we do the same comparison on approximately low rank matrices. This helps us understand how robust the estimator is against misspecified low-rank assumption.

A. Comparing square-root MC With Standard Approach (2)

For the non-square-root approach (2), as the sampling probability \(p\) and noise level \(\sigma\) is unknown, the regularization parameter needs to be carefully chosen. Here we compare square-root MC with (2) using oracle and \(k\)-fold cross-validated regularization parameters, namely
\[
\lambda_{\text{oracle}} := \arg \min_{\lambda} \left\|\mathbf{L}^* - \tilde{\mathbf{L}}_{\lambda, \Omega}\right\|_F,
\]
\[
\lambda_{\text{CV}} := \arg \min_{\lambda} \sum_{i=1}^{k} \left\|P_{\Omega_i} \left(\mathbf{L} - \tilde{\mathbf{L}}_{\lambda, \Omega_i}\right)\right\|_F^2,
\]
where
\[
\tilde{\mathbf{L}}_{\lambda, \Omega} := \arg \min_{\mathbf{L} \in \mathbb{R}^{n \times n}} \sum_{(i,j) \in \Omega} (L_{ij} - M_{ij})^2 + \lambda \|\mathbf{L}\|_*
\]
with \(\Omega_i\) being the \(i\)-th fold of the sampled entries and \(\Omega_{-i} := \Omega \setminus \Omega_i\). Due to computational limit, our experiment uses estimates \(\hat{\lambda}_{\text{oracle}}, \hat{\lambda}_{\text{CV}}\) obtained by taking minimum over a discrete set of parameters that is close to the true \(\lambda_{\text{oracle}}\). In practice \(\lambda_{\text{oracle}}\) is inaccessible as we do not know \(\mathbf{L}^*\).

B. Performance With Approximately Low-Rank Matrices

Another point of interest is whether square-root MC is robust to misspecification of the low-rank assumption. Here we conduct the experiment with approximately rank-\(r\) matrices \(\mathbf{L}^*\) that singular values \(\sigma_1, \cdots, \sigma_r = 1\) and \(\sigma_l \propto (n-l)^{-2}\) such that \(\sum_{l=r+1}^{n} \sigma_l = \gamma\). This parameter \(\gamma\) can be viewed as a measurement of deviation from the set rank-\(r\) matrices, as
\[
\gamma \equiv \min_{\mathbf{L} \text{ rank}(\mathbf{L}) = r} \left\|\mathbf{L}^* - \mathbf{L}\right\|_*.
\]

In each run of the experiment, we first generate an \(n \times n\) matrix \(\mathbf{M}\) as in (1) and calculate its estimator using square-root MC with fixed regularization parameter \(\lambda = 2/\sqrt{n}\) and (2) with \(\hat{\lambda}_{\text{oracle}}\) and 10-fold cross-validated \(\hat{\lambda}_{\text{CV}}\). Figure 3 shows the relative Frobenius errors of the different methods across varying matrix size \(n\) and varying noise level \(\sigma\). In both settings, we can see while square-root MC has very close estimation error to that of (2). Moreover their linear trends over problem size \(\sqrt{n}\) and noise size \(\sigma\) are similar, as we expect from their identical error rate. This shows that by using square-root MC, we achieve the tuning-free property with a minor sacrifice in the rate of estimation performance.
We then perform the same experiments as above, i.e., comparing square-root MC to (2) with oracle and cross-validated \( \lambda \). Figure 4 shows their respective estimation error vs \( \gamma \). We can see that the estimation error for all three methods increases when \( \gamma \) increases and the increments are small and comparable across the three methods. This shows that square-root MC and (2) to are somewhat robust to the violation of low rank assumption.

In addition, we showcase an interesting discovery which compares the robustness of convex and nonconvex version of square-root MC to approximate low-rankness. We generate the ground-truth matrices that is approximately low rank and calculate square-root MC and the nonconvex solution of (5) assuming the rank is \( r \). Figure 5 shows that the performance of square-root MC for approximately low rank matrices is close to the case with exact low-rankness (\( \gamma = 0 \)), while the nonconvex method suffers a much greater loss in estimation accuracy. The difference between convex and nonconvex method is close to 0 when \( \gamma = 0 \) and increases drastically as \( \gamma \) increases. To some extent, this is expected as the convex method does not require the input of rank information.

V. PRIOR ART
A. Matrix Completion

Convex relaxation has been extensively studied for the matrix completion problem both in the noiseless setting [1], [17], [18], [19], [20], and the noisy case [6], [7], [9], [14], [21]. In the noiseless setting, convex relaxation achieves exact recovery as soon as the number of observed entries \( n^2p \) exceeds \( nr \log n \log r \) [22]—roughly the degrees of freedom of a rank-\( r \) matrix, which is information-theoretically optimal. When it comes to the noisy setting, Candès and Plan [6] focuses on arbitrary noise (e.g., noise could be deterministic and adversarial), and proves that convex relaxation is stable w.r.t. the noise size. The theoretical guarantees for convex relaxation are strengthened by Chen et al. [7] in the stochastic noise case, which is the same setting we study in the current paper. Such a discrepancy between stochastic and deterministic noise for convex relaxation is also documented in [23].

Pioneered by the work [2], [24], nonconvex optimization has gained a lot of attentions during the past decade for solving matrix completion owing to its computational efficiency. Efficient computational and statistical guarantees have been provided for manifold optimization [2], [24], gradient descent [16], [25], projected gradient descent [15], [26], alternating minimization [27], [28], scaled gradient descent [29], singular value projection [22], etc. See the recent surveys [30], [31] for more related work on matrix completion.

