Journal of Machine Learning Research 19 (2018) 1-39 Submitted 8/17; Revised 10/18; Published 11/18

Robust PCA by Manifold Optimization

Teng Zhang TENG.ZHANG@UCF.EDU
Department of Mathematics
University of Central Florida
4000 Central Florida Blvd
Orlando, FL 32816, USA

Yi Yang YI.YANG6@MCGILL.CA
Department of Mathematics and Statistics
McGill University
805 Sherbrooke Street West
Montreal, QC H3A0B9, Canada

Editor: Michael Mahoney

Abstract

Robust PCA is a widely used statistical procedure to recover an underlying low-rank matrix with grossly corrupted observations. This work considers the problem of robust PCA as a nonconvex optimization problem on the manifold of low-rank matrices and proposes two algorithms based on manifold optimization. It is shown that, with a properly designed initialization, the proposed algorithms are guaranteed to converge to the underlying low-rank matrix linearly. Compared with a previous work based on the factorization of low-rank matrices Yi et al. (2016), the proposed algorithms reduce the dependence on the condition number of the underlying low-rank matrix theoretically. Simulations and real data examples confirm the competitive performance of our method.

Keywords: principal component analysis, low-rank modeling, manifold of low-rank matrices.

1. Introduction

In many problems, the underlying data matrix is assumed to be approximately low-rank. Examples include problems in computer vision Epstein et al. (1995); Ho et al. (2003), machine learning Deerwester et al. (1990), and bioinformatics Price et al. (2006). For such problems, principal component analysis (PCA) is a standard statistical procedure to recover the underlying low-rank matrix. However, PCA is highly sensitive to outliers in the data, and robust PCA Candès et al. (2011); Chandrasekaran et al. (2011); Clarkson and Woodruff (2013); Frieze et al. (2004); Bhojanapalli et al. (2015); Yi et al. (2016); Chen and Wainwright (2015); Gu et al. (2016); Cherapanamjeri et al. (2016); Netrapalli et al. (2014) is hence proposed as a modification to handle grossly corrupted observations. Mathematically, the robust PCA problem is formulated as follows: given a data matrix $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2}$ that can be written as the sum of a low-rank matrix $\mathbf{L}^*$ (signal) and a sparse matrix $\mathbf{S}^*$ (corruption) with only a few nonzero entries, can we recover both components accurately? Robust PCA has been shown to have applications in many real-life applications including background detection Li et al. (2004), face recognition Basri and Jacobs (2003), ranking, and collaborative filtering Candès et al. (2011).

Since the set of all low-rank matrices is nonconvex, it is generally difficult to obtain an algorithm with theoretical guarantee since there is no tractable optimization algorithm for the nonconvex problem. Here we review a few carefully designed algorithms such that the theoretical guarantee on the recovery of underlying low-rank matrix exists. The works Candès et al. (2011); Chandrasekaran et al. (2011) consider the convex relaxation of the original problem instead:

$$\min_{\mathbf{L},\mathbf{S}} \|\mathbf{L}\|_* + \|\mathbf{S}\|_1, \text{ s.t. } \mathbf{Y} = \mathbf{L} + \mathbf{S}, \tag{1}$$

where $\|\mathbf{L}\|_*$ represents the nuclear norm (i.e., Schatten 1-norm) of $\mathbf{L}$, defined by the sum of its singular values and $\|\mathbf{S}\|_1$ represents the sum of the absolute values of all entries of $\mathbf{S}$. Since this problem is convex, the solution to (1) can be solved in polynomial time. In addition, it is shown that the solution recovers the correct low-rank matrix when $\mathbf{S}^*$ has at most $\gamma^* = O(1/\mu^2 r)$ fraction of corrupted non-zero entries, where r is the rank of $\mathbf{L}^*$ and μ is the incoherence level of $\mathbf{L}^*$ Hsu et al. (2011). If the sparsity of $\mathbf{S}^*$ is assumed to be random, then Candès et al. (2011) shows that the algorithm succeeds with high probability, even when the percentage of corruption can be in the order of $O(1)$ while the rank $r = O(\min(n_1, n_2)/\mu \log^2 \max(n_1, n_2))$, where μ is a coherence parameter of the low-rank matrix $\mathbf{L}^*$ (this work defines μ slightly differently compared to Candès et al. (2011) and (16) in this work, but the value is comparable).

However, the aforementioned algorithms based on convex relaxation have a computational complexity of $O(n_1 n_2 \min(n_1, n_2))$ per iteration, which could be prohibitive when n_1 and n_2 are very large. Alternatively, some faster algorithms are proposed based on nonconvex optimization. In particular, the work by Kyrillidis and Cevher (2012) proposes a method based on the projected gradient method. However, it assumes that the sparsity pattern of $\mathbf{S}^*$ is random, and the algorithm still has the same computational complexity as the convex methods. Netrapalli et al. (2014) proposes a method based on the alternating projecting, which allows $\gamma^* \leq \frac{1}{\mu^2 r}$, with a computational complexity of $O(r^2 n_1 n_2)$ per iteration. Chen and Wainwright (2015) assumes that $\mathbf{L}^*$ is positive semidefinite and applies the gradient descent method on the Cholesky decomposition factor of $\mathbf{L}^*$, but the positive semidefinite assumption is not satisfied in many applications. Gu et al. (2016) factorizes $\mathbf{L}^*$ into the product of two matrices and performs alternating minimization over both matrices. It shows that the algorithm allows $\gamma^* = O(1/\mu^{2/3} r^{2/3} \min(n_1, n_2))$ and has the complexity of $O(r^2 n_1 n_2)$ per iteration. Yi et al. (2016) applies a similar factorization and applies an alternating gradient descent algorithm with a complexity of $O(r n_1 n_2)$ per iteration and allows $\gamma^* = O(1/\kappa^2 \mu r^{3/2})$, where κ is the condition number of the underlying low-rank matrix. There is another line of works that further reduces the complexity of the algorithm by subsampling the entries of the observation matrix $\mathbf{Y}$, including Mackey et al. (2011); Li and Haupt (2015); Rahmani and Atia (2017); Cherapanamjeri et al. (2016) and (Yi et al., 2016, Algorithm 2), which will also be discussed in this paper as the partially observed case.

The common idea shared by Gu et al. (2016) and Yi et al. (2016) is as follows. Since any low-rank matrix $\mathbf{L} \in \mathbb{R}^{n_1 \times n_2}$ with rank r can be written as the product of two low-rank matrices by $\mathbf{L} = \mathbf{U}\mathbf{V}^T$ with $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times r}$, we can optimize the pair $(\mathbf{U}, \mathbf{V})$ instead of $\mathbf{L}$, and a smaller computational cost is expected since $(\mathbf{U}, \mathbf{V})$ has $(n_1 + n_2)r$ parameters, which is smaller than $n_1 n_2$, the number of parameters in $\mathbf{L}$. In fact, such a re-parametrization technique has a long history Ruhe (1974), and has been popularized by Burer and Monteiro Burer and Monteiro (2003, 2005) for solving semi-definite programs (SDPs). The same idea has been used in other low-rank matrix estimation problems such as dictionary learning Sun et al. (2017), phase synchronization Boumal (2016), community detection Bandeira et al. (2016), matrix completion Jain et al. (2013), recovering matrix from linear measurements Tu et al. (2016), and even general problems Chen and Wainwright (2015); Wang et al. (2017); Park et al. (2016); Wang et al. (2017); Park et al. (2017). In addition, the property of associated stochastic gradient descent algorithm is studied in De Sa et al. (2015).

The main contribution of this work is a novel robust PCA algorithm based on the gradient descent algorithm on the manifold of low-rank matrices, with a theoretical guarantee on the exact recovery of the underlying low-rank matrix. Compared with Yi et al. (2016), the proposed algorithm utilizes the tool of manifold optimization, which leads to a simpler and more naturally structured algorithm with a stronger theoretical guarantee. In particular, with a proper initialization, our method can still succeed with $\gamma^* = O(1/\kappa\mu r^{3/2})$, which means that it can tolerate more corruption than Yi et al. (2016) by a factor of κ. In addition, the theoretical convergence rate is also faster than Yi et al. (2016) by a factor of κ. Simulations also verified the advantage of the proposed algorithm over Yi et al. (2016). We remark that while manifold optimization has been applied to robust PCA in Cambier and Absil (2016), our work studies a different algorithm and gives theoretical guarantees. Considering the popularity of the methods based on the factorization of low-rank matrices, it is expected that manifold optimization could be applied to other low-rank matrix estimation problems. In addition, we implement our method in an efficient and user-friendly R package `morpca`, which is available at `https://github.com/emeryyi/morpca`.

The paper is organized as follows. We first present the algorithm in Section 2, and explain how the proposed algorithms are derived in Section 3. Their theoretical properties are studied and compared with previous algorithms in Section 4. In Section 5, simulations and real data analysis on the `Shoppingmall` dataset show that the proposed algorithms are competitive in many scenarios and have superior performances to the algorithm based on matrix factorization. A discussion about the proposed algorithms is then presented in Section 6, followed by the proofs of the results in Appendix.

2. Algorithm

In this work, we consider the robust PCA problem in two settings: fully observed setting and partially observed setting. The problem under the fully observed setting can be formulated as follows: given $\mathbf{Y} = \mathbf{L}^* + \mathbf{S}^*$, where $\mathbf{L}^*$ is a low-rank matrix and $\mathbf{S}^*$ is a sparse matrix, then can we recover $\mathbf{L}^*$ from $\mathbf{Y}$? To recover $\mathbf{L}^*$, we solve the following optimization problem:

$$\widehat{\mathbf{L}} = \operatorname*{arg\,min}_{\operatorname{rank}(\mathbf{L})=r} f(\mathbf{L}), \text{ where } f(\mathbf{L}) = \frac{1}{2}\|F(\mathbf{L} - \mathbf{Y})\|_F^2, \tag{2}$$

where $F : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ is a hard thresholding procedure defined in (3):

$$F_{ij}(\mathbf{A}) = \begin{cases} 0, & \text{if } |\mathbf{A}_{ij}| > |\mathbf{A}_{i,\cdot}|^{[\gamma]} \text{ and } |\mathbf{A}_{ij}| > |\mathbf{A}_{\cdot,j}|^{[\gamma]} \\ \mathbf{A}_{ij}, & \text{otherwise.} \end{cases} \tag{3}$$

Here $\mathbf{A}_{i,\cdot}$ represents the i-th row of the matrix $\mathbf{A}$, and $\mathbf{A}_{\cdot,j}$ represents the j-th column of $\mathbf{A}$. $|\mathbf{A}_{i,\cdot}|^{[\gamma]}$ and $|\mathbf{A}_{\cdot,j}|^{[\gamma]}$ represent the $(1-\gamma)$-th percentile of the absolute values of the entries of $\mathbf{A}_{i,\cdot}$ and $\mathbf{A}_{\cdot,j}$ for $\gamma \in [0, 1)$. In other words, what are removed are the entries that are simultaneously among the largest γ-fraction in the corresponding row and column of $\mathbf{A}$ in terms of the absolute values. The threshold γ is set by users. If some entries of $\mathbf{A}_{i,\cdot}$ or $\mathbf{A}_{\cdot,j}$ have the entries with identical absolute values, the ties can be broken down arbitrarily.

The motivation is that, if $\mathbf{S}^*$ is sparse in the sense that the percentage of nonzero entries in each row and each column is smaller than γ, then $F(\mathbf{L}^* - \mathbf{Y}) = F(-\mathbf{S}^*)$ is zero by definition thus $f(\mathbf{L}^*)$ is zero. Since f is nonnegative, $\mathbf{L}^*$ is the solution to (2). To solve (2), we propose Algorithm 1 based on manifold optimization, with its derivation deferred to Section 3.3.1.

Algorithm 1 Gradient descent on the manifold under the fully observed setting.

```
Input: Observation $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2}$; Rank $r$; Thresholding value $\gamma$; Step size $\eta$.
Initialization: Set $k = 0$; Initialize $\mathbf{L}^{(0)}$ using the rank-$r$ approximation to $F(\mathbf{Y})$.
Loop: Iterate Steps 1–4 until convergence:
1: Let $\mathbf{L}^{(k)} = \mathbf{U}^{(k)}\mathbf{\Sigma}^{(k)}\mathbf{V}^{(k)T}$.
2: Let $\mathbf{D}^{(k)} = F(\mathbf{L}^{(k)} - \mathbf{Y})$.
3(a): (Option 1) Let $\mathbf{\Omega}^{(k)} = \mathbf{U}^{(k)}\mathbf{U}^{(k)T}\mathbf{D}^{(k)} + \mathbf{D}^{(k)}\mathbf{V}^{(k)}\mathbf{V}^{(k)T} - \mathbf{U}^{(k)}\mathbf{U}^{(k)T}\mathbf{D}^{(k)}\mathbf{V}^{(k)}\mathbf{V}^{(k)T}$, and let $\mathbf{U}^{(k+1)} \in \mathbb{R}^{n_1 \times r}$, $\mathbf{\Sigma}^{(k+1)} \in \mathbb{R}^{r \times r}$, and $\mathbf{V}^{(k+1)} \in \mathbb{R}^{n_2 \times r}$ be matrices consist of the top $r$ left singular vectors/singular values/right singular vectors of $\mathbf{L}^{(k)} - \eta\mathbf{\Omega}^{(k)}$.
3(b): (Option 2) Let $\mathbf{Q}_1, \mathbf{R}_1$ be the QR decomposition of $(\mathbf{L}^{(k)} - \eta\mathbf{D}^{(k)})^T\mathbf{U}^{(k)}$ and $\mathbf{Q}_2, \mathbf{R}_2$ be the QR decomposition of $(\mathbf{L}^{(k)} - \eta\mathbf{D}^{(k)})\mathbf{V}^{(k)}$. Then $\mathbf{U}^{(k+1)} = \mathbf{Q}_2$, $\mathbf{V}^{(k+1)} = \mathbf{Q}_1$ and $\mathbf{\Sigma}^{(k+1)} = \mathbf{R}_2[\mathbf{U}^{(k)T}(\mathbf{L}^{(k)} - \eta\mathbf{D}^{(k)})\mathbf{V}^{(k)}]^{-1}\mathbf{R}_1^T$.
4: $k := k + 1$.
Output: Estimation of the low-rank matrix $\mathbf{L}^*$, given by $\lim_{k\to\infty} \mathbf{L}^{(k)}$.
```

Under the partially observed setting, in addition to gross corruption $\mathbf{S}^*$, the observed matrix $\mathbf{Y}$ has a large number of missing values, i.e., many entries of $\mathbf{Y}$ are not observed. We denote the set of all observed entries by $\mathbf{\Phi} = \{(i,j) | \mathbf{Y}_{ij} \text{ is observed}\}$, and define $\tilde{F} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$

$$\tilde{F}_{ij}(\mathbf{A}) = \begin{cases} 0, & \text{if } |\mathbf{A}_{ij}| > |\mathbf{A}_{i,\cdot}|^{[\gamma,\mathbf{\Phi}]} \text{ and } |\mathbf{A}_{ij}| > |\mathbf{A}_{\cdot,j}|^{[\gamma,\mathbf{\Phi}]} \\ \mathbf{A}_{ij}, & \text{otherwise.} \end{cases} \tag{4}$$

Here $|\mathbf{A}_{i,\cdot}|^{[\gamma,\mathbf{\Phi}]}$ and $|\mathbf{A}_{\cdot,j}|^{[\gamma,\mathbf{\Phi}]}$ represent the $(1-\gamma)$-th percentile of the absolute values of the observed entries of $\mathbf{A}_{i,\cdot}$ and $\mathbf{A}_{\cdot,j}$ of the matrix $\mathbf{A}$ respectively.

As a generalization of Algorithm 1, we propose to solve

$$\mathop{\arg\min}_{\operatorname{rank}(\mathbf{L})=r} \tilde{f}(\mathbf{L}), \quad \tilde{f}(\mathbf{L}) = \frac{1}{2} \sum_{(i,j)\in\mathbf{\Phi}} \tilde{F}_{ij}(\mathbf{L} - \mathbf{Y})^2, \tag{5}$$

Algorithm 2 Gradient descent on the manifold under the partially observed setting.

```
Input: Observation $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2}$; Set of all observed entries by $\mathbf{\Phi}$; Rank $r$; Thresholding value $\gamma$; Step size $\eta$.
Initialization: Set $k = 0$; Initialize $\mathbf{L}^{(0)}$ using the rank-$r$ approximation to $\tilde{F}(\mathbf{Y})$.
Loop: Iterate Steps 1–4 until convergence:
1: Let $\mathbf{L}^{(k)}$ be a sparse matrix with support $\mathbf{\Phi}$, with nonzero entries given by the corresponding entries of $\mathbf{U}^{(k)}\mathbf{\Sigma}^{(k)}\mathbf{V}^{(k)\,T}$.
2: Let $\mathbf{D}^{(k)} = \tilde{F}(\mathbf{L}^{(k)} - \mathbf{Y})$.
3(a): (Option 1) Let $\mathbf{\Omega}^{(k)} = \mathbf{U}^{(k)}\mathbf{U}^{(k)\,T}\mathbf{D}^{(k)} + \mathbf{D}^{(k)}\mathbf{V}^{(k)}\mathbf{V}^{(k)\,T} - \mathbf{U}^{(k)}\mathbf{U}^{(k)\,T}\mathbf{D}^{(k)}\mathbf{V}^{(k)}\mathbf{V}^{(k)\,T}$, and let $\mathbf{U}^{(k+1)} \in \mathbb{R}^{n_1 \times r}$, $\mathbf{\Sigma}^{(k+1)} \in \mathbb{R}^{r \times r}$, and $\mathbf{V}^{(k+1)} \in \mathbb{R}^{n_2 \times r}$ be matrices consists of the top $r$ left singular vectors/singular values/right singular vectors of $\mathbf{L}^{(k)} - \eta\mathbf{\Omega}^{(k)}$.
3(b): (Option 2) Let $\mathbf{Q}_1, \mathbf{R}_1$ be the QR decomposition of $(\mathbf{L}^{(k)} - \eta\mathbf{D}^{(k)})^T\mathbf{U}^{(k)}$ and $\mathbf{Q}_2, \mathbf{R}_2$ be the QR decomposition of $(\mathbf{L}^{(k)} - \eta\mathbf{D}^{(k)})\mathbf{V}^{(k)}$. Then $\mathbf{U}^{(k+1)} = \mathbf{Q}_2$, $\mathbf{V}^{(k+1)} = \mathbf{Q}_1$ and $\mathbf{\Sigma}^{(k+1)} = \mathbf{R}_2[\mathbf{U}^{(k)\,T}(\mathbf{L}^{(k)} - \eta\mathbf{D}^{(k)})\mathbf{V}^{(k)}]^{-1}\mathbf{R}_1^T$.
4: $k := k + 1$.
Output: Estimation of the low-rank matrix $\mathbf{L}^*$, given by $\lim_{k\to\infty} \mathbf{L}^{(k)}$.
```

which is similar to (2) but only the observed entries are considered. The implementation is presented in Algorithm 2 and its derivation is deferred to Section 3.3.2.

For Algorithm 1, its memory usage is $O(n_1 n_2)$ due to the storage of $\mathbf{Y}$. For Algorithm 2, storing $\mathbf{Y}$ and $\mathbf{L}^{(k)}$ requires $O(|\mathbf{\Phi}|)$ and storing $\mathbf{U}^{(k)}$ and $\mathbf{V}^{(k)}$ requires $O(r(n_1+n_2))$. Adding them together, the memory usage is $O(|\mathbf{\Phi}| + r(n_1 + n_2))$.

