Burer-Monteiro factorizability of nuclear norm regularized optimization

Wenqing Ouyang[*] Ting Kei Pong[†] Man-Chung Yue[‡]

January 3, 2026

Abstract

This paper studies the relationship between the nuclear norm-regularized minimization problem, which minimizes the sum of a C^2 function h and a positive multiple of the nuclear norm, denoted by f, and its factorized problem obtained by the Burer-Monteiro technique. We are interested in deriving conditions that ensure every second-order stationary point of the factorized problem corresponds to a global minimizer of f, a property we call the r-factorizability of f in this paper. Under suitable restricted isometry property (RIP) type assumptions on h, we prove the r-factorizability of f. Moreover, the RIP constant in our paper is tight, in the sense that we can construct concrete examples of f that fail to be r-factorizable when the RIP constant is below the threshold. Our technique for constructing such examples is novel and may be of independent interest: specifically, we use a variant of the Von Neumann's trace inequality and relate the existence of such examples to the optimal value of a quadratic program involving the RIP constant, then we explicitly solve this optimization problem to detect all the possible counterexamples.

1 Introduction

Low-rank matrix estimation has been an extremely important and versatile problem that has attracted intense research over the last two decades and found many applications across a wide range of domains, such as network science [13], machine learning [12, 24], quantum physics [21], control [23] and imaging [34, 14], to name but a few. This paper focuses on the following low-rank optimization problem:

$$\min_{X\in\mathbb{R}^{m\times n}} f(X) := h(X) + \lambda\|X\|_*, \tag{1.1}$$

where h is assumed to be twice continuously differentiable, $\lambda > 0$, and $\|\cdot\|_*$ denotes the nuclear norm. Without loss of generality, we assume $m \le n$ throughout the paper. Problem (1.1) takes the form of the so-called composite minimization which has been heavily studied, especially when h is convex. Therefore, in principle it can be solved by many existing algorithms for composite

[*] Columbia University, New York, USA (wo2205@columbia.edu).

[†] The Hong Kong Polytechnic University, Hong Kong, People's Republic of China (tk.pong@polyu.edu.hk). The research of this author was partially supported by the Hong Kong Research Grants Council PolyU 15300122 and the PolyU internal grant 4-ZZPJ.

[‡] The Hong Kong Polytechnic University, Hong Kong, People's Republic of China (manchung.yue@polyu.edu.hk). The research of this author was partially supported by the Hong Kong Research Grants Council under the GRF project 17309423.

minimization, including in particular various proximal algorithms [26, 18, 30, 31], thanks to the closed-form expression of the proximal operator of the nuclear norm [8].

Nonetheless, in contemporary applications, the dimensions m and n of the decision variable X can potentially be extremely high, rendering methods working directly with the variable X impractical. For example, in collaborative filtering, which is a classical application of low-rank matrix estimation, the dimensions m and n could be of the order of millions or even higher [24]. Worse still, the computational cost of the proximal operator associated with the nuclear norm, which is a fundamental building block of many existing algorithms for solving problem (1.1), is a cubic function in m and n, as it involves the singular value decomposition of X. To circumvent this, researchers proposed to tackle problem (1.1) via the Burer-Monteiro factorization technique [6, 7, 27, 38, 35], which replaces the variable X by a low-rank approximation $UV^\top$ and solves the resulting problem:

$$\min_{U\in\mathbb{R}^{m\times r},V\in\mathbb{R}^{n\times r}} F_r(U,V) := h(UV^\top) + \frac{\lambda(\|U\|_F^2 + \|V\|_F^2)}{2}, \tag{1.2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $r \in [1,m]$ is an integer parameter specified by the modeler. The advantage of the factorized problem (1.2) over problem (1.1) is twofold. First, the objective function in the factorized problem (1.2) is differentiable as long as h is. In contrast, the objective function in problem (1.1) is nonsmooth because of the nuclear norm. Second, we often choose $r \ll \min\{m,n\}$ in practice. The total size $r(m+n)$ of the matrix variables (U,V) is therefore substantially smaller than the size mn of the variable X in the non-factorized counterpart (1.1).

Since our goal is to solve problem (1.1), the factorization rank r in (1.2) cannot be too small. Indeed, optimal solutions of problem (1.1) cannot be recovered by solving problem (1.2) through the correspondence $(U,V) \mapsto UV^\top$ if r is less than the minimum rank r^* of the optimal solutions of problem (1.1). This issue is currently addressed indirectly as follows. First, it can be readily shown that $U^*{V^*}^\top$ is a global minimizer of problem (1.1) for any global minimizer (U^*,V^*) of problem (1.2) if $r \geq r^*$ (*e.g.*, see [17, Lemma 1]). Second, despite the non-convexity due to the bilinear term $UV^\top$, the factorized problem (1.2) has no spurious local minimizers or second-order stationary points if r is *sufficiently large* and h satisfies certain technical conditions [19, 35], which implies that one can actually solve problem (1.1) with a strongly convex h by using any optimization algorithm with a second-order convergence guarantee on problem (1.2). However, a proper choice of the parameter r is highly nontrivial. As demonstrated by an example constructed in [35], merely having $r \geq r^*$ is not enough in general, let alone that the minimum solution rank r^* is often unknown in practice. Our paper also revolves around the choice of the factorization rank r by asking a different but more direct question:

> *When do all the second-order stationary points of problem* (1.2) *correspond to the global minimizers of problem* (1.1) *via the mapping* $(U,V) \mapsto UV^\top$*?*

This motivates the following definition.

Definition 1.1 (r-factorizability)**.** Let h be twice continuously differentiable. The function f in problem (1.1) is said to be r-factorizable if every second-order stationary point (U,V) of the function F_r in problem (1.2) satisfies that $UV^\top$ is a global minimizer of f.

With this definition, our problem is equivalent to the investigation of the r-factorizability of the objective function f of problem (1.1). Most previous works studying the r-factorizability, if not all, rely on the restricted isometry property of h [10, 9, 38, 33]. Here, we recall that for $\delta > 0$ and integers $s,t \geq 0$, a twice continuously differentiable function $h : \mathbb{R}^{m\times n} \to \mathbb{R}$ is said to satisfy δ-RIP$_{s,t}$ condition [19, 37, 33] if for all $X, H \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(X) \leq s$ and $\operatorname{rank}(H) \leq t$, it holds that

$$(1-\delta)\|H\|_F^2 \leq \nabla^2 h(X)[H,H] \leq (1+\delta)\|H\|_F^2.$$

A closely related subject concerns the factorizability in the symmetric, unregularized setting, in which case the analogue problem pairs are

$$\min_{X\in\mathbb{S}^n_+} h(X) \quad \text{and} \quad \min_{U\in\mathbb{R}^{n\times r}} \widetilde{h}(U) := h(UU^\top) \tag{1.3}$$

where $\mathbb{S}^n_+$ is the set of $n\times n$ symmetric positive semidefinite matrices and h is C^2. In [35, 36], the author considered the case where ∇h is Lipschitz differentiable and h is strongly convex, and showed that all second-order stationary points U^* of $\widetilde{h}$ satisfy that $U^*{U^*}^\top$ is the unique minimizer X^* of h over $\mathbb{S}^n_+$ under suitable conditions on the factorization rank r, solution rank $r^* = \operatorname{rank}(X^*)$ and the condition number κ (*i.e.*, the ratio between the Lipschitz constant of ∇h and the strong convexity modulus of h), namely (1) $r \geq r^*$ and $\kappa < 3$; or (2) $n > r \geq r^*$ and $r > \frac{1}{4}(\kappa-1)^2 r^*$; the author also constructed a function h with $\kappa = 3$ such that $\widetilde{h}$ has a second-order stationary point that does not correspond to any global minimizer of h over $\mathbb{S}^n_+$. The result was then extended to the class of non-strongly convex quadratic functions satisfying the δ-$\mathrm{RIP}_{r+r^*,r+r^*}$ condition in [35, Corollary 1.5] or [36, Theorem 1.4], with essentially the same bound on δ.

The asymmetric, unregularized case has also been studied in the literature, where the factorized problem is

$$\min_{U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{n\times r}} h(UV^\top), \tag{1.4}$$

It is known that $(U,V)\mapsto h(UV^\top)$ may have spurious second-order stationary points as long as $r < m$ even when h satisfies the (strongest) 0-$\mathrm{RIP}_{m,m}$ condition, as demonstrated by [36, Example 1.8]. This motivates the search of formulations that are equivalent to (1.4) but with better landscape properties. One example is

$$\min_{U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{n\times r}} \tilde{F}(U,V) := h(UV^\top) + \beta\|U^\top U - V^\top V\|_F^2, \tag{1.5}$$

which was first proposed in [28]. Upon defining $h_a : \mathbb{S}^{m+n} \to \mathbb{S}^{m+n}$ as

$$h_a(X) := h(X_2) + \beta(\|X_1\|_F^2 + \|X_3\|_F^2 - 2\|X_2\|_F^2), \quad X = \begin{bmatrix} X_1 & X_2 \\ X_2^\top & X_3 \end{bmatrix}, \ X_1 \in \mathbb{S}^m, \tag{1.6}$$

one can see that $\tilde{F}(U,V) = h_a\left(\begin{bmatrix} U \\ V \end{bmatrix} \begin{bmatrix} U^\top & V^\top \end{bmatrix}\right)$; in this case, it can be shown that the δ-$\mathrm{RIP}_{k,k}$ condition of h implies the 2δ-$\mathrm{RIP}_{k,k}$ condition of h_a, see [36, Fact 3.14]. Then one can reduce (1.5) to an instance of the second problem in (1.3) and apply [35, Corollary 1.5]. The best known RIP-type condition for ensuring the non-existence of spurious second-order stationary point of (1.5) was established in [36, Theorem 1.6].

For the assymetric, regularized problem (1.1), fewer works have been done. A common assumption in these works is that there exists an optimal solution to problem (1.1) and r is chosen to be at least the minimum solution rank r^*. In [22, Theorem 3], the author showed that if h is convex quadratic and satisfies the δ-$\mathrm{RIP}_{2r,2r}$ condition with $\delta < \frac{1}{3}$, then the corresponding f in problem (1.1) is r-factorizable. A similar result was established in [19, Theorem 2] for a general twice continuously differentiable convex function h with a more restrictive bound on δ. Later, in [16, Theorem 1],[1] it was shown that when h satisfies the δ-$\mathrm{RIP}_{2r,2r}$ condition with $\delta < \frac{1}{3}$, then the corresponding f in problem (1.1) is r-factorizable, and when $\delta \geq \frac{1}{3}$, a second-order stationary point of problem (1.2) corresponds to an approximate stationary point of problem (1.1). We are not aware

[1] The results in [16] were stated in terms of restricted strong convexity and restricted smoothness. The moduli α and β therein correspond to $1-\delta$ and $1+\delta$ in our discussion here, respectively.

of any tight result for (1.1), in the sense that once the proposed sufficient conditions fail, then an h can be explicitly constructed with the corresponding f in (1.1) being non-r-factorizable.

It is tempting to reduce the asymmetric problem to a symmetric problem, as in the unregularized case. Unfortunately, as we shall discuss in Section 1.1, this idea is inapplicable in the regularized case. To circumvent this, we adopt a different strategy to tackle the factorizability problem, which is outlined in Section 1.1. Our new techniques enable us to derive tight RIP-type conditions in the sense of Theorem 1.3 below.

To present our main results, we define the following classes.

Definition 1.2. Let $L \in (0, \infty)$, $\mu > 0$, $q, r^* \in [m] \cup \{0\}$. We define $\mathfrak{S}(L, \mu, q, r^*)$ to be the set of all $h \in C^2(\mathbb{R}^{m\times n})$ satisfying the following conditions:

(i) For all $X, H \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(X), \operatorname{rank}(H) \leq q + r^*$, it holds that:

$$\mu\|H\|_F^2 \leq \nabla^2 h(X)[H, H] \leq L\|H\|_F^2. \tag{1.7}$$

(ii) There exists a global minimizer $X^* \in \mathbb{R}^{m\times n}$ of f in (1.1) satisfying $\operatorname{rank}(X^*) = r^*$.

When $L = 1 + \delta$ and $\mu = 1 - \delta$, (1.7) can be viewed as the δ-RIP$_{q+r^*,q+r^*}$ condition. Let us also note that when $q + r^* \geq m$, (1.7) reduces to the μ-strong convexity of h and L-Lipschitz continuity of ∇h. To simplify the notation, we denote this latter kind of function classes by $\mathfrak{S}(L, \mu, r^*)$. Our first main result is the following theorem, which characterizes the r-factorizability of $\mathfrak{S}(L, \mu, r^*)$ in the following sense: when the conditions in Theorem 1.3 are satisfied, then for all h in $\mathfrak{S}(L, \mu, r^*)$ the corresponding f in (1.1) is r-factorizable, and if not, there exists $h \in \mathfrak{S}(L, \mu, r^*)$ such that the corresponding f in (1.1) is not r-factorizable.[2]

Theorem 1.3. *Let $r^* \in [m] \cup \{0\}$, $r \in [m]$, $\infty > L \geq \mu > 0$ and $\kappa := \frac{L}{\mu} \geq 1$. Suppose that r^*, r and κ satisfy any of the following conditions:*

(1) $r = m$.

(2) $r \geq r^*$ and $\min\{r, m - r^*\} > \frac{(\kappa-1)^2}{4} \min\{r^*, m - r\}$.

Then for all $h \in \mathfrak{S}(L, \mu, r^*)$, the corresponding f in (1.1) is r-factorizable. Otherwise, there exists a quadratic $h \in \mathfrak{S}(L, \mu, r^*)$ such that the corresponding f in (1.1) is not r-factorizable.

Next, by analyzing suitable subspaces, we also obtain the following corollary from Theorem 1.3.

Corollary 1.4. *Theorem 1.3 holds when $\mathfrak{S}(L, \mu, r^*)$ is replaced by $\mathfrak{S}(L, \mu, r, r^*)$.*

In the case of $r + r^* < m$, $L = 1 + \delta$ and $\mu = 1 - \delta$, Corollary 1.4 implies that f is r-factorizable if $r/r^* > \frac{\delta^2}{(1-\delta)^2}$ and h satisfies the δ-RIP$_{r+r^*,r+r^*}$ condition with a global minimizer $X^* \in \mathbb{R}^{m\times n}$ satisfying $\operatorname{rank}(X^*) = r^*$, and that an explicit counterexample can be constructed if $r/r^* \leq \frac{\delta^2}{(1-\delta)^2}$. This bound matches the necessity condition in [36, Theorem 1.6], which studied the different model (1.5). However, we would like to point out that the sufficient condition in [36, Theorem 1.6] requires $r/r^* > 4\delta^2/(1 - 2\delta)^2$, which does not match the necessary condition there.

Interestingly, our bound here also matches the bound in [35, Theorem 1.1], which is for (1.3) though. However, as we will discuss in Section 1.1 below, our results do not follow from [35] and cannot imply the results in [35] and [36], since we require the constant λ to be positive in (1.1), which yields special structures on the first-order stationary points of F_r in (1.2), c.f. Proposition 3.2.

[2] We will refer to this property as "tight in the sense of Theorem 1.3" for the rest of this paper.

1.1 Sketch of proof techniques

To the best of our knowledge, all the previous works [28, 9, 33, 36] on the non-existence of spurious second-order stationary point of asymmetric problem (1.5) rely on symmetric reduction. To study the factorizability of (1.1), naturally, one would want to invoke a similar reduction and then invoke results for the symmetric case, such as [35, 36]. However, such an idea will not work, as we now explain. First, note that the symmetric case counterpart of (1.1) is to minimize

$$h_s(X) = h(X_2) + \frac{\lambda}{2}\mathrm{tr}(X), \quad X = \begin{bmatrix} X_1 & X_2 \\ X_2^\top & X_3 \end{bmatrix}.$$

We therefore have $h_s\left(\begin{bmatrix} U \\ V \end{bmatrix}\begin{bmatrix} U^\top & V^\top \end{bmatrix}\right) = F_r(U, V)$. However, h_s loses the RIP property by Fact 5.1, since h_s is not even strongly convex on the one-dimensional linear subspace $\{tE_{11} : t \in \mathbb{R}\}$, where $E_{11} \in \mathbb{R}^{m\times n}$ is the matrix whose $(1,1)$-entry is 1 and is zero otherwise. Therefore, the classical techniques in the literature [28, 9, 33, 36] are not applicable here.