B. Tuning-Free Methods

A variety of tuning-free methods have been proposed to tackle high-dimensional linear regression. The seminal work [8] proposes the square-root Lasso estimator which does not rely on knowing the size of the noise and is also statistically optimal. Reference [32] proposes an equivalent method named scaled sparse linear regression, which originates from the concomitant scale estimation [33], [34]. Reference [35] proposes TREX, a method similar to square-root Lasso and is completely parameter-free. Reference [36] borrows ideas from non-parametric statistics and proposes Rank Lasso, whose optimal choice of tuning parameter can be simulated easily in the case with unknown variance of the noise. See [37] for a survey on the selection of tuning-parameters for high-dimensional regression and [38] for a survey on regression with unknown variance of noise.

C. Bridging Convex and Nonconvex Optimization

The connections between convex and nonconvex optimization has been extensively used in a recent line of work. Chen et al. [7] uses this to prove the optimality of the vanilla least-squares estimator for noisy matrix completion. Later, the papers [12], [39], [40] extend the technique to the robust PCA problem, the blind deconvolution problem, and matrix completion with heavy-tailed noise.
D. Leave-One-Out Analysis

Leave-one-out analysis is powerful in decoupling statistical dependence and obtain element-wise performance guarantees. It has been successfully applied to high-dimensional regression [41], [42], phase synchronization [43], ranking [44], matrix completion [16], [25], [46], reinforcement learning [47], high-dimensional inference [48], [49] to name a few. Interested readers are referred to a recent overview [50] for detailed discussions.

VI. DISCUSSIONS

Focusing on the noisy matrix completion problem, this paper shows that a tuning-free estimator—square-root MC achieves optimal statistical performance. This opens up several interesting avenues for future research. Below, we list a few of them.

- Extensions to robust PCA. While our work focuses on matrix completion, a natural extension is to further consider partial observations with outliers, i.e., robust PCA. As mentioned, Zhang et al. [10] has studied this problem (with full observation) and provides an error guarantee of order $O(n\sigma^2)$, which is sub-optimal in its dependency on the problem dimension. By contrast, a vanilla least-squares estimator with noise-size-dependent choice of $\lambda$ has been shown to be optimal [12]. It remains to be seen whether one can devise an optimal tuning-free method for robust PCA with noise and missing data.

- Inference for square-root MC estimator. The current paper discusses solely the estimation performance of the tuning-free estimator. As statistical inference for matrix completion is equally important, one wishes to develop inferential procedures around the square-root MC estimator as that has been done in the paper [48] for the vanilla least-squares estimator.

- Robustness to non-uniform design. In high-dimensional linear regression, optimal tuning-free methods have been developed to be adaptive to both the unknown noise size and the design matrix. In the matrix completion setting, the design is governed by the sampling pattern, which is assumed to be uniform in the current paper. It is of great interest to develop robust and tuning-free approaches for noisy matrix completion with non-uniform sampling that improve over the max-norm constrained estimator in [9].

APPENDIX

A. Proof of Lemma 3

We prove Lemma 3 via induction. Since all the algorithms start from the groundtruth, it is trivial to see that the hypotheses (10) hold for $t = 0$. We also record two important properties of the iterates at $t = 0$, namely,

$$\frac{1}{2}np^{1/2}\sigma \leq \theta_t \leq 2np^{1/2}\sigma$$  \hspace{1cm} (16)

and

$$\|X_t^T X_t - Y_t^T Y_t\|_F \leq C_3 \kappa \eta \sigma \sqrt{\frac{n}{\min}} \sqrt{\sigma_{\max}^2},$$ \hspace{1cm} (17)

where $C_3 > 0$ is a universal constant. Note that at $t = 0$, we have $\theta_0 = \|P_1(E)\|_F$, which concentrates sharply around $np^{1/2}\sigma$ under the noise assumption and uniform sampling.

Now suppose the hypotheses (10), (16), and (17) hold for the $t$-th iterates. We aim to show that the same set of hypotheses continue to hold for the $(t+1)$-th iterates. Sections B and C are devoted to this induction step. In addition, we prove the last claim (11) in Section D. In Section E we prove Lemma 6 which is a consequence of (10) and (16).

B. Induction on Hypotheses (10) and (17)

Define

$$\tilde{\lambda}_t := \lambda \theta_t \quad \text{and} \quad \tilde{\eta}_t := \eta / \theta_t.$$ \hspace{1cm}

We make a key observation that the $t$-th iterations of Algorithm 1 and 2 are exactly the same as the $t$-th iterations of Algorithm 1 (vanilla gradient descent) and 2 (construction of the leave-one-out sequence) in the paper [7] with the parameters $\tilde{\lambda}_t$ and $\tilde{\eta}_t$. Moreover, given the induction hypothesis (16) one has $\tilde{\eta}_t \sigma \leq \theta_{t-1} \leq 2n \sigma_\max$. Combine this with our choice of $\lambda = C_3 n^{-1/2}$ to see that

$$\tilde{\lambda}_t \leq \sqrt{\tilde{\eta}_t} \quad \text{and} \quad \tilde{\eta}_t \leq 1/(np\sigma^3\sigma_{\max}).$$

which are consistent with the choice of $\lambda$ and $\eta$ in [7]. These allow us to invoke Lemmas 10-15 in [7] to prove that claims (10) and (17) hold for the $(t+1)$-th iterates.

C. Induction on Hypotheses (16)

In this section, we aim to show that the claim (16) holds for the $(t+1)$-th iterates.

