Nonconvex Low Rank Matrix Factorization via Inexact First Order Oracle

Tuo Zhao* Zhaoran Wang† Han Liu‡

Abstract

We study the low rank matrix factorization problem via nonconvex optimization. Compared with the convex relaxation approach, nonconvex optimization exhibits superior empirical performance for large scale low rank matrix estimation. However, the understanding of its theoretical guarantees is limited. To bridge this gap, we exploit the notion of inexact first order oracle, which naturally appears in low rank matrix factorization problems such as matrix sensing and completion. Particularly, our analysis shows that a broad class of nonconvex optimization algorithms, including alternating minimization and gradient-type methods, can be treated as solving two sequences of convex optimization algorithms using inexact first order oracle. Thus we can show that these algorithms converge geometrically to the global optima and recover the true low rank matrices under suitable conditions. Numerical results are provided to support our theory.

1 Introduction

Let $M^* \in \mathbb{R}^{m\times n}$ be a rank k matrix with k much smaller than m and n. Our goal is to estimate M^* based on partial observations of its entries. For example, matrix completion is based on a subsample of M^*'s entries, while matrix sensing is based on linear measurements $\langle A_i, M^*\rangle$, where $i \in \{1,\ldots,d\}$ with d much smaller than mn and A_i is the sensing matrix. In the past decade, significant progress has been made on the recovery of low rank matrix [Candès and Recht, 2009, Candès and Tao, 2010, Candes and Plan, 2010, Recht et al., 2010, Lee and Bresler, 2010, Keshavan et al., 2010a,b, Jain et al., 2010, Cai et al., 2010, Recht, 2011, Gross, 2011, Chandrasekaran et al., 2011, Hsu et al., 2011, Rohde and Tsybakov, 2011, Koltchinskii et al., 2011, Negahban and Wainwright, 2011, Chen et al., 2011, Xiang et al., 2012, Negahban and Wainwright, 2012, Agarwal et al., 2012, Recht and Ré, 2013, Chen, 2013, Chen et al., 2013a,b, Jain et al., 2013, Jain and Netrapalli, 2014,

*Tuo Zhao is Affiliated with Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 USA and Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544; e-mail: tour@cs.jhu.edu.

†Zhaoran Wang is Affiliated with Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA; e-mail: zhaoran@princeton.edu.

‡Han Liu is Affiliated with Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA; e-mail: hanliu@princeton.edu.

Hardt, 2014, Hardt et al., 2014, Hardt and Wootters, 2014, Sun and Luo, 2014, Hastie et al., 2014, Cai and Zhang, 2015, Yan et al., 2015, Zhu et al., 2015, Wang et al., 2015]. Among these works, most are based upon convex relaxation with nuclear norm constraint or regularization. Nevertheless, solving these convex optimization problems can be computationally prohibitive in high dimensional regimes with large m and n [Hsieh and Olsen, 2014]. A computationally more efficient alternative is nonconvex optimization. In particular, we reparameterize the $m \times n$ matrix variable M in the optimization problem as $UV^\top$ with $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$, and optimize over U and V. Such a reparametrization automatically enforces the low rank structure and leads to low computational cost per iteration. Due to this reason, the nonconvex approach is widely used in large scale applications such as recommendation systems or collaborative filtering [Koren, 2009, Koren et al., 2009].

Despite the superior empirical performance of the nonconvex approach, the understanding of its theoretical guarantees is rather limited in comparison with the convex relaxation approach. The classical nonconvex optimization theory can only show its sublinear convergence to local optima. But many empirical results have corroborated its exceptional computational performance and convergence to global optima. Only until recently has there been theoretical analysis of the block coordinate descent-type nonconvex optimization algorithm, which is known as alternating minimization [Jain et al., 2013, Hardt, 2014, Hardt et al., 2014, Hardt and Wootters, 2014]. In particular, the existing results show that, provided a proper initialization, the alternating minimization algorithm attains a linear rate of convergence to a global optimum $U^* \in \mathbb{R}^{m\times k}$ and $V^* \in \mathbb{R}^{n\times k}$, which satisfy $M^* = U^*V^{*\top}$. Meanwhile, Keshavan et al. [2010a,b] establish the convergence of the gradient-type methods, and Sun and Luo [2014] further establish the convergence of a broad class of nonconvex optimization algorithms including both gradient-type and block coordinate descent-type methods. However, Keshavan et al. [2010a,b], Sun and Luo [2014] only establish the asymptotic convergence for an infinite number of iterations, rather than the explicit rate of convergence. Besides these works, Lee and Bresler [2010], Jain et al. [2010], Jain and Netrapalli [2014] consider projected gradient-type methods, which optimize over the matrix variable $M \in \mathbb{R}^{m\times n}$ rather than $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$. These methods involve calculating the top k singular vectors of an $m \times n$ matrix at each iteration. For k much smaller than m and n, they incur much higher computational cost per iteration than the aforementioned methods that optimize over U and V. All these works, except Sun and Luo [2014], focus on specific algorithms, while Sun and Luo [2014] do not establish the explicit optimization rate of convergence.

In this paper, we propose a new theory for analyzing a broad class of nonconvex optimization algorithms for low rank matrix estimation. The core of our theory is the notion of inexact first order oracle. Based on the inexact first order oracle, we establish sufficiently conditions under which the iteration sequences converge geometrically to the global optima. For both matrix sensing and completion, a direct consequence of our threoy is that, a broad family of nonconvex optimization algorithms, including gradient descent, block coordinate gradient descent, and block coordinate minimization, attain linear rates of convergence to the true low rank matrices

U^* and V^*. In particular, our proposed theory covers alternating minimization as a special case and recovers the results of Jain et al. [2013], Hardt [2014], Hardt et al. [2014], Hardt and Wootters [2014] under suitable conditions. Meanwhile, our approach covers gradient-type methods, which are also widely used in practice [Takács et al., 2007, Paterek, 2007, Koren et al., 2009, Gemulla et al., 2011, Recht and Ré, 2013, Zhuang et al., 2013]. To the best of our knowledge, our analysis is the first one that establishes exact recovery guarantees and geometric rates of convergence for a broad family of nonconvex matrix sensing and completion algorithms.

To achieve maximum generality, our unified analysis significantly differs from previous works. In detail, Jain et al. [2013], Hardt [2014], Hardt et al. [2014], Hardt and Wootters [2014] view alternating minimization as an approximate power method. However, their point of view relies on the closed form solution of each iteration of alternating minimization, which makes it difficult to generalize to other algorithms, e.g., gradient-type methods. Meanwhile, Sun and Luo [2014] take a geometric point of view. In detail, they show that the global optimum of the optimization problem is the unique stationary point within its neighborhood and thus a broad class of algorithms succeed. However, such geometric analysis of the objective function does not characterize the convergence rate of specific algorithms towards the stationary point. Unlike existing results, we analyze nonconvex optimization algorithms as approximate convex counterparts. For example, our analysis views alternating minimization on a nonconvex objective function as an approximate block coordinate minimization on some convex objective function. We use the key quantity, the inexact first order oracle, to characterize such a perturbation effect, which results from the local nonconvexity at intermediate solutions. This eventually allows us to establish explicit rate of convergence in an analogous way as existing convex optimization analysis.

Our proposed inexact first order oracle is closely related to a series previous work on inexact or approximate gradient descent algorithms: Güler [1992], Luo and Tseng [1993], Nedić and Bertsekas [2001], d'Aspremont [2008], Baes [2009], Friedlander and Schmidt [2012], Devolder et al. [2014]. Different from these existing results focusing on convex minimization, we show that the inexact first order oracle can also sharply captures the evolution of generic optimization algorithms even with the presence of nonconvexity. More recently, Candes et al. [2014], Balakrishnan et al. [2014], Arora et al. [2015] respectively analyze the Wirtinger Flow algorithm for phase retrieval, the expectation maximization (EM) Algorithm for latent variable models, and the gradient descent algorithm for sparse coding based on a similar idea to ours. Though their analysis exploits similar nonconvex structures, they work on completely different problems, and the delivered technical results are also fundamentally different.

A conference version of this paper was presented in the Annual Conference on Neural Information Processing Systems 2015 [Zhao et al., 2015]. During our conference version was under review, similar work was released on arXiv.org by Zheng and Lafferty [2015], Bhojanapalli et al. [2015], Tu et al. [2015], Chen and Wainwright [2015]. These works focus on symmetric positive semidefinite low rank matrix factorization problems. In contrast, our proposed methodologies and theory do not require the symmetry and positive semidefiniteness, and therefore can be ap-

plied to rectangular low rank matrix factorization problems.

The rest of this paper is organized as follows. In §2, we review the matrix sensing problems, and then introduce a general class of nonconvex optimization algorithms. In §3, we present the convergence analysis of the algorithms. In §4, we lay out the proof. In §5, we extend the proposed methodology and theory to the matrix completion problems. In §6, we provide numerical experiments and draw the conclusion.

Notation: For $v = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, we define the vector ℓ_q norm as $\|v\|_q^q = \sum_j v_j^q$. We define e_i as an indicator vector, where the i-th entry is one, and all other entries are zero. For a matrix $A \in \mathbb{R}^{m\times n}$, we use $A_{*j} = (A_{1j}, ..., A_{mj})^\top$ to denote the j-th column of A, and $A_{i*} = (A_{i1}, ..., A_{in})^\top$ to denote the i-th row of A. Let $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ be the largest and smallest nonzero singular values of A. We define the following matrix norms: $\|A\|_{\mathrm{F}}^2 = \sum_j \|A_{*j}\|_2^2$, $\|A\|_2 = \sigma_{\max}(A)$. Moreover, we define $\|A\|_*$ to be the sum of all singular values of A. We define as the Moore-Penrose pseudoinverse of A as $A^\dagger$. Given another matrix $B \in \mathbb{R}^{m\times n}$, we define the inner product as $\langle A, B\rangle = \sum_{i,j} A_{ij}B_{ij}$. For a bivariate function $f(u, v)$, we define $\nabla_u f(u, v)$ to be the gradient with respect to u. Moreover, we use the common notations of $\Omega(\cdot)$, $O(\cdot)$, and $o(\cdot)$ to characterize the asymptotics of two real sequences.

2 Matrix Sensing

We start with the matrix sensing problem. Let $M^* \in \mathbb{R}^{m\times n}$ be the unknown low rank matrix of interest. We have d sensing matrices $A_i \in \mathbb{R}^{m\times n}$ with $i \in \{1, \ldots, d\}$. Our goal is to estimate M^* based on $b_i = \langle A_i, M^*\rangle$ in the high dimensional regime with d much smaller than mn. Under such a regime, a common assumption is $\mathrm{rank}(M^*) = k \ll \min\{d, m, n\}$. Existing approaches generally recover M^* by solving the following convex optimization problem

$$\min_{M\in\mathbb{R}^{m\times n}} \|M\|_* \quad \text{subject to } b = \mathcal{A}(M), \tag{1}$$

where $b = [b_1, ..., b_d]^\top \in \mathbb{R}^d$, and $\mathcal{A}(M) : \mathbb{R}^{m\times n} \to \mathbb{R}^d$ is an operator defined as

$$\mathcal{A}(M) = [\langle A_1, M\rangle, ..., \langle A_i, M\rangle]^\top \in \mathbb{R}^d. \tag{2}$$

Existing convex optimization algorithms for solving (1) are computationally inefficient, since they incur high per-iteration computational cost and only attain sublinear rates of convergence to the global optimum [Jain et al., 2013, Hsieh and Olsen, 2014]. Therefore in large scale settings, we usually consider the following nonconvex optimization problem instead

$$\min_{U\in\mathbb{R}^{m\times k}, V\in\mathbb{R}^{n\times k}} \mathcal{F}(U, V), \quad \text{where } \mathcal{F}(U, V) = \frac{1}{2}\|b - \mathcal{A}(UV^\top)\|_2^2. \tag{3}$$

The reparametrization of $M = UV^\top$, though making the problem in (3) nonconvex, significantly improves the computational efficiency. Existing literature [Koren, 2009, Koren et al., 2009, Takács et al., 2007, Paterek, 2007, Koren et al., 2009, Gemulla et al., 2011, Recht and Ré, 2013, Zhuang

et al., 2013] has established convincing evidence that (3) can be effectively solved by a broad variety of gradient-based nonconvex optimization algorithms, including gradient descent, alternating exact minimization (i.e., alternating least squares or block coordinate minimization), as well as alternating gradient descent (i.e., block coordinate gradient descent), as illustrated in Algorithm 1.

Algorithm 1 A family of nonconvex optimization algorithms for matrix sensing. Here $(\overline{U}, D, \overline{V}) \leftarrow \mathsf{KSVD}(M)$ is the rank k singular value decomposition of M. D is a diagonal matrix containing the top k singular values of M in decreasing order, and $\overline{U}$ and $\overline{V}$ contain the corresponding top k left and right singular vectors of M. $(\overline{V}, R_{\overline{V}}) \leftarrow \mathsf{QR}(V)$ is the QR decomposition, where $\overline{V}$ is the corresponding orthonormal matrix and $R_{\overline{V}}$ is the corresponding upper triangular matrix.

Input: $\{b_i\}_{i=1}^d$, $\{A_i\}_{i=1}^d$

Parameter: Step size η, Total number of iterations T

$(\overline{U}^{(0)}, D^{(0)}, \overline{V}^{(0)}) \leftarrow \mathsf{KSVD}(\sum_{i=1}^d b_i A_i)$, $V^{(0)} \leftarrow \overline{V}^{(0)} D^{(0)}$, $U^{(0)} \leftarrow \overline{U}^{(0)} D^{(0)}$

For: $t = 0, \ldots, T-1$

Updating V:

- Alternating Exact Minimization : $V^{(t+0.5)} \leftarrow \mathrm{argmin}_V \mathcal{F}(\overline{U}^{(t)}, V)$
- $(\overline{V}^{(t+1)}, R_{\overline{V}}^{(t+0.5)}) \leftarrow \mathsf{QR}(V^{(t+0.5)})$
- Alternating Gradient Descent : $V^{(t+0.5)} \leftarrow V^{(t)} - \eta \nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)})$
- $(\overline{V}^{(t+1)}, R_{\overline{V}}^{(t+0.5)}) \leftarrow \mathsf{QR}(V^{(t+0.5)})$, $U^{(t)} \leftarrow \overline{U}^{(t)} R_{\overline{V}}^{(t+0.5)\top}$
- Gradient Descent : $V^{(t+0.5)} \leftarrow V^{(t)} - \eta \nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)})$
- $(\overline{V}^{(t+1)}, R_{\overline{V}}^{(t+0.5)}) \leftarrow \mathsf{QR}(V^{(t+0.5)})$, $U^{(t+1)} \leftarrow \overline{U}^{(t)} R_{\overline{V}}^{(t+0.5)\top}$

Updating U:

- Alternating Exact Minimization : $U^{(t+0.5)} \leftarrow \mathrm{argmin}_U \mathcal{F}(U, \overline{V}^{(t+1)})$
- $(\overline{U}^{(t+1)}, R_{\overline{U}}^{(t+0.5)}) \leftarrow \mathsf{QR}(U^{(t+0.5)})$
- Alternating Gradient Descent : $U^{(t+0.5)} \leftarrow U^{(t)} - \eta \nabla_U \mathcal{F}(U^{(t)}, \overline{V}^{(t+1)})$
- $(\overline{U}^{(t+1)}, R_{\overline{U}}^{(t+0.5)}) \leftarrow \mathsf{QR}(U^{(t+0.5)})$, $V^{(t+1)} \leftarrow \overline{V}^{t+1} R_{\overline{U}}^{(t+0.5)\top}$
- Gradient Descent : $U^{(t+0.5)} \leftarrow U^{(t)} - \eta \nabla_U \mathcal{F}(U^{(t)}, \overline{V}^{(t)})$
- $(\overline{U}^{(t+1)}, R_{\overline{U}}^{(t+0.5)}) \leftarrow \mathsf{QR}(U^{(t+0.5)})$, $V^{(t+1)} \leftarrow \overline{V}^{t} R_{\overline{U}}^{(t+0.5)\top}$

End for

Output: $M^{(T)} \leftarrow U^{(T-0.5)} \overline{V}^{(T)\top}$ (for gradient descent we use $\overline{U}^{(T)} V^{(T)\top}$)

It is worth noting that the QR decomposition and rank k singular value decomposition in Algorithm 1 can be accomplished efficiently. In particular, the QR decomposition can be accomplished in $O(k^2 \max\{m, n\})$ operations, while the rank k singular value decomposition can be accomplished in $O(kmn)$ operations. In fact, the QR decomposition is not necessary for particular update schemes, e.g., Jain et al. [2013] prove that the alternating exact minimization update schemes with or without the QR decomposition are equivalent.

3 Convergence Analysis

We analyze the convergence of the algorithms illustrated in §2. Before we present the main results, we first introduce a unified analytical framework based on a key quantity named the approximate first order oracle. Such a unified framework equips our theory with the maximum generality. Without loss of generality, we assume $m \leq n$ throughout the rest of this paper.

3.1 Main Idea

We first provide an intuitive explanation for the success of nonconvex optimization algorithms, which forms the basis of our later analysis of the main results in §4. Recall that (3) can be written as a special instance of the following optimization problem,

$$\min_{U \in \mathbb{R}^{m \times k}, V \in \mathbb{R}^{n \times k}} f(U, V). \tag{4}$$

A key observation is that, given fixed U, $f(U, \cdot)$ is strongly convex and smooth in V under suitable conditions, and the same also holds for U given fixed V correspondingly. For the convenience of discussion, we summarize this observation in the following technical condition, which will be later verified for matrix sensing and completion under suitable conditions.

Condition 1 (Strong Biconvexity and Bismoothness)**.** There exist universal constants $\mu_+ > 0$ and $\mu_- > 0$ such that

$$\frac{\mu_-}{2}\|U' - U\|_{\mathrm{F}}^2 \leq f(U', V) - f(U, V) - \langle U' - U, \nabla_U f(U, V)\rangle \leq \frac{\mu_+}{2}\|U' - U\|_{\mathrm{F}}^2 \text{ for all } U, U',$$
$$\frac{\mu_-}{2}\|V' - V\|_{\mathrm{F}}^2 \leq f(U, V') - f(U, V) - \langle V' - V, \nabla_V f(U, V)\rangle \leq \frac{\mu_+}{2}\|V' - V\|_{\mathrm{F}}^2 \text{ for all } V, V'.$$

3.1.1 Ideal First Order Oracle

To ease presentation, we assume that U^* and V^* are the unique global minimizers to the generic optimization problem in (4). Assuming that U^* is given, we can obtain V^* by

$$V^* = \operatorname*{argmin}_{V \in \mathbb{R}^{n \times k}} f(U^*, V). \tag{5}$$

Condition 1 implies the objective function in (5) is strongly convex and smooth. Hence, we can choose any gradient-based algorithm to obtain V^*. For example, we can directly solve for V^* in

$$\nabla_V f(U^*, V) = 0, \tag{6}$$

or iteratively solve for V^* using gradient descent, i.e.,

$$V^{(t)} = V^{(t-1)} - \eta \nabla_V f(U^*, V^{(t-1)}), \tag{7}$$

where η is a step size. Taking gradient descent as an example, we can invoke classical convex optimization results [Nesterov, 2004] to prove that

$$\|V^{(t)} - V^*\|_{\mathrm{F}} \leq \kappa \|V^{(t-1)} - V^*\|_{\mathrm{F}} \text{ for all } t = 0, 1, 2, \ldots,$$

where $\kappa \in (0, 1)$ and only depends on μ_+ and μ_- in Condition 1. For notational simplicity, we call $\nabla_V f(U^*, V^{(t-1)})$ the ideal first order oracle, since we do not know U^* in practice.

3.1.2 Inexact First Order Oracle

Though the ideal first order oracle is not accessible in practice, it provides us insights to analyze nonconvex optimization algorithms. Taking gradient descent as an example, at the t-th iteration, we take a gradient descent step over V based on $\nabla_V f(U, V^{(t-1)})$. Now we can treat $\nabla_V f(U, V^{(t-1)})$ as an approximation of $\nabla_V f(U^*, V^{(t-1)})$, where the approximation error comes from approximating U^* by U. Then the relationship between $\nabla_V f(U^*, V^{(t-1)})$ and $\nabla_V f(U, V^{(t-1)})$ is similar to that between gradient and approximate gradient in existing literature on convex optimization. For simplicity, we call $\nabla_V f(U, V^{(t-1)})$ the inexact first order oracle.

To characterize the difference between $\nabla_V f(U^*, V^{(t-1)})$ and $\nabla_V f(U, V^{(t-1)})$, we define the approximation error of the inexact first order oracle as

$$\mathcal{E}(V, V', U) = \|\nabla_V f(U^*, V') - \nabla_V f(U, V')\|_{\mathrm{F}}, \tag{8}$$

where V' is the current decision variable for evaluating the gradient. In the above example, it holds for $V' = V^{(t-1)}$. Later we will illustrate that $\mathcal{E}(V, V', U)$ is critical to our analysis. In the above example of alternating gradient descent, we will prove later that for $V^{(t)} = V^{(t-1)} - \eta \nabla_V f(U, V^{(t-1)})$, we have

$$\|V^{(t)} - V^*\|_{\mathrm{F}} \leq \kappa \|V^{(t-1)} - V^*\|_{\mathrm{F}} + \frac{2}{\mu_+} \mathcal{E}(V^{(t)}, V^{(t-1)}, U). \tag{9}$$

In other words, $\mathcal{E}(V^{(t)}, V^{(t-1)}, U)$ captures the perturbation effect by employing the inexact first order oracle $\nabla_V f(U, V^{(t-1)})$ instead of the ideal first order oracle $\nabla_V f(U^*, V^{(t-1)})$. For $V^{(t+1)} = \operatorname{argmin}_V f(U, V)$, we will prove that

$$\|V^{(t)} - V^*\|_{\mathrm{F}} \leq \frac{1}{\mu_-} \mathcal{E}(V^{(t)}, V^{(t)}, U). \tag{10}$$

According to the update schemes shown in Algorithms 1 and 2, for alternating exact minimization, we set $U = U^{(t)}$ in (10), while for gradient descent or alternating gradient descent, we set $U = U^{(t-1)}$ or $U = U^{(t)}$ in (9) respectively. Due to symmetry, similar results also hold for $\|U^{(t)} - U^*\|_{\mathrm{F}}$.

To establish the geometric rate of convergence towards the global minima U^* and V^*, it remains to establish upper bounds for the approximate error of the inexact first oder oracle. Taking gradient decent as an example, we will prove that given an appropriate initial solution, we have

$$\frac{2}{\mu_+} \mathcal{E}(V^{(t)}, V^{(t-1)}, U^{(t-1)}) \leq \alpha \|U^{(t-1)} - U^*\|_{\mathrm{F}} \tag{11}$$

for some $\alpha \in (0, 1-\kappa)$. Combining with (9) (where we take $U = U^{(t-1)}$), (11) further implies

$$\|V^{(t)} - V^*\|_F \le \kappa \|V^{(t-1)} - V^*\|_F + \alpha \|U^{(t-1)} - U^*\|_F. \tag{12}$$

Correspondingly, similar results hold for $\|U^{(t)} - U^*\|_F$, i.e.,

$$\|U^{(t)} - U^*\|_F \le \kappa \|U^{(t-1)} - U^*\|_F + \alpha \|V^{(t-1)} - V^*\|_F. \tag{13}$$

Combining (12) and (13) we then establish the contraction

$$\max\{\|V^{(t)} - V^*\|_F, \|U^{(t)} - U^*\|_F\} \le (\alpha + \kappa) \cdot \max\{\|V^{(t-1)} - V^*\|_F, \|U^{(t-1)} - U^*\|_F\},$$

which further implies the geometric convergence, since $\alpha \in (0, 1-\kappa)$. Respectively, we can establish similar results for alternating exact minimization and alternating gradient descent. Based upon such a unified analysis, we now present the main results.

3.2 Main Results

Before presenting the main results, we first introduce an assumption known as the restricted isometry property (RIP). Recall that k is the rank of the target low rank matrix M^*.