In contrast, our proofs do not rely on the symmetric reduction technique. We start with the following well-known fact, essentially a restatement of [25, Theorem 2.1.12]: if $h \in C^1(\mathbb{R}^{m\times n})$ is μ-strongly convex and ∇h is L-Lipschitz continuous, then for all $X, Y \in \mathbb{R}^{m\times n}$ it holds that

$$(L-\mu)\langle \nabla h(X) - \mu X - (\nabla h(Y) - \mu Y), X - Y\rangle \geq \|\nabla h(X) - \mu X - (\nabla h(Y) - \mu Y)\|_F^2. \qquad (1.8)$$

Next, by analyzing the first-order stationary point of F_r in (1.2), we show that if $(\bar{U}, \bar{V})$ is a stationary point of F_r, then $\bar{X} = \bar{U}\bar{V}^\top$ is a pseudo stationary point of f, in which case $\bar{X}$ and $\nabla h(\bar{X})$ admit simultaneous singular value decompositions; see Proposition 3.2 and Definition 3.4. Using a variant of Von Neumann's trace inequality (see Lemma 2.3), we transform (1.8) into a bound concerning the singular values of $\bar{X}, X^*, \nabla h(\bar{X}), \nabla h(X^*)$, when X^* and $\bar{X}$ are pseudo stationary points of f. Then, the existence of counterexamples can be transformed to the existence of a certain feasible solution to a quadratic program involving L and μ; see (4.16). We then solve this optimization problem analytically, which yields tight bounds in the sense of Theorem 1.3.

The generalization to the RIP case is based on the observation that, when (1.7) holds, the function h is μ-strongly convex and ∇h is L-Lipschitz continuous on any linear subspace of $\mathbb{R}^{m\times n}$ consisting of merely matrices whose rank is no more than $q + r^*$; see Fact 5.1. In particular, for $\mathrm{rank}(X^*) = r^*$ and $\mathrm{rank}(\bar{X}) = r$, we can restrict the function on a linear subspace of $\mathbb{R}^{(r+r^*)\times(r+r^*)}$ under proper orthogonal transformation, which contains X^* and $\bar{X}$ (see Lemma 5.2), and then apply the result for the strongly convex case (see Proposition 5.3).

Let us now compare our proof techniques to those used in the literature. In [35, 36], tight bounds on the RIP constant δ analogous to those in our Theorem 1.3 were obtained for (1.3), and their bounds are the same as ours despite the difference in the problems under study. Then the result in [36] for (1.3) was generalized to (1.5) by using symmetric reduction. Note that the best bounds on the RIP constant δ in [36, Theorem 1.6] for (1.5) are also not tight as in our Theorem 1.3, which may be attributed to their use of the reduction to the symmetric case, so that the δ-RIP$_{k,k}$ condition of h only yields the 2δ-RIP$_{k,k}$ condition on h_a in (1.6). Although the tight bound in [35, 36] for (1.3) aligns with our bound, their techniques are inapplicable in our context. In [35], the existence of counterexamples was formulated as an semidefinite programming (SDP) problem, then the author solved the optimization problem analytically to get a tight RIP constant bound. In [36], the author showed that the existence of (quadratic) counterexamples can be equivalently formulated as the existence of escape direction, and a sharp escape direction is constructed explicitly based on the tangent-normal decomposition of the manifold of rank-r positive semidefinite matrices. In contrast, we formulate a quadratic programming problem and solve it analytically. Our proof techniques are also different from those in [19, 16, 22], which focus on (1.2). Indeed, our proof is based on analyzing the singular values of the global minimizer X^* appeared in Definition 1.2 and other related matrices (see Proposition 4.4), which leads to the tight bounds in the sense of Theorem 1.3.

The remainder of the paper is organized as follows. In Section 2, we define the notation and present some preliminary results. The characterization of first- and second-order stationary points of problem (1.2) is presented in Section 3. In Section 4, we prove Theorem 1.3 by using the results in Section 2 and Section 3. The generalization to the RIP case, namely Corollary 1.4, is proved in Section 5.

2 Notation and preliminaries

Throughout this paper, we assume that $1 \leq r \leq m \leq n$ in problem (1.2). For a matrix $X \in \mathbb{R}^{m\times n}$, we let $\|X\|_*$, $\|X\|_2$ and $\|X\|_F$ denote its nuclear norm, spectral norm and Frobenius norm, respectively. The i-th largest singular value of X is denoted by $\sigma_i(X)$ for $i = 1, \ldots, m$. The vector of singular values is denoted by $\sigma(X) = \begin{bmatrix}\sigma_1(X) & \cdots & \sigma_m(X)\end{bmatrix}^\top$. The set of $n \times n$ orthogonal matrices is denoted by $\mathcal{O}^n$. For $x \in \mathbb{R}^s$, we denote by $\mathrm{Diag}(x) \in \mathbb{R}^{s\times s}$ the diagonal matrix with $(\mathrm{Diag}(x))_{ii} = x_i$ for $i = 1, \ldots, s$. Moreover, we define $\mathrm{diag} : \mathbb{R}^{s\times s} \to \mathbb{R}^s$ to be the adjoint operator of Diag. In this paper, to simplify the presentation, we also use $\widetilde{\mathrm{Diag}}$ and $\widetilde{\mathrm{diag}}$ to denote the possibly non-square versions of Diag and diag, respectively. Specifically, for $x \in \mathbb{R}^s$, $\widetilde{\mathrm{Diag}}(x)$ would be a diagonal matrix whose diagonal part is x, which is not necessarily square; the dimension of $\widetilde{\mathrm{Diag}}(x)$ is omitted when it can be understood from the context.[3] Also, for $X = [X_1 \;\; X_2] \in \mathbb{R}^{m\times n}$ with $X_1 \in \mathbb{R}^{m\times m}$ and $X_2 \in \mathbb{R}^{m\times(n-m)}$, we define $\widetilde{\mathrm{diag}}(X) := \mathrm{diag}(X_1) \in \mathbb{R}^m$. For $X \in \mathbb{R}^{m\times n}$, we define

$$\mathcal{O}_X := \{(R, P) \in \mathcal{O}^m \times \mathcal{O}^n : R\,\widetilde{\mathrm{Diag}}(\sigma(X))P^\top = X\}.$$

For a mapping $H : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$, we say H is Lipschitz continuous with modulus L if the following holds:

$$\|H(X) - H(Y)\|_F \leq L\|X - Y\|_F \qquad \forall X, Y \in \mathbb{R}^{m\times n}.$$

The strong convexity for an $h \in C^2(\mathbb{R}^{m\times n})$ is also defined with respect to the Frobenius norm. Namely, $h \in C^2(\mathbb{R}^{m\times n})$ is said to be μ-strongly convex if $\nabla^2 h(X)[Y, Y] \geq \mu\|Y\|_F^2$ for all $X, Y \in \mathbb{R}^{m\times n}$, where the Hessian $\nabla^2 h(X) : \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \to \mathbb{R}$ is regarded as a quadratic form on $\mathbb{R}^{m\times n}$. To avoid clutter, we sometimes use the notation $\nabla^2 h(X)[Y]^2$ to denote $\nabla^2 h(X)[Y, Y]$.

The set of nonnegative integers is denoted by $\mathbb{N}_0$. For a nonnegative integer r, we use $[r]$ to denote the set $\{1, \ldots, r\}$; in particular, $[0] := \emptyset$. The permutation group of order m is denoted by $\mathfrak{P}_m$, and the set of $m \times m$ permutation matrices is denoted by $\mathcal{P}_m$. Finally, for an $x \in \mathbb{R}$, we let $\lfloor x \rfloor$ denote the largest integer upper bounded by x.

We will need the following characterization of the subdifferential of the nuclear norm.

Proposition 2.1 ([29, Example 2])**.** *Let $X \in \mathbb{R}^{m\times n}$ be a matrix of rank s and $(R, P) \in \mathcal{O}_X$. Then,*

$$\partial\|X\|_* = \left\{R\begin{bmatrix} I & 0 \\ 0 & W\end{bmatrix}P^\top : \; W \in \mathbb{R}^{(m-s)\times(n-s)}, \;\; \|W\|_2 \leq 1\right\}.$$

Note that while the singular value decomposition of X is not unique, the subdifferential $\partial\|X\|_*$ is independent of the choice of the singular value decomposition.

Before ending this section, we present a variant of von Neumann's trace inequality. Its proof requires the following well-known result concerning doubly stochastic matrices.

[3] For example, if $R \in \mathbb{R}^{m\times m}$ and $P \in \mathbb{R}^{n\times n}$, then writing $R\,\widetilde{\mathrm{Diag}}(x)P^\top$ would imply that $\widetilde{\mathrm{Diag}}(x) \in \mathbb{R}^{m\times n}$.

Lemma 2.2. *Let $A \in \mathbb{R}^{m\times m}$ be a nonnegative matrix that satisfies*

$$\forall i \in [m], \quad \sum_{j=1}^{m} A_{ij} \le 1, \quad \sum_{j=1}^{m} A_{ji} \le 1.$$

Then, there exists a doubly stochastic matrix B such that $B_{ij} \ge A_{ij}$ for all i and j.

Proof. Let $\mathcal{R}$ and $\mathcal{C}$ be the sets consisting of the indices of the rows and columns of A whose sum is less than 1, respectively. Clearly, $\mathcal{R}$ and $\mathcal{C}$ must be simultaneously empty or nonempty. We modify the matrix A gradually in the following manner: at each step, we select $i \in \mathcal{R}$ and $j \in \mathcal{C}$, and enlarge A_{ij} until either the row sum of i-th row or the column sum of j-th column reaches 1. Then we update $\mathcal{R}$ and $\mathcal{C}$ and repeat this process. Since $\mathcal{R}$ and $\mathcal{C}$ are always simultaneously empty or nonempty, our algorithm is well defined. Moreover, after each step, the quantity $|\mathcal{R}| + |\mathcal{C}|$ is reduced by at least 1. Since this number is finite, we must end with $\mathcal{R} = \mathcal{C} = \emptyset$. Then the resulting matrix, denoted by B, is doubly stochastic, and it holds by construction that $B_{ij} \ge A_{ij}$ for all i and j. □

Below is the announced variant of von Neumann's trace inequality, which reduces to the classical von Neumann's inequality when C or D is a zero matrix.

Lemma 2.3. *Let A, B, C and D be nonnegative $m \times m$ diagonal matrices with diagonal vectors d^A, d^B, d^C and d^D, respectively. Then, we have*

$$\sup_{R\in\mathcal{O}^m, P\in\mathcal{O}^n} \operatorname{tr}(R \begin{bmatrix} A & 0 \end{bmatrix} P \begin{bmatrix} B \\ 0 \end{bmatrix}) + \operatorname{tr}(R \begin{bmatrix} C & 0 \end{bmatrix} P \begin{bmatrix} D \\ 0 \end{bmatrix}) = \max_{E\in\mathcal{P}_m} (d^A)^\top E d^B + (d^C)^\top E(d^D), \tag{2.1}$$

where $\mathcal{P}_m$ is the set of $m \times m$ permutation matrices.

Proof. For any $R \in \mathcal{O}^m$ and $P \in \mathcal{O}^n$, we have

$$\begin{aligned}
&\operatorname{tr}(R \begin{bmatrix} A & 0 \end{bmatrix} P \begin{bmatrix} B \\ 0 \end{bmatrix}) + \operatorname{tr}(R \begin{bmatrix} C & 0 \end{bmatrix} P \begin{bmatrix} D \\ 0 \end{bmatrix}) \\
&= \sum_{i,j=1}^{m} (d_i^A d_j^B + d_i^C d_j^D) P_{ij} R_{ji} \le \sum_{i,j=1}^{m} (d_i^A d_j^B + d_i^C d_j^D)(\frac{R_{ji}^2}{2} + \frac{P_{ij}^2}{2}) \\
&\overset{\text{(a)}}{=} \sum_{i,j=1}^{m} (d_i^A d_j^B + d_i^C d_j^D) Z_{ij} = (d^A)^\top Z d^B + (d^C)^\top Z d^D,
\end{aligned}$$

where in (a) we define $Z \in \mathbb{R}^{m\times m}$ such that $Z_{ij} = \frac{R_{ji}^2}{2} + \frac{P_{ij}^2}{2}$ for all i and j. Since $R \in \mathcal{O}^m$ and $P \in \mathcal{O}^n$, we see that all row sums and column sums of Z are at most 1. By Lemma 2.2, we know there is a doubly stochastic matrix Y such that $Y_{ij} \ge Z_{ij}$ for all i and j. Since d^A, d^B, d^C, d^D are all nonnegative, we have

$$(d^A)^\top Z d^B + (d^C)^\top Z d^D \le (d^A)^\top Y d^B + (d^C)^\top Y d^D.$$

Applying Birkhoff theorem (see, *e.g.*, [2, Theorem 1.2.5]), the matrix Y is a convex combination of permutation matrices, namely, $Y = \sum_{i=1}^{s} \lambda_i P_i$, where $P_i \in \mathcal{P}_m$, $\lambda_i \ge 0$ for each $i = 1, \dots, s$ with $\sum_{i=1}^{s} \lambda_i = 1$. Therefore, we see that

$$(d^A)^\top Y d^B + (d^C)^\top Y d^D = \sum_{i=1}^{s} \lambda_i [(d^A)^\top P_i d^B + (d^C)^\top P_i d^D] \le \sup_{E\in\mathcal{P}_m} (d^A)^\top E d^B + (d^C)^\top E(d^D).$$

This upper bound can be achieved by setting $R = E_*^\top$ and $P = \begin{bmatrix} E_* & 0 \\ 0 & I_{n-m} \end{bmatrix} \in \mathbb{R}^{n\times n}$, where E_* achieves the supremum in $\sup_{E\in\mathcal{P}_m} (d^A)^\top E d^B + (d^C)^\top E(d^D)$. □

3 First- and second-order stationary points of F_r

In this section, we present characterizations of first- and second-order stationary points of F_r in problem (1.2). Let us note that in the literature, the stationarity of F_r has already been studied in [19, 20]. However, for our purpose, we need to extract useful features that have not been documented in the literature. To begin with, we can derive directly from the first-order optimality condition of problem (1.2) (see (3.2) below; see also [19, Proposition 2]) that if (U, V) is a stationary point of F_r, then $U^\top U = V^\top V$, which is also called a balanced pair. The next lemma studies how the balancedness can affect the singular vectors of U and V, which serves as the basis to establish that X and $\nabla h(X)$ have SVDs with common orthonormal bases (possibly in different order) when $X = UV^\top$ with (U, V) being a first-order stationary point of F_r.

Lemma 3.1. *Let $V \in \mathbb{R}^{n\times r}$ and $U \in \mathbb{R}^{m\times r}$. Then, $U^\top U = V^\top V$ if and only if $\sigma(V) = \sigma(U)$ and for any $(P, Q) \in \mathcal{O}_V$ there exists R such that $(R, Q) \in \mathcal{O}_U$.*

Proof. To prove the "if" direction, we note that by the definitions of $\mathcal{O}_V$ and $\mathcal{O}_U$, $P\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top$ and $R\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top$ are singular value decompositions of V and U, respectively. Then,

$$\begin{aligned}U^\top U &= Q\widetilde{\mathrm{Diag}}(\sigma(U))^\top R^\top R\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top = Q\widetilde{\mathrm{Diag}}(\sigma(U))^\top \widetilde{\mathrm{Diag}}(\sigma(U))Q^\top \\ &= Q\widetilde{\mathrm{Diag}}(\sigma(V))^\top \widetilde{\mathrm{Diag}}(\sigma(V))Q^\top = Q\widetilde{\mathrm{Diag}}(\sigma(V))^\top P^\top P\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top = V^\top V.\end{aligned}$$

We next prove the "only if" direction. Suppose that $U^\top U = V^\top V$. The equality $\sigma(U) = \sigma(V)$ follows directly from the definition of singular values. For the remaining assertion, let $(P, Q) \in \mathcal{O}_V$. Then, $V = P\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top$ is a singular value decomposition. By the supposition $U^\top U = V^\top V$,

$$\begin{aligned}Q^\top U^\top UQ = Q^\top V^\top VQ &= Q^\top(P\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top)^\top(P\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top)Q \\ &= \mathrm{Diag}(\sigma_1^2(V), \ldots, \sigma_s^2(V), 0, \ldots, 0),\end{aligned} \tag{3.1}$$

where $s := \mathrm{rank}(V)$ and hence $\sigma_1(V), \ldots, \sigma_s(V) > 0$. Denote by $\hat{u}_i$ the i-th column of UQ for $i \in [m]$. It then follows from (3.1) that the vectors $\hat{u}_1/\sigma_1(V), \ldots, \hat{u}_s/\sigma_s(V)$ are orthonormal and that $\hat{u}_i = 0$ for $i = s+1, \ldots, m$. There must exist $m - s$ vectors $r_{s+1}, \ldots, r_m$ so that $R = [\hat{u}_1/\sigma_1(V), \ldots, \hat{u}_s/\sigma_s(V), r_{s+1}, \ldots, r_m] \in \mathcal{O}^m$. By the definition of R and the fact that $\sigma(V) = \sigma(U)$, we have

$$\begin{aligned}R\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top &= [\hat{u}_1/\sigma_1(V), \ldots, \hat{u}_s/\sigma_s(V), r_{s+1}, \ldots, r_m]\,\widetilde{\mathrm{Diag}}(\sigma_1(V), \ldots, \sigma_s(V), 0, \ldots, 0)\,Q^\top \\ &= [\hat{u}_1, \ldots, \hat{u}_s, 0, \ldots, 0]\,Q^\top = UQQ^\top = U,\end{aligned}$$

which implies $(R, Q) \in \mathcal{O}_U$ and thus completes the proof. □

Next, we are ready to establish the SVD structures of X and $\nabla h(X)$ when $X = UV^\top$ with (U, V) being a first-order stationary point of F_r. The proof is basically done by substituting the SVDs of (U, V) into the first-order optimality condition and then solving the stationary equations.