Observe that

$$\mathcal{P}_\Omega \left( X_{t+1} Y_{t+1}^T - M \right) = \mathcal{P}_\Omega \left( X_{t+1} Y_{t+1}^T - L^* \right) - \mathcal{P}_\Omega (E).$$

Similar to the proof of Lemma 1, using the incoherence assumption

$$\|F^*\|_{2,\infty} = \max \{ \|X^*\|_{2,\infty}, \|Y^*\|_{2,\infty} \} \leq \sqrt{\mu r \sigma_{\max}/n},$$

and

$$\|X_{t+1} Y_{t+1}^T - L^*\|_\infty \leq 3\sigma_{\max}/\sqrt{\mu r \sigma_{\max}/n},$$

Then

$$\mathcal{P}_\Omega \left( X_{t+1} Y_{t+1}^T - L^* \right) \lesssim \sqrt{\mu r \sigma_{\max}/n}.$$ \hspace{1cm}

As the sample size satisfies $n^2 p \gg \kappa^4 \mu^2 r^2 \log^3 n$, we have $\mathcal{P}_\Omega \left( X_{t+1} Y_{t+1}^T - L^* \right) \lesssim \sqrt{\mu r \sigma_{\max}/n}$. As mentioned before,
$\|P_\Omega(E)\|_F$ sharply concentrates around $np^{1/2}\sigma$. Therefore by the triangle inequality, we have

$$\frac{1}{2} \sigma n \sqrt{p} \leq \|P_\Omega (X_t Y_t^\top - M)\|_F \leq 2 \sigma n \sqrt{p}$$

for large enough $n$.

**D. Proof of Bound (11)**

Suppose for the moment that

$$f(X_t, Y_t, \theta_t) \leq f(X_{t-1}, Y_{t-1}, \theta_{t-1})$$

for all $t \geq 1$. Then a telescoping argument would yield the conclusion that

$$f(X_0, Y_0, \theta_0) - f(X_t, Y_t, \theta_t) \geq \frac{\eta}{2} \sum_{t=0}^{t_{t-1}} \|\nabla_{X,Y} f(X_t, Y_t, \theta_t)\|^2_F$$

holds for all $t \geq 1$. Expanding the left hand side, we see that it is upper bounded by

$$f(X_0, Y_0, \theta_0) - f(X_t, Y_t, \theta_t) \leq \|P_\Omega(E)\|_F - \|P_\Omega(X_t Y_t^\top - M)\|_F$$

$$+ \frac{\lambda}{2} \left( \|X^\top\|^2_F - \|X_{t-1}^\top\|^2_F + \|Y^\top\|^2_F - \|Y_{t-1}^\top\|^2_F \right)$$

$$\leq \|P_\Omega(E)\|_F + \frac{\lambda}{2} \left( \|X^\top\|^2_F - \|X_{t-1} H_{t-1}\|^2_F \right)$$

$$+ \|Y^\top\|^2_F - \|Y_{t-1} H_{t-1}\|^2_F,$$

where the last line uses the nonnegativity of norms and the invariance of Frobenius norm under rotation. In view of the properties (10) and the noise size assumption $\sigma_{\min} \sqrt{\frac{n}{p}} < 1$, we have

$$\|X^\top - X_{t-1} H_{t-1}\|_F \leq \|X^\top\|_F$$

and

$$\|X_{t-1}\|_F = \|X_{t-1} H_{t-1}\|_F \leq 2 \|X^\top\|_F.$$

Then,

$$\|X^\top\|^2_F - \|X_{t-1} H_{t-1}\|^2_F$$

$$\leq \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \|X^\top\|_F \|X^\top\|_F$$

$$\leq \sigma_{\text{rk}} \sqrt{\frac{n}{p}},$$

where the last line uses the fact that $\|X^\top\|_F \leq \sqrt{\frac{\sigma_{\text{max}}}{p}}$. Similarly, we have

$$\|Y^\top\|^2_F - \|Y_{t-1} H_{t-1}\|^2_F \leq \sigma_{\text{rk}} \sqrt{\frac{n}{p}}.$$

To simplify the expression we use $\kappa \lesssim n$ and $r \lesssim \sqrt{n}$ which are consequences of the sample size assumption $n^2 \geq n^2 p \gg \kappa^4 \mu^2 r^2 n \log n$.

**Proof of bound (18):** Define

$$h(X, Y) := \theta_t [f(X, Y, \theta_t) - \theta_t/2].$$

Then $h(X, Y)$ matches the form of the objective function in Lemma 16 of the paper [7]. Then Lemma 16 therein tells us that

$$h(X_{t+1}, Y_{t+1}) \leq h(X_t, Y_t) - \frac{\eta_\eta}{2} \|\nabla h(X_t, Y_t)\|^2_F,$$

where we recall $\eta_t = \eta/\theta_t$. Rewriting the bound in terms of $f$ yields

$$f(X_{t+1}, Y_{t+1}, \theta_t) \leq f(X_t, Y_t, \theta_t)$$

$$- \frac{\eta}{2} \|\nabla_{X,Y} f(X_t, Y_t, \theta_t)\|^2_F.$$

In addition, by the optimality of $\theta_{t+1}$, one has

$$f(X_{t+1}, Y_{t+1}, \theta_{t+1}) \leq f(X_{t+1}, Y_{t+1}, \theta_t).$$

Combining (20) and (21) completes the proof.

**E. Proof of Lemma 6**

By Lemma 3, we know that $X_{\text{ncvx}}$ satisfies

$$\|X_{\text{ncvx}} - X^\top\| \leq C_{\text{op}} \left( \frac{\sigma}{\sigma_{\text{min}} \sqrt{\frac{n}{p}}} \right) \|X^\top\| \ll \sqrt{\sigma_{\text{min}}},$$

where the last relation arises from the noise level assumption $\sigma_{\min} \sqrt{\frac{n}{p}} < 1/\sqrt{\kappa^4 \mu r n \log n}$. Therefore we can apply Weyl's inequality to obtain

$$\sigma_{\max}(X_{\text{ncvx}}) \leq \sqrt{\sigma_{\max}} + \|X_{\text{ncvx}} - X^\top\| \leq 2 \sigma_{\max};$$

$$\sigma_{\min}(X_{\text{ncvx}}) \geq \sqrt{\sigma_{\min}} - \|X_{\text{ncvx}} - X^\top\| \geq \sqrt{\sigma_{\min}/2},$$

for large enough $n$. These hold similarly for the singular values of $Y_{\text{ncvx}}$.