Assumption 1 (Restricted Isometry Property)**.** The linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{m \times n} \to \mathbb{R}^d$ defined in (2) satisfies $2k$-RIP with parameter $\delta_{2k} \in (0, 1)$, i.e., for all $\Delta \in \mathbb{R}^{m \times n}$ such that $\text{rank}(\Delta) \le 2k$, it holds that

$$(1 - \delta_{2k})\|\Delta\|_F^2 \le \|\mathcal{A}(\Delta)\|_2^2 \le (1 + \delta_{2k})\|\Delta\|_F^2.$$

Several random matrix ensembles satisfy $2k$-RIP for a sufficiently large d with high probability. For example, suppose that each entry of A_i is independently drawn from a sub-Gaussian distribution, $\mathcal{A}(\cdot)$ satisfies $2k$-RIP with parameter δ_{2k} with high probability for $d = \Omega(\delta_{2k}^{-2} k n \log n)$.

The following theorem establishes the geometric rate of convergence of the nonconvex optimization algorithms summarized in Algorithm 1.

Theorem 1. Assume there exists a sufficiently small constant C_1 such that $\mathcal{A}(\cdot)$ satisfies $2k$-RIP with $\delta_{2k} \le C_1/k$, and the largest and smallest nonzero singular values of M^* are constants, which do not scale with (d, m, n, k). For any pre-specified precision ϵ, there exist an η and universal constants C_2 and C_3 such that for all $T \ge C_2 \log(C_3/\epsilon)$, we have $\|M^{(T)} - M^*\|_F \le \epsilon$.

The proof of Theorems 1 is provided in §4.2, §4.3, and §4.4. Theorem 1 implies that all three nonconvex optimization algorithms converge geometrically to the global optimum. Moreover, assuming that each entry of A_i is independently drawn from a sub-Gaussian distribution with mean zero and variance proxy one, our result further suggests that, to achieve exact low rank matrix recovery, our algorithm requires the number of measurements d to satisfy

$$d = \Omega(k^3 n \log n), \tag{14}$$

since we assume that $\delta_{2k} \leq C_1/k$. This sample complexity result matches the state-of-the-art result for nonconvex optimization methods, which is established by Jain et al. [2013]. In comparison with their result, which only covers the alternating exact minimization algorithm, our results holds for a broader variety of nonconvex optimization algorithms.

Note that the sample complexity in (14) depends on a polynomial of $\sigma_{\max}(M^*)/\sigma_{\min}(M^*)$, which is treated as a constant in our paper. If we allow $\sigma_{\max}(M^*)/\sigma_{\min}(M^*)$ to increase, we can plug the nonconvex optimization algorithms into the multi-stage framework proposed by Jain et al. [2013]. Following similar lines to the proof of Theorem 1, we can derive a new sample complexity, which is independent of $\sigma_{\max}(M^*)/\sigma_{\min}(M^*)$. See more details in Jain et al. [2013].

4 Proof of Main Results

We sketch the proof of Theorems 1. The proof of all related lemmas are provided in the appendix. For notational simplicity, let $\sigma_1 = \sigma_{\max}(M^*)$ and $\sigma_k = \sigma_{\min}(M^*)$. Recall the nonconvex optimization algorithms are symmetric about the updates of U and V. Hence, the following lemmas for the update of V also hold for updating U. We omit some statements for conciseness. Theorem 2 can be proved in a similar manner, and its proof is provided in Appendix F.

Before presenting the proof, we first introduce the following lemma, which verifies Condition 1.

Lemma 1. Suppose that $\mathcal{A}(\cdot)$ satisfies $2k$-RIP with parameter δ_{2k}. Given an arbitrary orthonormal matrix $\overline{U} \in \mathbb{R}^{m\times k}$, for any $V,\ V' \in \mathbb{R}^{n\times k}$, we have

$$\frac{1+\delta_{2k}}{2}\|V'-V\|_{\mathrm{F}}^2 \geq \mathcal{F}(\overline{U},V') - \mathcal{F}(\overline{U},V) - \langle \nabla_V \mathcal{F}(\overline{U},V), V'-V\rangle \geq \frac{1-\delta_{2k}}{2}\|V'-V\|_{\mathrm{F}}^2.$$

The proof of Lemma 1 is provided in Appendix A.1. Lemma 1 implies that $\mathcal{F}(\overline{U},\cdot)$ is strongly convex and smooth in V given a fixed orthonormal matrix $\overline{U}$, as specified in Condition 1. Equipped with Lemma 1, we now lay out the proof for each update scheme in Algorithm 1.

4.1 Rotation Issue

Given a factorization of $M^* = \overline{U}^* V^{*\top}$, we can equivalently represent it as $M^* = \overline{U}^*_{\mathrm{new}} V^{*\top}_{\mathrm{new}}$, where

$$\overline{U}^*_{\mathrm{new}} = \overline{U}^* O_{\mathrm{new}} \quad \text{and} \quad V^{*\top}_{\mathrm{new}} = V^{*\top} O_{\mathrm{new}}$$

for an arbitrary unitary matrix $O_{\mathrm{new}} \in \mathbb{R}^{k\times k}$. This implies that directly calculating $\|\overline{U}-\overline{U}^*\|_{\mathrm{F}}$ is not desirable and the algorithm may converge to an arbitrary factorization of M^*.

To address this issue, existing analysis usually chooses subspace distances to evaluate the difference between subspaces spanned by columns of $\overline{U}^*$ and $\overline{U}$, because these subspaces are invariant to rotations [Jain et al., 2013]. For example, let $\overline{U}_\perp \in \mathbb{R}^{m\times(m-k)}$ denote the orthonormal

complement to $\overline{U}$, we can choose the subspace distance as $\|\overline{U}_\perp^\top \overline{U}^*\|_{\mathrm{F}}$. For any $O_{\mathrm{new}} \in \mathbb{R}^{k\times k}$ such that $O_{\mathrm{new}}^\top O_{\mathrm{new}} = I_k$, we have

$$\|\overline{U}_\perp^\top \overline{U}^*_{\mathrm{new}}\|_{\mathrm{F}} = \|\overline{U}_\perp^\top \overline{U}^* O_{\mathrm{new}}\|_{\mathrm{F}} = \|\overline{U}_\perp^\top \overline{U}^*\|_{\mathrm{F}}.$$

In this paper, we consider a different subspace distance defined as

$$\min_{O^\top O = I_k} \|\overline{U} - \overline{U}^* O\|_{\mathrm{F}}. \tag{15}$$

We can verify that (15) is also invariant to rotation. The next lemma shows that (15) is equivalent to $\|\overline{U}_\perp^\top \overline{U}^*\|_{\mathrm{F}}$.

Lemma 2. Given two orthonormal matrices $\overline{U} \in \mathbb{R}^{m\times k}$ and $\overline{U}^* \in \mathbb{R}^{m\times k}$, we have

$$\|\overline{U}_\perp^\top \overline{U}^*\|_{\mathrm{F}} \le \min_{O^\top O = I} \|\overline{U} - \overline{U}^* O\|_{\mathrm{F}} \le \sqrt{2}\|\overline{U}_\perp^\top \overline{U}^*\|_{\mathrm{F}}.$$

The proof of Lemma 2 is provided in Stewart et al. [1990], therefore omitted. Equipped with Lemma 2, our convergence analysis guarantees that there always exists a factorization of M^* satisfying the desired computational properties for each iteration (See Lemma 5, Corollaries 1 and 2). Similarly, the above argument can also be generalized to gradient descent and alternating gradient descent algorithms.

4.2 Proof of Theorem 1 (Alternating Exact Minimization)

Proof. Throughout the proof for alternating exact minimization, we define a constant $\xi \in (1, \infty)$ to simplify the notation. Moreover, we assume that at the t-th iteration, there exists a matrix factorization of M^*

$$M^* = \overline{U}^{*(t)} V^{*(t)\top},$$

where $\overline{U}^{*(t)} \in \mathbb{R}^{m\times k}$ is an orthonormal matrix. We define the approximation error of the inexact first order oracle as

$$\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}) = \|\nabla_V \mathcal{F}(\overline{U}^{*(t)}, V^{(t+0.5)}) - \nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t+0.5)})\|_{\mathrm{F}}.$$

The following lemma establishes an upper bound for the approximation error of the approximation first order oracle under suitable conditions.

Lemma 3. Suppose that δ_{2k} and $\overline{U}^{(t)}$ satisfy

$$\delta_{2k} \le \frac{(1-\delta_{2k})^2\sigma_k}{12\xi k(1+\delta_{2k})\sigma_1} \quad \text{and} \quad \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \le \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}. \tag{16}$$

Then we have

$$\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}) \le \frac{(1-\delta_{2k})\sigma_k}{2\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}.$$

The proof of Lemma 3 is provided in Appendix A.2. Lemma 3 shows that the approximation error of the inexact first order oracle for updating V diminishes with the estimation error of $\overline{U}^{(t)}$, when $\overline{U}^{(t)}$ is sufficiently close to $\overline{U}^{*(t)}$. The following lemma quantifies the progress of an exact minimization step using the inexact first order oracle.

Lemma 4. We have

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{1}{1-\delta_{2k}} \mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}).$$

The proof of Lemma 4 is provided in Appendix A.3. Lemma 4 illustrates that the estimation error of $V^{(t+0.5)}$ diminishes with the approximation error of the inexact first order oracle. The following lemma characterizes the effect of the renormalization step using QR decomposition, i.e., the relationship between $V^{(t+0.5)}$ and $\overline{V}^{(t+1)}$ in terms of the estimation error.

Lemma 5. Suppose that $V^{(t+0.5)}$ satisfies

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{4}. \tag{17}$$

Then there exists a factorization of $M^* = U^{*(t+1)}\overline{V}^{*(t+1)}$ such that $\overline{V}^{*(t+0.5)} \in \mathbb{R}^{n\times k}$ is an orthonormal matrix, and satisfies

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{2}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}.$$

The proof of Lemma 5 is provided in Appendix A.4. The next lemma quantifies the accuracy of the initialization $\overline{U}^{(0)}$.

Lemma 6. Suppose that δ_{2k} satisfies

$$\delta_{2k} \leq \frac{(1-\delta_{2k})^2\sigma_k^4}{192\xi^2 k(1+\delta_{2k})^2\sigma_1^4}. \tag{18}$$

Then there exists a factorization of $M^* = \overline{U}^{*(0)}V^{*(0)\top}$ such that $\overline{U}^{*(0)} \in \mathbb{R}^{m\times k}$ is an orthonormal matrix, and satisfies

$$\|\overline{U}^{(0)} - \overline{U}^{*}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}.$$

The proof of Lemma 6 is provided in Appendix A.5. Lemma 6 implies that the initial solution $\overline{U}^{(0)}$ attains a sufficiently small estimation error.

Combining Lemmas 3, 4, and 5, we obtain the following corollary for a complete iteration of updating V.

Corollary 1. Suppose that δ_{2k} and $\overline{U}^{(t)}$ satisfy

$$\delta_{2k} \leq \frac{(1-\delta_{2k})^2\sigma_k^4}{192\xi^2 k(1+\delta_{2k})^2\sigma_1^4} \quad \text{and} \quad \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}. \tag{19}$$

We then have

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}.$$

Moreover, we also have

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{1}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \quad \text{and} \quad \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{2\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}.$$

The proof of Corollary 1 is provided in Appendix A.6. Since the alternating exact minimization algorithm updates U and V in a symmetric manner, we can establish similar results for a complete iteration of updating U in the next corollary.

Corollary 2. Suppose that δ_{2k} and $\overline{V}^{(t+1)}$ satisfy

$$\delta_{2k} \leq \frac{(1-\delta_{2k})^2\sigma_k^4}{192\xi^2 k(1+\delta_{2k})^2\sigma_1^4} \quad \text{and} \quad \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}. \tag{20}$$

Then there exists a factorization of $M^* = \overline{U}^{*(t+1)}V^{*(t+1)\top}$ such $\overline{U}^{*(t+1)}$ is an orthonormal matrix, and satisfies

$$\|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}.$$

Moreover, we also have

$$\|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{1}{\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \quad \text{and} \quad \|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{2\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}.$$

The proof of Corollary 2 directly follows Appendix A.6, and is therefore omitted..

We then proceed with the proof of Theorem 1 for alternating exact minimization. Lemma 6 ensures that (19) of Corollary 1 holds for $\overline{U}^{(0)}$. Then Corollary 1 ensures that (20) of Corollary 2 holds for $\overline{V}^{(1)}$. By induction, Corollaries 1 and 2 can be applied recursively for all T iterations. Thus we obtain

$$\begin{aligned}\|\overline{V}^{(T)} - \overline{V}^{*(T)}\|_{\mathrm{F}} &\leq \frac{1}{\xi}\|\overline{U}^{(T-1)} - \overline{U}^{*(T-1)}\|_{\mathrm{F}} \leq \frac{1}{\xi^2}\|\overline{V}^{(T-1)} - \overline{V}^{*(T-1)}\|_{\mathrm{F}} \\ &\leq \cdots \leq \frac{1}{\xi^{2T-1}}\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi^{2T}(1+\delta_{2k})\sigma_1},\end{aligned} \tag{21}$$

where the last inequality comes from Lemma 6. Therefore, for a pre-specified accuracy ϵ, we need at most

$$T = \left\lceil \frac{1}{2}\log\left(\frac{(1-\delta_{2k})\sigma_k}{2\epsilon(1+\delta_{2k})\sigma_1}\right)\log^{-1}\xi \right\rceil \tag{22}$$

iterations such that

$$\|\overline{V}^{(T)} - \overline{V}^{*(T)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi^{2T}(1+\delta_{2k})\sigma_1} \leq \frac{\epsilon}{2}. \tag{23}$$

Moreover, Corollary 2 implies

$$\|U^{(T-0.5)} - U^{*(T)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{2\xi}\|\overline{V}^{(T)} - \overline{V}^{*(T)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k^2}{8\xi^{2T+1}(1+\delta_{2k})\sigma_1},$$

where the last inequality comes from (21). Therefore, we need at most

$$T = \left\lceil \frac{1}{2}\log\left(\frac{(1-\delta_{2k})\sigma_k^2}{4\xi\epsilon(1+\delta_{2k})}\right)\log^{-1}\xi \right\rceil \tag{24}$$

iterations such that

$$\|U^{(T-0.5)} - U^{*}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k^2}{8\xi^{2T+1}(1+\delta_{2k})\sigma_1} \leq \frac{\epsilon}{2\sigma_1}. \tag{25}$$

Then combining (23) and (25), we obtain

$$\begin{aligned}
\|M^{(T)} - M^*\| &= \|U^{(T-0.5)}\overline{V}^{(T)\top} - U^{*(T)}\overline{V}^{*(T)\top}\|_{\mathrm{F}} \\
&= \|U^{(T-0.5)}\overline{V}^{(T)\top} - U^{*(T)}\overline{V}^{(T)\top} + U^{*(T)}\overline{V}^{(T)\top} - U^{*(T)}\overline{V}^{*(T)\top}\|_{\mathrm{F}} \\
&\leq \|\overline{V}^{(T)}\|_2\|U^{(T-0.5)} - U^{*(T)}\|_{\mathrm{F}} + \|U^{*(T)}\|_2\|\overline{V}^{(T)} - \overline{V}^{*(T)}\|_{\mathrm{F}} \leq \epsilon,
\end{aligned} \tag{26}$$

where the last inequality comes from $\|\overline{V}^{(T)}\|_2 = 1$ (since $\overline{V}^{(T)}$ is orthonormal) and $\|U^*\|_2 = \|M^*\|_2 = \sigma_1$ (since $U^{*(T)}\overline{V}^{*(T)\top} = M^*$ and $\overline{V}^{*(T)}$ is orthonormal). Thus combining (22) and (24) with (26), we complete the proof. □

4.3 Proof of Theorem 1 (Alternating Gradient Descent)

Proof. Throughout the proof for alternating gradient descent, we define a sufficiently large constant ξ. Moreover, we assume that at the t-th iteration, there exists a matrix factorization of M^*

$$M^* = \overline{U}^{*(t)}V^{*(t)\top},$$

where $\overline{U}^{*(t)} \in \mathbb{R}^{m\times k}$ is an orthonormal matrix. We define the approximation error of the inexact first order oracle as

$$\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) = \|\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)}) - \nabla_V \mathcal{F}(\overline{U}^{*(t)}, V^{(t)})\|_{\mathrm{F}}.$$

The first lemma is parallel to Lemma 3 for alternating exact minimization.

Lemma 7. Suppose that δ_{2k}, $\overline{U}^{(t)}$, and $V^{(t)}$ satisfy

$$\delta_{2k} \leq \frac{(1-\delta_{2k})\sigma_k}{24\xi k\sigma_1}, \quad \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{4\xi\sigma_1^2}, \quad \text{and} \quad \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_1\sqrt{k}}{2}. \tag{27}$$

Then we have

$$\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \leq \frac{(1+\delta_{2k})\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}.$$

The proof of Lemma 7 is provided in Appendix B.1. Lemma 7 illustrates that the approximation error of the inexact first order oracle diminishes with the estimation error of $\overline{U}^{(t)}$, when $\overline{U}^{(t)}$ and $V^{(t)}$ are sufficiently close to $\overline{U}^{*(t)}$ and $V^{*(t)}$.

Lemma 8. Suppose that the step size parameter η satisfies

$$\eta = \frac{1}{1+\delta_{2k}}. \tag{28}$$

Then we have

$$\|V^{(t+0.5)} - V^*\|_{\mathrm{F}} \leq \sqrt{\delta_{2k}}\|V^{(t)} - V^*\|_{\mathrm{F}} + \frac{2}{1+\delta_{2k}}\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}).$$

The proof of Lemma 8 is in Appendix B.2. Lemma 8 characterizes the progress of a gradient descent step with a pre-specified fixed step size. A more practical option is adaptively selecting η using the backtracking line search procedure, and similar results can be guaranteed. See Nesterov [2004] for details. The following lemma characterizes the effect of the renormalization step using QR decomposition.

Lemma 9. Suppose that $V^{(t+0.5)}$ satisfies

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{4}. \tag{29}$$

Then there exists a factorization of $M^* = U^{*(t+1)}\overline{V}^{*(t+1)}$ such that $\overline{V}^{*(t+1)} \in \mathbb{R}^{n\times k}$ is an orthonormal matrix, and

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{2}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}},$$
$$\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \frac{3\sigma_1}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}},$$

The proof of Lemma 9 is provided in Appendix B.3. The next lemma quantifies the accuracy of the initial solutions.

Lemma 10. Suppose that δ_{2k} satisfies

$$\delta_{2k} \leq \frac{\sigma_k^6}{192\xi^2 k\sigma_1^6}. \tag{30}$$

Then we have

$$\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{4\xi\sigma_1^2} \quad \text{and} \quad \|V^{(0)} - V^{*(0)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{2\xi\sigma_1} \leq \frac{\sigma_1\sqrt{k}}{2}.$$

The proof of Lemma 10 is in Appendix B.4. Lemma 10 indicates that the initial solutions $\overline{U}^{(0)}$ and $V^{(0)}$ attain sufficiently small estimation errors.

Combining Lemmas 7, 8, 5, , we obtain the following corollary for a complete iteration of updating V.

Corollary 3. Suppose that δ_{2k}, $\overline{U}^{(t)}$, and $V^{(t)}$ satisfy

$$\delta_{2k} \le \frac{\sigma_k^6}{192\xi^2 k\sigma_1^6}, \quad \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{4\xi\sigma_1^2}, \quad \text{and} \quad \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{2\xi\sigma_1}. \tag{31}$$

We then have

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{4\xi\sigma_1^2} \quad \text{and} \quad \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{2\xi\sigma_1}.$$

Moreover, we have

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \le \sqrt{\delta_{2k}}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{32}$$

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \le \frac{2\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{4}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{33}$$

$$\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \le \frac{3\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \left(\frac{6}{\xi} + 1\right)\sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}. \tag{34}$$

The proof of Corollary 3 is provided in Appendix B.5. Since the alternating gradient descent algorithm updates U and V in a symmetric manner, we can establish similar results for a complete iteration of updating U in the next corollary.

Corollary 4. Suppose that δ_{2k}, $\overline{V}^{(t+1)}$, and $U^{(t)}$ satisfy

$$\delta_{2k} \le \frac{\sigma_k^6}{192\xi^2 k\sigma_1^6}, \ \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{4\xi\sigma_1^2}, \ \text{and} \ \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{2\xi\sigma_1}. \tag{35}$$

We then have

$$\|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{4\xi\sigma_1^2} \quad \text{and} \quad \|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{2\xi\sigma_1}.$$

Moreover, we have

$$\|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}} \le \sqrt{\delta_{2k}}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \frac{2\sigma_k}{\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{36}$$

$$\|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \le \frac{2\sqrt{\delta_{2k}}}{\sigma_k}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \frac{4}{\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{37}$$

$$\|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}} \le \frac{3\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \left(\frac{6}{\xi} + 1\right)\sigma_1\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}. \tag{38}$$

The proof of Corollary 4 directly follows Appendix B.5, and is therefore omitted..