Proposition 3.2 (First-order stationarity). *A pair $(U, V) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ is a stationary point of F_r in (1.2) if and only if there exist $R \in \mathcal{O}^m$, $P \in \mathcal{O}^n$ and $Q \in \mathcal{O}^r$ such that $(R, Q) \in \mathcal{O}_U$, $(P, Q) \in \mathcal{O}_V$, $\sigma(U) = \sigma(V)$, and $\nabla h(UV^\top) = -R\,\widetilde{\mathrm{Diag}}(d)\,P^\top$ for some $d \in \mathbb{R}^m$ satisfying $d_1 = \cdots = d_s = \lambda$ and $d_{s+1} \geq \cdots \geq d_m \geq 0$, where $s = \mathrm{rank}(U) = \mathrm{rank}(V)$.*

Remark 3.3. (i) Note that the decomposition $-\nabla h(UV^\top) = R\,\widetilde{\mathrm{Diag}}(d)\,P^\top$ in Proposition 3.2 is not a singular value decomposition in general because it is possible that $d_{s+1} > \lambda = d_1 = \cdots = d_s$. Nevertheless, the vector d contains all the singular values of $-\nabla h(UV^\top)$, *i.e.*, d is $\sigma(-\nabla h(UV^\top))$ up to a permutation of the entries.

(ii) For a stationary point (U, V) of F_r, Proposition 3.2 shows that $\mathrm{rank}(U) = \mathrm{rank}(V)$ and $UV^\top = R[\widetilde{\mathrm{Diag}}(\sigma(U))]^2 P^\top = R[\widetilde{\mathrm{Diag}}(\sigma(V))]^2 P^\top$ for some $R \in \mathcal{O}^m$ and $P \in \mathcal{O}^n$. Hence, $\sigma_i(UV^\top) = \sigma_i^2(U) = \sigma_i^2(V)$ for all $i \in [m]$.

Proof of Proposition 3.2. The first-order optimality condition of problem (1.2) reads

$$\begin{cases} \nabla h(UV^\top)V + \lambda U = 0, \\ \nabla h(UV^\top)^\top U + \lambda V = 0. \end{cases} \tag{3.2}$$

We first prove the "if" direction. By supposition, we have that

$$\begin{aligned} &\nabla h(UV^\top)V + \lambda U = -R\widetilde{\mathrm{Diag}}(d)P^\top P\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top + \lambda R\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top \\ =& -R\widetilde{\mathrm{Diag}}(d)P^\top P\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top + \lambda R\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top \\ =& -R\begin{bmatrix} \lambda I_s & 0 \\ 0 & \widetilde{\mathrm{Diag}}(d_{s+1},\ldots,d_m) \end{bmatrix}\begin{bmatrix} \mathrm{Diag}(\sigma_1(U),\ldots,\sigma_s(U)) & 0 \\ 0 & 0 \end{bmatrix} Q^\top + \lambda R\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top \\ =& -\lambda R\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top + \lambda R\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top = 0, \end{aligned}$$

which shows the first equality in (3.2). Similarly, we have

$$\begin{aligned} &\nabla h(UV^\top)^\top U + \lambda V = -P\widetilde{\mathrm{Diag}}(d)R^\top R\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top + \lambda P\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top \\ =& -P\widetilde{\mathrm{Diag}}(d)R^\top R\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top + \lambda P\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top \\ =& -P\begin{bmatrix} \lambda I_s & 0 \\ 0 & \widetilde{\mathrm{Diag}}(d_{s+1},\ldots,d_m) \end{bmatrix}\begin{bmatrix} \mathrm{Diag}(\sigma_1(V),\ldots,\sigma_s(V)) & 0 \\ 0 & 0 \end{bmatrix} Q^\top + \lambda P\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top \\ =& -\lambda P\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top + \lambda P\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top = 0, \end{aligned}$$

which shows the second equality in (3.2). This proves the "if" direction.

To prove the "only if" direction, we assume that (U, V) is a stationary point of F_r in (1.2), *i.e.*, (3.2) holds. A direct computation shows that $U^\top U = V^\top V$; see also [19, Proposition 2]. Fix a singular value decomposition of V:

$$V = P_1 \begin{bmatrix} \mathrm{Diag}(\sigma_1(V),\ldots,\sigma_s(V)) & 0 \\ 0 & 0 \end{bmatrix} Q_1^\top. \tag{3.3}$$

By Lemma 3.1, there exists some $R_1 \in \mathcal{O}^m$ such that

$$U = R_1 \begin{bmatrix} \mathrm{Diag}(\sigma_1(V),\ldots,\sigma_s(V)) & 0 \\ 0 & 0 \end{bmatrix} Q_1^\top. \tag{3.4}$$

Next, we write

$$\nabla h(UV^\top) = R_1 \begin{bmatrix} A & B \\ C & D \end{bmatrix} P_1^\top, \tag{3.5}$$

for some $A \in \mathbb{R}^{s\times s}$, $B \in \mathbb{R}^{s\times(n-s)}$, $C \in \mathbb{R}^{(m-s)\times s}$, $D \in \mathbb{R}^{(m-s)\times(n-s)}$. Substituting (3.3), (3.4) and (3.5) into (3.2), we get

$$\begin{aligned} A\,\mathrm{Diag}(\sigma_1(V),\ldots,\sigma_s(V)) + \lambda\,\mathrm{Diag}(\sigma_1(V),\ldots,\sigma_s(V)) &= 0,\\ C\,\mathrm{Diag}(\sigma_1(V),\ldots,\sigma_s(V)) &= 0,\\ A^\top\,\mathrm{Diag}(\sigma_1(V),\ldots,\sigma_s(V)) + \lambda\,\mathrm{Diag}(\sigma_1(V),\ldots,\sigma_s(V)) &= 0,\\ B^\top\,\mathrm{Diag}(\sigma_1(V),\ldots,\sigma_s(V)) &= 0, \end{aligned}$$

which imply that $B = C = 0$, $A = -\lambda I_s$ and D is unconstrained.

Finally, let $(R_2, P_2) \in \mathcal{O}_{-D}$ and define the following orthogonal matrices

$$P = P_1 \begin{bmatrix} I_s & 0 \\ 0 & P_2 \end{bmatrix} \quad \text{and} \quad R = R_1 \begin{bmatrix} I_s & 0 \\ 0 & R_2 \end{bmatrix}.$$

Using (3.3) and (3.4), one can check readily that $(P, Q_1) \in \mathcal{O}_V$ and $(R, Q_1) \in \mathcal{O}_U$. Moreover, using (3.5) together with the facts that $B = C = 0$ and $A = -\lambda I_s$, we see that

$$\nabla h(UV^\top) = -R_1 \begin{bmatrix} \lambda I_s & 0 \\ 0 & -D \end{bmatrix} P_1^\top = -R \begin{bmatrix} \lambda I_s & 0 \\ 0 & \widetilde{\mathrm{Diag}}(\sigma(-D)) \end{bmatrix} P^\top.$$

This completes the proof. □

In view of the first-order optimality condition of (1.1) and Proposition 2.1, one can see that the SVD structures of $UV^\top$ and $\nabla h(UV^\top)$ given by Proposition 3.2 share some features of the true stationary point X^* of f in (1.1) and the corresponding $\nabla h(X^*)$, even though $UV^\top$ is not necessarily a stationary point of f. To treat such points and the true stationary points under a unified framework, we define the following category of points of the function f.

Definition 3.4 (Pseudo-stationarity)**.** A matrix $X \in \mathbb{R}^{m\times n}$ is said to be a pseudo-stationary point of f in (1.1) if there exist $(R, P) \in \mathcal{O}_X$ and $d \in \mathbb{R}^m_+$ such that $-\nabla h(X) = R\widetilde{\mathrm{Diag}}(d)P^\top$ and $d_1 = \cdots = d_s = \lambda$, where $s = \mathrm{rank}(X)$.

Assume that X^* and $\bar{X}$ are two pseudo-stationary points of h. When h is μ-strongly convex and ∇h is L-Lipschitz continuous, we know that (1.8) holds with $X = X^*$ and $Y = \bar{X}$. We next leverage Proposition 3.2, which states that X^* and $\nabla h(X^*)$, as well as $\bar{X}$ and $\nabla h(\bar{X})$, admit simultaneous SVDs, to transform (1.8) to a bound involving only the singular values of X^*, $\nabla h(X^*)$, $\bar{X}$ and $\nabla h(\bar{X})$.

Lemma 3.5. *Let X_1, X_2 be two pseudo-stationary points of f in* (1.1) *in the sense of Definition 3.4, i.e., there exist R_1, $R_2 \in \mathcal{O}^m$ and P_1, $P_2 \in \mathcal{O}^n$ such that*

$$\begin{aligned} X_1 &= R_1 \begin{bmatrix} \Sigma_1 & 0_{m\times(n-m)} \end{bmatrix} P_1^\top, \; -\nabla h(X_1) = R_1 \begin{bmatrix} D_1 & 0_{m\times(n-m)} \end{bmatrix} P_1^\top,\\ X_2 &= R_2 \begin{bmatrix} \Sigma_2 & 0_{m\times(n-m)} \end{bmatrix} P_2^\top, \; -\nabla h(X_2) = R_2 \begin{bmatrix} D_2 & 0_{m\times(n-m)} \end{bmatrix} P_2^\top, \end{aligned} \tag{3.6}$$

where $\Sigma_i = \mathrm{diag}(\sigma_1(X_i), \ldots, \sigma_m(X_i)) \in \mathbb{R}^{m\times m}$ and $D_i = \mathrm{diag}(d_1^i, \ldots, d_m^i) \in \mathbb{R}^{m\times m}_+$ with $d_1^i = \cdots = d^i_{\mathrm{rank}(X_i)} = \lambda$ for $i = 1, 2$. Assume that h in (1.1) *satisfies that h is μ-strongly convex for some $\mu > 0$, and ∇h is Lipschitz continuous with modulus $L \geq \mu$. Then, we have*

$$\begin{aligned} \max_{\tau\in\mathfrak{P}_m} &\sum_{i=1}^m (L\sigma_i(X_1) + d_i^1)(\mu\sigma_{\tau(i)}(X_2) + d^2_{\tau(i)}) + \sum_{i=1}^m (\mu\sigma_i(X_1) + d_i^1)(L\sigma_{\tau(i)}(X_2) + d^2_{\tau(i)})\\ &- \sum_{i=1}^m (\mu\sigma_i(X_1) + d_i^1)(L\sigma_i(X_1) + d_i^1) - \sum_{i=1}^m (\mu\sigma_i(X_2) + d_i^2)(L\sigma_i(X_2) + d_i^2) \geq 0. \end{aligned} \tag{3.7}$$

Proof. Define $\phi(\cdot) := h(\cdot) - \frac{\mu}{2}\|\cdot\|_F^2$, and let us rewrite (1.8) as

$$0 \leq (L-\mu)\langle \nabla\phi(X_1) - \nabla\phi(X_2), X_1 - X_2\rangle - \|\nabla\phi(X_1) - \nabla\phi(X_2)\|_F^2. \tag{3.8}$$

For the right hand side of (3.8), a direct computation shows that

$$\begin{aligned}
&(L-\mu)\langle \nabla\phi(X_1) - \nabla\phi(X_2), X_1 - X_2\rangle - \|\nabla\phi(X_1) - \nabla\phi(X_2)\|_F^2 \\
&= \langle \nabla\phi(X_1) - \nabla\phi(X_2), (L-\mu)X_1 - \nabla\phi(X_1) - ((L-\mu)X_2 - \nabla\phi(X_2))\rangle \\
&= \langle -\nabla\phi(X_1), (L-\mu)X_2 - \nabla\phi(X_2)\rangle + \langle -\nabla\phi(X_2), (L-\mu)X_1 - \nabla\phi(X_1)\rangle \\
&\quad - \langle -\nabla\phi(X_1), (L-\mu)X_1 - \nabla\phi(X_1)\rangle - \langle -\nabla\phi(X_2), (L-\mu)X_2 - \nabla\phi(X_2)\rangle \\
&= \langle \mu X_1 - \nabla h(X_1), LX_2 - \nabla h(X_2)\rangle + \langle \mu X_2 - \nabla h(X_2), LX_1 - \nabla h(X_1)\rangle \\
&\quad - \langle \mu X_1 - \nabla h(X_1), LX_1 - \nabla h(X_1)\rangle - \langle \mu X_2 - \nabla h(X_2), LX_2 - \nabla h(X_2)\rangle \\
&=: S_1 + S_2,
\end{aligned} \tag{3.9}$$

where $S_1 := \langle \mu X_1 - \nabla h(X_1), LX_2 - \nabla h(X_2)\rangle + \langle \mu X_2 - \nabla h(X_2), LX_1 - \nabla h(X_1)\rangle$ and $S_2 := -\langle \mu X_1 - \nabla h(X_1), LX_1 - \nabla h(X_1)\rangle - \langle \mu X_2 - \nabla h(X_2), LX_2 - \nabla h(X_2)\rangle$.

We now rewrite S_1 and S_2. We start by noting that for S_2, its two summands can be rewritten as follows using (3.6): for $i = 1, 2$,

$$-\langle \mu X_i - \nabla h(X_i), LX_i - \nabla h(X_i)\rangle = -\sum_{j=1}^{m}(\mu\sigma_j(X_i) + d_j^i)(L\sigma_j(X_i) + d_j^i). \tag{3.10}$$

Next, for S_1, notice that

$$\begin{aligned}
S_1 &= \langle \mu X_1 - \nabla h(X_1), LX_2 - \nabla h(X_2)\rangle + \langle \mu X_2 - \nabla h(X_2), LX_1 - \nabla h(X_1)\rangle \\
&\overset{\text{(a)}}{=} \left\langle R_1 \begin{bmatrix}\mu\Sigma_1 + D_1 & 0\end{bmatrix} P_1^\top, R_2 \begin{bmatrix}L\Sigma_2 + D_2 & 0\end{bmatrix} P_2^\top \right\rangle \\
&\quad + \left\langle R_1 \begin{bmatrix}L\Sigma_1 + D_1 & 0\end{bmatrix} P_1^\top, R_2 \begin{bmatrix}\mu\Sigma_2 + D_2 & 0\end{bmatrix} P_2^\top \right\rangle \\
&= \left\langle R_2^\top R_1 \begin{bmatrix}\mu\Sigma_1 + D_1 & 0\end{bmatrix} P_1^\top P_2, \begin{bmatrix}L\Sigma_2 + D_2 & 0\end{bmatrix} \right\rangle \\
&\quad + \left\langle R_2^\top R_1 \begin{bmatrix}L\Sigma_1 + D_1 & 0\end{bmatrix} P_1^\top P_2, \begin{bmatrix}\mu\Sigma_2 + D_2 & 0\end{bmatrix} \right\rangle,
\end{aligned} \tag{3.11}$$

where in (a) we have used (3.6). Using the above display and Lemma 2.3, we see that

$$S_1 \leq \max_{\tau\in\mathfrak{P}_m} \sum_{i=1}^{m}(\mu\sigma_i(X_1) + d_i^1)(L\sigma_{\tau(i)}(X_2) + d_{\tau(i)}^2) + \sum_{i=1}^{m}(L\sigma_i(X_1) + d_i^1)(\mu\sigma_{\tau(i)}(X_2) + d_{\tau(i)}^2).$$

The desired conclusion now follows immediately upon combining the above displays. □

We already know that when $(\bar{U}, \bar{V})$ is a stationary point of F_r in (1.2), then $\bar{X} = \bar{U}\bar{V}^\top$ is a pseudo stationary point of f in (1.1). The next step is to identify more structural information of $\bar{X}$ and $\nabla h(\bar{X})$ when $(\bar{U}, \bar{V})$ is a second-order stationary point of F_r. The first task is to rewrite the second-order optimality conditions such that it is aligned with the SVD of $\bar{X}$. To be more precise, the second-order optimality condition can be written as $\nabla^2 F_r(\bar{U}, \bar{V})[U, V]^2 \geq 0$ for all $(U, V) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$. To be more aligned with the SVDs of $(\bar{U}, \bar{V})$ given in Proposition 3.2, we may rewrite it as

$$\forall (U, V) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}, \quad \nabla^2 F_r(\bar{U}, \bar{V})[(RUQ^\top, PVQ^\top), (RUQ^\top, PVQ^\top)] \geq 0, \tag{3.12}$$

where $(R,Q) \in \mathcal{O}_U$ and $(P,Q) \in \mathcal{O}_V$. Note that using [19, Equation (3.14)],[4] we have

$$\begin{aligned}&\nabla^2 F_r(\bar U,\bar V)[(RUQ^\top, PVQ^\top),(RUQ^\top, PVQ^\top)]\\ &= 2\langle R^\top \nabla h(\bar X)P, UV^\top\rangle + \lambda(\|U\|_F^2 + \|V\|_F^2) + \nabla^2 h(\bar X)[R\left(R^\top \bar U Q V^\top + UQ^\top \bar V^\top P\right)P^\top]^2.\end{aligned} \tag{3.13}$$

Next, we introduce a natural partition for matrices of sizes $\mathbb{R}^{m\times r}$ and $\mathbb{R}^{n\times r}$. Let $(\bar U,\bar V) \in \mathbb{R}^{m\times r}\times\mathbb{R}^{n\times r}$ satisfy $\operatorname{rank}(\bar U) = \operatorname{rank}(\bar V)$ (which holds in particular when $(\bar U,\bar V)$ is a stationary point of F_r, thanks to Proposition 3.2). Denote this common rank by $s = \operatorname{rank}(\bar U) = \operatorname{rank}(\bar V) \le r$. We can then partition any matrices $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ into the following block form:

$$U = \begin{bmatrix} U_{11} & U_{12}\\ U_{21} & U_{22}\end{bmatrix} \quad\text{and}\quad V = \begin{bmatrix} V_{11} & V_{12}\\ V_{21} & V_{22}\end{bmatrix}, \tag{3.14}$$

where $U_{11}, V_{11} \in \mathbb{R}^{s\times s}$, $U_{12}, V_{12} \in \mathbb{R}^{s\times(r-s)}$, $U_{21} \in \mathbb{R}^{(m-s)\times s}$, $V_{21} \in \mathbb{R}^{(n-s)\times s}$, $U_{22} \in \mathbb{R}^{(m-s)\times(r-s)}$, $V_{22} \in \mathbb{R}^{(n-s)\times(r-s)}$. Note that when $\bar U$ and $\bar V$ are of full rank, *i.e.*, $s = r$, the matrices U_{12}, U_{22}, V_{12} and V_{22} are null. Now, we are ready to state the following characterization of second-order stationarity of F_r in (1.2).

