Alternating Minimizations Converge to Second-Order Optimal Solutions

Qiuwei Li [*1] **Zhihui Zhu** [*2] **Gongguo Tang** [1]

Abstract

This work studies the second-order convergence for both standard alternating minimization and proximal alternating minimization. We show that under mild assumptions on the (nonconvex) objective function, both algorithms avoid strict saddles almost surely from random initialization. Together with known first-order convergence results, this implies both algorithms converge to a second-order stationary point. This solves an open problem for the second-order convergence of alternating minimization algorithms that have been widely used in practice to solve large-scale nonconvex problems due to their simple implementation, fast convergence, and superb empirical performance.

1. Introduction

We consider the following optimization problem over two sets of variables:

$$\underset{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{y}\in\mathbb{R}^m}{\text{minimize}}\ f(\boldsymbol{x},\boldsymbol{y}), \tag{1}$$

where $f:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ is a continuous (nonconvex) function and the partition of variables into $\boldsymbol{x}$ and $\boldsymbol{y}$ blocks typically reflect natural structures within the problem.

One approach to solve (1) is by concatenating $\boldsymbol{x}$ and $\boldsymbol{y}$ as a single variable $\boldsymbol{z}=(\boldsymbol{x},\boldsymbol{y})$ and then directly applying standard iterative algorithms like gradient descent (or its variants) for $f(\boldsymbol{z})$. Recent progress in nonconvex optimization has provided solid theoretical guarantees for gradient-type algorithms in solving nonconvex problems. In particular, the seminal work (Lee et al., 2016) shows that gradient descent with random initialization almost surely converges to a second-order stationary point. Meanwhile, recent results in landscape analysis show that many popular nonconvex optimization problems enjoy a nice landscape where all second-order stationary points are global minima, including low-rank matrix recovery (Ge et al., 2016; 2017; Bhojanapalli et al., 2016; Park et al., 2017; Li et al., 2016; 2018; Zhu et al., 2018; 2017), phase retrieval (Sun et al., 2018), dictionary learning (Sun et al., 2017), blind deconvolution (Zhang et al., 2017), and tensor decomposition (Ge et al., 2015); see (Jain et al., 2017; Chen & Chi, 2018; Chi et al., 2018) for an overview. This implies gradient-type algorithms can find a global minimum for many popular nonconvex problems.

An alternative approach to solve (1) is alternating minimization (cf. Algorithms 1; a.k.a. nonlinear Gauss-Seidel method or block-coordinate descent), which sequentially optimizes over one variable while fixes the other. Compared with gradient-type algorithms, alternating minimization has several advantages: (i) it is easy to implement as there is no need to tune optimization parameters like step sizes, (ii) it converges very fast in practice, and (iii) the subproblems are easy to solve as they usually have closed-form solutions. Thus, alternating minimization has been widely used in practice (Wang et al., 2008; Comon et al., 2009; Jain et al., 2013; Netrapalli et al., 2013; Hastie et al., 2015; Lu et al., 2019).

Algorithm 1 Alternating Minimization

1: **Initialization:** $\boldsymbol{x}_0$.
2: **For** $k=1,2,\ldots$, recursively generate $(\boldsymbol{x}_k,\boldsymbol{y}_k)$ by

$$\begin{aligned}\boldsymbol{y}_k&=\arg\min_{\boldsymbol{y}\in\mathbb{R}^m} f(\boldsymbol{x}_{k-1},\boldsymbol{y}),\\ \boldsymbol{x}_k&=\arg\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x},\boldsymbol{y}_k).\end{aligned} \tag{2}$$

Algorithm 2 Proximal Alternating Minimization

1: **Input:** $\beta>L_f$.
2: **Initialization:** $(\boldsymbol{x}_0,\boldsymbol{y}_0)$.
3: **For** $k=1,2,\ldots$, recursively generate $(\boldsymbol{x}_k,\boldsymbol{y}_k)$ by

$$\begin{aligned}\boldsymbol{y}_k&=\arg\min_{\boldsymbol{x}} f(\boldsymbol{x}_{k-1},\boldsymbol{y})+\frac{\beta}{2}\|\boldsymbol{y}-\boldsymbol{y}_{k-1}\|_2^2,\\ \boldsymbol{x}_k&=\arg\min_{\boldsymbol{y}} f(\boldsymbol{x},\boldsymbol{y}_k)+\frac{\beta}{2}\|\boldsymbol{x}-\boldsymbol{x}_{k-1}\|_2^2.\end{aligned} \tag{3}$$

[*]Equal contribution [1]Department of Electrical Engineering, Colorado School of Mines [2]Mathematical Institute for Data Science, Johns Hopkins University. Correspondence to: Gongguo Tang <gtang@mines.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

However, the empirical performance of alternating minimization is not sufficiently substantiated by solid convergence guarantees. In fact, although the idea of alternatingly updating the variables is quite straightforward, the convergence properties for alternating minimization are far more complicated. In particular, alternating minimization may not converge to first-order stationary points of the problem (Grippo & Sciandrone, 2000). When the function f is strongly bi-convex and satisfies the Kurdyka-Lojasiewicz (KL) property, (Xu & Yin, 2013) shows that alternating minimization converges to a first-order stationary point of f. The KL property is satisfied by a wide class of nonconvex (and even nonsmooth) functions, including all semi-algebraic functions and sub-analytic functions (Attouch et al., 2010). To relax the bi-convexity condition, Attouch et al. (Attouch et al., 2010) utilized a proximal method when updating each variable and proved that the corresponding proximal alternating minimization (cf. Algorithm 2) converges to a first-order stationary point of f under even mild conditions. We now summarize these results as follows.

Assumption 1. *f satisfies the KL property and ∇f is Lipschitz continuous on any bounded subset of domain $\mathbb{R}^n \times \mathbb{R}^m$.*

Theorem 1 (First-order Convergence, (Xu & Yin, 2013; Attouch et al., 2010))**.** *Under Assumption 1, let $(\boldsymbol{x}_0, \boldsymbol{y}_0)$ be any initialization and $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ be the sequence generated by Algorithm 1 (if f is further bi-convex) or by Algorithm 2. If the sequence $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ is bounded, then it converges to a first-order stationary point of f.*

The first-order convergence alone is not sufficient to explain the successful practical performance of alternating minimizations for a considerable body of machine learning problems, as it cannot exclude the case of getting stuck at saddle points, but showing the second-order convergence of the alternating minimizations remains open. The main contribution of this work is closing this gap between the power of alternating minimizations in solving nonconvex problems and its second-order convergence. More precisely, we study the second-order convergence of alternating minimizations by answering the following question:

> **Question**: Does (proximal) alternating minimization almost surely converge to a second-order optimal solution from random initialization?

We answer this question affirmatively for real analytic functions and establish the following main results on the second-order convergence of alternating minimizations:

Theorem 2 (Second-order convergence)**.** *Under Assumption 1, let $(\boldsymbol{x}_0, \boldsymbol{y}_0)$ be a random initialization and $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ be the sequence generated by Algorithm 1 (if f is further analytic and bi-convex with full-rank cross Hessian at strict saddles) or by Algorithm 2 (if f is further bi-smooth). If the sequence $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ is bounded[1], then it converges to a second-order stationary point of f almost surely.*

If additionally, the objective function of the problem satisfies the strict saddle property (i.e., a stationary point is either a strict saddle or a local minimum), then Theorem 2 implies that alternating minimizations with random initialization converge to local minima with probability one. Moreover, many popular machine learning and signal processing problems have no spurious local minimum and thus alternating minimizations converge to a global minimum, partially explaining the good empirical performance of alternating minimizations in achieving global optimality for these problems.