Now we proceed with the proof of Theorem 1 for alternating gradient descent. Recall that Lemma 10 ensures that (31) of Corollary 3 holds for $\overline{U}^{(0)}$ and $V^{(0)}$. Then Corollary 3 ensures that (35) of Corollary 4 holds for $U^{(0)}$ and $\overline{V}^{(1)}$. By induction, Corollaries 1 and 2 can be applied recursively for all T iterations. For notational simplicity, we write (32)-(38) as

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \alpha_1 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \gamma_1 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{39}$$

$$\sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_2 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \gamma_2 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{40}$$

$$\|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_3 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \gamma_3 \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{41}$$

$$\sigma_1 \|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_4 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \gamma_4 \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{42}$$

$$\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_5 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \gamma_5 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{43}$$

$$\|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_6 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \gamma_6 \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}. \tag{44}$$

Note that we have $\gamma_5, \gamma_6 \in (1, 2)$, but α_1,...,α_6, γ_1,..., and γ_4 can be sufficiently small as long as ξ is sufficiently large. We then have

$$\begin{aligned}
\|U^{(t+1)} - U^{*(t+2)}\|_{\mathrm{F}} &\overset{\text{(i)}}{\leq} \alpha_5 \|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}} + \gamma_5 \sigma_1 \|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \\
&\overset{\text{(ii)}}{\leq} \alpha_5 \alpha_6 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \alpha_5 \gamma_6 \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} + \gamma_5 \sigma_1 \|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \\
&\overset{\text{(iii)}}{\leq} (\alpha_5 \alpha_6 + \gamma_5 \alpha_4) \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + (\gamma_5 \gamma_4 \sigma_1 + \alpha_5 \gamma_6) \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \\
&\overset{\text{(iv)}}{\leq} (\alpha_5 \alpha_6 + \gamma_5 \alpha_4) \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + (\gamma_5 \gamma_4 \sigma_1 + \alpha_5 \gamma_6) \alpha_2 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} \\
&\qquad + (\gamma_5 \gamma_4 \sigma_1 + \alpha_5 \gamma_6) \gamma_2 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}},
\end{aligned} \tag{45}$$

where (i) comes from (43), (ii) comes from (44), (iii) comes from (42), and (iv) comes from (40). Similarly, we can obtain

$$\begin{aligned}
\|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_6 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \gamma_6 \alpha_2 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} \\
+ \gamma_6 \gamma_2 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}},
\end{aligned} \tag{46}$$

$$\begin{aligned}
\sigma_1 \|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_4 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \gamma_4 \alpha_2 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} \\
+ \gamma_4 \gamma_2 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}
\end{aligned} \tag{47}$$

$$\begin{aligned}
\|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \alpha_3 \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \gamma_3 \alpha_2 \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} \\
+ \gamma_3 \gamma_2 \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}.
\end{aligned} \tag{48}$$

For simplicity, we define

$$\begin{aligned}
&\phi_{V^{(t+1)}} = \|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}}, \ \phi_{V^{(t+0.5)}} = \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}, \ \phi_{\overline{V}^{(t+1)}} = \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \\
&\phi_{U^{(t+1)}} = \|U^{(t+1)} - U^{*(t+2)}\|_{\mathrm{F}}, \ \phi_{U^{(t+0.5)}} = \|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}}, \ \phi_{\overline{U}^{(t+1)}} = \sigma_1 \|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}}.
\end{aligned}$$

Then combining (39), (40) with (45)–(48), we obtain

$$\max\left\{\phi_{V^{(t+1)}}, \phi_{V^{(t+0.5)}}, \phi_{\overline{V}^{(t+1)}}, \phi_{U^{(t+1)}}, \phi_{U^{(t+0.5)}}, \phi_{\overline{U}^{(t+1)}}\right\} \leq \beta \max\left\{\phi_{V^{(t)}}, \phi_{U^{(t)}}, \phi_{\overline{U}^{(t)}}\right\}, \tag{49}$$

where β is a contraction coefficient defined as

$$\begin{aligned}\beta = \max\{\alpha_5\alpha_6 + \gamma_5\alpha_4, \alpha_6, \alpha_4, \alpha_3\} + \max\{\alpha_1, \alpha_2, (\gamma_5\gamma_4\sigma_1 + \alpha_5\gamma_6), \gamma_6\alpha_2, \gamma_4\alpha_2, \gamma_3\alpha_2\} \\ + \max\{\gamma_1, \gamma_2, (\gamma_5\gamma_4\sigma_1 + \alpha_5\gamma_6)\gamma_2, \gamma_6\gamma_2, \gamma_4\gamma_2, \gamma_3\gamma_2\}.\end{aligned}$$

Then we can choose ξ as a sufficiently large constant such that $\beta < 1$. By recursively applying (49) for $t = 0, ..., T$, we obtain

$$\begin{aligned}\max\left\{\phi_{V^{(T)}}, \phi_{V^{(T-0.5)}}, \phi_{\overline{V}^{(T)}}, \phi_{U^{(T)}}, \phi_{U^{(T-0.5)}}, \phi_{\overline{U}^{(T)}}\right\} &\leq \beta \max\left\{\phi_{V^{(T-1)}}, \phi_{U^{(T-1)}}, \phi_{\overline{U}^{(T-1)}}\right\} \\ &\leq \beta^2 \max\left\{\phi_{V^{(T-2)}}, \phi_{U^{(T-2)}}, \phi_{\overline{U}^{(T-2)}}\right\} \leq \ldots \leq \beta^T \max\left\{\phi_{V^{(0)}}, \phi_{U^{(0)}}, \phi_{\overline{U}^{(0)}}\right\}.\end{aligned}$$

By Corollary 3, we obtain

$$\begin{aligned}\|U^{(0)} - U^{*(1)}\|_{\mathrm{F}} &\leq \frac{3\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(0)} - V^{*(0)}\|_{\mathrm{F}} + \left(\frac{6}{\xi} + 1\right)\sigma_1\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \\ &\overset{\text{(i)}}{\leq} \frac{3\sigma_1}{\sigma_k} \cdot \frac{\sigma_k^3}{12\xi\sigma_1^3} \cdot \frac{\sigma_k^2}{2\xi\sigma_1} + \left(\frac{6}{\xi} + 1\right)\frac{\sigma_k^2}{4\xi\sigma_1} \\ &\overset{\text{(ii)}}{=} \frac{\sigma_k^4}{8\xi^2\sigma_1^3} + \frac{3\sigma_k^2}{2\xi^2\sigma_1} + \frac{\sigma_k^2}{4\xi\sigma_1} \overset{\text{(iii)}}{\leq} \frac{\sigma_k^2}{2\xi\sigma_1},\end{aligned} \tag{50}$$

where (i) and (ii) come from Lemma 10, and (iii) comes from the definition of ξ and $\sigma_1 \geq \sigma_k$. Combining (50) with Lemma 10, we have

$$\left\{\phi_{V^{(0)}}, \phi_{U^{(0)}}, \phi_{\overline{U}^{(0)}}\right\} \leq \max\left\{\frac{\sigma_k^2}{2\xi\sigma_1}, \frac{\sigma_k^2}{4\xi\sigma_1^2}\right\}.$$

Then we need at most

$$T = \left\lceil \log\left(\max\left\{\frac{\sigma_k^2}{\xi\sigma_1}, \frac{\sigma_k^2}{2\xi\sigma_1^2}, \frac{\sigma_k^2}{\xi}, \frac{\sigma_k^2}{2\xi\sigma_1}\right\} \cdot \frac{1}{\epsilon}\right) \log^{-1}(\beta^{-1}) \right\rceil$$

iterations such that

$$\|\overline{V}^{(T)} - \overline{V}^*\|_{\mathrm{F}} \leq \beta^T \max\left\{\frac{\sigma_k^2}{2\xi\sigma_1}, \frac{\sigma_k^2}{4\xi\sigma_1^2}\right\} \leq \frac{\epsilon}{2} \text{ and } \|U^{(T)} - U^*\|_{\mathrm{F}} \leq \beta^T \max\left\{\frac{\sigma_k^2}{2\xi\sigma_1}, \frac{\sigma_k^2}{4\xi\sigma_1^2}\right\} \leq \frac{\epsilon}{2\sigma_1}.$$

We then follow similar lines to (26) in §4.2, and show $\|M^{(T)} - M^*\|_{\mathrm{F}} \leq \epsilon$, which completes the proof. □

4.4 Proof of Theorem 1 (Gradient Descent)

Proof. The convergence analysis of the gradient descent algorithm is similar to that of the alternating gradient descent. The only difference is that for updating U, the gradient descent algorithm employs $V = \overline{V}^{(t)}$ instead of $V = \overline{V}^{(t+1)}$ to calculate the gradient at $U = U^{(t)}$. Then everything else directly follows §4.3, and is therefore omitted.. □

5 Extensions to Matrix Completion

We then extend our methodology and theory to matrix completion problems. Let $M^* \in \mathbb{R}^{m\times n}$ be the unknown low rank matrix of interest. We observe a subset of the entries of M^*, namely, $\mathcal{W} \subseteq \{1,\ldots,m\}\times\{1,\ldots,n\}$. We assume that $\mathcal{W}$ is drawn uniformly at random, i.e., $M^*_{i,j}$ is observed independently with probability $\bar{\rho} \in (0,1]$. To exactly recover M^*, a common assumption is the incoherence of M^*, which will be specified later. A popular approach for recovering M^* is to solve the following convex optimization problem

$$\min_{M\in\mathbb{R}^{m\times n}} \|M\|_* \quad \text{subject to } \mathcal{P}_{\mathcal{W}}(M^*) = \mathcal{P}_{\mathcal{W}}(M), \tag{51}$$

where $\mathcal{P}_{\mathcal{W}}(M):\ \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is an operator defined as

$$[\mathcal{P}_{\mathcal{W}}(M)]_{ij} = \begin{cases} M_{ij} & \text{if } (i,j)\in\mathcal{W}, \\ 0 & \text{otherwise.} \end{cases}$$

Similar to matrix sensing, existing algorithms for solving (51) are computationally inefficient. Hence, in practice we usually consider the following nonconvex optimization problem

$$\min_{U\in\mathbb{R}^{m\times k}, V\in\mathbb{R}^{n\times k}} \mathcal{F}_{\mathcal{W}}(U,V), \ \text{ where } \mathcal{F}_{\mathcal{W}}(U,V) = \frac{1}{2}\|\mathcal{P}_{\mathcal{W}}(M^*) - \mathcal{P}_{\mathcal{W}}(UV^\top)\|_{\mathrm{F}}^2. \tag{52}$$

Similar to matrix sensing, (52) can also be efficiently solved by gradient-based algorithms illustrated in Algorithm 2. For the convenience of later convergence analysis, we partition the observation set $\mathcal{W}$ into $2T+1$ subsets $\mathcal{W}_0$,...,$\mathcal{W}_{2T}$ by Algorithm 4. However, in practice we do not need the partition scheme, i.e., we simply set $\mathcal{W}_0 = \cdots = \mathcal{W}_{2T} = \mathcal{W}$.

Before we present the convergence analysis, we first introduce an assumption known as the incoherence property.

Assumption 2 (Incoherence Property)**.** The target rank k matrix M^* is incoherent with parameter μ, i.e., given the rank k singular value decomposition of $M^* = \overline{U}^*\Sigma^*\overline{V}^{*\top}$, we have

$$\max_i \|\overline{U}^*_{i*}\|_2 \le \mu\sqrt{\frac{k}{m}} \quad \text{and} \quad \max_j \|\overline{V}^*_{j*}\|_2 \le \mu\sqrt{\frac{k}{n}}.$$

Roughly speaking, the incoherence assumption guarantees that each entry of M^* contains similar amount of information, which makes it feasible to complete M^* when its entries are missing uniformly at random. The following theorem establishes the iteration complexity and the estimation error under the Frobenius norm.

Theorem 2. Suppose that there exists a universal constant C_4 such that $\bar{\rho}$ satisfies

$$\bar{\rho} \ge \frac{C_4\mu^2k^3\log n\log(1/\epsilon)}{m}, \tag{53}$$

where ϵ is the pre-specified precision. Then there exist an η and universal constants C_5 and C_6 such that for any $T \ge C_5\log(C_6/\epsilon)$, we have $\|M^{(T)} - M\|_{\mathrm{F}} \le \epsilon$ with high probability.

Algorithm 2 A family of nonconvex optimization algorithms for matrix completion. The incoherence factorization algorithm $\mathsf{IF}(\cdot)$ is illustrated in Algorithm 3, and the partition algorithm $\mathsf{Partition}(\cdot)$, which is proposed by Hardt and Wootters [2014], is provided in Algorithm 4 of Appendix C for the sake of completeness. The initialization procedures $\mathsf{INT}_{\overline{U}}(\cdot)$ and $\mathsf{INT}_{\overline{U}}(\cdot)$ are provided in Algorithm 5 and Algorithm 6 of Appendix D for the sake of completeness. Here $\mathcal{F}_{\mathcal{W}}(\cdot)$ is defined in (52).

Input: $\mathcal{P}_{\mathcal{W}}(M^*)$

Parameter: Step size η, Total number of iterations T

$(\{\mathcal{W}_t\}_{t=0}^{2T}, \widetilde{\rho}) \leftarrow \mathsf{Partition}(\mathcal{W})$, $\mathcal{P}_{\mathcal{W}_0}(\widetilde{M}) \leftarrow \mathcal{P}_{\mathcal{W}_0}(M^*)$, and $\widetilde{M}_{ij} \leftarrow 0$ for all $(i,j) \notin \mathcal{W}_0$

$(\overline{U}^{(0)}, V^{(0)}) \leftarrow \mathsf{INT}_{\overline{U}}(\widetilde{M})$, $(\overline{V}^{(0)}, U^{(0)}) \leftarrow \mathsf{INT}_{\overline{V}}(\widetilde{M})$

For: $t = 0, \ldots, T-1$

Updating V.

- Alternating Exact Minimization : $V^{(t+0.5)} \leftarrow \mathrm{argmin}_V \mathcal{F}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V)$
- $(\overline{V}^{(t+1)}, R_{\overline{V}}^{(t+0.5)}) \leftarrow \mathsf{IF}(V^{(t+0.5)})$
- Alternating Gradient Descent : $V^{(t+0.5)} \leftarrow V^{(t)} - \eta \nabla_V \mathcal{F}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V^{(t)})$
- $(\overline{V}^{(t+1)}, R_{\overline{V}}^{(t+0.5)}) \leftarrow \mathsf{IF}(V^{(t+0.5)})$, $U^{(t)} \leftarrow \overline{U}^{(t)} R_{\overline{V}}^{(t+0.5)\top}$
- Gradient Descent : $V^{(t+0.5)} \leftarrow V^{(t)} - \eta \nabla_V \mathcal{F}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V^{(t)})$
- $(\overline{V}^{(t+1)}, R_{\overline{V}}^{(t+0.5)}) \leftarrow \mathsf{IF}(V^{(t+0.5)})$, $U^{(t+1)} \leftarrow \overline{U}^{(t)} R_{\overline{V}}^{(t+0.5)\top}$

Updating U.

- Alternating Exact Minimization : $U^{(t+0.5)} \leftarrow \mathrm{argmin}_U \mathcal{F}_{\mathcal{W}_{2t+2}}(U, \overline{V}^{(t+1)})$
- $(\overline{U}^{(t+1)}, R_{\overline{U}}^{(t+0.5)}) \leftarrow \mathsf{IF}(U^{(t+0.5)})$
- Alternating Gradient Descent : $U^{(t+0.5)} \leftarrow U^{(t)} - \eta \nabla_U \mathcal{F}_{\mathcal{W}_{2t+2}}(U^{(t)}, \overline{V}^{(t+1)})$
- $(\overline{U}^{(t+1)}, R_{\overline{U}}^{(t+0.5)}) \leftarrow \mathsf{IF}(U^{(t+0.5)})$, $V^{(t+1)} \leftarrow \overline{V}^{(t+1)} R_{\overline{U}}^{(t+0.5)\top}$
- Gradient Descent : $U^{(t+0.5)} \leftarrow U^{(t)} - \eta \nabla_U \mathcal{F}_{\mathcal{W}_{2t+2}}(U^{(t)}, \overline{V}^{(t)})$
- $(\overline{U}^{(t+1)}, R_{\overline{U}}^{(t+0.5)}) \leftarrow \mathsf{IF}(U^{(t+0.5)})$, $V^{(t+1)} \leftarrow \overline{V}^{(t)} R_{\overline{U}}^{(t+0.5)\top}$

End for

Output: $M^{(T)} \leftarrow U^{(T-0.5)} \overline{V}^{(T)\top}$ (for gradient descent we use $\overline{U}^{(T)} V^{(T)\top}$)

The proof of Theorem 2 is provided in §F.1, §F.2, and §F.3. Theorem 2 implies that all three nonconvex optimization algorithms converge to the global optimum at a geometric rate. Furthermore, our results indicate that the completion of the true low rank matrix M^* up to ϵ-accuracy requires the entry observation probability $\bar{\rho}$ to satisfy

$$\bar{\rho} = \Omega(\mu^2 k^3 \log n \log(1/\epsilon)/m). \tag{54}$$

This result matches the result established by Hardt [2014], which is the state-of-the-art result for alternating minimization. Moreover, our analysis covers three nonconvex optimization algorithms.

In fact, the sample complexity in (54) depends on a polynomial of $\frac{\sigma_{\max}(M^*)}{\sigma_{\min}(M^*)}$, which is a constant since in this paper we assume that $\sigma_{\max}(M^*)$ and $\sigma_{\min}(M^*)$ are constants. If we allow $\frac{\sigma_{\max}(M^*)}{\sigma_{\min}(M^*)}$ to in-

Algorithm 3 The incoherence factorization algorithm for matrix completion. It guarantees that the solutions satisfy the incoherence condition throughout all iterations.

Input: W^{in}

$r \leftarrow$ Number of rows of W^{in}

Parameter: Incoherence parameter μ

$(\overline{W}^{\text{in}}, R^{\text{in}}_{\overline{W}}) \leftarrow \mathsf{QR}(W^{\text{in}})$

$\widetilde{W} \leftarrow \underset{W}{\operatorname{argmin}} \|W - \overline{W}^{\text{in}}\|_{\text{F}}^2$ subject to $\max_j \|W_{j*}\|_2 \leq \mu\sqrt{k/r}$

$(\overline{W}^{\text{out}}, R^{\text{tmp}}_{\overline{W}}) \leftarrow \mathsf{QR}(W^{\text{out}})$

$R^{\text{out}}_{\overline{W}} = \overline{W}^{\text{out}\top} W^{\text{in}}$

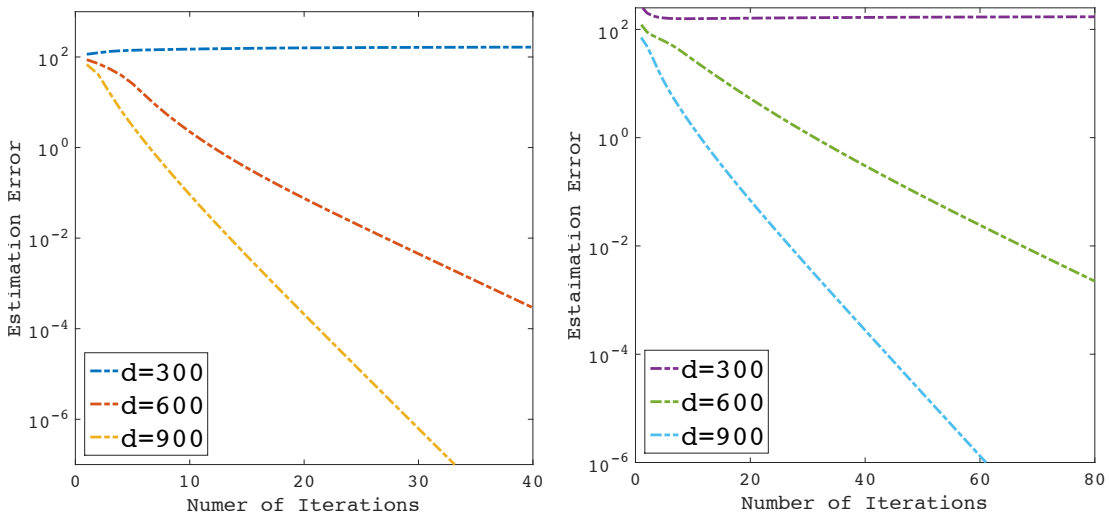
Output: $\overline{W}^{\text{out}}, R^{\text{out}}_{\overline{W}}$

crease, we can replace the QR decomposition in Algorithm 3 with the smooth QR decomposition proposed by Hardt and Wootters [2014] and achieve a dependency of $\log\left(\frac{\sigma_{\max}(M^*)}{\sigma_{\min}(M^*)}\right)$ on the condition number with a more involved proof. See more details in Hardt and Wootters [2014]. However, in this paper, our primary focus is on the dependency on k, n and m, rather than optimizing over the dependency on condition number.

6 Numerical Experiments

We present numerical experiments to support our theoretical analysis. We first consider a matrix sensing problem with $m = 30$, $n = 40$, and $k = 5$. We vary d from 300 to 900. Each entry of A_i's are independent sampled from $N(0,1)$. We then generate $M = UV^\top$, where $\widetilde{U} \in \mathbb{R}^{m\times k}$ and $\widetilde{V} \in \mathbb{R}^{n\times k}$ are two matrices with all their entries independently sampled from $N(0,1/k)$. We then generate d measurements by $b_i = \langle A_i, M\rangle$ for $i = 1,...,d$. Figure 1 illustrates the empirical performance of the alternating exact minimization and alternating gradient descent algorithms for a single realization. The step size for the alternating gradient descent algorithm is determined by the backtracking line search procedure. We see that both algorithms attain linear rate of convergence for $d = 600$ and $d = 900$. Both algorithms fail for $d = 300$, because $d = 300$ is below the minimum requirement of sample complexity for the exact matrix recovery.

We then consider a matrix completion problem with $m = 1000$, $n = 50$, and $k = 5$. We vary $\bar{\rho}$ from 0.025 to 0.1. We then generate $M = UV^\top$, where $\widetilde{U} \in \mathbb{R}^{m\times k}$ and $\widetilde{V} \in \mathbb{R}^{n\times k}$ are two matrices with all their entries independently sampled from $N(0,1/k)$. The observation set is generated uniformly at random with probability $\bar{\rho}$. Figure 2 illustrates the empirical performance of the alternating exact minimization and alternating gradient descent algorithms for a single realization. The step size for the alternating gradient descent algorithm is determined by the backtracking line search procedure. We see that both algorithms attain linear rate of convergence for $\bar{\rho} = 0.05$ and $\bar{\rho} = 0.1$. Both algorithms fail for $\bar{\rho} = 0.025$, because the entry observation probability is below



(a) Alternating Exact Minimization Algorithm (b) Alternating Gradient Descent Algorithm

Figure 1: Two illustrative examples for matrix sensing. The vertical axis corresponds to estimation error $\|M^{(t)} - M\|_{\mathrm{F}}$. The horizontal axis corresponds to numbers of iterations. Both the alternating exact minimization and alternating gradient descent algorithms attain linear rate of convergence for $d = 600$ and $d = 900$. But both algorithms fail for $d = 300$, because the sample size is not large enough to guarantee proper initial solutions.

the minimum requirement of sample complexity for the exact matrix recovery.

7 Conclusion

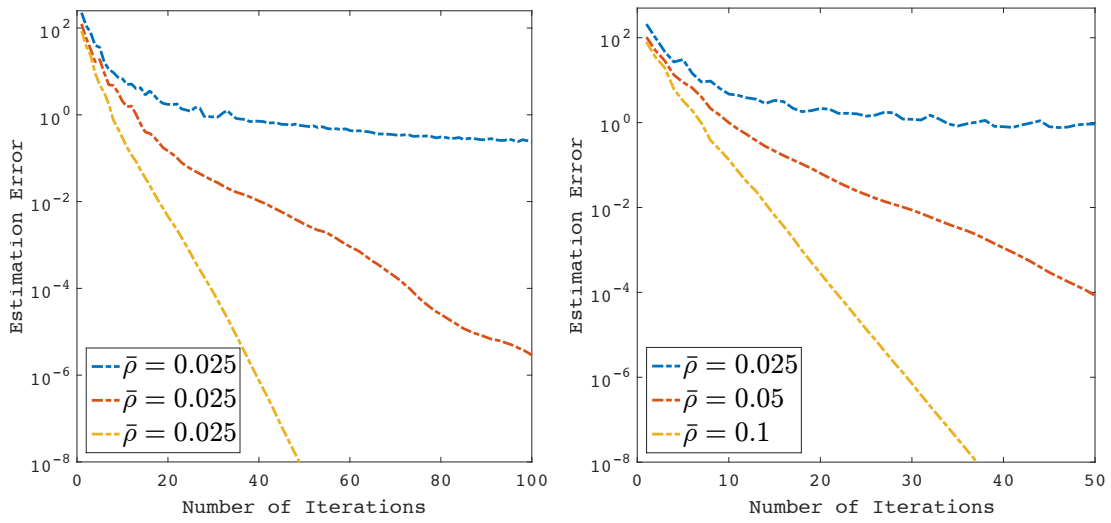
In this paper, we propose a generic analysis for characterizing the convergence properties of nonconvex optimization algorithms. By exploiting the inexact first order oracle, we prove that a broad class of nonconvex optimization algorithms converge geometrically to the global optimum and exactly recover the true low rank matrices under suitable conditions.

A Lemmas for Theorem 1 (Alternating Exact Minimization)

A.1 Proof of Lemma 1

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^{*}$ and V^{*} respectively. Then we define two $nk \times nk$ matrices

$$S^{(t)} = \begin{bmatrix} S_{11}^{(t)} & \cdots & S_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ S_{k1}^{(t)} & \cdots & S_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad S_{pq}^{(t)} = \sum_{i=1}^{d} A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{(t)\top} A_i^{\top},$$



(a) Alternating Exact Minimization Algorithm (b) Alternating Gradient Descent Algorithm

Figure 2: Two illustrative examples for matrix completion. The vertical axis corresponds to estimation error $\|M^{(t)} - M\|_{\mathrm{F}}$. The horizontal axis corresponds to numbers of iterations. Both the alternating exact minimization and alternating gradient descent algorithms attain linear rate of convergence for $\bar{\rho} = 0.05$ and $\bar{\rho} = 0.1$. But both algorithms fail for $\bar{\rho} = 0.025$, because the entry observation probability is not large enough to guarantee proper initial solutions.

$$G^{(t)} = \begin{bmatrix} G_{11}^{(t)} & \cdots & G_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ G_{k1}^{(t)} & \cdots & G_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad G_{pq}^{(t)} = \sum_{i=1}^{d} A_i \overline{U}_{*p}^{*} \overline{U}_{*q}^{*\top} A_i^{\top}$$

for $1 \leq p, q \leq k$. Note that $S^{(t)}$ and $G^{(t)}$ are essentially the partial Hessian matrices $\nabla_V^2 \mathcal{F}(\overline{U}^{(t)}, V)$ and $\nabla_V^2 \mathcal{F}(\overline{U}^{*}, V)$ for a vectorized V, i.e., $\mathrm{vec}(V) \in \mathbb{R}^{nk}$. Before we proceed with the main proof, we first introduce the following lemma.