For both Algorithm 1 and Algorithm 2 with Option 1, the singular value decomposition is the most computationally intensive step and as a result, the complexity per iteration is $O(rn_1n_2)$. For Algorithm 1 and Algorithm 2 with Option 2, their computational complexities per iteration are in the order of $O(rn_1n_2)$ and $O(r^2(n_1 + n_2) + r|\mathbf{\Phi}|)$ respectively.

3. Derivation of the Proposed Algorithms

This section gives the derivations of Algorithms 1 and 2. Since they are derived from manifold optimization, we first give a review of manifold optimization in Section 3.1 and the geometry of the manifold of low-rank matrices in Section 3.2.

3.1. Manifold optimization

The purpose of this section is to review the framework of the gradient descent method on manifolds. It summarizes mostly the framework used in Vandereycken (2013); Shalit et al. (2012); Absil et al. (2009), and we refer readers to these work for more details.

Given a smooth manifold $\mathcal{M} \subset \mathbb{R}^n$ and a differentiable function $f : \mathcal{M} \to \mathbb{R}$, the procedure of the gradient descent algorithm for solving $\min_{x\in\mathcal{M}} f(x)$ is as follows:

Step 1. Consider $f(x)$ as a differentiable function from $\mathbb{R}^n$ to $\mathbb{R}$ and calculate the Euclidean gradient $\nabla f(x)$.

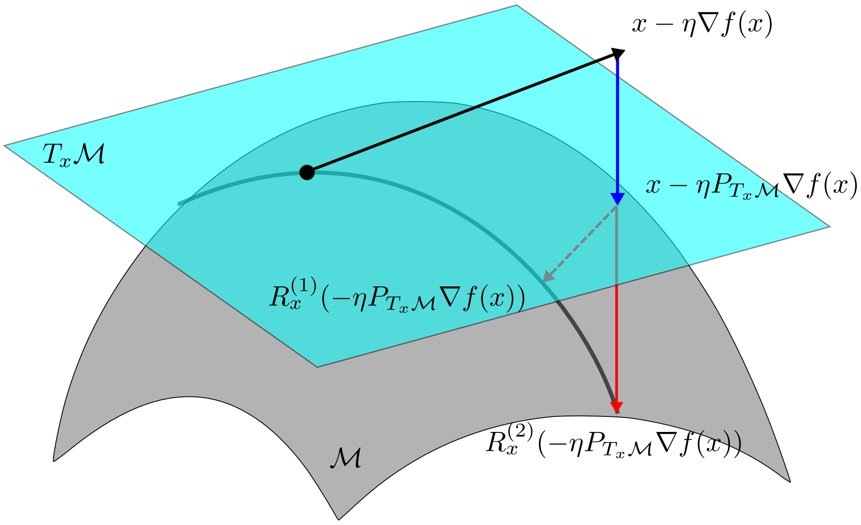
$x - \eta \nabla f(x)$
$T_x\mathcal{M}$
$x - \eta P_{T_x\mathcal{M}} \nabla f(x)$
$R_x^{(1)}(-\eta P_{T_x\mathcal{M}} \nabla f(x))$
$R_x^{(2)}(-\eta P_{T_x\mathcal{M}} \nabla f(x))$
$\mathcal{M}$

Figure 1: The visualization of gradient descent algorithms on the manifold $\mathcal{M}$. The black solid line is the Euclidean gradient. The blue solid line is the projection of the Euclidean gradient to the tangent space. The red solid line represents the orthographic retraction, while the red dashed line represents the projective retraction.

Step 2. Calculate its Riemannian gradient, which is the direction of steepest ascent of $f(x)$ among all directions in the *tangent space* $T_x\mathcal{M}$. This direction is given by $P_{T_x\mathcal{M}} \nabla f(x)$, where $P_{T_x\mathcal{M}}$ is the projection operator to the tangent space $T_x\mathcal{M}$.

Step 3. Define a *retraction* R_x that maps the tangent space back to the manifold, i.e. $R_x : T_x\mathcal{M} \to \mathcal{M}$, where R_x needs to satisfy the conditions in (Vandereycken, 2013, Definition 2.2). In particular, $R_x(0) = x$, $R_x(y) = x + y + O(\|y\|^2)$ as $y \to 0$, and R_x needs to be smooth. Then the update of the gradient descent algorithm x^+ is defined by

$$x^+ = R_x(-\eta P_{T_x\mathcal{M}} \nabla f(x)), \tag{6}$$

where η is the step size.

We remark that in differential geometry, the standard “retraction” is the exponential map from the tangent space to the manifold. However, in this work (as well as many works on manifold optimization) it is used to represent a generic mapping from the tangent plane to the manifold. As a result, the definition of retraction is not unique in this work. In Figure 1, we visualize the gradient descent method on the manifold $\mathcal{M}$ with two different kinds of retractions (orthographic and projective). We will discuss the details of those two retractions in Section 3.2.

3.2. The geometry of the manifold of low-rank matrices

To apply the gradient descent algorithm in Section 3.1 to the manifold of the low-rank matrices, the projection $P_{T_x\mathcal{M}}$ and the retraction R_x need to be defined. In this section, we let $\mathcal{M}$ be the manifold of all $\mathbb{R}^{n_1\times n_2}$ matrices with rank r and $\mathbf{X}\in\mathcal{M}$ be a matrix of rank r and will find the explicit expressions of $P_{T_x\mathcal{M}}$ and R_x.

The tangent space $T_{\mathbf{X}}\mathcal{M}$ and the retraction $R_{\mathbf{X}}$ of the manifold of the low-rank matrices have been well-studied Absil and Oseledets (2015): Assume that the SVD decomposition of $\mathbf{X}$ is $\mathbf{X}=\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, then the tangent space $T_{\mathbf{X}}\mathcal{M}$ can be defined by $T_{\mathbf{X}}\mathcal{M}=\{\mathbf{A}\mathbf{V}\mathbf{V}^T+\mathbf{U}\mathbf{U}^T\mathbf{B}:\text{ for }\mathbf{A}\in\mathbb{R}^{n_1\times n_1},\mathbf{B}\in\mathbb{R}^{n_2\times n_2}\}$ according to Absil and Oseledets (2015). The explicit formula for the projection $P_{T_{\mathbf{X}}\mathcal{M}}$ is given in (Absil and Oseledets, 2015, (9)):

$$P_{T_{\mathbf{X}}\mathcal{M}}(\mathbf{D})=\mathbf{U}\mathbf{U}^T\mathbf{D}+\mathbf{D}\mathbf{V}\mathbf{V}^T-\mathbf{U}\mathbf{U}^T\mathbf{D}\mathbf{V}\mathbf{V}^T,\qquad \mathbf{D}\in\mathbb{R}^{n_1\times n_2}. \tag{7}$$

For completeness, a proof of (7) is presented in Appendix.

There are various ways of defining retractions for the manifold of low-rank matrices, and we refer the reader to Absil and Oseledets (2015) for more details. In this work, we consider two types of retractions. One is called the *projective* retraction Shalit et al. (2012); Vandereycken (2013), Given any $\delta\in T_{\mathbf{X}}\mathcal{M}$, the retraction is defined as the nearest low-rank matrix to $\mathbf{X}+\delta$ in terms of Frobenius norm:

$$R^{(1)}_{\mathbf{X}}(\delta)=\arg\min_{\mathbf{Z}\in\mathcal{M}}\|\mathbf{X}+\delta-\mathbf{Z}\|_F. \tag{8}$$

The solution is the rank-r approximation of $\mathbf{X}+\delta$ (for any matrix $\mathbf{W}$, its rank-r approximation is given by $\sum_{i=1}^r\sigma_i\mathbf{u}_i\mathbf{v}_i^T$, where $\sigma_i,\mathbf{u}_i,\mathbf{v}_i$ are the ordered singular values and vectors of $\mathbf{W}$).

In order to further improve computation efficiency, we also consider the *orthographic* retraction Absil and Oseledets (2015). Denoted by $R^{(2)}_{\mathbf{X}}(\delta)$, it is the nearest rank-r matrix to $\mathbf{X}+\delta$ that their difference is orthogonal to the tangent space $T_{\mathbf{X}}\mathcal{M}$:

$$R^{(2)}_{\mathbf{X}}(\delta)=\arg\min_{\mathbf{Z}\in\mathcal{M}}\|\mathbf{X}+\delta-\mathbf{Z}\|_F,\ \text{ s.t. }\langle R^{(2)}_{\mathbf{X}}(\delta)-(\mathbf{X}+\delta),\mathbf{Z}\rangle_F=0\text{ for all }\mathbf{Z}\in T_{\mathbf{X}}\mathcal{M}, \tag{9}$$

and its explicit solution of (9) is given in (Absil and Oseledets, 2015, Section 3.2),

$$R^{(2)}_{\mathbf{X}}(\delta)=(\mathbf{X}+\delta)\mathbf{V}[\mathbf{U}^T(\mathbf{X}+\delta)\mathbf{V}]^{-1}\mathbf{U}^T(\mathbf{X}+\delta), \tag{10}$$

and a proof is given in Appendix.

3.3. Derivation of the proposed algorithms

3.3.1. Derivation of Algorithm 1

The gradient descent algorithm (6) for solving (2) can be written as

$$\mathbf{L}^{(k+1)}=R_{\mathbf{L}^{(k)}}(-\eta P_{T_{\mathbf{L}^{(k)}}}\nabla f(\mathbf{L}^{(k)})), \tag{11}$$

where $P_{T_{\mathbf{L}^{(k)}}}$ is defined in (7) and $R_{\mathbf{L}^{(k)}}$ is defined in (8) or (10). To derive the explicit algorithm, it remains to find the gradient ∇f.

If the absolute values of all entries of $\mathbf{A}$ are different, then we have

$$\nabla f(\mathbf{L}) = F(\mathbf{L} - \mathbf{Y}). \tag{12}$$

The proof of (12) is deferred to Appendix. When some entries of $\mathbf{A}$ are equivalent and there is a tie in generating $F(\mathbf{L} - \mathbf{Y})$, the objective function could be non-differentiable. However, it can be shown that by arbitrarily breaking the tie, $F(\mathbf{L} - \mathbf{Y})$ is a subgradient of $f(\mathbf{L})$.

The corresponding gradient descent method (or subgradient method when f is not differentiable) with projective retraction can be written as follows:

$$\mathbf{L}^{(k+1)} := \text{rank-}r \text{ approximation of } \left[\mathbf{L}^{(k)} - \eta P_{T_{\mathbf{L}^{(k)}}} F(\mathbf{L}^{(k)} - \mathbf{Y})\right], \tag{13}$$

where the rank-r approximation has been defined after (8). This leads to Algorithm 1 with Option 1.

For the orthographic retraction, i.e., $R_{\mathbf{L}^{(k)}}$ defined according to (10), by writing $\mathbf{D} = F(\mathbf{L}^{(k)} - \mathbf{Y})$, the update formula (11) can be simplified to

$$\mathbf{L}^{(k+1)} := (\mathbf{L}^{(k)} - \eta\mathbf{D})\mathbf{V}^{(k)}[\mathbf{U}^{(k)T}(\mathbf{L}^{(k)} - \eta\mathbf{D})\mathbf{V}^{(k)}]^{-1}\mathbf{U}^{(k)T}(\mathbf{L}^{(k)} - \eta\mathbf{D}), \tag{14}$$

where $\mathbf{U}^{(k)} \in \mathbb{R}^{n_1 \times r}$ is any matrix such that its column space is the same as the column space of $\mathbf{L}^{(k)}$; and $\mathbf{V}^{(k)} \in \mathbb{R}^{n_2 \times r}$ is any matrix such that its column space is the same as the row space of $\mathbf{L}^{(k)}$. The derivation of (14) is deferred to Appendix, and it can be shown that the implementation of (14) leads to Algorithm 1 with Option 2.

3.3.2. Derivation of Algorithm 2

By a similar argument as in the previous section, we can conclude that when all entries of $|\mathbf{L} - \mathbf{Y}|$ are different from each other, then applying the same procedure of deriving (12), we have

$$\nabla \tilde{f}(\mathbf{L}) = \tilde{F}(\mathbf{L} - \mathbf{Y});$$

and $\tilde{F}(\mathbf{L} - \mathbf{Y})$ is a subgradient when $\tilde{f}(\mathbf{L})$ is not differentiable. Based on this observation, the algorithm under the partially observed setting is identical to (13) or (14), with F replaced by $\tilde{F}$. This gives the implementation of Algorithm 2.

3.3.3. Basic convergence properties of Algorithms 1 and 2

An interesting topic is that, can we still expect the algorithm to have reasonable basic properties, such as convergence to a critical point? Unfortunately, it is impossible to have such a theoretical guarantee if a fixed step size η is chosen: in general, the subgradient method with fixed step size does not have the convergence guarantee if the objective function is non-differentiable. However, if we choose step size with line search, then any accumulation point of $\mathbf{L}^{(k)}$, $\widehat{\mathbf{L}}$, would have the property that either the objective function is not differentiable at $\widehat{\mathbf{L}}$, or it is a critical point in the sense that its Riemannian gradient is zero. For example, the line search strategy for Algorithm 1 can be described as follows: start the step size η_k with a relatively large value, and repeatedly shrinks it by a factor of $\beta \in (0,1)$ such that the following condition is satisfied: for $\mathbf{L}^{(k+1)} = R_{\mathbf{L}^{(k)}}(-\eta_k P_{T_{\mathbf{L}^{(k)}}} \nabla f(\mathbf{L}^{(k)}))$,

$$f(\mathbf{L}^{(k)}) - f(\mathbf{L}^{(k+1)}) > c\eta_k \|P_{T_{\mathbf{L}^{(k)}}} \nabla f(\mathbf{L}^{(k)})\|^2,$$

where $c \in (0,1)$ is prespecified. The argument for convergence follows from the same argument as the proof of (Absil et al., 2009, Theorem 4.3.1).

3.4. Prior works on manifold optimization

The idea of optimization on manifolds has been well investigated in the literature Vandereycken (2013); Shalit et al. (2012); Absil et al. (2009). For example, Absil et al. Absil et al. (2009) give a summary of many advances in the field of optimization on manifolds. Manifold optimization has been applied to many matrix estimation problems, including recovering a low rank matrix from its partial entries, i.e., matrix completion Keshavan et al. (2010); Vandereycken (2013); Wei et al. (2016) and robust matrix completion in Cambier and Absil (2016). In fact, the problem studied in this work can be reformulated to the problem analyzed in Cambier and Absil (2016). In comparison, our work studies a different algorithm and gives additional theoretical guarantees.

In another aspect, while Wei et al. (2016) studies matrix completion, it shares some similarities with this work: both works study manifold optimization algorithms and have theoretical guarantees showing that the proposed algorithms can recover the underlying low-rank matrix exactly. In fact, Wei et al. (2016) can be considered as our problem under the partially observed setting, without corruption $\mathbf{S}^*$. It proposes to solve

$$\operatorname*{arg\,min}_{\mathbf{L}\in\mathbb{R}^{n_1\times n_2},\mathrm{rank}(\mathbf{L})=r} \sum_{(i,j)\in\Phi} (\mathbf{Y}_{ij}-\mathbf{L}_{ij})^2,$$

which can be considered as $\tilde{f}$ in (5) when $\gamma = 0$.

4. Theoretical Analysis

In this section, we analyze the theoretical properties of Algorithms 1 and 2 and compare them with previous algorithms. Since the goal is to recover the low-rank matrix $\mathbf{L}^*$ and the sparse matrix $\mathbf{S}^*$ from $\mathbf{Y} = \mathbf{L}^* + \mathbf{S}^*$, to avoid identifiability issues, we need to assume that $\mathbf{L}^*$ can not be both low-rank and sparse. Specifically, we make the following standard assumptions on $\mathbf{L}^*$ and $\mathbf{S}^*$:

Assumption 1 *Each row of $\mathbf{S}^*$ contains at most $\gamma^* n_2$ nonzero entries and each column of $\mathbf{S}^*$ contains at most $\gamma^* n_1$ nonzero entries. In other words, for $\gamma^* \in [0,1)$, assume $\mathbf{S}^* \in \mathcal{S}_{\gamma^*}$ where*

$$\mathcal{S}_{\gamma^*} := \left\{A \in \mathbb{R}^{n_1\times n_2} \mid \|A_{i,\cdot}\|_0 \le \gamma^* n_2, \text{ for } 1\le i\le n_1; \|A_{\cdot,j}\|_0 \le \gamma^* n_1, \text{ for } 1 \le j \le n_2\right\}. \tag{15}$$

Assumption 2 *The low-rank matrix $\mathbf{L}^*$ is not near-sparse. To achieve this, we require that $\mathbf{L}^*$ must be μ-coherent. Given the singular value decomposition (SVD) $\mathbf{L}^* = \mathbf{U}^*\mathbf{\Sigma}^*\mathbf{V}^{*T}$, where $\mathbf{U}^* \in \mathbb{R}^{n_1\times r}$ and $\mathbf{V}^* \in \mathbb{R}^{n_2\times r}$, we assume there exists an incoherence parameter μ such that*

$$\|\mathbf{U}^*\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_1}}, \;\; \|\mathbf{V}^*\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_2}}, \tag{16}$$

where the norm $\|\cdot\|_{2,\infty}$ is defined by $\|\mathbf{A}\|_{2,\infty} = \max_{\|\mathbf{z}\|_2=1}\|\mathbf{A}\mathbf{z}\|_\infty$ and $\|\mathbf{x}\|_\infty = \max_i |\mathbf{x}_i|$.

4.1. Analysis of Algorithm 1

With Assumption 1 and 2, we have the following theoretical results regarding the convergence rate, initialization, and stability of Algorithm 1:

Theorem 1 (Linear convergence rate, fully observed case) *Suppose that* $\|\mathbf{L}^{(0)}-\mathbf{L}^*\|_F \leq a\sigma_r(\mathbf{L}^*)$*, where* $\sigma_r(\mathbf{L}^*)$ *is the* r*-th largest singular value of* $\mathbf{L}^*$*,* $a \leq 1/2$*,* $\gamma > 2\gamma^*$ *and* $C_1 = \sqrt{4(\gamma+2\gamma^*)\mu r + 4\frac{\gamma^*}{\gamma-\gamma^*} + a^2} < \frac{1}{2}$*, then there exists* $\eta_0 = \eta_0(C_1, a) > 0$ *that does not depend on* k*, such that for all* $\eta \leq \eta_0$*,*

$$\|\mathbf{L}^{(k)} - \mathbf{L}^*\|_F \leq \Big(1 - \frac{1-2C_1}{8}\eta\Big)^k \|\mathbf{L}^{(0)} - \mathbf{L}^*\|_F.$$

Remark 2 *(Choices of parameters). It is shown in the proof that* η_0 *can be set to the solution of the equation*

$$\eta_0(1+C_1)^2 \left[\frac{1}{2} + \frac{a^2}{1-\eta_0(1+C_1)a}\right] = \frac{1}{8}(1-2C_1).$$

Since the LHS is an increasing function of η_0 *and is zero when* $\eta_0 = 0$*, and its RHS is a positive number.*

While $C_1 < 1/2$ *requires* $\sqrt{4\gamma^*/(\gamma-\gamma^*)} < 1/2$*, i.e.,* $\gamma > 17\gamma^*$*. In practice a much smaller* γ *can be used. In Section 5,* $\gamma = 1.5\gamma^*$ *is used and works well for a large number of examples. It suggests that some constants in Theorem 1 might be due to the technicalities in the proof and can be potentially improved.*

Remark 3 *(Simplified choices of parameters) There exists* c_1 *and* c_2 *such that if* $a < c_1$*,* $\gamma^*\mu r < c_2$ *and* $\gamma = 65\gamma^*$*, then one can choose* $\eta_0 = 1/8$*. In this sense, if the initialization of the algorithm is good, then the algorithm can handle* γ^* *as large as* $O(1/\mu r)$*. In addition, it requires* $O(\log(1/\epsilon))$ *iterations to achieve* $\|\mathbf{L}^{(k)} - \mathbf{L}^*\|_F/\|\mathbf{L}^{(0)} - \mathbf{L}^*\|_F < \epsilon$*.*

Since the statements require proper initializations (i.e., small a), the question arises as to how to choose proper initializations. The work by Yi et al. (2016) shows that if the rank-r approximation to $F(\mathbf{Y})$ is used as the initialization $\mathbf{L}^{(0)}$, then such initialization has the upper bound $\|\mathbf{L}^{(0)} - \mathbf{L}^*\|$ according to the proofs of (Yi et al., 2016, Theorems 1 and 3) (we borrow this estimation along with the fact that $\|\mathbf{L}^{(0)} - \mathbf{L}^*\|_F \leq \sqrt{2r}\|\mathbf{L}^{(0)} - \mathbf{L}^*\|$).