2. Preliminary

Definition 1. *Let f be a twice continuously differentiable function and ∇ be the gradient operator. Then we say*

1. *$\boldsymbol{x}$ is a (first-order) stationary point (a.k.a. critical point) of f, if $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$;*
2. *$\boldsymbol{x}$ is a second-order stationary point of f, if it is a stationary point of f and $\nabla^2 f(\boldsymbol{x}) \succeq 0$;*
3. *$\boldsymbol{x}$ is a strict saddle of f, if it is a first-order stationary point but not a second-order stationary point of f, i.e., $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ and $\nabla^2 f(\boldsymbol{x})$ has a negative eigenvalue.*

Definition 1 (Unstable Fixed Point)**.** *For a mapping $g : \Omega \to \Omega$, the set of unstable fixed points is defined as*

$$\mathcal{A}_g = \{\boldsymbol{x} : g(\boldsymbol{x}) = \boldsymbol{x}, \max_i |\lambda_i(Dg(\boldsymbol{x}))| > 1\},$$

where D denotes the Jacobian operator.

Theorem 3 (Theorem 2 of (Lee et al., 2019))**.** *Let g be a $\mathcal{C}^1$ mapping from Ω to Ω and $\det(Dg(\boldsymbol{x})) \neq 0$ for all $\boldsymbol{x} \in \Omega$. Then the set of initial points that converge to unstable fixed points has measure zero, $\mu(\{\boldsymbol{x}_0 : \lim_{k\to\infty} g^k(\boldsymbol{x}_0) \in \mathcal{A}_g\}) = 0$. Here $\mu(\cdot)$ counts the Lebesgue measure.*

Theorem 3 is instrumental in establishing second-order convergence guarantees for many first-order algorithms (cf. (Lee et al., 2019)). However, the condition that $\det(Dg(\boldsymbol{x})) \neq 0$ for all $\boldsymbol{x} \in \Omega$ is a strong global condition that is difficult to satisfy and is challenging to verify theoretically. The rest of this section focuses on relaxing this global condition in Theorem 3 to a local one so that it can be applied to a larger class of mappings. More precisely, we will replace the global non-singularity condition on the whole domain by a local non-singularity condition around

[1]Note that the boundedness assumption is automatically satisfied when f is coercive, cf. (Xu & Yin, 2013; Attouch et al., 2010).

the stationary points. This is achieved by refining the arguments used in (Lee et al., 2019) and the main technical tools are the Zero-Property Theorem and the Maximum Rank Theorem.

Theorem 4 (Zero-Property Theorem, Theorem 3 of (Ponomarev, 1987))**.** *Let a mapping $g : \Omega \to \Omega$ is continuous and almost everywhere differentiable. Then g satisfies the zero-property (i.e., preimage of any zero-measure set has measure zero) if and only if $\operatorname{rank}(Dg(\boldsymbol{x})) = \dim(\Omega)$ for almost all $\boldsymbol{x} \in \Omega$.*

Theorem 5 (Maximum Rank Theorem, Proposition B.4 of (Bamber & Van Santen, 1985))**.** *Suppose $g : \Omega \to \Omega$ is an analytic mapping. $Dg(\boldsymbol{x})$ achieves the maximum rank almost everywhere in Ω. Here the maximum rank is defined as $\max_{\boldsymbol{x}\in\Omega} \operatorname{rank}(Dg(\boldsymbol{x}))$.*

Note the analytic assumption of Theorem 5 is stronger than infinite differentiability, but still covers a fairly large class of functions, including all elementary functions, most special functions, as well as their combinations and compositions. The Maximum Rank Theorem states that the Jacobian of any analytic mapping almost always achieves the maximum rank. Then as long as the Jacobian is of full-rank at some specific point, the mapping would satisfy the zero-property, which is indicated by Theorem 4. Now we present the main technical theorem.

Theorem 6. *Let g be an analytic mapping from Ω to Ω. Then the set of initial points that converge to nondegenerate unstable fixed points has measure zero.*

The proof is adapted from (Lee et al., 2019) and therefore the most important ingredient is the Stable Manifold Theorem Theorem III.7 (Shub, 2013).

Theorem 7 (Stable Manifold Theorem, Theorem III.7 of (Shub, 2013))**.** *Let $\boldsymbol{x}^\star$ be a fixed point for a C^r local diffeomorphism $g : U \to E$, where U is a neighborhood of $\boldsymbol{x}^\star$ in the Banach space E. Suppose that $E = E_s \oplus E_u$, where E_s is the span of the eigenvectors of $Dg(\boldsymbol{x}^\star)$ corresponding to eigenvalues of magnitude smaller than or equal to 1, and E_u is the span of the eigenvectors of $Dg(\boldsymbol{x}^\star)$ corresponding to eigenvalues of magnitude larger than 1. Then there exists a C^r embedded disk W^{cs}_{loc} that is tangent to E_s at $\boldsymbol{x}^\star$ called the local stable center manifold. Moreover, there exists a neighborhood $B_{\boldsymbol{x}^\star}$ of $\boldsymbol{x}^\star$, such that $g(W^{cs}_{loc}) \cap B_{\boldsymbol{x}^\star} \subset W^{cs}_{loc}$ and $\bigcap_{k=0}^{\infty} g^{-k}(B_{\boldsymbol{x}^\star}) \subset W^{cs}_{loc}$.*

Proof of Theorem 6. First, for any unstable fixed point $\boldsymbol{x}^\star \in \mathcal{A}_g$, if it is also non-degenerate, i.e., the Jacobian $Dg(\boldsymbol{x}^\star)$ is non-singular, then $Dg(\boldsymbol{x})$ is nonsingular in some neighborhood U of $\boldsymbol{x}^\star$. This shows $g : U \to g(U)$ is a local diffeomorphism. Then by Stable Manifold Theorem 7, for any $\boldsymbol{x}^\star \in \mathcal{A}_g$, there is an associated open neighborhood $B_{\boldsymbol{x}^\star}$ and thus the union $\bigcup_{\boldsymbol{x}^\star\in\mathcal{A}_g} B_{\boldsymbol{x}^\star}$ forms an open cover for $\mathcal{A}_g$. Clearly $\mathcal{A}_g \subset \mathbb{R}^n$, and since $\mathbb{R}^n$ is known to be second-countable (cf. (Lee et al., 2019)), we can extract a countable subcover $\bigcup_{i=1}^{\infty} B_{\boldsymbol{x}_i^\star}$ for $\mathcal{A}_g$. Let

$$W \doteq \{\boldsymbol{x}_0 \in \Omega : \lim_k g^k(\boldsymbol{x}_0) \in \mathcal{A}_g\}.$$

Because $\bigcup_{i=1}^{\infty} B_{\boldsymbol{x}_i^\star}$ forms a countable subcover of $\mathcal{A}_g$, $\boldsymbol{x}^\star \in B_{\boldsymbol{x}_i^\star}$ for some i, i.e., $\lim_{t\to\infty} g^t(\boldsymbol{x}_0) \in B_{\boldsymbol{x}_i^\star}$. That is to say, $g^t(\boldsymbol{x}_0) \in B_{\boldsymbol{x}_i^\star}$ for all $t \geq N$ for some sufficiently large N, or equivalently,

$$g^t(\boldsymbol{x}_0) \in \bigcap_{k=0}^{\infty} g^{-k}(B_{\boldsymbol{x}_i^\star}) \doteq S_i, \quad \forall t \geq N.$$

By Stable Manifold Theorem 7, we have $S_i \subset W^{cs}_{loc}$ with W^{cs}_{loc} of co-dimension at least one (since $\boldsymbol{x}^\star \in \mathcal{A}_g$). Therefore, S_i has measure zero. Since $g^N(\boldsymbol{x}_0) \in S_i$ with an unknown non-negative integer N and $\boldsymbol{x}_0$ is an arbitrary element in W, we must have

$$W \subset \bigcup_{i=1}^{\infty} \bigcup_{N=0}^{\infty} g^{-N}(S_i).$$

Now we show $g^{-N}(S_i)$ has measure zero for any non-negative numbers N and i. Then the proof follows from that any countable union of zero-measure sets has measure zero. Since g is analytic and $\boldsymbol{x}^\star$ is nondegenerate, i.e., $\operatorname{rank}(Dg(\boldsymbol{x}^\star)) = n$, which must be the maximum rank of the Jacobian $Dg(\boldsymbol{x})$ in Ω. Then Theorem 5 implies that the Jacobian $Dg(\boldsymbol{x})$ achieves the maximum rank n for almost all $\boldsymbol{x} \in \Omega$. Further because g is analytic (and hence continuous and almost everywhere differentiable), we can use the Zero-Property Theorem 4 to get $g^{-N}(S_i)$ has measure zero for all $N \geq 0$. Finally note that the above argument is independent of choice of i. □

3. Second-order Convergence of Alternating Minimization

For this case when f is strongly bi-convex, we will apply Theorem 6 to show that alternating minimization (cf. Algorithm 1) will not converge to a strict saddle point. Then combining this with the first-order convergence result (cf. Theorem 1), we can get the second-order convergence of the alternating minimization. We first provide some additional assumptions that are used to prove the avoiding-saddle property of alternating minimization in solving problem (1).