Lemma 11. Suppose that $\mathcal{A}(\cdot)$ satisfies $2k$-RIP with parameter δ_{2k}. We then have

$$1 + \delta_{2k} \geq \sigma_{\max}(S^{(t)}) \geq \sigma_{\min}(S^{(t)}) \geq 1 - \delta_{2k}.$$

The proof of Lemma 11 is provided in Appendix A.7. Note that Lemma 11 is also applicable $G^{(t)}$, since $G^{(t)}$ shares the same structure with $S^{(t)}$.

We then proceed with the proof of Lemma 1. Given a fixed $\overline{U}$, $\mathcal{F}(\overline{U}, V)$ is a quadratic function of V. Therefore we have

$$\mathcal{F}(\overline{U}, V') = \mathcal{F}(\overline{U}, V) + \langle \nabla_V \mathcal{F}(\overline{U}, V), V' - V \rangle + \langle \mathrm{vec}(V') - \mathrm{vec}(V), \nabla_V^2 F(\overline{U}, V)(\mathrm{vec}(V') - \mathrm{vec}(V)) \rangle,$$

which further implies implies

$$\mathcal{F}(\overline{U}, V') - \mathcal{F}(\overline{U}, V) - \langle \nabla F_V(\overline{U}, V), V' - V \rangle \leq \sigma_{\max}(\nabla_V^2 F(\overline{U}, V)) \|V' - V\|_{\mathrm{F}}^2$$
$$\mathcal{F}(\overline{U}, V') - \mathcal{F}(\overline{U}, V) - \langle \nabla F_V(\overline{U}, V), V' - V \rangle \geq \sigma_{\min}(\nabla_V^2 F(\overline{U}, V)) \|V' - V\|_{\mathrm{F}}^2.$$

Then we can verify that $\nabla_V^2 F(U, V)$ also shares the same structure with $S^{(t)}$. Thus applying Lemma 11 to the above two inequalities, we complete the proof. □

A.2 Proof of Lemma 3

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^*$ and V^* respectively. We define two $nk \times nk$ matrices

$$J^{(t)} = \begin{bmatrix} J_{11}^{(t)} & \cdots & J_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ J_{k1}^{(t)} & \cdots & J_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad J_{pq}^{(t)} = \sum_{i=1}^{d} A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{*\top} A_i^\top,$$

$$K^{(t)} = \begin{bmatrix} K_{11}^{(t)} & \cdots & K_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ K_{k1}^{(t)} & \cdots & K_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad K_{pq}^{(t)} = \overline{U}_{*p}^{(t)\top} \overline{U}_{*q}^{*} I_n$$

for $1 \leq p, q \leq k$. Before we proceed with the main proof, we first introduce the following lemmas.

Lemma 12. Suppose that $\mathcal{A}(\cdot)$ satisfies $2k$-RIP with parameter δ_{2k}. We then have

$$\|S^{(t)} K^{(t)} - J^{(t)}\|_2 \leq 3\delta_{2k} \sqrt{k} \|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}.$$

The proof of Lemma 12 is provided in Appendix A.8. Note that Lemma 12 is also applicable to $G^{(t)} K^{(t)} - J^{(t)}$, since $G^{(t)}$ and $S^{(t)}$ share the same structure.

Lemma 13. Given $F \in \mathbb{R}^{k \times k}$, we define a $nk \times nk$ matrix

$$\mathbb{F} = \begin{bmatrix} F_{11} I_n & \cdots & F_{1k} I_n \\ \vdots & \ddots & \vdots \\ F_{k1} I_n & \cdots & F_{kk} I_n \end{bmatrix}.$$

For any $V \in \mathbb{R}^{n \times k}$, let $v = \mathrm{vec}(V) \in \mathbb{R}^{nk}$, then we have $\|\mathbb{F} v\|_2 = \|F V^\top\|_{\mathrm{F}}$.

Proof. By linear algebra, we have

$$[FV]_{ij} = F_{i*}^\top V_{j*} = \sum_{\ell=1}^{k} F_{i\ell} V_{j\ell} = \sum_{\ell=1}^{k} F_{i\ell} I_{*\ell}^\top V_{*\ell},$$

which completes the proof. □

We then proceed with the proof of Lemma 3. Since $b_i = \mathrm{tr}(V^{*\top} A_i U^*)$, then we rewrite $\mathcal{F}(\overline{U}, V)$ as

$$\mathcal{F}(\overline{U}, V) = \frac{1}{2}\sum_{i=1}^{d}\Big(\mathrm{tr}(V^\top A_i \overline{U}) - b_i\Big)^2 = \frac{1}{2}\sum_{i=1}^{d}\bigg(\sum_{j=1}^{k} V_{j*}^\top A_i \overline{U}_{*j} - \sum_{j=1}^{k} V_{j*}^{*\top} A_i \overline{U}^*_{*j}\bigg)^2.$$

For notational simplicity, we define $v = \mathrm{vec}(V)$. Since $V^{(t+0.5)}$ minimizes $\mathcal{F}(\overline{U}^{(t)}, V)$, we have

$$\mathrm{vec}\Big(\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t+0.5)})\Big) = S^{(t)} v^{(t+0.5)} - J^{(t)} v^* = 0.$$

Solving the above system of equations, we obtain

$$v^{(t+0.5)} = (S^{(t)})^{-1} J^{(t)} v^*. \tag{55}$$

Meanwhile, we have

$$\begin{aligned}\mathrm{vec}(\nabla_V \mathcal{F}(\overline{U}^*, V^{(t+0.5)})) &= G^{(t)} v^{(t+0.5)} - G^{(t)} v^* \\ &= G^{(t)}(S^{(t)})^{-1} J^{(t)} v^* - G^{(t)} v^* = G^{(t)}\Big((S^{(t)})^{-1} J^{(t)} - I_{nk}\Big) v^*,\end{aligned} \tag{56}$$

where the second equality come from (55). By triangle inequality, (56) further implies

$$\begin{aligned}\|((S^{(t)})^{-1} J^{(t)} - I_{nk}) v^*\|_2 &\le \|(K^{(t)} - I_{nk}) v^*\|_2 + \|(S^{(t)})^{-1}(J^{(t)} - S^{(t)} K^{(t)}) v^*\|_2 \\ &\le \|(\overline{U}^{(t)\top}\overline{U}^* - I_k) V^{*\top}\|_{\mathrm{F}} + \|(S^{(t)})^{-1}\|_2 \|(J^{(t)} - S^{(t)} K^{(t)}) v^*\|_2 \\ &\le \|\overline{U}^{(t)\top}\overline{U}^* - I_k\|_{\mathrm{F}} \|V^*\|_2 + \|(S^{(t)})^{-1}\|_2 \|(J^{(t)} - S^{(t)} K^{(t)}) v^*\|_2,\end{aligned} \tag{57}$$

where the second inequality comes from Lemma 13. Plugging (57) into (56), we have

$$\begin{aligned}&\|\mathrm{vec}(\nabla_V \mathcal{F}(\overline{U}^*, V^{(t+0.5)}))\|_2 \le \|G^{(t)}\|_2 \|((S^{(t)})^{-1} J^{(t)} - I_{nk}) v^*\|_2 \\ &\overset{\text{(i)}}{\le} (1+\delta_{2k})(\sigma_1 \|\overline{U}^{(t)\top}\overline{U}^* - I_k\|_2 + \|(S^{(t)})^{-1}\|_2 \|S^{(t)} K^{(t)} - J^{(t)}\|_2 \sigma_1 \sqrt{k}) \\ &\overset{\text{(ii)}}{\le} (1+\delta_{2k})\sigma_1 \bigg(\|(\overline{U}^{(t)} - \overline{U}^*)^\top (\overline{U}^{(t)} - \overline{U}^*)\|_{\mathrm{F}} + \frac{3\delta_{2k} k}{1-\delta_{2k}} \|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}\bigg) \\ &\overset{\text{(iii)}}{\le} (1+\delta_{2k})\sigma_1 \bigg(\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}^2 + \frac{3\delta_{2k} k}{1-\delta_{2k}} \|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}\bigg) \overset{\text{(iv)}}{\le} \frac{(1-\delta_{2k})\sigma_k}{2\xi} \|\overline{U}^* - \overline{U}^{(t)}\|_{\mathrm{F}},\end{aligned}$$

where (i) comes from Lemma 11 and $\|V^*\|_2 = \|M^*\| = \sigma_1$ and $\|V^*\|_{\mathrm{F}} = \|v^*\|_2 \le \sigma_1\sqrt{k}$, (ii) comes from Lemmas 11 and 12, (iii) from Cauchy-Schwartz inequality, and (iv) comes from (16). Since we have $\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t+0.5)}) = \mathbf{0}$, we further btain

$$\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}) \le \frac{(1-\delta_{2k})\sigma_k}{2\xi} \|\overline{U}^* - \overline{U}^{(t)}\|_{\mathrm{F}},$$

which completes the proof. □

A.3 Proof of Lemma 4

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^*$ and V^* respectively. By the strong convexity of $\mathcal{F}(\overline{U}^*, \cdot)$, we have

$$\mathcal{F}(\overline{U}^*, V^*) - \frac{1-\delta_{2k}}{2}\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2 \geq \mathcal{F}(\overline{U}^*, V^{(t+0.5)}) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t+0.5)}), V^* - V^{(t+0.5)} \rangle. \tag{58}$$

By the strong convexity of $\mathcal{F}(\overline{U}^*, \cdot)$ again, we have

$$\begin{aligned}\mathcal{F}(\overline{U}^*, V^{(t+0.5)}) &\geq \mathcal{F}(\overline{U}^*, V^*) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^*), V^* - V^{(t+0.5)} \rangle + \frac{1-\delta_{2k}}{2}\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2 \\ &\geq \mathcal{F}(\overline{U}^*, V^*) + \frac{1-\delta_{2k}}{2}\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2, \end{aligned} \tag{59}$$

where the last inequality comes from the optimality condition of $V^* = \operatorname{argmin}_V \mathcal{F}(\overline{U}^*, V)$, i.e.

$$\langle \nabla_V \mathcal{F}(\overline{U}^*, V^*), V^{(t+0.5)} - V^* \rangle \geq 0.$$

Meanwhile, since $V^{(t+0.5)}$ minimizes $\mathcal{F}(\overline{U}^{(t)}, \cdot)$, we have the optimality condition

$$\langle \nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t+0.5)}), V^* - V^{(t+0.5)} \rangle \geq 0,$$

which further implies

$$\begin{aligned}&\langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t+0.5)}), V^* - V^{(t+0.5)} \rangle \\ &\qquad \geq \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t+0.5)}) - \nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t+0.5)}), V^* - V^{(t+0.5)} \rangle. \end{aligned} \tag{60}$$

Combining (58) and (59) with (60), we obtain

$$\|V^{(t+0.5)} - V^*\|_2 \leq \frac{1}{1-\delta_{2k}} \mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}),$$

which completes the proof. □

A.4 Proof of Lemma 5

Proof. Before we proceed with the proof, we first introduce the following lemma.

Lemma 14. Suppose that $A^* \in \mathbb{R}^{n\times k}$ is a rank k matrix. Let $E \in \mathbb{R}^{n\times k}$ satisfy $\|E\|_2\|A^{*\dagger}\|_2 < 1$. Then given a QR decomposition $(A^* + E) = QR$, there exists a factorization of $A^* = Q^*O^*$ such that $Q^* \in \mathbb{R}^{n\times k}$ is an orthonormal matrix, and satisfies

$$\|Q - Q^*\|_{\mathrm{F}} \leq \frac{\sqrt{2}\|A^{*\dagger}\|_2\|E\|_{\mathrm{F}}}{1 - \|E\|_2\|A^{*\dagger}\|_2}.$$

The proof of Lemma 14 is provided in Stewart et al. [1990], therefore omitted.

We then proceed with the proof of Lemma 5. We consider $A^* = V^{*(t)}$ and $E = V^{(t+0.5)} - V^{*(t)}$ in Lemma 14 respectively. We can verify that

$$\|V^{(t+0.5)} - V^{*(t)}\|_2 \|V^{*(t)\dagger}\|_2 \leq \frac{\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}}{\sigma_k} \leq \frac{1}{4}.$$

Then there exists a $V^{*(t)} = \overline{V}^{*(t+1)} O^*$ such that $\overline{V}^{*(t+1)}$ is an orthonormal matrix, and satisfies

$$\|\overline{V}^{*(t+0.5)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq 2\|V^{*(t)\dagger}\|_2 \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{2}{\sigma_k} \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}.$$

□

A.5 Proof of Lemma 6

Proof. Before we proceed with the main proof, we first introduce the following lemma.

Lemma 15. Let $b = \mathcal{A}(M^*)$, M is a rank-k matrix, and $\mathcal{A}$ is a linear measurement operator that satisfies $2k$-RIP with constant $\delta_{2k} < 1/3$. Let $X^{(t+1)}$ be the $(t+1)$-th step iterate of SVP, then we have

$$\|\mathcal{A}(X^{(t+1)}) - b\|_2^2 \leq \|\mathcal{A}(M^*) - b\|_2^2 + 2\delta_{2k} \|\mathcal{A}(X^{(t)}) - b\|_2^2$$

The proof of Lemma 15 is provided in Jain et al. [2010], therefore omitted. We then explain the implication of Lemma 15. Jain et al. [2010] show that $X^{(t+1)}$ is obtained by taking a projected gradient iteration over $X^{(t)}$ using step size $\frac{1}{1+\delta_{2k}}$. Then taking $X^{(t)} = 0$, we have

$$X^{(t+1)} = \frac{\overline{U}^{(0)} \overline{\Sigma}^{(0)} \overline{V}^{(0)\top}}{1 + \delta_{2k}}.$$

Then Lemma 15 implies

$$\left\| \mathcal{A}\left(\frac{\overline{U}^{(0)} \overline{\Sigma}^{(0)} \overline{V}^{(0)\top}}{1 + \delta_{2k}} - M^* \right) \right\|_2^2 \leq 4\delta_{2k} \|\mathcal{A}(M^*)\|_2^2. \tag{61}$$

Since $\mathcal{A}(\cdot)$ satisfies $2k$-RIP, then (61) further implies

$$\left\| \frac{\overline{U}^{(0)} \overline{\Sigma}^{(0)} \overline{V}^{(0)\top}}{1 + \delta_{2k}} - M^* \right\|_{\mathrm{F}}^2 \leq 4\delta_{2k}(1 + 3\delta_{2k}) \|M^*\|_{\mathrm{F}}^2. \tag{62}$$

We then project each column of M^* into the subspace spanned by $\{\overline{U}_{*i}^{(0)}\}_{i=1}^k$, and obtain

$$\|\overline{U}^{(0)} \overline{U}^{(0)\top} M^* - M^*\|_{\mathrm{F}}^2 \leq 6\delta_{2k} \|M^*\|_{\mathrm{F}}^2.$$

Let $\overline{U}_\perp^{(0)}$ denote the orthonormal complement of $\overline{U}^{(0)}$, i.e.,

$$\overline{U}_\perp^{(0)\top}\overline{U}_\perp^{(0)} = I_{n-k} \quad \text{and} \quad \overline{U}_\perp^{(0)\top}\overline{U}^{(0)} = 0.$$

Then given a compact singular value decomposition of $M^* = \widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top}$, we have

$$\frac{6\delta_{2k}k\sigma_1^2}{\sigma_k^2} \geq \|(\overline{U}^{(0)}\overline{U}^{(0)\top} - I_n)\widetilde{U}^*\|_{\mathrm{F}}^2 = \|\overline{U}_\perp^{(0)\top}\widetilde{U}^*\|_{\mathrm{F}}^2.$$

Thus Lemma 2 guarantees that for $O^* = \mathrm{argmin}_{O^\top O = I_k}\|\overline{U}^{(0)} - \widetilde{U}^* O\|_{\mathrm{F}}$, we have

$$\|\overline{U}^{(0)} - \widetilde{U}^* O^*\|_{\mathrm{F}} \leq \sqrt{2}\|\overline{U}_\perp^{(0)\top}\widetilde{U}^*\|_{\mathrm{F}} \leq 2\sqrt{3\delta_{2k}k}\cdot\frac{\sigma_1}{\sigma_k}.$$

We define $\overline{U}^{*(0)} = \widetilde{U}^* O^*$. Then combining the above inequality with (18), we have

$$\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}.$$

Meanwhile, we define $V^{*(0)} = \widetilde{V}^*\widetilde{D}^* O^*$. Then we have $\overline{U}^{*(0)}V^{*(0)\top} = \widetilde{U}^* O O^{*\top}\widetilde{D}^*\widetilde{V}^* = M^*$. □

A.6 Proof of Corollary 1

Proof. Since (19) ensures that (16) of Lemma 3 holds, then we have

$$\begin{aligned}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} &\leq \frac{1}{1-\delta_{2k}}\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}) \overset{\text{(i)}}{\leq} \frac{1}{1-\delta_{2k}}\cdot\frac{(1-\delta_{2k})\sigma_k}{2\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\overset{\text{(ii)}}{\leq} \frac{1}{1-\delta_{2k}}\cdot\frac{(1-\delta_{2k})\sigma_k}{2\xi}\cdot\frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1} \leq \left(\frac{(1-\delta_{2k})\sigma_k}{8\xi^2(1+\delta_{2k})\sigma_1}\right)\sigma_k \overset{\text{(iii)}}{\leq} \frac{\sigma_k}{4},\end{aligned} \tag{63}$$

where (i) comes from Lemma 4, (ii) comes from (19), and (iii) comes from the definition of ξ and $\sigma_k \leq \sigma_1$. Since (63) ensures that (17) of Lemma 5 holds for $V^{(t+0.5)}$, then we obtain

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{2}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \overset{\text{(i)}}{\leq} \frac{1}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \overset{\text{(ii)}}{\leq} \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1}, \tag{64}$$

where (i) comes from (63), and (ii) comes from the definition of ξ and (19). □

A.7 Proof of Lemma 11

Proof. We consider an arbitrary $W \in \mathbb{R}^{n\times k}$ such that $\|W\|_{\mathrm{F}} = 1$. Let $w = \mathrm{vec}(W)$. Then we have

$$\begin{aligned}w^\top B w &= \sum_{p,q=1}^k W_{*p}^\top S_{pq}^{(t)} W_{*p} = \sum_{p,q=1}^k W_{*p}^\top\left(\sum_{i=1}^d A_i \overline{U}_{*p}^{(t)}\overline{U}_{*q}^{(t)\top}A_i^\top\right)W_{*q} \\ &= \sum_{i=1}^d\left(\sum_{p=1}^k W_{*p}^\top A_i\overline{U}_{*p}^{(t)}\right)\left(\sum_{q=1}^k W_{*q}^\top A_i\overline{U}_{*q}^{(t)}\right) = \sum_{i=1}^n \mathrm{tr}(W^\top A_i\overline{U}^{(t)})^2 = \|\mathcal{A}(\overline{U}^{(t)}W^\top)\|_2^2.\end{aligned}$$

Since $\mathcal{A}(\cdot)$ satisfies $2k$-RIP, then we have

$$\|\mathcal{A}(\overline{U}^{(t)}W^\top)\|_2^2 \geq (1-\delta_{2k})\|\overline{U}^{(t)}W^\top\|_{\mathrm{F}} = (1-\delta_{2k})\|W\|_{\mathrm{F}} = 1-\delta_{2k},$$
$$\|\mathcal{A}(\overline{U}^{(t)}W^\top)\|_2^2 \leq (1+\delta_{2k})\|\overline{U}^{(t)}W^\top\|_{\mathrm{F}} = (1+\delta_{2k})\|W\|_{\mathrm{F}} = 1+\delta_{2k}.$$

Since W is arbitrary, then we have

$$\sigma_{\min}(S^{(t)}) = \min_{\|w\|_2=1} w^\top S^{(t)} w \geq 1-\delta_{2k} \quad \text{and} \quad \sigma_{\max}(S^{(t)}) = \max_{\|w\|_2=1} w^\top S^{(t)} w \leq 1+\delta_{2k}.$$

□

A.8 Proof of Lemma 12

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^*$ and V^* respectively. Before we proceed with the main proof, we first introduce the following lemma.

Lemma 16. Suppose $\mathcal{A}(\cdot)$ satisfies $2k$-RIP. For any $U,\ U' \in \mathbb{R}^{m\times k}$ and $V,\ V' \in \mathbb{R}^{n\times k}$, we have

$$|\langle \mathcal{A}(UV^\top), \mathcal{A}(U'V'^\top)\rangle - \langle U^\top U', V^\top V'\rangle| \leq 3\delta_{2k}\|UV^\top\|_{\mathrm{F}} \cdot \|U'V'^\top\|_{\mathrm{F}}.$$

The proof of Lemma 16 is provided in Jain et al. [2013], and hence omitted.

We now proceed with the proof of Lemma 12. We consider arbitrary $W, Z \in \mathbb{R}^{n\times k}$ such that $\|W\|_{\mathrm{F}} = \|Z\|_{\mathrm{F}} = 1$. Let $w = \mathrm{vec}(W)$ and $z = \mathrm{vec}(Z)$. Then we have

$$w^\top(S^{(t)}K^{(t)} - J^{(t)})z = \sum_{p,q=1}^{k} W_{*p}^\top [S^{(t)}K^{(t)} - J^{(t)}]_{pq} Z_{*q}.$$

We consider a decomposition

$$\begin{aligned}
[S^{(t)}K^{(t)} - J^{(t)}]_{pq} &= \sum_{\ell=1}^{k} S_{p\ell}^{(t)} K_{\ell q}^{(t)} - J_{pq}^{(t)} = \sum_{\ell=1}^{k} S_{p\ell}^{(t)} \overline{U}_{*\ell}^{(t)\top} \overline{U}_{*q}^{*} I_n - J_{pq}^{(t)} \\
&= \sum_{\ell=1}^{k} \overline{U}_{*q}^{*\top} \overline{U}_{*\ell}^{(t)} \sum_{i=1}^{d} A_i \overline{U}_{*p}^{(t)} \overline{U}_{*\ell}^{(t)} A_i^\top - J_{pq}^{(t)} = \sum_{\ell=1}^{k} A_i \overline{U}_{*q}^{*\top} \overline{U}_{*\ell}^{(t)} \sum_{i=1}^{d} \overline{U}_{*p}^{(t)} \overline{U}_{*\ell}^{(t)} A_i^\top - \sum_{i=1}^{d} A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{*} A_i^\top \\
&= \sum_{i=1}^{d} A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{*} (\overline{U}^{(t)} \overline{U}^{(t)\top} - I_n) A_i^\top.
\end{aligned}$$

which further implies

$$
\begin{aligned}
w^\top (S^{(t)}K^{(t)} - J^{(t)})z &= \sum_{p,q} W_{*p}^\top \Big(\sum_{i=1}^d A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{*} (\overline{U}^{(t)}\overline{U}^{(t)\top} - I_m) A_i^\top \Big) Z_{*q} \\
&= \sum_{i=1}^d \sum_{p,q} W_{*p}^\top A_i \overline{U}_{*p}^{(t)} \overline{U}_{*q}^{*} (\overline{U}^{(t)}\overline{U}^{(t)\top} - I_m) A_i^\top Z_{*q} \\
&= \sum_{i=1}^d \mathrm{tr}(W^\top A_i \overline{U}^{(t)}) \mathrm{tr}\Big(Z^\top A_i (\overline{U}^{(t)}\overline{U}^{(t)\top} - I_m)\overline{U}^{*} \Big). \qquad (65)
\end{aligned}
$$

Since $\mathcal{A}(\cdot)$ satisfies $2k$-RIP, then by Lemma 16, we obtain

$$
\begin{aligned}
w^\top (S^{(t)}K^{(t)} - J^{(t)})z &\le \mathrm{tr}\Big(\overline{U}^{*} (\overline{U}^{(t)}\overline{U}^{(t)\top} - I_m)\overline{U}^{(t)} W^\top Z \Big) + 3\delta_{2k} \|\overline{U}^{(t)} W^\top\|_{\mathrm{F}} \|(\overline{U}^{(t)}\overline{U}^{(t)\top} - I_m)\overline{U}^{*} Z^\top\|_{\mathrm{F}} \\
&\overset{\text{(i)}}{\le} 3\delta_{2k} \|W\|_{\mathrm{F}} \sqrt{\|\overline{U}^{*\top} (\overline{U}^{(t)}\overline{U}^{(t)\top} - I_m)\overline{U}^{*}\|_{\mathrm{F}} \|Z^\top Z\|_{\mathrm{F}}}, \qquad (66)
\end{aligned}
$$

where the last inequality comes from $(\overline{U}^{(t)}\overline{U}^{(t)\top} - I_m)\overline{U}^{(t)} = 0$. Let $\overline{U}_\perp^{(t)} \in \mathbb{R}^{m-k}$ denote the orthogonal complement to $\overline{U}^{(t)}$ such that $\overline{U}^{(t)\top}\overline{U}_\perp^{(t)} = 0$ and $\overline{U}_\perp^{(t)\top}\overline{U}_\perp^{(t)} = I_{m-k}$. Then we have

$$
I_m - \overline{U}^{(t)}\overline{U}^{(t)\top} = \overline{U}_\perp^{(t)}\overline{U}_\perp^{(t)\top},
$$

which implies

$$
\begin{aligned}
\sqrt{\|\overline{U}^{*\top} (\overline{U}^{(t)}\overline{U}^{(t)\top} - I_m)\overline{U}^{*}\|_{\mathrm{F}}} &= \sqrt{\|\overline{U}^{*\top}\overline{U}_\perp^{(t)}\overline{U}_\perp^{(t)\top}\overline{U}^{*}\|_{\mathrm{F}}} \le \|\overline{U}_\perp^{(t)\top}\overline{U}^{*}\|_{\mathrm{F}} \\
&= \|\overline{U}_\perp^{(t)\top}\overline{U}^{(t)} - \overline{U}_\perp^{(t)\top}\overline{U}^{*}\|_{\mathrm{F}} \le \|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}. \qquad (67)
\end{aligned}
$$

Combining (66) with (67), we obtain

$$
w^\top (S^{(t)}K^{(t)} - J^{(t)})z \le 3\delta_{2k}\sqrt{k}\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}. \qquad (68)
$$

Since W and Z are arbitrary, then (68) implies

$$
\sigma_{\max}(S^{(t)}K^{(t)} - J^{(t)}) = \max_{\|w\|_2=1, \|z\|_2=1} w^\top (S^{(t)}K^{(t)} - J^{(t)})w \le 3\delta_{2k}\sqrt{k}\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}},
$$

which completes the proof.

