Noname manuscript No.
(will be inserted by the editor)

Provable Accelerated Gradient Method for Nonconvex Low Rank Optimization

Huan Li · Zhouchen Lin

Received: date / Accepted: date

Abstract Optimization over low rank matrices has broad applications in machine learning. For large scale problems, an attractive heuristic is to factorize the low rank matrix to a product of two much smaller matrices. In this paper, we study the nonconvex problem $\min_{\mathbf{U}\in\mathbb{R}^{n\times r}} g(\mathbf{U}) = f(\mathbf{U}\mathbf{U}^T)$ under the assumptions that $f(\mathbf{X})$ is restricted μ-strongly convex and L-smooth on the set $\{\mathbf{X} : \mathbf{X} \succeq 0, \mathrm{rank}(\mathbf{X}) \leq r\}$. We propose an accelerated gradient method with alternating constraint that operates directly on the $\mathbf{U}$ factors and show that the method has local linear convergence rate with the optimal dependence on the condition number of $\sqrt{L/\mu}$. Globally, our method converges to the critical point with zero gradient from any initializer. Our method also applies to the problem with the asymmetric factorization of $\mathbf{X} = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T$ and the same convergence result can be obtained. Extensive experimental results verify the advantage of our method.

1 Introduction

Low rank matrix estimation has broad applications in machine learning, computer vision and signal processing. In this paper, we consider the problem of the form:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times n}} f(\mathbf{X}), \quad s.t. \quad \mathbf{X} \succeq 0, \tag{1}$$

Huan Li
School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China.
E-mail: lihuanss@pku.edu.cn

Zhouchen Lin
Key Lab. of Machine Perception (MOE), School of EECS, Peking University, Beijing, China.
Z. Lin is the corresponding author.
E-mail: zlin@pku.edu.cn

where there exists minimizer $\mathbf{X}^*$ of rank-r. We consider the case of $r \ll n$. Optimizing problem (1) in the $\mathbf{X}$ space often requires computing at least the top-r singular value/vectors in each iteration and $O(n^2)$ memory to store a large n by n matrix, which restricts the applications with huge size matrices. To reduce the computational cost as well as the storage space, many literatures exploit the observation that a positive semidefinite low rank matrix can be factorized as a product of two much smaller matrices, i.e., $\mathbf{X} = \mathbf{U}\mathbf{U}^T$, and study the following nonconvex problem instead:

$$\min_{\mathbf{U}\in\mathbb{R}^{n\times r}} g(\mathbf{U}) = f(\mathbf{U}\mathbf{U}^T). \tag{2}$$

A wide family of problems can be cast as problem (2), including matrix sensing [1], matrix completion [2], one bit matrix completion [3], sparse principle component analysis [4] and factorization machine [5]. In this paper, we study problem (2) and aim to propose an accelerated gradient method that operates on the $\mathbf{U}$ factors directly. The factorization in problem (2) makes $g(\mathbf{U})$ nonconvex, even if $f(\mathbf{X})$ is convex. Thus, proving the acceleration becomes a harder task than the analysis for convex programming.

1.1 Related Work

Recently, there is a trend to study the nonconvex problem (2) in the machine learning and optimization community. Recent developments come from two aspects: (1). The geometric aspect which proves that there is no spurious local minimum for some special cases of problem (2), e.g., matrix sensing [1], matrix completion [6], and [7, 8, 9, 10] for a unified analysis. (2). The algorithmic aspect which analyzes the local linear convergence of some efficient schemes such as the gradient descent method. Examples include [11, 12, 13, 14, 15, 16] for semidefinite programs, [17, 18, 19, 20, 21] for matrix completion, [21, 18] for matrix sensing and [22, 23] for Robust PCA. The local linear convergence rate of the gradient descent method is proved for problem (2) in a unified framework in [24, 25, 26]. However, no acceleration scheme is studied in these literatures. It remains an open problem on how to analyze the accelerated gradient method for nonconvex problem (2).

Nesterov's acceleration technique [27, 28, 29] has been empirically verified efficient on some nonconvex problems, e.g., Deep Learning [30]. Several literatures studied the accelerated gradient method and the inertial gradient descent method for the general nonconvex programming [31, 32, 33]. However, they only proved the convergence and had no guarantee on the acceleration for nonconvex problems. Carmon et al. [34, 35], Agarwal et al. [36] and Jin et al. [37] analyzed the accelerated gradient method for the general nonconvex optimization and proved the complexity of $O(\epsilon^{-7/4}\log(1/\epsilon))$ to escape saddle points or achieve critical points. They studied the general problem and did not exploit the specification of problem (2). Thus, their complexity is sublinear. Necoara et al. [38] studied several conditions under which the gradient descent

and accelerated gradient method converge linearly for non-strongly convex optimization. Their conclusion of the gradient descent method can be extended to nonconvex problem (2). For the accelerated gradient method, Necoara et al. required a strong assumption that all $\mathbf{y}^k, k = 0, 1, \cdots,$[1] have the same projection onto the optimum solution set. It does not hold for problem (2).

1.2 Our Contributions

In this paper, we use Nesterov's acceleration scheme for problem (2) and an efficient accelerated gradient method with alternating constraint is proposed, which operates on the $\mathbf{U}$ factors directly. We back up our method with provable theoretical results. Specifically, our contributions can be summarized as follows:

1. We establish the curvature of local restricted strong convexity along a certain trajectory by restricting the problem onto a constraint set, which allows us to use the classical accelerated gradient method for convex programs to solve the constrained problem. We build our result with the tool of polar decomposition.
2. In order to reduce the negative influence of the constraint and ensure the convergence to the critical point of the original unconstrained problem, rather than the reformulated constrained problem, we propose a novel alternating constraint strategy and combine it with the classical accelerated gradient method.
3. When f is restricted μ-strongly convex and restricted L-smooth, our method has the local linear convergence to the optimum solution, which has the same dependence on $\sqrt{L/\mu}$ as convex programming. As far as we know, we are the first to establish the convergence matching the optimal dependence on $\sqrt{L/\mu}$ for this kind of nonconvex problems. Globally, our method converges to a critical point of problem (2) from any initializer.

1.3 Notations and Assumptions

For matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n\times r}$, we use $\|\mathbf{U}\|_F$ as the Frobenius norm, $\|\mathbf{U}\|_2$ as the spectral norm and $\langle \mathbf{U}, \mathbf{V}\rangle = \text{trace}(\mathbf{U}^T\mathbf{V})$ as their inner products. We denote $\sigma_r(\mathbf{U})$ as the smallest singular value of $\mathbf{U}$ and $\sigma_1(\mathbf{U}) = \|\mathbf{U}\|_2$ as the largest one. We use $\mathbf{U}_S \in \mathbb{R}^{r\times r}$ as the submatrix of $\mathbf{U}$ with the rows indicated by the index set $S \subseteq \{1, 2, \cdots, n\}$, $\mathbf{U}_{-S} \in \mathbb{R}^{(n-r)\times r}$ as the submatrix with the rows indicated by the indexes out of S and $\mathbf{X}_{S,S} \in \mathbb{R}^{r\times r}$ as the submatrix of $\mathbf{X}$ with the rows and columns indicated by S. $\mathbf{X} \succeq 0$ means that $\mathbf{X}$ is symmetric and positive semidefinite. Let $I_{\Omega_S}(\mathbf{U})$ be the indicator function of set Ω_S. For the objective function $g(\mathbf{U})$, its gradient w.r.t. $\mathbf{U}$ is $\nabla g(\mathbf{U}) = 2\nabla f(\mathbf{U}\mathbf{U}^T)\mathbf{U}$. We assume that $\nabla f(\mathbf{U}\mathbf{U}^T)$ is symmetric for simplicity.

[1] Necoara et al. [38] analyzed the method with recursions of $\mathbf{y}^k = \mathbf{x}^k + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}(\mathbf{x}^k - \mathbf{x}^{k-1})$ and $\mathbf{x}^{k+1} = \mathbf{y}^k - \eta\nabla f(\mathbf{y}^k)$.

Our conclusions for the asymmetric case naturally generalize since $\nabla g(\mathbf{U}) = \nabla f(\mathbf{U}\mathbf{U}^T)\mathbf{U} + \nabla f(\mathbf{U}\mathbf{U}^T)^T\mathbf{U}$ in this case. Denote the optimum solution set of problem (2) as

$$\mathcal{X}^* = \{\mathbf{U}^* : \mathbf{U}^* \in \mathbb{R}^{n\times r}, \mathbf{U}^*\mathbf{U}^{*T} = \mathbf{X}^*\}. \tag{3}$$

where $\mathbf{X}^*$ is a minimizer of problem (1). An important issue in minimizing $g(\mathbf{U})$ is that its optimum solution is not unique, i.e., if $\mathbf{U}^*$ is the optimum solution of problem (2), then $\mathbf{U}^*\mathbf{R}$ is also an optimum solution for any orthogonal matrix $\mathbf{R} \in \mathbb{R}^{r\times r}$. Given $\mathbf{U}$, we define the optimum solution that is closest to $\mathbf{U}$ as

$$P_{\mathcal{X}^*}(\mathbf{U}) = \mathbf{U}^*\mathbf{R}, \text{ where } \mathbf{R} = \text{argmin}_{\mathbf{R}\in\mathbb{R}^{r\times r}, \mathbf{R}\mathbf{R}^T=\mathbf{I}}\|\mathbf{U}^*\mathbf{R} - \mathbf{U}\|_F^2. \tag{4}$$

1.3.1 Assumptions

In this paper, we assume that f is restricted μ-strongly convex and L-smooth on the set $\{\mathbf{X} : \mathbf{X} \succeq 0, \text{rank}(\mathbf{X}) \leq r\}$. We state the standard definitions below.

Definition 1 Let $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ be a convex differentiable function. Then, f is restricted μ-strongly convex on the set $\{\mathbf{X} : \mathbf{X} \succeq 0, \text{rank}(\mathbf{X}) \leq r\}$ if, for any $\mathbf{X}, \mathbf{Y} \in \{\mathbf{X} : \mathbf{X} \succeq 0, \text{rank}(\mathbf{X}) \leq r\}$, we have

$$f(\mathbf{Y}) \geq f(\mathbf{X}) + \langle \nabla f(\mathbf{X}), \mathbf{Y} - \mathbf{X}\rangle + \tfrac{\mu}{2}\|\mathbf{Y} - \mathbf{X}\|_F^2.$$

Definition 2 Let $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ be a convex differentiable function. Then, f is restricted L-smooth on the set $\{\mathbf{X} : \mathbf{X} \succeq 0, \text{rank}(\mathbf{X}) \leq r\}$ if, for any $\mathbf{X}, \mathbf{Y} \in \{\mathbf{X} : \mathbf{X} \succeq 0, \text{rank}(\mathbf{X}) \leq r\}$, we have

$$f(\mathbf{Y}) \leq f(\mathbf{X}) + \langle \nabla f(\mathbf{X}), \mathbf{Y} - \mathbf{X}\rangle + \tfrac{L}{2}\|\mathbf{Y} - \mathbf{X}\|_F^2$$

and

$$\|\nabla f(\mathbf{Y}) - \nabla f(\mathbf{X})\|_F \leq L\|\mathbf{Y} - \mathbf{X}\|_F.$$

1.3.2 Polar decomposition

Polar decomposition is a powerful tool for matrix analysis. We briefly review it in this section. We only describe the left polar decomposition of a square matrix.

Definition 3 The polar decomposition of a matrix $\mathbf{A} \in \mathbb{R}^{r\times r}$ has the form $\mathbf{A} = \mathbf{H}\mathbf{Q}$ where $\mathbf{H} \in \mathbb{R}^{r\times r}$ is positive semidefinite and $\mathbf{Q} \in \mathbb{R}^{r\times r}$ is an orthogonal matrix.

If $\mathbf{A} \in \mathbb{R}^{r\times r}$ is of full rank, then $\mathbf{A}$ has the unique polar decomposition with positive definite $\mathbf{H}$. In fact, since a positive semidefinite Hermitian matrix has a unique positive semidefinite square root, $\mathbf{H}$ is uniquely given by $\mathbf{H} = \sqrt{\mathbf{A}\mathbf{A}^T}$. $\mathbf{Q} = \mathbf{H}^{-1}\mathbf{A}$ is also unique.

In this paper, we use the tool of polar decomposition's perturbation theorem to build the restricted strong convexity of $g(\mathbf{U})$. It is described below.

Lemma 1 *[39] Let* $\mathbf{A} \in \mathbb{R}^{r\times r}$ *be of full rank and* $\mathbf{HQ}$ *be its unique polar decomposition,* $\mathbf{A}+\triangle\mathbf{A}$ *be of full rank and* $(\mathbf{H}+\triangle\mathbf{H})(\mathbf{Q}+\triangle\mathbf{Q})$ *be its unique polar decomposition. Then, we have*

$$\|\triangle\mathbf{Q}\|_F \leq \frac{2}{\sigma_r(\mathbf{A})}\|\triangle\mathbf{A}\|_F.$$

2 The Restricted Strongly Convex Curvature

Function $g(\mathbf{U})$ is a special kind of nonconvex function and the non-convexity only comes from the factorization of $\mathbf{U}\mathbf{U}^T$. Based on this observation, we exploit the special curvature of $g(\mathbf{U})$ in this section.

The existing works proved the local linear convergence of the *gradient descent method* for problem (2) by exploiting curvatures such as the local second order growth property [17, 25] or the (α,β) regularity condition [40, 24, 1, 26]. The former is described as

$$g(\mathbf{U}) \geq g(\mathbf{U}^*) + \frac{\alpha}{2}\|P_{\mathcal{X}^*}(\mathbf{U})-\mathbf{U}\|_F^2, \forall \mathbf{U} \tag{5}$$

while the later is defined as

$$\langle \nabla g(\mathbf{U}), \mathbf{U}-P_{\mathcal{X}^*}(\mathbf{U})\rangle \geq \frac{\alpha}{2}\|P_{\mathcal{X}^*}(\mathbf{U})-\mathbf{U}\|_F^2 + \frac{1}{2\beta}\|\nabla g(\mathbf{U})\|_F^2, \forall \mathbf{U}, \tag{6}$$

where $\mathbf{U}^* \in \mathcal{X}^*$ and $P_{\mathcal{X}^*}(\mathbf{U})$ is defined in (4). Both (5) and (6) can be derived by the local weakly strongly convex condition [38] combing with the smoothness of $g(\mathbf{U})$. The former is described as

$$g(\mathbf{U}^*) \geq g(\mathbf{U}) + \langle \nabla g(\mathbf{U}), P_{\mathcal{X}^*}(\mathbf{U})-\mathbf{U}\rangle + \frac{\alpha}{2}\|P_{\mathcal{X}^*}(\mathbf{U})-\mathbf{U}\|_F^2, \tag{7}$$

where $\alpha = \mu\sigma_r^2(\mathbf{U}^*)$. As discussed in Section 1.3, the optimum solution of problem (2) is not unique. This non-uniqueness makes the difference between the weakly strong convexity and strong convexity, e.g., on the right hand side of (7), we use $P_{\mathcal{X}^*}(\mathbf{U})$, rather than $\mathbf{U}^*$. Moreover, the weakly strongly convex condition cannot infer convexity and $g(\mathbf{U})$ is not convex even around a small neighborhood of the global optimum solution [41].

Necoara, Nesterov and Glineur [38] studied several conditions under which the linear convergence of the *gradient descent method* is guaranteed for general convex programming without strong convexity. The weakly strongly convex condition is the strongest one and can derive all the other conditions. However, it is not enough to analyze the accelerated gradient method only with the weakly strongly convex condition. Necoara et al. [38] proved the acceleration of the classical *accelerated gradient method* under an additional assumption that all the iterates $\{\mathbf{y}^k, k=0,1,\cdots\}$ have the same projection onto the optimum solution set besides the weakly strongly convex condition and the smoothness condition. From the proof in [38, Section 5.2.1], we can see that the non-uniqueness of the optimum solution makes the main trouble to analyze

the accelerated gradient method[2]. The additional assumption made in [38] somehow aims to reduce this non-uniqueness. Since this assumption is not satisfied for problem (2), only (7) is not enough to prove the acceleration for problem (2) and it requires us to exploit stronger curvature than (7) to analyze the accelerated gradient method.

Motivated by [38], we should remove the non-uniqueness in problem (2). Our intuition is based on the following observation. Suppose that we can find an index set $S \subseteq \{1, 2, \cdots, n\}$ with size r such that $\mathbf{X}^*_{S,S}$ is of r full rank, then there exists a unique decomposition $\mathbf{X}^*_{S,S} = \mathbf{U}^*_S(\mathbf{U}^*_S)^T$ where we require $\mathbf{U}^*_S \succ 0$. Thus, we can easily have that there exists a unique $\mathbf{U}^*$ such that $\mathbf{U}^*\mathbf{U}^{*T} = \mathbf{X}^*$ and $\mathbf{U}^*_S \succ 0$. To verify it, consider $S = \{1, \cdots, r\}$ for simplicity. Then $\mathbf{U}\mathbf{U}^T = \begin{pmatrix} \mathbf{U}_S\mathbf{U}_S^T & \mathbf{U}_S\mathbf{U}_{-S}^T \\ \mathbf{U}_{-S}\mathbf{U}_S^T & \mathbf{U}_{-S}\mathbf{U}_{-S}^T \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{S,S} & \mathbf{X}_{S,-S} \\ \mathbf{X}_{-S,S} & \mathbf{X}_{-S,-S} \end{pmatrix}$. The uniqueness of $\mathbf{U}_S$ comes from $\mathbf{X}_{S,S} \succ 0$ and $\mathbf{U}_S \succ 0$ and the uniqueness of $\mathbf{U}_{-S}$ comes from $\mathbf{U}_{-S} = \mathbf{X}_{-S,S}\mathbf{U}_S^{-T}$.

Based on the above observation, we can reformulate problem (2) as

$$\min_{\mathbf{U} \in \Omega_S} g(\mathbf{U}) \tag{8}$$

where

$$\Omega_S = \{\mathbf{U} \in \mathbb{R}^{n \times r} : \mathbf{U}_S \succeq \epsilon \mathbf{I}\}$$

and ϵ is a small enough constant such that $\epsilon \ll \sigma_r(\mathbf{U}^*_S)$. We require $\mathbf{U}_S \succeq \epsilon\mathbf{I}$ rather than $\mathbf{U}_S \succ 0$ to make the projection onto Ω_S computable. Due to the additional constraint of $\mathbf{U} \in \Omega_S$, we observe that the optimum solution of problem (8) is unique. Moreover, the minimizer of (8) minimizes also (2).