Theorem 4 (Initialization, fully observed case) *If* $\gamma > \gamma^*$ *and we initialize* $\mathbf{L}^{(0)}$ *using the rank-*r *approximation to* $F(\mathbf{Y})$*, then*

$$\|\mathbf{L}^{(0)} - \mathbf{L}^*\|_F \leq 8\gamma\mu r\sqrt{2r}\sigma_1(\mathbf{L}^*).$$

The combination of Theorem 1, 4 and the fact that $\gamma = O(\gamma^*)$ implies that under the fully observed setting, the tolerance of the proposed algorithms to corruption is at most $\gamma^* = O(\frac{1}{\mu r\sqrt{r}\kappa})$, where $\kappa = \sigma_1(\mathbf{L}^*)/\sigma_r(\mathbf{L}^*)$ is the condition number of $\mathbf{L}^*$. We also study the stability of Algorithm 1 in the following statement.

Theorem 5 (Stability, fully observed case) *Let* $\mathbf{L}$ *be the current value, and let* $\mathbf{L}^{+}$ *be the next update by applying Algorithm 1 to* $\mathbf{L}$ *for one iteration. Assuming* $\mathbf{Y} = \mathbf{L}^{*} + \mathbf{S}^{*} + \mathbf{N}^{*}$*, where* $\mathbf{N}^{*}$ *is a random Gaussian noise i.i.d. sampled from* $N(0, \sigma^2)$*,* $\gamma > 10\gamma^{*}$ *and* $(\gamma + 2\gamma^{*})\mu r < 1/64$*, then there exist* $C, a, c, \eta_0 > 0$ *such that when* $\eta < \eta_0$*,*

$$P\left(\|\mathbf{L}^{+} - \mathbf{L}^{*}\|_F \leq (1 - c\eta)\,\|\mathbf{L} - \mathbf{L}^{*}\|_F \text{ for all } \mathbf{L} \in \Gamma\right) \to 1, \quad \text{as } n_1, n_2 \to \infty, \tag{17}$$

where

$$\Gamma = \{\mathbf{L} \in \mathbb{R}^{n_1 \times n_2} : \text{rank}(\mathbf{L}) = r, C\sigma\sqrt{(n_1 + n_2) r \ln(n_1 n_2)} \leq \|\mathbf{L} - \mathbf{L}^{*}\|_F \leq a\sigma_r(\mathbf{L}^{*})\|,$$

Since $1 - c\eta < 1$, and Theorem 5 shows that when the observation $\mathbf{Y}$ is contaminated with a random Gaussian noise, if $\mathbf{L}^{(0)}$ is properly initialized such that $\|\mathbf{L}^{(0)} - \mathbf{L}^{*}\|_F < a\sigma_r(\mathbf{L}^{*})$, Algorithms 1 will converge to a neighborhood of $\mathbf{L}^{*}$ given by

$$\{\mathbf{L} : \|\mathbf{L} - \mathbf{L}^{*}\|_F \leq C\sigma\sqrt{(n_1 + n_2) r \ln(n_1 n_2)}\}$$

in $[-\log(\|\mathbf{L}^{(0)} - \mathbf{L}^{*}\|_F) + \log(C\sigma\sqrt{(n_1 + n_2) r \ln(n_1 n_2)})]/\log(1 - c\eta)$ iterations, with probability goes to 1 as $n_1, n_2 \to \infty$.

4.2. Analysis of Algorithm 2

For the partially observed setting, we assume that each entry of $\mathbf{Y} = \mathbf{L}^{*} + \mathbf{S}^{*}$ is observed with probability p. That is, for any $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$, $\Pr((i,j) \in \mathbf{\Phi}) = p$. Then we have the following statement on convergence:

Theorem 6 (Linear convergence rate, partially observed case) *There exists* $c > 0$ *such that for* $n = \max(n_1, n_2)$*, if* $p \geq \max(c\mu r \log(n)/\epsilon^2 \min(n_1, n_2), \frac{56}{3}\frac{\log n}{\gamma \min(n_1, n_2)})$*, then with probability* $1 - 2n^{-3} - 6n^{-1}$*,*

$$\frac{\|\mathbf{L}^{(k)} - \mathbf{L}^{*}\|_F}{\|\mathbf{L}^{(0)} - \mathbf{L}^{*}\|_F} \leq \Bigg[\sqrt{1 - p^2(1-\epsilon)^2\left(2\eta\left(1 - \tilde{C}_1 - \frac{ap(1+\epsilon)}{2(1-a)}(1 + \tilde{C}_1)\right) - \eta^2(1 + \tilde{C}_1)^2\right)} \\ + \frac{\eta^2 a^2 (p + p\epsilon)^2 (1 + \tilde{C}_1)^2}{1 - \eta a(p + p\epsilon)(1 + \tilde{C}_1)}\Bigg]^k \tag{18}$$

for

$$\tilde{C}_1 = \frac{1}{p(1-\epsilon)}\left[6(\gamma + 2\gamma^{*})p\mu r + 4\frac{3\gamma^{*}}{\gamma - 3\gamma^{*}}(\sqrt{p(1+\epsilon)} + \frac{a}{2})^2 + a^2\right].$$

Remark 7 *(Choice of parameters) Note that when* η *is small, the RHS of* (18) *is in the order of*

$$1 - p^2(1-\epsilon)^2\left(1 - \tilde{C}_1 - \frac{ap(1+\epsilon)}{2(1-a)}(1 + \tilde{C}_1)\right)\eta + O(\eta^2).$$

As a result, to make sure that the RHS of (18) *to be smaller than* 1 *for small* η*, we assume that*

$$1 - \tilde{C}_1 - \frac{ap(1+\epsilon)}{2(1-a)}(1 + \tilde{C}_1) > 0. \tag{19}$$

For example, when $ap(1+\epsilon) = 4(1-a)$*, it requires that* $\tilde{C}_1 < 1/3$*. If* (19) *holds, then there exists* $\eta_0 = \eta_0(\tilde{C}_1, p, \epsilon, a)$ *such that for all* $\eta \leq \eta_0$*, the RHS of* (18) *is smaller than* 1*.*

The practical choices of η *and* γ *will be discussed in Section 5.*

Remark 8 *(Simplified choice of parameters) There exists $\{c_i\}_{i=1}^4 > 0$ such that when $\epsilon < 1/2$, $a < c_1 p$, $\gamma^* \mu r < c_2$ and $\gamma = c_3 \gamma^*$, then when $\eta < 1/8$,*

$$\frac{\|\mathbf{L}^{(k)} - \mathbf{L}^*\|_F}{\|\mathbf{L}^{(0)} - \mathbf{L}^*\|_F} \leq (1 - c_4 \eta p^2)^k.$$

Compared with the result in Theorem 1, the addition parameter p appears in both the initialization requirement $a < c_1 p$ as well as the convergence rate. This makes the result weaker, but we suspect that the dependence on the subsampling ratio p could be improved through a better estimation in (39) and the estimation of $\tilde{C}_1$ in Lemma 16, and we leave it as a possible future direction.

We present a method of obtaining a proper initialization in Theorem 9. Combining it with Theorem 6, Algorithm 2 allows the corruption level γ^* to be in the order of $O(\frac{p}{\mu r \sqrt{r\kappa}})$.

Theorem 9 (Initialization, partially observed case) *There exists $c_1, c_2, c_3 > 0$ such that if $\gamma > 2\gamma^*$, and $p \geq c_2(\frac{\mu r^2}{\epsilon^2} + \frac{1}{\alpha}) \log n / \min(n_1, n_2)$, and we initialize $\mathbf{L}^{(0)}$ using the rank-r approximation to $F(\mathbf{Y})$, then*

$$\|\mathbf{L}^{(0)} - \mathbf{L}^*\|_F \leq 16\gamma\mu r \sigma_1(\mathbf{L}^*)\sqrt{2r} + 2\sqrt{2} c_1 \epsilon \sigma_1(\mathbf{L}^*)$$

with probability at least $1 - c_3 n^{-1}$, where $\sigma_1(\mathbf{L}^*)$ is the largest singular value of $\mathbf{L}^*$.

4.3. Comparison with Alternating Gradient Descent

Since our objective functions are equivalent to the objective functions of the alternating gradient descent (AGD) in Yi et al. (2016), it would be interesting to compare these two works. The only difference of these two works lies in the algorithmic implementation: our methods use the gradient descent on the manifold of low-rank matrices, while the methods in Yi et al. (2016) use alternating gradient descent on the factors of the low-rank matrix. In the following we compare the results of both works from four aspects:

1. **Accuracy of initialization.** What is the largest value t that the algorithm can tolerate, such that for any initialization $\mathbf{L}^{(0)}$ satisfying $\|\mathbf{L}^{(0)} - \mathbf{L}^*\|_F \leq t$, the algorithm is guaranteed to converge to $\mathbf{L}^*$?

2. **Convergence rate.** What is the smallest number of iteration steps k such that the algorithm reaches a given convergence criterion ϵ, i.e. $\|\mathbf{L}^{(k)} - \mathbf{L}^*\|_F / \|\mathbf{L}^{(0)} - \mathbf{L}^*\|_F < \epsilon$?

3. **Corruption level (perfect initialization).** Suppose that the initialization is in a sufficiently small neighborhood of $\mathbf{L}^*$ (i.e. there exists a very small $\epsilon_0 > 0$ such that $\mathbf{L}^{(0)}$ satisfies $\|\mathbf{L}^{(0)} - \mathbf{L}^*\|_F < \epsilon_0$), what is the maximum corruption level that can be tolerated in the convergence analysis?

4. **Corruption level (proper initialization).** Suppose that the initialization is given by the procedure in Theorem 4 (for the partially observed case) and 9 (for the fully observed case), what is the maximum corruption level that can be tolerated?

These comparisons are summarized in Table 1. We can see that under the full observed setting, our results remove or reduce the dependence on the condition number κ, while keeping other values unchanged. Under the partially observed setting our results still have the advantage of less dependence on κ, but sometimes require an additional dependence on the subsampling ratio p. The simulation results discussed in the next section also verify that when κ is large our algorithms have better performance, while that the slowing effect of p under the partially observed setting is not significant. As discussed after Theorem 6, we suspect that this dependence could be removed after a more careful analysis (or more assumptions).

Criterion	Accuracy of initialization	Convergence rate	Max corruption (perfect init.)	Max corruption (proper init.)
Algorithm 1	$O(\sigma_r(\mathbf{L}^*))$	$O(\log(\frac{1}{\epsilon}))$	$O(\frac{1}{\mu r})$	$O(\frac{1}{\mu r^{1.5}\kappa})$
APG (full)	$O(\frac{\sigma_r(\mathbf{L}^*)}{\sqrt{\kappa}})$	$O(\kappa\log(\frac{1}{\epsilon}))$	$O(\frac{1}{\kappa^2\mu r})$	$O(\frac{1}{\max(\mu r^{1.5}\kappa^{1.5},\kappa^2\mu r)})$
Algorithm 2	$O(\sqrt{p}\sigma_r(\mathbf{L}^*))$	$O(\log(\frac{1}{\epsilon})/p^2)$	$O(\frac{1}{\mu r})$	$O(\frac{\sqrt{p}}{\mu r^{1.5}\kappa})$
APG (partial)	$O(\frac{\sigma_r(\mathbf{L}^*)}{\kappa})$	$O(\kappa\mu r\log(\frac{1}{\epsilon})$	$O(\frac{1}{\kappa^2\mu r})$	$O(\frac{1}{\max(\mu r^{1.5}\kappa^{1.5},\kappa^2\mu r)})$

Table 1: Comparison of the theoretical guarantees in our work and the alternating gradient descent algorithm in Yi et al. (2016). The four criteria are explained in details in Section 4.3.

Here we use a simple example to give some intuition of why our proposed methods work better than gradient descent method based on factorization. Let us consider the following simple optimization problem:

$$\arg\min_{\mathbf{z}\in\mathbb{R}^m} f(\mathbf{z}), \text{ where } \mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y},$$

where $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{A},\mathbf{B} \in \mathbb{R}^{m\times n}$. In this example, $(\mathbf{x},\mathbf{y})$ can be considered as the "factors" of $\mathbf{z}$. The gradient descent method on the factors $(\mathbf{x},\mathbf{y})$ is then given by

$$\mathbf{x}^+ = \mathbf{x} - \eta\mathbf{A}^T f'(\mathbf{z}), \;\; \mathbf{y}^+ = \mathbf{y} - \eta\mathbf{B}^T f'(\mathbf{z}). \tag{20}$$

Writing the update formula (20) in terms of $\mathbf{z}$, it becomes

$$\mathbf{z}^+ = \mathbf{A}\mathbf{x}^+ + \mathbf{B}\mathbf{y}^+ = \mathbf{z} - \eta(\mathbf{A}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T)f'(\mathbf{z}).$$

As a result, the "gradient descent on factors $(\mathbf{x},\mathbf{y})$" has a direction of $-(\mathbf{A}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T)f'(\mathbf{z})$. In comparison, the gradient descent on the variable $\mathbf{z}$ has a direction of $-f'(\mathbf{z})$, which is the direction that f decreases fastest. If $\mathbf{A}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T$ is a matrix with a large condition number, then we expect that the gradient descent method on factors $(\mathbf{x},\mathbf{y})$ would not work well. This example shows that generally, compared to applying the gradient descent method to the factors of a variable, it is better to apply it to the variable itself. Similar to this example, our method applies gradient descent on the $\mathbf{L}$ itself while Yi et al. (2016) applies gradient descent to the factors of $\mathbf{L}$.

4.4. Comparison with other robust PCA algorithms

In this section we compare our result with other robust PCA methods and summarize them in Table 2. Some criterion in Table 1 are not included since they do not apply. For example, (Netrapalli et al., 2014, Alternating Projection) only analyzes the algorithm with specific initialization, and the criterion 1 and 3 in Table 1 do not apply to this work. As a result, we only compare the maximum corruption ratio that these methods can handle, and the computational complexity per iteration in Table 2. As for the convergence rate, it depends on assumptions on parameters such as the coherence parameter, rank, and the size of the matrix: The alternating projection Netrapalli et al. (2014) requires $10\log(4n_1\mu^2 r\|\mathbf{Y}-\mathbf{L}^{(0)}\|_2/\epsilon\sqrt{n_1 n_2})$ iterations to achieve an accuracy ϵ, under the assumptions that $\gamma^* < 1/512\mu r^2$ and a tuning parameter is chosen to be $4\mu^2 r\sqrt{n_1 n_2}$. The alternating minimization method Gu et al. (2016) have the guarantee that if $\|\mathbf{L}^{(1)}-\mathbf{L}^*\|_2 \leq \sigma_1(\mathbf{L}^*)$, then

$$\|\mathbf{L}^{(k+1)}-\mathbf{L}^*\|_F \leq \sigma_1\left(\frac{96\sqrt{2}\nu\mu\sqrt{r}(s^*/d)^{3/2}\kappa\sigma_1}{1-24\sqrt{2}\nu\mu\sqrt{r}(s^*/d)^{3/2}\kappa\sigma_1}\right)^k \|\mathbf{L}^{(1)}-\mathbf{L}^*\|_F,$$

where ν is a parameter concerning the coherence of $\mathbf{L}^*$, s^* is the number of nonzero entries in $\mathbf{S}^*$, $d=\min(n_1,n_2)$. As a result s^*/d is approximately $\gamma^*\max(n_1,n_2)$ in our notation. If ν is in the order of $O(1)$, then this results requires that $\mu\sqrt{r}\kappa\sigma_1\max(n_1,n_2)^{3/2}\gamma^{*\,3/2} \leq O(1)$, which is more restrictive than our assumption in Theorem 1 that $\gamma^*\mu r \leq O(1)$. Convex methods usually have convergence rate guarantees based on convexity, for example, the accelerated proximal gradient method Toh and Yun (2010) has a convergence rate of $O(1/k^2)$. While it is a slower convergence rate compared to the result in Theorem 1 in this work or the results in Netrapalli et al. (2014); Gu et al. (2016) and it does not necessarily converge to the correct solution, this result does not depend on any assumption on the low-rank matrix and the corruption ratio.

Criterion	Maximum corruption level	Complexity per iteration
Algorithm 1	$O(1/\kappa\mu r^{3/2})$	$O(rn_1n_2)$
Convex methods	$O(1/\mu^2 r)$	$O(n_1n_2\min(n_1,n_2))$
Netrapalli et al. (2014)	$1/512\mu^2 r$	$O(r^2n_1n_2)$
Gu et al. (2016)	$O(1/\mu^{2/3}r^{2/3}\min(n_1,n_2))$	$O(r^2n_1n_2)$
Yi et al. (2016)	$O(1/\kappa^2\mu r^{3/2})$	$O(rn_1n_2)$

Table 2: Comparison of the theoretical guarantees in our work and some other robust PCA algorithms.

The stability in Theorem 5 is comparable to analysis in Netrapalli et al. (2014) (the works Gu et al. (2016) and (Yi et al., 2016, Alternating Gradient Descent) do not have stability analysis). The work Netrapalli et al. (2014) assumes that $\|\mathbf{N}^*\|_\infty < \sigma_r(\mathbf{L}^*)/100n_2$ and proves that the output of their algorithm $\widehat{\mathbf{L}}$ satisfies

$$\|\widehat{\mathbf{L}}-\mathbf{L}^*\|_F \leq \epsilon + 2\mu^2 r\left(7\|\mathbf{N}^*\|_2 + \frac{8\sqrt{n_1n_2}}{\sqrt{r}}\|\mathbf{N}^*\|_\infty\right),$$

where ϵ is the error of the algorithm when there is no noise. If $\mathbf{N}^*$ is i.i.d. sampled from $N(0, \sigma^2)$, this result suggests that $\|\widehat{\mathbf{L}} - \mathbf{L}^*\|_F$ is bounded above by $\epsilon + O(\mu^2\sqrt{rn_1n_2}\sigma)$. In comparison, Theorem 5 suggests that after a few iterations, $\|\mathbf{L}^{(k)} - \mathbf{L}^*\|_F$ is bounded above by $C\sigma\sqrt{(n_1+n_2)r\ln(n_1n_2)}$ with high probability, which is a tighter upper bound.

5. Simulations

In this section, we test the performance of the proposed algorithms by simulated data sets and real data sets. The MATLAB implementation of our algorithm used in this section is available at `https://sciences.ucf.edu/math/tengz/`. For simulated data sets, we generate $\mathbf{L}^*$ by $\mathbf{U\Sigma V}^T$, where $\mathbf{U} \in \mathbb{R}^{n_1\times r}$ and $\mathbf{V} \in \mathbb{R}^{n_2\times r}$ are random matrices that i.i.d. sampled from $N(0,1)$, and $\mathbf{\Sigma} \in \mathbb{R}^{n_1\times r}$ is an diagonal matrix. As for $\mathbf{S}^* \in \mathbb{R}^{n_1\times n_2}$, each entry is sampled from $N(0,100)$ with probability q, and is zero with probability $1-q$. That is, q represents the level of sparsity in the sparse matrix $\mathbf{S}^*$. It measures the overall corruption level of $\mathbf{Y}$ and is associated with the corruption level γ^* (γ^* measures the row and column-wise corruption level). For the partially observed case, we assume that each entry of $\mathbf{Y}$ is observed with probability p.