□

B Lemmas for Theorem 1 (Alternating Gradient Descent)

B.1 Proof of Lemma 7

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^{*}$ and V^{*} respectively. We have

$$
\mathrm{vec}(\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)})) = S^{(t)}v^{(t)} - J^{(t)}v^{*} \ \text{ and } \ \mathrm{vec}(\nabla_V \mathcal{F}(\overline{U}^{*}, V^{(t)})) = G^{(t)}v^{(t)} - G^{(t)}v^{*}.
$$

Therefore, we further obtain

$$\begin{aligned}
&\|\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)}) - \nabla_V \mathcal{F}(\overline{U}^{*}, V^{(t)})\|_{\mathrm{F}} \\
&\quad = \|(S^{(t)} - J^{(t)})(v^{(t)} - v^{*}) + (S^{(t)} - J^{(t)})v^{*} + (J^{(t)} - G^{(t)})(v^{(t)} - v^{*})\|_2 \\
&\quad \le \|(S^{(t)} - J^{(t)})(v^{(t)} - v^{*})\|_2 + \|(S^{(t)} - J^{(t)})v^{*}\|_2 + \|(J^{(t)} - G^{(t)})(v^{(t)} - v^{*})\|_2 \\
&\quad \le \|S^{(t)}\|_2 \cdot \|((S^{(t)})^{-1}J^{(t)} - I_{nk})(v^{(t)} - v^{*})\|_2 + \|S^{(t)}\|_2 \cdot \|((S^{(t)})^{-1}J^{(t)} - I_{nk})v^{*}\|_2 \\
&\qquad\qquad + \|G\|_2 \cdot \|((G^{(t)})^{-1}J^{(t)} - I_{nk})(v^{(t)} - v^{*})\|_2. \qquad (69)
\end{aligned}$$

Recall that Lemma 12 is also applicable to $G^{(t)}K^{(t)} - J^{(t)}$. Since we have

$$\|V^{(t)} - V^{*}\|_2 \le \|V^{(t)} - V^{*}\|_{\mathrm{F}} = \|v^{(t)} - v^{*}\|_2 \le \sigma_1,$$

following similar lines to Appendix A.2, we can show

$$\begin{aligned}
\|((S^{(t)})^{-1}J^{(t)} - I_{mn})v^{*}\|_2 &\le \sigma_1\bigg(\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}^2 + \frac{3\delta_{2k}k}{1-\delta_{2k}}\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}\bigg), \\
\|((G^{(t)})^{-1}J^{(t)} - I_{mn})(v^{(t)} - v^{*})\|_2 &\le \sigma_1\bigg(\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}^2 + \frac{3\delta_{2k}k}{1-\delta_{2k}}\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}\bigg), \\
\|((S^{(t)})^{-1}J^{(t)} - I_{mn})(v^{(t)} - v^{*})\|_2 &\le \sigma_1\bigg(\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}^2 + \frac{3\delta_{2k}k}{1-\delta_{2k}}\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}\bigg).
\end{aligned}$$

Combining the above three inequalities with (69), we have

$$\begin{aligned}
&\|\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)}) - \nabla_V \mathcal{F}(\overline{U}^{*}, V^{(t)})\|_{\mathrm{F}} \\
&\qquad \le 2(1+\delta_{2k})\sigma_1\bigg(\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}^2 + \frac{3\delta_{2k}k}{1-\delta_{2k}}\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}}\bigg). \qquad (70)
\end{aligned}$$

Since $\overline{U}^{(t)}$, δ_{2k}, and ξ satisfy (27), then (70) further implies

$$\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) = \|\nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)}) - \nabla_V \mathcal{F}(\overline{U}^{*}, V^{(t)})\|_{\mathrm{F}} \le \frac{(1+\delta_{2k})\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*}\|_{\mathrm{F}},$$

which completes the proof. □

B.2 Proof of Lemma 8

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^{*}$ and V^{*} respectively. By the strong convexity of $\mathcal{F}(\overline{U}^{*}, \cdot)$, we have

$$\begin{aligned}
&\mathcal{F}(\overline{U}^{*}, V^{*}) - \frac{1-\delta_{2k}}{2}\|V^{(t)} - V^{*}\|_{\mathrm{F}}^2 \ge \mathcal{F}(\overline{U}^{*}, V^{(t)}) + \langle \nabla_V \mathcal{F}(\overline{U}^{*}, V^{(t)}), V^{*} - V^{(t)} \rangle \\
&= \mathcal{F}(\overline{U}^{*}, V^{(t)}) + \langle \nabla_V \mathcal{F}(\overline{U}^{*}, V^{(t)}), V^{(t+0.5)} - V^{(t)} \rangle + \langle \nabla_V \mathcal{F}(\overline{U}^{*}, V^{(t)}), V^{*} - V^{(t+0.5)} \rangle. \qquad (71)
\end{aligned}$$

Meanwhile, we define

$$\mathcal{Q}(V; \overline{U}^{*}, V^{(t)}) = \mathcal{F}(\overline{U}^{*}, V^{(t)}) + \langle \nabla_V \mathcal{F}(\overline{U}^{*}, V^{(t)}), V - V^{(t)} \rangle + \frac{1}{2\eta}\|V - V^{(t)}\|_{\mathrm{F}}^2.$$

Since η satisfies (28) and $\mathcal{F}(\overline{U}^*, V)$ is strongly smooth in V for a fixed orthonormal $\overline{U}^*$, we have

$$\mathcal{Q}(V; \overline{U}^*, V^{(t)}) \geq \mathcal{F}(\overline{U}^*, V^{(t)}).$$

Combining the above two inequalities, we obtain

$$\begin{aligned}\mathcal{F}(\overline{U}^*, V^{(t)}) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)}), V^{(t+0.5)} - V^{(t)} \rangle &= \mathcal{Q}(V^{(t+0.5)}; \overline{U}^*, V^{(t)}) - \frac{1}{2\eta}\|V^{(t+0.5)} - V^{(t)}\|_{\mathrm{F}}^2 \\ &\geq \mathcal{F}(\overline{U}^*, V^{(t+0.5)}) - \frac{1}{2\eta}\|V^{(t+0.5)} - V^{(t)}\|_{\mathrm{F}}^2. \end{aligned} \tag{72}$$

Moreover, by the strong convexity of $\mathcal{F}(\overline{U}^*, \cdot)$ again, we have

$$\begin{aligned}\mathcal{F}(\overline{U}^*, V^{(t+0.5)}) &\geq \mathcal{F}(\overline{U}^*, V^*) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^*), V^{(t+0.5)} - V^* \rangle + \frac{1-\delta_{2k}}{2}\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2 \\ &\geq \mathcal{F}(\overline{U}^*, V^*) + \frac{1-\delta_{2k}}{2}\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2, \end{aligned} \tag{73}$$

where the second equalities comes from the optimality condition of $V^* = \operatorname{argmin}_V \mathcal{F}(\overline{U}^*, V)$, i.e.

$$\langle \nabla_V \mathcal{F}(\overline{U}^*, V^*), V^{(t+0.5)} - V^* \rangle \geq 0.$$

Combining (71) and (72) with (73), we obtain

$$\begin{aligned}&\mathcal{F}(\overline{U}^*, V^{(t)}) + \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)}), V^{(t+0.5)} - V^{(t)} \rangle \\ &\qquad \geq \mathcal{F}(\overline{U}^*, V^*) + \frac{1-\delta_{2k}}{2}\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2 - \frac{1}{2\eta}\|V^{(t+0.5)} - V^{(t)}\|_{\mathrm{F}}^2. \end{aligned} \tag{74}$$

On the other hand, since $V^{(t+0.5)}$ minimizes $\mathcal{Q}(V; \overline{U}^*, V^{(t)})$, we have

$$\begin{aligned}0 &\leq \langle \nabla \mathcal{Q}(V^{(t+0.5)}; \overline{U}^*, V^{(t)}), V^* - V^{(t+0.5)} \rangle \\ &\leq \langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)}), V^* - V^{(t+0.5)} \rangle + (1+\delta_{2k}) \langle V^{(t+0.5)} - V^{(t)}, V^* - V^{(t+0.5)} \rangle. \end{aligned} \tag{75}$$

Meanwhile, we have

$$\begin{aligned}&\langle \nabla_V \mathcal{F}(\overline{U}^*, V^{(t)}), V^* - V^{(t+0.5)} \rangle \\ &\quad = \langle \nabla_V \mathcal{F}(\overline{U}^{(t)}, V^{(t)}), V^* - V^{(t+0.5)} \rangle - \mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \|V^* - V^{(t+0.5)}\|_2 \\ &\quad \geq (1+\delta_{2k}) \langle V^{(t)} - V^{(t+0.5)}, V^* - V^{(t+0.5)} \rangle - \mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \|V^* - V^{(t+0.5)}\|_2 \\ &\quad = (1+\delta_{2k}) \langle V^{(t)} - V^{(t+0.5)}, V^* - V^{(t)} \rangle + \frac{1}{2\eta}\|V^{(t)} - V^{(t+0.5)}\|_{\mathrm{F}}^2 \\ &\qquad\qquad - \mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \|V^* - V^{(t+0.5)}\|_2. \end{aligned} \tag{76}$$

Combining (75) with (76), we obtain

$$\begin{aligned}2\langle V^{(t)} - V^{(t+0.5)}, V^* - V^{(t)} \rangle &\leq -\eta(1-\delta_{2k})\|V^{(t)} - V^*\|_2^2 - \eta(1-\delta_{2k})\|V^{(t+0.5)} - V^*\|_2^2 \\ &\quad - \|V^{(t+0.5)} - V^{(t)}\|_2^2 + \mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \|V^* - V^{(t+0.5)}\|_2. \end{aligned} \tag{77}$$

Therefore, combining (74) with (77), we obtain

$$\begin{aligned}\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2 &\le \|V^{(t+0.5)} - V^{(t)} + V^{(t)} - V^*\|_{\mathrm{F}}^2 \\ &= \|V^{(t+0.5)} - V^{(t)}\|_{\mathrm{F}}^2 + \|V^{(t)} - V^*\|_{\mathrm{F}}^2 + 2\langle V^{(t+0.5)} - V^{(t)}, V^{(t)} - V^*\rangle \\ &\le 2\eta\|V^{(t)} - V^*\|_{\mathrm{F}}^2 - \eta(1-\delta_{2k})\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}^2 \\ &\qquad - \mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)})\|V^* - V^{(t+0.5)}\|_2.\end{aligned}$$

Rearranging the above inequality, we obtain

$$\|V^{(t+0.5)} - V^*\|_{\mathrm{F}} \le \sqrt{\delta_{2k}}\|V^{(t)} - V^*\|_{\mathrm{F}} + \frac{2}{1+\delta_{2k}}\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}),$$

which completes the proof. □

B.3 Proof of Lemma 9

Proof. Before we proceed with the main proof, we first introduce the following lemma.

Lemma 17. For any matrix $U, \widetilde{U} \in \mathbb{R}^{m\times k}$ and $V, \widetilde{V} \in \mathbb{R}^{n\times k}$, we have

$$\|UV^\top - \widetilde{U}\widetilde{V}^\top\|_{\mathrm{F}} \le \|U\|_2\|V - \widetilde{V}\| + \|\widetilde{V}\|_2\|U - \widetilde{U}\|_{\mathrm{F}}.$$

Proof. By linear algebra, we have

$$\begin{aligned}\|UV^\top - \widetilde{U}\widetilde{V}^\top\|_{\mathrm{F}} &= \|UV^\top - U\widetilde{V}^\top + U\widetilde{V}^\top - \widetilde{U}\widetilde{V}^\top\|_{\mathrm{F}} \\ &\le \|UV^\top - U\widetilde{V}^\top\|_{\mathrm{F}} + \|U\widetilde{V}^\top - \widetilde{U}\widetilde{V}^\top\|_{\mathrm{F}} \le \|U\|_2\|V - \widetilde{V}\|_{\mathrm{F}} + \|\widetilde{V}\|_2\|U - \widetilde{U}\|_{\mathrm{F}}. \qquad (78)\end{aligned}$$

□

We then proceed with the proof of Lemma 9. By Lemma 17, we have

$$\begin{aligned}\|R_{\overline{V}}^{(t+0.5)} - \overline{V}^{*(t+1)\top}V^{*(t)}\|_{\mathrm{F}} &= \|\overline{V}^{(t+0.5)\top}V^{(t+0.5)} - \overline{V}^{*(t+1)\top}V^{*(t)}\|_{\mathrm{F}} \\ &\le \|\overline{V}^{(t+0.5)}\|_2\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \|V^{*(t)}\|_2\|\overline{V}^{(t+0.5)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \\ &\le \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2\sigma_1}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}, \qquad (79)\end{aligned}$$

where the last inequality comes from Lemma 5. Moreover, we define $U^{*(t+1)} = \overline{U}^{*(t)}(\overline{V}^{*(t+1)\top}V^{*(t)})^\top$. Then we can verify

$$U^{*(t+1)}\overline{V}^{*(t+1)} = \overline{U}^{*(t)}V^{*(t)\top}\overline{V}^{*(t+1)}\overline{V}^{*(t+1)\top} = M^*\overline{V}^{*(t+1)}\overline{V}^{*(t+1)\top} = M^*,$$

where the last equality holds, since $\overline{V}^{*(t+1)}\overline{V}^{*(t+1)\top}$ is exactly the projection matrix for the row space of M^*. Thus by Lemma 17, we have

$$\begin{aligned}\|U^{(t+1)} - U^{*(t+1)}\|_{\mathrm{F}} &= \|\overline{U}^{(t)}R_{\overline{V}}^{(t+0.5)\top} - \overline{U}^{*(t)}(\overline{V}^{*(t+1)\top}V^{*(t)})^\top\|_{\mathrm{F}} \\ &\le \|\overline{U}^{(t)}\|_2\|R_{\overline{V}}^{(t+0.5)} - \overline{V}^{*(t+1)\top}V^{*(t)}\|_{\mathrm{F}} + \|\overline{V}^{*(t+1)\top}V^{*(t)}\|_2\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\ &\le \left(1 + \frac{2\sigma_1}{\sigma_k}\right)\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}},\end{aligned}$$

where the last inequality comes from (79), $\|\overline{V}^{*(t+1)}\|_2 = 1$, $\|\overline{U}^{(t)}\|_2 = 1$, and $\|V^{*(t)}\|_2 = \sigma_1$. □

B.4 Proof of Lemma 10

Proof. Following similar lines to Appendix A.5, we have

$$\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{4\xi\sigma_1^2}. \tag{80}$$

In Appendix A.5, we have already shown

$$\left\|\frac{\overline{U}^{(0)}\overline{\Sigma}^{(0)}\overline{V}^{(0)\top}}{1+\delta_{2k}} - M^*\right\|_{\mathrm{F}} \le 2\sqrt{\delta_{2k}(1+3\delta_{2k})}\|\overline{\Sigma}^*\|_{\mathrm{F}}. \tag{81}$$

Then by Lemma 17 we have

$$\begin{aligned}
\left\|\frac{\overline{U}^{(0)}\overline{\Sigma}^{(0)}}{1+\delta_{2k}} - V^{*(0)}\right\|_{\mathrm{F}} &= \left\|\frac{\overline{U}^{(0)\top}\overline{U}^{(0)}\overline{\Sigma}^{(0)}\overline{V}^{(0)\top}}{1+\delta_{2k}} - \overline{U}^{*(0)\top}M^*\right\|_{\mathrm{F}} \\
&\le \|\overline{U}^{(0)}\|_2\left\|\frac{\overline{U}^{(0)}\overline{\Sigma}^{(0)}\overline{V}^{(0)\top}}{1+\delta_{2k}} - M^*\right\|_{\mathrm{F}} + \|M^*\|_2\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \\
&\le 2\sqrt{\delta_{2k}k(1+3\delta_{2k})}\sigma_1 + \frac{\sigma_k^2}{4\xi\sigma_1^1},
\end{aligned} \tag{82}$$

where the last inequality comes from (80), (81), $\|M^*\|_2 = \sigma_1$, and $\|\overline{U}^{(0)}\|_2 = 1$. By triangle inequality, we further have

$$\begin{aligned}
\|\overline{U}^{(0)}\overline{\Sigma}^{(0)} - V^{*(0)}\|_{\mathrm{F}} &\le (1+\delta_{2k})\left\|\frac{\overline{U}^{(0)}\overline{\Sigma}^{(0)}}{1+\delta_{2k}} - V^{*(0)}\right\|_{\mathrm{F}} + \delta_{2k}\|V^{*(0)}\|_{\mathrm{F}} \\
&\overset{\text{(i)}}{\le}(1+\delta_{2k})\left(2\sqrt{\delta_{2k}k(1+3\delta_{2k})}\sigma_1 + \frac{\sigma_k^2}{4\xi\sigma_1}\right) + \delta_{2k}\sigma_1\sqrt{k} \\
&\overset{\text{(ii)}}{\le}\left(\frac{\sigma_k^3}{9\sigma_1^3\xi} + \frac{\sigma_k^2}{3\sigma_1^3\xi^2} + \frac{\sigma_k^3}{192\xi^3\sigma_1^2}\right)\sigma_1 \overset{\text{(iii)}}{\le} \frac{\sigma_k^2}{2\xi\sigma_1},
\end{aligned}$$

where (i) comes from (82) and $\|V^{*(0)}\|_{\mathrm{F}} = \|M^*\|_{\mathrm{F}} \le \sigma_1\sqrt{k}$, (ii) comes from (30), and (iii) comes from the definition of ξ and $\sigma_1 \ge \sigma_k$. □

B.5 Proof of Corollary 3

Proof. Since (31) ensures that (27) of Lemma 7 holds, we have

$$\begin{aligned}\|V^{(t+0.5)}-V^{*(t)}\|_{\mathrm{F}} &\le \sqrt{\delta_{2k}}\|V^{(t)}-V^{*(t)}\|_{\mathrm{F}}+\frac{2}{1+\delta_{2k}}\mathcal{E}(V^{(t+0.5)},V^{(t)},\overline{U}^{(t)})\\ &\overset{\text{(i)}}{\le}\sqrt{\delta_{2k}}\|V^{(t)}-V^{*(t)}\|_{\mathrm{F}}+\frac{2}{1+\delta_{2k}}\cdot\frac{(1+\delta_{2k})\sigma_k}{\xi}\|\overline{U}^{(t)}-\overline{U}^{*(t)}\|_{\mathrm{F}}\\ &\overset{\text{(ii)}}{\le}\frac{\sigma_k^2}{12\xi\sigma_1^2}\|V^{(t)}-V^{*(t)}\|_{\mathrm{F}}+\frac{2\sigma_k}{\xi}\|\overline{U}^{(t)}-\overline{U}^{*(t)}\|_{\mathrm{F}}\\ &\overset{\text{(iii)}}{\le}\frac{\sigma_k^2}{12\xi\sigma_1^2}\cdot\frac{\sigma_k^2}{2\xi\sigma_1}+\frac{2\sigma_k}{\xi}\cdot\frac{\sigma_k^2}{4\xi\sigma_1^2}\overset{\text{(iv)}}{\le}\frac{13\sigma_k^3}{24\xi^2\sigma_1^2}\overset{\text{(v)}}{\le}\frac{\sigma_k}{4},\end{aligned}\tag{83}$$

where (i) comes from Lemma 8, (ii) and (iii) come from (31), and (iv) and (v) come from the definition of ξ and $\sigma_k\le\sigma_1$. Since (83) ensures that (17) of Lemma 5, then we obtain

$$\begin{aligned}\|\overline{V}^{(t+1)}-\overline{V}^{*(t+1)}\|_{\mathrm{F}} &\le \frac{2}{\sigma_k}\|V^{(t+0.5)}-V^{*(t)}\|_{\mathrm{F}}\overset{\text{(i)}}{\le}\frac{2\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)}-V^{*(t)}\|_{\mathrm{F}}+\frac{4}{\xi}\|\overline{U}^{(t)}-\overline{U}^{*(t)}\|_{\mathrm{F}}\\ &\overset{\text{(ii)}}{\le}\left(\frac{\sigma_k}{3\xi\sigma_1}+\frac{4}{\xi}\right)\cdot\frac{\sigma_k^2}{4\xi\sigma_1^2}\overset{\text{(iii)}}{\le}\frac{\sigma_k^2}{4\xi\sigma_1^2},\end{aligned}\tag{84}$$

where (i) and (ii) come from (83), and (iii) comes from the definition of ξ and $\sigma_1>\sigma_k$. Moreover, since (83) ensures that (29) of Lemma 9 holds, then we have

$$\begin{aligned}\|U^{(t)}-U^{*(t+1)}\|_{\mathrm{F}} &\le \frac{3\sigma_1}{\sigma_k}\|V^{(t+0.5)}-V^{*(t)}\|_{\mathrm{F}}+\sigma_1\|\overline{U}^{(t)}-\overline{U}^{*(t)}\|_{\mathrm{F}}\\ &\overset{\text{(i)}}{\le}\frac{3\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)}-V^{*(t)}\|_{\mathrm{F}}+\left(\frac{6}{\xi}+1\right)\sigma_1\|\overline{U}^{(t)}-\overline{U}^{*(t)}\|_{\mathrm{F}}\\ &\overset{\text{(ii)}}{\le}\frac{3\sigma_1}{\sigma_k}\cdot\frac{\sigma_k^3}{12\xi\sigma_1^3}\cdot\frac{\sigma_k^2}{2\xi\sigma_1}+\left(\frac{6}{\xi}+1\right)\cdot\frac{\sigma_k^2}{4\xi\sigma_1}=\left(\frac{\sigma_k^2}{4\xi^2\sigma_1^2}+\frac{3}{\xi}+\frac{1}{2}\right)\frac{\sigma_k^2}{2\xi\sigma_1}\overset{\text{(iii)}}{\le}\frac{\sigma_k^2}{2\xi\sigma_1},\end{aligned}$$

where (i) comes from (83), (ii) comes from (31), and (iii) comes from the definition of ξ and $\sigma_1\ge\sigma_k$. □

C Partition Algorithm for Matrix Computation

Algorithm 4 The observation set partition algorithm for matrix completion. It guarantees the independence among all $2T+1$ output observation sets.