Proposition 3.6 (Second-order stationarity)**.** *A pair $(\bar U,\bar V) \in \mathbb{R}^{m\times r}\times\mathbb{R}^{n\times r}$ is a second-order stationary point of F_r in* (1.2) *if and only if both of the following two conditions hold:*

(i) There exist $R \in \mathcal{O}^m$, $P \in \mathcal{O}^n$ and $Q \in \mathcal{O}^r$ such that $(R,Q) \in \mathcal{O}_{\bar U}$, $(P,Q) \in \mathcal{O}_{\bar V}$, $\sigma(\bar U) = \sigma(\bar V)$ and $\nabla h(\bar U\bar V^\top) = -R\,\widetilde{\operatorname{Diag}}(d)\,P^\top$ for some $d \in \mathbb{R}^m$ satisfying $d_1 = \cdots = d_s = \lambda$ and $d_{s+1} \ge \cdots \ge d_m \ge 0$, where $s = \operatorname{rank}(\bar U) = \operatorname{rank}(\bar V)$.

(ii) For any $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$, it holds that[5]

$$\begin{aligned}&-2\lambda\operatorname{tr}(U_{11}^\top V_{11} + U_{12}V_{12}^\top) - 2\operatorname{tr}(D^\top(U_{21}V_{21}^\top + U_{22}V_{22}^\top))\\ &+\lambda(\|U\|_F^2 + \|V\|_F^2) + \nabla^2 h(\bar U\bar V^\top)\left[R\begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top\\ U_{21}\Sigma & 0\end{bmatrix}P^\top\right]^2 \ge 0,\end{aligned} \tag{3.15}$$

where $\Sigma = \operatorname{Diag}(\sigma_1(\bar U),\ldots,\sigma_s(\bar U)) \in \mathbb{R}^{s\times s}$, and $D = \widetilde{\operatorname{Diag}}(d_{s+1},\ldots,d_m) \in \mathbb{R}^{(m-s)\times(n-s)}$ with d_i given in (i).

Moreover, if $s = \operatorname{rank}(\bar U) < r$, then (i) and (ii) imply that[6] $\|\nabla h(\bar U\bar V^\top)\|_2 \le \lambda$.

Proof. A pair $(\bar U,\bar V) \in \mathbb{R}^{m\times r}\times\mathbb{R}^{n\times r}$ is a second-order stationary point of F_r if and only if it satisfies both the first- and second-order optimality conditions. By Proposition 3.2, the first-order optimality condition is equivalent to (i).

Using (3.13), the second-order condition in (3.12) can be rewritten as that for all $(U,V) \in \mathbb{R}^{m\times r}\times\mathbb{R}^{n\times r}$,

$$\begin{aligned}&2\langle R^\top\nabla h(\bar X)P, UV^\top\rangle + \lambda(\|U\|_F^2 + \|V\|_F^2) + \nabla^2 h(\bar X)[\bar U Q V^\top P^\top + RUQ^\top \bar V^\top]^2 \ge 0\\ \overset{\text{(a)}}{\Longleftrightarrow}\;& -2\left\langle \begin{bmatrix}\lambda I_s & 0\\ 0 & D\end{bmatrix}, UV^\top\right\rangle + \lambda(\|U\|_F^2 + \|V\|_F^2)\end{aligned}$$

[4] Note that [19] assumed h is convex, but the derivation of Equation (3.14) there does not rely on the convexity assumption.

[5] Here, we use the partition (3.14) with respect to $(\bar U,\bar V)$; this is well defined because $\sigma(\bar U) = \sigma(\bar V)$ holds in (i).

[6] We would like to point out this last claim (i.e., "if $s = \operatorname{rank}(\bar U) < r$, then (i) and (ii) imply that $\|\nabla h(\bar U\bar V^\top)\|_2 \le \lambda$.") has been proved in [3, Theorem 1]. We include its proof for completeness.

$$+ \nabla^2 h(\bar{X}) \left[R \left(\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^\top + U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \right) P^\top \right]^2 \geq 0 \tag{3.16}$$

$$\overset{(b)}{\iff} - 2\lambda \mathrm{tr}(U_{11}V_{11}^\top + U_{12}V_{12}^\top) - 2\mathrm{tr}(D^\top (U_{21}V_{21}^\top + U_{22}V_{22}^\top))$$

$$+ \lambda(\|U\|_F^2 + \|V\|_F^2) + \nabla^2 h(\bar{X}) \left[R \begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0 \end{bmatrix} P^\top \right]^2 \geq 0, \tag{3.17}$$

where (a) follows from the decomposition for $-\nabla h(\bar{X})$ in (i) and the definition of D and Σ, and (b) from the definition of the blocks in (3.14). This shows that (i) and (ii) together form an equivalent characterization of the second-order stationary points.

We next prove the second claim. Assume that $s = \mathrm{rank}(\bar{U}) < r$. Note that this implies that U_{22} and V_{22} are not null. Hence, we can take U and V in (3.15) to be the matrices with the blocks $U_{22} = [e_1\ 0]$ and $V_{22} = [e_1\ 0]$ and U_{11}, U_{12}, U_{21}, V_{11}, V_{12}, V_{21} all being zero matrices to deduce that

$$-2\sigma_1(D) + 2\lambda \geq 0.$$

The desired conclusion now follows immediately from the above display and the decomposition of $\nabla h(\bar{U}\bar{V}^\top)$ in (i). □

Using the second-order condition in Proposition 3.6, we now derive the information on the singular values of $\nabla h(\bar{X})$, when $\bar{X} = \bar{U}\bar{V}^\top$ with $(\bar{U}, \bar{V})$ being a second-order stationary point of F_r in (1.2). The technique used here is not new, and similar arguments can be found in [11] and [3].

Theorem 3.7 (Singular values for 2nd stationary points). *Let $(\bar{U}, \bar{V}) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ be a second-order stationary point of F_r in* (1.2)*, $s = \mathrm{rank}(\bar{U})$ and $d \in \mathbb{R}^m$ be given in Proposition 3.6. Then,*[7] *$d_{s+1} \leq \lambda + \tilde{L}\sigma_r(\bar{U}\bar{V}^\top)$, where*

$$\tilde{L} := \sup_{\substack{Y\in\mathbb{R}^{m\times n},\, \|Y\|_F=1, \\ \mathrm{rank}(Y)=2}} \nabla^2 h(\bar{U}\bar{V}^\top)[Y, Y].$$

Proof. We first consider the case when $\mathrm{rank}(\bar{U}) < r$. Since $\mathrm{rank}(\bar{U}) < r$, we have $\sigma_r(\bar{U}\bar{V}^\top) = 0$. Therefore,

$$d_{s+1} \leq \|d\|_\infty = \|\nabla h(\bar{U}\bar{V}^\top)\|_2 \leq \lambda = \lambda + \tilde{L}\sigma_r(\bar{U}\bar{V}^\top),$$

where the first equality follows from Remark 3.3(i) and the second inequality follows from Proposition 3.6.

We next consider the case where $\mathrm{rank}(\bar{U}) = r$. If $r = m$, then $d_{m+1} = 0$ and the desired conclusions then hold trivially. Thus, from now on, we assume $r < m$.

By Proposition 3.2, $\mathrm{rank}(\bar{V}) = \mathrm{rank}(\bar{U}) = r$. Therefore, in this case, the blocks $U_{12}, U_{22}, V_{12}, V_{22}$ in (3.14) are null. Since $(\bar{U}, \bar{V})$ is a second-order stationary point of F_r in (1.2), by Proposition 3.6, it satisfies the following inequality for any matrices $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$:

$$0 \leq - 2\lambda \mathrm{tr}(U_{11}^\top V_{11}) - 2\mathrm{tr}(D^\top U_{21}V_{21}^\top) + \lambda(\|U\|_F^2 + \|V\|_F^2)$$

$$+ \nabla^2 h(\bar{U}\bar{V}^\top) \left[R \begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0_{(m-r)\times(n-r)} \end{bmatrix} P^\top \right]^2. \tag{3.18}$$

Taking U and V to be the matrices with U_{11} and V_{11} being zero and $U_{21} = [e_r\ 0]^\top = e_1 e_r^\top \in \mathbb{R}^{(m-r)\times r}$ and $V_{21} = [e_r\ 0]^\top = e_1 e_r^\top \in \mathbb{R}^{(n-r)\times r}$, we have that $\Sigma V_{21}^\top = \sigma_r(\bar{U}) e_r e_1^\top$ and $U_{21}\Sigma = \sigma_r(\bar{U}) e_1 e_r^\top$ and

[7] We define $d_{m+1} = 0$.

that $\mathrm{tr}(D^\top U_{21} V_{21}^\top) = e_1^\top D e_1 = d_{r+1}$. Substituting these into (3.18) yields

$$\begin{aligned} 0 &\le -2d_{r+1} + 2\lambda + \nabla^2 h(\bar U \bar V^\top)\left[\sigma_r(\bar U) R \begin{bmatrix} 0_{r\times r} & e_r e_1^\top \\ e_1 e_r^\top & 0_{(m-r)\times(n-r)} \end{bmatrix} P^\top\right]^2 \\ &\le -2d_{r+1} + 2\lambda + 2\sigma_r^2(\bar U)\tilde L = -2d_{r+1} + 2\lambda + 2\sigma_r(\bar U \bar V^\top)\tilde L, \end{aligned}$$

where the second inequality follows from the fact that $\left\|\sigma_r(\bar U) R \begin{bmatrix} 0_{r\times r} & e_r e_1^\top \\ e_1 e_r^\top & 0_{(m-r)\times(n-r)} \end{bmatrix} P^\top\right\|_F = \sqrt{2}\sigma_r(\bar U)$ and the definition of $\tilde L$, and the equality follows from Remark 3.3(ii). Hence, $d_{r+1} \le \lambda + \tilde L\, \sigma_r(\bar U \bar V^\top)$. □

In view of Theorem 3.7, the definition of d in Proposition 3.6 and Proposition 2.1, for a second-order stationary point $(\bar U, \bar V)$ of F_r in (1.2), by setting $\bar X = \bar U \bar V^\top$, we see that if $\sigma_r(\bar X) = 0$, then $\bar X$ is a stationary point of f, as proved in [3]; see, also [32, Section 3], and [15, 4, 5] for similar results. This is formally presented as the next corollary.

Corollary 3.8. *For any second-order stationary point $(\bar U, \bar V) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ of F_r in* (1.2) *satisfying* $\mathrm{rank}(\bar U) < r$, *$\bar U \bar V^\top$ is a stationary point of f in* (1.1).

Intuitively, based on the bound on d_{s+1} in Theorem 3.7, we can say that the smaller $\sigma_r(\bar X)$ is, the closer $\bar X$ is to being a stationary point of f in (1.1).

4 Characterization of r-factorizability for strongly convex functions

Suppose we have an $h \in \mathfrak{S}(L, \mu, r^*)$ in hand such that the corresponding f in (1.1) is not r-factorizable. By definition, we know there exists $X^* \in \mathbb{R}^{m\times n}$ such that X^* is a stationary point of f. Let $(\bar U, \bar V)$ be the second-order stationary point of F_r in (1.1) such that $\bar X = \bar U \bar V^\top$ is not a global minimizer of f (or equivalently, not a stationary point of f since f is strongly convex thanks to $h \in \mathfrak{S}(L, \mu, r^*)$). Then according to Proposition 3.2, we know $\bar X$ is a pseudo stationary point of f. In this case, we have the bound (3.7) on the singular values of X^*, $\nabla h(X^*)$, $\bar X$ and $\nabla h(\bar X)$. In addition, for a pseudo stationary point X of f, it is a stationary point of f if and only if $\|\nabla h(X)\|_2 \le \lambda$ in view of Definition 3.4 and Proposition 2.1. Finally, the singular value of $\nabla h(\bar X)$ is further constrained by Theorem 3.7. Putting all these conditions together, we have the following result.

Proposition 4.1. *Let $L \ge \mu > 0$, $r \in [m]$ and $r^* \in [m] \cup \{0\}$. Assume that there exists an $h \in \mathfrak{S}(L, \mu, r^*)$ (see Definition 1.2) such that f in* (1.1) *is not r-factorizable. Then, there exist $x, g, y, v \in \mathbb{R}^m$ with $\|g\|_\infty > \lambda$ and $\tau \in \mathfrak{P}_m$ such that*

$$\begin{aligned} &\sum_{i=1}^m (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i + g_i)(L y_{\tau(i)} + v_{\tau(i)}) \\ &- \sum_{i=1}^m (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^m (L y_i + v_i)(\mu y_i + v_i) \ge 0, \end{aligned} \tag{4.1a}$$

and

$$\forall i \in [r],\ x_i > 0,\ g_i = \lambda, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = \lambda, \tag{4.1b}$$

$$\forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, \lambda + L \min_{j\in[r]} x_j], \quad \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, \lambda]. \tag{4.1c}$$

Proof. By assumption, we can select $h \in \mathfrak{S}(L, \mu, r^*)$ and $X_2 \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(X_2) = r^*$, $-\nabla h(X_2) \in \lambda\partial\|X_2\|_*$, and f is not r-factorizable. The latter means we can find $(\bar{U}, \bar{V})$ being a second-order stationary point of F_r in (1.2) and $X_1 = \bar{U}\bar{V}^\top$ is not a stationary point of f.

Applying Proposition 3.2 (see also Remark 3.3) to $(\bar{U}, \bar{V})$ and using Proposition 2.1 and the condition that $-\nabla h(X_2) \in \lambda\partial\|X_2\|_*$, we can write

$$\begin{aligned} X_1 = R_1 \begin{bmatrix}\Sigma_1 & 0\end{bmatrix} P_1^\top,\ -\nabla h(X_1) = R_1 \begin{bmatrix}D_1 & 0\end{bmatrix} P_1^\top, \\ X_2 = R_2 \begin{bmatrix}\Sigma_2 & 0\end{bmatrix} P_2^\top,\ -\nabla h(X_2) = R_2 \begin{bmatrix}D_2 & 0\end{bmatrix} P_2^\top, \end{aligned} \tag{4.2}$$

for some $R_i \in \mathcal{O}^m$ and $P_i \in \mathcal{O}^n$ and $m \times m$ diagonal matrices Σ_i and D_i, $i = 1, 2$, where $\operatorname{diag}(\Sigma_i) \in \mathbb{R}^m_+$ consisting of all the singular values of X_i in descending order, $d^i := \operatorname{diag}(D_i) \in \mathbb{R}^m_+$ with $d^i_1 = \cdots = d^i_{\operatorname{rank}(X_i)} = \lambda$, for $i = 1, 2$. Clearly, we have $\operatorname{rank}(X_1) = r$, otherwise by Corollary 3.8 we can conclude that X_1 is a stationary point of f, leading to a contradiction. Applying Lemma 3.5, there exists $\bar{\tau} \in \mathfrak{P}_m$ such that

$$\begin{aligned} &\sum_{i=1}^m (L\sigma_i(X_1) + d^1_i)(\mu\sigma_{\bar{\tau}(i)}(X_2) + d^2_{\bar{\tau}(i)}) + \sum_{i=1}^m (\mu\sigma_i(X_1) + d^1_i)(L\sigma_{\bar{\tau}(i)}(X_2) + d^2_{\bar{\tau}(i)}) \\ &- \sum_{i=1}^m (L\sigma_i(X_1) + d^1_i)(\mu\sigma_i(X_1) + d^1_i) - \sum_{i=1}^m (L\sigma_i(X_2) + d^2_i)(\mu\sigma_i(X_2) + d^2_i) \geq 0, \end{aligned} \tag{4.3}$$

where L and μ are defined in Definition 1.2 for the h we selected. Next, applying Theorem 3.7, we know for all $i \geq r + 1$, it holds that $d^1_i \leq \lambda + L\sigma_r(X_1)$; in addition, it must hold that $\|d^1\|_\infty > \lambda$ for otherwise, (4.2) and Proposition 2.1 would imply that X_1 is a stationary point of f, which is a contradiction. On the other hand, using the fact that $-\nabla h(X_2) \in \lambda\partial\|X_2\|_*$, (4.2) and Proposition 2.1, we know $d^2_i \leq \lambda$ for all $i \in [m]$. This means that $(x, g, y, v, \tau) = (\operatorname{diag}(\Sigma_1), \operatorname{diag}(D_1), \operatorname{diag}(\Sigma_2), \operatorname{diag}(D_2), \bar{\tau})$ satisfies (4.1a)–(4.1c) and $\|g\|_\infty > \lambda$. □

To provide a complete characterization, it is important to know whether the converse of Proposition 4.1 holds. The first issue we need to solve is that, a feasible pair (x, g, y, v) satisfying (4.1b) and (4.1c) does not necessarily satisfy that x and y are sorted in descending order, which means that they cannot be the singular values of any matrix. Nevertheless, we can reduce to the case where (x, y) are sorted in descending order by doing an additional permutation.