Until now, we are ready to establish a stronger curvature than (7) by restricting the variables of $g(\mathbf{U})$ on the set Ω_S. We should lower bound $\|P_{\mathcal{X}^*}(\mathbf{U}) - \mathbf{U}\|_F^2$ in (7) by $\|\mathbf{U}^* - \mathbf{U}\|_F^2$. Our result is built upon polar decomposition's perturbation theorem [39]. Based on Lemma 1, we first establish the following critical lemma.

Lemma 2 *For any $\mathbf{U} \in \Omega_S$ and $\mathbf{V} \in \Omega_S$, let $\mathbf{R} = argmin_{\mathbf{R} \in \mathbb{R}^{r \times r}, \mathbf{R}\mathbf{R}^T = \mathbf{I}} \|\mathbf{V}\mathbf{R} - \mathbf{U}\|_F^2$ and $\hat{\mathbf{V}} = \mathbf{V}\mathbf{R}$. Then, we have*

$$\|\mathbf{V} - \mathbf{U}\|_F \leq \frac{3\|\mathbf{U}\|_2}{\sigma_r(\mathbf{U}_S)} \|\hat{\mathbf{V}} - \mathbf{U}\|_F.$$

Proof Since the conclusion is not affected by permutating the rows of $\mathbf{U}$ and $\mathbf{V}$ under the same permutation, we can consider the case of $S = \{1, \cdots, r\}$ for simplicity. Let $\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix}$, $\mathbf{V} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{pmatrix}$ and $\hat{\mathbf{V}} = \begin{pmatrix} \hat{\mathbf{V}}_1 \\ \hat{\mathbf{V}}_2 \end{pmatrix}$, where $\mathbf{U}_1, \mathbf{V}_1, \hat{\mathbf{V}}_1 \in \mathbb{R}^{r \times r}$. Then, we have $\hat{\mathbf{V}}_1 = \mathbf{V}_1\mathbf{R}$. From $\mathbf{U} \in \Omega_S$ and $\mathbf{V} \in \Omega_S$, we know $\mathbf{U}_1 \succ 0$

[2] [38] used induction to prove [38, Lemma 1]. When the optimum solution is not unique, $\mathbf{y}^*$ in [38, Equation (57)] should be replaced by $P_{\mathcal{X}^*}(\mathbf{y}^k)$ and they have different values for different k. Thus, the induction is not correct any more.

and $\mathbf{V}_1 \succ 0$. Thus, $\mathbf{U}_1\mathbf{I}$ and $\mathbf{V}_1\mathbf{R}$ are the unique polar decompositions of $\mathbf{U}_1$ and $\hat{\mathbf{V}}_1$, respectively. From Lemma 1, we have

$$\|\mathbf{R}-\mathbf{I}\|_F \leq \frac{2}{\sigma_r(\mathbf{U}_1)}\|\hat{\mathbf{V}}_1-\mathbf{U}_1\|_F.$$

With some simple computations, we can have

$$\begin{aligned}\|\mathbf{V}-\mathbf{U}\|_F =&\|\hat{\mathbf{V}}\mathbf{R}^T-\mathbf{U}\|_F\\ =&\|\hat{\mathbf{V}}\mathbf{R}^T-\mathbf{U}\mathbf{R}^T+\mathbf{U}\mathbf{R}^T-\mathbf{U}\|_F\\ \leq&\|\hat{\mathbf{V}}\mathbf{R}^T-\mathbf{U}\mathbf{R}^T\|_F+\|\mathbf{U}\mathbf{R}^T-\mathbf{U}\|_F\\ \leq&\|\hat{\mathbf{V}}-\mathbf{U}\|_F+\|\mathbf{U}\|_2\|\mathbf{R}-\mathbf{I}\|_F\\ \leq&\|\hat{\mathbf{V}}-\mathbf{U}\|_F+\frac{2\|\mathbf{U}\|_2}{\sigma_r(\mathbf{U}_1)}\|\hat{\mathbf{V}}_1-\mathbf{U}_1\|_F\\ \leq&\frac{3\|\mathbf{U}\|_2}{\sigma_r(\mathbf{U}_1)}\|\hat{\mathbf{V}}-\mathbf{U}\|_F,\end{aligned} \tag{9}$$

where we use $\sigma_r(\mathbf{U}_1) \leq \|\mathbf{U}\|_2$ and $\|\hat{\mathbf{V}}_1-\mathbf{U}_1\|_F \leq \|\hat{\mathbf{V}}-\mathbf{U}\|_F$ in the last inequality. Replacing $\mathbf{U}_1$ with $\mathbf{U}_S$, we can have the conclusion. □

Built upon Lemma 2, we can give the local restricted strong convexity of $g(\mathbf{U})$ on the set Ω_S in the following theorem. There are two differences between the restricted strong convexity and the weakly strong convexity: (i) the restricted strong convexity removes the non-uniqueness and (ii) the restricted strong convexity establishes the curvature between any two points $\mathbf{U}$ and $\mathbf{V}$ in a local neighborhood of $\mathbf{U}^*$, while (7) only exploits the curvature between $\mathbf{U}$ and the optimum solution.

Theorem 1 *Let $\mathbf{U}^* = \Omega_S \cap \mathcal{X}^*$ and assume that $\mathbf{U} \in \Omega_S$ and $\mathbf{V} \in \Omega_S$ with $\|\mathbf{U}-\mathbf{U}^*\|_F \leq C$ and $\|\mathbf{V}-\mathbf{U}^*\|_F \leq C$, where $C = \frac{\mu\sigma_r^2(\mathbf{U}^*)\sigma_r^2(\mathbf{U}_S^*)}{100L\|\mathbf{U}^*\|_2^3}$. Then, we have*

$$g(\mathbf{U}) \geq g(\mathbf{V}) + \langle\nabla g(\mathbf{V}), \mathbf{U}-\mathbf{V}\rangle + \frac{\mu\sigma_r^2(\mathbf{U}^*)\sigma_r^2(\mathbf{U}_S^*)}{50\|\mathbf{U}^*\|_2^2}\|\mathbf{U}-\mathbf{V}\|_F^2.$$

Proof From the restricted convexity of $f(\mathbf{X})$, we have

$$\begin{aligned}&f(\mathbf{V}\mathbf{V}^T)-f(\mathbf{U}\mathbf{U}^T)\\ \leq&\left\langle\nabla f(\mathbf{V}\mathbf{V}^T), \mathbf{V}\mathbf{V}^T-\mathbf{U}\mathbf{U}^T\right\rangle - \frac{\mu}{2}\|\mathbf{V}\mathbf{V}^T-\mathbf{U}\mathbf{U}^T\|_F^2\\ =&\left\langle\nabla f(\mathbf{V}\mathbf{V}^T), (\mathbf{V}-\mathbf{U})\mathbf{V}^T\right\rangle + \left\langle\nabla f(\mathbf{V}\mathbf{V}^T), \mathbf{V}(\mathbf{V}-\mathbf{U})^T\right\rangle\\ &-\left\langle\nabla f(\mathbf{V}\mathbf{V}^T), (\mathbf{V}-\mathbf{U})(\mathbf{V}-\mathbf{U})^T\right\rangle - \frac{\mu}{2}\|\mathbf{V}\mathbf{V}^T-\mathbf{U}\mathbf{U}^T\|_F^2\\ =&2\left\langle\nabla f(\mathbf{V}\mathbf{V}^T)\mathbf{V}, \mathbf{V}-\mathbf{U}\right\rangle - \left\langle\nabla f(\mathbf{V}\mathbf{V}^T), (\mathbf{V}-\mathbf{U})(\mathbf{V}-\mathbf{U})^T\right\rangle\\ &-\frac{\mu}{2}\|\mathbf{V}\mathbf{V}^T-\mathbf{U}\mathbf{U}^T\|_F^2\\ \leq&2\left\langle\nabla f(\mathbf{V}\mathbf{V}^T)\mathbf{V}, \mathbf{V}-\mathbf{U}\right\rangle - \left\langle\nabla f(\mathbf{V}\mathbf{V}^T)-\nabla f(\mathbf{X}^*), (\mathbf{V}-\mathbf{U})(\mathbf{V}-\mathbf{U})^T\right\rangle\\ &-\frac{\mu}{2}\|\mathbf{V}\mathbf{V}^T-\mathbf{U}\mathbf{U}^T\|_F^2.\end{aligned} \tag{10}$$

where we use $\nabla f(\mathbf{X}^*) \succeq 0$ proved in Lemma 7 and the fact that the inner product of two positive semidefinite matrices is nonnegative in the last inequality, i.e., $\left\langle \nabla f(\mathbf{X}^*), (\mathbf{V}-\mathbf{U})(\mathbf{V}-\mathbf{U})^T \right\rangle \geq 0$. Applying Von Neumann's trace inequality and Lemma 10 to bound the second term, applying Lemmas 2 and 8 to bound the third term, we can have

$$\begin{aligned}
&f(\mathbf{V}\mathbf{V}^T) - f(\mathbf{U}\mathbf{U}^T) \\
\leq\ & 2\left\langle \nabla f(\mathbf{V}\mathbf{V}^T)\mathbf{V}, \mathbf{V}-\mathbf{U} \right\rangle + L(\|\mathbf{U}^*\|_2 + \|\mathbf{V}\|_2)\|\mathbf{V}-\mathbf{U}^*\|_F\|\mathbf{V}-\mathbf{U}\|_F^2 \\
& - \frac{(\sqrt{2}-1)\mu\sigma_r^2(\mathbf{U})\sigma_r^2(\mathbf{U}_S)}{9\|\mathbf{U}\|_2^2}\|\mathbf{V}-\mathbf{U}\|_F^2 \\
\leq\ & \langle \nabla g(\mathbf{V}), \mathbf{V}-\mathbf{U}\rangle - \left(\frac{\mu\sigma_r^2(\mathbf{U}^*)\sigma_r^2(\mathbf{U}_S^*)}{23.1\|\mathbf{U}^*\|_2^2} - 2.01L\|\mathbf{U}^*\|_2\|\mathbf{V}-\mathbf{U}^*\|_F\right)\|\mathbf{V}-\mathbf{U}\|_F^2,
\end{aligned}$$

where we use Lemma 9 in the last inequality. From the assumption of $\|\mathbf{V}-\mathbf{U}^*\|_F \leq C$, we can have the conclusion. We leave Lemmas 7, 8, 9 and 10 in Appendix A. □

2.1 Smoothness of Function $g(\mathbf{U})$

Besides the local restricted strong convexity, we can also prove the smoothness of $g(\mathbf{U})$, which is built in the following theorem.

Theorem 2 *Let $\hat{L} = 2\|\nabla f(\mathbf{V}\mathbf{V}^T)\|_2 + L(\|\mathbf{V}\|_2 + \|\mathbf{U}\|_2)^2$. Then, we can have*

$$g(\mathbf{U}) \leq g(\mathbf{V}) + \langle \nabla g(\mathbf{V})\mathbf{V}, \mathbf{U}-\mathbf{V}\rangle + \frac{\hat{L}}{2}\|\mathbf{U}-\mathbf{V}\|_F^2.$$

Proof From the restricted Lipschitz smoothness of f and a similar induction to (10), we have

$$\begin{aligned}
&f(\mathbf{U}\mathbf{U}^T) - f(\mathbf{V}\mathbf{V}^T) \\
\leq\ & \left\langle \nabla f(\mathbf{V}\mathbf{V}^T), \mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T \right\rangle + \frac{L}{2}\|\mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T\|_F^2 \\
=\ & \left\langle \nabla f(\mathbf{V}\mathbf{V}^T), (\mathbf{U}-\mathbf{V})(\mathbf{U}-\mathbf{V})^T \right\rangle \\
& + 2\left\langle \nabla f(\mathbf{V}\mathbf{V}^T)\mathbf{V}, \mathbf{U}-\mathbf{V} \right\rangle + \frac{L}{2}\|\mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T\|_F^2.
\end{aligned}$$

Applying Von Neumann's trace inequality to the first term, applying Lemma 10 to the third term, we can have the conclusion. □

When restricted in a small neighborhood of $\mathbf{U}^*$, we can give a better estimate for the smoothness parameter $\hat{L}$, as follows. The proof is provided in Appendix A.

Corollary 1 *Let* $\mathbf{U}^* = \Omega_S \cap \mathcal{X}^*$ *and assume that* $\mathbf{U}^k, \mathbf{V}^k, \mathbf{Z}^k \in \Omega_S$ *with* $\|\mathbf{V}^k - \mathbf{U}^*\|_F \leq C$, $\|\mathbf{U}^k - \mathbf{U}^*\|_F \leq C$ *and* $\|\mathbf{Z}^k - \mathbf{U}^*\|_F \leq C$, *where* C *is defined in Theorem 1 and* $\mathbf{U}^k, \mathbf{V}^k, \mathbf{Z}^k$ *are generated in Algorithm 1, which will be described later. Let* $L_g = 38L\|\mathbf{U}^*\|_2^2 + 2\|\nabla f(\mathbf{X}^*)\|_2$ *and* $\eta = \frac{1}{L_g}$. *Then, we have*

$$g(\mathbf{U}^{k+1}) \leq g(\mathbf{V}^k) + \left\langle \nabla g(\mathbf{V}^k), \mathbf{U}^{k+1} - \mathbf{V}^k \right\rangle + \frac{L_g}{2}\|\mathbf{U}^{k+1} - \mathbf{V}^k\|_F^2.$$

3 Accelerated Gradient Method with Alternating Constraint

From Theorem 1 and Corollary 1, we know that the objective $g(\mathbf{U})$ behaves locally like a strongly convex and smooth function when restricted on the set Ω_S. Thus, we can use the classical method for convex programming to solve problem (8), e.g., the accelerated gradient method[3].

However, there remains a practical issue that when solving problem (8), we may get stuck at a critical point of problem (8) at the boundary of the constraint $\mathbf{U} \in \Omega_S$, which is not the optimum solution of problem (2). In other words, we may halt before reaching the acceleration region, i.e., the local neighborhood of the optimum solution of problem (2). To overcome this trouble, we propose a novel alternating trajectory strategy. Specifically, we define two sets Ω_{S^1} and Ω_{S^2} as follows

$$\Omega_{S^1} = \{\mathbf{U} \in \mathbb{R}^{n\times r} : \mathbf{U}_{S^1} \succeq \epsilon\mathbf{I}\}, \quad \Omega_{S^2} = \{\mathbf{U} \in \mathbb{R}^{n\times r} : \mathbf{U}_{S^2} \succeq \epsilon\mathbf{I}\}$$

and minimize the objective $g(\mathbf{U})$ along the trajectories of Ω_{S^1} and Ω_{S^2} alternatively, i.e., when the iteration number t is odd, we minimize $g(\mathbf{U})$ with the constraint of $\mathbf{U} \in \Omega_{S^1}$, and when t is even, we minimize $g(\mathbf{U})$ with the constraint of $\mathbf{U} \in \Omega_{S^2}$. Intuitively, when the iterates approach the boundary of Ω_{S^1}, we cancel the constraint of positive definiteness on $\mathbf{U}_{S^1}$ and put it on $\mathbf{U}_{S^2}$. Fortunately, with this strategy we can cancel the negative influence of the constraint. We require that both the two index sets S^1 and S^2 are of size r and $S^1 \cap S^2 = \emptyset$ such that $\mathbf{U}^*_{S^1}$ and $\mathbf{U}^*_{S^2}$ are of full rank. Given proper S^1 and S^2, we can prove that the method globally converges to a critical point of problem (2). i.e., a point with $\nabla g(\mathbf{U}) = 0$, rather than a critical point of problem (8) at the boundary of the constraint.

We describe our method in Algorithm 1. We use Nesterov's acceleration scheme in the inner loop with finite K iterations and restart the acceleration scheme at each outer iteration. At the end of each outer iteration, we change the constraint and transform $\mathbf{U}^{t,K+1} \in \Omega_S$ to a new point $\mathbf{U}^{t+1,0} \in \Omega_{S'}$ via polar decomposition such that $g(\mathbf{U}^{t,K+1}) = g(\mathbf{U}^{t+1,0})$. At step (13), we need to project $\mathbf{Z} \equiv \mathbf{Z}^{t,k} - \frac{\eta}{\theta_k}\nabla g(\mathbf{V}^{t,k})$ onto Ω_S. Let $\mathbf{A}\Sigma\mathbf{A}^T$ be the eigenvalue

[3] However, it is still more challenging than convex programming since we should guarantee that all the variables in Theorem 1 belong to Ω_S, while it is not required in convex programming. So the conclusion in [38] cannot be applied to problem (8) since we cannot obtain $\mathbf{y}^{k+1} \in \Omega_S$ given $\mathbf{x}^{k+1} \in \Omega_S$ and $\mathbf{x}^k \in \Omega_S$ because $\mathbf{y}^{k+1}$ is not a convex combination of $\mathbf{x}^{k+1}$ and $\mathbf{x}^k$.

Algorithm 1 Accelerated Gradient Descent with Alternating Constraint

Initialize $\mathbf{Z}^{0,0} = \mathbf{U}^{0,0} \in \Omega_{S^2}$, η, K, ϵ.
for $t = 0, 1, 2, \cdots$ **do**
  $\theta_0 = 1$.
  for $k = 0, 1, \ldots, K$ **do**

$$\mathbf{V}^{t,k} = (1-\theta_k)\mathbf{U}^{t,k} + \theta_k \mathbf{Z}^{t,k}. \tag{12}$$

$$\mathbf{Z}^{t,k+1} = \mathrm{argmin}_{\mathbf{Z}\in\Omega_S} \left\langle \nabla g(\mathbf{V}^{t,k}), \mathbf{Z} \right\rangle + \frac{\theta_k}{2\eta} \left\| \mathbf{Z} - \mathbf{Z}^{t,k} \right\|_F^2, S = \begin{cases} S^1, & \text{if } t \text{ is odd,} \\ S^2, & \text{if } t \text{ is even.} \end{cases} \tag{13}$$

$$\mathbf{U}^{t,k+1} = (1-\theta_k)\mathbf{U}^{t,k} + \theta_k \mathbf{Z}^{t,k+1}. \tag{14}$$

$$\text{compute } \theta_{k+1} \text{ from } \frac{1-\theta_{k+1}}{\theta_{k+1}^2} = \frac{1}{\theta_k^2}. \tag{15}$$

  end for
  Let $\mathbf{HQ} = \mathbf{U}_{S'}^{t,K+1}$ be its polar decomposition and $\mathbf{Z}^{t+1,0} = \mathbf{U}^{t+1,0} = \mathbf{U}^{t,K+1}\mathbf{Q}^T$, where $S' = \begin{cases} S^2, & \text{if } S = S^1, \\ S^1, & \text{if } S = S^2. \end{cases}$
end for

decomposition of $\frac{\mathbf{Z}_S + \mathbf{Z}_S^T}{2}$ and $\hat{\Sigma} = \mathrm{diag}([\max\{\epsilon, \Sigma_{1,1}\}, \cdots, \max\{\epsilon, \Sigma_{r,r}\}])$, then $\mathbf{Z}_S^{t,k+1} = \mathbf{A}\hat{\Sigma}\mathbf{A}^T$ and $\mathbf{Z}_{-S}^{t,k+1} = \mathbf{Z}_{-S}$. At step (15), θ_{k+1} is computed by $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$. At the end of each outer iteration, we need to compute the polar decomposition. Let $\mathbf{A}\Sigma\mathbf{B}^T$ be the SVD of $\mathbf{U}_{S'}^{t,K+1}$, then we can set $\mathbf{H} = \mathbf{A}\Sigma\mathbf{A}^T$ and $\mathbf{Q} = \mathbf{A}\mathbf{B}^T$. In Algorithm 1, we predefine S^1 and S^2 and fix them during the iterations. In Section 3.1 we will discuss how to find S^1 and S^2 using some local information.