Input: $\mathcal{W}$, $\bar{\rho}$

$\widetilde{\rho} = 1 - (1-\bar{\rho})^{\frac{1}{2T+1}}$.

For: $t = 0, ..., 2T$

$$\widetilde{\rho}_t = \frac{(mn)!\bar{\rho}^{t+1}(1-\bar{\rho})^{mn-t-1}}{\bar{\rho}(mn-t-1)!(t+1)!}$$

End for

$\mathcal{W}_0 = \emptyset$, ..., $\mathcal{W}_{2T} = \emptyset$

For every $(i,j) \in \mathcal{W}$

Sample t from $\{0, ..., 2T\}$ with probability $\{\widetilde{\rho}_0, ..., \widetilde{\rho}_{2T}\}$

Sample (w/o replacement) a set $\mathcal{B}$ such that $|\mathcal{B}| = t$ from $\{0, ..., 2T\}$ with equal probability

Add (i,j) to $\mathcal{W}_\ell$ for all $\ell \in \mathcal{B}$

End for

Output: $\{\mathcal{W}_t\}_{t=0}^{2T}$, $\widetilde{\rho}$

D Initialization Procedures for Matrix Computation

Algorithm 5 The initialization procedure $\mathsf{INT}_{\overline{U}}(\cdot)$ for matrix completion. It guarantees that the initial solutions satisfy the incoherence condition throughout all iterations.

Input: $\widetilde{M}$

Parameter: Incoherence parameter μ

$(\widetilde{U}, \widetilde{D}, \widetilde{V}) \leftarrow \mathsf{KSVD}(\widetilde{M})$

$\widetilde{U}^{\text{tmp}} \leftarrow \underset{U}{\operatorname{argmin}} \|U - \widetilde{U}\|_{\text{F}}^2$ subject to $\max_i \|U_{i*}\|_2 \le \mu\sqrt{k/m}$

$(\overline{U}^{\text{out}}, R_{\overline{U}}^{\text{out}}) \leftarrow \mathsf{QR}(\widetilde{U}^{\text{tmp}})$

$\widetilde{V}^{\text{tmp}} \leftarrow \underset{V}{\operatorname{argmin}} \|V - \widetilde{V}^{\text{tmp}}\|_{\text{F}}^2$ subject to $\max_j \|V_{j*}\|_2 \le \mu\sqrt{k/n}$

$(\overline{V}^{\text{out}}, R_{\overline{V}}^{\text{out}}) \leftarrow \mathsf{QR}(\widetilde{V}^{\text{tmp}})$

$V^{\text{out}} = \overline{V}^{\text{out}}(\overline{U}^{\text{out}\top}\widetilde{M}\overline{V}^{\text{out}})^\top$

Output: $\overline{U}^{\text{out}}$, V^{out}

Algorithm 6 The initialization procedure $\mathsf{INT}_{\overline{V}}(\cdot)$ for matrix completion. It guarantees that the initial solutions satisfy the incoherence condition throughout all iterations.

Input: $\widetilde{M}$
Parameter: Incoherence parameter μ
$(\widetilde{U}, \widetilde{D}, \widetilde{V}) \leftarrow \mathsf{KSVD}(\widetilde{M})$
$\widetilde{V}^{\mathrm{tmp}} \leftarrow \underset{V}{\operatorname{argmin}} \|V - \widetilde{V}\|_{\mathrm{F}}^2$ subject to $\max_j \|V_{j*}\|_2 \le \mu\sqrt{k/n}$
$(\overline{V}^{\mathrm{out}}, R_{\overline{V}}^{\mathrm{out}}) \leftarrow \mathsf{QR}(\widetilde{V}^{\mathrm{tmp}})$
$\widetilde{U}^{\mathrm{tmp}} \leftarrow \underset{U}{\operatorname{argmin}} \|U - \widetilde{U}^{\mathrm{tmp}}\|_{\mathrm{F}}^2$ subject to $\max_i \|U_{i*}\|_2 \le \mu\sqrt{k/m}$
$(\overline{U}^{\mathrm{out}}, R_{\overline{U}}^{\mathrm{out}}) \leftarrow \mathsf{QR}(\widetilde{U}^{\mathrm{tmp}})$
$U^{\mathrm{out}} = \overline{U}^{\mathrm{out}}(\overline{U}^{\mathrm{out}\top} \widetilde{M} \overline{V}^{\mathrm{out}})$
Output: $\overline{V}^{\mathrm{out}}$, U^{out}

E Lemmas for Theorem 2 (Alternating Exact Minimization)

E.1 Proof of Lemma 21

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^{*}$ and V^{*} respectively. Before we proceed with the main proof, we first introduce the following lemma.

Lemma 18. Suppose that $\widetilde{\rho}$ satisfies (93). For any $z \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$ such that $\sum_i z_i = 0$, and a $t \in \{0, ..., 2T\}$, there exists a universal constant C such that

$$\sum_{(i,j)\in\mathcal{W}_t} z_i w_j \le C m^{1/4} n^{1/4} \widetilde{\rho}^{1/2} \|z\|_2 \|w\|_2$$

with probability at least $1 - n^{-3}$.

The proof of Lemma 18 is provided in Keshavan et al. [2010a], and therefore omitted.

We then proceed with the proof of Lemma 21. For $j = 1, ..., k$, we define $S^{(j,t)}$, $J^{(j,t)}$, and $K^{(j,t)}$ as

$$S^{(j,t)} = \frac{1}{\widetilde{\rho}} \sum_{i:(i,j)\in\mathcal{W}_{2t+1}} \overline{U}_{i*}^{(t)} \overline{U}_{i*}^{(t)\top}, \quad J^{(j,t)} = \frac{1}{\widetilde{\rho}} \sum_{i:(i,j)\in\mathcal{W}_{2t+1}} \overline{U}_{i*}^{(t)} \overline{U}_{i*}^{*\top}, \quad \text{and} \quad K^{(j,t)} = \overline{U}^{(t)\top} \overline{U}^{*}.$$

We then consider an arbitrary $W \in \mathbb{R}^{n\times k}$ such that $\|W\|_{\mathrm{F}} = 1$ and $w = \mathrm{vec}(W)$. Then we have

$$\max_j \sigma_{\max}(S^{(j,t)}) \ge w^\top S^{(t)} w = \sum_{j=1} W_{j*}^\top S^{(j,t)} W_{j*} \ge \min_j \sigma_{\min}(S^{(j,t)}). \tag{85}$$

Since $\mathcal{W}_{2t+1}$ is drawn uniformly at random, we can use mn independent Bernoulli random variables δ_{ij}'s to describe $\mathcal{W}_{2t+1}$, i.e., $\delta_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\widetilde{\rho})$ with $\delta_{ij} = 1$ denoting $(i,j) \in \mathcal{W}_{2t+1}$ and 0

denoting $(i,j) \notin \mathcal{W}_{2t+1}$. We then consider an arbitrary $z \in \mathbb{R}^k$ with $\|z\|_2 = 1$, and define

$$Y = z^\top S^{(j,t)} z = \frac{1}{\widetilde{\rho}} \sum_{i:(i,j)\in\mathcal{W}_{2t+1}} \Big(z^\top \overline{U}_{i*}^{(t)}\Big)^2 = \frac{1}{\widetilde{\rho}} \sum_i \delta_{ij} \Big(z^\top \overline{U}_{i*}^{(t)}\Big)^2.$$

Then we have

$$\mathbb{E}Y = z^\top \overline{U}^{(t)} \overline{U}^{(t)\top} z = 1 \quad \text{and} \quad \mathbb{E}Y^2 = \frac{1}{\widetilde{\rho}} \sum_i \delta_{ij} \Big(z^\top \overline{U}_{i*}^{(t)}\Big)^4 \le \frac{4\mu^2 k}{m\widetilde{\rho}} \sum_i \Big(z^\top \overline{U}_{i*}^{(t)}\Big)^2 = \frac{4\mu^2 k}{m\widetilde{\rho}},$$

where the last inequality holds, since $\overline{U}^{(t)}$ is incoherent with parameter 2μ. Similarly, we can show

$$\max_i \Big(z^\top \overline{U}_{i*}^{(t)}\Big)^2 \le \frac{4\mu^2 k}{m\widetilde{\rho}}.$$

Thus by Bernstein's inequality, we obtain

$$\mathbb{P}\left(|Y - \mathbb{E}Y| \ge \delta_{2k}\right) \le \exp\left(-\frac{3\delta_{2k}^2}{6 + 2\delta_{2k}} \frac{m\widetilde{\rho}}{4\mu^2 k}\right).$$

Since $\widetilde{\rho}$ and δ_{2k} satisfy (53) and (95), then for a sufficiently large C_7, we have

$$\mathbb{P}\left(|Y - \mathbb{E}Y| \ge \delta_{2k}\right) \le \frac{1}{n^3},$$

which implies that for any z and j, we have

$$\mathbb{P}\Big(1 + \delta_{2k} \ge z^\top S^{(j,t)} z \ge 1 - \delta_{2k}\Big) \ge 1 - n^{-3}. \tag{86}$$

Combining (86) with (85), we complete the proof. □

E.2 Proof of Lemma 22

Proof. For notational convenience, we omit the index t in $\overline{U}^{*(t)}$ and $V^{*(t)}$, and denote them by $\overline{U}^*$ and V^* respectively. Let $H^{(j,t)} = S^{(j,t)} K^{(j,t)} - J^{(j,t)}$. We have

$$H^{(j,t)} = \frac{1}{\widetilde{\rho}} \sum_{i:(i,j)\in\mathcal{W}_{2t+1}} \overline{U}_{i*}^{(t)} \overline{U}_{i*}^{(t)\top} \overline{U}^{(t)\top} \overline{U}^* - \overline{U}_{i*} \overline{U}_{i*}^\top = \frac{1}{\widetilde{\rho}} \sum_{i:(i,j)\in\mathcal{W}_{2t+1}} H^{(i,j,t)}.$$

We consider an arbitrary $Z \in \mathbb{R}^{n\times k}$ such that $\|Z\|_{\mathrm{F}} = 1$. Let $z = \mathrm{vec}(Z)$ and $v = \mathrm{vec}(V)$. Since

$$\sum_i H^{(i,j,t)} = \overline{U}^{(t)\top} \overline{U}^{(t)} \overline{U}^{(t)\top} \overline{U}^* - \overline{U}^{(t)\top} \overline{U}^* = 0,$$

then by Lemma 18, we have

$$z^\top (S^{(t)} K^{(t)} - J^{(t)}) v = \sum_j Z_{*j}^\top (S^{(j,t)} K^{(t)} - J^{(j,t)}) V_{j*} \le \frac{1}{\rho} \sum_{p,q} \sqrt{\sum_j Z_{pj}^2 (V_{jq})^2} \sqrt{\sum_i [H^{(i,j,t)}]_{pq}^2}.$$

Meanwhile, we have

$$\begin{aligned}\sum_i [H^{(i,j,t)}]_{pq}^2 &= \sum_i (\overline{U}_{ip}^{(t)})^2 (\overline{U}_{i*}^{(t)\top}\overline{U}^{(t)\top}\overline{U}_{*q}^* - \overline{U}_{iq}^*)^2 \le \max_i (\overline{U}_{ip}^{(t)})^2 \sum_i (\overline{U}_{i*}^{(t)\top}\overline{U}^{(t)\top}\overline{U}_{*q}^* - \overline{U}_{iq}^*)^2 \\ &= \max_i (\overline{U}_{ip}^{(t)})^2 (1 - \|\overline{U}^{(t)\top}\overline{U}_{*q}^*\|_2^2) \le \max_i \|\overline{U}_{i*}^{(t)}\|_2^2 (1 - (\overline{U}_{*q}^{(t)\top}\overline{U}_{*q}^*)^2) \\ &\overset{\text{(i)}}{\le} \frac{4\mu^2 k}{m}(1 - \overline{U}_{*q}^{(t)\top}\overline{U}_{*q}^*)(1 + \overline{U}_{*q}^{(t)\top}\overline{U}_{*q}^*) \\ &\overset{\text{(ii)}}{\le} \frac{4\mu^2 k}{m}\|\overline{U}_{*q} - \overline{U}_{*q}^*\|_2^2 \le \frac{4\sqrt{2}\mu^2 k}{m}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}^2,\end{aligned}$$

where (i) comes from the incoherence of $\overline{U}^{(t)}$, and (ii) comes from $\overline{U}_{*q}^{(t)\top}\overline{U}_{*q}^* \le \|\overline{U}_{*q}^{(t)}\|_2\|\overline{U}_{*q}^*\|_2 \le 1$.

Combining the above inequalities, by the incoherence of V and Bernstein's inequality, we have

$$z^\top(S^{(t)}K^{(t)} - J^{(t)})v \le \sum_{p,q} \frac{4\sigma_1\mu^2 k}{m\widetilde{\rho}}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}\|Z_{*p}\|_2 \le 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}$$

with probability at least $1 - n^{-3}$, where the last inequality comes from the incoherence of V, $\sum_p \|Z_{*p}\|_2 \le \sqrt{k}$, and a sufficiently large $\widetilde{\rho}$. Since z is arbitrary, then we have

$$\mathbb{P}\Big(\|(S^{(t)}K^{(t)} - J^{(t)})v\|_2 \le 3\delta_{2k}k\sigma_1\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}\Big) \ge 1 - n^{-3},$$

which completes the proof. □

E.3 Proof of Lemma 26

Proof. Recall that we have $W^{\text{in}} = V^{(t+0.5)}$ and $\overline{V}^{(t+1)} = W^{\text{out}}$ in Algorithm 3. Since $V^{(t+0.5)}$ satisfies (5) of Lemma 5, then there exists a factorization of $M^* = U^{*(t+0.5)}\overline{V}^{*(t+0.5)\top}$ such that $\overline{V}^{*(t+0.5)}$ is an orthonormal matrix, and satisfies

$$\|\overline{W}^{\text{in}} - \overline{V}^{*(t+0.5)}\|_{\mathrm{F}} \le \frac{2}{\sigma_k}\|W^{\text{in}} - V^{*(t)}\|_{\mathrm{F}} \le \frac{2}{\sigma_k}\cdot\frac{\sigma_k}{8} = \frac{1}{4}. \tag{87}$$

Since the Frobenius norm projection is contractive, then we have

$$\|\widetilde{W} - \overline{V}^{*(t+0.5)}\|_{\mathrm{F}} \le \|\overline{W}^{\text{in}} - \overline{V}^{*(t+0.5)}\|_{\mathrm{F}} \le \frac{1}{4}. \tag{88}$$

Since $\overline{V}^{*(t+0.5)}$ is an orthonormal matrix, by Lemma 14, we have

$$\|\overline{W}^{\text{out}} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \le \frac{\sqrt{2}\|\overline{V}^{*(t+0.5)\dagger}\|_2\|\widetilde{W} - \overline{V}^{*(t+0.5)}\|_{\mathrm{F}}}{1 - \|\overline{W}^{\text{in}} - \overline{V}^{*(t+0.5)}\|_{\mathrm{F}}\|\overline{V}^{*(t+0.5)\dagger}\|_2} \le 2\|\widetilde{W} - \overline{V}^{*(t+0.5)}\|_{\mathrm{F}} \le \frac{1}{2}, \tag{89}$$

where $\overline{V}^{*(t+1)} = \overline{V}^{*(t+0.5)}O$ for some unitary matrix $O \in \mathbb{R}^{k\times k}$, and the last inequality comes from (88). Moreover, since $\overline{V}^{*(t+1)}$ is an orthonormal matrix, then we have

$$\sigma_{\min}(\widetilde{W}) \ge \sigma_{\min}(\overline{V}^{*(t+1)}) - \|\widetilde{W} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \ge 1 - \frac{1}{2} = \frac{1}{2}.$$

where the last inequality comes from (89). Since $\overline{W}^{\text{out}} = \widetilde{W}(R_{\widetilde{W}}^{\text{tmp}})^{-1}$, then we have

$$\|\overline{W}_{i*}^{\text{out}}\|_2 \le \|\overline{W}^{\text{out}\top} e_i\|_2 = \|(R_{\widetilde{W}})^{-1}\|_2 \|\widetilde{W}^\top e_i\|_2 \le \sigma_{\min}^{-1}(\widetilde{W})\mu\sqrt{\frac{k}{n}} \le 2\mu\sqrt{\frac{k}{n}}.$$

□

E.4 Proof of Lemma 27

Proof. Before we proceed with the main proof, we first introduce the following lemma.

Lemma 19. Suppose that $\widetilde{\rho}$ satisfies (93). Recall that $\widetilde{U}$, $\widetilde{\Sigma}$, and $\widetilde{V}$ are defined in Algorithm 2. There exists a universal constant C such that

$$\|\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top - M^*\|_2 = C\sqrt{\frac{k}{\widetilde{\rho}\sqrt{mn}}}$$

with high probability.

The proof of Lemma 19 is provided in Keshavan et al. [2010a], therefore omitted.

We then proceed with the proof of Lemma 27. Since both $\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top$ and M^* are rank k matrices, then $\widetilde{U}\widetilde{\Sigma}\widetilde{V} - M^*$ has at most rank $2k$. Thus by Lemma 19, we have

$$\|\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top - M^*\|_{\text{F}}^2 \le 2k\|\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top - M^*\|_2^2 \le \frac{2Ck^2}{\rho\sqrt{mn}}\|M^*\|_{\text{F}}^2 \le \frac{2Ck^3\sigma_1^2}{\widetilde{\rho}\sqrt{mn}} \le \frac{\sigma_k^6(1-\delta_{2k})}{1024(1+\delta_{2k})\sigma_1^4\xi^2} \tag{90}$$

with high probability, where the last inequality comes from (93) with

$$C_7 \ge \frac{2048(1+\delta_{2k})^2\sigma_1^6\xi^2}{\mu^2\sigma_k^6(1-\delta_{2k})^2}.$$

Suppose that M^* has a rank k singular value decomposition $M^* = \widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top}$. Then we have

$$\begin{aligned}
\|\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top - M^*\|_{\text{F}}^2 &= \|\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top} - \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top\|_{\text{F}}^2 \\
&= \|\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top} - \widetilde{U}\widetilde{U}^\top\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top} + \widetilde{U}\widetilde{U}^\top\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top} - \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top\|_{\text{F}}^2 \\
&= \|(I_m - \widetilde{U}\widetilde{U}^\top)\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top} + \widetilde{U}(\widetilde{U}^\top\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top} - \widetilde{\Sigma}\widetilde{V}^\top)\|_{\text{F}}^2 \\
&\ge \|(I_m - \widetilde{U}\widetilde{U}^T)\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top}\|_{\text{F}}^2.
\end{aligned}$$

Let $\widetilde{U}_\perp \in \mathbb{R}^{m\times(m-k)}$ denote the orthogonal complement to $\widetilde{U}$. Then we have

$$\|(I_m - \widetilde{U}\widetilde{U}^T)\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top}\|_{\text{F}}^2 = \|(\widetilde{U}_\perp\widetilde{U}_\perp^\top)\widetilde{U}^*\widetilde{D}^*\widetilde{V}^{*\top}\|_2^2 = \|\widetilde{U}_\perp^\top\widetilde{U}^*\widetilde{D}^*\|_{\text{F}}^2 \ge \frac{\sigma_k^2}{2}\|\widetilde{U}_\perp^\top\widetilde{U}^*\|_{\text{F}}^2.$$

Thus Lemma 2 guarantees that for $\widetilde{O} = \text{argmin}_{O^\top O = I_k}\|\widetilde{U} - \widetilde{U}^*O\|_{\text{F}}$, we have

$$\|\widetilde{U} - \widetilde{U}^*\widetilde{O}\|_{\text{F}} \le \sqrt{2}\|\widetilde{U}_\perp^\top\widetilde{U}^*\|_{\text{F}} \le \frac{2}{\sigma_k}\|\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top - M^*\|_{\text{F}}.$$

We define $\widetilde{U}^{*\mathrm{tmp}} = \widetilde{U}^{*}\widetilde{O}$. Then combining the above inequality with (90), we have

$$\|\widetilde{U} - \widetilde{U}^{*\mathrm{tmp}}\|_{\mathrm{F}} \leq \frac{2}{\sigma_k}\|\widetilde{U}\widetilde{\Sigma}\widetilde{V}^{\top} - M^{*}\|_{\mathrm{F}} \leq \frac{\sigma_k^2(1-\delta_{2k})}{16(1+\delta_{2k})\sigma_1^2\xi}.$$

Since the Frobenius norm projection is contractive, then we have

$$\|\widetilde{U}^{\mathrm{tmp}} - \widetilde{U}^{*\mathrm{tmp}}\|_{\mathrm{F}} \leq \|\widetilde{U} - \widetilde{U}^{*\mathrm{tmp}}\|_{\mathrm{F}} \leq \frac{\sigma_k^2(1-\delta_{2k})}{16(1+\delta_{2k})\sigma_1^2\xi} \leq \frac{1}{16}, \tag{91}$$

where the last inequality comes from the definition of ξ and $\sigma_1 \geq \sigma_k$. Since $\widetilde{U}^{*\mathrm{tmp}}$ is an orthonormal matrix, by Lemma 14, we have

$$\begin{aligned}\|\overline{U}^{\mathrm{out}} - \overline{U}^{*(0)}\|_{\mathrm{F}} &\leq \frac{\sqrt{2}\widetilde{U}^{*\mathrm{tmp}\dagger}\|\widetilde{U}^{\mathrm{tmp}} - \widetilde{U}^{*\mathrm{tmp}}\|_{\mathrm{F}}}{1 - \|\widetilde{U}^{\mathrm{tmp}} - \widetilde{U}^{*\mathrm{tmp}}\|_{\mathrm{F}}\|\widetilde{U}^{*\mathrm{tmp}\dagger}\|_2} \\ &\leq 2\|\widetilde{U}^{\mathrm{tmp}} - \widetilde{U}^{*\mathrm{tmp}}\|_{\mathrm{F}} \leq \frac{\sigma_k^2(1-\delta_{2k})}{8(1+\delta_{2k})\sigma_1^2\xi} \leq \frac{1}{8},\end{aligned} \tag{92}$$

where $\overline{U}^{*(0)} = \widetilde{U}^{*\mathrm{tmp}}\widetilde{O}^{\mathrm{tmp}}$ for some unitary matrix $\widetilde{O}^{\mathrm{tmp}} \in \mathbb{R}^{k\times k}$ such that $\widetilde{O}^{\mathrm{tmp}}\widetilde{O}^{\mathrm{tmp}\top} = I_k$, and the last inequality comes from (91). Moreover, since $\overline{U}^{*(0)}$ is an orthonormal matrix, then we have

$$\sigma_{\min}(\widetilde{U}^{\mathrm{tmp}}) \geq \sigma_{\min}(\overline{U}^{*(0)}) - \|\widetilde{U}^{\mathrm{tmp}} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \geq 1 - \frac{1}{8} = \frac{7}{8},$$

where the last inequality comes from (91). Since $\overline{U}^{\mathrm{out}} = \widetilde{U}^{\mathrm{tmp}}(R_{\overline{U}}^{\mathrm{out}})^{-1}$, then we have

$$\|\overline{U}_{i*}^{\mathrm{out}}\|_2 \leq \|\overline{U}^{\mathrm{out}\top}e_i\|_2 = \|(R_{\overline{U}}^{\mathrm{out}})^{-1}\|_2\|\widetilde{U}^{\top}e_i\|_2 \leq \sigma_{\min}^{-1}(\widetilde{U}^{\mathrm{tmp}})\mu\sqrt{\frac{k}{m}} \leq \frac{8\mu}{7}\sqrt{\frac{k}{m}}.$$

Moreover, we define $V^{*(0)} = M^{*\top}\overline{U}^{*(0)}$. Then we have $\overline{U}^{*(0)}V^{*(0)\top} = \overline{U}^{*(0)}\overline{U}^{*(0)\top}M^{*} = M^{*}$, where the last inequality comes from the fact that $\overline{U}^{*(0)}\overline{U}^{*(0)\top}$ is exactly the projection matrix for the column space of M^{*}. □

E.5 Proof of Corollary 5

Proof. Since $\mathcal{E}_U^{(t)}$ implies that $\mathcal{E}_{U,1}^{(t)}$,..., and $\mathcal{E}_{U,4}^{(t)}$ hold with probability at least $1-4n^{-3}$, then combining Lemmas 24 and 25, we obtain

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{2\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \overset{\text{(i)}}{\leq} \frac{\sigma_k}{2\xi\sigma_1}\cdot\frac{\sigma_k(1-\delta_{2k})}{4(1+\delta_{2k})\sigma_1} = \frac{\sigma_k^2(1-\delta_{2k})}{8(1+\delta_{2k})\sigma_1^2\xi^2} \overset{\text{(ii)}}{\leq} \frac{\sigma_k}{8}$$

with probability at least $1-4n^{-3}$, where (i) comes from the definition of $\mathcal{E}_U^{(t)}$, and (ii) comes from the definition of ξ and $\sigma_1 \geq \sigma_k$. Therefore Lemma 26 implies that $\overline{V}^{(t+1)}$ is incoherent with parameter 2μ, and

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{4}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{2}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k(1-\delta_{2k})}{4(1+\delta_{2k})\sigma_1}$$

with probability at least $1-4n^{-3}$, where the last inequality comes from the definition of ξ and $\mathcal{E}_U^{(t)}$. □

F Proof of Theorem 2

We present the technical proof for matrix completion. Before we proceed with the main proof, we first introduce the following lemma.