Assumption 2. *f is a strongly bi-convex[2] analytic function.*

Assumption 3. *$\nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ has full row rank for all strict saddles $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$.*

Theorem 8 (Avoiding Strict Saddles)**.** *Suppose f satisfies Assumptions 2 and 3. Then solving (1) using Algorithm 1*

[2] $\nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) \succ 0$ and $\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \succ 0$ in the whole domain.

with random initialization will not converge to a strict saddle of f almost surely.

Therefore, together with the first-order convergence Theorem 1 and noting that any analytic function satisfies Assumption 1, we have the second-order convergence property of the alternating minimization.

Corollary 1. *Suppose f satisfies Assumptions 2 and 3 and the sequence $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ generated by Algorithm 1 is bounded. Then solving (1) using Algorithm 1 with random initialization will converge to a second-order stationary point of f almost surely.*

3.1. The Mapping Function

First note that the alternating minimization (cf. Algorithm 1) is well defined under the strong bi-convexity condition in Assumption 2, since each subproblem minimizes a strongly convex function and thus has a unique optimal solution.

Proposition 1. *Under Assumption 2, the following two mappings are well-defined in the whole domain:*

$$\begin{aligned} \boldsymbol{\phi}(\boldsymbol{x}) &\doteq \arg\min_{\boldsymbol{y}\in\mathbb{R}^m} f(\boldsymbol{x}, \boldsymbol{y}), \\ \boldsymbol{\psi}(\boldsymbol{y}) &\doteq \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}, \boldsymbol{y}). \end{aligned} \tag{4}$$

Proposition 1 immediately implies that Algorithm 1 is well-defined. That is, each subproblem in the k-th iteration has a unique minimizer:

$$\begin{aligned} \boldsymbol{y}_{k+1} &= \boldsymbol{\phi}(\boldsymbol{x}_k), \\ \boldsymbol{x}_{k+1} &= \boldsymbol{\psi}(\boldsymbol{y}_{k+1}). \end{aligned}$$

By defining the composition $g = \boldsymbol{\psi} \circ \boldsymbol{\phi}$ from $\mathbb{R}^n$ to $\mathbb{R}^n$, we can view the alternating minimization process (2) as iteratively performing the following mapping:

$$\boldsymbol{x}^k = g(\boldsymbol{x}^{k-1}) = g^k(\boldsymbol{x}_0) \quad \text{for } k = 1, 2, \ldots \tag{5}$$

By the first-order convergence of the alternating minimization, the iterative process (5) is continuing until reaching a fixed point $\boldsymbol{x}^\star$ of the mapping g

$$\boldsymbol{x}^\star = g(\boldsymbol{x}^\star). \tag{6}$$

In view of (4), this is equivalent to

$$\begin{aligned} \boldsymbol{y}^\star &= \arg\min_{\boldsymbol{y}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star), \\ \boldsymbol{x}^\star &= \arg\min_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \end{aligned}$$

with $\boldsymbol{y}^\star \doteq \boldsymbol{\phi}(\boldsymbol{x}^\star)$. Then together with the strong bi-convexity and the sufficient differentiability (by analytic property) of f, we immediately have that there is a one-to-one correspondence between the fixed points of g and the first-order stationary points of f.

Lemma 1. *A point $\boldsymbol{x}^\star$ is a fixed point of g if and only if*

$$\nabla f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) = \boldsymbol{0}, \tag{7}$$

where we have defined $\boldsymbol{y}^\star = \boldsymbol{\phi}(\boldsymbol{x}^\star)$ and $\nabla f(\boldsymbol{x}, \boldsymbol{y}) = [\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})^{\mathrm{T}} \ \ \nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})^{\mathrm{T}}]^{\mathrm{T}}$. For simplifying notations, we will also often informally write $\nabla f(\boldsymbol{x}, \boldsymbol{y}) = (\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), \nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}))$.

3.2. Proof of Theorem 8

According to Theorem 6, it is sufficient to show that 1) the mapping g is analytic; 2) all strict saddles of f correspond to unstable fixed points of g; 3) the Jacobian Dg at any strict saddle is full rank. Without loss of generality, we also assume $n \leq m$. This assumption can always be satisfied since otherwise, we can exchange the coordinates of f. We will see this assumption helps to show the non-degenerate property at unstable fixed points of g.

(1) Showing analytic mapping Towards that end, we now derive the closed-form expression of the Jacobian Dg which will also be useful for the remaining proof. To begin, we present an immediate consequence of Proposition 1.

Proposition 2. *There exist two well-defined and unique mappings $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{\psi} : \mathbb{R}^m \to \mathbb{R}^n$ such that*

$$\begin{aligned} \nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x})) &= \boldsymbol{0}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n, \\ \nabla_{\boldsymbol{x}} f(\boldsymbol{\psi}(\boldsymbol{y}), \boldsymbol{y}) &= \boldsymbol{0}, \quad \forall \boldsymbol{y} \in \mathbb{R}^m. \end{aligned} \tag{8}$$

Then we use the Analytic Implicit Function Theorem 9.

Theorem 9 (Analytic Implicit Function Theorem, Theorem 7.6 of (Fritzsche & Grauert, 2012))**.** *Let the function $h(\boldsymbol{x}, \boldsymbol{y}) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be analytic. Assume $h(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{0}_m$ for some point $(\boldsymbol{a}, \boldsymbol{b}) \in \mathbb{R}^n \times \mathbb{R}^m$. If the partial Jacobian $D_{\boldsymbol{y}} h(\boldsymbol{a}, \boldsymbol{b})$ is invertible, then there exists an open set U of $\mathbb{R}^n$ containing $\boldsymbol{a}$ such that there exists a unique analytic function $\boldsymbol{\phi} : U \to \mathbb{R}^m$ such that*

$$\boldsymbol{\phi}(\boldsymbol{a}) = \boldsymbol{b}$$

and

$$h(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x})) = \boldsymbol{0}_m \quad \text{for all } \boldsymbol{x} \in U.$$

Moreover, the Jacobian of $\boldsymbol{\phi}$ in U is given by

$$D\boldsymbol{\phi}(\boldsymbol{x}) = -D_{\boldsymbol{y}} h(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x}))^{-1} D_{\boldsymbol{x}} h(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x})).$$

We now prove g is analytic.