5.1. Choice of parameters

We first investigate the performance of the proposed algorithms, in particular, the dependence on the parameters η and γ. In simulations, we let $[n_1, n_2] = [500, 600]$, $r = 3$, $\mathbf{\Sigma} = \mathbf{I}$, and $q = 0.02$. For the partially observed case, we let $p = 0.2$.

The first simulation investigates the following questions:

- Should we use the Algorithms 1 and 2 with Option 1 or Option 2?
- What is the appropriate choice of the step size η?

The simulation results for Option 1 an 2 with various step sizes are visualized in Figure 2, which show that the two options perform similarly. Usually the algorithms converge faster when the step size is larger. However, if the step size is too large then it might diverge. As a result, we use the step size $\eta = 0.7$ for Algorithm 1 and $0.7/p$ for Algorithm 2 for the following simulations.

The second simulation concerns the choice of γ. We test $\gamma = c\gamma^*$ for a few choices of c (γ^* can be calculated from the zero pattern of $\mathbf{S}$). Figure 5.1 shows that if γ is too small, for example, $0.5\gamma^*$, then the algorithm fail to converge to the correct solution; and if γ is too large, then the convergence is slow. Following these observations, we use $\gamma = 1.5\gamma^*$ as the default choice of the following experiments, which is also used in Yi et al. (2016).

5.2. Performance of the proposed algorithm

In this section, we analyze the convergence behavior as the parameters (overall ratio of corrupted entries q, condition number κ, rank r, subsampling ratio p) changes and visualized the result in Figure 4.

Figure 4(a) shows the simulation for corruptions level q, we use the setting in Section 5.1, but replace the corruption level q by $q = 0.1, 0.2, 0.3, 0.4$. Figure 4 shows that the algorithm

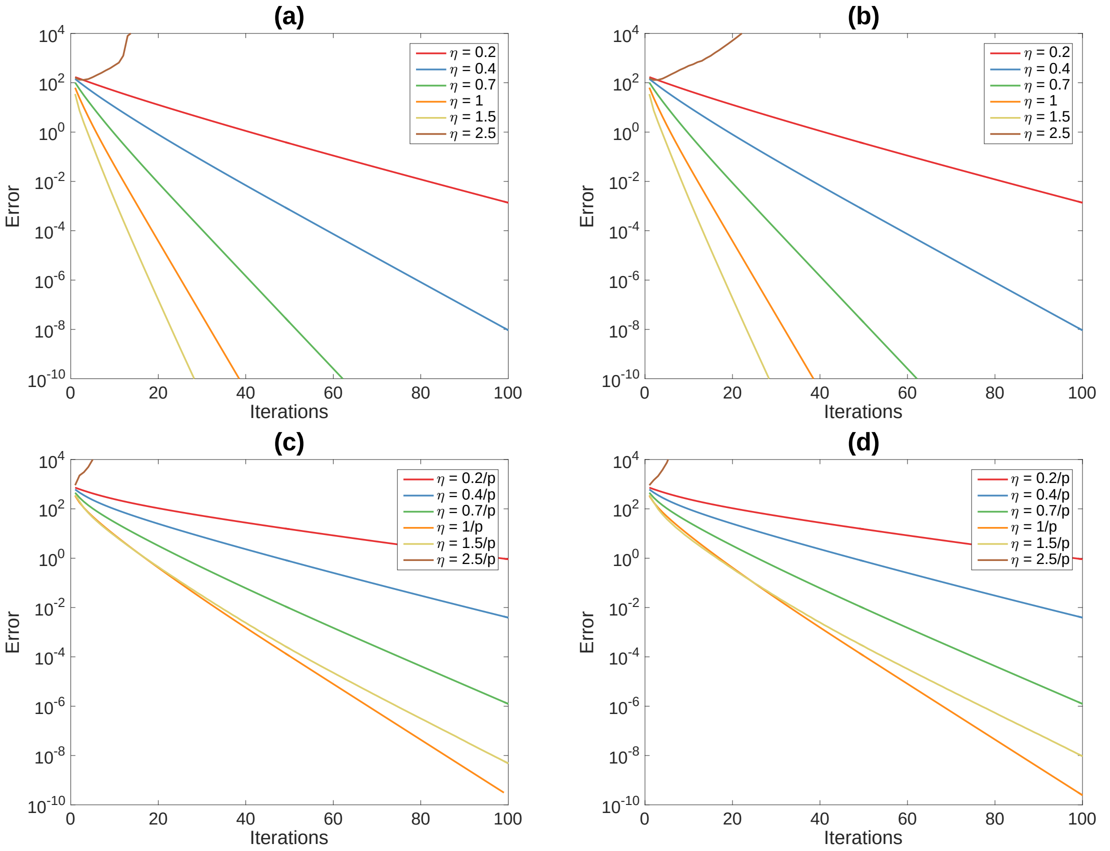

Figure 2: The dependence of the estimation error on the number of iterations for different step sizes η (a) Algorithm 1 (Option 1); (b) Algorithm 1 (Option 2); (c) Algorithm 2 (Option 1); (d) Algorithm 2 (Option 2).

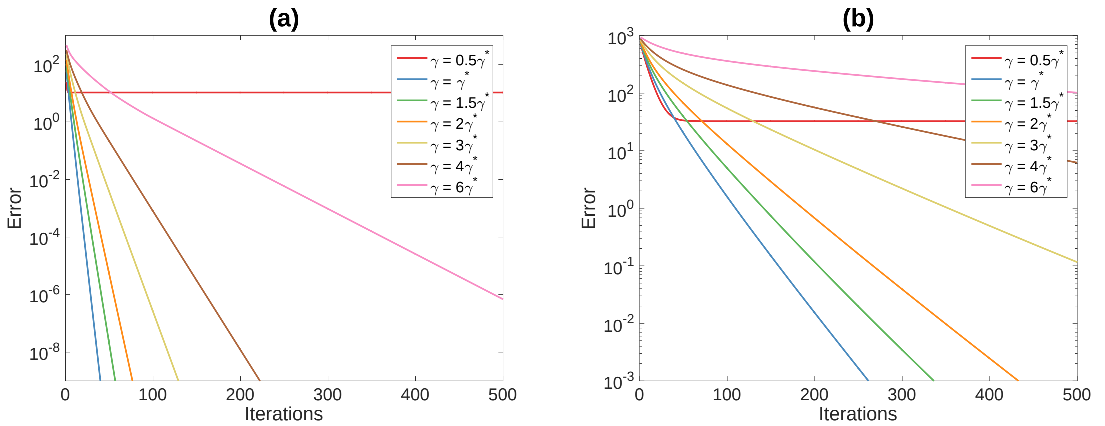

Figure 3: The convergence of the algorithm depending on the choice of γ. (a) fully observed setting; (b) partially observed setting.

converges slower with more corruption, which is expected since there is fewer information available. However, the algorithm still converges even with an overall corruption level at 0.4.

Figure 4(b) shows the simulation for rank r, we use the setting in Section 5.1, but replace r by $r = 3, 10, 30, 100, 300$ respectively. Simulations show that the algorithm works fine for rank $r = 3, 10, 30, 100$ and it converges slower for rank $r = 300$.

Figure 4(c) shows the simulation for condition number κ of $\mathbf{L}$, we use the setting in Section 5.1, but replace Σ by $\Sigma = \mathrm{diag}(1, 1, 1/\kappa)$ and try various values of κ. While the algorithm converges for κ up to 30 in the simulation, for larger κ the algorithm converges slowly at the beginning, and then decreases quickly to zero. We suspect that the initialization is not sufficiently good and it takes a while for the algorithm to reach the "local neighborhood of convergence". We also remark that $\mathbf{L}$ with a very large condition number, e.g. $\kappa = 100$, is generally challenging for any nonconvex optimization algorithm, as shown in Figure 5, Setting 4. It is because that when κ is large, the solution is close to a matrix with rank less than r – a singular point on the manifold of the matrices of rank r, which gives a geometry of manifold that is not "smooth" enough. We observe that when $\kappa = 100$, our algorithm performs well if the rank r is set to 2 (instead of the true value 3)—in fact, when $\kappa = 100$, the underlying matrix is approximately of rank 2 since the third singular value is very small.

We test the algorithm with various matrix sizes using the setting in Section 5.1 and set $[n_1, n_2] = [1000, 1200], [5000, 6000], [10000, 12000]$. Figure 4(d) shows that Algorithm 1 converges quickly for all of the choices within a few iterations.

In the last simulation, we test Algorithm 2 under the setting in Section 5.1 with various choices of the subsampling ratio p. Figure 4(e) shows suggest that the algorithm converges for p as small as 0.1, though the convergence rate is slow for small p.

5.3. Comparison with other robust PCA algorithms

In this section, we compare our algorithm with the accelerated proximal gradient method (APG) and the alternating direction method of multiplier (ADMM) based on convex relaxation (1); the robust matrix completion algorithm (RMC) Cambier and Absil (2016) based on manifold optimization problem

$$\operatorname*{arg\,min}_{\mathrm{rank}(\mathbf{L})=r} \sum_{(i,j)\in\Phi} \|\mathbf{L}_{ij} - \mathbf{Y}_{ij}\| + \lambda \sum_{(i,j)\notin\Phi} \mathbf{L}_{ij}^2,$$

as well as the alternating gradient descent method (AGD) in Yi et al. (2016) that solves the same optimization as this work, but with an implementation based on matrix factorization rather than manifold optimization. We use the implementation of APG from Toh and Yun (2010) and the implementation of ADMM from `https://github.com/dlaptev/RobustPCA`. In these two algorithms, we use the choice of parameter $\lambda = 1/\sqrt{\max(n_1, n_2)}$, which is the default choice in the implementation Toh and Yun (2010) and the theoretical analysis in Candès et al. (2011). For ADMM, the augmented Lagrangian parameter is set by default as 10λ. For RMC and GD, we use their default setting of parameters. Since the setting of Algorithm 2 does not apply to the implementations of APG/ADMM, we compare them under the fully observed setting. We compare them in the following four settings:

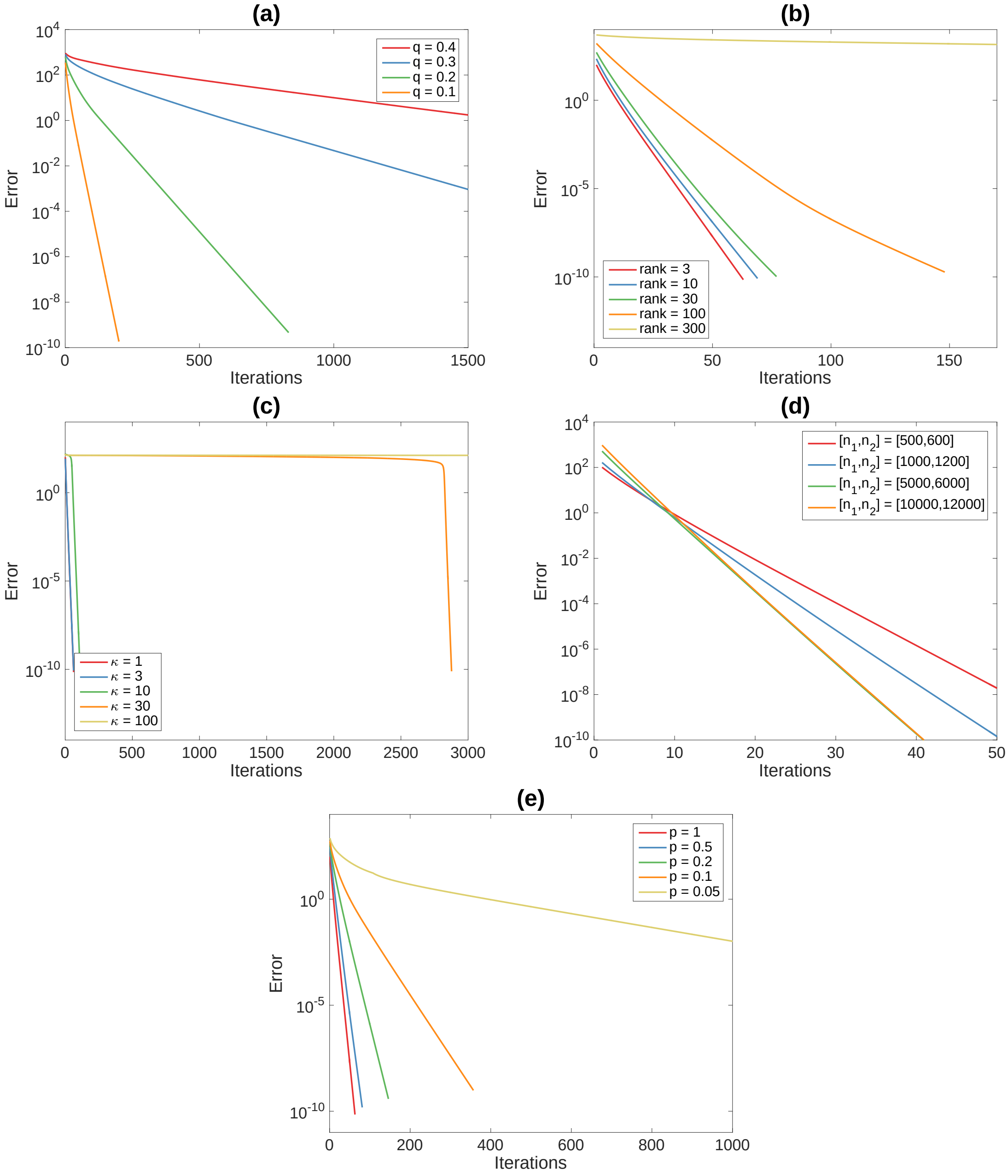

Figure 4: Dependence of the estimation error on the number of iterations for different (a) Overall ratios of corrupted entries q (Algorithm 1); (b) Ranks r (Algorithm 1); (c) Condition numbers κ (Algorithm 1); (d) Matrix sizes $[n_1, n_2]$ (Algorithm 1); (e) Subsampling ratio p (Algorithm 2).

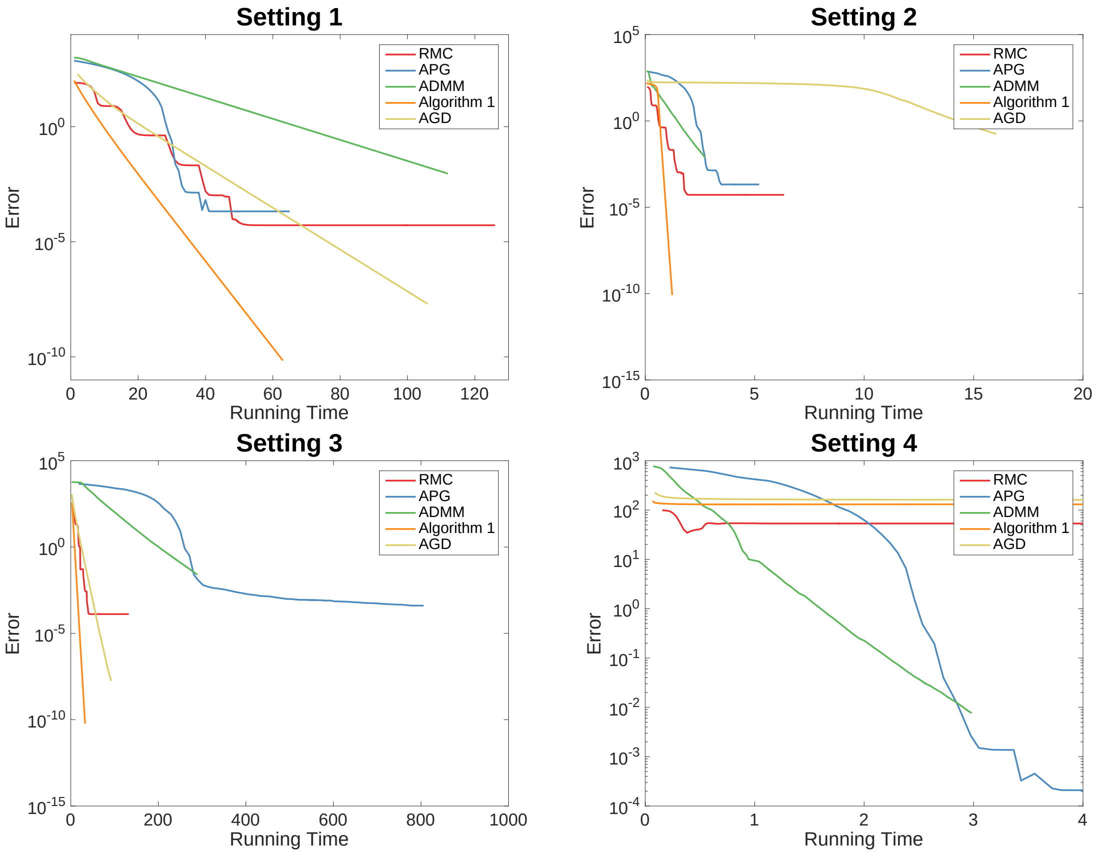

Figure 5: The comparison of the performance of the algorithms under the fully observed setting. The running time is measured in seconds.

- Setting 1: same setting as in Section 5.1.
- Setting 2 (large condition number): replace Σ by $\mathrm{diag}(1, 1, 0.1)$ in Setting 1.
- Setting 3 (large matrix): replace n_1 and n_2 by 3000 and 4000 in Setting 1.
- Setting 4 (large condition number): replace Σ by $\mathrm{diag}(1, 1, 0.01)$ in Setting 1.

Figure 5 shows that under Setting 1, 2 and 3, Algorithm 1 converges faster than other algorithms. In particular, the advantage over the AGD algorithm is very clear under Setting 2, where the condition number is larger. This verifies our theoretical analysis, where the convergence rate is faster than the analysis in Yi et al. (2016) by a factor of $\sqrt{\kappa}$. In Setting 3, the algorithms RMC, AGD and Algorithm 1 converge much faster than APG and ADMM, which verifies the computational advantage of nonconvex algorithms when the matrix size is large. However, in Setting 4, the convex algorithms converge to the correct solution while the nonconvex algorithms converge to a local minimizer that is different than the correct solution. This is due to the fact that the nonconvex algorithms have more than one minimizer, and if it is not initialized well then it could get trapped in local minimizers. In

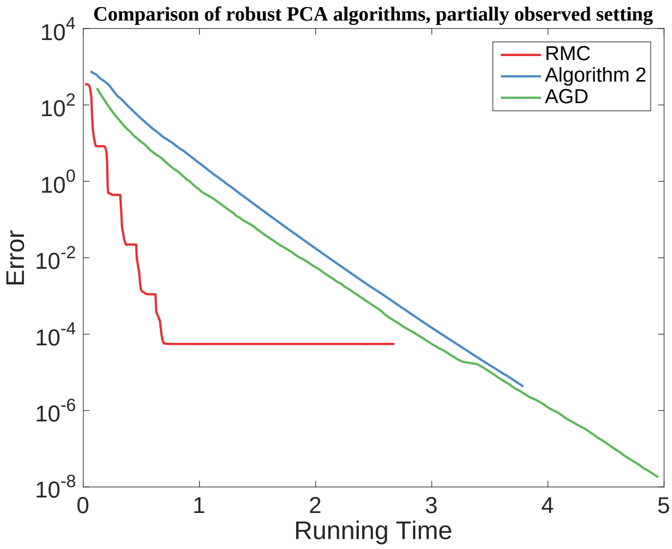

Figure 6: The comparison of the performances of the algorithms under the partially observed setting. The running time is measured in seconds.

practice, we observed that if the initialization is well chosen and close to the true $\mathbf{L}^*$, then Algorithm 1 converges quickly to the correct solution.