Lemma 4.2. *Let $L \geq \mu > 0$, $r \in [m]$ and $r^* \in [m] \cup \{0\}$. Assume that (x, g, y, v, τ) satisfies* (4.1a)*–*(4.1c) *with $\|g\|_\infty > \lambda$, and we pick $\tau_1, \tau_2 \in \mathfrak{P}_m$ such that*[8]

$$\bar{x} = \tau_1(x),\ \bar{y} = \tau_2(y) \text{ are sorted in descending order, and we define } \bar{g} = \tau_1(g),\ \bar{v} = \tau_2(v), \tag{4.4}$$

Then $(\bar{x}, \bar{g}, \bar{y}, \bar{v}, \rho)$ also satisfies (4.1a)*–*(4.1c) *with $\|\bar{g}\|_\infty > \lambda$, where $\rho := \tau_2^{-1}\tau\tau_1$.*

Proof. According to the constraints on x, y in (4.1b) and (4.1c), we know that $x_i > 0$ if and only if $i \in [r]$, and $y_j > 0$ if and only if $j \in [r^*]$. Consequently, according to the definition of τ_1 and τ_2, we have

$$\tau_1([r]) = [r], \quad \tau_1([m] \setminus [r]) = [m] \setminus [r], \quad \tau_2([r^*]) = [r^*], \quad \tau_2([m] \setminus [r^*]) = [m] \setminus [r^*]. \tag{4.5}$$

Therefore, $(\bar{x}, \bar{g}, \bar{y}, \bar{v})$ satisfies (4.1b) and (4.1c). To verify (4.1a), notice that the four sums in (4.1a) takes the form $\sum_{i=1}^m a_i b_{\tau(i)}$ for some $a, b \in \mathbb{R}^m$. We also note that for any $a, b, c, d \in \mathbb{R}^m$ with

[8] Here for a vector $x \in \mathbb{R}^n$ and a permutation τ, $\tau(x) \in \mathbb{R}^n$ is defined as $\tau(x)_i = x_{\tau(i)}$ for all $i \in [n]$.

$a = \tau_1(c)$ and $b = \tau_2(d)$, it holds that

$$\sum_{i=1}^{m} a_i b_{\rho(i)} \overset{\text{(a)}}{=} \sum_{i=1}^{m} c_{\tau_1(i)} d_{\tau_2\rho(i)} \overset{\text{(b)}}{=} \sum_{i=1}^{m} c_{\tau_1(i)} d_{\tau\tau_1(i)} \overset{\text{(c)}}{=} \sum_{i=1}^{m} c_i d_{\tau(i)}, \tag{4.6}$$

where (a) holds because $a = \tau_1(c)$ and $b = \tau_2(d)$, (b) holds as $\rho = \tau_2^{-1}\tau\tau_1$, and in (c) we have used the substitution $i \leftarrow \tau_1^{-1}(i)$. Consequently, to verify (4.1a) for $(\bar{x}, \bar{g}, \bar{y}, \bar{v}, \rho)$, it suffices to use (4.4), (4.6) and (4.1a) for (x, g, y, v, τ). Finally, the condition $\|\bar{g}\|_\infty > \lambda$ is clear since $\bar{g}$ is a permutation of g and $\|g\|_\infty > \lambda$. □

Consequently, if there exists (x, y, g, v, τ) satisfying (4.1a)–(4.1c) with $\|g\|_\infty > \lambda$, then there is no loss of generality if we assume x and y are sorted in descending order. In this case, we would like to construct $\bar{X}$, $\nabla h(\bar{X})$, X^* and $\nabla h(X^*)$ such that $\sigma(\bar{X}) = x$, $\sigma(X^*) = y$ and the bound in (1.8) is satisfied. To this end, we first define

$$\bar{X} := \widetilde{\mathrm{Diag}}(x),\ X^* := \widetilde{\mathrm{Diag}}(\tau(y)),\ \bar{G} := \widetilde{\mathrm{Diag}}(g),\ G^* := \widetilde{\mathrm{Diag}}(\tau(v)), \tag{4.7}$$

In this case, by direct calculation, we have

$$\begin{aligned} &(L-\mu)\langle (G^* + \mu X^*) - (\bar{G} + \mu\bar{X}), \bar{X} - X^* \rangle - \|(\bar{G} + \mu\bar{X}) - (G^* + \mu X^*)\|_F^2 \\ &= \langle (G^* + \mu X^*) - (\bar{G} + \mu\bar{X}), L\bar{X} + \bar{G} - (LX^* + G^*) \rangle \\ &= \langle G^* + \mu X^*, L\bar{X} + \bar{G} \rangle + \langle \bar{G} + \mu\bar{X}, LX^* + G^* \rangle \\ &\quad - \langle G^* + \mu X^*, LX^* + G^* \rangle - \langle \bar{G} + \mu\bar{X}, L\bar{X} + \bar{G} \rangle \\ &\overset{\text{(a)}}{=} \sum_{i=1}^{m} (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^{m} (\mu x_i + g_i)(L y_{\tau(i)} + v_{\tau(i)}) \\ &\quad - \sum_{i=1}^{m} (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^{m} (Ly_i + v_i)(\mu y_i + v_i) \overset{\text{(b)}}{\geq} 0 \end{aligned} \tag{4.8}$$

where (a) follows from the diagonal structures in (4.7), and (b) follows from (4.1a). The display (4.8) implies that letting $\nabla h(\bar{X}) = -\bar{G}$ and $\nabla h(X^*) = -G^*$ does not violate (1.8) when h is μ-strongly convex and ∇h is L-Lipschitz. The next question is, does there exist such a function $h \in C^2(\mathbb{R}^{m\times n})$? Notice that $-\nabla h(X^*) \in \lambda\partial\|X^*\|_*$ if such a function exists, which means X^* is a stationary point of f in (1.1), and $h \in \mathfrak{S}(L, \mu, r^*)$. This question is a special case of the extension of strongly convex function [1]. However, for our purpose, we need to further ensure that there exists a second-order stationary point $(\bar{U}, \bar{V})$ of the corresponding F_r in (1.2) such that $\bar{U}\bar{V}^\top = \bar{X}$, which will be achieved by adding quadratic penalty on the off diagonal entries due to the structure of the second-order optimality condition in Section 3. Let us now provide the construction of such example.

Lemma 4.3. *Let $L \geq \mu > 0$, $r \in [m]$ and $r^* \in [m] \cup \{0\}$. Assume that (x, g, y, v, τ) satisfies* (4.1a)*–*(4.1c)*, and x and y are sorted in descending order. Let $\bar{X}$, X^*, $\bar{G}$, $G^* \in \mathbb{R}^{m\times n}$ be defined as* (4.7)*. If $G^* + \mu X^* \neq \bar{G} + \mu\bar{X}$, we define a quadratic function h as follows:*

$$\begin{aligned} h(X) =& \frac{L}{2}\sum_{i=1}^{m}\sum_{j\neq i}^{n} X_{ij}^2 + \frac{\mu}{2}\sum_{i=1}^{m}(X_{ii} - (\bar{X})_{ii})^2 - \langle \bar{G}, X \rangle \\ &+ \frac{(\langle X - \bar{X}, -G^* - \mu X^* + \bar{G} + \mu\bar{X} \rangle)^2}{2\langle X^* - \bar{X}, -G^* - \mu X^* + \bar{G} + \mu\bar{X} \rangle}, \end{aligned} \tag{4.9}$$

Otherwise, we set

$$h(X) = \frac{L}{2}\sum_{i=1}^{m}\sum_{j\neq i}^{n} X_{ij}^2 + \frac{\mu}{2}\sum_{i=1}^{m}(X_{ii} - (\bar{X})_{ii})^2 - \langle \bar{G}, X\rangle. \tag{4.10}$$

Then h is well defined and μ-strongly convex, ∇h is Lipschitz continuous with modulus L, $\nabla h(\bar{X}) = -\bar{G}$, and $\nabla h(X^*) = -G^*$. Moreover, if we define F_r as in (1.2) *with the above h and define $(\bar{U}, \bar{V}) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ as*

$$\bar{U} := \widetilde{\mathrm{Diag}}(\sqrt{\sigma_1(\bar{X})}, \ldots, \sqrt{\sigma_r(\bar{X})}) \quad \text{and} \quad \bar{V} := \widetilde{\mathrm{Diag}}(\sqrt{\sigma_1(\bar{X})}, \ldots, \sqrt{\sigma_r(\bar{X})}), \tag{4.11}$$

then $(\bar{U}, \bar{V})$ is a second-order stationary point of F_r. In particular, if $\|g\|_\infty > \lambda$, then f in (1.1) *is not r-factorizable given such an h.*

Proof. We first consider the case where $\bar{G} + \mu\bar{X} = G^* + \mu X^*$. In this case, the function h in (4.10) is clearly well defined, and one can verify that h is μ-strongly convex, and ∇h is Lipschitz continuous with modulus L. Moreover, $\nabla h(\bar{X}) = -\bar{G}$ and

$$\nabla h(X^*) = \mu(X^* - \bar{X}) - \bar{G} = -G^*.$$

Now it remains to show that $(\bar{U}, \bar{V})$ is a second-order stationary point of F_r.

We start by noticing from Proposition 3.2 that $(\bar{U}, \bar{V})$ is a stationary point of F_r (with $R = I_m$, $Q = I_r$ and $P = I_n$ in Proposition 3.2). Consequently, by Proposition 3.6, we know that $(\bar{U}, \bar{V})$ is a second-order stationary point of F_r if and only if for all $U_{11}, V_{11} \in \mathbb{R}^{r\times r}$, $U_{21} \in \mathbb{R}^{(m-r)\times r}$, $V_{21} \in \mathbb{R}^{(n-r)\times r}$,[9] it holds that

$$\begin{aligned} &-2\lambda \mathrm{tr}(U_{11}^\top V_{11}) - 2\mathrm{tr}(D^\top U_{21}V_{21}^\top) + \lambda(\|U_{11}\|_F^2 + \|V_{11}\|_F^2 + \|U_{21}\|_F^2 + \|V_{21}\|_F^2) \\ &+ \nabla^2 h(\bar{X})\left[\begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0 \end{bmatrix}\right]^2 \geq 0, \end{aligned} \tag{4.12}$$

where

$$\begin{aligned} &\Sigma = \mathrm{diag}(\sqrt{\sigma_1(\bar{X})}, \ldots, \sqrt{\sigma_r(\bar{X})}) \overset{(a)}{=} \mathrm{diag}(\sqrt{x_1}, \ldots, \sqrt{x_r}) \in \mathbb{R}^{r\times r}, \\ &D = \widetilde{\mathrm{Diag}}(g_{r+1}, \ldots, g_m) \in \mathbb{R}^{(m-r)\times(n-r)}, \end{aligned} \tag{4.13}$$

and in (a) we have used the fact that x is a nonnegative vector sorted in descending order. We will verify (4.12).

To this end, we first use the representation of h in (4.10) to deduce that

$$\begin{aligned} \nabla^2 h(\bar{X})\left[\begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0 \end{bmatrix}\right]^2 &\geq L\|\Sigma V_{21}^\top\|_F^2 + L\|U_{21}\Sigma\|_F^2 \\ &\overset{\text{(a)}}{\geq} Lx_r(\|U_{21}\|_F^2 + \|V_{21}\|_F^2), \end{aligned} \tag{4.14}$$

where in (a) we have used the fact that $\|AB\|_F \geq \sigma_{\min}(A)\|B\|_F$ for any square matrix A. Therefore, it holds that

$$-2\lambda \mathrm{tr}(U_{11}^\top V_{11}) - 2\mathrm{tr}(D^\top U_{21}V_{12}^\top) + \lambda(\|U_{11}\|_F^2 + \|V_{11}\|_F^2 + \|U_{21}\|_F^2 + \|V_{21}\|_F^2)$$

[9] Here, we use the partition in (3.14), which is well defined because $\sigma(\bar{U}) = \sigma(\bar{V})$. We also note that $U_{12}, V_{12}, U_{22}, V_{22}$ are void because $\mathrm{rank}(\bar{U}) = r$.

$$
\begin{aligned}
&+ \nabla^2 h(\bar{X})\left[\begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0 \end{bmatrix}\right]^2 \\
&\overset{\text{(a)}}{\geq} -2\mathrm{tr}(D^\top U_{21} V_{21}^\top) + \lambda(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) + \nabla^2 h(\bar{X})\left[\begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0 \end{bmatrix}\right]^2 \\
&\overset{\text{(b)}}{\geq} -2\|D^\top\|_2 \|U_{21}\|_F \|V_{21}^\top\|_F + \lambda(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) + \nabla^2 h(\bar{X})\left[\begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0 \end{bmatrix}\right]^2 \\
&\overset{\text{(c)}}{\geq} -2(\lambda + Lx_r)\|U_{21}\|_F \|V_{21}\|_F + \lambda(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) + Lx_r(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) \\
&\overset{\text{(d)}}{\geq} -(\lambda + Lx_r)(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) + \lambda(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) + Lx_r(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) = 0,
\end{aligned}
$$

where in (a) we have used the Cauchy-Schwartz inequality to show that $\mathrm{tr}(U_{11}^\top V_{11}) \leq \frac{1}{2}(\|U_{11}\|_F^2 + \|V_{11}\|_F^2)$, in (b) we have used the fact $\mathrm{tr}(ABC) = \mathrm{tr}(CAB) \leq \|C\|_F \|AB\|_F \leq \|A\|_2 \|C\|_F \|B\|_F$, in (c) we have used (4.1c), (4.14) and the definition of D in (4.13), and in (d) we have used the fact $\|A\|_F \|B\|_F \leq \frac{1}{2}(\|A\|_F^2 + \|B\|_F^2)$. This verifies (4.12) and hence $(\bar{U}, \bar{V})$ is a second-order stationary point of F_r.

Next, we consider the case where $\bar{G} + \mu\bar{X} \neq G^* + \mu X^*$. By (4.8) and the fact that $\bar{G} + \mu\bar{X} \neq G^* + \mu X^*$, we see that $(L-\mu)\langle (G^* + \mu X^*) - (\bar{G} + \mu\bar{X}), \bar{X} - X^* \rangle \geq \|(\bar{G} + \mu\bar{X}) - (G^* + \mu X^*)\|_F^2 > 0$. In particular, this implies $L > \mu$ and $\langle (G^* + \mu X^*) - (\bar{G} + \mu\bar{X}), \bar{X} - X^* \rangle > 0$, showing that h in (4.9) is well defined. Furthermore, we have

$$
L - \mu \geq \frac{\|(\bar{G} + \mu\bar{X}) - (G^* + \mu X^*)\|_F^2}{\langle (G^* + \mu X^*) - (\bar{G} + \mu\bar{X}), \bar{X} - X^* \rangle}. \tag{4.15}
$$

Now, it is routine to check that h is μ-strongly convex, $\nabla h(\bar{X}) = -\bar{G}$, and $\nabla h(X^*) = -G^*$. Moreover, the relation in (4.14) and hence the second-order stationarity of $(\bar{U}, \bar{V})$ can be verified similarly to that in the case where $\bar{G} + \mu\bar{X} = G^* + \mu X^*$. Thus, it remains to show that ∇h is Lipschitz continuous with modulus L.