Lemma 2. *The mapping g is analytic and its Jacobian $Dg(\boldsymbol{x})$ is given by*

$$\begin{aligned} Dg(\boldsymbol{x}) = &\nabla^2_{\boldsymbol{x}} f(g(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}))^{-1} \nabla^2_{\boldsymbol{xy}} f(g(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x})) \times \\ &\nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x}))^{-1} \nabla^2_{\boldsymbol{yx}} f(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x})). \end{aligned} \tag{9}$$

Proof of Lemma 2. From Corollary 2, we know that there have been two well-defined and unique mappings that satisfy (8):

$$\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x})) = \mathbf{0}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n,$$
$$\nabla_{\boldsymbol{x}} f(\boldsymbol{\psi}(\boldsymbol{y}), \boldsymbol{y}) = \mathbf{0}, \quad \forall \boldsymbol{y} \in \mathbb{R}^m.$$

Now denote $h_{\boldsymbol{y}} = \nabla_{\boldsymbol{y}} f$ and $h_{\boldsymbol{x}} = \nabla_{\boldsymbol{x}} f$, which are both analytic as f is analytic. Then the above equations read that

$$\begin{aligned} h_{\boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x})) &= \mathbf{0}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n, \\ h_{\boldsymbol{x}}(\boldsymbol{\psi}(\boldsymbol{y}), \boldsymbol{y}) &= \mathbf{0}, \quad \forall \boldsymbol{y} \in \mathbb{R}^m. \end{aligned} \tag{10}$$

Further note that both $D_{\boldsymbol{y}} h_{\boldsymbol{y}} = \nabla^2_{\boldsymbol{y}} f$ and $D_{\boldsymbol{x}} h_{\boldsymbol{x}} = \nabla^2_{\boldsymbol{x}} f$ are both nonsingular by assumption of strong bi-convexity. Then we can apply Analytic Implicit Function Theorem 9 to (10) to get that $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are the unique analytic mappings satisfying (10). Further, using Analytic Implicit Function Theorem 9, we can compute their Jacobians as

$$D\boldsymbol{\phi}(\boldsymbol{x}) = -\nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x}))^{-1} \nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x})),$$
$$D\boldsymbol{\psi}(\boldsymbol{y}) = -\nabla^2_{\boldsymbol{x}} f(\boldsymbol{\psi}(\boldsymbol{y}), \boldsymbol{y})^{-1} \nabla^2_{\boldsymbol{yx}} f(\boldsymbol{\psi}(\boldsymbol{y}), \boldsymbol{y}).$$

Therefore, $g = \boldsymbol{\psi} \circ \boldsymbol{\phi}$ is analytic, as it is a composition of two analytic mappings $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$. Also, the Jacobian Dg is given by the chain rule:

$$\begin{aligned} Dg(\boldsymbol{x}) =& D\boldsymbol{\psi}(\boldsymbol{\phi}(\boldsymbol{x})) \times D\boldsymbol{\phi}(\boldsymbol{x}) \\ =& \nabla^2_{\boldsymbol{x}} f(g(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}))^{-1} \nabla^2_{\boldsymbol{xy}} f(g(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x})) \times \\ & \nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x}))^{-1} \nabla^2_{\boldsymbol{yx}} f(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x})). \end{aligned}$$

□

(2) Showing unstable fixed point First of all, by (8), for any strict saddle $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ of f, $\boldsymbol{x}^\star = g(\boldsymbol{x}^\star)$, i.e., $\boldsymbol{x}^\star$ is a fixed point of g. It remains to show that the maximal magnitude of the eigenvalues of $Dg(\boldsymbol{x}^\star)$ is greater than 1. Using the fixed point equation $\boldsymbol{x}^\star = g(\boldsymbol{x}^\star)$, we can simplify the Jacobian expression (9) as

$$\begin{aligned} Dg(\boldsymbol{x}^\star) =& \nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{-1} \nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \times \\ & \nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{-1} \nabla^2_{\boldsymbol{yx}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star). \end{aligned} \tag{11}$$

Define a new matrix that is similar to $Dg(\boldsymbol{x}^\star)$:

$$\boldsymbol{\Gamma} \doteq \nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{1/2} Dg(\boldsymbol{x}^\star) \nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{-1/2}.$$

Hence by matrix similarity, they have the same eigenvalues. Plugging $Dg(\boldsymbol{x}^\star)$ into $\boldsymbol{\Gamma}$, we have

$$\begin{aligned} \boldsymbol{\Gamma} =& (\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{-\frac{1}{2}} \nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{-\frac{1}{2}}) \\ & \times (\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{-\frac{1}{2}} \nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{-\frac{1}{2}})^{\mathrm{T}} \\ =& \mathbf{L}\mathbf{L}^{\mathrm{T}}, \end{aligned}$$

where

$$\mathbf{L} \doteq \nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{-\frac{1}{2}} \nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{-\frac{1}{2}}. \tag{12}$$

Therefore, it reduces to show that $\boldsymbol{\Gamma} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ has at least an eigenvalue of magnitude greater than 1, since this can imply $Dg(\boldsymbol{x}^\star)$ has at least an eigenvalue of magnitude greater than 1. Note that $\boldsymbol{\Gamma} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ has at least an eigenvalue of magnitude greater than 1 if and only if the spectral norm of $\|\mathbf{L}\| > 1$.

Now we prove $\|\mathbf{L}\| > 1$ via contradiction. For the sake of contradiction, suppose $\|\mathbf{L}\| \leq 1$. We can represent Hessian $\nabla^2 f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ (which is known to have a negative eigenvalue since $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ is a strict saddle of f) as

$$\begin{aligned} & \nabla^2 f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \\ =& \begin{bmatrix} \nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) & \nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \\ \nabla^2_{\boldsymbol{yx}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) & \nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \end{bmatrix} \\ =& \begin{bmatrix} \nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{1/2} & \\ & \nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & \mathbf{L} \\ \mathbf{L}^{\mathrm{T}} & \mathbf{I}_m \end{bmatrix} \\ & \begin{bmatrix} \nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{1/2} & \\ & \nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{1/2} \end{bmatrix}. \end{aligned}$$

Further note that $\begin{bmatrix} \mathbf{I}_n & \mathbf{L} \\ \mathbf{L}^{\mathrm{T}} & \mathbf{I}_m \end{bmatrix}$ is semi-positive definite, since

$$\begin{aligned} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{I}_n & \mathbf{L} \\ \mathbf{L}^{\mathrm{T}} & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} &= \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 + 2\boldsymbol{x}^{\mathrm{T}} \mathbf{L} \boldsymbol{y} \\ &\geq \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2\|\boldsymbol{x}\|_2 \|\mathbf{L}\| \|\boldsymbol{y}\|_2 \\ &\geq \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \geq 0, \end{aligned}$$

which holds for all $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{y} \in \mathbb{R}^m$. Consequently, $\nabla^2 f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ is semi-positive definite, leading to a contradiction. Therefore, we have proved that for any strict saddle $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$, $\boldsymbol{x}^\star$ is an unstable fixed point of the mapping g.

(3) Showing non-degenerate property First recall that the Jacobian $Dg(\boldsymbol{x}^\star)$ at any strict saddles point $\boldsymbol{x}^\star$ is given by (11). Due to the strict positive-definiteness of $\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ and $\nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$, we know $Dg(\boldsymbol{x}^\star)$ is similar to a semi-positive definite matrix:

$$Dg(\boldsymbol{x}^\star) = \nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{1/2} \mathbf{L}\mathbf{L}^{\mathrm{T}} \nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)^{-1/2}$$

where $\mathbf{L} \in \mathbb{R}^{n \times m}$ is defined in (12). Therefore, the non-degenerateness follows from Assumption 3 and that $n \leq m$.

Combining all, we complete the proof of Theorem 8. □

3.3. Stylized Application of Algorithm 1

We use a simple example to illustrate our result.