Lemma 20. [Hardt and Wootters [2014]] Suppose that the entry observation probability $\bar{\rho}$ of $\mathcal{W}$ satisfies (53). Then the output sets $\{\mathcal{W}_t\}_{t=0}^{2T}$, of Algorithm 4 are equivalent to $2T+1$ observation sets, which are independently generated with the entry observation probability

$$\widetilde{\rho} \geq \frac{C_7 \mu^2 k^3 \log n}{m} \tag{93}$$

for some constant C_7.

See Hardt and Wootters [2014] for the proof of Lemma 20. Lemma 20 ensures the independence among all observation sets generated by Algorithm 4. To make the convergence analysis for matrix completion comparable to that for matrix sensing, we rescale both the objective function $\mathcal{F}_{\mathcal{W}}$ and step size η by the entry observation probability $\widetilde{\rho}$ of each individual set, which is also obtained by Algorithm 4. In particular, we define

$$\widetilde{\mathcal{F}}_{\mathcal{W}}(U,V) = \frac{1}{2\widetilde{\rho}} \|\mathcal{P}_{\mathcal{W}}(UV^\top) - \mathcal{P}_{\mathcal{W}}(M^*)\|_{\mathrm{F}}^2 \quad \text{and} \quad \widetilde{\eta} = \widetilde{\rho}\eta. \tag{94}$$

For notational simplicity, we assume that at the t-th iteration, there exists a matrix factorization of M^* as

$$M^* = \overline{U}^{*(t)} V^{*(t)\top},$$

where $\overline{U}^{*(t)} \in \mathbb{R}^{m\times k}$ is an orthonormal matrix. Then we define several $nk \times nk$ matrices

$$S^{(t)} = \begin{bmatrix} S_{11}^{(t)} & \cdots & S_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ S_{k1}^{(t)} & \cdots & S_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad S_{pq}^{(t)} = \begin{bmatrix} \frac{1}{\widetilde{\rho}} \sum_{i:(i,1)\in\mathcal{W}_{2t+1}} \overline{U}_{ip}^{(t)} \overline{U}_{iq}^{(t)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\widetilde{\rho}} \sum_{i:(i,n)\in\mathcal{W}_{2t+1}} \overline{U}_{ip}^{(t)} \overline{U}_{iq}^{(t)} \end{bmatrix},$$

$$G^{(t)} = \begin{bmatrix} G_{11}^{(t)} & \cdots & G_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ G_{k1}^{(t)} & \cdots & G_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad G_{pq}^{(t)} = \begin{bmatrix} \frac{1}{\widetilde{\rho}} \sum_{i:(i,1)\in\mathcal{W}_{2t+1}} \overline{U}_{ip}^{*(t)} \overline{U}_{iq}^{*(t)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\widetilde{\rho}} \sum_{i:(i,n)\in\mathcal{W}_{2t+1}} \overline{U}_{ip}^{*(t)} \overline{U}_{iq}^{*(t)} \end{bmatrix},$$

$$J^{(t)} = \begin{bmatrix} J_{11}^{(t)} & \cdots & J_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ J_{k1}^{(t)} & \cdots & J_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad J_{pq}^{(t)} = \begin{bmatrix} \frac{1}{\widetilde{\rho}} \sum_{i:(i,1)\in\mathcal{W}_{2t+1}} \overline{U}_{ip}^{(t)} \overline{U}_{iq}^{*(t)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\widetilde{\rho}} \sum_{i:(i,n)\in\mathcal{W}_{2t+1}} \overline{U}_{ip}^{(t)} \overline{U}_{iq}^{*(t)} \end{bmatrix},$$

$$K^{(t)} = \begin{bmatrix} K_{11}^{(t)} & \cdots & K_{1k}^{(t)} \\ \vdots & \ddots & \vdots \\ K_{k1}^{(t)} & \cdots & K_{kk}^{(t)} \end{bmatrix} \quad \text{with} \quad K_{pq}^{(t)} = \overline{U}_{*p}^{(t)\top} \overline{U}_{*q}^{*(t)} I_n,$$

where $1 \le p, q \le k$. Note that $S^{(t)}$ and $G^{(t)}$ are the partial Hessian matrices $\nabla_V^2 \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V)$ and $\nabla_V^2 \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)}, V)$ with respect to a vectorized V, i.e., $\mathrm{vec}(V)$.

F.1 Proof of Theorem 2 (Alternating Exact Minimization)

Proof. Throughout the proof for alternating exact minimization, we define a constant $\xi \in (2, \infty)$ to simplify the notation. We define the approximation error of the inexact first order oracle as

$$\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}) = \|\nabla_V \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)}, V^{(t+0.5)}) - \nabla_V \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V^{(t+0.5)})\|_{\mathrm{F}}.$$

To simplify our later analysis, we first introduce the following event.

$$\mathcal{E}_U^{(t)} = \left\{ \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \le \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1} \quad \text{and} \quad \max_i \|\overline{U}_{i*}^{(t)}\|_2 \le 2\mu\sqrt{\frac{k}{m}} \right\}.$$

We then present two important consequences of $\mathcal{E}_U^{(t)}$.

Lemma 21. Suppose that $\mathcal{E}_U^{(t)}$ holds, and $\widetilde{\rho}$ satisfies (93). Then we have

$$\mathbb{P}\Big(1 + \delta_{2k} \ge \sigma_{\max}(S^{(t)}) \ge \sigma_{\min}(S^{(t)}) \ge 1 - \delta_{2k}\Big) \ge 1 - n^{-3},$$

where δ_{2k} is some constant satisfying

$$\delta_{2k} \le \frac{\sigma_k^6}{192\xi^2 k \sigma_1^6}. \tag{95}$$

The proof of Lemma 21 is provided in Appendix E.1. Lemma 21 is also applicable to $G^{(t)}$, since $G^{(t)}$ shares the same structure with $S^{(t)}$, and $\overline{U}^{*(t)}$ is incoherent with parameter μ.

Lemma 22. Suppose that $\mathcal{E}_U^{(t)}$ holds, and $\widetilde{\rho}$ satisfies (93). Then for an incoherent V with parameter $3\sigma_1\mu$, we have

$$\mathbb{P}\Big(\|(S^{(t)}K^{(t)} - J^{(t)}) \cdot \mathrm{vec}(V)\|_2 \le 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - U^{*(t)}\|_{\mathrm{F}}\Big) \ge 1 - n^{-3},$$

where δ_{2k} is defined in (95).

The proof of Lemma 22 is provided in Appendix E.2. Note that Lemma 22 is also applicable to $\|(G^{(t)}K^{(t)} - J^{(t)}) \cdot \mathrm{vec}(V)\|_2$, since $G^{(t)}$ shares the same structure with $S^{(t)}$, and $\overline{U}^{*(t)}$ is incoherent with parameter μ.

We then introduce another two events:

$$\mathcal{E}_{U,1}^{(t)} = \{1 + \delta_{2k} \geq \sigma_{\max}(S^{(t)}) \geq \sigma_{\min}(S^{(t)}) \geq 1 - \delta_{2k}\},$$
$$\mathcal{E}_{U,2}^{(t)} = \{1 + \delta_{2k} \geq \sigma_{\max}(G^{(t)}) \geq \sigma_{\min}(G^{(t)}) \geq 1 - \delta_{2k}\},$$

where δ_{2k} is defined in $\mathcal{E}_U^{(t)}$. By Lemmas 21, we can verify that $\mathcal{E}_U^{(t)}$ implies $\mathcal{E}_{U,1}^{(t)}$ and $\mathcal{E}_{U,2}^{(t)}$ with probability at least $1 - 2n^{-3}$. The next lemma shows that $\mathcal{E}_{U,1}^{(t)}$ and $\mathcal{E}_{U,2}^{(t)}$ imply the strong convexity and smoothness of $\widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(U, V)$ in V at $U = \overline{U}^{(t)}$ and $\overline{U}^{*(t)}$.

Lemma 23. Suppose that $\mathcal{E}_{U,1}^{(t)}$ and $\mathcal{E}_{U,2}^{(t)}$ hold. Then for any $V,\ V' \in \mathbb{R}^{n\times k}$, we have

$$\begin{aligned}
\frac{1+\delta_{2k}}{2}\|V' - V\|_{\mathrm{F}}^2 \geq \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V') - \mathcal{F}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V) \\
- \langle \nabla_V \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V), V' - V\rangle \geq \frac{1-\delta_{2k}}{2}\|V' - V\|_{\mathrm{F}}^2, \\
\frac{1+\delta_{2k}}{2}\|V' - V\|_{\mathrm{F}}^2 \geq \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)}, V') - \mathcal{F}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)}, V) \\
- \langle \nabla_V \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)}, V), V' - V\rangle \geq \frac{1-\delta_{2k}}{2}\|V' - V\|_{\mathrm{F}}^2.
\end{aligned}$$

Since $S^{(t)}$ and $G^{(t)}$ are essentially the partial Hessian matrices $\nabla_V^2 \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V)$ and $\nabla_V^2 \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)}, V)$, the proof of 23 directly follows Appendix A.1, and is therefore omitted.

We then introduce another two events:

$$\mathcal{E}_{U,3}^{(t)} = \{\|(S^{(t)}K^{(t)} - J^{(t)}) \cdot \mathrm{vec}(V^{*(t)})\|_2 \leq 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - U^{*(t)}\|_{\mathrm{F}}\},$$
$$\mathcal{E}_{U,4}^{(t)} = \{\|(G^{(t)}K^{(t)} - J^{(t)}) \cdot \mathrm{vec}(V^{*(t)})\|_2 \leq 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - U^{*(t)}\|_{\mathrm{F}}\},$$

where δ_{2k} is defined in $\mathcal{E}_U^{(t)}$. We can verify that $\mathcal{E}_U^{(t)}$ implies $\mathcal{E}_{U,3}^{(t)}$ and $\mathcal{E}_{U,4}^{(t)}$ with probability at least $1 - 2n^{-3}$ by showing the incoherence of $V^{*(t)}$. More specifically, let $V^{*(t)} = \overline{V}^{*(t)}R_V^{*(t)}$ denote the QR decomposition of $V^{*(t)}$. We have

$$\|V_{j*}^{*(t)}\|_2 = \|R_V^{*(t)\top}V^{*(t)\top}e_j\|_2 \leq \|R_V^{*(t)}\|_2\|V^{*(t)\top}e_j\|_2 \leq \sigma_1\|V_{j*}^{*(t)}\|_2 \leq \sigma_1\mu\sqrt{\frac{k}{n}}. \tag{96}$$

Then Lemma 22 are applicable to $\mathcal{E}_{U,3}^{(t)}$ and $\mathcal{E}_{U,4}^{(t)}$.

We then introduce the following key lemmas, which will be used in the main proof.

Lemma 24. Suppose that $\mathcal{E}_U^{(t)}$, $\mathcal{E}_{U,1}^{(t)}$,..., and $\mathcal{E}_{U,4}^{(t)}$ hold. We then have

$$\mathcal{E}(V^{(t+0.5)}, V^{(t+0.5)}, \overline{U}^{(t)}) \leq \frac{(1-\delta_{2k})\sigma_k}{2\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}.$$

Lemma 24 shows that the approximation error of the inexact first order oracle for updating V diminishes with the estimation error of $U^{(t)}$, when $U^{(t)}$ is sufficiently close to $U^{*(t)}$. It is analogous to Lemma 3 in the analysis of matrix sensing, and its proof directly follows A.2, and is therefore omitted.

Lemma 25. Suppose that $\mathcal{E}_U^{(t)}$, $\mathcal{E}_{U,1}^{(t)}$,..., and $\mathcal{E}_{U,4}^{(t)}$ hold. We then have

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{1}{1-\delta_{2k}} \mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}).$$

Lemma 25 illustrates that the estimation error of $V^{(t+0.5)}$ diminishes with the approximation error of the inexact first order oracle. It is analogous to Lemma 4 in the analysis of matrix sensing, and its proof directly follows Appendix A.3, and is therefore omitted.

Lemma 26. Suppose that $V^{(t+0.5)}$ satisfies

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{8}. \tag{97}$$

Then there exists a factorization of $M^* = U^{*(t+1)}\overline{V}^{*(t+1)\top}$ such that $\overline{V}^{*(t+1)} \in \mathbb{R}^{n\times k}$ is an orthonormal matrix, and satisfies

$$\max_j \|\overline{V}_{j*}^{(t+1)}\|_2 \leq 2\mu\sqrt{\frac{2k}{n}} \quad \text{and} \quad \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{4}{\sigma_k}\|V^{(t+0.5)} - V^*\|_{\mathrm{F}}.$$

The proof of Lemma 26 is provided in Appendix E.3. Lemma 26 ensures that the incoherence factorization enforces $\overline{V}^{(t+1)}$ to be incoherent with parameter 2μ.

Lemma 27. Suppose that $\widetilde{\rho}$ satisfies (93). Then $\mathcal{E}_U^{(0)}$ holds with high probability.

The proof of Lemma 27 is provided in §E.4. Lemma 27 shows that the initial solution $\overline{U}^{(0)}$ is incoherent with parameter 2μ, while achieving a sufficiently small estimation error with high probability. It is analogous to Lemma 6 for matrix sensing.

Combining Lemmas 24, 25, and 26, we obtain the next corollary for a complete iteration of updating V.

Corollary 5. Suppose that $\mathcal{E}_U^{(t)}$ holds. Then

$$\mathcal{E}_V^{(t)} = \left\{ \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \leq \frac{(1-\delta_{2k})\sigma_k}{4\xi(1+\delta_{2k})\sigma_1} \quad \text{and} \quad \max_j \|\overline{V}_{j*}^{(t+1)}\|_2 \leq 2\mu\sqrt{\frac{k}{m}} \right\}$$

holds with probability at least $1 - 4n^{-3}$. Moreover, we have

$$\|\overline{V}^{(t+1)} - \overline{V}^*\|_{\mathrm{F}} \leq \frac{2}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \text{ and } \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{2\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}$$

with probability at least $1 - 4n^{-3}$.

The proof of Corollary 5 is provided in Appendix E.5. Since the alternating exact minimization algorithm updates U and V in a symmetric manner, we can establish similar results for a complete iteration of updating U in the next corollary.

Corollary 6. Suppose $\mathcal{E}_V^{(t)}$ holds. Then $\mathcal{E}_U^{(t+1)}$ holds with probability at least $1-4n^{-3}$. Moreover, we have

$$\|\overline{U}^{(t+1)}-\overline{U}^{*(t+1)}\|_{\mathrm{F}} \le \frac{2}{\xi}\|\overline{V}^{(t+1)}-\overline{V}^{*(t+1)}\|_{\mathrm{F}} \text{ and } \|U^{(t+0.5)}-U^{*(t+1)}\|_{\mathrm{F}} \le \frac{\sigma_k}{2\xi}\|\overline{V}^{(t+1)}-\overline{V}^{*(t+1)}\|_{\mathrm{F}}$$

with probability at least $1-4n^{-3}$.

The proof of Lemma 6 directly follows Appendix E.5, and is therefore omitted.

We proceed with the proof of Theorem 2 conditioning on $\mathcal{E}_U^{(0)}$. Similar to Appendix 1, we can recursively apply Corollaries 5 and 6, and show that $\{\mathcal{E}_U^{(t)}\}_{t=1}^T$ and $\{\mathcal{E}_V^{(t)}\}_{t=0}^T$ simultaneously hold with probability at least $1-8Tn^{-3}$. Then conditioning on all $\{\mathcal{E}_U^{(t)}\}_{t=0}^T$ and $\{\mathcal{E}_V^{(t)}\}_{t=0}^T$, we have

$$\begin{aligned}\|\overline{V}^{(T)}-\overline{V}^{*(T)}\|_{\mathrm{F}} &\le \frac{2}{\xi}\|\overline{U}^{(T-1)}-\overline{U}^{*(T-1)}\|_{\mathrm{F}} \le \left(\frac{2}{\xi}\right)^2\|\overline{V}^{(T-1)}-\overline{V}^{*(T-1)}\|_{\mathrm{F}} \\ &\le \left(\frac{2}{\xi}\right)^{2T-1}\|\overline{U}^{(0)}-\overline{U}^{*(0)}\|_{\mathrm{F}} \le \left(\frac{2}{\xi}\right)^{2T}\frac{(1-\delta_{2k})\sigma_k}{8(1+\delta_{2k})\sigma_1}, \end{aligned} \tag{98}$$

where the last inequality comes from the definition of $\mathcal{E}_U^{(0)}$. Thus we only need

$$T=\left\lceil \frac{1}{2}\log^{-1}\left(\frac{\xi}{2}\right)\log\left(\frac{(1-\delta_{2k})\sigma_k}{4(1+\delta_{2k})\sigma_1}\cdot\frac{1}{\epsilon}\right)\right\rceil$$

iterations such that

$$\|\overline{V}^{(T)}-\overline{V}^{*(T)}\|_{\mathrm{F}} \le \left(\frac{2}{\xi}\right)^{2T}\frac{(1-\delta_{2k})\sigma_k}{8(1+\delta_{2k})\sigma_1} \le \frac{\epsilon}{2}. \tag{99}$$

Meanwhile, by (99) and Corollary 6, we have

$$\|U^{(T-0.5)}-U^{*(T)}\|_{\mathrm{F}} \le \frac{\sigma_k}{2\xi}\|\overline{V}^{(T)}-\overline{V}^{*(T)}\|_{\mathrm{F}} \le \left(\frac{2}{\xi}\right)^{2T}\frac{(1-\delta_{2k})\sigma_k^2}{16\xi(1+\delta_{2k})\sigma_1},$$

where the last inequality comes from (98). Thus we only need

$$T=\left\lceil \frac{1}{2}\log^{-1}\left(\frac{\xi}{2}\right)\log\left(\frac{(1-\delta_{2k})\sigma_k^2}{8\xi(1+\delta_{2k})}\cdot\frac{1}{\epsilon}\right)\right\rceil$$

iterations such that

$$\|U^{(T-0.5)}-U^{*(T)}\|_{\mathrm{F}} \le \left(\frac{2}{\xi}\right)^{2T}\frac{(1-\delta_{2k})\sigma_k^2}{16\xi(1+\delta_{2k})\sigma_1} \le \frac{\epsilon}{2\sigma_1}. \tag{100}$$

We then combine (99) and (100) by following similar lines to §4.2, and show

$$\|M^{(T)}-M^*\|_{\mathrm{F}} \le \epsilon. \tag{101}$$

The above analysis only depends on $\mathcal{E}_U^{(0)}$. Because Lemma 27 guarantees that $\mathcal{E}_U^{(0)}$ holds with high probability, given $T \ll n^3$, (101) also holds with high probability. □

F.2 Proof of Theorem 2 (Alternating Gradient Descent)

Proof. Throughout the proof for alternating gradient descent, we define a sufficiently large constant ξ. Moreover, we assume that at the t-th iteration, there exists a matrix factorization of M^* as

$$M^* = \overline{U}^{*(t)} V^{*(t)\top},$$

where $\overline{U}^{*(t)} \in \mathbb{R}^{m\times k}$ is an orthonormal matrix. We define the approximation error of the inexact first order oracle as

$$\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \le \|\widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{(t)}, V^{(t)}) - \nabla_V \widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(\overline{U}^{*(t)}, V^{(t)})\|_{\mathrm{F}}$$

To simplify our later analysis, we introduce the following event.

$$\mathcal{E}_U^{(t)} = \Bigg\{ \max_i \|\overline{U}_{i*}^{(t)}\|_2 \le 2\mu\sqrt{\frac{k}{n}}, \quad \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{4\xi\sigma_1^2} \\ \max_i \|V_{j*}^{(t)}\|_2 \le 2\sigma_1\mu\sqrt{\frac{k}{n}}, \quad \text{and} \quad \|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{2\xi\sigma_1} \Bigg\}.$$

As has been shown in §F.1, $\mathcal{E}_U^{(t)}$ implies the following four events with probability at least $1-4n^{-3}$,

$$\begin{aligned}
\mathcal{E}_{U,1}^{(t)} &= \Big\{1 + \delta_{2k} \ge \sigma_{\max}(S^{(t)}) \ge \sigma_{\min}(S^{(t)}) \ge 1 - \delta_{2k}\Big\}, \\
\mathcal{E}_{U,2}^{(t)} &= \Big\{1 + \delta_{2k} \ge \sigma_{\max}(G^{(t)}) \ge \sigma_{\min}(G^{(t)}) \ge 1 - \delta_{2k}\Big\}, \\
\mathcal{E}_{U,3}^{(t)} &= \Big\{\|(S^{(t)}K^{(t)} - J^{(t)}) \cdot \mathrm{vec}(V^{*(t)})\|_2 \le 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - U^{*(t)}\|_{\mathrm{F}}\Big\}, \\
\mathcal{E}_{U,4}^{(t)} &= \Big\{\|(G^{(t)}K^{(t)} - J^{(t)}) \cdot \mathrm{vec}(V^{*(t)})\|_2 \le 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - U^{*(t)}\|_{\mathrm{F}}\Big\},
\end{aligned}$$

where δ_{2k} is defined in (95). In §F.1, we also show that $\mathcal{E}_{U,1}^{(t)}$ and $\mathcal{E}_{U,2}^{(t)}$ imply the strong convexity and smoothness of $\widetilde{\mathcal{F}}_{\mathcal{W}_{2t+1}}(U, V)$ at $U = \overline{U}^{(t)}$ and $\overline{U}^{*(t)}$.

