Burer-Monteiro factorizability of nuclear norm regularized optimization

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

May 21, 2025

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, and its factorized problem obtained by the Burer-Monteiro technique. We first prove that every second-order stationary point of the factorized problem corresponds to an approximate stationary point of its non-factorized counterpart, and those rank-deficient ones correspond to global minimizers of the latter problem when h is additionally convex, conforming with the observations in [2, 15]. Next, discarding the rank condition on the second-order stationary points but assuming the convexity and Lipschitz differentiability of h, we *characterize*, with respect to some natural problem parameters, when every second-order stationary point of the factorized problem is a global minimizer of the corresponding nuclear norm-regularized problem. More precisely, we subdivide the class of Lipschitz differentiable convex C^2 functions into subclasses according to those natural parameters and characterize when each subclass consists solely of functions h such that every second-order stationary point of the associated factorized model is a global minimizer of the nuclear norm regularized model. In particular, explicit counterexamples are established when the characterizing condition on the said parameters is violated.

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 [12], machine learning [11, 26], quantum physics [22], control [25] and imaging [43, 13], to name a few. Natural formulations of the problem include the rank-constrained minimization problem (*e.g.*, see [21, 46]):

$$\begin{array}{cl}\min\limits_{X\in\mathbb{R}^{m\times n}} & h(X)\\ \text{s.t.} & \operatorname{rank}(X)\le r,\end{array}\tag{1.1}$$

the constrained rank minimization problem (*e.g.*, see [8, 39]):

$$\begin{array}{cl}\min\limits_{X\in\mathbb{R}^{m\times n}} & \operatorname{rank}(X)\\ \text{s.t.} & h(X)\le c,\end{array}\tag{1.2}$$

[*] 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 PolyU153001/22p and the PolyU internal grant 4-ZZPJ.

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

or the rank-regularized minimization problem (*e.g.*, see [23, 20]):

$$\min_{X \in \mathbb{R}^{m \times n}} h(X) + \lambda \cdot \operatorname{rank}(X). \tag{1.3}$$

In the above formulations, $r, c, \lambda > 0$ are constants prescribed by the modelers and h is a function representing the misfit between the output predicted by X and the true observations.

Unfortunately, due to the non-convexity and combinatorial nature of the rank function, the optimization problems (1.1), (1.2) and (1.3) are difficult to solve in general. For computational tractability, many convex and non-convex surrogates were proposed [33, 35, 27, 39]. This paper focuses on the following surrogate:

$$\min_{X \in \mathbb{R}^{m \times n}} f(X) := h(X) + \lambda \|X\|_*, \tag{1.4}$$

where h is assumed to be twice continuously differentiable for the theoretical analysis, $\lambda > 0$, and $\|\cdot\|_*$ denotes the nuclear norm. Problem (1.4) can be seen as an approximation to problem (1.3). Indeed, it was shown that the nuclear norm is the convex envelope of the rank function [8]. The upshot of problem (1.4) is that it is in the form of the so-called composite minimization that has been heavily studied in the literature, especially when h is convex. Therefore, in principle it can be solved by many existing algorithms for composite minimization, including in particular various proximal algorithms [31, 18, 38, 40] in view of the closed-form expression of the proximal operator of the nuclear norm [7].

Nonetheless, in contemporary applications, the dimensions m and n of the decision variable X can potentially be extremely high, rendering existing methods inapplicable. 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 [26]. 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 (1.4), is a cubic function in m and n, as it involves the singular value decomposition of X. To circumvent this, researchers proposed to solve problem (1.4) via the Burer-Monteiro factorization technique [5, 6, 32, 46, 44], which replaces the variable X by a low-rank approximation $UV^\top$ and solves the resulting problem:

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

where $\|\cdot\|_F$ denotes the Frobenius norm and r is an integer parameter specified by the modeler. Any optimal solution (U^*, V^*) of problem (1.5) corresponds to an approximately optimal solution $U^*{V^*}^\top$ to problem (1.4). The advantage of the factorized problem (1.5) over problem (1.4) is twofold. First, the objective function in the factorized problem (1.5) is differentiable as soon as h is. In contrast, the objective function in problem (1.4) 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 (1.4).

Since our goal is to solve problem (1.4), the factorization rank r cannot be too small. Indeed, optimal solutions of problem (1.4) cannot be recovered by solving problem (1.5) through the correspondence $(U, V) \mapsto UV^\top$ if r is less than the minimum rank r^* of the optimal solutions of problem (1.4). This issue is currently addressed indirectly as follows. First, it can be readily shown that $U^*{V^*}^\top$ is a global minimizer of problem (1.4) for any global minimizer (U^*, V^*) of problem (1.5) if $r \geq r^*$ (*e.g.*, see [16, Lemma 1]). Second, despite the non-convexity due to the bilinear term $UV^\top$, the factorized problem (1.5) has no spurious local minimizers or second-order stationary points if r is *sufficiently large* [19, 44], which implies that one can actually solve problem (1.4) with a convex h

by using any optimization algorithm with a second-order convergence guarantee on problem (1.5). However, a proper choice of the parameter r is highly nontrivial. In particular, as demonstrated by an example constructed in [44], 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 (1.5) *correspond to the global minimizers of problem* (1.4) *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* (1.4) *is said to be r-factorizable if every second-order stationary point (U, V) of the function F_r in problem* (1.5) *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 (1.4).