We also compare the performance of RMC, AGD and Algorithm 2 under the partially observed setting. We use Setting 1 with $p = 0.3$ and visualize the result in Figure 6. The results are similar to that of the fully observed setting: AGD and Algorithm 2 are comparable and RMC converges faster at the beginning, but then does not achieve higher accuracy, possibly due to their choice of the regularization parameter.

We also test the proposed algorithms in a real data application for video background subtraction. We adopt the public data set `Shoppingmall` studied in Yi et al. (2016),[1] A few frames are visualized in the first column of Figure 7. There are 1000 frames in this video sequence, represented by a matrix of size 81920 × 1000, where each column corresponds to a frame of the video and each row corresponds to a pixel of the video. We apply our algorithms with $r = 3$ and $\gamma^* = 0.1$, $p = 0.5$ for the partially observed case, the step size $\eta = 0.7$. We stop the algorithm after 100 iterations. Figure 7 shows that our algorithms obtain desirable low-rank approximations within 100 iterations.

In Figure 8, we compare our algorithms with APG in terms of the convergence of the objective function value. In this figure, the relative error is defined as $\|F(\mathbf{L}-\mathbf{Y})\|_F/\|\mathbf{Y}\|_F$, a scaled objective value. A smaller relative error implies a better low-rank approximation. Figure 8 shows out that our algorithms can obtain smaller objective values within 100 iterations under both fully observed and partially observed cases.

[1] The data set is originally from `http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html`, and is available at `https://sciences.ucf.edu/math/tengz/`.

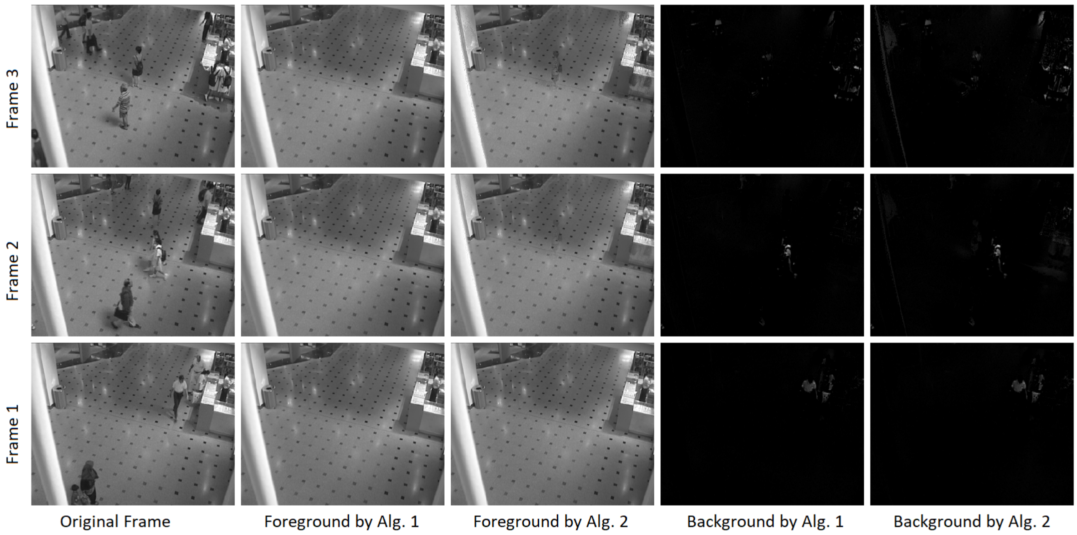

Figure 7: The performance of Algorithms 1 and 2 in video background subtraction, with three rows representing three frames in the video sequence. For Algorithm 2, a subsampling ratio of $p = 0.5$ is used.

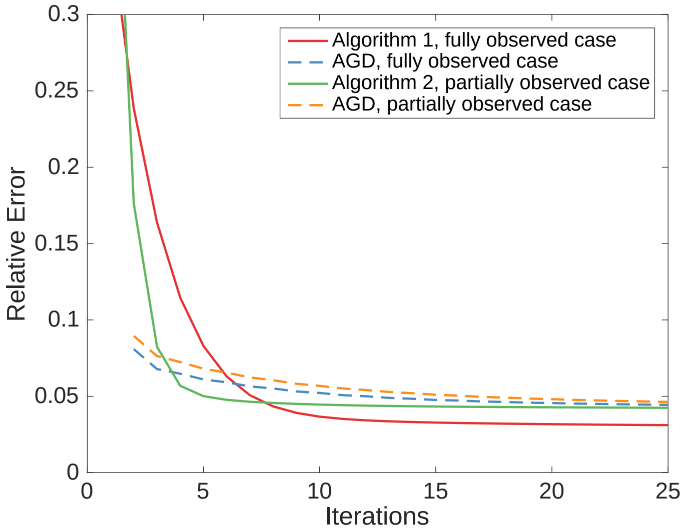

Figure 8: The relative error of Algorithms 1, 2, and AGD with respect to the iterations, for both fully observed case and partially observed case in the experiment of background subtraction.

6. Conclusion

This paper proposes two robust PCA algorithms (one for fully observed case and one for partially observed case) based on the gradient descent algorithm on the manifold of low-rank matrices. Theoretically, compared with the gradient descent algorithm with matrix factorization, our approach has a faster convergence rate, better tolerance of the initialization accuracy and corruption level. The approach removes or reduces the dependence of the algorithms on the condition number of the underlying low-rank matrix. Numerically, the proposed algorithms performance is less sensitive to the choice of step sizes. We also find that under the partially observed setting, the performance of the proposed algorithm is not significantly affected by the presence of the additional dependence on the observation probability. Considering the popularity of the methods based on matrix factorization, it is an interesting future direction to apply manifold optimization to other low-rank matrix estimation problems.

Acknowledgements

The authors thank the editor, an associate editor and referee for their helpful comments and suggestions. The authors also thank David Fleischer for providing helps on the coding. Yang's research is partially supported by NSERC RGPIN-2016-05174. Zhang's research is partially supported by National Science Foundation (NSF) grant CNS-1739736.

Appendix for “Robust PCA by Manifold Optimization"

A. Technical Derivations in Section 3

Verification of (7). Formula (7) can be verified as follows. Let $\langle\cdot\rangle_F$ be the Frobenius inner product of two matrices, then

$$\begin{aligned}&\langle \mathbf{D}-P_{T_{\mathbf{X}}\mathcal{M}}(\mathbf{D}),\mathbf{A}\mathbf{V}\mathbf{V}^T\rangle_F=\langle(\mathbf{I}-\mathbf{U}\mathbf{U}^T)\mathbf{D}(\mathbf{I}-\mathbf{V}\mathbf{V}^T),\mathbf{A}\mathbf{V}\mathbf{V}^T\rangle_F\\=&\langle(\mathbf{I}-\mathbf{U}\mathbf{U}^T)\mathbf{D}(\mathbf{I}-\mathbf{V}\mathbf{V}^T)\mathbf{V}\mathbf{V}^T,\mathbf{A}\rangle_F=\langle\mathbf{0},\mathbf{A}\rangle_F=0\end{aligned}$$

and similarly $\langle\mathbf{D}-P_{T_{\mathbf{X}}\mathcal{M}}(\mathbf{D}),\mathbf{U}\mathbf{U}^T\mathbf{B}\rangle_F=0$. As a result, $\langle\mathbf{D}-P_{T_{\mathbf{X}}\mathcal{M}}(\mathbf{D}),\mathbf{A}\mathbf{V}\mathbf{V}^T+\mathbf{U}\mathbf{U}^T\mathbf{B}\rangle_F=0$ for all $\mathbf{A}\in\mathbb{R}^{n_1\times n_1}$ and $\mathbf{B}\in\mathbb{R}^{n_2\times n_2}$, which verifies formula (7) by showing that $\mathbf{D}-P_{T_{\mathbf{X}}\mathcal{M}}(\mathbf{D})$ is orthogonal to $T_{\mathbf{X}}\mathcal{M}$.

Verification of (10). It is clear that $R^{(2)}_{\mathbf{X}}(\delta)$ defined in (10) has rank r; and to show that $\langle R^{(2)}_{\mathbf{X}}(\delta)-(\mathbf{X}+\delta),\mathbf{Z}\rangle_F=0$ for all $\mathbf{Z}\in T_{\mathbf{X}}\mathcal{M}$, we first write this property as $[R^{(2)}_{\mathbf{X}}(\delta)-(\mathbf{X}+\delta)]\perp T_{\mathbf{X}}\mathcal{M}$ for simplicity, and since $T_{\mathbf{X}}\mathcal{M}=\{\mathbf{A}\mathbf{V}\mathbf{V}^T+\mathbf{U}\mathbf{U}^T\mathbf{B}:\text{for }\mathbf{A}\in\mathbb{R}^{n_1\times n_1},\mathbf{B}\in\mathbb{R}^{n_2\times n_2}\}$, we just need to show that $\langle R^{(2)}_{\mathbf{X}}(\delta)-(\mathbf{X}+\delta),\mathbf{A}\mathbf{V}\mathbf{V}^T+\mathbf{U}\mathbf{U}^T\mathbf{B}\rangle_F=0$ for all $\mathbf{A}\in\mathbb{R}^{n_1\times n_1}$ and $\mathbf{B}\in\mathbb{R}^{n_2\times n_2}$. This is easy to verify, because we have $R^{(2)}_{\mathbf{X}}(\delta)\mathbf{V}=(\mathbf{X}+\delta)\mathbf{V}$,

$$\langle R^{(2)}_{\mathbf{X}}(\delta)-(\mathbf{X}+\delta),\mathbf{A}\mathbf{V}\mathbf{V}^T\rangle_F=\langle(R^{(2)}_{\mathbf{X}}(\delta)\mathbf{V}-(\mathbf{X}+\delta)\mathbf{V})\mathbf{V}^T,\mathbf{A}\rangle_F=\langle\mathbf{0},\mathbf{A}\rangle_F=0,\tag{21}$$

Similarly, we can easily verify that $\mathbf{U}^TR^{(2)}_{\mathbf{X}}(\delta)=\mathbf{U}^T(\mathbf{X}+\delta)$, we have $\langle R^{(2)}_{\mathbf{X}}(\delta)-(\mathbf{X}+\delta),\mathbf{U}\mathbf{U}^T\mathbf{B}\rangle_F=0$, and therefore $[R^{(2)}_{\mathbf{X}}(\delta)-(\mathbf{X}+\delta)]\perp T_{\mathbf{X}}\mathcal{M}$. As a result, there exists a unique $R^{(2)}_{\mathbf{X}}$ such that $\text{rank}(R^{(2)}_{\mathbf{X}})=r$ and $[R^{(2)}_{\mathbf{X}}(\delta)-(\mathbf{X}+\delta)]\perp T_{\mathbf{X}}\mathcal{M}$.

Verification of (12). We first define the operator $S:\mathbb{R}^{n_1\times n_2}\to\mathbb{R}^{n_1\times n_2}$ such that $F(\mathbf{A})=\mathbf{S}(\mathbf{A})\circ\mathbf{A}$ ($\circ$ represents the elementwise product), i.e.,

$$\mathbf{S}(\mathbf{A})=\begin{cases}0, & \text{if }|\mathbf{A}_{ij}|>|\mathbf{A}_{i,\cdot}|^{[\gamma]}\text{ and }|\mathbf{A}_{ij}|>|\mathbf{A}_{\cdot,j}|^{[\gamma]},\\ 1, & \text{otherwise.}\end{cases}$$

Then if the absolute values of all entries of $\mathbf{A}$ are different, the sparsity pattern does not change under a small perturbation, i.e., $\mathbf{S}(\mathbf{A})=\mathbf{S}(\mathbf{A}+\mathbf{\Delta})$. Then by definition of $f(\cdot)$,

$$\begin{aligned}&f(\mathbf{L}+\mathbf{\Delta})-f(\mathbf{L})=\frac{1}{2}\|\mathbf{S}(\mathbf{L}-\mathbf{Y}+\mathbf{\Delta})\circ(\mathbf{L}-\mathbf{Y}+\mathbf{\Delta})\|_F^2-\frac{1}{2}\|\mathbf{S}(\mathbf{L}-\mathbf{Y})\circ(\mathbf{L}-\mathbf{Y})\|_F^2\\=&\frac{1}{2}\|\mathbf{S}(\mathbf{L}-\mathbf{Y})\circ(\mathbf{L}-\mathbf{Y}+\mathbf{\Delta})\|_F^2-\frac{1}{2}\|\mathbf{S}(\mathbf{L}-\mathbf{Y})\circ(\mathbf{L}-\mathbf{Y})\|_F^2\\=&\langle\mathbf{S}(\mathbf{L}-\mathbf{Y})\circ(\mathbf{L}-\mathbf{Y}),\mathbf{\Delta}\rangle_F+O(\|\mathbf{\Delta}\|_F^2),\end{aligned}$$

where $\circ$ represents the Hadamard product, i.e., the elementwise product between matrices.

Verification of (14). It is sufficient to prove the case where $\mathbf{U}^{(k)}$ and $\mathbf{V}^{(k)}$ are given by the SVD decomposition $\mathbf{L}^{(k)}=\mathbf{U}^{(k)}\mathbf{\Sigma}^{(k)}\mathbf{V}^{(k)T}$. Denote $\mathbf{D}=\nabla f(\mathbf{L}^{(k)})=F(\mathbf{L}^{(k)}-\mathbf{Y})$. Set $\mathbf{X}=\mathbf{L}^{(k)}$ and $\delta=-\eta P_{T_{\mathbf{L}^{(k)}}\mathcal{M}}(\mathbf{D})$ in (10), we have

$$\begin{aligned}\mathbf{L}^{(k+1)}\ :=\ &(\mathbf{L}^{(k)}-\eta P_{T_{\mathbf{L}^{(k)}}\mathcal{M}}(\mathbf{D}))\mathbf{V}^{(k)}[\mathbf{U}^{(k)T}(\mathbf{L}^{(k)}-\eta P_{T_{\mathbf{L}^{(k)}}\mathcal{M}}(\mathbf{D}))\mathbf{V}^{(k)}]^{-1}\\&\mathbf{U}^{(k)T}(\mathbf{L}^{(k)}-\eta P_{T_{\mathbf{L}^{(k)}}\mathcal{M}}(\mathbf{D}))\end{aligned}\tag{22}$$

On the other hand, from (7) we have the projection

$$P_{T_{\mathbf{L}^{(k)}}\mathcal{M}}(\mathbf{D}) = \mathbf{U}^{(k)}\mathbf{U}^{(k)T}\mathbf{D} + \mathbf{D}\mathbf{V}^{(k)}\mathbf{V}^{(k)T} - \mathbf{U}^{(k)}\mathbf{U}^{(k)T}\mathbf{D}\mathbf{V}^{(k)}\mathbf{V}^{(k)T}.$$

As a result

$$\begin{aligned}&P_{T_{\mathbf{L}^{(k)}}\mathcal{M}}(\mathbf{D})\mathbf{V}^{(k)} = [\mathbf{U}^{(k)}\mathbf{U}^{(k)T}\mathbf{D} + \mathbf{D}\mathbf{V}^{(k)}\mathbf{V}^{(k)T} - \mathbf{U}^{(k)}\mathbf{U}^{(k)T}\mathbf{D}\mathbf{V}^{(k)}\mathbf{V}^{(k)T}]\mathbf{V}^{(k)}\\ &= \mathbf{U}^{(k)}\mathbf{U}^{(k)T}\mathbf{D}\mathbf{V}^{(k)} + \mathbf{D}\mathbf{V}^{(k)}\mathbf{V}^{(k)T}\mathbf{V}^{(k)} - \mathbf{U}^{(k)}\mathbf{U}^{(k)T}\mathbf{D}\mathbf{V}^{(k)}\mathbf{V}^{(k)T}\mathbf{V}^{(k)} = \mathbf{D}\mathbf{V}^{(k)} \qquad (23)\end{aligned}$$

and similarly,

$$\mathbf{U}^{(k)T}P_{T_{\mathbf{L}^{(k)}}\mathcal{M}}(\mathbf{D}) = \mathbf{U}^{(k)T}\mathbf{D}. \qquad (24)$$

Combining (23), (24) with (22), the update formula (14) is verified.

B. Proof of Theorem 1

In this proof, we will investigate $\|\mathbf{L}^+ - \mathbf{L}^*\|_F$, where

$$\mathbf{L}^+ = R_{\mathbf{L}}(-\eta P_{T_{\mathbf{L}}}F(\mathbf{L}-\mathbf{Y})).$$

It is sufficient to prove that when $\|\mathbf{L}-\mathbf{L}^*\| \leq a\sigma_r(\mathbf{L}^*)$ with the value a satisfying the conditions in Theorem 1, then

$$\|\mathbf{L}^+ - \mathbf{L}^*\|_F \leq \Big(1 - \frac{1-2C_1}{8}\eta\Big)\|\mathbf{L}-\mathbf{L}^*\|_F. \qquad (25)$$

To prove (25), we first introduce three auxiliary lemmas.

Lemma 10 **(a)** *Let* $\mathbf{D} = \mathbf{L}-\mathbf{L}^* - F(\mathbf{L}-\mathbf{Y}) = \mathbf{L}-\mathbf{L}^* - F(\mathbf{L}-\mathbf{L}^*-\mathbf{S}^*)$*, then*

$$\|\mathbf{D}\|_F^2 \leq C_1^2\|\mathbf{L}-\mathbf{L}^*\|_F^2. \qquad (26)$$

(b) *For the noisy setting where* $\mathbf{Y} = \mathbf{L}^* + \mathbf{S}^* + \mathbf{N}^*$*, and* $\mathbf{D}' = \mathbf{L}-\mathbf{L}^*-\mathbf{N}^*-F(\mathbf{L}-\mathbf{Y})$*, we have*

$$\|\mathbf{D}'\|_F^2 \leq 2C_1^2\|\mathbf{L}-\mathbf{L}^*\|_F^2 + 2(\gamma+5\gamma^*)N_1, \qquad (27)$$

where $N_1 = n_2\sum_{i=1}^{n_1}|\mathbf{N}^*_{i,\cdot}|^{\max} + n_1\sum_{j=1}^{n_2}|\mathbf{N}^*_{\cdot,j}|^{\max}$.