At last, let's compare the per-iteration cost of Algorithm 1 with the methods operating on $\mathbf{X}$ space. Both the eigenvalue decomposition and polar decomposition required in Algorithm 1 perform on the submatrices of size $r \times r$, which need $O(r^3)$ operations. Thus, the per-iteration complexity of Algorithm 1 is $O(nr + r^3)$. As a comparison, the methods operating on $\mathbf{X}$ space require at least the top-r singular value/vectors, which need $O(n^2 r)$ operations for the deterministic algorithms and $O(n^2 \log r)$ for randomized algorithms [42]. Thus, our method is more efficient at each iteration when $r \ll n$, especially when r is upper bounded by a constant independent on n.

3.1 Finding the Index Sets S^1 and S^2

In this section, we consider how to find the index sets S^1 and S^2. $S^1 \cap S^2 = \emptyset$ can be easily satisfied and we only need to ensure that $\mathbf{U}_{S^1}^*$ and $\mathbf{U}_{S^2}^*$ are of full rank. Suppose that we have some initializer $\mathbf{U}^0$ close to $\mathbf{U}^*$. We want to use $\mathbf{U}^0$ to find such S^1 and S^2. We first discuss how to select one index set S based on $\mathbf{U}^0$. We can use the volume sampling subset selection algorithm [43, 44], which can select S such that $\sigma_r(\mathbf{U}_S^0) \geq \frac{\sigma_r(\mathbf{U}^0)}{\sqrt{2r(n-r+1)}}$ with probability

of $1-\delta'$ in $O(nr^3\log(1/\delta'))$ operations. Then, we can bound $\sigma_r(\mathbf{U}^*_S)$ in the following lemma since $\mathbf{U}^0$ is close to $\mathbf{U}^*$.

Lemma 3 *If* $\|\mathbf{U}^0-\mathbf{U}^*\|_F \le 0.01\sigma_r(\mathbf{U}^*)$ *and* $\|\mathbf{U}^0_S-\mathbf{U}^*_S\|_F \le \frac{0.99\sigma_r(\mathbf{U}^*)}{2\sqrt{2r(n-r+1)}}$, *then for the index set* S *returned by the volume sampling subset selection algorithm performed on* $\mathbf{U}^0$ *after* $O(nr^3\log(1/\delta'))$ *operations, we have* $\sigma_r(\mathbf{U}^*_S) \ge \frac{0.99\sigma_r(\mathbf{U}^*)}{2\sqrt{2r(n-r+1)}}$ *with probability of* $1-\delta'$.

Proof Form Theorem 3.11 in [44], we have $\sigma_r(\mathbf{U}^0_S) \ge \frac{\sigma_r(\mathbf{U}^0)}{\sqrt{2r(n-r+1)}}$ with probability of $1-\delta'$ after $O(nr^3\log(1/\delta'))$ operations. So we can obtain

$$\sigma_r(\mathbf{U}^0_S)-\sigma_r(\mathbf{U}^*_S)\le\|\mathbf{U}^0_S-\mathbf{U}^*_S\|_F\le\frac{0.99\sigma_r(\mathbf{U}^*)}{2\sqrt{2r(n-r+1)}}\le\frac{\sigma_r(\mathbf{U}^0)}{2\sqrt{2r(n-r+1)}}\le\frac{\sigma_r(\mathbf{U}^0_S)}{2},$$

which leads to

$$\sigma_r(\mathbf{U}^*_S)\ge\frac{\sigma_r(\mathbf{U}^0_S)}{2}\ge\frac{\sigma_r(\mathbf{U}^0)}{2\sqrt{2r(n-r+1)}}\ge\frac{0.99\sigma_r(\mathbf{U}^*)}{2\sqrt{2r(n-r+1)}},$$

where we use $0.99\sigma_r(\mathbf{U}^*)\le\sigma_r(\mathbf{U}^0)$, which is proved in Lemma 9 in Appendix A. □

In the column selection problem and its variants, existing algorithms (please see [44] and the references therein) can only find one index set. Our purpose is to find both S^1 and S^2. We believe that this is a challenging target in the theoretical computer science community. In our applications, since $n \gg r$, we may expect that the rank of $\mathbf{U}^0_{-S^1}$ is not influenced after dropping r rows from $\mathbf{U}^0$. Thus, we can use the procedure discussed above again to find S^2 from $\mathbf{U}^0_{-S^1}$. From Lemma 3, we have $\sigma_r(\mathbf{U}^0_{S^1}) \ge \frac{\sigma_r(\mathbf{U}^0)}{\sqrt{2r(n-r+1)}}$ and $\sigma_r(\mathbf{U}^0_{S^2}) \ge \frac{\sigma_r(\mathbf{U}^0_{-S^1})}{\sqrt{2r(n-2r+1)}}$. In the asymmetric case, this challenge disappears. Please see the details in Section 7. We show in experiments that Algorithm 1 works well even for the simple choice of $S^1=\{1,\cdots,r\}$ and $S^2=\{r+1,\cdots,2r\}$. The discussion of finding S^1 and S^2 in this section is only for the theoretical purpose.

3.2 Initialization

Our theorem ensures the accelerated linear convergence given that the initial point $\mathbf{U}^0 \in \Omega_{S^2}$ is within the local neighborhood of the optimum solution, with radius C defined in Theorem 1. We use the initialization strategy in [24]. Specifically, let $\mathbf{X}^0 = \text{Project}_+\left(\frac{-\nabla f(0)}{\|\nabla f(0)-\nabla f(11^T)\|_F}\right)$ and $\mathbf{V}^0\mathbf{V}^{0T}$ be the best rank-r approximation of $\mathbf{X}_0$, where Project_+ means the projection operator onto the semidefinite cone. Then, [24] proved $\|\mathbf{V}^0-P_{\mathcal{X}^*}(\mathbf{V}^0)\|_F \le \frac{4\sqrt{2}r\|\mathbf{U}^*\|_2^2}{\sigma_r(\mathbf{U}^*)}\sqrt{\frac{L^2}{\mu^2}-\frac{2\mu}{L}+1}$. Let $\mathbf{HQ}=\mathbf{V}^0_{S^2}$ be its polar decomposition and

$\mathbf{U}^0 = \mathbf{V}^0\mathbf{Q}^T$. Then, $\mathbf{U}^0$ belongs to Ω_{S^2}. Although this strategy does not produce an initial point close enough to the target, we show in experiments that our method performs well in practice. It should be noted that for the gradient descent method to solve the general problem (2), the initialization strategy in [24] also does not satisfy the requirement of the theorems in [24] for the general objective f.

4 Accelerated Convergence Rate Analysis

In this section, we prove the local accelerated linear convergence rate of Algorithm 1. We first consider the inner loop. It uses the classical accelerated gradient method to solve problem (8) with fixed index set S for finite K iterations. Thanks to the stronger curvature built in Theorem 1 and the smoothness in Corollary 1, we can use the standard proof framework to analyze the inner loop, e.g., [45]. Some slight modifications are needed since we should ensure that all the iterates belong to the local neighborhood of $\mathbf{U}^*$. We present the result in the following lemma and give its proof sketch. For simplicity, we omit the outer iteration number t.

Lemma 4 *Let $\mathbf{U}^* = \Omega_S \cap \mathcal{X}^*$ and assume that $\mathbf{U}^0 \in \Omega_S$ with $\epsilon \leq 0.99\sigma_r(\mathbf{U}^*_{S'})$ and $\|\mathbf{U}^0 - \mathbf{U}^*\|_F \leq C$. Let $\eta = \frac{1}{L_g}$, where C is defined in Theorem 1 and L_g is defined in Corollary 1. Then, we have $\sigma_r(\mathbf{U}^{K+1}_{S'}) \geq \epsilon$, $\|\mathbf{U}^{K+1} - \mathbf{U}^*\|_F \leq C$ and*

$$g(\mathbf{U}^{K+1}) - g(\mathbf{U}^*) \leq \frac{2}{(K+1)^2\eta}\left\|\mathbf{U}^* - \mathbf{U}^0\right\|_F^2.$$

Proof We follow four step to prove the lemma.

Step 1: We can easily check that if $\mathbf{U}^0 \in \Omega_S$, then all the iterates of $\{\mathbf{U}^k, \mathbf{V}^k, \mathbf{Z}^k\}$ belong to Ω_S by $0 \leq \theta_k \leq 1$, the convexity of Ω_S and the convex combinations in (12) and (14).

Step 2: Consider the k-th iteration. If $\|\mathbf{V}^k - \mathbf{U}^*\|_F \leq C$, $\|\mathbf{Z}^k - \mathbf{U}^*\|_F \leq C$ and $\|\mathbf{U}^k - \mathbf{U}^*\|_F \leq C$, then Theorem 1 and Corollary 1 hold. From the standard analysis of the accelerated gradient method for convex programming, e.g., Proposition 1 in [45], we can have

$$\begin{aligned}&\frac{1}{\theta_k^2}\left(g(\mathbf{U}^{k+1}) - g(\mathbf{U}^*)\right) + \frac{1}{2\eta}\|\mathbf{Z}^{k+1} - \mathbf{U}^*\|_F^2\\ \leq&\frac{1}{\theta_{k-1}^2}\left(g(\mathbf{U}^k) - g(\mathbf{U}^*)\right) + \frac{1}{2\eta}\|\mathbf{Z}^k - \mathbf{U}^*\|_F^2.\end{aligned} \tag{16}$$

Step 3: Since Theorem 1 and Corollary 1 hold only in a local neighbourhood of $\mathbf{U}^*$, we need to check that $\{\mathbf{U}^k, \mathbf{V}^k, \mathbf{Z}^k\}$ belongs to this neighborhood for all the iterations, which can be easily done via induction. In fact, from (16)

and the convexity combinations in (12) and (14), we know that if the following conditions hold,

$$\|\mathbf{V}^k - \mathbf{U}^*\|_F \le C, \quad \|\mathbf{U}^k - \mathbf{U}^*\|_F \le C, \quad \|\mathbf{Z}^k - \mathbf{U}^*\|_F \le C,$$
$$\frac{1}{\theta_{k-1}^2}\left(g(\mathbf{U}^k) - g(\mathbf{U}^*)\right) + \frac{1}{2\eta}\|\mathbf{Z}^k - \mathbf{U}^*\|_F^2 \le \frac{C^2}{2\eta},$$

then we can have

$$\|\mathbf{V}^{k+1} - \mathbf{U}^*\|_F \le C, \quad \|\mathbf{U}^{k+1} - \mathbf{U}^*\|_F \le C, \quad \|\mathbf{Z}^{k+1} - \mathbf{U}^*\|_F \le C,$$
$$\frac{1}{\theta_k^2}\left(g(\mathbf{U}^{k+1}) - g(\mathbf{U}^*)\right) + \frac{1}{2\eta}\|\mathbf{Z}^{k+1} - \mathbf{U}^*\|_F^2 \le \frac{C^2}{2\eta}.$$

Step 4: From $\frac{1}{\theta_{-1}} = 0$ and Step 3, we know (16) holds for all the iterations. Thus, we have

$$g(\mathbf{U}^{K+1}) - g(\mathbf{U}^*) \le \frac{\theta_K^2}{2\eta}\left\|\mathbf{Z}^0 - \mathbf{U}^*\right\|_F^2 \le \frac{2}{(K+1)^2\eta}\|\mathbf{Z}^0 - \mathbf{U}^*\|_F^2,$$

where we use $\theta_k \le \frac{2}{k+1}$ from $\frac{1-\theta_{k+1}}{\theta_{k+1}^2} = \frac{1}{\theta_k^2}$ and $\theta_0 = 1$.

On the other hand, from the perturbation theorem of singular values, we have

$$\sigma_r(\mathbf{U}_{S'}^*) - \sigma_r(\mathbf{U}_{S'}^{K+1}) \le \|\mathbf{U}_{S'}^{K+1} - \mathbf{U}_{S'}^*\|_F \le \|\mathbf{U}^{K+1} - \mathbf{U}^*\|_F \le C \le 0.01\sigma_r(\mathbf{U}_{S'}^*),$$

which leads to $\sigma_r(\mathbf{U}_{S'}^{K+1}) \ge 0.99\sigma_r(\mathbf{U}_{S'}^*) \ge \epsilon$. □

Now we consider the outer loop of Algorithm 1. Based on Lemma 4, the second order growth property (5) and the perturbation theory of polar decomposition, we can establish the exponentially decreasing of $\|\mathbf{U}^{t,0} - \mathbf{U}^{t,*}\|_F$ in the following lemma.

Lemma 5 *Let* $\mathbf{U}^{t,*} = \Omega_S \cap \mathcal{X}^*$ *and* $\mathbf{U}^{t+1,*} = \Omega_{S'} \cap \mathcal{X}^*$ *and assume that* $\mathbf{U}^{t,0} \in \Omega_S$ *with* $\epsilon \le 0.99\sigma_r(\mathbf{U}_{S'}^{t,*})$ *and* $\|\mathbf{U}^{t,0} - \mathbf{U}^{t,*}\|_F \le C$. *Let* $K+1 = \frac{28\|\mathbf{U}^*\|_2}{\sqrt{\eta\mu}\sigma_r(\mathbf{U}^*)\min\{\sigma_r(\mathbf{U}_{S1}^*),\sigma_r(\mathbf{U}_{S^2}^*)\}}$. *Then, we can have* $\mathbf{U}^{t+1,0} \in \Omega_{S'}$ *and*

$$\|\mathbf{U}^{t+1,0} - \mathbf{U}^{t+1,*}\|_F \le \frac{1}{4}\|\mathbf{U}^{t,0} - \mathbf{U}^{t,*}\|_F. \tag{17}$$

Proof We follow four steps to prove the lemma.

Step 1. From Lemma 4, we have $\sigma_r(\mathbf{U}_{S'}^{t,K+1}) \ge \epsilon$, $\|\mathbf{U}^{t,K+1} - \mathbf{U}^{t,*}\|_F \le C$ and

$$g(\mathbf{U}^{t,K+1}) - g(\mathbf{U}^{t,*}) \le \frac{2}{(K+1)^2\eta}\|\mathbf{U}^{t,0} - \mathbf{U}^{t,*}\|_F^2. \tag{18}$$

From Algorithm 1, we have $\sigma_r(\mathbf{U}_{S'}^{t+1,0}) = \sigma_r(\mathbf{U}_{S'}^{t,K+1})$. So $\mathbf{U}_{S'}^{t+1,0} \succeq \epsilon\mathbf{I}$ and $\mathbf{U}^{t+1,0} \in \Omega_{S'}$.

Step 2. From Lemma 11 in Appendix B, we have

$$g(\mathbf{U}^{t,K+1}) - g(\mathbf{U}^{t,*}) \geq 0.4\mu\sigma_r^2(\mathbf{U}^{t,*})\|\mathbf{U}^{t,K+1} - \hat{\mathbf{U}}^{t,*}\|_F^2, \tag{19}$$

where $\hat{\mathbf{U}}^{t,*} = P_{\mathcal{X}^*}(\mathbf{U}^{t,K+1}) = \mathbf{U}^{t,*}\mathbf{R}$ and $\mathbf{R} = \operatorname{argmin}_{\mathbf{R}\mathbf{R}^T=\mathbf{I}}\|\mathbf{U}^{t,*}\mathbf{R}-\mathbf{U}^{t,K+1}\|_F^2$.

Step 3. Given (18) and (19), in order to prove (17), we only need to lower bound $\|\mathbf{U}^{t,K+1} - \hat{\mathbf{U}}^{t,*}\|_F$ by $\|\mathbf{U}^{t+1,0} - \mathbf{U}^{t+1,*}\|_F$.