On the other hand, the relations (14a) come directly from (16), and (14b) follows from Lemma 4 in [7].
F. Proof of Lemma 7

To simplify the notation, we denote
\[ \theta := \|P_{\Omega}(L_{\text{ncvx}} - M)\|_F \]
and \( \Delta := L_{\text{ncvx}} - L_{\text{ncvx}} \) throughout this section. In view of Lemma 6, we know that \( \theta \neq 0 \), and hence \( \theta^{-1} \) is well defined.

Recall that \( U\Sigma V^T \) is the SVD for \( L_{\text{ncvx}} \), and \( T \) is the tangent space at \( L_{\text{ncvx}} \). The following lemma is useful in controlling the size of \( \Delta \).

**Lemma 8:** Under the notations and assumptions of Lemma 7, we have
\[ \frac{1}{\theta} P_{\Omega}(L_{\text{ncvx}} - M) = -\lambda(UV^T + R), \] (22)
where \( R \) is a residual matrix such that
\[ \|P_T(R)\|_F \leq 70\kappa \sigma_{\min}^{1/2} \|\nabla g(X, Y)\|_F, \]
and
\[ \|P_{T^\perp}(R)\| < 1/2. \]

See Section G for the proof.

We decompose the proof into three steps. In Step 1, we show that the difference matrix \( \Delta \) mainly lies in the tangent space \( T \). In Step 2, the previous fact is leveraged to show an upper bound on \( P_{\Omega}(\Delta) \). In the last step (Step 3), we connect the previous steps with the injectivity property (cf. Lemma 5) to reach the desired conclusion.

**Step 1:** Showing That \( \Delta \) Lies Primarily in the Tangent Space \( T \): By the optimality of \( L_{\text{ncvx}} \), we have
\[
0 \geq \left\|P_{\Omega}(L_{\text{ncvx}} - M)\right\|_F - \|P_{\Omega}(L_{\text{ncvx}} - M)\|_F^\alpha + \lambda \left(\|L_{\text{ncvx}}\|_* - \|L_{\text{ncvx}}\|_*\right). \] (23)

Use the convexity of \( \|\cdot\|_F \) and \( \|\cdot\|_* \), and the decomposition \( L_{\text{ncvx}} = U\Sigma V^T \) to see that
\[
0 \geq \left\{\frac{1}{\theta} P(L_{\text{ncvx}} - M), \Delta\right\} + \lambda \left(UV^T + W_0, \Delta\right)
\]
holds for any \( W_0 \in T^\perp \) with \( \|W_0\| < 1 \). Apply Lemma 8 to further obtain
\[
0 \geq -\lambda \left(R, \Delta\right) + \lambda \left(W_0, \Delta\right) .
\]

In particular, one can choose \( W_0 \in T^\perp \) such that \( \|P_{T^\perp}(\Delta)\|_* = (W_0, \Delta) \), which yields the inequality
\[
0 \geq \lambda \|P_{T^\perp}(\Delta)\|_* - \lambda \left(R, \Delta\right) = \lambda \|P_{T^\perp}(\Delta)\|_* - \lambda \left(P_T(R), \Delta\right) - \lambda \left(P_{T^\perp}(R), \Delta\right) \geq \lambda \|P_{T^\perp}(\Delta)\|_* - \lambda \left(P_T(R)\right) \|_F \|P_T(\Delta)\|_F - \lambda \left(P_{T^\perp}(R)\right) \|P_{T^\perp}(\Delta)\|_* .
\]

Here the last line arises from Holder’s inequality.

Again, by Lemma 8, we have the bounds
\[
\|P_T(R)\|_F \leq 70\kappa \sigma_{\min}^{1/2} \|\nabla g(X, Y)\|_F \]
and \( \|P_{T^\perp}(R)\| < 1/2 \), which allow us to further arrive at
\[
0 \geq \frac{1}{2} \|P_{T^\perp}(\Delta)\|_* - 70\lambda \sigma_{\min}^{1/2} \|\nabla g(X, Y)\|_F \|P_T(\Delta)\|_F .
\]

This further implies
\[
\|P_{T^\perp}(\Delta)\|_F \leq \|P_{T^\perp}(\Delta)\|_* \leq 140\kappa \sigma_{\min}^{-1/2} \|\nabla g(X, Y)\|_F \|P_T(\Delta)\|_F . \] (24)

As an immediate consequence, under the assumed upper bound (15) for \( \|\nabla g(X, Y)\|_F \), we have
\[
\|\Delta\|_F \leq 140\kappa \sigma_{\min}^{-1/2} \|\nabla g(X, Y)\|_F \|P_T(\Delta)\|_F \]
\[
\leq 2 \|P_T(\Delta)\|_F . \] (25)

**Step 2:** Bounding \( \|P_{\Omega}(\Delta)\|_F^2 \): We start with presenting an identity involving \( \|P_{\Omega}(\Delta)\|_F^2 \):
\[
\|P_{\Omega}(\Delta)\|_F^2 = \alpha \cdot \left(\|P_{\Omega}(L_{\text{ncvx}} - M)\|_F + \|P_{\Omega}(L_{\text{ncvx}} - M)\|_F \right)
\]
\[
+ 2\lambda \left(P_{\Omega}(L_{\text{ncvx}} - M)\right) \|P_T(\Delta)\|_F \]
\[
+ \left(\alpha - \frac{1}{\theta} P_{\Omega}(L_{\text{ncvx}} - M), \Delta\right) . \] (26)

Lemma 8 and (23) tell us that
\[
\alpha - \left\{\frac{1}{\theta} P_{\Omega}(L_{\text{ncvx}} - M), \Delta\right\} \leq \lambda \|L_{\text{ncvx}}\|_* - \|L_{\text{ncvx}}\|_* + \lambda \left(UV^T + R, \Delta\right) .
\]