Example 1 (Best Rank-1 Matrix PCA)**.** *Consider the problem of computing the best rank-1 approximation of a given*

matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ *with* $\operatorname{rank}(\mathbf{A}) = n$*:*

$$f(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}\|\mathbf{A} - \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\|_F^2 + \frac{\lambda}{2}(\|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2), \quad (13)$$

which is an analytic, strongly bi-convex function (cf. Assumption 2). Note that there are efficient closed-form solutions when using the alternating minimization (cf. Algorithm 1) to solve (13)*: given any initialization* $\boldsymbol{x}_0 \in \mathbb{R}^n$*, alternating minimization recursively generates the following sequence: for* $k = 0, 1, 2, \ldots$

$$\begin{aligned}\boldsymbol{y}_{k+1} &\doteq \boldsymbol{\phi}(\boldsymbol{x}_k) = \mathbf{A}^{\mathrm{T}}\boldsymbol{x}_k / \left(\lambda + \|\boldsymbol{x}_k\|_2^2\right),\\ \boldsymbol{x}_{k+1} &\doteq \boldsymbol{\psi}(\boldsymbol{y}_{k+1}) = \mathbf{A}\boldsymbol{y}_{k+1} / \left(\lambda + \|\boldsymbol{y}_{k+1}\|_2^2\right).\end{aligned}$$

To apply Corollary 1, one still needs to verify the full-rankness of $\nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ at any strict saddle $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ of f, where $\boldsymbol{y}^\star = \boldsymbol{\phi}(\boldsymbol{x}^\star)$. Direct computations give that

$$\nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) = 2\boldsymbol{x}^\star \boldsymbol{\phi}(\boldsymbol{x}^\star)^{\mathrm{T}} - \mathbf{A} = \left(2\frac{\boldsymbol{x}^\star \boldsymbol{x}^{\star\mathrm{T}}}{\lambda + \|\boldsymbol{x}^\star\|_2^2} - \mathbf{I}\right)\mathbf{A}.$$

Clearly, when $\boldsymbol{x}^\star = \mathbf{0}$, we have $\nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) = -\mathbf{A}$ and the full-rankness assumption automatically holds and for $\boldsymbol{x}^\star \neq \mathbf{0}$, $\operatorname{rank}(\nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)) = \operatorname{rank}(\mathbf{A})$ provided $\lambda \neq \|\boldsymbol{x}^\star\|_2^2$. Therefore:

Corollary 2. *Assume* $\mathbf{A}$ *is nonsingular in* (13)*. Then Alternating Minimization Algorithm 1 from random initialization almost surely converges to a second-order stationary point, provided* $\lambda \neq \|\boldsymbol{x}^\star\|_2^2$*.*

4. Second-order Convergence of Proximal Alternating Minimization

We begin with the following bi-smoothness assumption.

Assumption 4. *$f \in \mathcal{C}^2$ is L_f bi-smooth in the domain, i.e.,* $\max\{\|\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{y})\|, \|\nabla^2_{\boldsymbol{y}} f(\boldsymbol{x},\boldsymbol{y})\|\} \leq L_f$ *in the domain.*[3]

In the case where $f(\boldsymbol{x},\boldsymbol{y})$ is L_f bi-smooth, we note that Algorithm 2 requires even minor assumptions for it to avoid the strict saddle points.

Theorem 10 (Avoiding Strict Saddles)**.** *Suppose f satisfies Assumption 4. Choose $\beta > L_f$ in Algorithm 2. Then solving* (1) *using Algorithm 2 with random initialization will not converge to a strict saddle of f almost surely.*

Therefore, together with the first-order convergence Theorem 1, we have the second-order convergence property of Algorithm 2.

Corollary 3. *Suppose f satisfies Assumptions 1 and 4 and the sequence $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ generated by Algorithm 2 is bounded. Choose $\beta > L_f$ in Algorithm 2. Then solving* (1) *using Algorithm 2 with random initialization will return a second-order stationary point of f for almost sure.*

[3] Any globally smooth function f with $\|\nabla^2 f(\boldsymbol{x},\boldsymbol{y})\| \leq L_f$ satisfies Assumption 4.

4.1. The Mapping Function

First from (3), we know under the assumptions of $\beta > L_f$ and the L_f bi-smoothness of f, then each subproblem in any iteration of Algorithm 2 is well-defined, since the objective function of each subproblem is strongly convex.

Proposition 3. *Under Assumption 4, choose $\beta > L_f$. Then the following two mappings are analytic and well-defined for any $(\boldsymbol{x},\boldsymbol{y})$:*

$$\begin{aligned}\boldsymbol{p}_\beta(\boldsymbol{x},\boldsymbol{y}) &\doteq \operatorname*{arg\,min}_{\boldsymbol{y}'\in\mathbb{R}^m} f(\boldsymbol{x},\boldsymbol{y}') + \frac{\beta}{2}\|\boldsymbol{y}' - \boldsymbol{y}\|_2^2,\\ \boldsymbol{q}_\beta(\boldsymbol{x},\boldsymbol{y}) &\doteq \operatorname*{arg\,min}_{\boldsymbol{x}'\in\mathbb{R}^n} f(\boldsymbol{x}',\boldsymbol{y}) + \frac{\beta}{2}\|\boldsymbol{x}' - \boldsymbol{x}\|_2^2.\end{aligned} \quad (14)$$

With (14), each iteration of Algorithm 2 is equivalent to

$$\begin{aligned}\boldsymbol{y}_k &= \boldsymbol{p}_\beta(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1}),\\ \boldsymbol{x}_k &= \boldsymbol{q}_\beta(\boldsymbol{x}_{k-1}, \boldsymbol{y}_k).\end{aligned} \quad (15)$$

We define a mapping $g_\beta : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^m$ such that

$$g_\beta(\boldsymbol{x},\boldsymbol{y}) = \left(\boldsymbol{q}_\beta(\boldsymbol{x}, \boldsymbol{p}_\beta(\boldsymbol{x},\boldsymbol{y})), \boldsymbol{p}_\beta(\boldsymbol{x},\boldsymbol{y})\right), \quad (16)$$

with which we can rewrite (15) as

$$\begin{aligned}(\boldsymbol{x}_k, \boldsymbol{y}_k) &= \left(\boldsymbol{q}_\beta(\boldsymbol{x}_{k-1}, \boldsymbol{p}_\beta(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1})), \boldsymbol{p}_\beta(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1})\right)\\ &= g_\beta(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1}).\end{aligned}$$

With the implicit function theorem, the following result establishes the expression of the Jacobian of g_β.

Lemma 3. *For any $(\boldsymbol{x},\boldsymbol{y})$, denote $(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) = g_\beta(\boldsymbol{x},\boldsymbol{y})$, and assume* $\max\{\|\nabla^2_{\boldsymbol{x}} f(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}})\|, \|\nabla^2_{\boldsymbol{y}} f(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}})\|\} \leq L_f$*. Set $\beta > L_f$ in Algorithm 2. Then the mapping function g_β is continuous at a neighborhood of $(\boldsymbol{x},\boldsymbol{y})$ and the Jacobian Dg_β is nonsingular at $(\boldsymbol{x},\boldsymbol{y})$ and is given by*

$$Dg_\beta(\boldsymbol{x},\boldsymbol{y}) = \begin{bmatrix}\nabla^2_{\boldsymbol{x}} f(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) + \beta\mathbf{I}_n & \nabla^2_{\boldsymbol{xy}} f(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}})\\ \mathbf{0} & \nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}, \widetilde{\boldsymbol{y}}) + \beta\mathbf{I}_m\end{bmatrix}^{-1} \begin{bmatrix}\beta\mathbf{I}_n & \mathbf{0}\\ -\nabla^2_{\boldsymbol{yx}} f(\boldsymbol{x}, \widetilde{\boldsymbol{y}}) & \beta\mathbf{I}_m\end{bmatrix}. \quad (17)$$

Proof. Since $\widetilde{\boldsymbol{y}} = \boldsymbol{p}_\beta(\boldsymbol{x},\boldsymbol{y}), \widetilde{\boldsymbol{x}} = \boldsymbol{q}_\beta(\boldsymbol{x}, \widetilde{\boldsymbol{y}})$, both $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$ can be viewed as functions of $(\boldsymbol{x},\boldsymbol{y})$. Note that $(\boldsymbol{x},\boldsymbol{y})$ and $(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}})$ satisfy the first-order optimality condition of (15):

$$\begin{aligned}\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \widetilde{\boldsymbol{y}}) + \beta(\widetilde{\boldsymbol{y}} - \boldsymbol{y}) &= \mathbf{0},\\ \nabla_{\boldsymbol{x}} f(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) + \beta(\widetilde{\boldsymbol{x}} - \boldsymbol{x}) &= \mathbf{0}.\end{aligned} \quad (18)$$

We now compute the expression of the Jacobian

$$Dg_\beta(\boldsymbol{x},\boldsymbol{y}) = \begin{bmatrix}\frac{\partial\widetilde{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{y})}{\partial\boldsymbol{x}} & \frac{\partial\widetilde{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{y})}{\partial\boldsymbol{y}}\\ \frac{\partial\widetilde{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})}{\partial\boldsymbol{x}} & \frac{\partial\widetilde{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})}{\partial\boldsymbol{y}}\end{bmatrix}.$$