From Algorithm 1, we know that $\mathbf{H}\mathbf{Q} = \mathbf{U}_{S'}^{t,K+1}$ is the unique polar decomposition of $\mathbf{U}_{S'}^{t,K+1}$ and $\mathbf{U}^{t+1,0} = \mathbf{U}^{t,K+1}\mathbf{Q}^T$. Let $\mathbf{H}^*\mathbf{Q}^* = \hat{\mathbf{U}}_{S'}^{t,*}$ be its unique polar decomposition and $\mathbf{U}^{t+1,*} = \hat{\mathbf{U}}^{t,*}(\mathbf{Q}^*)^T$, then $\mathbf{U}^{t+1,*} \in \Omega_{S'} \cap \mathcal{X}^*$. From the perturbation theorem of polar decomposition in Lemma 1, we have

$$\|\mathbf{Q} - \mathbf{Q}^*\|_F \leq \frac{2}{\sigma_r(\hat{\mathbf{U}}_{S'}^{t,*})}\|\mathbf{U}_{S'}^{t,K+1} - \hat{\mathbf{U}}_{S'}^{t,*}\|_F.$$

Similar to (9), we have

$$\begin{aligned}
&\|\mathbf{U}^{t+1,0} - \mathbf{U}^{t+1,*}\|_F \\
=&\|\mathbf{U}^{t,K+1}\mathbf{Q}^T - \hat{\mathbf{U}}^{t,*}(\mathbf{Q}^*)^T\|_F \\
=&\|\mathbf{U}^{t,K+1}\mathbf{Q}^T - \hat{\mathbf{U}}^{t,*}\mathbf{Q}^T + \hat{\mathbf{U}}^{t,*}\mathbf{Q}^T - \hat{\mathbf{U}}^{t,*}(\mathbf{Q}^*)^T\|_F \\
\leq&\|\mathbf{U}^{t,K+1} - \hat{\mathbf{U}}^{t,*}\|_F + \|\hat{\mathbf{U}}^{t,*}\|_2\|\mathbf{Q} - \mathbf{Q}^*\|_F \\
\leq&\frac{3\|\mathbf{U}^{t,*}\|_2}{\sigma_r(\mathbf{U}_{S'}^{t,*})}\|\mathbf{U}^{t,K+1} - \hat{\mathbf{U}}^{t,*}\|_F.
\end{aligned} \tag{20}$$

Step 4. Combining (18), (19) and (20), we have

$$\begin{aligned}
&\|\mathbf{U}^{t+1,0} - \mathbf{U}^{t+1,*}\|_F \\
\leq& \frac{3\|\mathbf{U}^{t,*}\|_2}{\sigma_r(\mathbf{U}_{S'}^{t,*})}\|\mathbf{U}^{t,K+1} - \hat{\mathbf{U}}^{t,*}\|_F \\
\leq& \frac{3\|\mathbf{U}^{t,*}\|_2}{\sigma_r(\mathbf{U}_{S'}^{t,*})}\frac{\sqrt{5}}{\sqrt{\eta\mu}(K+1)\sigma_r(\mathbf{U}^{t,*})}\|\mathbf{U}^{t,0} - \mathbf{U}^{t,*}\|_F \\
\leq& \frac{7\|\mathbf{U}^{t,*}\|_2}{\sqrt{\eta\mu}(K+1)\sigma_r(\mathbf{U}^{t,*})\min\{\sigma_r(\mathbf{U}_S^{t,*}), \sigma_r(\mathbf{U}_{S'}^{t,*})\}}\|\mathbf{U}^{t,0} - \mathbf{U}^{t,*}\|_F.
\end{aligned}$$

From the setting of $K+1$, we can have the conclusion. □

From Lemma 5, we can give the accelerated convergence rate in the following theorem. The proof is provided in Appendix B. We contain several assumptions in Theorem 3. For the trajectories, we assume that we can find two disjoint sets S^1 and S^2 such that $\sigma_r(\mathbf{U}_{S^1}^*)$ and $\sigma_r(\mathbf{U}_{S^2}^*)$ are as large as possible (please see Section 3.1 for the discussion). For the initialization, we assume that we can find an initial point $\mathbf{U}^{0,0}$ close enough to $\mathbf{U}^{0,*}$ (please see Section 3.2 for the discussion). Then, we can prove that when the outer iteration number t is odd, $\mathbf{U}^{t,k}$ belongs to Ω_{S^1} and the iterates converge to the optimum solution of $\Omega_{S^1} \cap \mathcal{X}^*$. When t is even, the iterates belong to Ω_{S^2} and converge to another

optimum solution of $\Omega_{S^2} \cap \mathcal{X}^*$. In our algorithm, we set η and K based on a reliable knowledge on $\|\mathbf{U}^*\|_2$, $\sigma_r(\mathbf{U}^*)$ and $\sigma_r(\mathbf{U}^*_S)$. As suggested by [24, 46], they can be estimated by $\|\mathbf{U}^0\|_2$, $\sigma_r(\mathbf{U}^0)$ and $\sigma_r(\mathbf{U}^0_S)$−up to constants−since $\mathbf{U}^0$ is close to $\mathbf{U}^*$.

Theorem 3 *Let $\mathbf{U}^{t,*} = \Omega_{S^1} \cap \mathcal{X}^*$ when t is odd and $\mathbf{U}^{t,*} = \Omega_{S^2} \cap \mathcal{X}^*$ when t is even. Assume that $\mathbf{U}^* \in \mathcal{X}^*$ and $\mathbf{U}^{0,0} \in \Omega_{S^2}$ with $\|\mathbf{U}^{0,0} - \mathbf{U}^{0,*}\|_F \leq C$ and $\epsilon \leq \min\{0.99\sigma_r(\mathbf{U}^*_{S^1}), 0.99\sigma_r(\mathbf{U}^*_{S^2})\}$. Then, we have*

$$\|\mathbf{U}^{t+1,0} - \mathbf{U}^{t+1,*}\|_F \leq \left(1 - \frac{1}{6}\sqrt{\frac{\mu_g}{L_g}}\right)^{(t+1)(K+1)} \|\mathbf{U}^{0,0} - \mathbf{U}^*\|_F,$$

and

$$g(\mathbf{U}^{t+1,0}) - g(\mathbf{U}^*) \leq L_g \left(1 - \frac{1}{6}\sqrt{\frac{\mu_g}{L_g}}\right)^{2(t+1)(K+1)} \|\mathbf{U}^{0,0} - \mathbf{U}^*\|_F^2,$$

where $\mu_g = \frac{\mu\sigma_r^2(\mathbf{U}^*)\min\{\sigma_r^2(\mathbf{U}^*_{S^1}),\sigma_r^2(\mathbf{U}^*_{S^2})\}}{25\|\mathbf{U}^*\|_2^2}$, $L_g = 38L\|\mathbf{U}^*\|_2^2 + 2\|\nabla f(\mathbf{X}^*)\|_2$ and $C = \frac{\mu\sigma_r^2(\mathbf{U}^*)\min\{\sigma_r^2(\mathbf{U}^*_{S^1}),\sigma_r^2(\mathbf{U}^*_{S^2})\}}{100L\|\mathbf{U}^*\|_2^3}$.

4.1 Comparison to the Gradient Descent

Bhojanapalli et al. [24] used the gradient descent to solve problem (2), which consists of the following recursion:

$$\mathbf{U}^{k+1} = \mathbf{U}^k - \eta\nabla g(\mathbf{U}^k).$$

With the restricted strong convexity and smoothness of $f(\mathbf{X})$, Bhojanapalli et al. [24] proved the linear convergence of gradient descent in the form of

$$\begin{aligned} &\|\mathbf{U}^{N+1} - P_{\mathcal{X}^*}(\mathbf{U}^{N+1})\|_F^2 \\ &\leq \left(1 - \frac{\sigma_r^2(\mathbf{U}^*)}{\|\mathbf{U}^*\|_2^2}\frac{\mu}{L + \|\nabla f(\mathbf{X}^*)\|_2/\|\mathbf{U}^*\|_2^2}\right)^N \|\mathbf{U}^0 - P_{\mathcal{X}^*}(\mathbf{U}^0)\|_F^2. \end{aligned} \tag{21}$$

As a comparison, from Theorem 3, our method converges linearly within the error of $\left(1 - \frac{\sigma_r(\mathbf{U}^*)\min\{\sigma_r(\mathbf{U}^*_{S^1}),\sigma_r(\mathbf{U}^*_{S^2})\}}{\|\mathbf{U}^*\|_2^2}\sqrt{\frac{\mu}{L+\|\nabla f(\mathbf{X}^*)\|_2/\|\mathbf{U}^*\|_2^2}}\right)^N$, where N is the total number of inner iterations. From Lemma 3, we know $\sigma_r(\mathbf{U}^*_S) \approx \frac{1}{\sqrt{rn}}\sigma_r(\mathbf{U}^*)$ in the worst case and it is tight [44]. Thus, our method has the convergence rate of $\left(1 - \frac{\sigma_r^2(\mathbf{U}^*)}{\|\mathbf{U}^*\|_2^2}\sqrt{\frac{\mu}{nr(L+\|\nabla f(\mathbf{X}^*)\|_2/\|\mathbf{U}^*\|_2^2)}}\right)^N$ in the worst case. When the function f is ill-conditioned, i.e., $\frac{L}{\mu} \geq nr$, our method outperforms the gradient descent. This phenomenon is similar to the case observed in the stochastic optimization community: the non-accelerated methods such as SDCA [47], SVRG [48] and SAG [49] have the complexity of $O\left(\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$ while

Table 1 Convergence rate comparisons of the gradient descent method (GD) and accelerated gradient descent method (AGD).

Method	Convex problem	Nonconvex problem (2)
GD	$\left(\frac{L-\mu}{L+\mu}\right)^N$ [29]	$\left(1-\frac{\sigma_r^2(\mathbf{U}^*)}{\|\mathbf{U}^*\|_2^2}\frac{\mu}{L+\|\nabla f(\mathbf{X}^*)\|_2/\|\mathbf{U}^*\|_2^2}\right)^N$ [24]
AGD	$\left(1-\sqrt{\frac{\mu}{L}}\right)^N$ [29]	$\left(1-\frac{\sigma_r(\mathbf{U}^*)\min\{\sigma_r(\mathbf{U}^*_{S1}),\sigma_r(\mathbf{U}^*_{S2})\}}{\|\mathbf{U}^*\|_2^2}\sqrt{\frac{\mu}{L+\|\nabla f(\mathbf{X}^*)\|_2/\|\mathbf{U}^*\|_2^2}}\right)^N$ $=\left(1-\frac{\sigma_r^2(\mathbf{U}^*)}{\|\mathbf{U}^*\|_2^2}\sqrt{\frac{\mu}{nr(L+\|\nabla f(\mathbf{X}^*)\|_2/\|\mathbf{U}^*\|_2^2)}}\right)^N$

the accelerated methods such as Accelerated SDCA [50], Catalyst [51] and Katyusha [52] have the complexity of $O\left(\sqrt{\frac{mL}{\mu}}\log\frac{1}{\epsilon}\right)$, where m is the sample size. The latter is tight when $\frac{L}{\mu}\geq m$ for stochastic programming [53]. In matrix completion, the optimal sample complexity is $O(rn\log n)$ [54]. It is unclear whether our convergence rate for problem (2) is tight or there exists a faster method. We leave it as an open problem.

For better reference, we summarize the comparisons in Table 1. We can see that our method has the same optimal dependence on $\sqrt{\frac{L}{\mu}}$ as convex programming.

4.1.1 Dropping the Dependence on n

Our convergence rate has an additional dependence on n compared with the gradient descent method. It comes from $\sigma_r(\mathbf{U}_S^*)$, i.e., Lemma 2. In fact, we use a loose relaxation in the last inequality of (9), i.e., $\frac{2\|\mathbf{U}\|_2}{\sigma_r(\mathbf{U}_S)}\|\hat{\mathbf{V}}_S-\mathbf{U}_S\|_F\leq\frac{2\|\mathbf{U}\|_2}{\sigma_r(\mathbf{U}_S)}\|\hat{\mathbf{V}}-\mathbf{U}\|_F$. Since $\mathbf{U}_S\in\mathbb{R}^{r\times r}$ and $\mathbf{U}\in\mathbb{R}^{n\times r}$, a more suitable estimation should be

$$\frac{2\|\mathbf{U}\|_2}{\sigma_r(\mathbf{U}_S)}\|\hat{\mathbf{V}}_S-\mathbf{U}_S\|_F\approx\frac{2\|\mathbf{U}\|_2}{\sigma_r(\mathbf{U}_S)}\sqrt{\frac{r}{n}}\|\hat{\mathbf{V}}-\mathbf{U}\|_F\approx\frac{2r\|\mathbf{U}\|_2}{\sigma_r(\mathbf{U})}\|\hat{\mathbf{V}}-\mathbf{U}\|_F. \quad (22)$$

In practice, (22) holds when the entries of $\mathbf{U}^{t,k}$ and $\mathbf{V}^{t,k}$ converge nearly equally fast to those of $\mathbf{U}^{t,*}$, which may be expected in practice. Thus, under the condition of (22), our convergence rate can be improved to

$$\left(1-\frac{\sigma_r^2(\mathbf{U}^*)}{r\|\mathbf{U}^*\|_2^2}\sqrt{\frac{\mu}{L+\|\nabla f(\mathbf{X}^*)\|_2/\|\mathbf{U}^*\|_2^2}}\right)^N.$$

We numerically verify (22) in Section 8.4.

4.1.2 Examples with Ill-conditioned Objective f

Although the condition number $\frac{L}{\mu}$ approximate to 1 for some famous problems in machine learning, e.g., matrix regression and matrix completion [25], we

can still find many problems with ill-conditioned objective, especially in the computer vision applications. We give the example of low rank representation (LRR) [55]. The LRR model is a famous model in computer vision. It can be formulated as

$$\min_{\mathbf{X}} \operatorname{rank}(\mathbf{X}) \quad s.t. \quad \mathbf{DX} = \mathbf{A},$$

where $\mathbf{A}$ is the observed data and $\mathbf{D}$ is a dictionary that linearly spans the data space. We can reformulate the problem as follows:

$$\min_{\mathbf{X}} \|\mathbf{DX} - \mathbf{A}\|_F^2 \quad s.t. \quad \operatorname{rank}(\mathbf{X}) \leq r.$$

We know $L/\mu = \kappa(\mathbf{D}^T\mathbf{D})$, i.e., the condition number of $\mathbf{D}^T\mathbf{D}$. If we generate $\mathbf{D} \in \mathbb{R}^{n\times n}$ as a random matrix with normal distribution, then $E\left[\log\kappa(\mathbf{D})\right] \sim \log n$ as $n \to \infty$ [56] and thus $E\left[\frac{L}{\mu}\right] \sim n^2$. We numerically verify on MATLAB that if $n = 1000$, then $\frac{L}{\mu}$ is of the order 10^7, which is much larger than $O(n)$.

Another example is the reduced rank logistic generalized linear model (RR-LGLM) [57, 58]. Assume that $\mathbf{A}$ are all binary and denote $\mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_n]^T$ and $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$. RR-LGLAM minimizes

$$\min_{\mathbf{X}} -\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbf{A}_{i,j}\mathbf{d}_i^T\mathbf{x}_j - \log\left(1+\exp(\mathbf{d}_i^T\mathbf{x}_j)\right)\right) \quad s.t. \quad \operatorname{rank}(\mathbf{X}) \leq r.$$

The Hessian of the objective is $\operatorname{diag}(\mathbf{D}^T\mathbf{G}_1\mathbf{D}, \cdots, \mathbf{D}^T\mathbf{G}_n\mathbf{D})$, where $\mathbf{G}_j$ is the $n \times n$ diagonal matrix whose i-th component is $\frac{\exp(\mathbf{d}_i^T\mathbf{x}_j)}{(1+\exp(\mathbf{d}_i^T\mathbf{x}_j))^2}$. Thus, L/μ is at least $\kappa(\mathbf{D}^T\mathbf{D})$. As discussed above, it may be much larger than n. Other similar examples can be found in [59, 60].

5 Global Convergence

In this section, we study the global convergence of Algorithm 1 without the assumption that $f(\mathbf{X})$ is restricted strongly convex. We allow the algorithm to start from any initializer. Since we have no information about $\mathbf{U}^*$ when $\mathbf{U}^0$ is far from $\mathbf{U}^*$, we use an adaptive index sets selection procedure for Algorithm 1. That is to say, after each inner loop, we check whether $\sigma_r(\mathbf{U}_{S'}^{t,K+1}) \geq \epsilon$ holds. If not, we select the new index set S' using the volume sampling subset selection algorithm.

We first consider the inner loop and establish Lemma 6. We drop the outer iteration number t for simplicity and leave the proof in Appendix C.

Lemma 6 *Assume that* $\{\mathbf{U}^k, \mathbf{V}^k\}$ *is bounded and* $\mathbf{U}^0 \in \Omega_S$. *Let* $\eta \leq \frac{1-\beta_{\max}^2}{\hat{L}(2\beta_{\max}+1)+2\gamma}$, *where* $\hat{L} = 2D + 4LM^2$, $D = \max\{\|\nabla f(\mathbf{U}^k(\mathbf{U}^k)^T)\|_2, \|\nabla f(\mathbf{V}^k(\mathbf{V}^k)^T)\|_2, \forall k\}$,

$M = \max\{\|\mathbf{U}^k\|_2, \|\mathbf{V}^k\|_2, \forall k\}$, $\beta_{\max} = \max\{\beta_k, k = 0, \cdots, K\}$, $\beta_k = \frac{\theta_k(1-\theta_{k-1})}{\theta_{k-1}}$ *and* γ *is a small constant. Then, we have*

$$g(\mathbf{U}^{K+1}) - g(\mathbf{U}^0) \leq -\sum_{k=0}^{K} \gamma \|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2.$$

Now we consider the outer loop. As discussed in Section 3, when solving problem (8) directly, we may get stuck at the boundary of the constraint. Thanks to the alternating constraint strategy, we can cancel the negative influence of the constraint and establish the global convergence to a critical point of problem (2), which is described in Theorem 4. It establishes that after at most $O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$ operations, $\mathbf{U}^{T,K+1}$ is an approximate zero gradient point in the precision of ε. Briefly speaking, since the projection operation in (13) only influences the rows indicated by the index set S, a simple calculation yields that $\|(\nabla g(\mathbf{Z}^{t,K+1}))_{-S^1}\|_F \leq O(\varepsilon)$ and $\|(\nabla g(\mathbf{Z}^{t,K+1}))_{-S^2}\|_F \leq O(\varepsilon)$. From $S^1 \cap S^2 = \emptyset$, we have $\|\nabla g(\mathbf{Z}^{t,K+1})\|_F \leq \|(\nabla g(\mathbf{Z}^{t,K+1}))_{-S^1}\|_F + \|(\nabla g(\mathbf{Z}^{t,K+1}))_{-S^2}\|_F \leq O(\varepsilon)$, which explains why the alternating constraint strategy avoids the boundary of the constraint.

Theorem 4 *Assume that* $\{\mathbf{U}^{t,k}, \mathbf{V}^{t,k}\}$ *is bounded and* $\sigma_r(\mathbf{U}_{S'}^{t,K+1}) \geq \epsilon, \forall t$. *Let* η *be the one defined in Lemma 6. Then, after at most* $T = 2\frac{f(\mathbf{U}^{t,0}(\mathbf{U}^{t,0})^T) - f(\mathbf{X}^*)}{\varepsilon^2}$ *outer iterations, we have*

$$\left\|\nabla g(\mathbf{U}^{T,K+1})\right\|_F \leq \frac{35\varepsilon}{\eta\theta_K}$$

with probability of $1-\delta$. *The volume sampling subset selection algorithm needs* $O\left(nr^3 \log\left(\frac{f(\mathbf{U}^{t,0}(\mathbf{U}^{t,0})^T) - f(\mathbf{X}^*))}{\delta\varepsilon^2}\right)\right)$ *operations for each running.*

Proof We follow three steps to prove the theorem.

Step 1. Firstly, we bound the difference of two consecutive variables, i.e., $\mathbf{U}^{t,k+1} - \mathbf{U}^{t,k}$.