Moreover, we introduce the following two events,

$$\begin{aligned}
\mathcal{E}_{U,5}^{(t)} &= \Big\{\|(S^{(t)}K^{(t)} - J^{(t)}) \cdot \mathrm{vec}(V^{(t)} - V^{*(t)})\|_2 \le 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - U^{*(t)}\|_{\mathrm{F}}\Big\}, \\
\mathcal{E}_{U,6}^{(t)} &= \Big\{\|(G^{(t)}K^{(t)} - J^{(t)}) \cdot \mathrm{vec}(V^{(t)} - V^{*(t)})\|_2 \le 3k\sigma_1\delta_{2k}\|\overline{U}^{(t)} - U^{*(t)}\|_{\mathrm{F}}\Big\},
\end{aligned}$$

where δ_{2k} is defined in (95). We can verify that $\mathcal{E}_U^{(t)}$ implies $\mathcal{E}_{U,5}^{(t)}$ and $\mathcal{E}_{U,6}^{(t)}$ with probability at least $1-2n^{-3}$ by showing the incoherence of $V^{(t)} - V^{*(t)}$. More specifically, we have

$$\max_j \|V_{j*}^{(t)} - V_{j*}^{*(t)}\|_2 \le \max_i \|V_{j*}^{(t)}\|_2 + \max_j \|V_{j*}^{*(t)}\|_2 \le 3\sigma_1\mu\sqrt{\frac{k}{n}},$$

where the last inequality follows the definition of $\mathcal{E}_U^{(t)}$ and the incoherence of $V^{*(t)}$ as shown in (96). Then Lemma 22 are applicable to $\mathcal{E}_{U,5}^{(t)}$ and $\mathcal{E}_{U,6}^{(t)}$

We then introduce the following key lemmas, which will be used in the main proof.

Lemma 28. Suppose that $\mathcal{E}_U^{(t)}$, $\mathcal{E}_{U,1}^{(t)}$, ..., and $\mathcal{E}_{U,6}^{(t)}$ hold. Then we have

$$\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}) \leq \frac{(1+\delta_{2k})\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^*\|_{\mathrm{F}}.$$

Lemma 28 shows that the approximation error of the inexact first order oracle for updating V diminishes with the estimation error of $U^{(t)}$, when $U^{(t)}$ is sufficiently close to $U^{*(t)}$. It is analogous to Lemma 7 in the analysis of matrix sensing, and its proof directly follows B.1, and is therefore omitted.

Lemma 29. Suppose that $\mathcal{E}_U^{(t)}$, $\mathcal{E}_{U,1}^{(t)}$, ..., and $\mathcal{E}_{U,6}^{(t)}$ hold. Meanwhile, the rescaled step size parameter $\widetilde{\eta}$ satisfies

$$\widetilde{\eta} = \frac{1}{1+\delta_{2k}}.$$

Then we have

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \sqrt{\delta_{2k}}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2}{1+\delta_{2k}}\mathcal{E}(V^{(t+0.5)}, V^{(t)}, \overline{U}^{(t)}).$$

Lemma 29 illustrates that the estimation error of $V^{(t+0.5)}$ diminishes with the approximation error of the inexact first order oracle. It is analogous to Lemma 8 in the analysis of matrix sensing. Its proof directly follows Appendix B.2, and is therefore omitted.

Lemma 30. Suppose that $V^{(t+0.5)}$ satisfies

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \leq \frac{\sigma_k}{8}.$$

We then have

$$\max_j \|\overline{V}_{j*}^{(t+1)}\|_2 \leq 2\mu\sqrt{\frac{2k}{n}} \quad \text{and} \quad \max_i \|U_{i*}^{(t+1)}\|_2 \leq 2\sigma_1\mu\sqrt{\frac{2k}{m}}$$

Moreover, there exists a factorization of $M^* = U^{*(t+1)}\overline{V}^{*(t+1)\top}$ such that $\overline{V}^{*(t+1)}$ is an orthonormal matrix, and

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t)}\|_{\mathrm{F}} \leq \frac{4}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}},$$
$$\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \leq \frac{5\sigma_1}{\sigma_k}\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}.$$

The proof of Lemma 26 is provided in Appendix G.1. Lemma 30 guarantees that the incoherence factorization enforces $\overline{V}^{(t+1)}$ and $U^{(t)}$ to be incoherent with parameters 2μ and $2\sigma_1\mu$ respectively. The next lemma characterizes the estimation error of the initial solutions.

Lemma 31. Suppose that $\widetilde{\rho}$ satisfies (93). Then $\mathcal{E}_U^{(0)}$ holds with high probability.

The proof of Lemma 31 is provided in Appendix G.2. Lemma 31 ensures that the initial solutions $U^{(0)}$, and $V^{(0)}$ are incoherent with parameters 2μ and $2\sigma_1\mu$ respectively, while achieving sufficiently small estimation errors with high probability. It is analogous to Lemma 10 in the analysis of matrix sensing.

Combining Lemmas 28, 29, and 26, we obtain the following corollary for a complete iteration of updating V.

Corollary 7. Suppose that $\mathcal{E}_U^{(t)}$ holds. Then

$$\mathcal{E}_V^{(t)} = \Bigg\{ \max_j \|\overline{V}_{j*}^{(t)}\|_2 \le 2\mu\sqrt{\frac{k}{m}}, \quad \|\overline{V}^{(t)} - \overline{V}^{*(t)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{4\xi\sigma_1^2}, \\ \max_i \|U_{i*}^{(t)}\|_2 \le 2\sigma_1\mu\sqrt{\frac{k}{m}}, \text{ and } \|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \le \frac{\sigma_k^2}{2\xi\sigma_1} \Bigg\}$$

holds with probability at least $1 - 6n^{-3}$. Moreover, we have

$$\|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} \le \sqrt{\delta_{2k}}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{2\sigma_k}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{102}$$

$$\|\overline{V}^{(t+1)} - \overline{V}^{*(t)}\|_{\mathrm{F}} \le \frac{4\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \frac{8}{\xi}\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{103}$$

$$\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} \le \frac{5\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)} - V^{*(t)}\|_{\mathrm{F}} + \left(\frac{10}{\xi} + 1\right)\sigma_1\|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}}, \tag{104}$$

with probability at least $1 - 6n^{-3}$.

The proof of Corollary 7 is provided in Appendix G.3. Since the algorithm updates U and V in a symmetric manner, we can establish similar results for a complete iteration of updating U in the next corollary.

Corollary 8. Suppose $\mathcal{E}_V^{(t)}$ holds. Then $\mathcal{E}_U^{(t+1)}$ holds with probability at least $1 - 6n^{-3}$. Moreover, we have

$$\|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}} \le \sqrt{\delta_{2k}}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \frac{2\sigma_k}{\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{105}$$

$$\|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}} \le \frac{4\sqrt{\delta_{2k}}}{\sigma_k}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \frac{8}{\xi}\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{106}$$

$$\|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}} \le \frac{5\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} + \left(\frac{10\sigma_1}{\xi} + 1\right)\|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}}, \tag{107}$$

with probability at least $1 - 6n^{-3}$.

The proof of Corollary 8 directly follows Appendix G.3, and is therefore omitted.

We then proceed with the proof of Theorem 2 conditioning on $\mathcal{E}_U^{(0)}$. Similar to Appendix 1, we can recursively apply Corollaries 7 and 8, and show that $\{\mathcal{E}_U^{(t)}\}_{t=1}^T$ and $\{\mathcal{E}_V^{(t)}\}_{t=0}^T$ simultaneously

hold with probability at least $1 - 12Tn^{-3}$. For simplicity, we define

$$\phi_{V^{(t+1)}} = \|V^{(t+1)} - V^{*(t+1)}\|_{\mathrm{F}}, \quad \phi_{V^{(t+0.5)}} = \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}, \quad \phi_{\overline{V}^{(t+1)}} = \sigma_1 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}},$$
$$\phi_{U^{(t+1)}} = \|U^{(t+1)} - U^{*(t+2)}\|_{\mathrm{F}}, \quad \phi_{U^{(t+0.5)}} = \|U^{(t+0.5)} - U^{*(t+1)}\|_{\mathrm{F}}, \quad \phi_{\overline{U}^{(t+1)}} = \sigma_1 \|\overline{U}^{(t+1)} - \overline{U}^{*(t+1)}\|_{\mathrm{F}}.$$

We then follow similar lines to §4.3 and §F.1, and show that $\|M^{(T)} - M\|_{\mathrm{F}} \le \epsilon$ with high probability. □

F.3 Proof of Theorem 2 (Gradient Descent)

Proof. The convergence analysis of the gradient descent algorithm is similar to alternating gradient descent. The only difference is, for updating U, gradient descent uses $V = \overline{V}^{(t)}$ instead of $V = \overline{V}^{(t+1)}$ to calculate the gradient at $U = U^{(t)}$. Then everything else directly follows Appendix F.2, and is therefore omitted.

□

G Lemmas for Theorem 2 (Alternating Gradient Descent)

G.1 Proof of Lemma 30

Proof. Recall that we have $W^{\text{in}} = V^{(t+0.5)}$ and $\overline{V}^{(t+1)} = W^{\text{out}}$ in Algorithm 3. By Lemma 26, we can show

$$\|\overline{W}^{\text{out}} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \le \frac{4}{\sigma_k} \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}}. \tag{108}$$

By Lemma 17, we have

$$\begin{aligned}
\|R_{\overline{V}}^{(t+0.5)} - \overline{V}^{*(t+1)\top} V^{*(t)}\|_{\mathrm{F}} &= \|\overline{V}^{(t+1)\top} V^{(t+0.5)} - \overline{V}^{*(t+1)\top} V^{*(t)}\|_{\mathrm{F}} \\
&\le \|\overline{V}^{(t+1)}\|_2 \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \|V^{*(t)}\|_2 \|\overline{V}^{(t+1)} - \overline{V}^{*(t+1)}\|_{\mathrm{F}} \\
&\le \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \frac{4\sigma_1}{\sigma_k} \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}},
\end{aligned} \tag{109}$$

where the last inequality comes from (108), $\|\overline{V}^{(t+1)}\|_2 = 1$, and $\|V^{*(t)}\|_2 = \sigma_1$. Moreover, we define $U^{*(t+1)} = \overline{U}^{*(t)} (\overline{V}^{*(t+1)\top} V^{*(t)})^\top$. Then we can verify

$$U^{*(t+1)} \overline{V}^{*(t+1)} = \overline{U}^{*(t)} V^{*(t)\top} \overline{V}^{*(t+1)} \overline{V}^{*(t+1)\top} = M^*,$$

where the last equality holds since $\overline{V}^{*(t+1)} \overline{V}^{*(t+1)}$ is exactly the projection matrix for the row space of M^*. Thus we further have

$$\begin{aligned}
\|U^{(t)} - U^{*(t+1)}\|_{\mathrm{F}} &= \|\overline{U}^{(t)} (\overline{V}^{(t+1)\top} V^{(t+0.5)})^\top - \overline{U}^{*(t)} (\overline{V}^{*(t+1)\top} V^{*(t)})^\top\|_{\mathrm{F}} \\
&\le \|\overline{U}^{(t)}\|_2 \|R_{\overline{V}}^{(t+0.5)} - \overline{V}^{*(t+1)\top} V^{*(t)}\|_{\mathrm{F}} + \|\overline{V}^{*(t+1)\top} V^{*(t)}\|_2 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}} \\
&\le \frac{5\sigma_1}{\sigma_k} \|V^{(t+0.5)} - V^{*(t)}\|_{\mathrm{F}} + \sigma_1 \|\overline{U}^{(t)} - \overline{U}^{*(t)}\|_{\mathrm{F}},
\end{aligned}$$

where the last inequality comes from (109), $\|\overline{U}^{(t)}\|_2 = 1$, $\|\overline{V}^{*(t+1)\top} V^{*(t)}\|_2 = \sigma_1$, and $\sigma_1 \geq \sigma_k$. □

G.2 Proof of Lemma 31

Proof. Following similar lines to Appendix E.4, we can obtain

$$\max_i \|\overline{U}_{i*}^{(0)}\|_2 \leq \frac{8\mu}{7}\sqrt{\frac{k}{m}}, \quad \|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{8\xi\sigma_1^2}, \tag{110}$$

$$\max_j \|\overline{V}_{j*}^{(0)}\|_2 \leq \frac{8\mu}{7}\sqrt{\frac{k}{n}}, \quad \|\overline{V}^{(0)} - \overline{V}^{*(0)}\|_{\mathrm{F}} \leq \frac{\sigma_k^2}{8\xi\sigma_1^2}. \tag{111}$$

Then by Lemma 17, we have

$$\begin{aligned}\|\overline{U}^{(0)\top}\widetilde{M} - \overline{U}^{*(0)\top}M^*\|_{\mathrm{F}} &\leq \|\overline{U}^{(0)}\|_2\|\widetilde{M} - M^*\|_{\mathrm{F}} + \|M^*\|_2\|\overline{U}^{(0)} - \overline{U}^{*(0)}\|_{\mathrm{F}} \\ &\leq \frac{\sigma_1^3}{32\xi\sigma_k^2} + \frac{\sigma_k^2}{8\xi\sigma_1} \leq \frac{5\sigma_k^2}{32\xi\sigma_1}.\end{aligned} \tag{112}$$

By Lemma 17 again, we have

$$\begin{aligned}&\|\overline{V}^{(0)\top}\widetilde{M}^\top\overline{U}^{(0)} - \overline{V}^{*(0)\top}M^{*\top}\overline{U}^{*(0)}\|_{\mathrm{F}} \\ &\quad\leq \|\overline{V}^{(0)}\|_2\|\widetilde{M}^\top\overline{U}^{(0)} - M^{*\top}\overline{U}^{*(0)}\|_{\mathrm{F}} + \|\overline{U}^{*(0)\top}M^*\|_2\|\overline{V}^{(0)} - \overline{V}^{*(0)}\|_{\mathrm{F}} \\ &\quad\leq \frac{5\sigma_k^2}{32\xi\sigma_1} + \frac{\sigma_k^2}{8\xi\sigma_1} \leq \frac{9\sigma_k^2}{32\xi\sigma_1},\end{aligned} \tag{113}$$

where the last inequality comes from (111) and (112), and $\|M^*\|_2 = \sigma_1$. By Lemma 17 again, we have

$$\begin{aligned}\|V^{(0)} - V^{*(0)}\|_{\mathrm{F}} &\leq \|\overline{V}^{(0)}\overline{V}^{(0)\top}\widetilde{M}^\top\overline{U}^{(0)} - \overline{V}^{*(0)}\overline{V}^{*(0)\top}M^{*\top}\overline{U}^{*(0)}\|_{\mathrm{F}} \\ &\leq \|\overline{V}^{(0)}\|_2\|\overline{V}^{(0)\top}\widetilde{M}^\top\overline{U}^{(0)} - \overline{V}^{*(0)\top}M^{*\top}\overline{U}^{*(0)}\|_{\mathrm{F}} + \|\overline{U}^{*(0)\top}M^*\overline{V}^{*(0)}\|_2\|\overline{V}^{(0)} - \overline{V}^{*(0)}\|_{\mathrm{F}} \\ &\leq \frac{9\sigma_k^2}{32\xi\sigma_1} + \frac{\sigma_k^2}{8\xi\sigma_1} \leq \frac{13\sigma_k^2}{32\xi\sigma_1},\end{aligned} \tag{114}$$

where the last two inequalities come from (111), (114), and $\|\overline{U}^{*(0)\top}M^*\overline{V}^{*(0)}\|_2 \leq \sigma_1$, the definition of ξ, and $\sigma_1 \geq \sigma_k$. Moreover, by the incoherence of $V^{(0)}$, we have

$$\begin{aligned}\|V_{j*}^{(0)}\|_2 &\leq \|V^{(0)\top}e_j\|_2 = \|\overline{V}^{(0)\top}\widetilde{M}^\top\overline{U}^{(0)}\|_2\|\overline{V}^{(0)\top}e_i\|_2 \\ &\leq \left(\|\overline{V}^{*(0)}\overline{V}^{*(0)\top}M^{*\top}\overline{U}^{*(0)}\|_2 + \|\overline{V}^{(0)\top}\widetilde{M}^\top\overline{U}^{(0)} - \overline{V}^{*(0)}\overline{V}^{*(0)\top}M^{*\top}\overline{U}^{*(0)}\|_{\mathrm{F}}\right)\frac{6\mu}{5}\sqrt{\frac{k}{m}} \\ &\leq \left(1 + \frac{9\sigma_k^2}{32\xi\sigma_1^2}\right)\frac{6\sigma_1\mu}{5}\sqrt{\frac{k}{m}} \leq \frac{41\sigma_1\mu}{28}\sqrt{\frac{k}{m}},\end{aligned}$$

where the last two inequalities come from (111), (114), the definition of ξ, and $\sigma_1 \geq \sigma_k$

□

G.3 Proof of Corollary 7

Proof. Since $\mathcal{E}_U^{(t)}$ implies that $\mathcal{E}_{U,1}^{(t)},\ldots,\mathcal{E}_{U,6}^{(t)}$ hold with probability $1-6n^{-3}$, then combining Lemmas 28 and 29, we obtain

$$\begin{aligned}\|V^{(t+0.5)}-V^{*(t)}\|_{\mathrm{F}} &\le \sqrt{\delta_{2k}}\|V^{(t)}-V^{*(t)}\|_{\mathrm{F}}+\frac{2}{1+\delta_{2k}}\mathcal{E}(V^{(t+0.5)},V^{(t)},\overline{U}^{(t)})\\ &\le \sqrt{\delta_{2k}}\|V^{(t)}-V^{*(t)}\|_{\mathrm{F}}+\frac{2\sigma_k}{\xi}\|\overline{U}^{(t)}-\overline{U}^{*(t)}\|_{\mathrm{F}}\\ &\le \frac{\sigma_k^3}{12\xi\sigma_1^3}\cdot\frac{\sigma_k^2}{2\xi\sigma_1}+\frac{2\sigma_k^2}{\xi}\frac{\sigma_k}{4\xi\sigma_1^2}=\frac{\sigma_k^5}{24\xi^2\sigma_1^4}+\frac{\sigma_k^3}{2\xi^2\sigma_1^2}\le\frac{\sigma_k}{8}\end{aligned}$$

with probability $1-6n^{-3}$, where the last inequality comes from the definition of ξ and $\sigma_1\ge\sigma_k$. Thus by Lemma 30, we have

$$\begin{aligned}\|\overline{V}^{(t+1)}-\overline{V}^{*(t+1)}\|_{\mathrm{F}} &\le \frac{4\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)}-V^{*(t)}\|_{\mathrm{F}}+\frac{8}{\xi}\|\overline{U}^{(t)}-\overline{U}^{*(t)}\|_{\mathrm{F}}\\ &\le \frac{4}{\sigma_k}\left(\frac{\sigma_k^5}{24\xi^2\sigma_1^4}+\frac{\sigma_k^3}{2\xi^2\sigma_1^2}\right)=\frac{\sigma_k^4}{6\xi^2\sigma_1^4}+\frac{2\sigma_k^2}{\xi^2\sigma_1^2}\le\frac{\sigma_k^2}{4\xi\sigma_1^2},\end{aligned}$$

with probability $1-6n^{-3}$, where the last inequality comes from the definition of ξ and $\sigma_1\ge\sigma_k$. Moreover, by Lemma 30 again, we have

$$\begin{aligned}\|U^{(t)}-U^{*(t+1)}\|_{\mathrm{F}} &\le \frac{5\sigma_1\sqrt{\delta_{2k}}}{\sigma_k}\|V^{(t)}-V^{*(t)}\|_{\mathrm{F}}+\left(\frac{10}{\xi}+1\right)\sigma_1\|\overline{U}^{(t)}-\overline{U}^{*(t)}\|_{\mathrm{F}}\\ &\le \frac{5\sigma_1}{\sigma_k}\cdot\frac{\sigma_k^3}{12\xi\sigma_1^3}\cdot\frac{\sigma_k^2}{2\xi\sigma_1}+\left(\frac{10}{\xi}+1\right)\sigma_1\frac{\sigma_k^2}{4\xi\sigma_1^2}=\frac{5\sigma_k^4}{24\xi^2\sigma_1^3}+\frac{\sigma_k^2}{3\xi\sigma_1}\le\frac{\sigma_k^2}{2\xi\sigma_1}\end{aligned}$$

with probability $1-6n^{-3}$, where the last inequality comes from the definition of ξ and $\sigma_1\ge\sigma_k$. □

References

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

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. *arXiv preprint arXiv:1503.00778*, 2015.

Michel Baes. Estimate sequence methods: extensions and approximations. *Institute for Operations Research, ETH, Zürich, Switzerland*, 2009.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *arXiv preprint arXiv:1408.2156*, 2014.

Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. *arXiv preprint arXiv:1509.03917*, 2015.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

T Tony Cai and Anru Zhang. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

Emmanuel Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *arXiv preprint arXiv:1407.1065*, 2014.

Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6): 925–936, 2010.

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

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

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.

Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 42–50. ACM, 2011.

Yudong Chen. Incoherence-optimal matrix completion. *arXiv preprint arXiv:1310.0154*, 2013.

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

Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Coherent matrix completion. *arXiv preprint arXiv:1306.2979*, 2013a.

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

Alexandre d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

Michael P Friedlander and Mark Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, 2012.

Rainer Gemulla, Erik Nijkamp, Peter J Haas, and Yannis Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 69–77, 2011.

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

Osman Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

Moritz Hardt. Understanding alternating minimization for matrix completion. In *Symposium on Foundations of Computer Science*, pages 651–660, 2014.

Moritz Hardt and Mary Wootters. Fast matrix completion without the condition number. *arXiv preprint arXiv:1407.4070*, 2014.

Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. *arXiv preprint arXiv:1402.2331*, 2014.

Trevor Hastie, Rahul Mazumder, Jason Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *arXiv preprint arXiv:1410.2596*, 2014.

Cho-Jui Hsieh and Peder Olsen. Nuclear norm minimization via active subspace selection. In *International Conference on Machine Learning*, pages 575–583, 2014.

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

Prateek Jain and Praneeth Netrapalli. Fast exact matrix completion with finite samples. *arXiv preprint arXiv:1411.1087*, 2014.

Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Symposium on Theory of Computing*, pages 665–674, 2013.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010a.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010b.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Yehuda Koren. The bellkor solution to the netflix grand prize. *Netflix Prize Documentation*, 81, 2009.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 18:30–37, 2009.

Kiryung Lee and Yoram Bresler. Admira: Atomic decomposition for minimum rank approximation. *IEEE Transactions on Information Theory*, 56(9):4402–4416, 2010.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Angelia Nedić and Dimitri Bertsekas. Convergence rate of incremental subgradient algorithms. In *Stochastic optimization: algorithms and applications*, pages 223–264. Springer, 2001.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

Arkadiusz Paterek. Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD Cup and workshop*, volume 2007, pages 5–8, 2007.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

Benjamin Recht and Christopher Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

Gilbert W Stewart, Ji-guang Sun, and Harcourt Brace Jovanovich. *Matrix perturbation theory*, volume 175. Academic press New York, 1990.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *arXiv preprint arXiv:1411.8003*, 2014.

Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Major components of the gravity recommendation system. *ACM SIGKDD Explorations Newsletter*, 9(2):80–83, 2007.

Stephen Tu, Ross Boczar, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

Zheng Wang, Ming-Jun Lai, Zhaosong Lu, Wei Fan, Hasan Davulcu, and Jieping Ye. Orthogonal rank-one matrix pursuit for low rank matrix completion. *SIAM Journal on Scientific Computing*, 37(1):A488–A514, 2015.

Shuo Xiang, Yunzhang Zhu, Xiaotong Shen, and Jieping Ye. Optimal exact least squares rank minimization. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 480–488. ACM, 2012.

Qi Yan, Jieping Ye, and Xiaotong Shen. Simultaneous pursuit of sparseness and rank structures for matrix decomposition. *Journal of Machine Learning Research*, 16:47–75, 2015.

Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix factorization. *Advances in Neural Information Processing Systems*, 2015. To appear.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. *arXiv preprint arXiv:1506.06081*, 2015.

Yunzhang Zhu, Xiaotong Shen, and Changqing Ye. Personalized prediction and sparsity pursuit in latent factor models. *Journal of the American Statistical Association*, 2015. To appear.

Yong Zhuang, Wei-Sheng Chin, Yu-Chin Juan, and Chih-Jen Lin. A fast parallel sgd for matrix factorization in shared memory systems. In *ACM Conference on Recommender Systems*, pages 249–256, 2013.