To this end, notice that for all $X, Y \in \mathbb{R}^{m\times n}$ and the function h defined in (4.9), it holds that

$$
\begin{aligned}
&\nabla^2 h(X)[Y,Y] = L\sum_{i=1}^m \sum_{j\neq i}^n Y_{ij}^2 + \mu\sum_{i=1}^n Y_{ii}^2 + \frac{(\langle Y, -G^* - \mu X^* + \bar{G} + \mu\bar{X}\rangle)^2}{\langle X^* - \bar{X}, -G^* - \mu X^* + \bar{G} + \mu\bar{X}\rangle} \\
&\overset{\text{(a)}}{=} L\sum_{i=1}^m \sum_{j\neq i}^n Y_{ij}^2 + \mu\sum_{i=1}^n Y_{ii}^2 + \frac{(\langle \widetilde{\mathrm{Diag}}(Y_{11},\ldots,Y_{mm}), -G^* - \mu X^* + \bar{G} + \mu\bar{X}\rangle)^2}{\langle X^* - \bar{X}, -G^* - \mu X^* + \bar{G} + \mu\bar{X}\rangle} \\
&\overset{\text{(b)}}{\leq} L\sum_{i=1}^m \sum_{j\neq i}^n Y_{ij}^2 + \mu\sum_{i=1}^n Y_{ii}^2 + \frac{\|\widetilde{\mathrm{Diag}}(Y_{11},\ldots,Y_{mm})\|_F^2 \, \|-G^* - \mu X^* + \bar{G} + \mu\bar{X}\|_F^2}{\langle X^* - \bar{X}, -G^* - \mu X^* + \bar{G} + \mu\bar{X}\rangle} \overset{\text{(c)}}{\leq} L\|Y\|_F^2,
\end{aligned}
$$

where in (a) we have used the fact that $\bar{X}, X^*, \bar{G}, G^*$ are diagonal (see (4.7)), in (b) we have used the Cauchy-Schwartz inequality, and in (c) we have used (4.15). This proves that ∇h is Lipschitz continuous with modulus L.

Finally, if $\|g\|_\infty > \lambda$, then by Proposition 2.1 we know $\bar{X}$ is not a stationary point of f in (1.1) given such an h, and hence cannot be a global minimizer. Therefore, in this case, f in (1.1) is not r-factorizable. □

Combining Proposition 4.1, Lemma 4.2 and Lemma 4.3, we have the following result concerning the existence of the function f in (1.1) when $h \in \mathfrak{S}(L, \mu, r^*)$, which is not r-factorizable.

Proposition 4.4. *Let $L \geq \mu > 0$, $r \in [m]$ and $r^* \in [m] \cup \{0\}$. Consider the following optimization problem:*

$$
\begin{aligned}
\sup_{\substack{x,g,y,v \in \mathbb{R}^m \\ \tau \in \mathfrak{P}_m}} \quad & \sum_{i=1}^m (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i + g_i)(Ly_{\tau(i)} + v_{\tau(i)}) \\
& - \sum_{i=1}^m (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^m (Ly_i + v_i)(\mu y_i + v_i) \\
s.t. \quad & \forall i \in [r],\ x_i > 0,\ g_i = \lambda, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = \lambda, \\
& \forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, \lambda + L \min_{j \in [r]} x_j], \\
& \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, \lambda].
\end{aligned}
\tag{4.16}
$$

Then, the existence of $h \in \mathfrak{S}(L, \mu, r^*)$ such that the corresponding f in (1.1) *is not r-factorizable, is equivalent to the existence of a feasible solution (x, g, y, v, τ) to* (4.16) *with $\|g\|_\infty > \lambda$ and a nonnegative objective function value.*

Proof. If there exists $h \in \mathfrak{S}(L, \mu, r^*)$ such that the corresponding f in (1.1) is not r-factorizable, then by Proposition 4.1 we know there exists a feasible solution (x, g, y, v, τ) with $\|g\|_\infty > \lambda$ to (4.16) having nonnegative function value. Conversely, if there exists a feasible solution (x, g, y, v, τ) with $\|g\|_\infty > \lambda$ to (4.16) having nonnegative function value, by Lemma 4.2 we may assume that x and y are sorted in descending order, and then by Lemma 4.3 we know there exists $h \in \mathfrak{S}(L, \mu, r^*)$ such that f in (1.1) is not r-factorizable. □

Consequently, to prove Theorem 1.3, it suffices to solve the optimization problem (4.16) analytically. This is very technically and contains less insight, and hence we defer it to Section A. Finally, let us note that Theorem 1.3 follows from Proposition 4.4, Proposition A.1 and Proposition A.3.

5 Generalization to the RIP case

When the condition in Theorem 1.3 fails, we can find a function $h \in \mathfrak{S}(L, \mu, r^*)$ such that the corresponding f in (1.1) is not r-factorizable. Since $\mathfrak{S}(L, \mu, r^*) \subseteq \mathfrak{S}(L, \mu, r, r^*)$, this means the condition in Theorem 1.3 is also necessary for the function class $\mathfrak{S}(L, \mu, r, r^*)$, and it suffices to prove sufficiency in Corollary 1.4. Our proof is largely motivated by the following observation:

Fact 5.1. *If $h \in C^2(\mathbb{R}^{m \times n})$ and satisfies* (1.7), *and the linear subspace $\mathcal{S} \subseteq \mathbb{R}^{m \times n}$ contains only matrices of rank at most $q + r^*$, then the restriction $h|_{\mathcal{S}}$ of h on $\mathcal{S}$ satisfies that $h|_{\mathcal{S}} \in C^2(\mathcal{S})$ is μ-strongly convex, and $\nabla h|_{\mathcal{S}}$ is L-Lipschitz continuous.*

Roughly speaking, to prove the sufficiency of the conditions in Theorem 1.3 for the function class $\mathfrak{S}(L, \mu, r, r^*)$, we argue by contradiction. Suppose there exists $h \in \mathfrak{S}(L, \mu, r, r^*)$ such that the corresponding f in (1.1) is not r-factorizable, then there exists a second-order stationary point $(\bar{U}, \bar{V}) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$ such that $\bar{X} = \bar{U}\bar{V}^\top$ is not a global minimizer of f. If we can find a linear subspace $\mathcal{S}$ such that $\mathcal{S}$ contains $\bar{X}$ and X^*, and $\mathcal{S}$ consists of matrices of rank no more than $r + r^*$, then we can apply Theorem 1.3 to arrive at a contradiction by restricting h and f on $\mathcal{S}$ and utilizing Fact 5.1. This is achieved by the following lemma.

Lemma 5.2. *Let $A, B \in \mathbb{R}^{m \times n}$ with $\operatorname{rank}(A) + \operatorname{rank}(B) = k$. Then, there exists $U \in \mathcal{O}_m$ and $V \in \mathcal{O}_n$ such that*

$$
A = U \widetilde{\operatorname{Diag}}(\sigma(A)) V^\top, \quad B = U \begin{bmatrix} B_1 & 0 \\ 0 & 0 \end{bmatrix} V^\top,
$$

where $B_1 \in \mathbb{R}^{k\times k}$.

Proof. Suppose $\begin{bmatrix} A & B \end{bmatrix} = U_1 D_1 V_1^\top$ and $\begin{bmatrix} A \\ B \end{bmatrix} = U_2 D_2 V_2^\top$ are the corresponding SVDs of $\begin{bmatrix} A & B \end{bmatrix}$ and $\begin{bmatrix} A \\ B \end{bmatrix}$, respectively. Notice that both rank $\begin{bmatrix} A & B \end{bmatrix}$ and rank $\begin{bmatrix} A \\ B \end{bmatrix}$ are no more than k. Then, the last $m-k$ rows of $U_1^\top A$ and $U_1^\top B$ are zero because $U_1^\top \begin{bmatrix} A & B \end{bmatrix} = D_1 V_1^\top$, and the last $n-k$ columns of AV_2 and BV_2 are 0 because $\begin{bmatrix} A \\ B \end{bmatrix} V_2 = U_2 D_2$. This implies that

$$U_1^\top A V_2 = \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad U_1^\top B V_2 = \begin{bmatrix} B_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad A_1, B_1 \in \mathbb{R}^{k\times k}.$$

Let an SVD of A_1 be given by $A_1 = U_3 D_3 V_3^\top$. Now, it suffices to take

$$U = U_1 \begin{bmatrix} U_3 & 0 \\ 0 & I_{m-k} \end{bmatrix} \quad \text{and} \quad V = V_2 \begin{bmatrix} V_3 & 0 \\ 0 & I_{n-k} \end{bmatrix}.$$

□

We are now ready to prove in the next proposition the sufficiency of the conditions in Theorem 1.3 for the function class $\mathfrak{S}(L,\mu,r,r^*)$. With Lemma 5.2 in hand, the remaining task is to ensure that we are also able to restrict F_r in (1.2) on a corresponding linear subspace to preserve the second-order stationarity of $(\bar U, \bar V)$, which is the key ingredient in the proof.

Proposition 5.3. *Assume* $h \in \mathfrak{S}(L,\mu,r,r^*)$ *and let* $X^* \in \mathbb{R}^{m\times n}$ *with* $\operatorname{rank}(X^*) = r^*$ *being a global minimizer of* f *in* (1.1)*. Set* $\kappa = \frac{L}{\mu} \geq 1$*. Assume that* F_r *in* (1.2) *has a second-order stationary point* $(\bar U, \bar V)$ *with* $\bar X := \bar U \bar V^\top \neq X^*$*, then both of the following conditions fail:*

(1) $r = m$*;*

(2) $r \geq r^*$ *and* $\min\{r, m-r^*\} > \frac{(\kappa-1)^2}{4}\min\{r^*, m-r\}$.

Proof. Let $k = r + r^*$. If $k \geq m$, then the result follows immediately from Theorem 1.3. Therefore, we assume $k < m$ in the following. Then $r < m$, which proves that $r = m$ is false. Since $k \geq \operatorname{rank}(\bar X) + \operatorname{rank}(X^*)$, we can apply Lemma 5.2 to show that there exist $R \in \mathcal{O}_m$ and $Q \in \mathcal{O}_n$ such that

$$R^\top X^* Q = \begin{bmatrix} X_1^* & 0 \\ 0 & 0 \end{bmatrix}, \quad R^\top \bar X Q = \begin{bmatrix} \bar X_1 & 0 \\ 0 & 0 \end{bmatrix},$$

where $X_1^*, \bar X_1 \in \mathbb{R}^{k\times k}$. Then $\begin{bmatrix} X_1^* & 0 \\ 0 & 0 \end{bmatrix}$ is a stationary point of $\tilde f$, where $\tilde f(W) := f(RWQ^\top) = \tilde h(W) + \lambda\|W\|_*$ with $\tilde h(W) := h(RWQ^\top)$. Clearly, we have $\tilde h \in \mathfrak{S}(L,\mu,r,r^*)$ and $\tilde F_r(U,V) := \tilde h(UV^\top) + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) = F_r(RU, QV)$ for all $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$. In particular, $(R^\top \bar U, Q^\top \bar V)$ is a second-order stationary point of $\tilde F_r$ with $R^\top \bar U \bar V^\top Q = R^\top \bar X Q \neq \begin{bmatrix} X_1^* & 0 \\ 0 & 0 \end{bmatrix}$. Consequently, replacing f by $\tilde f$ if necessary, we may assume that

$$X^* = \begin{bmatrix} X_1^* & 0 \\ 0 & 0 \end{bmatrix}, \quad \bar X = \begin{bmatrix} \bar X_1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{5.1}$$

Let us also note that since $\dim \mathcal{R}(\bar{U}) = \operatorname{rank}(\bar{U}) = \operatorname{rank}(\bar{V}) = \operatorname{rank}(\bar{X}) = \dim \mathcal{R}(\bar{X})$ by Remark 3.3(ii), and $\mathcal{R}(\bar{X}) \subseteq \mathcal{R}(\bar{U})$, we know that $\mathcal{R}(\bar{U}) = \mathcal{R}(\bar{X})$. This implies that $\bar{U} = \begin{bmatrix} \bar{U}_1 \\ 0 \end{bmatrix}$ with $\bar{U}_1 \in \mathbb{R}^{k\times r}$. Similar arguments on $\bar{X}^\top$ show that $\bar{V} = \begin{bmatrix} \bar{V}_1 \\ 0 \end{bmatrix}$ with $\bar{V}_1 \in \mathbb{R}^{k\times r}$.

Define a function $\widehat{f} : \mathbb{R}^{k\times k} \to \mathbb{R}$ and $\widehat{h} : \mathbb{R}^{k\times k} \to \mathbb{R}$ as:

$$\begin{aligned} \widehat{h}(K) &:= h\left(\begin{bmatrix} K & 0 \\ 0 & 0 \end{bmatrix}\right), \\ \widehat{f}(K) &:= f\left(\begin{bmatrix} K & 0 \\ 0 & 0 \end{bmatrix}\right) = \widehat{h}(K) + \lambda \left\| \begin{bmatrix} K & 0 \\ 0 & 0 \end{bmatrix} \right\|_* = \widehat{h}(K) + \lambda \|K\|_* \end{aligned}$$

The function $\widehat{f}$ can be viewed as the restriction of f to the linear subspace $\mathcal{S} := \begin{bmatrix} \mathbb{R}^{k\times k} & 0 \\ 0 & 0 \end{bmatrix}$. Using Fact 5.1, we know that $\widehat{h}$ is μ-strongly convex, and $\nabla \widehat{h}$ is Lipschitz continuous with modulus L. Next, observe that X_1^* is a global minimizer of $\widehat{f}$ and hence we have $0 \in \partial \widehat{f}(X_1^*)$. Since $\widehat{f}(\cdot) = \widehat{h}(\cdot) + \lambda \|\cdot\|_*$, we see that $\widehat{f}$ is μ-strongly convex, and hence X_1^* is the unique stationary point of $\widehat{f}$. Moreover, using (5.1), we know that $\operatorname{rank}(X_1^*) = \operatorname{rank}(X^*) = r^*$. Consequently, $\widehat{h} \in C^2(\mathbb{R}^{k\times k})$ satisfies that $\widehat{h} \in \mathfrak{S}(L, \mu, r^*)$. Next, by direct calculation, the corresponding $\widehat{F}_r$ in (1.2) can be written as:

$$\widehat{F}_r(U, V) := \widehat{h}(UV^\top) + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) = h\left(\begin{bmatrix} UV^\top & 0 \\ 0 & 0 \end{bmatrix}\right) + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right), \tag{5.2}$$

which can be viewed as the restriction of F_r on the linear subspace $\mathcal{S}_1 := \begin{bmatrix} \mathbb{R}^{k\times r} \\ 0 \end{bmatrix} \times \begin{bmatrix} \mathbb{R}^{k\times r} \\ 0 \end{bmatrix}$. Since $(\bar{U}, \bar{V}) \in \mathcal{S}_1$ is a second-order stationary point of F_r, we immediately see that $(\bar{U}_1, \bar{V}_1)$ is a second-order stationary point of $\widehat{F}_r$, since restricting a function on a linear subspace would not change second-order stationarity. The assumption $\bar{X} \neq X^*$ gives that $\bar{X}_1 \neq X_1^*$ given the representation in (5.1), and hence $\bar{X}_1$ is not a global minimizer of $\widehat{f}$, since $\widehat{f}$ is strongly convex and its unique global minimizer is X_1^*. Consequently, $\widehat{f}$ is not r-factorizable. By Theorem 1.3, we know that both of the following conditions fail:

(1') $r = k = r + r^*$;

(2') $r \geq r^*$ and $r > \frac{(\kappa-1)^2}{4} r^*$.

Since $r + r^* = k < m$, we know that $\min\{r, m - r^*\} = r$ and $\min\{r^*, m - r\} = r^*$. This proves that (2') is equivalent to (2), and the proof has been completed. □

A Proof of Theorem 1.3

In this section, we finish the key step of the proof of Theorem 1.3 by solving (4.16). We observe that the objective can be rewritten as a sum with the ith summand depending only on $(x_i, g_i, y_{\tau(i)}, v_{\tau(i)})$.[10] In addition, from the structure of the constraints in (4.16), we see that the terms $\{(x_i, g_i, y_{\tau(i)}, v_{\tau(i)})\}_{i\in[m]}$ can be divided into 4 groups depending on whether $i \in [r]$ and

[10] Specifically, notice that the fourth sum in the objective can be rewritten as $\sum_{i=1}^m (Ly_{\tau(i)} + v_{\tau(i)})(\mu y_{\tau(i)} + v_{\tau(i)})$.