To obtain the expressions for these partial derivatives $\frac{\partial \widetilde{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}}, \frac{\partial \widetilde{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{y}}, \frac{\partial \widetilde{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}}, \frac{\partial \widetilde{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{y}}$, we apply the implicit function theorem to the first-order optimality condition of (18) and obtain

$$
\begin{aligned}
&(\nabla_{\boldsymbol{y}}^2 f(\boldsymbol{x},\widetilde{\boldsymbol{y}})+\beta\mathbf{I}_m)\frac{\partial \widetilde{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}} = -\nabla_{\boldsymbol{yx}}^2 f(\boldsymbol{x},\widetilde{\boldsymbol{y}}),\\
&(\nabla_{\boldsymbol{y}}^2 f(\boldsymbol{x},\widetilde{\boldsymbol{y}})+\beta\mathbf{I}_m)\frac{\partial \widetilde{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{y}} = \beta\mathbf{I}_m,\\
&(\nabla_{\boldsymbol{x}}^2 f(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}})+\beta\mathbf{I}_n)\frac{\partial \widetilde{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}} + \nabla_{\boldsymbol{xy}}^2 f(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}})\frac{\partial \widetilde{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}} = \beta\mathbf{I}_n,\\
&(\nabla_{\boldsymbol{x}}^2 f(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}})+\beta\mathbf{I}_n)\frac{\partial \widetilde{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{y}} + \nabla_{\boldsymbol{xy}}^2 f(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}})\frac{\partial \widetilde{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{y}} = \mathbf{0},
\end{aligned}
$$

which can be rearranged into matrix multiplications as

$$
\begin{aligned}
&\begin{bmatrix} \nabla_{\boldsymbol{x}}^2 f(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}}) + \beta\mathbf{I}_n & \nabla_{\boldsymbol{xy}}^2 f(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}}) \\ \mathbf{0} & \nabla_{\boldsymbol{y}}^2 f(\boldsymbol{x},\widetilde{\boldsymbol{y}}) + \beta\mathbf{I}_m \end{bmatrix} \begin{bmatrix} \frac{\partial \widetilde{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}} & \frac{\partial \widetilde{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{y}} \\ \frac{\partial \widetilde{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}} & \frac{\partial \widetilde{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{y}} \end{bmatrix}\\
&= \begin{bmatrix} \beta\mathbf{I}_n & \mathbf{0} \\ -\nabla_{\boldsymbol{yx}}^2 f(\boldsymbol{x},\widetilde{\boldsymbol{y}}) & \beta\mathbf{I}_m \end{bmatrix} \iff \boldsymbol{\Gamma}_1 D g_\beta(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{\Gamma}_2.
\end{aligned}
$$

We now show that the matrix $\boldsymbol{\Gamma}_1$ is nonsingular. Towards that end, suppose there exists $\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix}$ such that $\boldsymbol{\Gamma}_1 \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$, which is equivalent to

$$
\begin{aligned}
&\begin{bmatrix} (\nabla_{\boldsymbol{x}}^2 f(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}}) + \beta\mathbf{I}_n)\boldsymbol{u} \\ (\nabla_{\boldsymbol{y}}^2 f(\boldsymbol{x},\widetilde{\boldsymbol{y}}) + \beta\mathbf{I}_m)\boldsymbol{v} \end{bmatrix} = \begin{bmatrix} -\nabla_{\boldsymbol{xy}}^2 f(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}})\boldsymbol{v} \\ \mathbf{0} \end{bmatrix}\\
\iff &\begin{bmatrix} (\nabla_{\boldsymbol{x}}^2 f(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}}) + \beta\mathbf{I}_n)\boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \iff \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix},
\end{aligned}
$$

where we have used the strict positive-definiteness of $\nabla_{\boldsymbol{x}}^2 f(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}})+\beta\mathbf{I}_n$ and $\nabla_{\boldsymbol{y}}^2 f(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}})+\beta\mathbf{I}_m$ by the assumption. Thus, the matrix $\boldsymbol{\Gamma}_1$ is nonsingular. Therefore, by the implicit function theorem, $D g_\beta(\boldsymbol{x},\boldsymbol{y})$ is a continuous function at some neighborhood of $\boldsymbol{x},\boldsymbol{y}$. With similar argument, we obtain that the matrix $\boldsymbol{\Gamma}_2$ is also nonsingular. Therefore, we have $D g_\beta(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{\Gamma}_1^{-1}\boldsymbol{\Gamma}_2$ is nonsingular at $\boldsymbol{x},\boldsymbol{y}$. □

4.2. Proof of Theorem 10

By Theorem 3, it suffices to show the following conditions:

(1) g_β is a $\mathcal{C}^1$ mapping. Because $D g_\beta$ is continuous in the whole domain by Lemma 3 and Assumption 4.

(2) $\det(D g_\beta) \neq 0$ in the whole domain. This directly follows from Lemma 3 and Assumption 4.

(3) Any strict saddle of f is an unstable fixed point of g_β. Assume $(\boldsymbol{x}^\star,\boldsymbol{y}^\star)$ is an arbitrary strict saddle point of f. First of all, we show $(\boldsymbol{x}^\star,\boldsymbol{y}^\star)$ is a fixed point of g_β. Since a strict saddle point must be a stationary point, here we show every stationary point of f is a fixed point of g_β. Towards that end, first note that any stationary point $(\boldsymbol{x},\boldsymbol{y})$ satisfies $\nabla f(\boldsymbol{x},\boldsymbol{y}) = (\nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{y}), \nabla_{\boldsymbol{y}} f(\boldsymbol{x},\boldsymbol{y})) = (\mathbf{0},\mathbf{0})$, which implies the first optimality condition (18). Then noting that Proposition 3 which states that the mapping g_β is well-defined in the whole domain, we conclude that $(\boldsymbol{x},\boldsymbol{y}) = g_\beta(\boldsymbol{x},\boldsymbol{y})$, i.e., $(\boldsymbol{x},\boldsymbol{y})$ is a fixed point of g_β.

Now we show that the maximum magnitude of eigenvalues of $D g_\beta(\boldsymbol{x}^\star,\boldsymbol{y}^\star)$ is greater than 1 at any strict saddle $(\boldsymbol{x}^\star,\boldsymbol{y}^\star)$.

Lemma 4. *Let $(\boldsymbol{x}^\star,\boldsymbol{y}^\star)$ be any strict saddle of f with $\max\{\nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x}^\star,\boldsymbol{y}^\star), \nabla_{\boldsymbol{y}}^2 f(\boldsymbol{x}^\star,\boldsymbol{y}^\star)\} \leq L_f$. Set $\beta > L_f$ in Algorithm 2. Then $\lambda_{\max}(D g_\beta(\boldsymbol{x}^\star,\boldsymbol{y}^\star)) > 1$, where $\lambda_{\max}$ denotes the largest eigenvalue.*

Proof. To simplify notations, denote

$$
\begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} \doteq \begin{bmatrix} \nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x}^\star,\boldsymbol{y}^\star) & \nabla_{\boldsymbol{xy}}^2 f(\boldsymbol{x}^\star,\boldsymbol{y}^\star) \\ \nabla_{\boldsymbol{yx}}^2 f(\boldsymbol{x}^\star,\boldsymbol{y}^\star) & \nabla_{\boldsymbol{y}}^2 f(\boldsymbol{x}^\star,\boldsymbol{y}^\star) \end{bmatrix}.
$$