Lemma 11 *If* $\|\mathbf{L}-\mathbf{L}^*\|_F \leq a\sigma_r(\mathbf{L}^*)$ *and* $a \leq 1$*, then*

$$\|(\mathbf{L}-\mathbf{L}^*) - \mathbf{P}_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\|_F \leq \frac{a}{2(1-a)}\|\mathbf{L}-\mathbf{L}^*\|_F, \qquad (28)$$

$$\|(\mathbf{L}-\mathbf{L}^*) - \mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)\|_F \leq \frac{a}{2}\|\mathbf{L}-\mathbf{L}^*\|_F. \qquad (29)$$

Lemma 12 *For* $\mathbf{X} \in T_{\mathbf{L}}\mathcal{M}$*, then*

$$\|R^{(i)}_{\mathbf{L}}(\mathbf{X}) - (\mathbf{L}+\mathbf{X})\|_F \leq \frac{\|\mathbf{X}\|_F^2}{2(\sigma_r(\mathbf{L}) - \|\mathbf{X}\|)}, \textit{ for either } i = 1 \textit{ or } 2.$$

To prove (25), first we note that

$$\begin{aligned}
&\|\mathbf{L}-\mathbf{L}^*\|_F^2-\|\mathbf{L}-\eta P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})-\mathbf{L}^*\|_F^2\\
=&\|\mathbf{L}-\mathbf{L}^*\|_F^2-\|\mathbf{L}-\mathbf{L}^*\|_F^2+2\eta\langle\mathbf{L}-\mathbf{L}^*,P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\rangle_F-\|\eta P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\|_F^2\\
=&2\eta\langle\mathbf{L}-\mathbf{L}^*,P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\rangle_F-\|\eta P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\|_F^2\\
=&2\eta\langle\mathbf{L}-\mathbf{L}^*,P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*)-P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*-F(\mathbf{L}-\mathbf{Y}))\rangle_F-\eta^2\|P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\|_F^2\\
=&2\eta\langle P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*),P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*)-P_{T_\mathbf{L}}\mathbf{D}\rangle_F-\eta^2\|P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\|_F^2\\
\geq&2\eta(\|P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*)\|_F^2-\|\mathbf{D}\|_F\|P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*)\|_F)-\eta^2(\|\mathbf{L}-\mathbf{L}^*\|_F+\|\mathbf{D}\|_F)^2. \qquad (30)
\end{aligned}$$

The fourth line is obtained by $P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*-F(\mathbf{L}-\mathbf{Y}))=\mathbf{L}-\mathbf{L}^*-P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\rangle_F$. The fifth line is because $\mathbf{L}-\mathbf{L}^*=P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*)+P_{T_\mathbf{L}}^{\perp}(\mathbf{L}-\mathbf{L}^*)$. The last line uses Cauchy-Schwarz inequality $\langle P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*),P_{T_\mathbf{L}}\mathbf{D}\rangle_F\leq\|\mathbf{D}\|_F\|P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*)\|_F$ and triangular inequality $\|P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\|_F\leq\|\mathbf{L}-\mathbf{L}^*\|_F+\|P_{T_\mathbf{L}}(\mathbf{D})\|_F\leq\|\mathbf{L}-\mathbf{L}^*\|_F+\|\mathbf{D}\|_F$. Lemma 11 and the assumptions $\|\mathbf{L}-\mathbf{L}^*\|_F\leq a\sigma_r(\mathbf{L}^*)$ and $\sqrt{1-(\frac{a}{2(1-a)})^2}>\frac{1}{2}$ imply

$$\|P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*)\|_F\geq\frac{1}{2}\|\mathbf{L}-\mathbf{L}^*\|_F. \qquad (31)$$

Combining it with the estimation of $\|\mathbf{D}\|_F$ in Lemma 10, we have

$$\begin{aligned}
&\|\mathbf{L}-\mathbf{L}^*\|_F^2-\|\mathbf{L}-\eta P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})-\mathbf{L}^*\|_F^2\\
\geq&\ \eta(\frac{1}{2}-C_1)\|\mathbf{L}-\mathbf{L}^*\|_F^2-\eta^2(1+C_1)^2\|\mathbf{L}-\mathbf{L}^*\|_F^2. \qquad (32)
\end{aligned}$$

When ths RHS of (32) is positive (i.e., when $(1-2C_1)\geq2\eta(1+C_1)^2$), (32) implies $\|\mathbf{L}-\mathbf{L}^*\|_F>\|\mathbf{L}-\eta P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})-\mathbf{L}^*\|_F$ and

$$\begin{aligned}
&\|\mathbf{L}-\mathbf{L}^*\|_F-\|\mathbf{L}-\eta P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})-\mathbf{L}^*\|_F\\
\geq&\frac{\eta(\frac{1}{2}-C_1)\|\mathbf{L}-\mathbf{L}^*\|_F^2-\eta^2(1+C_1)^2\|\mathbf{L}-\mathbf{L}^*\|_F^2}{\|\mathbf{L}-\mathbf{L}^*\|_F+\|\mathbf{L}-\eta P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})-\mathbf{L}^*\|_F}\\
\geq&\frac{1}{2}\left(\eta(\frac{1}{2}-C_1)-\eta^2(1+C_1)^2\right)\|\mathbf{L}-\mathbf{L}^*\|_F. \qquad (33)
\end{aligned}$$

In addition,

$$\|P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\|_F\leq\|F(\mathbf{L}-\mathbf{Y})\|_F=\|\mathbf{L}-\mathbf{L}^*\|_F+\|\mathbf{D}\|_F\leq(1+C_1)\|\mathbf{L}-\mathbf{L}^*\|_F \qquad (34)$$

and Lemma 12 give

$$\begin{aligned}
&\|\mathbf{L}^+-\mathbf{L}^*\|_F-\|\mathbf{L}-\eta P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})-\mathbf{L}^*\|_F\leq\|\mathbf{L}-\eta P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})-\mathbf{L}^+\|_F\\
\leq&\frac{\eta^2\|P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\|_F^2}{\sigma_r(\mathbf{L}^*)-\eta\|P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\|_F}\leq\frac{\eta^2a^2(1+C_1)^2}{1-\eta a(1+C_1)}\|\mathbf{L}-\mathbf{L}^*\|_F. \qquad (35)
\end{aligned}$$

Combining (33) and (35),

$$\frac{\|\mathbf{L}-\mathbf{L}^*\|_F-\|\mathbf{L}^+-\mathbf{L}^*\|_F}{\|\mathbf{L}-\mathbf{L}^*\|_F}\geq\frac{1}{4}\eta(1-2C_1)-\eta^2(1+C_1)^2\left[\frac{1}{2}+\frac{a^2}{1-\eta(1+C_1)a}\right].$$

Therefore, Theorem 1 is proved when $C_1<1/2$, and η_0 is chosen such that

$$\eta_0(1+C_1)^2\left[\frac{1}{2}+\frac{a^2}{1-\eta_0(1+C_1)a}\right]\leq\frac{1}{8}(1-2C_1).$$

C. Proof of Theorem 5

The proof of the noisy case also follows similarly from the proofs of Theorem 1 and 6. Note that

$$F(\mathbf{L}-\mathbf{Y})=\mathbf{L}-\mathbf{L}^*-\mathbf{N}^*-\mathbf{D}',$$

and define $\mathbf{Q}=P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*)$, then following the proof of Theorem 1 and applying Lemma 10 (b), we have

$$\begin{aligned}&\|\mathbf{L}-\mathbf{L}^*\|_F^2-\|\mathbf{L}-\eta P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})-\mathbf{L}^*\|_F^2\\ =&2\eta\langle\mathbf{L}-\mathbf{L}^*,P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\rangle_F+O(\eta^2)=2\eta\langle P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*),P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})\rangle_F+O(\eta^2)\\ =&2\eta\langle P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*),P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*-\mathbf{N}^*-\mathbf{D}')\rangle_F+O(\eta^2)\\ \geq&2\eta\left(\|\mathbf{Q}\|_F^2-\langle\mathbf{N}^*,\mathbf{Q}\rangle_F-\|\mathbf{Q}\|_F\sqrt{2C_1^2\|\mathbf{L}-\mathbf{L}^*\|_F^2+2(\gamma+5\gamma^*)N_1}\right)+O(\eta^2).\end{aligned}$$

In addition, (35) gives

$$\big|\|\mathbf{L}^+-\mathbf{L}\|_F-\|\mathbf{L}-\eta P_{T_\mathbf{L}}F(\mathbf{L}-\mathbf{Y})-\mathbf{L}^*\|_F\big|=O(\eta^2).$$

Combining it with the estimation of C_1, N_1, and $\langle\mathbf{N}^*,\mathbf{Q}\rangle_F$ in Lemma 13 and the fact that $(1-\frac{a}{2(1-a)})\|\mathbf{L}-\mathbf{L}^*\|_F\leq\|\mathbf{Q}\|_F\leq(1+\frac{a}{2(1-a)})\|\mathbf{L}-\mathbf{L}^*\|_F$ (which follows from Lemma 11), the Theorem is proved.

Lemma 13 *If $\mathbf{N}^*\in\mathbb{R}^{n_1\times n_2}$ is elementwisely i.i.d. sampled from $N(0,\sigma^2)$, then*
(a) with probability $1-\frac{4}{n_1^7n_2^7}$, $\sum_{i=1}^{n_1}(|\mathbf{N}^*_{i,\cdot}|^{\max})^2\leq16\sigma^2n_1\ln(n_1n_2)$, and $\sum_{j=1}^{n_2}(|\mathbf{N}^*_{\cdot,j}|^{\max})^2\leq16\sigma^2n_2\ln(n_1n_2)$, and as a result, $N_1\leq32\sigma^2n_1n_2\ln(n_1n_2)$.
(b) There exists $C_6>0$ such that as $n_1+n_2\to\infty$, the probability that

$$\langle\mathbf{N}^*,P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*)\rangle_F\leq\frac{1}{4}\|P_{T_\mathbf{L}}(\mathbf{L}-\mathbf{L}^*)\|_F^2\tag{36}$$

holds for all $\{\mathbf{L}:C_6\sigma\sqrt{(n_1+n_2)r\ln(n_1n_2)}\leq\|\mathbf{L}-\mathbf{L}^*\|_F\leq a\sigma_r(\mathbf{L}^*)\}$ converges to 1.

D. Proof of Theorem 6

This proof borrows two lemmas from (Yi et al., 2016, Lemmas 9, 10) as follows.

Lemma 14 *(Yi et al., 2016, Lemma 9) There exists $c>0$ such that for all $0<\epsilon<1$, if $p\geq c\mu r\log(n)/\epsilon^2\min(n_1,n_2)$, then with probability at least $1-2n^{-3}$, for all $\mathbf{X}$ in the tangent plane $T_{\mathbf{L}^*}$, i.e., all $\mathbf{X}$ that can be written as $\mathbf{L}^*\mathbf{A}+\mathbf{B}\mathbf{L}^*$, where $\mathbf{A}\in\mathbb{R}^{n_2\times n_2}$ and $\mathbf{B}\in\mathbb{R}^{n_1\times n_1}$,*

$$(1-\epsilon)\|\mathbf{X}\|_F^2\leq\frac{1}{p}\|P_\mathbf{\Phi}\mathbf{X}\|_F^2\leq(1+\epsilon)\|\mathbf{X}\|_F^2.$$

Lemma 15 *(Yi et al., 2016, Lemma 10) If $p\geq\frac{56}{3}\frac{\log n}{\gamma\min(n_1,n_2)}$, the with probability at least $1-6n^{-1}$, the number of entries in $\mathbf{\Phi}$ per row is in the interval $[pn_2/2,3pn_2/2]$, and the number of entries in $\mathbf{\Phi}$ per column is in $[pn_1/2,3pn_1/2]$.*

Then we introduce the following lemma parallel to Lemma 10:

Lemma 16 *When the events in Lemmas 14 and 15 hold, for* $\tilde{\mathbf{D}} = P_{\mathbf{\Phi}}[\mathbf{L}-\mathbf{L}^* - \tilde{F}(\mathbf{L}-\mathbf{Y})]$ *we have*

$$\|\tilde{\mathbf{D}}\|_F^2 \leq \tilde{C}_1^2 \|\mathbf{L}-\mathbf{L}^*\|_F^2, \tag{37}$$

with

$$\tilde{C}_1 = \frac{1}{p(1-\epsilon)}\Big[6(\gamma+2\gamma^*)p\mu r + 4\frac{3\gamma^*}{\gamma-3\gamma^*}(\sqrt{p(1+\epsilon)}+\frac{a}{2})^2 + a^2\Big].$$

The proof of Theorem 6 is parallel to the proof of Theorem 1, with $\mathbf{L}^+$ defined slightly differently by

$$\mathbf{L}^+ = R_{\mathbf{L}}(-\eta P_{T_{\mathbf{L}}}\tilde{F}(\mathbf{L}-\mathbf{Y})).$$

Defining $P_{\mathbf{\Phi}}: \mathbb{R}^{n_1\times n_2} \rightarrow \mathbb{R}^{n_1\times n_2}$ by

$$[P_{\mathbf{\Phi}}\mathbf{X}]_{ij} = \begin{cases}\mathbf{X}_{ij}, \text{ if } (i,j)\in\mathbf{\Phi},\\ 0, \text{ if } (i,j)\notin\mathbf{\Phi}.\end{cases}$$

Then $\tilde{F}(\mathbf{L}-\mathbf{Y}) = P_{\mathbf{\Phi}}\tilde{F}(\mathbf{L}-\mathbf{Y})$. Following a similar analysis as (30),

$$\begin{aligned}&\|\mathbf{L}-\mathbf{L}^*\|_F^2 - \|\mathbf{L}-\eta P_{T_{\mathbf{L}}}P_{\mathbf{\Phi}}\tilde{F}(\mathbf{L}-\mathbf{Y})-\mathbf{L}^*\|_F^2\\ =&2\eta\langle\mathbf{L}-\mathbf{L}^*, P_{T_{\mathbf{L}}}P_{\mathbf{\Phi}}\tilde{F}(\mathbf{L}-\mathbf{Y})\rangle_F - \|\eta P_{T_{\mathbf{L}}}P_{\mathbf{\Phi}}\tilde{F}(\mathbf{L}-\mathbf{Y})\|_F^2\\ \geq&2\eta\langle P_{\mathbf{\Phi}}P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*), P_{\mathbf{\Phi}}\tilde{F}(\mathbf{L}-\mathbf{Y})\rangle_F - \|\eta P_{\mathbf{\Phi}}\tilde{F}(\mathbf{L}-\mathbf{Y})\|_F^2\\ \geq&2\eta\langle P_{\mathbf{\Phi}}(\mathbf{L}-\mathbf{L}^*) - P_{\mathbf{\Phi}}P_{T_{\mathbf{L}}}^{\perp}(\mathbf{L}-\mathbf{L}^*), P_{\mathbf{\Phi}}(\mathbf{L}-\mathbf{L}^*)-\tilde{\mathbf{D}}\rangle_F - \eta^2(\|P_{\mathbf{\Phi}}(\mathbf{L}-\mathbf{L}^*)\|_F + \|\tilde{\mathbf{D}}\|_F)^2,\end{aligned} \tag{38}$$

here $P_{T_{\mathbf{L}}}^{\perp}$ represents the projector to the subspace orthogonal to $T_{\mathbf{L}}$. Lemma 11 and Lemma 14 imply

$$\frac{\|P_{\mathbf{\Phi}}P_{T_{\mathbf{L}}}^{\perp}(\mathbf{L}-\mathbf{L}^*)\|_F}{\|P_{\mathbf{\Phi}}(\mathbf{L}-\mathbf{L}^*)\|_F} \leq \frac{\|P_{T_{\mathbf{L}}}^{\perp}(\mathbf{L}-\mathbf{L}^*)\|_F}{\|P_{\mathbf{\Phi}}(\mathbf{L}-\mathbf{L}^*)\|_F} \leq \frac{ap(1+\epsilon)}{2(1-a)}, \tag{39}$$

and combining it with the estimation of $\tilde{\mathbf{D}}$ in Lemma 16, the RHS of (38) is larger than

$$\|P_{\mathbf{\Phi}}(\mathbf{L}-\mathbf{L}^*)\|_F^2\left(2\eta\left(1-\tilde{C}_1-\frac{ap(1+\epsilon)}{2(1-a)}(1+\tilde{C}_1)\right)-\eta^2(1+\tilde{C}_1)^2\right). \tag{40}$$

In addition, Lemma 14 implies

$$\begin{aligned}\|P_{\mathbf{\Phi}}\tilde{F}(\mathbf{L}-\mathbf{Y})\|_F &\leq \|P_{\mathbf{\Phi}}(\mathbf{L}-\mathbf{L}^*)\|_F + \|P_{\mathbf{\Phi}}\tilde{\mathbf{D}}\|_F\\ &\leq (1+\tilde{C}_1)\|P_{\mathbf{\Phi}}(\mathbf{L}-\mathbf{L}^*)\|_F,\\ &\leq (1+\tilde{C}_1)p(1+\epsilon)\|\mathbf{L}-\mathbf{L}^*\|\end{aligned}$$

and combining it with Lemma 12,

$$\|\mathbf{L}^+-\mathbf{L}^*\|_F - \|\mathbf{L}-\eta P_{T_{\mathbf{L}}}P_{\mathbf{\Phi}}\tilde{F}(\mathbf{L}-\mathbf{Y})-\mathbf{L}^*\|_F \leq \frac{\eta^2a^2(p+p\epsilon)^2(1+\tilde{C}_1)^2}{1-\eta a(p+p\epsilon)(1+\tilde{C}_1)}\|\mathbf{L}-\mathbf{L}^*\|_F.$$

Combining it with (40) and Lemma 11, we have

$$\frac{\|\mathbf{L}^+ - \mathbf{L}^*\|_F}{\|\mathbf{L} - \mathbf{L}^*\|_F} \leq \sqrt{1 - p^2(1-\epsilon)^2 \left(2\eta\left(1 - \tilde{C}_1 - \frac{ap(1+\epsilon)}{2(1-a)}(1+\tilde{C}_1)\right) - \eta^2(1+\tilde{C}_1)^2\right)} + \frac{\eta^2 a^2 (p+p\epsilon)^2 (1+\tilde{C}_1)^2}{1 - \eta a(p+p\epsilon)(1+\tilde{C}_1)},$$

and Theorem 6 is proved.

E. Proof of Lemmas

Lemma 10(a) **Proof** By the definition of F, for any matrix $\mathbf{A}$, $\mathbf{A} - F(\mathbf{A})$ is a sparse matrix, therefore

$$\mathbf{D} = \mathbf{L} - \mathbf{L}^* - \mathbf{S}^* - F(\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*) + \mathbf{S}^*$$

is a sparse matrix. Denote the locations of the nonzero entries of $\mathbf{D}$ by $\mathcal{S}$, and divide it into two sets $\mathcal{S}_1 \cup \mathcal{S}_2$ defined as follows:

$$\mathcal{S}_1 = \{(i,j) : |[\mathbf{L}-\mathbf{L}^*-\mathbf{S}^*]_{ij}| > |[\mathbf{L}-\mathbf{L}^*-\mathbf{S}^*]_{i,\cdot}|^{[\gamma]} \text{ and } |[\mathbf{L}-\mathbf{L}^*-\mathbf{S}^*]_{ij}| > |[\mathbf{L}-\mathbf{L}^*-\mathbf{S}^*]_{\cdot,j}|^{[\gamma]}\},$$

and

$$\mathcal{S}_2 = \{(i,j) \notin \mathcal{S}_1 : \mathbf{D}_{ij} = [\mathbf{L} - \mathbf{L}^*]_{ij} - F(\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*)_{ij} \neq 0\}.$$

For $(i,j) \in \mathcal{S}_1$, $[F(\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*)]_{ij} = 0$. As a result, $\mathbf{D}_{ij} = [\mathbf{L} - \mathbf{L}^*]_{ij}$. In addition, by definition of $F(\cdot)$, each row or column of $\mathbf{D}$ has at most γ percentage of points in $\mathcal{S}_1$.

For $(i,j) \in \mathcal{S}_2$, since $[F(\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*)]_{ij} = [\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*]_{ij}$, we have $\mathbf{D}_{ij} = \mathbf{S}^*_{ij} \neq 0$. By Assumption 1, therefore, for each row or column of $\mathbf{D}$, at most γ^* percentage of points lie in $\mathcal{S}_2$.