$\tau(i) \in [r^*]$. These motivate the definitions of the next four associated index sets, for any fixed $\tau \in \mathfrak{P}_m$:

$$\mathcal{J}_1^\tau := [r] \cap \tau^{-1}[r^*],\ \mathcal{J}_2^\tau := [r] \setminus \mathcal{J}_1^\tau,\ \mathcal{J}_3^\tau := ([m] \setminus [r]) \cap \tau^{-1}[r^*],\ \mathcal{J}_4^\tau := ([m] \setminus [r]) \setminus \mathcal{J}_3^\tau. \tag{A.1}$$

Let $d_\tau := |\mathcal{J}_2^\tau|$. Then, by the definition of $\{\mathcal{J}_i^\tau\}_{i\in[4]}$ in (A.1), we have

$$|\mathcal{J}_1| = r - |\mathcal{J}_2^\tau| = r - d_\tau,\ |\mathcal{J}_3^\tau| = r^* - |\mathcal{J}_1^\tau| = r^* - r + d_\tau,\ |\mathcal{J}_4^\tau| = m - r - |\mathcal{J}_3^\tau| = m - r^* - d_\tau. \tag{A.2}$$

To solve (4.16), our strategy is to introduce an auxiliary variable $w \in \mathbb{R}$ to transform (4.16) to the next equivalent form:

$$\begin{aligned}
\sup_{\substack{x,g,y,v\in\mathbb{R}^m \\ \tau\in\mathfrak{P}_m, w\in\mathbb{R}}} \ & \sum_{i=1}^m (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i + g_i)(L y_{\tau(i)} + v_{\tau(i)}) \\
& - \sum_{i=1}^m (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^m (L y_i + v_i)(\mu y_i + v_i) \\
s.t. \quad & \forall i \in [r],\ x_i \geq w > 0,\ g_i = \lambda, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = \lambda, \\
& \forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, \lambda + Lw], \\
& \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, \lambda].
\end{aligned} \tag{A.3}$$

Problem (4.16) is equivalent to (A.3) in the following sense: for any feasible solution (x, g, y, v, τ) of (4.16), $(x, g, y, v, \tau, \min_{i\in[r]} x_i)$ is a feasible solution of (4.16) having the same objective function value; for any feasible solution (x, g, y, v, τ, w) of (A.3), (x, g, y, v, τ) is a feasible solution of (4.16) having the same objective function value. Consequently, we have the following result.

Proposition A.1. *There exists a feasible solution (x, g, y, v, τ) with $\|g\|_\infty > \lambda$ to* (4.16) *having nonnegative objective function value if and only if there exists a feasible solution (x, g, y, v, τ, w) with $\|g\|_\infty > \lambda$ to* (A.3) *having nonnegative objective function value.*

Next, we plan to fix τ and w to analyze the optimal value and the optimal solution of (A.3). The reason to do so is that the optimization problem can be made separable when τ and w are fixed. To simplify the calculation, we only consider the case where $\lambda = 1$ in the next lemma.

Lemma A.2. *Let $r^* \in [m] \cup \{0\}$, $r \in [m]$, and $\infty > L \geq \mu > 0$. Let $\tau \in \mathfrak{P}_m$ and $w > 0$. Let $\{\mathcal{J}_i^\tau\}_{i\in[4]}$ be defined in* (A.1)*. Consider the following optimization problem:*

$$\begin{aligned}
\sup_{x,g,y,v\in\mathbb{R}^m} \ & \sum_{i=1}^m (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i + g_i)(L y_{\tau(i)} + v_{\tau(i)}) \\
& - \sum_{i=1}^m (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^m (L y_i + v_i)(\mu y_i + v_i) \\
s.t. \quad & \forall i \in [r],\ x_i \geq w,\ g_i = 1, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = 1, \\
& \forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, 1 + Lw], \quad \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, 1].
\end{aligned} \tag{A.4}$$

Then, the optimization problem (A.4) *has optimal solutions, and the optimal value is*

$$\left(-L\mu |\mathcal{J}_2^\tau| + |\mathcal{J}_3^\tau| \frac{L(L-\mu)^2}{4\mu} \right) w^2. \tag{A.5}$$

Moreover, the following statements are equivalent:

- $|\mathcal{J}_3^\tau| > 0$.
- *For all the optimal solutions* $(\bar{x}, \bar{g}, \bar{y}, \bar{v})$ *of* (A.4)*, we have* $\|\bar{g}\|_\infty > 1$.
- *There exists one optimal solution* $(\bar{x}, \bar{g}, \bar{y}, \bar{v})$ *of* (A.4) *such that* $\|\bar{g}\|_\infty > 1$.

Proof. Using the definition of permutation, we notice that

$$\sum_{i=1}^{m}(Ly_i + v_i)(\mu y_i + v_i) = \sum_{i=1}^{m}(Ly_{\tau(i)} + v_{\tau(i)})(\mu y_{\tau(i)} + v_{\tau(i)}). \tag{A.6}$$

Substituting (A.6) into (A.4), we get that

$$\begin{aligned} \sup_{x,g,y,v\in\mathbb{R}^m} \; & \sum_{i=1}^{m}(Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^{m}(\mu x_i + g_i)(Ly_{\tau(i)} + v_{\tau(i)}) \\ & - \sum_{i=1}^{m}(Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^{m}(Ly_{\tau(i)} + v_{\tau(i)})(\mu y_{\tau(i)} + v_{\tau(i)}) \\ s.t. \quad & \forall i \in [r],\ x_i \geq w,\ g_i = 1, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = 1, \\ & \forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, 1 + Lw], \quad \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, 1]. \end{aligned} \tag{A.7}$$

Observe that (A.7) can be decomposed into the next m subproblems for each $i \in [m]$:

$$\begin{aligned} \sup_{x_i,g_i,y_i,v_i\in\mathbb{R}} \quad & (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + (\mu x_i + g_i)(Ly_{\tau(i)} + v_{\tau(i)}) \\ & - (Lx_i + g_i)(\mu x_i + g_i) - (Ly_{\tau(i)} + v_{\tau(i)})(\mu y_{\tau(i)} + v_{\tau(i)}) \\ s.t. \quad & \begin{cases} x_i \geq w,\ g_i = 1,\ y_{\tau(i)} > 0,\ v_{\tau(i)} = 1, & \text{if } i \in \mathcal{J}_1^\tau, \\ x_i \geq w,\ g_i = 1,\ y_{\tau(i)} = 0,\ v_{\tau(i)} \in [0,1], & \text{if } i \in \mathcal{J}_2^\tau, \\ x_i = 0,\ g_i \in [0, 1 + Lw],\ y_{\tau(i)} > 0,\ v_{\tau(i)} = 1, & \text{if } i \in \mathcal{J}_3^\tau, \\ x_i = 0,\ g_i \in [0, 1 + Lw],\ y_{\tau(i)} = 0,\ v_{\tau(i)} \in [0,1], & \text{if } i \in \mathcal{J}_4^\tau, \end{cases} \end{aligned} \tag{A.8}$$

where we recall the definition of $\{\mathcal{J}_i^\tau\}_{i\in[4]}$ in (A.1). We now consider the solution and optimal value of each subproblem (A.8) for fixed i:

1. $i \in \mathcal{J}_1^\tau$. Then (A.8) takes the following form:

$$\begin{aligned} \sup_{x_i, y_{\tau(i)}} \; & (Lx_i + 1)(\mu y_{\tau(i)} + 1) + (\mu x_i + 1)(Ly_{\tau(i)} + 1) \\ & - (Lx_i + 1)(\mu x_i + 1) - (Ly_{\tau(i)} + 1)(\mu y_{\tau(i)} + 1) \\ s.t. \; & x_i \geq w,\ y_{\tau(i)} > 0, \end{aligned} \tag{A.9}$$

   where we used the fact that g_i and $v_{\tau(i)}$ are 1. Denote the objective function of (A.9) by S_1. By direct calculation, we can rewrite $S_1(x_i, y_{\tau(i)})$ as:

$$S_1(x_i, y_{\tau(i)}) = -L\mu(x_i - y_{\tau(i)})^2.$$

   Clearly, the optimal value of (A.9) is 0, and it is achieved if and only if

$$x_i = y_{\tau(i)} \geq w. \tag{A.10}$$

2. $i \in \mathcal{J}_2^\tau$. Then (A.8) takes the following form:

$$\begin{aligned} \sup_{x_i, v_{\tau(i)}} \ & (Lx_i + 1)v_{\tau(i)} + (\mu x_i + 1)v_{\tau(i)} - (Lx_i + 1)(\mu x_i + 1) - v_{\tau(i)}^2 \\ s.t. \ \ & x_i \geq w, \ v_{\tau(i)} \in [0, 1], \end{aligned} \tag{A.11}$$

where we used the fact that g_i and $y_{\tau(i)}$ are 1 and 0, respectively. Denote the objective function of (A.11) by S_2. By direct calculation, we can rewrite $S_2(x_i, v_{\tau(i)})$ as:

$$S_2(x_i, v_{\tau(i)}) = -L\mu x_i^2 + (L + \mu)x_i(v_{\tau(i)} - 1) - (v_{\tau(i)} - 1)^2.$$

First, we notice that S_2 is strictly decreasing on $[0, \infty)$ as a function of x_i when $v_{\tau(i)}$ is fixed to be any value in $[0, 1]$. This means that

$$\sup_{x_i \geq w} S_2(x_i, v_{\tau(i)}) = -L\mu w^2 + (L + \mu)w(v_{\tau(i)} - 1) - (v_{\tau(i)} - 1)^2, \tag{A.12}$$

where the optimal value is achieved if and only if $x_i = w$. Let $\tilde{S}$ denote the function on the right hand side of (A.12). Then we see $\tilde{S}$ is strictly increasing as a function of $v_{\tau(i)}$ on $(-\infty, 1]$ by using the elementary properties of quadratic functions. Therefore, the optimal value of (A.11) is $-L\mu w^2$, and it is achieved if and only if

$$x_i = w, \ v_{\tau(i)} = 1. \tag{A.13}$$

3. $i \in \mathcal{J}_3^\tau$. Then (A.8) has the following form:

$$\begin{aligned} \sup_{g_i, y_{\tau(i)}} \quad & g_i(\mu y_{\tau(i)} + 1) + g_i(Ly_{\tau(i)} + 1) - g_i^2 - (Ly_{\tau(i)} + 1)(\mu y_{\tau(i)} + 1) \\ s.t. \quad & g_i \in [0, 1 + Lw], \ y_{\tau(i)} > 0, \end{aligned} \tag{A.14}$$

where we used the fact that x_i and $v_{\tau(i)}$ are 0 and 1, respectively. Denote the objective function of (A.14) by S_3. By direct calculation we can rewrite S_3 as follows

$$\begin{aligned} S_3(g_i, y_{\tau(i)}) &= \frac{(L - \mu)^2}{4L\mu}(g_i - 1)^2 - L\mu \left(y_{\tau(i)} - \frac{(L + \mu)(g_i - 1)}{2L\mu} \right)^2 \\ &= \frac{L(L - \mu)^2 w^2}{4\mu} + \frac{(L - \mu)^2}{4L\mu}(g_i - (1 + Lw))(Lw + g_i - 1) \\ &\quad - L\mu \left(y_{\tau(i)} - \frac{(L + \mu)(g_i - 1)}{2L\mu} \right)^2 . \end{aligned} \tag{A.15}$$

Notice that $g_i - (1 + Lw) \leq 0$ and $Lw + g_i - 1 > 0$ when $g_i \in (1, 1 + Lw]$. We can thus see from the second expression in the above display that the optimal value of S_3 when $g_i \in (1, 1 + Lw]$ is $\frac{L(L-\mu)^2 w^2}{4\mu}$; moreover, when $L > \mu$, the optimal value is achieved if and only if

$$g_i = 1 + Lw, \ y_{\tau(i)} = \frac{(L + \mu)w}{2\mu}, \tag{A.16}$$

while when $L = \mu$, the optimal value is achieved if and only if

$$g_i \in (1, 1 + Lw], \ y_{\tau(i)} = \frac{(L + \mu)(g_i - 1)}{2L\mu}. \tag{A.17}$$

On the other hand, when $g_i \leq 1$, notice that $y_{\tau(i)} > 0$, and hence $y_{\tau(i)} - \frac{(L+\mu)(g_i-1)}{2L\mu} > |\frac{(L+\mu)(g_i-1)}{2L\mu}|$. Then we have from the first expression of S_3 in (A.15) that

$$S_3(g_i, y_{\tau(i)}) < \frac{(L-\mu)^2}{4L\mu}(g_i - 1)^2 - L\mu \left(\frac{(L+\mu)(g_i-1)}{2L\mu} \right)^2 = -(g_i - 1)^2 \leq 0.$$

Consequently, the optimal value of (A.14) is $\frac{L(L-\mu)^2 w^2}{4\mu}$, and is achieved as described in (A.16) and (A.17).

4. $i \in \mathcal{J}_4^\tau$. Then (A.8) has the following form:

$$\begin{aligned} \sup_{g_i, v_{\tau(i)}} \quad & 2g_i v_{\tau(i)} - g_i^2 - v_{\tau(i)}^2 \\ s.t. \quad & g_i \in [0, 1 + Lw],\ v_{\tau(i)} \in [0, 1], \end{aligned} \tag{A.18}$$

where we used the fact that x_i and $y_{\tau(i)}$ are 0. Notice that the objective of the above problem is $-(g_i - v_{\tau(i)})^2$. Clearly, the optimal value of (A.18) is 0, and is achieved if and only if

$$g_i = v_{\tau(i)} \in [0, 1]. \tag{A.19}$$

Consequently, by the solution sets given in (A.10), (A.13), (A.16), (A.17) and (A.19), we know the solution set of (A.4) is nonempty. The optimal value is obtained by summing all the optimal values given in the four cases. Moreover, every solution $(\bar{x}, \bar{g}, \bar{y}, \bar{v})$ of (A.4) satisfies $\|\bar{g}\|_\infty > 1$ if and only if $|\mathcal{J}_3^\tau| > 0$, according to the structure of $\bar{g}$ given in (A.16), (A.17) and (A.19). □

Proposition A.3. *Let $r^* \in [m] \cup \{0\}, r \in [m]$, and $\infty > L \geq \mu > 0$. Let G be the objective function of* (A.3) *and let $\kappa := \frac{L}{\mu} \geq 1$. If r^*, r and κ satisfy any of the following conditions, then there is no feasible $(\bar{x}, \bar{g}, \bar{y}, \bar{v}, \bar{\tau}, \bar{w})$ to* (A.3) *satisfying $\|\bar{g}\|_\infty > \lambda$ and $G(\bar{x}, \bar{g}, \bar{y}, \bar{v}, \bar{\tau}, \bar{w}) \geq 0$.*

(1) $r = m$.

(2) $r \geq r^*$ and $\min\{r, m - r^*\} > \frac{(\kappa-1)^2}{4} \min\{r^*, m - r\}$.

Otherwise, such a feasible solution exists.

Proof. By the change of variables $(x, g, y, v, \tau, w) \leftarrow (x/\lambda, g/\lambda, y/\lambda, v/\lambda, \tau, w/\lambda)$, we see that (A.3) can be reduced to the following problem:

$$\begin{aligned} \sup_{\substack{x,g,y,v \in \mathbb{R}^m \\ \tau \in \mathfrak{P}_m, w \in \mathbb{R}}} \quad & \lambda^2 \Bigg[\sum_{i=1}^m (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i + g_i)(L y_{\tau(i)} + v_{\tau(i)}) \\ & \qquad - \sum_{i=1}^m (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^m (Ly_i + v_i)(\mu y_i + v_i) \Bigg] \\ s.t. \quad & \forall i \in [r],\ x_i \geq w > 0,\ g_i = 1, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = 1, \\ & \forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, 1 + Lw], \\ & \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, 1]. \end{aligned} \tag{A.20}$$

Since dropping the constant λ^2 won't affect the sign of the function value and our claim only concerns the feasible set of (A.3) and the *sign* of its objective value, we shall consider the following optimization problem instead:

$$\begin{aligned}\sup_{\substack{x,g,y,v\in\mathbb{R}^m\\ \tau\in\mathfrak{P}_m, w\in\mathbb{R}}} & \sum_{i=1}^m (Lx_i+g_i)(\mu y_{\tau(i)}+v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i+g_i)(Ly_{\tau(i)}+v_{\tau(i)}) \\ & - \sum_{i=1}^m (Lx_i+g_i)(\mu x_i+g_i) - \sum_{i=1}^m (Ly_i+v_i)(\mu y_i+v_i) \\ s.t.\ & \forall i\in[r],\ x_i\geq w>0,\ g_i=1, \quad \forall i\in[r^*],\ y_i>0,\ v_i=1, \\ & \forall i\in[m]\setminus[r],\ x_i=0,\ g_i\in[0,1+Lw], \\ & \forall i\in[m]\setminus[r^*],\ y_i=0,\ v_i\in[0,1].\end{aligned} \tag{A.21}$$

Notice that when τ and w are fixed, (A.21) becomes (A.4). Applying Lemma A.2, setting $d_\tau=|\mathcal{J}_2^\tau|$ and recalling the definition of $\mathcal{J}_3^\tau$ in (A.2), we see that the solution set $\Omega_{w,\tau}$ of (A.4) is nonempty, and for all $(\bar{x},\bar{g},\bar{y},\bar{v})\in\Omega_{w,\tau}$, it holds that $\|\bar{g}\|_\infty>1$ if and only if $r^*-r+d_\tau>0$.