From Lemma 6 we have

$$\gamma \sum_{k=0}^{K} \|\mathbf{U}^{t,k+1} - \mathbf{U}^{t,k}\|_F^2 \leq g(\mathbf{U}^{t,0}) - g(\mathbf{U}^{t,K+1}).$$

Summing over $t = 0, \cdots, T$ yields

$$\begin{aligned}\gamma \sum_{t=0}^{T}\sum_{k=0}^{K} \|\mathbf{U}^{t,k+1} - \mathbf{U}^{t,k}\|_F^2 &\leq \sum_{t=0}^{T} \left(g(\mathbf{U}^{t,0}) - g(\mathbf{U}^{t,K+1})\right) \\ = \sum_{t=0}^{T} \left(g(\mathbf{U}^{t,0}) - g(\mathbf{U}^{t+1,0})\right) &\leq g(\mathbf{U}^{t,0}) - f(\mathbf{U}^*\mathbf{U}^{*T}).\end{aligned}$$

So after $T = 2\frac{g(\mathbf{U}^{t,0}) - f(\mathbf{X}^*)}{\varepsilon^2}$ outer iterations, we must have

$$\sum_{k=0}^{K} \|\mathbf{U}^{t,k+1} - \mathbf{U}^{t,k}\|_F^2 + \sum_{k=0}^{K} \|\mathbf{U}^{t+1,k+1} - \mathbf{U}^{t+1,k}\|_F^2 \leq \varepsilon^2 \tag{23}$$

for some $t < T$. Thus, we can bound $\|\mathbf{U}^{t',k+1} - \mathbf{U}^{t',k}\|_F$ by ε, where $t' = t$ or $t' = t+1$. Moreover, from Lemma 13 in Appendix C, we can bound $\|\mathbf{U}^{t',k+1} - \mathbf{Z}^{t',k+1}\|_F$, $\|\mathbf{Z}^{t',k+1} - \mathbf{Z}^{t',k}\|_F$ and $\|\mathbf{Z}^{t',k+1} - \mathbf{V}^{t',k}\|_F$ by $\frac{\varepsilon}{\theta_k}$.

Step 2. Secondly, we bound parts of elements of the gradient, i.e., $\left(\nabla g(\mathbf{Z}^{t,K+1})\right)_{-S^1}$ and $\left(\nabla g(\mathbf{Z}^{t,K+1})\right)_{-S^2}$.

From the optimality condition of (13), we have

$$-\frac{\theta_k}{\eta}\left(\mathbf{Z}^{t',k+1} - \mathbf{Z}^{t',k}\right) + \nabla g(\mathbf{Z}^{t',k+1}) - \nabla g(\mathbf{V}^{t',k}) \in \nabla g(\mathbf{Z}^{t',k+1}) + \partial I_{\Omega_{S^j}}(\mathbf{Z}^{t',k+1})$$

for $j = 1$ when $t' = t$ and $j = 2$ when $t' = t+1$. From Lemmas 10 and 13, we can easily check that

$$\left\| -\frac{\theta_k}{\eta}\left(\mathbf{Z}^{t',k+1} - \mathbf{Z}^{t',k}\right) + \nabla g(\mathbf{Z}^{t',k+1}) - \nabla g(\mathbf{V}^{t',k}) \right\|_F \leq \frac{14\varepsilon}{\eta\theta_k}.$$

Thus, we obtain

$$\operatorname{dist}\left(0, \nabla g(\mathbf{Z}^{t',k+1}) + \partial I_{\Omega_{S^j}}(\mathbf{Z}^{t',k+1})\right) \leq \frac{14\varepsilon}{\eta\theta_k}, \forall k = 0, \cdots, K.$$

Since $\partial I_{\Omega_{S^j}}(\mathbf{Z}^{t',k+1})$ has zero elements for the rows indicated by the indexes out of S^j, we can have

$$\left\| \left(\nabla g(\mathbf{Z}^{t,K+1})\right)_{-S^1} \right\|_F \leq \frac{14\varepsilon}{\eta\theta_K}, \tag{24}$$

and

$$\left\| \left(\nabla g(\mathbf{Z}^{t+1,1})\right)_{-S^2} \right\|_F \leq \frac{14\varepsilon}{\eta\theta_K}, \tag{25}$$

On the other hand,

$$\begin{aligned} &\left\| \left(\nabla g(\mathbf{Z}^{t+1,0})\right)_{-S^2} \right\|_F - \left\| \left(\nabla g(\mathbf{Z}^{t+1,1})\right)_{-S^2} \right\|_F \\ \leq &\left\| \left(\nabla g(\mathbf{Z}^{t+1,0}) - \nabla g(\mathbf{Z}^{t+1,1})\right)_{-S^2} \right\|_F \\ \leq &\left\| \nabla g(\mathbf{Z}^{t+1,0}) - \nabla g(\mathbf{Z}^{t+1,1}) \right\|_F \leq \hat{L}\|\mathbf{Z}^{t+1,0} - \mathbf{Z}^{t+1,1}\|_F \leq \frac{5\hat{L}\varepsilon}{\theta_K}, \end{aligned} \tag{26}$$

where we use Lemma 13 in the last inequality. Combing (25) and (26), we can obtain

$$\left\| \left(\nabla g(\mathbf{Z}^{t+1,0})\right)_{-S^2} \right\|_F \leq \frac{19\varepsilon}{\eta\theta_K}.$$

Since $\mathbf{Z}^{t+1,0} = \mathbf{Z}^{t,K+1}\mathbf{Q}^T$ for some orthogonal $\mathbf{Q}$, we can have

$$\begin{aligned} \frac{19\varepsilon}{\eta\theta_K} &\geq \left\| \left(\nabla g(\mathbf{Z}^{t+1,0})\right)_{-S^2} \right\|_F = \left\| \left(\nabla g(\mathbf{Z}^{t,K+1})\mathbf{Q}^T\right)_{-S^2} \right\|_F \\ &= \left\| \left(\nabla g(\mathbf{Z}^{t,K+1})\right)_{-S^2} \mathbf{Q}^T \right\|_F = \left\| \left(\nabla g(\mathbf{Z}^{t,K+1})\right)_{-S^2} \right\|_F. \end{aligned} \tag{27}$$

Step 3. We bound all the elements of the gradient. Recall that we require $S^1 \cap S^2 = \emptyset$. Thus, we have $-S^1 \cup -S^2 = \{1, 2, \cdots, n\}$. Then, from (24) and (27), we have

$$\left\|\nabla g(\mathbf{Z}^{t,K+1})\right\|_F \leq \left\|\left(\nabla g(\mathbf{Z}^{t,K+1})\right)_{-S^1}\right\|_F + \left\|\left(\nabla g(\mathbf{Z}^{t,K+1})\right)_{-S^2}\right\|_F \leq \frac{33\varepsilon}{\eta\theta_K}.$$

At last, we can bound $\left\|\nabla g(\mathbf{U}^{t,K+1})\right\|_F$ from Lemmas 10 and 13.

From the Algorithm, we know that the index set is selected at most T times. The volume sampling subset selection algorithm succeeds with the probability of $1-\delta'$. So the Algorithm succeeds with the probability at least of $1-T\delta' = 1-\delta$. On the other hand, the volume sampling subset selection algorithm needs $O\left(nr^3 \log\left(\frac{1}{\delta'}\right)\right) = O\left(nr^3 \log\left(\frac{T}{\delta}\right)\right) = O\left(nr^3 \log\left(\frac{f(\mathbf{U}^{t,0}(\mathbf{U}^{t,0})^T) - f(\mathbf{U}^*\mathbf{U}^{*T}))}{\delta\varepsilon^2}\right)\right)$ operations. □

6 Minimizing (2) Directly without the Constraint

Someone may doubt the necessity of the constraint in problem (8) and they wonder the performance of the classical accelerated gradient method to minimize problem (2) directly. In this case, the classical accelerated gradient method [28, 27, 45] becomes

$$\mathbf{V}^k = (1-\theta_k)\mathbf{U}^k + \theta_k\mathbf{Z}^k, \tag{28}$$
$$\mathbf{Z}^{k+1} = \mathbf{Z}^k - \eta\nabla g(\mathbf{V}^k), \tag{29}$$
$$\mathbf{U}^{k+1} = (1-\theta_k)\mathbf{U}^k + \theta_k\mathbf{Z}^{k+1}, \tag{30}$$

and it is equivalent to

$$\mathbf{V}^k = \mathbf{U}^k + \beta_k(\mathbf{U}^k - \mathbf{U}^{k-1}), \tag{31}$$
$$\mathbf{U}^{k+1} = \mathbf{V}^k - \eta\nabla g(\mathbf{V}^k). \tag{32}$$

where β_k is defined in Lemma 6. Another choice is a constant of $\beta < 1$. Theorem 5 establishes the convergence rate for the above two recursions. We leave the proof in Appendix D.

Theorem 5 *Assume that* $\mathbf{U}^* \in \mathcal{X}^*$ *and* $\mathbf{V}^k \in \mathbb{R}^{n\times r}$ *satisfy* $\|\mathbf{V}^k - P_{\mathcal{X}^*}(\mathbf{V}^k)\|_F \leq \min\left\{0.01\sigma_r(\mathbf{U}^*), \frac{\mu\sigma_r^2(\mathbf{U}^*)}{6L\|\mathbf{U}^*\|_2}\right\}$. *Let* η *be the one in Lemma 6. Then, we can have*

$$\begin{aligned} &g(\mathbf{U}^{k+1}) + \nu\|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2 - g(\mathbf{U}^*) \\ \leq\ &\frac{1}{1 + \frac{\gamma}{\frac{5}{\eta^2\mu\sigma_r^2(\mathbf{U}^*)} + \nu}} \left[g(\mathbf{U}^k) + \nu\|\mathbf{U}^k - \mathbf{U}^{k-1}\|_F^2 - g(\mathbf{U}^*)\right]. \end{aligned}$$

where $\gamma = \frac{1-\beta_{\max}^2}{4\eta} - \frac{\beta_{\max}\hat{L}}{2} - \frac{\hat{L}}{4} > 0$ *and* $\nu = \frac{1+\beta_{\max}^2}{4\eta} - \frac{\hat{L}}{4} > 0$.

Consider the case that β_k is a constant. Then, we know that all of the constants $\gamma, \nu, \hat{L}$ and $\frac{1}{\eta}$ are of the order $O\left(L\|\mathbf{U}^*\|_2^2 + \|\nabla f(\mathbf{X}^*)\|_2\right)$. Thus, the convergence rate of recursion (31)-(32) is in the form of

$$\left(1 - \frac{\mu\sigma_r^2(\mathbf{U}^*)}{L\|\mathbf{U}^*\|_2^2 + \|\nabla f(\mathbf{X}^*)\|_2}\right)^N,$$

which is the same as that of the gradient descent method in (21). Thus, although the convergence of the classical accelerated gradient method for problem (2) can be proved, it is not easy to build the acceleration upon the gradient descent. As a comparison, Algorithm 1 has a theoretical better dependence on the condition number of $\frac{L}{\mu}$. Thus, the reformulation of problem (2) to a constrained one is necessary to prove acceleration.

7 The Asymmetric Case

In this section, we consider the asymmetric case of problem (1):

$$\min_{\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times m}} f(\widetilde{\mathbf{X}}), \tag{33}$$

where there exists a minimizer $\widetilde{\mathbf{X}}^*$ of rank-r. We follow [46] to assume $\nabla f(\widetilde{\mathbf{X}}^*) = 0$. In the asymmetric case, we can factorize $\widetilde{\mathbf{X}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T$ and reformulate problem (33) as a similar problem to (2). Moreover, we follow [46, 26] to regularize the objective and force the solution pair $(\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}})$ to be balanced. Otherwise, the problem may be ill-conditioned since $\left(\frac{1}{\delta}\widetilde{\mathbf{U}}\right)(\delta\widetilde{\mathbf{V}})$ is also a factorization of $\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T$ for any large δ [46]. Specifically, we consider the following problem

$$\min_{\widetilde{\mathbf{U}} \in \mathbb{R}^{n\times r}, \widetilde{\mathbf{V}} \in \mathbb{R}^{m\times r}} f(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T) + \frac{\mu}{8}\|\widetilde{\mathbf{U}}^T\widetilde{\mathbf{U}} - \widetilde{\mathbf{V}}^T\widetilde{\mathbf{V}}\|_F^2. \tag{34}$$

Let $\widetilde{\mathbf{X}}^* = \mathbf{A}\Sigma\mathbf{B}^T$ be its SVD. Then, $(\widetilde{\mathbf{U}}^* = \mathbf{A}\sqrt{\Sigma}, \widetilde{\mathbf{V}}^* = \mathbf{B}\sqrt{\Sigma})$ is a minimizer of problem (34). Define a stacked matrix $\mathbf{U} = \begin{pmatrix} \widetilde{\mathbf{U}} \\ \widetilde{\mathbf{V}} \end{pmatrix}$ and let $\mathbf{X} = \mathbf{U}\mathbf{U}^T = \begin{pmatrix} \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T & \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T \\ \widetilde{\mathbf{V}}\widetilde{\mathbf{U}}^T & \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^T \end{pmatrix}$. Then we can write the objective in (34) in the form of $\hat{f}(\mathbf{X})$, defined as $\hat{f}(\mathbf{X}) = f(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T) + \frac{\mu}{8}\|\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T\|_F^2 + \frac{\mu}{8}\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^T\|_F^2 - \frac{\mu}{4}\|\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T\|_F^2$. Since f is restricted μ-strongly convex, we can easily check that $\hat{f}(\mathbf{X})$ is restricted $\frac{\mu}{4}$-strongly convex. On the other hand, we know that $\hat{f}(\mathbf{X})$ is restricted $\left(L + \frac{\mu}{2}\right)$-smooth. Applying the conclusions on the symmetric case to $\hat{f}(\mathbf{X})$, we can apply Algorithm 1 to the asymmetric case. From Theorem 3, we can get the

convergence rate. Moreover, since $\sigma_i(\mathbf{X}^*) = 2\sigma_i(\widetilde{\mathbf{X}}^*)$,

$$\begin{aligned}\nabla \hat{f}(\mathbf{X}^*) =& \begin{pmatrix} 0 & \nabla f(\widetilde{\mathbf{X}}^*) \\ \nabla f(\widetilde{\mathbf{X}}^*)^T & 0 \end{pmatrix} + \frac{\mu}{4} \begin{pmatrix} \widetilde{\mathbf{U}}^* \\ -\widetilde{\mathbf{V}}^* \end{pmatrix} \left(\widetilde{\mathbf{U}}^{*T}, -\widetilde{\mathbf{V}}^{*T} \right) \\ =& \frac{\mu}{4} \begin{pmatrix} \widetilde{\mathbf{U}}^* \\ -\widetilde{\mathbf{V}}^* \end{pmatrix} \left(\widetilde{\mathbf{U}}^{*T}, -\widetilde{\mathbf{V}}^{*T} \right)\end{aligned}$$

and $\|\nabla \hat{f}(\mathbf{X}^*)\|_2 = \frac{\mu}{4}\|\mathbf{X}^*\|_2$, where $\mathbf{X}^* = \begin{pmatrix} \widetilde{\mathbf{U}}^* \widetilde{\mathbf{U}}^{*T} & \widetilde{\mathbf{U}}^* \widetilde{\mathbf{V}}^{*T} \\ \widetilde{\mathbf{V}}^* \widetilde{\mathbf{U}}^{*T} & \widetilde{\mathbf{V}}^* \widetilde{\mathbf{V}}^{*T} \end{pmatrix}$, we can simplify the worst case convergence rate to $\left(1 - \frac{\sigma_r(\widetilde{\mathbf{X}}^*)}{\|\widetilde{\mathbf{X}}^*\|_2} \sqrt{\frac{\mu}{(m+n)rL}}\right)^N$. As a comparison, the rate of the gradient descent is $\left(1 - \frac{\sigma_r(\widetilde{\mathbf{X}}^*)}{\|\widetilde{\mathbf{X}}^*\|_2} \frac{\mu}{L}\right)^N$ [46].

In the asymmetric case, both $\widetilde{\mathbf{U}}^*$ and $\widetilde{\mathbf{V}}^*$ are of full rank. Otherwise, $\mathrm{rank}(\widetilde{\mathbf{X}}^*) < r$. Thus, we can select the index set S^1 from $\widetilde{\mathbf{U}}^0$ and select S^2 from $\widetilde{\mathbf{V}}^0$ with the guarantee of $\sigma_r(\widetilde{\mathbf{U}}^0_{S^1}) \geq \frac{\sigma_r(\widetilde{\mathbf{U}}^0)}{\sqrt{2r(n-r+1)}}$ and $\sigma_r(\widetilde{\mathbf{V}}^0_{S^2}) \geq \frac{\sigma_r(\widetilde{\mathbf{V}}^0)}{\sqrt{2r(m-r+1)}}$.

8 Experiments

In this section, we test the efficiency of the proposed Accelerated Gradient Descent (AGD) method on Matrix Completion, One Bit Matrix Completion and Matrix Regression.

8.1 Matrix Completion

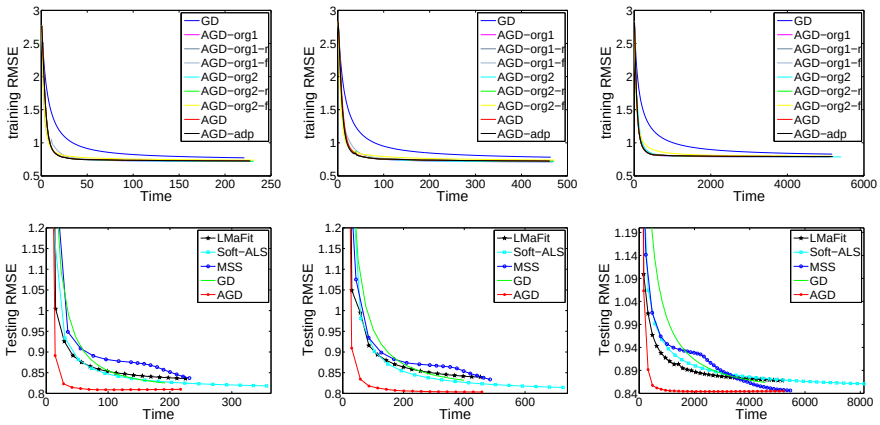
In matrix completion [61, 62, 63], the goal is to recover the low rank matrix $\mathbf{X}^*$ based on a set of randomly observed entries $\mathbf{O}$ from $\mathbf{X}^*$. The traditional matrix completion problem is to solve the following model:

$$\min_{\mathbf{X}} \frac{1}{2} \sum_{(i,j)\in\mathbf{O}} (\mathbf{X}_{i,j} - \mathbf{X}^*_{i,j})^2, \quad s.t. \quad \mathrm{rank}(\mathbf{X}) \leq r.$$

We consider the asymmetric case and solve the following model:

$$\min_{\widetilde{\mathbf{U}}\in\mathbb{R}^{n\times r}, \widetilde{\mathbf{V}}\in\mathbb{R}^{m\times r}} \frac{1}{2} \sum_{(i,j)\in\mathbf{O}} ((\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T)_{i,j} - \mathbf{X}^*_{i,j})^2 + \frac{1}{200}\|\widetilde{\mathbf{U}}^T\widetilde{\mathbf{U}} - \widetilde{\mathbf{V}}^T\widetilde{\mathbf{V}}\|_F^2.$$

We set $r = 10$ and test the algorithms on the Movielen-10M, Movielen-20M and Netflix data sets. The corresponding observed matrices are of size 69878×10677 with $o\% = 1.34\%$, 138493×26744 with $o\% = 0.54\%$ and 480189×17770 with $o\% = 1.18\%$, respectively, where $o\%$ means the percentage of the observed entries. We compare AGD and AGD-adp (AGD with adaptive index sets selection) with GD and several variants of the original AGD:



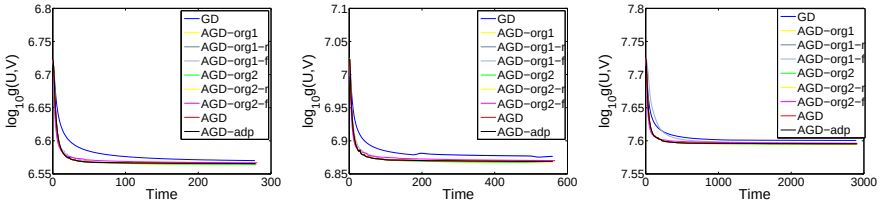
(a) MovieLens-10M (b) MovieLens-20M (c) Netflix

Fig. 1 Top: Compare the training RMSE of GD, AGD, AGD-adp and several variants of the original AGD. Bottom: Compare the testing RMSE of GD, AGD, LMaFit, Soft-ALS and MSS.