Combine the results

$$|[\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*]_{i,\cdot}|^{[\gamma]} \leq \{|[\mathbf{L} - \mathbf{L}^*]_{i,\cdot}| + |[\mathbf{S}^*]_{i,\cdot}|\}^{[\gamma]} \leq |[\mathbf{L} - \mathbf{L}^*]_{i,\cdot}|^{[\gamma-\gamma^*]},$$

$$|[\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*]_{j,\cdot}|^{[\gamma]} \leq \{|[\mathbf{L} - \mathbf{L}^*]_{j,\cdot}| + |[\mathbf{S}^*]_{j,\cdot}|\}^{[\gamma]} \leq |[\mathbf{L} - \mathbf{L}^*]_{j,\cdot}|^{[\gamma-\gamma^*]},$$

with $[F(\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*)]_{ij} = [\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*]_{ij}$, we have for $(i,j) \in \mathcal{S}_2$

$$\begin{aligned}
|\mathbf{D}_{ij}| =& |[\mathbf{L} - \mathbf{L}^* - F(\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*)]_{ij}| \\
\leq& |[\mathbf{L} - \mathbf{L}^*]_{ij}| + |F(\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*)]_{ij}| \\
\leq& |[\mathbf{L} - \mathbf{L}^*]_{ij}| + \max(|[\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*]_{i,\cdot}|^{[\gamma]}, |[\mathbf{L} - \mathbf{L}^* - \mathbf{S}^*]_{\cdot,j}|^{[\gamma]}) \\
\leq& |[\mathbf{L} - \mathbf{L}^*]_{ij}| + \max(|[\mathbf{L} - \mathbf{L}^*]_{i,\cdot}|^{[\gamma-\gamma^*]}, |[\mathbf{L} - \mathbf{L}^*]_{\cdot,j}|^{[\gamma-\gamma^*]}).
\end{aligned}$$

Applying the estimations above, and repeatedly use the fact that $(x+y)^2 \leq 2x^2 + 2y^2$, we have

$$\begin{aligned}
\|\mathbf{D}\|_F^2 = &\sum_{(i,j)\in\mathcal{S}} \mathbf{D}_{ij}^2 = \sum_{(i,j)\in\mathcal{S}_1} \mathbf{D}_{ij}^2 + \sum_{(i,j)\in\mathcal{S}_2} \mathbf{D}_{ij}^2 \leq \sum_{(i,j)\in\mathcal{S}_1} [\mathbf{L}-\mathbf{L}^*]_{ij}^2 \\
&+ \sum_{(i,j)\in\mathcal{S}_2} \Big\{ |[\mathbf{L}-\mathbf{L}^*]_{ij}| + \max(|[\mathbf{L}-\mathbf{L}^*]_{i,\cdot}|^{[\gamma-\gamma^*]}, |[\mathbf{L}-\mathbf{L}^*]_{\cdot,j}|^{[\gamma-\gamma^*]}) \Big\}^2 \\
\leq &\sum_{(i,j)\in\mathcal{S}_1} [\mathbf{L}-\mathbf{L}^*]_{ij}^2 + 2\sum_{(i,j)\in\mathcal{S}_2} [\mathbf{L}-\mathbf{L}^*]_{ij}^2 + 2\sum_{(i,j)\in\mathcal{S}_2} \max\{|[\mathbf{L}-\mathbf{L}^*]_{i,\cdot}|^{[\gamma-\gamma^*]}, |[\mathbf{L}-\mathbf{L}^*]_{\cdot,j}|^{[\gamma-\gamma^*]}\}^2 \\
\leq &\sum_{(i,j)\in\mathcal{S}_1} [\mathbf{L}-\mathbf{L}^*]_{ij}^2 + 2\sum_{(i,j)\in\mathcal{S}_2} [\mathbf{L}-\mathbf{L}^*]_{ij}^2 + 2\sum_{(i,j)\in\mathcal{S}_2} \{|[\mathbf{L}-\mathbf{L}^*]_{i,\cdot}|^{[\gamma-\gamma^*]} + |[\mathbf{L}-\mathbf{L}^*]_{\cdot,j}|^{[\gamma-\gamma^*]}\}^2 \\
\leq &\sum_{(i,j)\in\mathcal{S}_1} [\mathbf{L}-\mathbf{L}^*]_{ij}^2 + 2\sum_{(i,j)\in\mathcal{S}_2} [\mathbf{L}-\mathbf{L}^*]_{ij}^2 + 4\sum_{(i,j)\in\mathcal{S}_2} \{|[\mathbf{L}-\mathbf{L}^*]_{i,\cdot}|^{[\gamma-\gamma^*]}\}^2 + \{|[\mathbf{L}-\mathbf{L}^*]_{\cdot,j}|^{[\gamma-\gamma^*]}\}^2 \\
\leq &\sum_{(i,j)\in\mathcal{S}} [\mathbf{L}-\mathbf{L}^*]_{ij}^2 + \sum_{(i,j)\in\mathcal{S}_2} [\mathbf{L}-\mathbf{L}^*]_{ij}^2 + 4\frac{\gamma^*}{\gamma-\gamma^*}\|\mathbf{L}-\mathbf{L}^*\|_F^2 \\
\leq &2\sum_{(i,j)\in\mathcal{S}} [\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)]_{ij}^2 + 2\sum_{(i,j)\in\mathcal{S}_2} [\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)]_{ij}^2 + 4\frac{\gamma^*}{\gamma-\gamma^*}\|\mathbf{L}-\mathbf{L}^*\|_F^2 \\
&+ 2\sum_{(i,j)\in\mathcal{S}} [\mathbf{L}-\mathbf{L}^*-\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)]_{ij}^2 + 2\sum_{(i,j)\in\mathcal{S}_2} [\mathbf{L}-\mathbf{L}^*-\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)]_{ij}^2 \\
\leq &2\sum_{(i,j)\in\mathcal{S}} [\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)]_{ij}^2 + 2\sum_{(i,j)\in\mathcal{S}_2} [\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)]_{ij}^2 + 4\frac{\gamma^*}{\gamma-\gamma^*}\|\mathbf{L}-\mathbf{L}^*\|_F^2 \\
&+ 4\|\mathbf{L}-\mathbf{L}^*-\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)\|_F^2. \qquad (41)
\end{aligned}$$

Note that from line 5 to line 6, we used the fact that for $\mathbf{x}\in\mathbb{R}^n$, and $k \leq n$

$$k(x_{(k)})^2 \leq (x_{(k)})^2 + (x_{(k+1)})^2 + \cdots + (x_{(n-1)}) + (x_{(n)}) \leq (x_{(1)}) + \cdots + (x_{(n)}) = \|x\|_F^2,$$

where $x_{(k)}$ is the k-th order statistics of $x_1, \ldots, x_n$, i.e. the k-th smallest value. This gives us

$$(\gamma-\gamma^*)n_2|[\mathbf{L}-\mathbf{L}^*]_{i,\cdot}|^{[\gamma-\gamma^*]} \leq \|[\mathbf{L}-\mathbf{L}^*]_{i,\cdot}\|_2^2; \quad (\gamma-\gamma^*)n_2|[\mathbf{L}-\mathbf{L}^*]_{\cdot,j}|^{[\gamma-\gamma^*]} \leq \|[\mathbf{L}-\mathbf{L}^*]_{\cdot,j}\|_2^2.$$

Therefore

$$\sum_{(i,j)\in\mathcal{S}_2} |[\mathbf{L}-\mathbf{L}^*]_{i,\cdot}|^{[\gamma-\gamma^*]} \leq \frac{\gamma^* n_2}{(\gamma-\gamma^*)n_2}\|[\mathbf{L}-\mathbf{L}^*]_{i,\cdot}\|_F^2; \qquad (42)$$

$$\sum_{(i,j)\in\mathcal{S}_2} |[\mathbf{L}-\mathbf{L}^*]_{\cdot,j}|^{[\gamma-\gamma^*]} \leq \frac{\gamma^* n_1}{(\gamma-\gamma^*)n_1}\|[\mathbf{L}-\mathbf{L}^*]_{\cdot,j}\|_F^2. \qquad (43)$$

The values $\gamma^* n_2$ and $\gamma^* n_1$ in the numerator of the right hand sides in 42 and 43 are due to the fact that, in each row or column of $\mathbf{D}$, at most γ^* percentage of points lie in $\mathcal{S}_2$.

On the other hand, Lemma 11 implies

$$\|\mathbf{L}-\mathbf{L}^*-\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)\|_F \leq \frac{a}{2}\|\mathbf{L}-\mathbf{L}^*\|_F. \qquad (44)$$

In addition, using the fact that there exists $\mathbf{A} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{B} \in \mathbb{R}^{n_2 \times r}$, such that $\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*) = \mathbf{A}\mathbf{V}^T + \mathbf{U}\mathbf{B}^T$ and $\|\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)\|_F^2 = \|\mathbf{A}\mathbf{V}^T\|_F^2 + \|\mathbf{U}\mathbf{B}^T\|_F^2$, and that for each row or column, at most $\gamma+\gamma^*$ percentage of points lie in $\mathcal{S}$, we have

$$\begin{aligned}
&\sum_{(i,j)\in\mathcal{S}} [\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)]_{ij}^2 \le 2 \sum_{(i,j)\in\mathcal{S}} [\|(\mathbf{A}\mathbf{V}^T)_{ij}\|^2 + \|(\mathbf{U}\mathbf{B}^T)_{ij}\|^2] \\
\le& 2(\gamma+\gamma^*)\mu r \sum_{1\le i\le n_1, 1\le j\le n_2} [\|(\mathbf{A}\mathbf{V}^T)_{ij}\|^2 + \|(\mathbf{U}\mathbf{B}^T)_{ij}\|^2] = 2(\gamma+\gamma^*)\mu r \|\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)\|_F^2 \\
\le& 2(\gamma+\gamma^*)\mu r \|\mathbf{L}-\mathbf{L}^*\|_F^2.
\end{aligned} \tag{45}$$

Similarly, $\|\mathbf{A}\|_{2,\infty} = \max_{\|\mathbf{z}\|_2=1}\|\mathbf{A}\mathbf{z}\|_\infty$

$$\sum_{(i,j)\in\mathcal{S}_2} [\mathbf{P}_{T_{\mathbf{L}^*}}(\mathbf{L}-\mathbf{L}^*)]_{ij}^2 \le 2\gamma^*\mu r \|\mathbf{L}-\mathbf{L}^*\|_F^2, \tag{46}$$

Combining (41)-(46), (26) is proved. ■

Lemma 10(b) Proof Let $\mathbf{L}' = \mathbf{L} - \mathbf{N}^*$, then applying the fact that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$|[\mathbf{x}+\mathbf{y}]|^{[\gamma]} \le |[\mathbf{x}]|^{[\gamma]} + |[\mathbf{x}]|^{\max},$$

where $|[\mathbf{x}]|^{\max}$ represents the largest value of $|[\mathbf{x}]|$. We have

$$\begin{aligned}
\|\mathbf{D}'\|_F^2 \le& \sum_{(i,j)\in\mathcal{S}} [\mathbf{L}'-\mathbf{L}^*]_{ij}^2 + \sum_{(i,j)\in\mathcal{S}_2} [\mathbf{L}'-\mathbf{L}^*]_{ij}^2 \\
&+ 2\sum_{(i,j)\in\mathcal{S}_2} \{(|[\mathbf{L}'-\mathbf{L}^*]_{i,\cdot}|^{[\gamma-\gamma^*]})^2 + (|[\mathbf{L}'-\mathbf{L}^*]_{\cdot,j}|^{[\gamma-\gamma^*]})^2\} \\
\le& 2\left(\sum_{(i,j)\in\mathcal{S}} [\mathbf{L}-\mathbf{L}^*]_{ij}^2 + |\mathbf{N}_{ij}^*|^2 + \sum_{(i,j)\in\mathcal{S}_2} [\mathbf{L}-\mathbf{L}^*]_{ij}^2 + |\mathbf{N}_{ij}^*|^2\right) \\
&+ 4\sum_{(i,j)\in\mathcal{S}_2} \{(|[\mathbf{L}'-\mathbf{L}^*]_{i,\cdot}|^{[\gamma-\gamma^*]})^2 + (|\mathbf{N}_{i,\cdot}^*|^{\max})^2 + (|[\mathbf{L}'-\mathbf{L}^*]_{\cdot,j}|^{[\gamma-\gamma^*]})^2 + (|\mathbf{N}_{\cdot,j}^*|^{\max})^2\} \\
\le& 2C_1^2\|\mathbf{L}-\mathbf{L}^*\|_F^2 + 2(\gamma+5\gamma^*)N_1,
\end{aligned}$$

where the last inequality follows from the proof of part (a) and the definition of N_1. ■

Lemma 11 Proof Let the SVD decomposition of $\mathbf{L}^*$ be $\mathbf{L}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}$, $\mathbf{U}^\perp$ and $\mathbf{V}^\perp$ be orthogonal matrices of sizes $\mathbb{R}^{n_1\times(n_1-r)}$ and $\mathbb{R}^{n_2\times(n_2-r)}$ such that $\mathrm{Col}(\mathbf{U}^\perp) \perp \mathrm{Col}(\mathbf{U})$ and $\mathrm{Col}(\mathbf{V}^\perp) \perp \mathrm{Col}(\mathbf{V})$ (here $\mathrm{Col}(\mathbf{U})$ represents the subspace spanned by the columns of $\mathbf{U}$). Let

$$\mathbf{L}_{(1,1)}^* \equiv \mathbf{U}^T\mathbf{L}^*\mathbf{V}, \qquad \mathbf{L}_{(1,2)}^* \equiv \mathbf{U}^T\mathbf{L}^*\mathbf{V}^\perp,$$

$$\mathbf{L}_{(2,1)}^* \equiv \mathbf{U}^{\perp T}\mathbf{L}^*\mathbf{V}, \qquad \mathbf{L}_{(2,2)}^* \equiv \mathbf{U}^{\perp T}\mathbf{L}^*\mathbf{V}^\perp.$$

Since $\mathrm{rank}(\mathbf{L}^*) = r$, we have

$$\mathbf{L}^*_{(2,2)} = \mathbf{L}^*_{(2,1)}{\mathbf{L}^*_{(1,1)}}^{-1}\mathbf{L}^*_{(1,2)}.$$

Since all singular values of $\mathbf{L}^*_{(1,1)}$ are larger than $(1-a)\sigma_r(\mathbf{L}^*)$, if the singular value decomposition of ${\mathbf{L}^*_{(1,1)}}^{-1}$ is given by

$${\mathbf{L}^*_{(1,1)}}^{-1} = \mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^T,$$

then the $\|\mathbf{\Sigma}_0\| \leq 1/(1-a)\sigma_r(\mathbf{L}^*)$. Applying

$$\|\mathbf{A}\mathbf{B}\|_F^2 \leq \|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2$$

and the fact that for a square, diagonal matrix Σ, $|[\mathbf{X}\mathbf{\Sigma}]_{ij}| = |\mathbf{X}_{ij}\mathbf{\Sigma}_{jj}| \leq \|\mathbf{\Sigma}\||\mathbf{X}_{ij}|$, we have

$$\begin{aligned}
\|\mathbf{L}^*_{(2,2)}\|_F &= \|\mathbf{L}^*_{(2,1)}\mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^T\mathbf{L}^*_{(1,2)}\|_F \\
&\leq \|\mathbf{L}^*_{(2,1)}\mathbf{U}_0\mathbf{\Sigma}_0\|_F\|\mathbf{V}_0^T\mathbf{L}^*_{(1,2)}\|_F \\
&\leq \frac{1}{(1-a)\sigma_r(\mathbf{L}^*)}\|\mathbf{L}^*_{(2,1)}\mathbf{U}_0\|_F\|\mathbf{V}_0^T\mathbf{L}^*_{(1,2)}\|_F \\
&\leq \frac{1}{(1-a)\sigma_r(\mathbf{L}^*)}\|\mathbf{L}^*_{(2,1)}\|_F\|\mathbf{L}^*_{(1,2)}\|_F \\
&\leq \frac{1}{(1-a)\sigma_r(\mathbf{L}^*)}\Big(\frac{\|\mathbf{L}^*_{(2,1)}\|_F^2 + \|\mathbf{L}^*_{(1,2)}\|_F^2}{2}\Big) \\
&\leq \frac{1}{(1-a)\sigma_r(\mathbf{L}^*)}\Big(\frac{a^2\sigma_r(\mathbf{L}^*)^2}{2}\Big) \\
&\leq \frac{a^2}{2(1-a)}\sigma_r(\mathbf{L}^*),
\end{aligned} \tag{47}$$

and (28) is proved. The proof of (29) is similar. ■

Lemma 12 **Proof** Let the SVD decomposition of $\mathbf{L}$ be $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}$, and

$$\begin{aligned}
\mathbf{L}_{(1,1)} = \mathbf{U}^T(\mathbf{X}+\mathbf{L})\mathbf{V}, \mathbf{L}_{(1,2)} &= \mathbf{U}^T(\mathbf{X}+\mathbf{L})\mathbf{V}^\perp \\
&= \mathbf{U}^T\mathbf{X}\mathbf{V}^\perp, \mathbf{L}_{(2,1)} = \mathbf{U}^{\perp T}(\mathbf{X}+\mathbf{L})\mathbf{V} = \mathbf{U}^{\perp T}\mathbf{X}\mathbf{V},
\end{aligned}$$

then it is clear that

$$R^{(2)}_{\mathbf{L}}(\mathbf{X}) = \mathbf{L} + \mathbf{X} + \mathbf{U}^\perp\mathbf{L}_{(2,1)}{\mathbf{L}_{(1,1)}}^{-1}\mathbf{L}_{(1,2)}\mathbf{V}^{\perp T},$$

and using the same argument as in (47),

$$\begin{aligned}
\|\mathbf{L}_{(2,1)}{\mathbf{L}_{(1,1)}}^{-1}\mathbf{L}_{(1,2)}\|_F \leq& \frac{1}{\sigma_r(\mathbf{L}^*_{(1,1)})}\|\mathbf{L}_{(1,2)}\|_F\|\mathbf{L}_{(2,1)}\|_F \\
\leq& \frac{1}{\sigma_r(\mathbf{L}) - \|\mathbf{X}\|}\Big(\frac{\|\mathbf{L}_{(2,1)}\|_F^2 + \|\mathbf{L}_{(1,2)}\|_F^2}{2}\Big) \\
\leq& \frac{1}{\sigma_r(\mathbf{L}) - \|\mathbf{X}\|}\frac{\|\mathbf{X}\|_F^2}{2}.
\end{aligned}$$

So Lemma 12 is proved for $R^{(2)}_{\mathbf{L}}(\mathbf{X})$.

By definition, $R^{(1)}_{\mathbf{L}}(\mathbf{X})$ is the closest matrix to $\mathbf{L}+\mathbf{X}$ that has rank r, so $\|R^{(1)}_{\mathbf{L}}(\mathbf{X}) - (\mathbf{L}+\mathbf{X})\|_F \leq R^{(2)}_{\mathbf{L}}(\mathbf{X}) - (\mathbf{L}+\mathbf{X})\|_F$ and Lemma 12 is also proved for $R^{(1)}_{\mathbf{L}}(\mathbf{X})$. ■

Lemma 13 Proof WLOG, we assume $\sigma = 1$ and the generic cases can be proved similarly.
(a) It follows from the estimation of the distribution of the maximum of n_1 i.i.d. Gaussian variables $\{g_i\}_{i=1}^{n_1}$:

$$\begin{aligned}&\Pr\{\max_{1\leq i\leq n_1} |g_i| \leq 4\sqrt{\ln(n_1n_2)}\} \geq \Big(1 - 2\exp\Big(-\frac{(4\sqrt{\ln(n_1n_2)})^2}{2}\Big)\Big)^{n_1}\\ \geq&1-2n_1\exp\Big(-\frac{(4\sqrt{\ln(n_1n_2)})^2}{2}\Big) = 1-2n_1^{-7}n_2^{-8},\end{aligned}$$

where the first inequality applies the estimation of the cumulative distribution function of the Gaussian distribution (Ledoux and Talagrand, 1991, pg 8).