Next, define the following function $H:\mathbb{N}_0\times\mathbb{R}_+\to\mathbb{R}$:

$$H(d,w):=\left(-L\mu d+(r^*-r+d)\frac{L(L-\mu)^2}{4\mu}\right)w^2. \tag{A.22}$$

Then in view of (A.2) and (A.5), the optimal value of (A.4) is given by $H(|\mathcal{J}_2^\tau|,w)$. Moreover, we see that (A.21) is equivalent to the following problem

$$\sup_{w\in\mathbb{R},\ d\in\mathbb{N}_0} H(d,w) \quad s.t.\ w>0,\ r-r^*\leq d\leq\min\{r,m-r^*\}, \tag{A.23}$$

where the bound for d comes from the requirement that $|\mathcal{J}_i^\tau|\geq 0$ for $i\in[4]$ in (A.2).

We consider the following scenarios:

(S1) The optimal value of (A.23) is nonpositive, and (A.23) has no feasible solution (d,w) satisfying $H(d,w)\geq 0$ and $r^*-r+d>0$.

In this scenario, we claim that (A.3) has no feasible solution $(\bar{x},\bar{g},\bar{y},\bar{v},\bar{\tau},\bar{w})$ with $\|\bar{g}\|_\infty>\lambda$ and $G(\bar{x},\bar{g},\bar{y},\bar{v},\bar{\tau},\bar{w})\geq 0$, where G is the objective of (A.3). A short proof is provided below.

Suppose such a feasible solution $(\bar{x},\bar{g},\bar{y},\bar{v},\bar{\tau},\bar{w})$ of (A.3) exists, then either $(\bar{x},\bar{g},\bar{y},\bar{v})/\lambda$ is optimal for (A.4) with $w=\bar{w}/\lambda$ and $\tau=\bar{\tau}$, or $(\bar{x},\bar{g},\bar{y},\bar{v})/\lambda$ is not optimal. In the latter case, the optimal value of (A.23) must be positive. In the former case, we see from Lemma A.2 that $|\mathcal{J}_3^{\bar{\tau}}|>0$, and hence (A.23) has a feasible solution $(\tilde{d},\tilde{w})=(|\mathcal{J}_2^{\bar{\tau}}|,\bar{w}/\lambda)$ with $H(\tilde{d},\tilde{w})\geq 0$ and $r^*-r+\tilde{d}=r^*-r+|\mathcal{J}_2^{\bar{\tau}}|=|\mathcal{J}_3^{\bar{\tau}}|>0$ (see (A.2)). Both cases yield a contradiction.

(S2) There exists a feasible solution $(d,\tilde{w})$ of (A.23) satisfying $H(d,\tilde{w})\geq 0$ and $r^*-r+d>0$.

In this scenario, (A.3) has a feasible solution $(\bar{x},\bar{g},\bar{y},\bar{v},\bar{\tau},\bar{w})$ with $G(\bar{x},\bar{g},\bar{y},\bar{v},\bar{\tau},\bar{w})\geq 0$ and $\|\bar{g}\|_\infty>\lambda$. Indeed, we just need to take $\bar{\tau}\in\mathfrak{P}_m$ satisfying $|\mathcal{J}_2^{\bar{\tau}}|=d$, and then take $(\tilde{x},\tilde{g},\tilde{y},\tilde{v})$ to be the optimal solution of (A.4) with $\tau=\bar{\tau}$ and $w=\tilde{w}$, and set $(\bar{x},\bar{g},\bar{y},\bar{v},\bar{w})=\lambda(\tilde{x},\tilde{g},\tilde{y},\tilde{v},\tilde{w})$.

We note that the classification in (S1) and (S2) is *not* complete, since we cannot say anything if the optimal value of (A.23) is positive and there is no feasible solution (d,w) of (A.23) satisfying $H(d,w)\geq 0$ and $r^*-r+d>0$. Nevertheless, the two scenarios in (S1) and (S2) are enough for our proof. Consider the following cases on r, r^* and $\kappa:=L/\mu$.

Case 1: $r = m$ or $r^* = 0$. If $r = m$, then every feasible point (d, w) to (A.23) must satisfy $d = r - r^*$ and $H(d, w) = -L\mu d w^2 \leq 0$. Hence, (S1) holds. If $r^* = 0$, we see that every feasible point (d, w) to (A.23) must satisfy $d = r$. Then, in view of (A.22), we can rewrite (A.23) as:

$$\sup_{w>0} -L\mu r w^2.$$

This means that every feasible solution of (A.23) has a negative objective function value. Then (S1) holds.

Case 2: $r < r^*$. Setting $d = 0$ and selecting $w > 0$, we see that

$$H(d, w) = \left(-L\mu d + (r^* - r + d)\frac{L(L-\mu)^2}{4\mu}\right) w^2 = (r^* - r)\frac{L(L-\mu)^2}{4\mu} w^2 \geq 0,$$

and $r^* - r + d > 0$. Then (S2) holds.

Case 3: $m > r \geq r^*$ and $L = \mu$ (*i.e.*, $\kappa = 1$). Then, we have

$$H(d, w) = \left(-L\mu d + (r^* - r + d)\frac{L(L-\mu)^2}{4\mu}\right) w^2 = -L^2 d w^2.$$

Then the optimal value of (A.23) is 0, and for all feasible solution (d, w) of (A.23) with $r^* - r + d > 0$ it holds that $d > r - r^* \geq 0$, and hence $H(d, w) < 0$. Then (S1) holds.

Case 4: $m > r \geq r^* > 0$ and $L > \mu$ (*i.e.*, $\kappa > 1$). If $\kappa = \frac{L}{\mu} = 3$, then we have $-L\mu + \frac{L(L-\mu)^2}{4\mu} = L\mu(\frac{(\kappa-1)^2}{4} - 1) = 0$. Therefore, (A.23) is equivalent to that:

$$\sup_{w\in\mathbb{R},\ d\in\mathbb{N}_0} (r^* - r)\frac{L(L-\mu)^2}{4\mu} w^2 \qquad s.t.\ w > 0,\ r - r^* \leq d \leq \min\{r, m - r^*\}.$$

In this case, we clearly see that the optimal value of (A.23) is 0, and is achievable if and only if $r = r^*$. If $r > r^*$, then (S1) holds. If $r = r^*$, then any feasible solution to (A.23) is optimal. Since $r = r^* < m$, for $\hat{d} := \min\{r, m - r^*\}$ and any $w > 0$, the point $(w, \hat{d})$ is feasible (because $\min\{r, m - r^*\} > 0$), and in this case we have $r^* - r + \hat{d} = \hat{d} > 0$, which means (S2) holds. Next we assume $\kappa \neq 3$. Let $\alpha := \frac{r - r^*}{1 - \frac{4}{(\kappa-1)^2}} = \frac{r - r^*}{1 - \frac{4\mu^2}{(L-\mu)^2}}$. We now rewrite H in (A.22) as:

$$\begin{aligned}
H(d, w) &= \left(-L\mu d + (r^* - r + d)\frac{L(L-\mu)^2}{4\mu}\right) w^2 \\
&= \frac{L(L-\mu)^2}{4\mu}\left(\frac{-4\mu^2 d}{(L-\mu)^2} + r^* - r + d\right) w^2 = \frac{L(L-\mu)^2}{4\mu}\left(\frac{-4d}{(\kappa-1)^2} + r^* - r + d\right) w^2 \\
&= \frac{L(L-\mu)^2}{4\mu}\left(\left(1 - \frac{4}{(\kappa-1)^2}\right) d + r^* - r\right) w^2 = \frac{L(L-\mu)^2}{4\mu}\left(1 - \frac{4}{(\kappa-1)^2}\right)(d - \alpha) w^2.
\end{aligned}$$

Then, we can rewrite (A.23) as:

$$\begin{aligned}
\sup_{w\in\mathbb{R},\ d\in\mathbb{N}_0} \quad & \frac{L(L-\mu)^2}{4\mu}\left(1 - \frac{4}{(\kappa-1)^2}\right)(d - \alpha) w^2 \\
s.t. \quad & w > 0,\ r - r^* \leq d \leq \min\{r, m - r^*\},\ \alpha = \frac{r - r^*}{1 - \frac{4}{(\kappa-1)^2}}.
\end{aligned}$$

If $\kappa < 3$ and $r \geq r^*$, then $\alpha \leq 0$, and we see that the optimal value of (A.23) is 0. Moreover, for all feasible solution (d, w) of (A.23) with $r^* - r + d > 0$, we have $d > r - r^* \geq 0$ and $H(d, w) < 0$. Then (S1) holds.

Finally, we assume $\kappa > 3$. If $\alpha > \min\{r, m - r^*\}$, then the optimal value of (A.23) is 0, and for all (d, w) that is feasible to (A.23), it holds that $H(d, w) < 0$. Then (S1) holds. If $\alpha \leq \min\{r, m - r^*\}$, then we can select $d = \min\{r, m - r^*\}$ and any $w > 0$, which is feasible for (A.23) and $r^* - r + d = \min\{r^*, m - r\} > 0$, and we have $H(d, w) \geq 0$. Then (S2) holds.

In summary, note that we have argued that we have either (S1) or (S2). Moreover, (S1) holds if and only if any of the following is true, and (S2) holds otherwise.

(1) $r = m$ or $r^* = 0$.

(2) $m > r \geq r^* > 0$ and $\kappa = 1$.

(3) $m > r \geq r^* > 0$, $\kappa > 1$, $\kappa = 3$ and $r > r^*$.

(4) $m > r \geq r^* > 0$, $\kappa > 1$, $\kappa < 3$.

(5) $m > r \geq r^* > 0$, $\kappa > 1$, $\kappa > 3$ and $\alpha > \min\{r, m - r^*\}$.

Upon integrating (1) into the other conditions and regrouping (2), (3) and (4), we can further rewrite the above conditions as follows:

(1) $r = m$ or $r^* = 0$.

(2) $r > r^*$ and $\kappa = 3$.

(3) $r \geq r^*$ and $\kappa < 3$.

(4) $r \geq r^*$, $\kappa > 3$ and $\alpha > \min\{r, m - r^*\}$.

Next, notice that the condition $r^* = 0$ is not ruled out by (2), (3) and (4); in particular, observe that the condition $\alpha > \min\{r, m - r^*\}$ holds trivially when $\kappa > 3$ and $r^* = 0$, because $r \geq 1$. Thus, we can further rewrite the above conditions as follows:

(1) $r = m$.

(2) $r > r^*$ and $\kappa = 3$.

(3) $r \geq r^*$ and $\kappa < 3$.

(4) $r \geq r^*$, $\kappa > 3$ and $\alpha > \min\{r, m - r^*\}$.

Finally, we notice that $r - r^* = \min\{r, m - r^*\} - \min\{r^*, m - r\}$, which can be proved by discussing the two cases $r + r^* \leq m$ and $r + r^* > m$ separately. Then, we have $\alpha = (\min\{r, m - r^*\} - \min\{r^*, m - r\})/(1 - \frac{4}{(\kappa-1)^2})$. Thus, when $\kappa > 3$, we can rewrite the condition $\alpha > \min\{r, m - r^*\}$ as

$$\min\{r, m - r^*\} > \frac{(\kappa - 1)^2}{4} \min\{r^*, m - r\}. \tag{A.24}$$

Moreover, when $\kappa = 3$, the condition (A.24) is the same as $\min\{r, m - r^*\} - \min\{r^*, m - r\} > 0$, i.e., $r > r^*$. Furthermore, when $\kappa < 3$ and $r^* \leq r < m$, we have $\frac{(\kappa-1)^2}{4} < 1$ and hence we always have

$$\min\{r, m - r^*\} - \min\{r^*, m - r\} = r - r^* > \left(\frac{(\kappa - 1)^2}{4} - 1\right) \min\{r^*, m - r\};$$

this is because the strict inequality holds trivially when $r > r^*$, and if $r = r^*$, we have $\min\{r^*, m - r\} = \min\{r, m-r\} > 0$, which also implies the strict inequality above. Hence, we see from the above display that (A.24) holds trivially when $\kappa < 3$ and $r^* \leq r < m$. Consequently, the above four cases now can be recapped as the following two cases:

(1) $r = m$.

(2) $r \geq r^*$ and $\min\{r, m - r^*\} > \frac{(\kappa-1)^2}{4} \min\{r^*, m - r\}$.

□

References

[1] Daniel Azagra and Carlos Mudarra. Whitney extension theorems for convex functions of the classes C^1 and $C^{1,\omega}$. *Proceedings of the London Mathematical Society*, 114(1):133–158, 2017.

[2] Jonathan Borwein and Adrian Lewis. *Convex Analysis*. Springer, 2006.

[3] Nicolas Boumal and Andrew D. McRae. The usual smooth lift of the nuclear norm regularizer enjoys 2⇒1. `www.racetothebottom.xyz/posts/lift-regularizer-nuclear/`.

[4] Nicolas Boumal, Vladislav Voroninski, and Afonso S. Bandeira. The non-convex Burer–Monteiro approach works on smooth semidefinite programs. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NeurIPS'16, page 2765–2773, Red Hook, NY, USA, 2016. Curran Associates Inc.

[5] Nicolas Boumal, Vladislav Voroninski, and Afonso S Bandeira. Deterministic guarantees for Burer-Monteiro factorizations of smooth semidefinite programs. *Communications on Pure and Applied Mathematics*, 73(3):581–608, 2020.

[6] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[7] Samuel Burer and Renato DC Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

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

[9] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

[10] Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NeurIPS'16, page 2981–2989, Red Hook, NY, USA, 2016. Curran Associates Inc.

[11] Wooseok Ha, Haoyang Liu, and Rina Foygel Barber. An equivalence between critical points for rank constraints versus low-rank factorizations. *SIAM Journal on Optimization*, 30(4):2927–2955, 2020.

[12] Zaid Harchaoui, Matthijs Douze, Mattis Paulin, Miroslav Dudik, and Jérôme Malick. Large-scale image classification with trace-norm regularization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3386–3393. IEEE, 2012.

[13] Cho-Jui Hsieh, Kai-Yang Chiang, and Inderjit S Dhillon. Low rank modeling of signed networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 507–515, 2012.

[14] Yue Hu, Xiaohan Liu, and Mathews Jacob. A generalized structured low-rank matrix completion algorithm for mr image recovery. *IEEE Transactions on Medical Imaging*, 38(8):1841–1851, 2018.

[15] Michel Journée, Francis Bach, P-A Absil, and Rodolphe Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

[16] Junsu Kim, Jaeyeon Kim, and Ernest K Ryu. LoRA training provably converges to a low-rank global minimum or it fails loudly (but it probably won't fail). *arXiv preprint arXiv:2502.09376*, 2025.

[17] Ching-pei Lee, Ling Liang, Tianyun Tang, and Kim-Chuan Toh. Accelerating nuclear-norm regularized low-rank matrix optimization through Burer-Monteiro decomposition. *Journal of Machine Learning Research*, 25(379):1–52, 2024.

[18] Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

[19] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. Geometry of factored nuclear norm regularization. *arXiv preprint arXiv:1704.01265*, 2017.

[20] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2019.

[21] Yi-Kai Liu. Universal low-rank matrix recovery from pauli measurements. *Advances in Neural Information Processing Systems*, 24, 2011.

[22] Andrew D McRae. Low solution rank of the matrix lasso under rip with consequences for rank-constrained algorithms: Ad mcrae. *Mathematical Programming*, pages 1–25, 2025.

[23] Karthik Mohan and Maryam Fazel. Reweighted nuclear norm minimization with application to system identification. In *Proceedings of the 2010 American Control Conference*, pages 2953–2959. IEEE, 2010.

[24] Jacob Munson, Breschine Cummins, and Dominique Zosso. An introduction to collaborative filtering through the lens of the Netflix Prize. *Knowledge and Information Systems*, pages 1–50, 2025.

[25] Yurii Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

[26] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

[27] Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74. PMLR, 2017.

[28] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International conference on machine learning*, pages 964–973. PMLR, 2016.

[29] G Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and Its Applications*, 170(1):33–45, 1992.

[30] Bo Wen, Xiaojun Chen, and Ting Kei Pong. Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *SIAM Journal on Optimization*, 27(1):124–145, 2017.

[31] Man-Chung Yue, Zirui Zhou, and Anthony Man-Cho So. A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo–Tseng error bound property. *Mathematical Programming*, 174(1):327–358, 2019.

[32] Gavin Zhang, Salar Fattahi, and Richard Y Zhang. Preconditioned gradient descent for overparameterized nonconvex Burer–Monteiro factorization with global optimality certification. *Journal of Machine Learning Research*, 24(163):1–55, 2023.

[33] Haixiang Zhang, Yingjie Bi, and Javad Lavaei. General low-rank matrix optimization: geometric analysis and sharper bounds. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NeurIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

[34] Hongyan Zhang, Wei He, Liangpei Zhang, Huanfeng Shen, and Qiangqiang Yuan. Hyperspectral image restoration using low-rank matrix recovery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8):4729–4743, 2013.

[35] Richard Y Zhang. Improved global guarantees for the nonconvex Burer–Monteiro factorization via rank overparameterization. To appear in *Mathematical Programming*, 2024.

[36] Richard Y Zhang. Sharp global guarantees for nonconvex low-rank recovery in the noisy overparameterized regime. *SIAM Journal on Optimization*, 35(3):2128–2154, 2025.

[37] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.

[38] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. The global optimization geometry of low-rank matrix optimization. *IEEE Transactions on Information Theory*, 67(2):1308–1331, 2021.