1. AGD-original1: The classical AGD with recursions of (31)-(32).
2. AGD-original1-r: AGD-original1 with restart.
3. AGD-original1-f: AGD-original1 with fixed β_k of $\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$.
4. AGD-original2: The classical AGD with recursions of (28)-(30).
5. AGD-original2-r: AGD-original2 with restart.
6. AGD-original2-f: AGD-original2 with fixed θ.

Let $\mathbf{X_O}$ be the observed data and $\mathbf{A\Sigma B}^T$ be its SVD. We initialize $\widetilde{\mathbf{U}} = \mathbf{A}_{:,1:r}\sqrt{\boldsymbol{\Sigma}_{1:r,1:r}}$ and $\widetilde{\mathbf{V}} = \mathbf{B}_{:,1:r}\sqrt{\boldsymbol{\Sigma}_{1:r,1:r}}$ for all the compared methods. Since $\mathbf{X_O}$ is sparse, it is efficient to find the top r singular values and the corresponding singular vectors for large scale matrices [64]. We tune the best step sizes of $\eta = 5\times10^{-5}, 4\times10^{-5}$ and 1×10^{-5} for all the compared methods on the three data sets, respectively. For AGD, we set $\epsilon = 10^{-10}$, $S^1 = \{1:r\}$ and $S^2 = \{r+1:2r\}$ for simplicity. We set $K = 100$ for AGD, AGD-adp and the original AGD with restart. We run the compared methods 500 iterations for the Movielen-10M and Movielen-20M data sets and 1000 iterations for the Netflix data set.

The top part of Figure 1 plots the curves of the training RMSE v.s. time (seconds). We can see that AGD is faster than GD. The performances of AGD, AGD-adp and the original AGD are similar. In fact, in AGD-adp, we observe that the index sets do not change during the iterations. Thus, the condition of $\sigma_r(\mathbf{U}_{S'}^{t,K+1}) \geq \epsilon\ \forall t$ in Theorem 4 holds. The original AGD performs almost equally fast as our modified AGD in practice. However, it has an inferior convergence rate theoretically. The bottom part of Figure 1 plots the curves of the testing RMSE v.s. time. Besides GD, we also compare AGD with LMaFit [65], Soft-ALS [66] and MSS [67]. They all solve a factorization based nonconvex



(a) MovieLens-10M (b) MovieLens-20M (c) Netflix

Fig. 2 Compare AGD and AGD-adp with GD and several variants of the original AGD on the One Bit Matrix Completion problem.

model. From Figure 1 we can see that AGD achieves the lowest testing RMSE with the fastest speed.

8.2 One Bit Matrix Completion

In one bit matrix completion [3], the sign of a random subset from the unknown low rank matrix $\mathbf{X}^*$ is observed, instead of observing the actual entries. Given a probability density function f, e.g., the logistic function $f(\mathbf{x}) = \frac{e^x}{1+e^x}$, we observe the sign of $\mathbf{x}$ as $+1$ with probability $f(\mathbf{x})$ and observe the sign as -1 with probability $1 - f(\mathbf{x})$. The training objective is to minimize the negative log-likelihood:

$$\min_{\mathbf{X}} - \sum_{(i,j)\in\mathbf{O}} \{\mathbf{1}_{\mathbf{Y}_{i,j}=1}\log(f(\mathbf{X}_{i,j})) + \mathbf{1}_{\mathbf{Y}_{i,j}=-1}\log(1 - f(\mathbf{X}_{i,j}))\}, s.t. \mathrm{rank}(\mathbf{X}) \leq r.$$

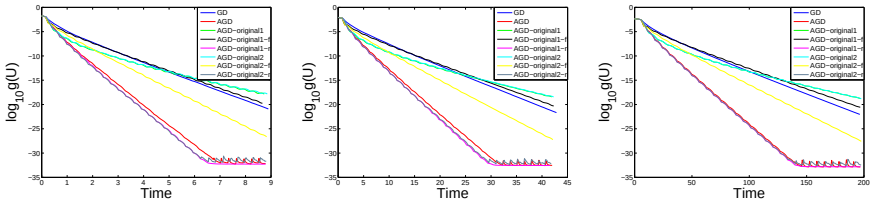
In this section, we solve the following model:

$$\min_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}} - \sum_{(i,j)\in\mathbf{O}} \left\{\mathbf{1}_{\mathbf{Y}_{i,j}=1}\log(f((\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T)_{i,j})) + \mathbf{1}_{\mathbf{Y}_{i,j}=-1}\log(1 - f((\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T)_{i,j}))\right\} + \frac{1}{200}\|\widetilde{\mathbf{U}}^T\widetilde{\mathbf{U}} - \widetilde{\mathbf{V}}^T\widetilde{\mathbf{V}}\|_F^2.$$

We use the data sets of Movielen-10M, Movielen-20M and Netflix. We set $\mathbf{Y}_{i,j} = 1$ if the (i,j)-th observation is larger than the average of all observations and $\mathbf{Y}_{i,j} = -1$, otherwise. We set $r = 5$ and $\eta = 0.001,\ 0.001, 0.0005$ for all the compared methods on the three data sets. The other experimental setting is the same as Matrix Completion. We run all the methods for 500 iterations. Figure 2 plots the curves of the objective value v.s. time (seconds) and we can see that AGD is also faster than GD. The performances of AGD, AGD-adp and the original AGD are nearly the same.

8.3 Matrix Regression

In matrix regression [68, 69], the goal is to estimate the unknown low rank matrix $\mathbf{X}^*$ from a set of measurements $\mathbf{y} = \mathbf{A}(\mathbf{X}^*) + \varepsilon$, where $\mathbf{A}$ is a linear



(a) n=512 (b) n=1024 (c) n=2048

Fig. 3 Compare AGD with GD and several variants of the original AGD on the Matrix Regression problem.

operator and ε is the noise. A reasonable estimation of $\mathbf{X}^*$ is to solve the following rank constrained problem:

$$\min_{\mathbf{X}} f(\mathbf{X}) = \frac{1}{2}\|\mathbf{A}(\mathbf{X}) - \mathbf{y}\|_F^2, \quad s.t. \quad \text{rank}(\mathbf{X}) \leq r.$$

We consider the symmetric case of $\mathbf{X}$ and solve the following nonconvex model:

$$\min_{\mathbf{U}\in\mathbb{R}^{n\times r}} f(\mathbf{U}) = \frac{1}{2}\|\mathbf{A}(\mathbf{U}\mathbf{U}^T) - \mathbf{y}\|_F^2.$$

We follow [24] to use the permuted and sub-sampled noiselets [70] for the linear operator $\mathbf{A}$ and $\mathbf{U}^*$ is generated from the normal Gaussian distribution without noise. We set $r = 10$ and test different n with n=512, 1024 and 2048. We fix the number of measurements to $4nr$ and follow [24] to use the initializer from the eigenvalue decomposition of $\frac{\mathbf{X}^0+(\mathbf{X}^0)^T}{2}$ for all the compared methods, where $\mathbf{X}^0 = \text{Project}_+\left(\frac{-\nabla f(0)}{\|\nabla f(0)-\nabla f(11^T)\|_F}\right)$. We set $\eta = 5, 10$ and 20 for all the compared methods for $n = 512, 1024$ and 2048, respectively. In AGD, we set $\epsilon = 10^{-10}$, $K = 10$, $S^1 = \{1 : r\}$ and $S^2 = \{r + 1 : 2r\}$. Figure 3 plots the curves of the objective value v.s. time (seconds). We run all the compared methods for 300 iterations. We can see that AGD and the original AGD with restart perform almost equally fast. AGD runs faster than GD and the original AGD without restart.

8.4 Verifying (22) in Practice

In this section, we verify that the conditions of $\|\hat{\mathbf{U}}_S^{t,k} - \mathbf{U}_S^*\|_F \leq c\sqrt{\frac{r}{n}}\|\hat{\mathbf{U}}^{t,k} - \mathbf{U}^*\|_F$ and $\|\hat{\mathbf{V}}_S^{t,k} - \mathbf{U}_S^*\|_F \leq c\sqrt{\frac{r}{n}}\|\hat{\mathbf{V}}^{t,k} - \mathbf{U}^*\|_F$ in (22) hold in our experiments, where $\hat{\mathbf{U}}^{t,k} = \mathbf{U}^{t,k}\mathbf{R}$ with $\mathbf{R} = \text{argmin}_{\mathbf{R}\in\mathbb{R}^{r\times r},\mathbf{R}\mathbf{R}^T=\mathbf{I}}\|\mathbf{U}^{t,k}\mathbf{R} - \mathbf{U}^*\|_F^2$ and $\hat{\mathbf{V}}^{t,k}$ is defined similarly. We use the final output $\mathbf{U}^{T,K+1}$ as $\mathbf{U}^*$. Table 2 lists the results. We can see that $\frac{\|\hat{\mathbf{U}}_S^{t,k}-\mathbf{U}_S^*\|_F}{\|\hat{\mathbf{U}}^{t,k}-\mathbf{U}^*\|_F}$ and $\frac{\|\hat{\mathbf{V}}_S^{t,k}-\mathbf{U}_S^*\|_F}{\|\hat{\mathbf{V}}^{t,k}-\mathbf{U}^*\|_F}$ have the same order as $\sqrt{\frac{r}{n}}$.

Table 2 Testing the order of $\frac{\|\hat{\mathbf{U}}_S^{t,k}-\mathbf{U}_S^*\|_F}{\|\hat{\mathbf{U}}^{t,k}-\mathbf{U}^*\|_F}$ and $\frac{\|\hat{\mathbf{V}}_S^{t,k}-\mathbf{U}_S^*\|_F}{\|\hat{\mathbf{V}}^{t,k}-\mathbf{U}^*\|_F}$.

Problem	Data	$\frac{\|\hat{\mathbf{U}}_S^{t,k}-\mathbf{U}_S^*\|_F}{\|\hat{\mathbf{U}}^{t,k}-\mathbf{U}^*\|_F}$			$\frac{\|\hat{\mathbf{V}}_S^{t,k}-\mathbf{U}_S^*\|_F}{\|\hat{\mathbf{V}}^{t,k}-\mathbf{U}^*\|_F}$			$\sqrt{\frac{r}{n}}$
		max	average	min	max	average	min	
MR	512	0.1536	0.1521	0.1505	0.1536	0.1521	0.1505	0.1398
	1024	0.0984	0.0939	0.0894	0.0984	0.0939	0.0894	0.0988
	2048	0.0715	0.0681	0.0648	0.0715	0.0681	0.0648	0.0699
1bit-MC	512	0.0344	0.0086	0.0021	0.0344	0.0086	0.0017	0.0079
	1024	0.0330	0.0077	0.0022	0.0329	0.0077	0.0020	0.0055
	2048	0.0189	0.0103	0.0068	0.0151	0.0103	0.0062	0.0032
MC	512	0.0664	0.0280	0.0191	0.0664	0.0280	0.0191	0.0111
	1024	0.0569	0.0230	0.0151	0.0569	0.0230	0.0139	0.0078
	2048	0.0346	0.0191	0.0105	0.0346	0.0190	0.0104	0.0045

9 Conclusions

In this paper, we study the factorization based low rank optimization. A linearly convergent accelerated gradient method with alternating constraint is proposed with the optimal dependence on the condition number of $\sqrt{L/\mu}$ as convex programming. As far as we know, this is the first work with the provable optimal dependence on $\sqrt{L/\mu}$ for this kind of nonconvex problems. Globally, the convergence to a critical point is proved.

There are two problems unsolved in this paper. 1. How to find two distinct sets S^1 and S^2 such that $\sigma_r(\mathbf{U}_{S^1})$ and $\sigma_r(\mathbf{U}_{S^2})$ are as large as possible? 2. How to find the initial point close enough to the optimum solution for the general problems with large condition number?

Acknowledgement

Zhouchen Lin is supported by National Basic Research Program of China (973 Program) (grant no. 2015CB352502), National Natural Science Foundation (NSF) of China (grant nos. 61625301 and 61731018), Qualcomm and Microsoft Research Asia.

Appendix A

Lemma 7 *For problem (1) and its minimizer* $\mathbf{X}^*$*, we have*

$$\nabla f(\mathbf{X}^*) \succeq 0.$$

Proof Introduce the Lagrange function

$$L(\mathbf{X}, \boldsymbol{\Lambda}) = f(\mathbf{X}) + \langle \boldsymbol{\Lambda}, \mathbf{X} \rangle .$$

Since $\mathbf{X}^*$ is the minimizer of problem (1), we know that there exists $\boldsymbol{\Lambda}^*$ such that

$$\begin{aligned}&\nabla f(\mathbf{X}^*) + \boldsymbol{\Lambda}^* = 0,\\ &\langle \boldsymbol{\Lambda}^*, \mathbf{X}^*\rangle = 0, \quad \mathbf{X}^* \succeq 0, \quad \boldsymbol{\Lambda}^* \preceq 0.\end{aligned}$$

Thus, we can have the conclusion. □

Lemma 8 *[14] For any $\mathbf{U} \in \mathbb{R}^{n\times r}, \mathbf{V} \in \mathbb{R}^{n\times r}$, let $\mathbf{R} = argmin_{\mathbf{R}\mathbf{R}^T=\mathbf{I}}\|\mathbf{V}\mathbf{R} - \mathbf{U}\|_F^2$ and $\hat{\mathbf{V}} = \mathbf{V}\mathbf{R}$. Then, we can have*

$$\|\mathbf{V}\mathbf{V}^T - \mathbf{U}\mathbf{U}^T\|_F^2 \geq (2\sqrt{2}-2)\sigma_r^2(\mathbf{U})\|\hat{\mathbf{V}} - \mathbf{U}\|_F^2.$$

Lemma 9 *[24] Assume that $\|\mathbf{U} - \mathbf{U}^*\|_F \leq 0.01\sigma_r(\mathbf{U}^*)$. Then, we can have*

$$\begin{aligned}0.99\sigma_r(\mathbf{U}^*) \leq \sigma_r(\mathbf{U}) \leq 1.01\sigma_r(\mathbf{U}^*),\\ 0.99\|\mathbf{U}^*\|_2 \leq \|\mathbf{U}\|_2 \leq 1.01\|\mathbf{U}^*\|_2.\end{aligned}$$

Lemma 10 *For any $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n\times r}$, we have*

$$\|\mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T\|_F \leq (\|\mathbf{U}\|_2 + \|\mathbf{V}\|_2)\|\mathbf{U} - \mathbf{V}\|_F, \tag{35}$$

$$\|\nabla f(\mathbf{V}\mathbf{V}^T) - \nabla f(\mathbf{U}^*\mathbf{U}^{*T})\|_2 \leq L(\|\mathbf{V}\|_2 + \|\mathbf{U}^*\|_2)\|\mathbf{V} - \mathbf{U}^*\|_F, \tag{36}$$

$$\|\nabla g(\mathbf{U}) - \nabla g(\mathbf{V})\|_F \leq \big(L\|\mathbf{U}\|_2(\|\mathbf{U}\|_2 + \|\mathbf{V}\|_2) + \|\nabla f(\mathbf{V}\mathbf{V}^T)\|_2\big)\|\mathbf{U} - \mathbf{V}\|_F \tag{37}$$

Proof For the first inequality, we have

$$\begin{aligned}&\|\mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T\|_F\\ \leq&\|\mathbf{U}\mathbf{U}^T - \mathbf{U}\mathbf{V}^T\|_F + \|\mathbf{U}\mathbf{V}^T - \mathbf{V}\mathbf{V}^T\|_F\\ \leq&\|\mathbf{U}\|_2\|\mathbf{U} - \mathbf{V}\|_F + \|\mathbf{V}\|_2\|\mathbf{U} - \mathbf{V}\|_F\\ =&(\|\mathbf{U}\|_2 + \|\mathbf{V}\|_2)\|\mathbf{U} - \mathbf{V}\|_F.\end{aligned}$$

For the second one, we have

$$\begin{aligned}\|\nabla f(\mathbf{V}\mathbf{V}^T) - f(\mathbf{U}^*\mathbf{U}^{*T})\|_F &\leq L\|\mathbf{V}\mathbf{V}^T - \mathbf{U}^*\mathbf{U}^{*T}\|_F\\ &\leq L(\|\mathbf{V}\|_2 + \|\mathbf{U}^*\|_2)\|\mathbf{V} - \mathbf{U}^*\|_F,\end{aligned}$$

where we use (35). For the third one, we have

$$\begin{aligned}&\|\nabla f(\mathbf{U}\mathbf{U}^T)\mathbf{U} - \nabla f(\mathbf{V}\mathbf{V}^T)\mathbf{V}\|_F\\ \leq&\|\nabla f(\mathbf{U}\mathbf{U}^T)\mathbf{U} - \nabla f(\mathbf{V}\mathbf{V}^T)\mathbf{U}\|_F + \|\nabla f(\mathbf{V}\mathbf{V}^T)\mathbf{U} - \nabla f(\mathbf{V}\mathbf{V}^T)\mathbf{V}\|_F\\ \leq&\|\mathbf{U}\|_2\|\nabla f(\mathbf{U}\mathbf{U}^T) - \nabla f(\mathbf{V}\mathbf{V}^T)\|_F + \|\nabla f(\mathbf{V}\mathbf{V}^T)\|_2\|\mathbf{U} - \mathbf{V}\|_F\\ \leq&L\|\mathbf{U}\|_2(\|\mathbf{U}\|_2 + \|\mathbf{V}\|_2)\|\mathbf{U} - \mathbf{V}\|_F + \|\nabla f(\mathbf{V}\mathbf{V}^T)\|_2\|\mathbf{U} - \mathbf{V}\|_F,\end{aligned}$$

where we use the restricted smoothness of f and (35) in the last inequality. □

Now we give the proof of Corollary 1.