Combining this estimation for each column of $\mathbf{N}^*$ and applying the union bound, the second inequality in part (a) holds with probability $1-2n_1^{-7}n_2^{-7}$. Similarly, the first inequality in part (a) holds with the same probability.

(b) First, we parameterize $\mathbf{L}$ by $g(\mathbf{L}) = P_{\mathbf{L}^*}(\mathbf{L}-\mathbf{L}^*)$. Then we claim that, for any $\mathbf{L}$ and $\mathbf{L}'$ such that $\|\mathbf{L}-\mathbf{L}^*\|_F, \|\mathbf{L}'-\mathbf{L}^*\|_F \leq a\sigma_r(\mathbf{L}^*)$, there exists C_0 depending on a such that

$$\|P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*) - P_{T_{\mathbf{L}'}}(\mathbf{L}'-\mathbf{L}^*)\|_F \leq C_0\|g(\mathbf{L}) - g(\mathbf{L}')\|_F. \tag{48}$$

To prove (48), apply (29) and obtain

$$\|\mathbf{L}-\mathbf{L}'\|_F \leq \frac{1}{1-\frac{a}{2}}\|g(\mathbf{L}) - g(\mathbf{L}')\|_F. \tag{49}$$

Since $P_{T_{\mathbf{L}}} = \mathbf{U}_{\mathbf{L}}\mathbf{U}_{\mathbf{L}}^T + \mathbf{V}_{\mathbf{L}}\mathbf{V}_{\mathbf{L}}^T - \mathbf{U}_{\mathbf{L}}\mathbf{U}_{\mathbf{L}}^T\mathbf{V}_{\mathbf{L}}\mathbf{V}_{\mathbf{L}}^T$, and using Davis-Kahan theorem Davis and Kahan (1970) and the assumption $\|\mathbf{L}-\mathbf{L}^*\|_F \leq a\sigma_r(\mathbf{L}^*)$, there exists c_1, c_2 depending on a such that

$$\|\mathbf{U}_{\mathbf{L}}\mathbf{U}_{\mathbf{L}}^T - \mathbf{U}_{\mathbf{L}'}\mathbf{U}_{\mathbf{L}'}^T\|_F \leq c_1,\ \|\mathbf{V}_{\mathbf{L}}\mathbf{V}_{\mathbf{L}}^T - \mathbf{V}_{\mathbf{L}'}\mathbf{V}_{\mathbf{L}'}^T\|_F \leq c_2,$$

so there exists C' depending on a such that

$$\begin{aligned}&\|P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*) - P_{T_{\mathbf{L}'}}(\mathbf{L}'-\mathbf{L}^*)\|_F \\ =&\|[P_{T_{\mathbf{L}'}}(\mathbf{L}-\mathbf{L}^*) - P_{T_{\mathbf{L}'}}(\mathbf{L}'-\mathbf{L}^*)] + [P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*) - P_{T_{\mathbf{L}'}}(\mathbf{L}-\mathbf{L}^*)]\|_F\\ \leq&\|\mathbf{L}-\mathbf{L}'\|_F + C'\|\mathbf{L}-\mathbf{L}'\|_F.\end{aligned} \tag{50}$$

Combining (49) and (50), (48) is proved.

Second, based on (48), we will apply an ϵ-net covering argument to finish the proof that combines probabilistic estimation for each $\mathbf{L}$ and a union bound (ϵ-net covering argument is a standard argument in probabilistic estimation Vershynin (2012)). Use the estimation of the cumulative distribution function of the Gaussian distribution (Ledoux and Talagrand, 1991, pg 8), for any $\mathbf{L}'$,

$$\Pr\left\{\langle \mathbf{N}^*, P_{T_{\mathbf{L}'}}(\mathbf{L}'-\mathbf{L}^*)\rangle_F \geq t\|P_{T_{\mathbf{L}'}}(\mathbf{L}'-\mathbf{L}^*)\|_F\right\} \leq \frac{1}{2}\exp\left(-\frac{t^2}{2}\right).$$

For any $\mathbf{L}$ such that $\|g(\mathbf{L}')-g(\mathbf{L})\|_F<\epsilon$, applying (48),

$$\Pr\left\{\langle \mathbf{N}^*, P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\rangle_F \geq t\|P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\|_F + C_0\epsilon(\|\mathbf{N}^*\|_F+t)\right\} \leq \frac{1}{2}\exp\left(-\frac{t^2}{2}\right).$$

Using union bound, there is an ϵ-net of the set $\{g(\mathbf{L}) : \|g(\mathbf{L})\|_F = x\}$ with at most $(C_5x/\epsilon)^{n_1r+n_2r-r^2}$ points. Therefore, for all $\mathbf{L}$ such that $x-\epsilon \leq \|P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\|_F \leq x+\epsilon$,

$$\begin{aligned}&\Pr\left\{\langle \mathbf{N}^*, P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\rangle_F \geq t\|P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\|_F + 2C_0\epsilon(\|\mathbf{N}^*\|_F+t)\right\}\\ \leq&\frac{1}{2}\exp\left(-\frac{t^2}{2}\right)\cdot\left(\frac{C_5x}{\epsilon}\right)^{n_1r+n_2r-r^2}.\end{aligned}\tag{51}$$

Let $t=x/8$ and $\epsilon = x/16C_0\|\mathbf{N}^*\|_F$, then when $\|\mathbf{N}^*\|_F\geq 1$ (which holds with high probability as n_1n_2 goes to infinity), then using $C_0\geq 1$ we have $\epsilon\leq x/16$, and when $x\geq 4$,

$$\begin{aligned}&t\|P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\|_F + 2C_0\epsilon(\|\mathbf{N}^*\|_F+t) \leq \frac{x}{8}(x+\epsilon)+\frac{x}{8\|\mathbf{N}^*\|_F}(\|\mathbf{N}^*\|_F+t)\\ =&\frac{x}{8}(x+\epsilon)+\frac{x}{8}+\frac{x^2}{64\|\mathbf{N}^*\|_F}\leq\frac{x^2}{8}\frac{17}{16}+\frac{x}{8}+\frac{x^2}{64}\leq\frac{x^2}{8}\frac{17}{16}+\frac{x^2}{32}+\frac{x^2}{64}\leq\frac{1}{4}(x-\epsilon)^2\\ \leq&\frac{1}{4}\|P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\|_F^2,\end{aligned}\tag{52}$$

where the last inequality applies the assumption $x-\epsilon\leq\|P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\|_F$. Combining (51) and (52) and recall that $t=x/8$, we have that for all $\mathbf{L}$ such that $x-x/16C_0\|\mathbf{N}^*\|_F\leq \|P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\|_F\leq x+x/16C_0\|\mathbf{N}^*\|_F$,

$$\begin{aligned}&\Pr\Bigg\{\langle \mathbf{N}^*, P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\rangle_F \geq \frac{1}{4}\|P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\|_F^2,\\ &\qquad\text{for all } \mathbf{L} \text{ s.t. } \Big|\|P_{T_{\mathbf{L}}}(\mathbf{L}-\mathbf{L}^*)\|_F - x\Big| \leq \frac{x}{16C_0\|\mathbf{N}^*\|_F}\Bigg\}\\ \leq&\frac{1}{2}\exp\left(-\frac{x^2}{128}\right)\cdot\Big(16C_5C_0\|\mathbf{N}^*\|_F\Big)^{n_1r+n_2r-r^2}.\end{aligned}\tag{53}$$

Let $x_i = \sqrt{n_1 + n_2 + 128(n_1 r + n_2 r - r^2)\ln(16C_5C_0\|\mathbf{N}^*\|_F)}(1 + 1/16C_0\|\mathbf{N}^*\|_F)^i$ with $i = 1, 2, ...$, then

$$\begin{aligned}
&\sum_{i=1}^{\infty} \exp\left(-\frac{x_i^2}{128}\right) \cdot \Big(16C_5C_0\|\mathbf{N}^*\|_F\Big)^{n_1r+n_2r-r^2} \\
\leq& \exp(-\frac{n_1+n_2}{128}) \sum_{i=1}^{\infty} \exp(-(1 + 1/16C_0\|\mathbf{N}^*\|_F)^{2i}) \\
\leq& \exp(-\frac{n_1+n_2}{128}) \sum_{i=1}^{\infty} \exp(-1 - i/8C_0\|\mathbf{N}^*\|_F) \\
=& \exp(-\frac{n_1+n_2}{128} - 1) \frac{\exp(-1/8C_0\|\mathbf{N}^*\|_F)}{1 - \exp(-1/8C_0\|\mathbf{N}^*\|_F)} \leq 8C_0\|\mathbf{N}^*\|_F \exp(-\frac{n_1+n_2}{128} - 1), \qquad (54)
\end{aligned}$$

where the last inequality uses $\exp(-c) \leq 1 - c$ when $c \geq 0$. Clearly, the RHS goes to 0 as $n_1 + n_2 \to \infty$.

Combining the estimation (53) for $\{x_i\}_{i=1}^{\infty}$, with probability $1 - 8C_0\|\mathbf{N}^*\|_F \exp(-\frac{n_1+n_2}{128} - 1)$, the event (36) holds for all $\mathbf{L}$ such that

$$\|g(\mathbf{L})\|_F \geq \max(\sqrt{n_1 + n_2 + 128(n_1r + n_2r - r^2)\ln(16C_5C_0\|\mathbf{N}^*\|_F)}, 4).$$

Combining it with (29), the event (36) holds for all for all $\mathbf{L}$ such that

$$\begin{aligned}
a\sigma_r(\mathbf{L}^*) &\geq \|\mathbf{L} - \mathbf{L}^*\|_F \\
&\geq \frac{1}{1 - \frac{a}{2}} \max(\sqrt{n_1 + n_2 + 128(n_1r + n_2r - r^2)\ln(16C_5C_0\|\mathbf{N}^*\|_F)}, 4).
\end{aligned}$$

Considering that $\sqrt{n_1 + n_2 + 128(n_1r + n_2r - r^2)\ln(16C_5C_0\|\mathbf{N}^*\|_F)}$ is the dominant term when $n_1, n_2 \to \infty$, Lemma 13(b) is proved. ■

Lemma 16 **Proof** Following (41) and the proof of Lemma 10[a], and note that Lemma 15 means that γ^* and γ are replaced by arbitrary numbers in the intervals $[0.5p\gamma^*, 1.5p\gamma^*]$ and $[0.5p\gamma, 1.5p\gamma]$, we have

$$\|\tilde{\mathbf{D}}\|_F^2 \leq 6(\gamma + 2\gamma^*)p\mu r\|\mathbf{L} - \mathbf{L}^*\|_F^2 + 4\frac{3\gamma^*}{\gamma - 3\gamma^*}\|P_{\mathbf{\Phi}}(\mathbf{L} - \mathbf{L}^*)\|_F^2 + a^2\|\mathbf{L} - \mathbf{L}^*\|_F^2.$$

Applying Lemma 11 and (44), we have

$$\begin{aligned}
\|P_{\mathbf{\Phi}}(\mathbf{L} - \mathbf{L}^*)\|_F &\leq \|P_{\mathbf{\Phi}}P_{T_{\mathbf{L}^*}}(\mathbf{L} - \mathbf{L}^*)\|_F + \|P_{\mathbf{\Phi}}P_{T_{\mathbf{L}^*}}^{\perp}(\mathbf{L} - \mathbf{L}^*)\|_F \\
&\leq \sqrt{p(1+\epsilon)}\|\mathbf{L} - \mathbf{L}^*\|_F + \frac{a}{2}\|\mathbf{L} - \mathbf{L}^*\|_F.
\end{aligned}$$

Combining it with the estimation of $\|P_{\mathbf{\Phi}}(\mathbf{L} - \mathbf{L}^*)\|_F$ in Lemma 11, we have $\|\tilde{\mathbf{D}}\|_F \leq \tilde{C}_1\|P_{\mathbf{\Phi}}(\mathbf{L} - \mathbf{L}^*)\|_F$ with

$$\tilde{C}_1 = \frac{1}{p(1-\epsilon)}\Big[6(\gamma + 2\gamma^*)p\mu r + 4\frac{3\gamma^*}{\gamma - 3\gamma^*}(\sqrt{p(1+\epsilon)} + \frac{a}{2})^2 + a^2\Big].$$

■

References

P.-A. Absil and I V Oseledets. Low-rank retractions: a survey and new results. *Computational Optimization and Applications*, 62(1):5–29, 2015. ISSN 1573-2894. doi: 10.1007/s10589-014-9714-4. URL `http://dx.doi.org/10.1007/s10589-014-9714-4`.

P.A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009. ISBN 9781400830244. URL `http://books.google.com/books?id=NSQGQeLN3NcC`.

Afonso S. Bandeira, Nicolas Boumal, and Vladislav Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 361–382, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL `http://proceedings.mlr.press/v49/bandeira16.html`.

R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, February 2003.

Srinadh Bhojanapalli, Prateek Jain, and Sujay Sanghavi. Tighter low-rank approximation via sampling the leveraged element. In *Proceedings of the Twenty-sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 902–920, Philadelphia, PA, USA, 2015. Society for Industrial and Applied Mathematics. URL `http://dl.acm.org/citation.cfm?id=2722129.2722191`.

Nicolas Boumal. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26 (4):2355–2377, 2016. doi: 10.1137/16M105808X. URL `http://dx.doi.org/10.1137/16M105808X`.

Samuel Burer and Renato D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003. ISSN 1436-4646. doi: 10.1007/s10107-002-0352-8. URL `http://dx.doi.org/10.1007/s10107-002-0352-8`.

Samuel Burer and Renato D.C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005. ISSN 1436-4646. doi: 10.1007/s10107-004-0564-1. URL `http://dx.doi.org/10.1007/s10107-004-0564-1`.

Léopold Cambier and P.-A. Absil. Robust low-rank matrix completion by riemannian optimization. *SIAM Journal on Scientific Computing*, 38(5):S440–S460, 2016. doi: 10.1137/15M1025153. URL `https://doi.org/10.1137/15M1025153`.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395. URL `http://doi.acm.org/10.1145/1970392.1970395`.

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. doi: 10.1137/090761793. URL http://dx.doi.org/10.1137/090761793.

Yudong Chen and Martin J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *CoRR*, abs/1509.03025, 2015.

Yeshwanth Cherapanamjeri, Kartik Gupta, and Prateek Jain. Nearly-optimal robust matrix completion. *CoRR*, abs/1606.07315, 2016. URL http://arxiv.org/abs/1606.07315.

Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 81–90, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2029-0. doi: 10.1145/2488608.2488620. URL http://doi.acm.org/10.1145/2488608.2488620.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001. URL http://dx.doi.org/10.1137/0707001.

Christopher De Sa, Kunle Olukotun, and Christopher Ré. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2332–2341. JMLR.org, 2015. URL http://dl.acm.org/citation.cfm?id=3045118.3045366.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. ISSN 1097-4571. doi: 10.1002/(SICI)1097-4571(199009)41:6⟨391::AID-ASI1⟩3.0.CO;2-9. URL http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

R. Epstein, P. Hallinan, and A. Yuille. 5 ± 2 eigenimages suffice: An empirical investigation of low-dimensional lighting models. In *IEEE Workshop on Physics-based Modeling in Computer Vision*, pages 108–116, June 1995.

Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, November 2004. ISSN 0004-5411. doi: 10.1145/1039488.1039494. URL http://doi.acm.org/10.1145/1039488.1039494.

Quanquan Gu, Zhaoran Wang, and Han Liu. Low-rank and sparse structure pursuit via alternating minimization. In Arthur Gretton and Christian C. Robert, editors, *AISTATS*, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 600–609. JMLR.org, 2016. URL http://dblp.uni-trier.de/db/conf/aistats/aistats2016.html#GuWL16.

J. Ho, M. Yang, J. Lim, K. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 11–18, 2003.

D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, Nov 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2158250.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 665–674, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2029-0. doi: 10.1145/2488608.2488693. URL `http://doi.acm.org/10.1145/2488608.2488693`.

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. ISSN 0018-9448. doi: 10.1109/TIT.2010.2046205.

A. Kyrillidis and V. Cevher. Matrix alps: Accelerated low rank and sparse matrix reconstruction. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 185–188, Aug 2012. doi: 10.1109/SSP.2012.6319655.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. A Series of Modern Surveys in Mathematics Series. Springer, 1991. ISBN 9783540520139. URL `https://books.google.com/books?id=cyKYDfvxRjsC`.

L. Li, W. Huang, I. Gu, and Q. Tian. Statistical modeling of complex backgrounds for foreground object detection. *Image Processing, IEEE Transactions on*, 13(11):1459 – 1472, nov. 2004. ISSN 1057-7149. doi: 10.1109/TIP.2004.836169.

X. Li and J. Haupt. Identifying outliers in large matrices via randomized adaptive compressive sampling. *IEEE Transactions on Signal Processing*, 63(7):1792–1807, April 2015. ISSN 1053-587X. doi: 10.1109/TSP.2015.2401536.

Lester W. Mackey, Michael I. Jordan, and Ameet Talwalkar. Divide-and-conquer matrix factorization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1134–1142. Curran Associates, Inc., 2011. URL `http://papers.nips.cc/paper/4486-divide-and-conquer-matrix-factorization.pdf`.

Praneeth Netrapalli, Niranjan U N, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1107–1115. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5430-non-convex-robust-pca.pdf`.

Dohyung Park, Anastasios Kyrillidis, Srinadh Bhojanapalli, Constantine Caramanis, and Sujay Sanghavi. Provable burer-monteiro factorization for a class of norm-constrained matrix problems. *arXiv preprint arXiv:1606.01316*, 2016.

Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference* *on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 65–74, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL `http://proceedings.mlr.press/v54/park17a.html`.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, aug 2006. ISSN 1061-4036 (Print). doi: 10.1038/ng1847.

M. Rahmani and G. K. Atia. High dimensional low rank plus sparse matrix decomposition. *IEEE Transactions on Signal Processing*, 65(8):2004–2019, April 2017. ISSN 1053-587X. doi: 10.1109/TSP.2017.2649482.

A. Ruhe. *Numerical Computation of Principal Components when Several Observations are Missing*. Univ., 1974. URL `https://books.google.com/books?id=CgbyjgEACAAJ`.

Uri Shalit, Daphna Weinshall, and Gal Chechik. Online learning in the embedded manifold of low-rank matrices. *J. Mach. Learn. Res.*, 13(1):429–458, February 2012. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2503308.2188399`.

J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, Feb 2017. ISSN 0018-9448. doi: 10.1109/TIT.2016.2632162.

Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of optimization*, 6(615-640):15, 2010.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 964–973, 2016. URL `http://jmlr.org/proceedings/papers/v48/tu16.html`.

Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. doi: 10.1137/110845768. URL `http://dx.doi.org/10.1137/110845768`.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and GittaEditors Kutyniok, editors, *Compressed Sensing: Theory and Practice*, pages 210–268. Cambridge University Press, 2012. ISBN 9780511794308. doi: 10.1017/CBO9780511794308.006.

Lingxiao Wang, Xiao Zhang, and Quanquan Gu. A Unified Computational and Statistical Framework for Nonconvex Low-rank Matrix Estimation. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 981–990, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL `http://proceedings.mlr.press/v54/wang17b.html`.

Ke Wei, Jian-Feng Cai, Tony F. Chan, and Shingyu Leung. Guarantees of riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016. doi: 10.1137/15M1050525. URL `https://doi.org/10.1137/15M1050525`.

Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4152–4160, 2016. URL `http://papers.nips.cc/paper/6445-fast-algorithms-for-robust-pca-via-gradient-descent`.