Then plugging $(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}}) = (\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}^\star,\boldsymbol{y}^\star)$ to (17), we can compute the Jacobian $D g_\beta$ at $(\boldsymbol{x}^\star,\boldsymbol{y}^\star)$ as

$$
\begin{aligned}
D g_\beta(\boldsymbol{x}^\star,\boldsymbol{y}^\star) &= \begin{bmatrix} \mathbf{F}_{11}+\beta\mathbf{I}_n & \mathbf{F}_{12} \\ \mathbf{0} & \mathbf{F}_{22}+\beta\mathbf{I}_m \end{bmatrix}^{-1} \begin{bmatrix} \beta\mathbf{I}_n & \mathbf{0} \\ -\mathbf{F}_{21} & \beta\mathbf{I}_m \end{bmatrix}\\
&= \mathbf{I} - \begin{bmatrix} \mathbf{F}_{11}+\beta\mathbf{I}_n & \mathbf{F}_{12} \\ \mathbf{0} & \mathbf{F}_{22}+\beta\mathbf{I}_m \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} \doteq \mathbf{I} - \boldsymbol{\Phi}.
\end{aligned}
$$

Therefore, to show that $D g_\beta(\boldsymbol{x}^\star,\boldsymbol{y}^\star)$ has an eigenvalue larger than 1, it suffices to show $\boldsymbol{\Phi}$ has a negative eigenvalue. We prove this by showing the event that $\det(\boldsymbol{\Phi}+\mu\mathbf{I}) = 0$ for some $\mu > 0$, where $\det(\cdot)$ denotes the determinant of a matrix. Then with some algebra on the properties of determinant, we have $\det(\boldsymbol{\Phi}+\mu\mathbf{I}) = 0$ is equivalent to

$$
\begin{aligned}
&\det\left(\begin{bmatrix} (1+\mu)\mathbf{F}_{11}+\mu\beta\mathbf{I} & (1+\mu)\mathbf{F}_{12} \\ \mathbf{F}_{21} & (1+\mu)\mathbf{F}_{22}+\mu\beta\mathbf{I} \end{bmatrix}\right) = 0\\
\iff &\det\left(\begin{bmatrix} (1+\mu)\mathbf{F}_{11}+\mu\beta\mathbf{I} & \sqrt{1+\mu}\,\mathbf{F}_{12} \\ \sqrt{1+\mu}\,\mathbf{F}_{21} & (1+\mu)\mathbf{F}_{22}+\mu\beta\mathbf{I} \end{bmatrix}\right) = 0,
\end{aligned}
$$

where the second line has used the property that $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$ and the matrix similarity transform.

Thus, the whole proof now reduces to show that

$$
\mathbf{J}(\mu) \doteq \begin{bmatrix} (1+\mu)\mathbf{F}_{11}+\mu\beta\mathbf{I} & \sqrt{1+\mu}\,\mathbf{F}_{12} \\ \sqrt{1+\mu}\,\mathbf{F}_{21} & (1+\mu)\mathbf{F}_{22}+\mu\beta\mathbf{I} \end{bmatrix}
$$

has a zero eigenvalue for some $\mu > 0$. Note that $\mathbf{J}(\mu)$ is a symmetric matrix (with real eigenvalues) and is a continuous matrix function of μ. Then by Theorem 5.1 of (Kato, 2013), all the eigenvalues of $\mathbf{J}(\mu)$ (including the minimum eigenvalue $\lambda_{\min}(\mathbf{J}(\mu))$) are continuous functions of μ. We will show the real continuous function $\lambda_{\min}(\mathbf{J}(\mu))$ equals zero for some $\mu > 0$. Towards that end, we observe that

$$
\begin{aligned}
\mathbf{J}(0) &= \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} = \nabla^2 f(\boldsymbol{x}^\star,\boldsymbol{y}^\star),\\
\lim_{\mu\to\infty} \frac{\mathbf{J}(\mu)}{\mu} &= \begin{bmatrix} \mathbf{F}_{11}+\beta\mathbf{I} & \\ & \mathbf{F}_{22}+\beta\mathbf{I} \end{bmatrix} \succ 0.
\end{aligned}
$$

First, since $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ is a strict saddle of $f(\boldsymbol{x}, \boldsymbol{y})$, by definition of strict saddle, we have $\lambda_{\min}(\mathbf{J}(0)) < 0$. Second, since $\beta > L_f \geq \max\{\|\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)\|, \|\nabla^2_{\boldsymbol{y}} f(\boldsymbol{x}^\star, \boldsymbol{y}^\star)\|\}$ by the assumption, we have both $\mathbf{F}_{11} + \beta\mathbf{I}$ and $\mathbf{F}_{22} + \beta\mathbf{I}_m$ are positive definite and hence $\lambda_{\min}(\mathbf{J}(N)) > 0$ for some sufficiently large N. Finally, since $\lambda_{\min}(\mathbf{J}(\mu))$ is a continuous real-valued function of $\mathbf{X}$, we claim that there must exist a $\mu > 0$ such that $\lambda_{\min}(\mathbf{J}(\mu)) = 0$. □

4.3. Stylized Applications of Algorithm 2

We can apply the proximal alternating minimization (cf. Algorithm 2) to a popular large-scale matrix optimization method — the Burer-Monteiro Factorization (BMF) method (Burer & Monteiro, 2003). Given a large-scale matrix optimization problem

$$\underset{\mathbf{M}\in\mathbb{R}^{n\times m}}{\text{minimize}}\ q(\mathbf{M}), \tag{19}$$

where n and m are super large, BMF factorizes the large matrix variable $\mathbf{M}$ into two smaller matrices $\mathbf{X}\mathbf{Y}^{\mathrm{T}}$ with $\mathbf{X} \in \mathbb{R}^{n\times r}$, $\mathbf{Y} \in \mathbb{R}^{m\times r}$, and focuses on

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times r}, \mathbf{Y}\in\mathbb{R}^{m\times r}}{\text{minimize}}\ q(\mathbf{X}\mathbf{Y}^{\mathrm{T}}). \tag{20}$$

It has been shown in (Ge et al., 2016; 2017; Bhojanapalli et al., 2016; Park et al., 2017; Li et al., 2018; Zhu et al., 2018) that when $q(\mathbf{M})$ some restricted well-conditioned property (e.g., RIP), then any second-order stationary point of (20) is a global minimum point of (19). So, the second-order convergence of the proximal alternating minimization will imply the global optimality convergence.

Example 2 (Matrix Sensing)**.** *Consider a regularized matrix sensing problem*

$$\underset{\mathbf{M}}{\text{minimize}}\ \|\mathcal{A}(\mathbf{M}) - \boldsymbol{y}\|_2^2 + \lambda\|\mathbf{M}\|_*,$$

where $\boldsymbol{y}$ is the observation vector and $\mathcal{A} : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ is the linear sampling operator with a bounded spectral norm $\|\mathcal{A}\| \leq \mathsf{L}$. Then the BMF approach focuses on

$$\underset{\mathbf{X},\mathbf{Y}}{\text{minimize}}\ \|\mathcal{A}(\mathbf{X}\mathbf{Y}^{\mathrm{T}}) - \boldsymbol{y}\|_2^2 + \frac{\lambda}{2}(\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2). \tag{21}$$

Denote $f(\mathbf{X}, \mathbf{Y})$ as the objective function of (21). Note that for this case, Assumption 4 is not satisfied because $\|\nabla^2 f\|$ is not globally upper bounded in the whole domain. We can bypass this issue by using the forward-invariant property of the proximal alternating minimization mapping g_β.