By convexity of \( \|\cdot\|_* \), this further simplifies to
\[
\lambda \|L_{\text{ncvx}}\|_* - \lambda \|L_{\text{ncvx}}\|_* + \lambda \left(UV^T + R, \Delta\right) \leq -\lambda \left(UV^T + W, \Delta\right) + \lambda \left(UV^T + R, \Delta\right) \]
\[
= \lambda \Delta R - W , \] (27)
for any \( W \in T^\perp \) with \( \|W\| < 1 \). Combine (26) and (27) to reach
\[
\|P_{\Omega}(\Delta)\|_F^2 \leq \alpha^2 + 2\lambda \|P_{\Omega}(L_{\text{ncvx}} - M)\|_F \|\Delta R - W\| . \]

We prove in the end of this section that the two terms \( \alpha^2 \) and \( \beta \) obey
\[
\alpha^2 \leq \lambda^2 \left(\sqrt{\beta} + 140\kappa \sigma_{\min}^{-1/2} \|\nabla g(X, Y)\|_F^2\right) \cdot \|P_T(\Delta)\|_F^2 , \] (28a)
\[
\beta \leq 560\lambda \kappa \sigma_{\min}^{-1/2} \lambda \|\nabla g(X, Y)\|_F \|P_T(\Delta)\|_F , \] (28b)
which yields the upper bound on \( \|P_{\Omega}(\Delta)\|_F^2 \) in terms of \( \|P_T(\Delta)\|_F^2 \):
\[
\|P_{\Omega}(\Delta)\|_F^2 \leq \lambda^2 \left(\sqrt{\beta} + 560\lambda \sigma_{\min}^{-1/2} \|\nabla g(X, Y)\|_F^2 \|P_T(\Delta)\|_F^2 \right)
\]
\[
+ 560\lambda \sigma_{\min}^{-1/2} \|\nabla g(X, Y)\|_F \|P_T(\Delta)\|_F \|P_T(\Delta)\|_F . \]

**Step 3:** Final Calculations: Using the decomposition \( P_{\Omega}(\Delta) = P_{\Omega} P_T(\Delta) + P_{\Omega} P_{T^\perp}(\Delta) \), we obtain
\[
\|P_{\Omega}(\Delta)\|_F = \|P_{\Omega} P_T(\Delta) + P_{\Omega} P_{T^\perp}(\Delta)\|_F \]
\[
\geq \|P_{\Omega} P_T(\Delta)\|_F - \|P_{\Omega} P_{T^\perp}(\Delta)\|_F . \]
Together with Lemma 5 and (24), we have
\[\|P_\lambda(\Delta)\|_F \geq (\sqrt{p}C_{\text{inj}} - 140\kappa\sigma_{\text{min}}^{-1/2}\|\nabla g(X, Y)\|_F)\|P_\lambda(\Delta)\|_F \geq \sqrt{p}C_{\text{inj}}\|P_\lambda(\Delta)\|_F,\]
where the last line uses (15). As a result, we arrive at the sandwich formula
\[\frac{1}{4}pC_{\text{inj}}^2\|P_\lambda(\Delta)\|_F^2 \leq \|P_\lambda(\Delta)\|_F^2 \leq \lambda^2(\sqrt{\tau} + 140\kappa\sigma_{\text{min}}^{-1/2}\|\nabla g(X, Y)\|_F)^2\|P_\lambda(\Delta)\|_F^2 + 560\lambda\kappa\sigma_{\text{min}}^{-1/2}\|\nabla g(X, Y)\|_F\|P_\lambda(\Delta)\|_F,\]
which further implies
\[\left\{\frac{pC_{\text{inj}}^2}{4} - \lambda^2(\sqrt{\tau} + 140\kappa\sigma_{\text{min}}^{-1/2}\|\nabla g(X, Y)\|_F)^2\right\}\|P_\lambda(\Delta)\|_F^2 \leq 560\lambda\kappa\sigma_{\text{min}}^{-1/2}\|\nabla g(X, Y)\|_F\|P_\lambda(\Delta)\|_F.\]
Reorganize and substitute in (15) to see that for large enough n,
\[\frac{pC_{\text{inj}}^2}{4} - \lambda^2(\sqrt{\tau} + 140\kappa\sigma_{\text{min}}^{-1/2}\|\nabla g(X, Y)\|_F)^2 \geq \frac{pC_{\text{inj}}^2}{8}.\]
Combine the above two relations to reach
\[\frac{pC_{\text{inj}}^2}{8}\|P_\lambda(\Delta)\|_F^2 \leq 560\lambda\kappa\sigma_{\text{min}}^{-1/2}\|\nabla g(X, Y)\|_F\|P_\lambda(\Delta)\|_F,\]
which together with \(C_{\text{inj}} = (32\kappa)^{-1/2}\) and (14a) implies
\[\|P_\lambda(\Delta)\|_F \leq \frac{\lambda\kappa^2}{\sqrt{\rho}^\theta_{\text{min}}}n\sigma\|\nabla g(X, Y)\|_F.\]
Use (25), we obtain the bound on \(\|\Delta\|_F\),
\[\|\Delta\|_F \leq 2\|P_\lambda(\Delta)\|_F \leq \frac{\lambda\kappa^2}{\sqrt{\rho}^\theta_{\text{min}}}n\sigma\|\nabla g(X, Y)\|_F.\]

**Proof of the Bound (28a):** For \(\alpha^2\) we consider the cases when \(\alpha\) is positive and non-positive separately.