Proof From Lemma 9 and the assumptions, we have

$$\begin{gathered}\|\mathbf{U}-\mathbf{U}^*\|_F \le 0.01\sigma_r(\mathbf{U}^*),\\ \|\mathbf{U}_S-\mathbf{U}_S^*\|_F \le 0.01\sigma_r(\mathbf{U}_S^*),\\ 0.99\sigma_r(\mathbf{U}^*) \le \sigma_r(\mathbf{U}) \le 1.01\sigma_r(\mathbf{U}^*),\\ 0.99\|\mathbf{U}^*\|_2 \le \|\mathbf{U}\|_2 \le 1.01\|\mathbf{U}^*\|_2,\\ 0.99\sigma_r(\mathbf{U}_S^*) \le \sigma_r(\mathbf{U}_S) \le 1.01\sigma_r(\mathbf{U}_S^*),\\ 0.99\|\mathbf{U}_S^*\|_2 \le \|\mathbf{U}_S\|_2 \le 1.01\|\mathbf{U}_S^*\|_2,\end{gathered}$$

where $\mathbf{U}$ can be $\mathbf{U}^k$, $\mathbf{V}^k$ and $\mathbf{Z}^k$. From (36), we have

$$\begin{aligned}&\|\nabla f(\mathbf{V}^k(\mathbf{V}^k)^T)\|_2 \le \|\nabla f(\mathbf{V}^k(\mathbf{V}^k)^T)-\nabla f(\mathbf{X}^*)\|_2 + \|\nabla f(\mathbf{X}^*)\|_2\\ &\le 2.01L\|\mathbf{U}^*\|_2\|\mathbf{V}^k-\mathbf{U}^*\|_F + \|\nabla f(\mathbf{X}^*)\|_2 \le 0.0201L\|\mathbf{U}^*\|_2^2 + \|\nabla f(\mathbf{X}^*)\|_2,\end{aligned} \tag{38}$$

where we use $\|\mathbf{V}^k-\mathbf{U}^*\|_F \le 0.01\|\mathbf{U}^*\|_2$. On the other hand, let

$$\hat{\mathbf{Z}}^{k+1} = \mathbf{Z}^k - \frac{\eta}{\theta_k}\nabla g(\mathbf{V}^k)$$

then we have $\mathbf{Z}^{k+1} = \mathrm{Project}_{\Omega_S}(\hat{\mathbf{Z}}^{k+1})$ and

$$\begin{aligned}\|\hat{\mathbf{Z}}^{k+1}\|_2 &\le \|\mathbf{Z}^k\|_2 + \frac{2\eta}{\theta_k}\|\nabla f(\mathbf{V}^k(\mathbf{V}^k)^T)\|_2\|\mathbf{V}^k\|_2\\ &\le 1.01\|\mathbf{U}^*\|_2\left(1+\frac{(0.0402L\|\mathbf{U}^*\|_2^2+2\|\nabla f(\mathbf{X}^*)\|_2)\eta}{\theta_k}\right)\\ &\le 1.01\|\mathbf{U}^*\|_2\left(1+\frac{1}{\theta_k}\right),\end{aligned}$$

where we use $\|\mathbf{Z}^k\|_2 \le 1.01\|\mathbf{U}^*\|_2$, $\|\mathbf{V}^k\|_2 \le 1.01\|\mathbf{U}^*\|_2$, (38) and the setting of η. Let $\hat{\Omega}_S = \{\mathbf{U}_S \in \mathbb{R}^{r\times r} : \mathbf{U}_S \succeq \epsilon\mathbf{I}\}$, then

$$\begin{aligned}&\mathrm{Project}_{\hat{\Omega}_S}(\hat{\mathbf{Z}}_S^{k+1})\\ &= \mathrm{argmin}_{\mathbf{U}\in\hat{\Omega}_S}\|\mathbf{U}-\hat{\mathbf{Z}}_S^{k+1}\|_F^2\\ &= \mathrm{argmin}_{\mathbf{U}\in\hat{\Omega}_S}\left\|\mathbf{U}-\frac{\hat{\mathbf{Z}}_S^{k+1}+(\hat{\mathbf{Z}}_S^{k+1})^T}{2}-\frac{\hat{\mathbf{Z}}_S^{k+1}-(\hat{\mathbf{Z}}_S^{k+1})^T}{2}\right\|_F^2\\ &= \mathrm{argmin}_{\mathbf{U}\in\hat{\Omega}_S}\left\|\mathbf{U}-\frac{\hat{\mathbf{Z}}_S^{k+1}+(\hat{\mathbf{Z}}_S^{k+1})^T}{2}\right\|_F^2+\left\|\frac{\hat{\mathbf{Z}}_S^{k+1}-(\hat{\mathbf{Z}}_S^{k+1})^T}{2}\right\|_F^2,\end{aligned}$$

where we use $\mathrm{trace}(\mathbf{AB})=0$ if $\mathbf{A}=\mathbf{A}^T$ and $\mathbf{B}=-\mathbf{B}^T$, and $\mathbf{U}=\mathbf{U}^T$ from $\mathbf{U}\in\hat{\Omega}_S$. Let $\mathbf{U}\Sigma\mathbf{U}^T$ be the eigenvalue decomposition of $\frac{\hat{\mathbf{Z}}_S^{k+1}+(\hat{\mathbf{Z}}_S^{k+1})^T}{2}$ and $\hat{\Sigma}_{i,i}=\max\{\epsilon,\Sigma_{i,i}\}$. Then $\mathrm{Project}_{\hat{\Omega}_S}(\hat{\mathbf{Z}}_S^{k+1}) = \mathbf{U}\hat{\Sigma}\mathbf{U}^T$ and

$$\begin{aligned}&\|\mathrm{Project}_{\hat{\Omega}_S}(\hat{\mathbf{Z}}_S^{k+1})\|_2 = \max\{\epsilon,\Sigma_{1,1}\}\\ &\le \max\left\{\epsilon,\left\|\frac{\hat{\mathbf{Z}}_S^{k+1}+(\hat{\mathbf{Z}}_S^{k+1})^T}{2}\right\|_2\right\} \le \max\left\{\epsilon,\|\hat{\mathbf{Z}}_S^{k+1}\|_2\right\},\end{aligned}$$

where $\Sigma_{1,1}$ is the largest eigenvalue of $\frac{\hat{\mathbf{Z}}_S^{k+1}+(\hat{\mathbf{Z}}_S^{k+1})^T}{2}$. Then, we have

$$\begin{aligned}\|\mathbf{Z}^{k+1}\|_2 &\le \|\mathbf{Z}_S^{k+1}\|_2 + \|\mathbf{Z}_{-S}^{k+1}\|_2\\ &= \|\mathrm{Project}_{\hat{\Omega}_S}(\hat{\mathbf{Z}}_S^{k+1})\|_2 + \|\hat{\mathbf{Z}}_{-S}^{k+1}\|_2\\ &\le \max\left\{\epsilon, \|\hat{\mathbf{Z}}_S^{k+1}\|_2\right\} + \|\hat{\mathbf{Z}}_{-S}^{k+1}\|_2\\ &\le \max\left\{\epsilon, \|\hat{\mathbf{Z}}^{k+1}\|_2\right\} + \|\hat{\mathbf{Z}}^{k+1}\|_2\\ &\le 2\max\left\{\epsilon, \|\hat{\mathbf{Z}}^{k+1}\|_2\right\}\end{aligned}$$

and

$$\begin{aligned}\|\mathbf{U}^{k+1}\|_2 &\le (1-\theta_k)\|\mathbf{U}^k\|_2 + \theta_k\|\mathbf{Z}^{k+1}\|_2\\ &\le 1.01(1-\theta_k)\|\mathbf{U}^*\|_2 + 2\theta_k\max\left\{\epsilon, 1.01\|\mathbf{U}^*\|_2\left(1+\frac{1}{\theta_k}\right)\right\}\\ &\le 1.01(1-\theta_k)\|\mathbf{U}^*\|_2 + \max\left\{2\epsilon, 1.01\|\mathbf{U}^*\|_2\,(2+2)\right\}\\ &\le 5.05\|\mathbf{U}^*\|_2,\end{aligned}$$

where we use (14) in the first inequality, $0 \le \theta_k \le 1$ in the third and forth inequality and $\|\mathbf{U}^*\|_2 \ge \|\mathbf{U}_S^*\|_2 \ge \sigma_r(\mathbf{U}_S^*) \ge \epsilon$ in the last inequality. So

$$\begin{aligned}&\|\nabla f(\mathbf{V}^k(\mathbf{V}^k)^T)\|_2 + \frac{L(\|\mathbf{V}^k\|_2 + \|\mathbf{U}^{k+1}\|_2)^2}{2}\\ \le\ & 0.0201L\|\mathbf{U}^*\|_2^2 + \|\nabla f(\mathbf{X}^*)\|_2 + \frac{L(6.06\|\mathbf{U}^*\|_2)^2}{2} \le \frac{L_g}{2}.\end{aligned}$$

From Theorem 2, we can have the conclusion. □

Appendix B

Lemma 11 *Assume that $\mathbf{U}^* \in \mathcal{X}^*$. Then, for any $\mathbf{U}$, we have*

$$g(\mathbf{U}) - g(\mathbf{U}^*) \ge 0.4\mu\sigma_r^2(\mathbf{U}^*)\|P_{\mathcal{X}^*}(\mathbf{U}) - \mathbf{U}\|_F^2.$$

Proof From (10), we have

$$\begin{aligned}&f(\mathbf{U}^*\mathbf{U}^{*T}) - f(\mathbf{U}\mathbf{U}^T)\\ \le\ & 2\left\langle\nabla f(\mathbf{U}^*\mathbf{U}^{*T})\mathbf{U}^*, \mathbf{U}^* - \mathbf{U}\right\rangle - \left\langle\nabla f(\mathbf{U}^*\mathbf{U}^{*T}), (\mathbf{U}^*-\mathbf{U})(\mathbf{U}^*-\mathbf{U})^T\right\rangle\\ &-\frac{\mu}{2}\|\mathbf{U}^*\mathbf{U}^{*T} - \mathbf{U}\mathbf{U}^T\|_F^2.\end{aligned}$$

Since $\mathbf{U}^*$ is a minimizer of problem (2), we have $\nabla f(\mathbf{U}^*\mathbf{U}^{*T})\mathbf{U}^* = 0$. From $\left\langle\nabla f(\mathbf{U}^*\mathbf{U}^{*T}), (\mathbf{U}^*-\mathbf{U})(\mathbf{U}^*-\mathbf{U})^T\right\rangle \ge 0$ and Lemma 8, we can have the conclusion. □

Now we give the proof of Theorem 3.

Proof From (17), we have

$$\begin{aligned}
&\|\mathbf{U}^{t+1,0}-\mathbf{U}^{t+1,*}\|_F\\
\le&\left(\frac{1}{4}\right)^{t+1}\|\mathbf{U}^{0,0}-\mathbf{U}^{0,*}\|_F\\
=&\left(4^{-\frac{\sqrt{\eta\mu}\sigma_r(\mathbf{U}^*)\min\{\sigma_r(\mathbf{U}^*_{S^1}),\sigma_r(\mathbf{U}^*_{S^2})\}}{28\|\mathbf{U}^*\|_2}}\right)^{(t+1)(K+1)}\|\mathbf{U}^{0,0}-\mathbf{U}^{0,*}\|_F\\
\le&\left(1-\frac{\sqrt{\eta\mu}\sigma_r(\mathbf{U}^*)\min\{\sigma_r(\mathbf{U}^*_{S^1}),\sigma_r(\mathbf{U}^*_{S^2})\}}{28\|\mathbf{U}^*\|_2}\right)^{(t+1)(K+1)}\|\mathbf{U}^{0,0}-\mathbf{U}^{0,*}\|_F,
\end{aligned}$$

where we use $4^{-x}\le e^{-x}\le 1-x$.

From Theorem 2 and $\nabla g(\mathbf{U}^{t+1,*})=0$, we have

$$g(\mathbf{U}^{t+1,0})-g(\mathbf{U}^*)=g(\mathbf{U}^{t+1,0})-g(\mathbf{U}^{t+1,*})\le\frac{L_g}{2}\|\mathbf{U}^{t+1,0}-\mathbf{U}^{t+1,*}\|_F^2,$$

which leads to the conclusion. □

Appendix C

Proof of Lemma 6.

Proof We can easily check that $\beta_{\max}<1$ due to $\beta_k\le 1-\theta_{k-1}$ and the fact that K a finite constant. From Theorem 2, we have

$$\begin{aligned}
g(\mathbf{U}^{k+1})&\le g(\mathbf{U}^k)+\left\langle\nabla g(\mathbf{U}^k),\mathbf{U}^{k+1}-\mathbf{U}^k\right\rangle+\frac{\hat{L}}{2}\|\mathbf{U}^{k+1}-\mathbf{U}^k\|_F^2\\
&=g(\mathbf{U}^k)+\left\langle\nabla g(\mathbf{U}^k)-\nabla g(\mathbf{V}^k),\mathbf{U}^{k+1}-\mathbf{U}^k\right\rangle\\
&\quad+\left\langle\nabla g(\mathbf{V}^k),\mathbf{U}^{k+1}-\mathbf{U}^k\right\rangle+\frac{\hat{L}}{2}\|\mathbf{U}^{k+1}-\mathbf{U}^k\|_F^2.
\end{aligned}$$

Applying the inequality of $\langle\mathbf{u},\mathbf{v}\rangle\le\|\mathbf{u}\|\|\mathbf{v}\|$, Lemma 10 and the inequality of $2\|\mathbf{u}\|\|\mathbf{v}\|\le\alpha\|\mathbf{u}\|^2+\frac{1}{\alpha}\|\mathbf{v}\|^2$ to the second term, we can have

$$\begin{aligned}
g(\mathbf{U}^{k+1})&\le g(\mathbf{U}^k)+\frac{\hat{L}}{2}\left(\alpha\|\mathbf{U}^k-\mathbf{V}^k\|_F^2+\frac{1}{\alpha}\|\mathbf{U}^{k+1}-\mathbf{U}^k\|_F^2\right)\\
&\quad+\left\langle\nabla g(\mathbf{V}^k),\mathbf{U}^{k+1}-\mathbf{U}^k\right\rangle+\frac{\hat{L}}{2}\|\mathbf{U}^{k+1}-\mathbf{U}^k\|_F^2.
\end{aligned}$$

Applying Lemma 12 in Appendix C to bound the third term, we can have

$$\begin{aligned}
&g(\mathbf{U}^{k+1}) - g(\mathbf{U}^k) \\
\leq &\frac{\hat{L}\alpha}{2}\|\mathbf{U}^k - \mathbf{V}^k\|_F^2 + \frac{\hat{L}}{2\alpha}\|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2 + \frac{1}{2\eta}\|\mathbf{U}^k - \mathbf{V}^k\|_F^2 \\
&- \frac{1}{2\eta}\|\mathbf{U}^k - \mathbf{U}^{k+1}\|_F^2 + \frac{\hat{L}}{2}\|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2 \\
= &\beta_k^2\left(\frac{1}{2\eta} + \frac{\hat{L}\alpha}{2}\right)\|\mathbf{U}^k - \mathbf{U}^{k-1}\|_F^2 - \left(\frac{1}{2\eta} - \frac{\hat{L}}{2} - \frac{\hat{L}}{2\alpha}\right)\|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2
\end{aligned} \tag{39}$$

for all $k = 1, 2, \cdots, K$, where we use $\mathbf{V}^k - \mathbf{U}^k = \beta_k(\mathbf{U}^k - \mathbf{U}^{k-1})$ proved in Lemma 12. Specially, from $\mathbf{U}^0 = \mathbf{V}^0$ we have

$$g(\mathbf{U}^1) \leq g(\mathbf{U}^0) - \left(\frac{1}{2\eta} - \frac{\hat{L}}{2} - \frac{\hat{L}}{2\alpha}\right)\|\mathbf{U}^1 - \mathbf{U}^0\|_F^2. \tag{40}$$

Summing (39) over $k = 1, 2, \ldots, K$ and (40), we have

$$\begin{aligned}
&g(\mathbf{U}^{K+1}) - g(\mathbf{U}^0) \\
\leq &-\sum_{k=0}^{K}\left(\left(\frac{1}{2\eta} - \frac{\hat{L}}{2} - \frac{\hat{L}}{2\alpha}\right) - \beta_{k+1}^2\left(\frac{1}{2\eta} + \frac{\hat{L}\alpha}{2}\right)\right)\|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2.
\end{aligned}$$

Letting $\alpha = 1/\beta_{\max}$, from the setting of η, we have the desired conclusion. □

Lemma 12 *For Algorithm 1, we have*

$$\left\langle \nabla g(\mathbf{V}^k), \mathbf{U}^{k+1} - \mathbf{U}^k \right\rangle \leq \frac{1}{2\eta}\|\mathbf{U}^k - \mathbf{V}^k\|_F^2 - \frac{1}{2\eta}\|\mathbf{U}^k - \mathbf{U}^{k+1}\|_F^2.$$

and

$$\mathbf{V}^k = \mathbf{U}^k + \beta_k(\mathbf{U}^k - \mathbf{U}^{k-1}).$$

Proof From the optimality condition of (13), we have

$$0 \in \frac{\theta_k}{\eta}\left(\mathbf{Z}^{k+1} - \mathbf{Z}^k\right) + \nabla g(\mathbf{V}^k) + \partial I_{\Omega_S}(\mathbf{Z}^{k+1}).$$