Lemma 5. *g_β is a forward-invariant mapping on any sublevel set $\Omega \doteq \mathrm{Lev}_f(\mathbf{U}, \mathbf{V})$ for any $\mathbf{U}, \mathbf{V}$, i.e., $g(\Omega) \subseteq \Omega$.*

Proof. In one way, for any $(\mathbf{X}, \mathbf{Y}) \in \Omega$, we have $f(\mathbf{X}, \mathbf{V}) \leq f(\mathbf{U}, \mathbf{V})$ by definition of Ω. In another way, letting $(\mathbf{X}_+, \mathbf{Y}_+) = g_\beta(\mathbf{X}, \mathbf{V})$, we have $f(\mathbf{X}_+, \mathbf{Y}_+) \leq f(\mathbf{X}, \mathbf{V})$ by the sufficient decrease property of Algorithm 2 (cf. (Attouch et al., 2010)). Therefore, $(\mathbf{X}_+, \mathbf{Y}_+) \in \Omega$. □

Proposition 4. *Choosing $\beta > L_f(\Omega)$ for some constant $L_f(\Omega)$ depending on $\Omega \doteq \mathrm{Lev}_f(\mathbf{U}, \mathbf{V})$, we have: (i) $\det(Dg_\beta) \neq 0$ on Ω, and (ii) all strict saddles of f in Ω are unstable fixed points of g_β. Then by Theorem 3, the set of all initialization points in Ω that will let g_β converge to strict saddles is of zero Lebesgue measure. Thus together with the first-order convergence (cf. Theorem 1), Algorithm 2 from random initialization in Ω almost surely converges to a second-order stationary solution of (21).*

Proof. By Theorem 3 and knowing g_β is forward-invariant in Ω, to prove Proposition 4, it suffices to show the terms (i) and (ii). To show these two, we first prove a local Lipschitz-gradient condition for f: $\|\nabla^2 f(\mathbf{X}, \mathbf{Y})\| \leq L_f(\Omega)$ for all $(\mathbf{X}, \mathbf{Y}) \in \Omega$. By definition of Ω, $(\mathbf{X}, \mathbf{Y}) \in \Omega$ gives that

$$f(\mathbf{X}, \mathbf{Y}) \leq f(\mathbf{U}, \mathbf{V}) \overset{①}{\Longrightarrow} \begin{cases} \|\mathcal{A}(\mathbf{X}\mathbf{Y}^{\mathrm{T}}) - \boldsymbol{y}\|_2^2 \leq f(\mathbf{U}, \mathbf{V}), \\ \frac{\lambda}{2}\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2 \ \leq f(\mathbf{U}, \mathbf{V}). \end{cases}$$

Now deonote $\mathbf{D} \doteq (\mathbf{D}_X, \mathbf{D}_Y)$, $\Lambda \doteq \lambda\|\mathbf{D}\|_F^2$, and compute

$$\begin{aligned}
&[\nabla^2 f(\mathbf{X}, \mathbf{Y})](\mathbf{D}, \mathbf{D}) \\
&= 2\|\mathcal{A}(\mathbf{X}\mathbf{D}_Y^{\mathrm{T}} + \mathbf{D}_X\mathbf{Y}^{\mathrm{T}})\|_2^2 + 4\langle\mathcal{A}(\mathbf{D}_X\mathbf{D}_Y^{\mathrm{T}}), \mathcal{A}(\mathbf{X}\mathbf{Y}^{\mathrm{T}}) - \boldsymbol{y}\rangle + \Lambda \\
&\leq \left(4\mathsf{L}^2(\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2) + 4\mathsf{L}\|\mathcal{A}(\mathbf{X}\mathbf{Y}^{\mathrm{T}}) - \boldsymbol{y}\|_2 + \lambda\right)\|\mathbf{D}\|_F^2.
\end{aligned}$$

Together with the definition of spectral norm, this implies

$$\begin{aligned}
\|\nabla^2 f(\mathbf{X}, \mathbf{Y})\| &= \underset{\mathbf{D}}{\text{maximize}}[\nabla^2 f(\mathbf{X}\mathbf{Y})](\mathbf{D}, \mathbf{D})/\|\mathbf{D}\|_F^2 \\
&\leq 4\mathsf{L}^2(\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2) + 4\mathsf{L}\|\mathcal{A}(\mathbf{X}\mathbf{Y}^{\mathrm{T}}) - \boldsymbol{y}\|_2 + \lambda \\
&\leq 8\mathsf{L}^2 f(\mathbf{U}, \mathbf{V})/\lambda + 4\mathsf{L}\sqrt{f(\mathbf{U}, \mathbf{V})} + \lambda \doteq L_f(\Omega),
\end{aligned}$$

where the second inequality follows from ①. Now given the local Lipschitz condition in Ω and the forward-invariant property $g(\Omega) \subseteq \Omega$, (i) and (ii) immediately follow from Lemma 3 and Lemma 4, respectively. □

Example 3 (Matrix Completion)**.** *Consider the matrix completion problem*

$$\underset{\mathbf{M}}{\text{minimize}}\ \|\mathbf{M} - \mathbf{M}^\star\|_{\boldsymbol{\Omega}}^2 + \lambda\|\mathbf{M}\|_*,$$

where $\mathbf{M}^\star$ is the ground-truth matrix, $\boldsymbol{\Omega}$ is the binary mask matrix, and $\|\mathbf{M}\|_{\boldsymbol{\Omega}} \doteq \|\boldsymbol{\Omega} \odot \mathbf{M}\|_F$. Then BMF focuses on

$$\underset{\mathbf{X},\mathbf{Y}}{\text{minimize}}\ \|\mathbf{X}\mathbf{Y}^{\mathrm{T}} - \mathbf{M}^\star\|_{\boldsymbol{\Omega}}^2 + \frac{\lambda}{2}(\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2). \tag{22}$$

We remark that the same result (cf. Proposition 4) applies to the matrix completion problem (22), because binary sampling operator $\boldsymbol{\Omega}$ is essentially a bounded linear operator.

Acknowledgements

This work was supported by the DARPA Lagrange Program under ON- R/SPAWAR contract N660011824020. The authors gratefully acknowledge Waheed Bajwa, Haroon Raja, Clement Royer, Yue Xie, Xinshuo Yang, Michael Wakin, and Stephen J. Wright for helpful discussions.

References

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Bamber, D. and Van Santen, J. P. How many parameters can a model have and still be testable? *Journal of Mathematical Psychology*, 29(4):443–473, 1985.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pp. 3873–3881, 2016.

Burer, S. and Monteiro, R. D. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Chen, Y. and Chi, Y. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31, 2018.

Chi, Y., Lu, Y. M., and Chen, Y. Nonconvex optimization meets low-rank matrix factorization: An overview. *arXiv preprint arXiv:1809.09573*, 2018.

Comon, P., Luciani, X., and De Almeida, A. L. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 23(7-8): 393–405, 2009.

Fritzsche, K. and Grauert, H. *From holomorphic functions to complex manifolds*, volume 213. Springer Science & Business Media, 2012.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pp. 797–842, 2015.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pp. 2973–2981, 2016.

Ge, R., Jin, C., and Zheng, Y. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1233–1242. PMLR, 2017.

Grippo, L. and Sciandrone, M. On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. *Operations research letters*, 26(3):127–136, 2000.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. Matrix completion and low-rank SVD via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, 2015.

Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674. ACM, 2013.

Jain, P., Kar, P., et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10 (3-4):142–336, 2017.

Kato, T. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pp. 1246–1257, 2016.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid saddle points. *Mathematical Programming*, 2019.

Li, Q., Zhu, Z., and Tang, G. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2018.

Li, X., Wang, Z., Lu, J., Arora, R., Haupt, J., Liu, H., and Zhao, T. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296*, 2016.

Lu, S., Hong, M., and Wang, Z. PA-GD: On the convergence of perturbed alternating gradient descent to second-order stationary points for structured nonconvex optimization. In *International Conference on Machine Learning*, 2019.

Netrapalli, P., Jain, P., and Sanghavi, S. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pp. 2796–2804, 2013.

Park, D., Kyrillidis, A., Carmanis, C., and Sanghavi, S. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, pp. 65–74, 2017.

Ponomarev, S. P. Submersions and preimages of sets of measure zero. *Siberian Mathematical Journal*, 28(1):153–163, 1987.

Shub, M. *Global stability of dynamical systems*. Springer Science & Business Media, 2013.

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5): 1131–1198, 2018.

Wang, Y., Yang, J., Yin, W., and Zhang, Y. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 1(3):248–272, 2008.

Xu, Y. and Yin, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

Zhang, Y., Lau, Y., Kuo, H.-w., Cheung, S., Pasupathy, A., and Wright, J. On the global geometry of sphere-constrained sparse blind deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4894–4902, 2017.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. The global optimization geometry of low-rank matrix optimization. *arXiv preprint arXiv:1703.01256*, 2017.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.