**Case of \(\alpha \leq 0\):** By convexity of \(\|\cdot\|_F\),
\[0 \geq \alpha > \left\{\frac{1}{\theta}P_{\text{convex}}(X Y^T - M), \Delta \right\}.\]
Using the representation in Lemma 8, the last term can be written as \(\langle UV^T + R, \Delta \rangle\). Splitting the parts into T and \(T^\perp\), we have
\[\alpha^2 \leq \lambda^2\left\langle UV^T + R, \Delta \right\rangle^2 \leq \lambda^2\left(\|U V^T\|_F\|P_\lambda(\Delta)\|_F + \|P_\lambda(R)\|_F\|P_\lambda(\Delta)\|_F + \|P_{\lambda^+}(R)\|\|P_{\lambda^+}(\Delta)\|_{\text{opt}}\right)^2.\]
Together with (24) and Lemma 8, we arrive at
\[\alpha^2 \leq \lambda^2(\sqrt{\tau} + 140\kappa\sigma_{\text{min}}^{-1/2}\|\nabla g(X, Y)\|_F)^2\|P_\lambda(\Delta)\|_F^2.\]

**Case of \(\alpha > 0\):** By optimality of \(L_{\text{convex}}\) and convexity of \(\|\cdot\|_F\),
\[0 < \alpha \leq -\lambda\langle L_{\text{convex}}\|s\| - L_{\text{ncvx}}\|s\rangle \leq -\lambda\langle UV^T, \Delta \rangle.\]
Then similar to the case of \(\alpha \leq 0\),
\[\alpha^2 \leq \lambda^2\|P_\lambda(\Delta)\|_F^2.\]
Combining the two cases yields (28a).

**Proof of the Bound (28b):** For \(\beta\), we can split the parts into T and \(T^\perp\) similar to the proof for (28a). Using (24) and Lemma 8, we have
\[2\theta \cdot \lambda(\Delta, R - W) \leq 2\lambda(\|\Delta\|_F + \|\Delta W\|) \leq 2\lambda\theta\left(\|P_{\lambda^+}(R)\|_F\|P_{\lambda^+}(\Delta)\|_F + \|P_{\lambda^+}(W)\|\|P_{\lambda^+}(\Delta)\|_{\text{opt}}\right) \leq 560\lambda\kappa\sigma_{\text{min}}^{-1/2}\|\nabla g(X, Y)\|_F\|P_{\lambda^+}(\Delta)\|_F.\]

**G. Proof of Lemma 8**

The proof relies on the following representation of the low-rank factors \(X, Y\) of the nonconvex solution \(L_{\text{ncvx}}\).

**Lemma 9:** Under the assumptions and notations of Lemma 7, there exists an invertible matrix \(Q \in \mathbb{R}^{r \times r}\) such that \(X = U \Sigma^{1/2} Q Y = V \Sigma^{1/2} Q^{-1}\), \(\|Q\| \leq 2\) and
\[\|\Sigma^{1/2} Q Q^T \| \Sigma^{-1/2} - I_r \| \leq 32\kappa^{-\theta}/\sigma_{\text{min}}\|\nabla g(X, Y)\|_F \leq 1/3.\]
where \(U Q \Sigma V Q^T\) is the SVD of \(Q\).
See Section H for the proof.

Denote the partial gradients of \(g(X, Y)\) as \(B_1, B_2\), i.e.,
\[B_1 := \nabla_X g(X, Y) = \frac{1}{\theta}Q_1(X Y^T - M)Y + \lambda X; \quad (31a)\]
\[B_2 := \nabla_Y g(X, Y) = \frac{1}{\theta}Q_1(X Y^T - M)^T X + \lambda Y, \quad (31b)\]
where we recall \(\theta = \|P_{\lambda^+}(X Y^T - M)\|_F\). By definition, we know that \(\max\{\|B_1\|_F, \|B_2\|_F\} \leq \|\nabla g(X, Y)\|_F\).
Let \(R\) be the matrix that is defined by (22). We now control its component in T and \(T^\perp\) separately.

**Part 1: Bounding \(\|P_{\lambda^+}(R)\|_F\):** By the definition of the projection operator \(P_{\lambda^+}\), we have
\[\|P_{\lambda^+}(R)\|_F = \|U U^T R(I - V V^T) + R V V^T\|_F \leq \|U U^T R(I - V V^T)\|_F + \|R V V^T\|_F \leq \|U^T R\|_F + \|R V\|_F.\]
For the term \(R V\), we use the definitions of \(B_1\) and \(R\) to see that
\[\lambda U V^T Y + \lambda R Y = \lambda X - B_1,\]
which together with the representations in Lemma 9 implies
\[R V = U \Sigma^{1/2} (Q Q^T - I_r) \Sigma^{-1/2} - B_1 Q^T \Sigma^{-1/2}.\]
In view of the relation (30), we have
\[ \|RV\|_F \leq \|\Sigma^{1/2}QQ^T - I_r\|_F\Sigma^{-1/2}\|_F \\
+ \|\Sigma^{-1/2}\|_F Q\|_F B_1 \|_F \leq \frac{32\kappa}{\sqrt{\sigma_{\min}}} \|\nabla g(X, Y)\|_F + 2\sqrt{\frac{2}{\sigma_{\min}}} \|\nabla g(X, Y)\|_F \leq \frac{35\kappa}{\sqrt{\sigma_{\min}}} \|\nabla g(X, Y)\|_F, \]
where we have used the fact that \(\|\Sigma^{-1}\| \leq \sigma_{\min}/2\). Similarly we can establish that
\[ \|U^T R\|_F \leq \frac{35\kappa}{\sqrt{\sigma_{\min}}} \|\nabla g(X, Y)\|_F. \]
Combine the two inequalities to arrive at
\[ \|P_T(R)\|_F \leq \frac{70\kappa}{\sqrt{\sigma_{\min}}} \|\nabla g(X, Y)\|_F. \]