Since Ω_S is a convex set, we have

$$I_{\Omega_S}(\mathbf{U}) \geq I_{\Omega_S}(\mathbf{Z}^{k+1}) - \left\langle \frac{\theta_k}{\eta}\left(\mathbf{Z}^{k+1} - \mathbf{Z}^k\right) + \nabla g(\mathbf{V}^k), \mathbf{U} - \mathbf{Z}^{k+1} \right\rangle, \forall \mathbf{U} \in \Omega_S$$

and

$$\frac{\theta_k}{\eta}\left\langle \mathbf{Z}^{k+1} - \mathbf{Z}^k, \mathbf{U} - \mathbf{Z}^{k+1} \right\rangle \geq -\left\langle \nabla g(\mathbf{V}^k), \mathbf{U} - \mathbf{Z}^{k+1} \right\rangle, \forall \mathbf{U} \in \Omega_S. \tag{41}$$

With some simple computations, we have

$$\begin{aligned}
&\left\langle \nabla g(\mathbf{V}^k), \mathbf{U}^{k+1} - \mathbf{U}^k \right\rangle \\
=& \theta_k \left\langle \nabla g(\mathbf{V}^k), \mathbf{Z}^{k+1} - \mathbf{U}^k \right\rangle (\text{ from (14)}) \\
\leq& \frac{\theta_k^2}{\eta} \left\langle \mathbf{Z}^{k+1} - \mathbf{Z}^k, \mathbf{U}^k - \mathbf{Z}^{k+1} \right\rangle (\text{ from (41)}) \\
=& \frac{\theta_k}{\eta} \left\langle \mathbf{Z}^{k+1} - \mathbf{Z}^k, \mathbf{U}^k - \mathbf{U}^{k+1} \right\rangle (\text{ from (14)}) \\
=& \frac{1}{\eta} \left\langle \mathbf{U}^{k+1} - \mathbf{V}^k, \mathbf{U}^k - \mathbf{U}^{k+1} \right\rangle (\text{ from (12) and (14)}) \\
=& \frac{1}{2\eta} \left[\|\mathbf{U}^k - \mathbf{V}^k\|_F^2 - \|\mathbf{U}^{k+1} - \mathbf{V}^k\|_F^2 - \|\mathbf{U}^k - \mathbf{U}^{k+1}\|_F^2 \right] \\
\leq& \frac{1}{2\eta} \|\mathbf{U}^k - \mathbf{V}^k\|_F^2 - \frac{1}{2\eta} \|\mathbf{U}^k - \mathbf{U}^{k+1}\|_F^2.
\end{aligned}$$

From (12) and (14), we have

$$\mathbf{V}^k = (1-\theta_k)\mathbf{U}^k + \frac{\theta_k}{\theta_{k-1}}(\mathbf{U}^k - (1-\theta_{k-1})\mathbf{U}^{k-1}) = \mathbf{U}^k + \frac{\theta_k(1-\theta_{k-1})}{\theta_{k-1}}(\mathbf{U}^k - \mathbf{U}^{k-1}),$$

which leads to the second conclusion. □

Lemma 13 *Under the assumptions in Theorem 4, if (23) holds, then we have* $\|\mathbf{U}^{t',k+1} - \mathbf{Z}^{t',k+1}\|_F \leq \frac{2\varepsilon}{\theta_k}$, $\|\mathbf{Z}^{t',k+1} - \mathbf{Z}^{t',k}\|_F \leq \frac{5\varepsilon}{\theta_k}$ *and* $\|\mathbf{Z}^{t',k+1} - \mathbf{V}^{t',k}\|_F \leq \frac{9\varepsilon}{\theta_k}$ *for* $t' = t$ *or* $t' = t+1$.

Proof From (23), for $t' = t$ or $t+1$ and $\forall k = 0, \cdots, K$, we can have the following easy-to-check inequalities.

$$\mathbf{U}^{t',0} = \mathbf{V}^{t',0} = \mathbf{Z}^{t',0}, \tag{42}$$

$$\|\mathbf{U}^{t',k+1} - \mathbf{U}^{t',k}\|_F \leq \varepsilon, \tag{43}$$

$$\|\mathbf{Z}^{t',k+1} - \mathbf{U}^{t',k}\|_F \leq \frac{\varepsilon}{\theta_k}, (\text{ from (14)}) \tag{44}$$

$$\|\mathbf{U}^{t',k+1} - \mathbf{Z}^{t',k+1}\|_F \leq \varepsilon + \frac{\varepsilon}{\theta_k}, (\text{ from (43) and (44)}) \tag{45}$$

$$\|\mathbf{V}^{t',k+1} - \mathbf{U}^{t',k+1}\|_F \leq \theta_{k+1}\left(\varepsilon + \frac{\varepsilon}{\theta_k}\right), (\text{ from (12) and (45)}) \tag{46}$$

$$\|\mathbf{Z}^{t',k+1} - \mathbf{Z}^{t',k}\|_F \leq \varepsilon + \frac{\varepsilon}{\theta_k} + \varepsilon + \frac{\varepsilon}{\theta_{k-1}} + \varepsilon, (\text{ from (43) and (45)}) \tag{47}$$

$$\|\mathbf{Z}^{t',k+1} - \mathbf{V}^{t',k}\|_F \leq \varepsilon + \frac{\varepsilon}{\theta_k} + (2+\theta_k)\left(\varepsilon + \frac{\varepsilon}{\theta_{k-1}}\right) + \varepsilon, ((47), (45), (46)) \tag{48}$$

From $\theta_k \leq \theta_{k-1} \leq 1$, we can have the conclusions. □

Appendix D

Lemma 14 *Assume that* $\mathbf{U}^* \in \mathcal{X}^*$ *and* $\mathbf{V} \in \mathbb{R}^{n\times r}$ *satisfy* $\|\mathbf{V} - P_{\mathcal{X}^*}(\mathbf{V})\|_F \leq \min\left\{0.01\sigma_r(\mathbf{U}^*), \frac{\mu\sigma_r^2(\mathbf{U}^*)}{6L\|\mathbf{U}^*\|_2}\right\}$. *Then, we have*

$$\|\mathbf{V} - P_{\mathcal{X}^*}(\mathbf{V})\|_F \leq \frac{5}{\mu\sigma_r^2(\mathbf{U}^*)}\|\nabla g(\mathbf{V})\|_F.$$

Proof Similar to the proof of Theorem 1, we have

$$\begin{aligned} &g(\mathbf{U}^*) = g(P_{\mathcal{X}^*}(\mathbf{V})) \\ &\geq g(\mathbf{V}) + \langle\nabla g(\mathbf{V}), P_{\mathcal{X}^*}(\mathbf{V}) - \mathbf{V}\rangle + 0.2\mu\sigma_r^2(\mathbf{U}^*)\|\mathbf{V} - P_{\mathcal{X}^*}(\mathbf{V})\|_F^2, \end{aligned} \tag{49}$$

where we use Lemma 8 to bound $\|\mathbf{V}\mathbf{V}^T - P_{\mathcal{X}^*}(\mathbf{V})(P_{\mathcal{X}^*}(\mathbf{V}))^T\|_F^2$. Since $g(\mathbf{U}^*) \leq g(\mathbf{V})$, we can have

$$\begin{aligned} 0.2\mu\sigma_r^2(\mathbf{U}^*)\|\mathbf{V} - P_{\mathcal{X}^*}(\mathbf{V})\|_F^2 &\leq \langle\nabla g(\mathbf{V}), \mathbf{V} - P_{\mathcal{X}^*}(\mathbf{V})\rangle \\ &\leq \|\nabla g(\mathbf{V})\|_F\|\mathbf{V} - P_{\mathcal{X}^*}(\mathbf{V})\|_F, \end{aligned}$$

which leads to the conclusion. □

Lemma 15 *Under the assumptions of Lemma 6, we have*

$$\begin{aligned} &g(\mathbf{U}^{k+1}) + \nu\|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2 \\ &\leq g(\mathbf{U}^k) + \nu\|\mathbf{U}^k - \mathbf{U}^{k-1}\|_F^2 - \gamma\left(\|\mathbf{U}^k - \mathbf{U}^{k-1}\|_F^2 + \|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2\right), \end{aligned}$$

where $\gamma = \frac{1-\beta_{\max}^2}{4\eta} - \frac{\beta_{\max}\hat{L}}{2} - \frac{\hat{L}}{4} > 0$ *and* $\nu = \frac{1+\beta_{\max}^2}{4\eta} - \frac{\hat{L}}{4} > 0$.

Proof Letting $\alpha = \frac{1}{\beta_{\max}}$ in (39), we can have the conclusion. □

Now we give the proof of Theorem 5.

Proof Denote $\hat{\mathbf{U}}^* = P_{\mathcal{X}^*}(\mathbf{V}^k)$. From Theorem 2, we can have

$$\begin{aligned}
&g(\mathbf{U}^{k+1})\\
&\le g(\mathbf{V}^k) + \left\langle \nabla g(\mathbf{V}^k), \mathbf{U}^{k+1} - \mathbf{V}^k \right\rangle + \frac{\hat{L}}{2}\|\mathbf{U}^{k+1} - \mathbf{V}^k\|_F^2\\
&= g(\mathbf{V}^k) + \left\langle \nabla g(\mathbf{V}^k), \hat{\mathbf{U}}^* - \mathbf{V}^k \right\rangle + \left\langle \nabla g(\mathbf{V}^k), \mathbf{U}^{k+1} - \hat{\mathbf{U}}^* \right\rangle + \frac{\hat{L}}{2}\|\mathbf{U}^{k+1} - \mathbf{V}^k\|_F^2\\
&\le g(\hat{\mathbf{U}}^*) + \left\langle \nabla g(\mathbf{V}^k), \mathbf{U}^{k+1} - \hat{\mathbf{U}}^* \right\rangle + \frac{\hat{L}}{2}\|\mathbf{U}^{k+1} - \mathbf{V}^k\|_F^2\\
&= g(\hat{\mathbf{U}}^*) + \frac{1}{\eta}\left\langle \mathbf{V}^k - \mathbf{U}^{k+1}, \mathbf{U}^{k+1} - \hat{\mathbf{U}}^* \right\rangle + \frac{\hat{L}}{2}\|\mathbf{U}^{k+1} - \mathbf{V}^k\|_F^2\\
&\le g(\hat{\mathbf{U}}^*) + \frac{1}{\eta}\left\langle \mathbf{V}^k - \mathbf{U}^{k+1}, \mathbf{V}^k - \hat{\mathbf{U}}^* \right\rangle\\
&\le g(\hat{\mathbf{U}}^*) + \frac{1}{\eta}\|\mathbf{V}^k - \mathbf{U}^{k+1}\|_F\|\mathbf{V}^k - \hat{\mathbf{U}}^*\|_F\\
&\le g(\hat{\mathbf{U}}^*) + \frac{5}{\eta\mu\sigma_r^2(\mathbf{U}^*)}\|\mathbf{V}^k - \mathbf{U}^{k+1}\|_F\|\nabla g(\mathbf{V}^k)\|_F\\
&= g(\hat{\mathbf{U}}^*) + \frac{5}{\eta^2\mu\sigma_r^2(\mathbf{U}^*)}\|\mathbf{U}^{k+1} - \mathbf{V}^k\|_F^2\\
&= g(\hat{\mathbf{U}}^*) + \frac{5}{\eta^2\mu\sigma_r^2(\mathbf{U}^*)}\|\mathbf{U}^{k+1} - \mathbf{U}^k - \beta_k(\mathbf{U}^k - \mathbf{U}^{k-1})\|_F^2\\
&\le g(\hat{\mathbf{U}}^*) + \frac{5}{\eta^2\mu\sigma_r^2(\mathbf{U}^*)}\left(\|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2 + \|\mathbf{U}^k - \mathbf{U}^{k-1}\|_F^2\right),
\end{aligned}$$

where we use (49) in the second inequality, (32) in the second equality, $\eta < \frac{1}{\hat{L}}$ in the third inequality, Lemma 14 in the fifth inequality, (31) in the forth equality and $\beta_{\max} < 1$ in the last inequality. So we have

$$\begin{aligned}
&g(\mathbf{U}^{k+1}) + \nu\|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2 - g(\mathbf{U}^*)\\
&\le \left(\frac{5}{\eta^2\mu\sigma_r^2(\mathbf{U}^*)} + \nu\right)\left(\|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2 + \|\mathbf{U}^k - \mathbf{U}^{k-1}\|_F^2\right).
\end{aligned} \tag{50}$$

Combing Lemma 15 and (50), we can have the conclusion. □

References

1. S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *NIPS*, 2016.
2. P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC*, 2013.
3. M. Davenport, Y. Plan, E. Van Den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.
4. T. Cai, Z. Ma, and Y. Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.
5. M. Lin and J. Ye. A non-convex one-pass framework for generalized factorization machines and rank-one matrix sensing. In *NIPS*, 2016.

6. R. Ge, J. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *NIPS*, 2016.
7. R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *ICML*, 2017.
8. Q. Li, Z. Zhu, and G. Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2018.
9. Z. Zhu, Q. Li, G. Tang, and M. Wakin. The global optimization geometry on low-rank matrix optimization. *arxiv:1703.01256*, 2018.
10. X. Zhang, L. Wang, Y. Yu, and Q. Gu. A primal-dual analysis of global optimality in nonconvex low-rank matrix recovery. In *ICML*, 2018.
11. S. Burer and R. monteiro. A nonlinaer programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.
12. S. Burer and R. monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.
13. N. Boumal, V. Voroninski, and A. Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *NIPS*, 2016.
14. S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *ICML*, 2016.
15. Q. Zhang and J. Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *NIPS*, 2015.
16. D. Park, A. Kyrillidis, S. Bhojanapalli, C. Caramanis, and S. Sanghavi. Provable Burer-Monteiro factorization for a class of norm-constrained matrix problems. *arxiv:1606.01316*, 2016.
17. R. Sun and Z. Luo. Guaranteed matrix completion via nonconvex factorization. In *FOCS*, 2015.
18. D. Park, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC*, 2013.
19. M. Hardt and M. Wootters. Fast matrix completion without the condition number. In *COLT*, 2014.
20. Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arxiv:1605.07051*, 2016.
21. T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In *NIPS*, 2015.
22. X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust PCA via gradient descent. In *NIPS*, 2016.
23. Q. Gu, Z. Wang, and H. Liu. Low-rank and sparse structure pursuit via alternating minimization. In *AISTATS*, 2016.
24. S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi. Dropping convexity for faster semi-definite optimization. In *COLT*, 2016.
25. Y. Chen and M. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algoritmic guarantees. *arxiv:1509.03025*, 2015.
26. L. Wang, X. Zhang, and Q. Gu. A unified computational and statistical framework for nonconvex low rank matrix estimation. In *AISTATS*, 2017.
27. Yu. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.
28. Yu. Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex functions. *Èkonom. i. Mat. Metody*, 24:509–517, 1988.
29. Yu. Nesterov, editor. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2004.
30. I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013.
31. S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.
32. H. Li and Z. Lin. Accelerated proximal gradient methods for nonconvex programming. In *NIPS*, 2015.
33. Y. Xu and W. Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72(2):700–734, 2017.

34. Y. Carmon, J. Duchi, O. Hinder, and A. Sidford. Accelerated methods for non-convex optimization. *SIAM J. on Optimization*, 28(2):1751–1772, 2018.
35. Y. Carmon, O. Hinder, J. Duchi, and A. Sidford. Convex until proven guilty: Dimension-free acceleration of gradient descent on non-convex functions. In *ICML*, 2017.
36. N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma. Finding approximate local minima faster than gradient descent. In *STOC*, 2017.
37. C. Jin, P. Netrapalli, and M. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *COLT*, 2018.
38. I. Necoara, Yu. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1-2):69–107, 2019.
39. R. Li. New perturbation bounds for the unitary polar factor. *SIAM J. on Matrix Analysis and Applications*, 16(1):327–332, 1995.
40. C. Jin, R. Ge, P. Netrapalli, and S. Kakade. How to escape saddle points efficiently. In *ICML*, 2018.
41. X. Li, Z. Wang, J. Lu, R. Arora, J. Haupt, H. Liu, and T. Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arxiv:1612.09296*, 2016.
42. N. Halko, P. Martinsson, and J. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
43. V. Guruswami and A. Sinop. Optimal column-based low-rank matrix reconstruction. In *SODA*, 2012.
44. H. Avron and C. Boutsidis. Faster subset selection for matrices and applications. *SIAM J. on Matrix Analysis and Applications*, 34(4):1464–1499, 2013.
45. P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington, Seattle, 2008.
46. D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Finding low-rank solutions via non-convex matrix factorization, efficiently and provably. *SIAM J. on Image Science*, 11(4):333–361, 2018.
47. S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.
48. L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM J. on Optimization*, 24(4):2057–2075, 2014.
49. M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017.
50. S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1):105–145, 2016.
51. H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *NIPS*, 2015.
52. Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *STOC*, 2017.
53. B. Woodworth and N. Srebro. Tight complexity bounds for optimizing composite objectives. In *NIPS*, 2016.
54. E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.
55. G. Liu, Z. Lin, S. Yan, J. Sun, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.
56. A. Edelman. eigenvalues and condition numbers of random matrices. *SIAM J. on Matrix Analysis and Applications*, 9(4):543–560, 1988.
57. T. Yee and T. Hastie. Reduced-rank vector generalized linear models. *Statistical Modelling*, 3(1):15–41, 2000.
58. Y. She. Reduced rank vector generalized linear models for feature extraction. *Statistics and Its Interface*, 6(2):197–209, 2013.
59. A. Wagner and O. Zuk. Low-rank matrix recovery from row-and-column affine measurements. In *ICML*, 2015.
60. G. Liu and P. Li. Low-rank matrix completion in the presence of high coherence. *IEEE Trans. on Signal Processing*, 64(21):5623–5633, 2016.

61. A. Rohde and A. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.
62. V. Koltchinsii, K. Lounici, and A. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.
63. S. Negahban and M. Wainwright. Resticted strong convexity and weighted matrix completion: optimal bounds with noise. *Journal of Machine Learning Research*, 13(May):1665–1697, 2012.
64. R. Larsen. Lanczos bidiagonalization with partial reorthogonalization. Technical report, Aarhus University, 1998.
65. Z. Wen, W. Yin, and Y. Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012.
66. T. Hastie, R. Mazumder, J. Lee, and R. Zadeh. Matrix completion and low-rank SVD via fast alternating least squares. *Journal of Machine Learning Research*, 16(1):3367–3402, 2015.
67. C. Xu, Z. Lin, and H. Zha. A unified convex surrogate for the Schatten-p norm. In *AAAI*, 2017.
68. B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.
69. S. Negahban and M. Wainwright. Estimation of (near) low-rank maatrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):109–1097, 2011.
70. A. Waters, A. Sankaranarayanan, and R. Baraniuk. SpaRCS: Recovering low rank and sparse matrices from compressive measurements. In *NIPS*, 2011.