**Part 2: Bounding \(\|P_{T^\perp}(R)\|\):** For any matrix \(A\), define \(P_{\Omega\Omega}(A) := \Omega\Omega(A) - pA\). We can rewrite the identities (31a) and (31b) as
\[ \frac{1}{\theta} \left[pL^* + P_{\Omega}(E) - P_{\Omega}(\text{debias}(XY^T - L^*))\right] Y = \frac{p}{\theta} XY^T Y + \lambda X - B_1; \]
\[ \frac{1}{\theta} \left[pL^* + P_{\Omega}(E) - P_{\Omega}(\text{debias}(XY^T - L^*))\right]^T X = \frac{p}{\theta} XY^T X + \lambda Y - B_2. \]
Again, using the representations in Lemma 9, we have the following two identities
\[ \frac{1}{\theta} \left[pL^* + P_{\Omega}(E) - P_{\Omega}(\text{debias}(XY^T - L^*))\right]^T V \]
\[ = \frac{1}{\theta} pU\Sigma + \lambda U\Sigma^{1/2}QQ^T\Sigma^{-1/2} - B_1 Q^T\Sigma^{-1/2}; \]
\[ \frac{1}{\theta} \left[pL^* + P_{\Omega}(E) - P_{\Omega}(\text{debias}(XY^T - L^*))\right] U = \frac{1}{\theta} p\Sigma + \lambda \Sigma^{1/2}QQ^T\Sigma^{-1/2} - B_2 Q^T\Sigma^{-1/2}. \]
These two equations motivate us to define a matrix \(\hat{R}\) using
\[ \frac{1}{\theta} \left[pL^* + P_{\Omega}(E) - P_{\Omega}(\text{debias}(XY^T - L^*))\right] V \]
\[ = \frac{1}{\theta} pU\Sigma V^T + \lambda U\Sigma^{1/2}QQ^T\Sigma^{-1/2}V^T + \lambda\hat{R}, \]
where \(\hat{R}\) obeys \(P_{T^\perp}(R) = P_{T^\perp}(\hat{R})\). To see this, we use the definition of \(\hat{R}\) to write
\[ P_{T^\perp}(R) = -\frac{1}{\lambda} P_{T^\perp} \left(\theta^{-1}P_{\Omega}(XY^T - M)\right) \]
\[ = -\frac{1}{\lambda} P_{T^\perp} \left[P_{\Omega}(XY^T - L^*) - P_{\Omega}(E)\right]. \]
Since \(P_{T^\perp}(XY^T) = 0\), by definition of \(\hat{R}\), we obtain
\[ P_{T^\perp}(R) = \frac{1}{\lambda} P_{T^\perp} \left[p(L^* - XY^T) + P_{\Omega}(E) \right. \]
\[ - \left. P_{\Omega}(XY^T - L^*)\right] = P_{T^\perp}(\hat{R}). \]
Therefore from now on, we concentrate on bounding \(\|P_{T^\perp}(\hat{R})\|\).

To this end, we rewrite (33) as
\[ \frac{1}{\theta} \left[pL^* + P_{\Omega}(E) - P_{\Omega}(\text{debias}(XY^T - L^*))\right] - \lambda P_{T^\perp}(\hat{R}) \]
\[ = \frac{1}{\theta} pU\Sigma V^T + \lambda U\Sigma^{1/2}QQ^T\Sigma^{-1/2}V^T \]
\[ + \lambda P_{T^\perp}(\hat{R}). \]
Suppose that
\[ \|P_{T^\perp}(\hat{R})\| \leq \frac{\lambda}{4}\theta, \]
which together with Lemma 4 and Lemma 6 implies that
\[ \frac{1}{\theta} \left\|P_{\Omega}(E) - P_{\Omega}(\text{debias}(XY^T - L^*))\right\| \leq \frac{\lambda}{8} + \frac{\lambda}{8} + \frac{\lambda}{4} = \frac{\lambda}{2}. \]
By Weyl’s inequality and the fact that \(L^*\) is of rank \(r\), for each \(i = r + 1, \ldots, n\), one has
\[ \sigma_i \left(\frac{1}{\theta} pU\Sigma V^T + \lambda U\Sigma^{1/2}QQ^T\Sigma^{-1/2}V^T \right. \]
\[ \left. + \lambda P_{T^\perp}(\hat{R})\right) \]
\[ \leq \frac{1}{\theta} \left\|P_{\Omega}(E) - P_{\Omega}(\text{debias}(XY^T - L^*))\right\| \]
\[ \leq \frac{\lambda}{2}. \] (35)
At the same time, for each \(i = 1, \ldots, r\), we have
\[ \sigma_i \left(\frac{1}{\theta} pU\Sigma V^T + \lambda U\Sigma^{1/2}QQ^T\Sigma^{-1/2}V^T \right) \]
\[ \geq \sigma_r \left[U \left(\frac{1}{\theta} p\Sigma + \lambda I_r\right) \right. \]
\[ \left. + \lambda (\Sigma^{1/2}QQ^T\Sigma^{-1/2} - I_r)\right) V^T \]
\[ \geq \sigma_r \left(\frac{1}{\theta} p\Sigma + \lambda I_r\right) - \lambda \left\|\Sigma^{1/2}QQ^T\Sigma^{-1/2} - I_r\right\| \]
\[ \geq \lambda - \lambda/3 > \lambda/2, \]
where the last line uses the claim (30). As a result, the singular values of \(\lambda P_{T^\perp}(\hat{R})\) must fall below \(\lambda/2\), i.e.,
\[ \|P_{T^\perp}(\hat{R})\| = \|P_{T^\perp}(\hat{R})\| < 1/2. \]
We are left with controlling \(\|P_{T^\perp}(\hat{R})\|\). Similar to bounding \(\|P_{T^\perp}(\hat{R})\|\), using (32a) and (32b) we have
\[ \|\hat{R}V\|_F = \frac{1}{\lambda} \|B_1 Q^T\Sigma^{-1/2}V\|_F \]
\[ \leq \frac{1}{\lambda} \|Q\|\|\Sigma^{-1/2}\|\|B_1\|_F \]
\[ \leq \frac{2q}{\lambda} \|\nabla g(X, Y)\|_F \]
\[ \Vert R^T U \Vert_F = \Vert V (\Sigma^{-1/2} Q Q^T \Sigma^{-1/2} - \Sigma^{-1/2} Q^{-1} Q^{-1} \Sigma^{-1/2}) \]

\[ - \frac{1}{\lambda} B_2 Q^T \Sigma^{-1/2} U \Vert_F \]

then we have

\[ \Vert \Sigma^{-1/2} (Q Q^T - I_r) \Sigma^{-1/2} \Vert_F \]

\[ + \Vert \Sigma^{-1/2} (Q^T Q^{-1} - I_r) \Sigma^{-1/2} \Vert_F \]

\[ + \frac{1}{\lambda} \Vert B_2 Q^T \Sigma^{-1/2} U \Vert_F \]

\[ \leq 64 \kappa \sqrt{\sigma_{\min}} \Vert \nabla g(X, Y) \Vert_F \]

\[ + \frac{2}{\lambda \sqrt{\sigma_{\min}/2}} \Vert \nabla g(X, Y) \Vert_F . \]

Combining the two bounds we have

\[ \Vert P_{\tilde{T}} (\tilde{R}) \Vert \leq \Vert P_{\tilde{T}} (\tilde{R}) \Vert_F \leq \Vert R^T U \Vert_F + \Vert \tilde{R} V \Vert_F \]

\[ \leq \frac{64 \kappa + 8/\lambda}{\sqrt{\sigma_{\min}}} \Vert \nabla g(X, Y) \Vert_F \]

\[ \leq \frac{\lambda \beta}{4}, \]

where the last line comes from (15) and Lemma 6.

**H. Proof of Lemma 9**

Reusing the definitions of \( B_1, B_2 \) in (31a) and (31b). We can then write

\[ X^T X - Y^T Y \]

\[ = \frac{1}{\lambda} \left[ X^T \left( B_1 - \frac{1}{\theta} \mathcal{P}(XY^T - M) Y \right) \right] \]

\[ - \left( B_2 - \frac{1}{\theta} \mathcal{P}(XY^T - M) X^T \right) Y \]

\[ = \frac{1}{\lambda} \left( X^T B_1 - B_2^T Y \right), \]

which further implies

\[ \Vert X^T X - Y^T Y \Vert_F \leq \frac{1}{\lambda} \Vert X^T B_1 - B_2^T Y \Vert_F \]

\[ \leq \frac{1}{\lambda} \left( \Vert X \Vert_F \Vert B_1 \Vert_F + \Vert B_2 \Vert_F \Vert Y \Vert_F \right) \]

\[ \leq 2 \sqrt{2 \sigma_{\max}} \Vert \nabla g(X, Y) \Vert_F . \]

Here, the last inequality uses the fact that

\[ \max \{ \Vert B_1 \Vert_F, \Vert B_2 \Vert_F \} \leq \Vert \nabla g(X, Y) \Vert_F \]

and that max\( \{ \Vert X \Vert, \Vert Y \Vert \} \) \leq \( \sqrt{2 \sigma_{\max}} \).

In addition, since

\[ \min \{ \sigma_{\min}(X), \sigma_{\min}(Y) \} \geq \sqrt{\sigma_{\min}/2}, \]

we have \( \sigma_{\min}(XY^T) \geq \sigma_{\min}/2 \), which together with Lemma 20 in the paper [7] implies the existence of an invertible \( Q \in \mathbb{R}^{r \times r} \) such that

\[ X = U \Sigma^{1/2} Q, Y = V \Sigma^{1/2} Q^{-T}, \]

and

\[ \Vert \Sigma_Q - \Sigma_Q^{-1} \Vert_F \leq \frac{2}{\sigma_{\min}} \Vert X^T X - Y^T Y \Vert_F \]

\[ \leq 4 \sqrt{2 \sigma_{\max}} \Vert \nabla g(X, Y) \Vert_F \]

\[ \leq 4 \sqrt{2 \kappa} / \sqrt{n^2 p} \leq 1, \]

and hence \( \Vert Q \Vert = \Vert \Sigma_Q \Vert = \sigma_{\max}(\Sigma_Q) \leq 2 \). As a result, we have

\[ \Vert \Sigma^{1/2} Q Q^T \Sigma^{-1/2} - I_r \Vert \]

\[ \leq 2 \sqrt{2 \sigma_{\max}} \Vert \nabla g(X, Y) \Vert_F \leq 1 / 3, \]

where the last inequality again uses the assumed bound (15).

**ACKNOWLEDGMENT**

The authors would like to thank the associate editor and two reviewers for their careful reviews and helpful comments that improve the article.

**REFERENCES**

[1] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Dec. 2009.

[2] R. H. Keshavan, A. Montanari, and S. Oh, “Matrix completion from a few entries,” *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

[3] J. D. M. Rennie and N. Srebro, “Fast maximum margin matrix factorization,” in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, Aug. 2005, pp. 713–719.

[4] S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi, “Matrix completion methods for causal panel data models,” *J. Amer. Stat. Assoc.*, vol. 116, no. 536, pp. 1716–1730, Oct. 2021.

[5] P. Biswas, T.-C. Lian, T.-C. Wang, and Y. Ye, “Semidefinite programming based algorithms for sensor network localization,” *ACM Trans. Sensor Netw.*, vol. 2, no. 2, pp. 188–220, May 2006.

[6] E. J. Candès and Y. Plan, “Matrix completion with noise,” *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[7] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan, “Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization,” *SIAM J. Optim.*, vol. 30, no. 4, pp. 3098–3121, Jan. 2020.

[8] A. Belloni, V. Chernozhukov, and L. Wang, “Square-root lasso: Pivotal recovery of sparse signals via conic programming,” *Biometrika*, vol. 98, no. 4, pp. 791–806, Dec. 2011.

[9] O. Klopp, “Noisy low-rank matrix completion with general sampling distribution,” *Bernoulli*, vol. 20, no. 1, pp. 282–303, Feb. 2014.
Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.