Most of the existing works on this question focus on the unregularized counterpart:

$$\min_{U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r}} h(UV^\top), \tag{1.6}$$

and rely on the restricted isometry property of h. 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, 45, 42] 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

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

The state-of-the-art results were established in [42, Corollary 2], which showed that all second-order stationary points of problem (1.6) are global minimizers of problem (1.6) if h satisfies the $\frac{1}{3}$-RIP$_{2r,2r}$ condition and that there exists a function h satisfying the $\frac{1}{2}$-RIP$_{2r,2r}$ condition but possessing a spurious second-order stationary point. Other theoretical results can be found in [10, 9, 46]. Interestingly, the question about the choice of the factorization rank r has also been investigated in the symmetric case, where the non-factorized and factorized problems are to minimize $h(X)$ and $h(UU^\top)$, respectively, where X is a symmetric positive semidefinite matrix. In the special case of semidefinite programming, it has been resolved rather satisfactorily [3, 4, 29].

For asymmetric, regularized problem (1.4), fewer works have been done. A common assumption in these works is that there exists an optimal solution to problem (1.4) and r is chosen to be at least the minimum solution rank r^*. In [24, Theorem 3], the author showed that if h is convex quadratic and satisfies the δ-RIP$_{2r,2r}$ condition with $\delta < \frac{1}{3}$, then the corresponding f in problem (1.4) 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 [15, Theorem 1],[1] it was shown that when h satisfies the δ-RIP$_{2r,2r}$ condition with $\delta < \frac{1}{3}$, then the corresponding f in problem (1.4) is r-factorizable, and when $\delta \geq \frac{1}{3}$, a second-order stationary point of problem (1.5) corresponds to an approximate stationary point of problem (1.4). Notice that all these results rely on RIP-type conditions. It remains unclear whether these conditions are necessary for the r-factorizability. It is also worth mentioning that verifying RIP-type conditions is NP-hard [37].

[1] The results in [15] 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.

Unlike most of the existing works, we do not invoke any RIP-type assumptions. The work closest to our non-RIP approach is [44], which considered the following pair of optimization problems:

$$\min_{X\in\mathbb{S}^n_+} h(X) \qquad \text{and} \qquad \min_{U\in\mathbb{R}^{n\times r}} \widetilde{h}(U) := h(UU^\top) \tag{1.7}$$

where $\mathbb{S}^n_+$ is the set of $n\times n$ symmetric matrices and h is C^2, Lipschitz differentiable and 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 $\operatorname{rank}(X^*)$ and the condition number κ (*i.e.*, the ratio between the Lipschitz constant of ∇h and the strong convexity modulus of h), namely (1) $r \geq \operatorname{rank}(X^*)$ and $\kappa < 3$; or (2) $n > r \geq \operatorname{rank}(X^*)$ and $r > \frac{1}{4}(\kappa-1)^2\operatorname{rank}(X^*)$; 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 $\mathbb{S}^n_+$. Our work can be seen as an extension of the studies in [44] to the asymmetric, regularized case, and is more general in the sense that we consider not only strongly but also non-strongly convex h. More precisely, instead of the problem pairs (1.7), we focus on the problems (1.4) and (1.5) and study the r-factorizability of f in connection with a set of natural problem parameters, including the condition number κ of h, the Lipschitz constant of ∇h, the solution rank of problem (1.4) and the rank of a second-order stationary point of (1.5).

We now summarize our technical contributions. First, we prove that every second-order stationary point of problem (1.5) corresponds to an approximate stationary point of problem (1.4) via the mapping $(U,V)\mapsto UV^\top$. In particular, when the second-order stationary point is rank-deficient, it corresponds to a stationary point of problem (1.4), which is in accordance with the findings of [2, 15]. Consequently, if all second-order stationary points of problem (1.5) are rank-deficient and h is convex, then f is r-factorizable. Next, discarding the rank condition on the second-order stationary points but assuming the convexity and Lipschitz differentiability of h, we characterize, with respect to some natural problem parameters, when every second-order stationary point of the factorized problem (1.5) is a global minimizer of non-factorized problem (1.4). More precisely, we subdivide the class of Lipschitz differentiable convex C^2 functions into subclasses according to those natural parameters and characterize when each subclass consists solely of functions h such that every second-order stationary point of problem (1.5) is a global minimizer of problem (1.4). Furthermore, explicit counterexamples are established when the characterizing condition on the said parameters is violated. To our knowledge, our results are the first characterizations of r-factorizability in terms of these natural parameters.

The remainder of the paper is organized as follows. In Section 2, we define the notation and prepare some preliminary results. The characterization of first- and second-order stationary points of problem (1.5) is presented in Section 3. We show that second-order stationary points of problem (1.5) correspond to approximate stationary points of problem (1.4) in Section 4. In Section 5, we derive the characterization of the r-factorizability of the objective function f in problem (1.4).

2 Notation and preliminaries

Throughout this paper, we assume that $1 \leq r \leq m \leq n$ in problem (1.5). 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 $\operatorname{Diag}(x)\in\mathbb{R}^{s\times s}$ the diagonal matrix with $(\operatorname{Diag}(x))_{ii} = x_i$ for $i = 1,\ldots,s$. Moreover, we define $\operatorname{diag}:\mathbb{R}^{s\times s}\to\mathbb{R}^s$ 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.[2] 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:

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

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 $\nabla^2 h(X)[Y, Y] \ge \mu\|Y\|_F^2$ 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 $\{1, \dots, r\}$; 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 ([36, Example 2])**.** *Let $X \in \mathbb{R}^{m\times n}$ be a matrix of rank s and $(R, P) \in \mathcal{O}_X$. Then,*

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

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.

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. □

[2] 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 $\widetilde{\mathrm{Diag}}(x) \in \mathbb{R}^{m\times n}$.

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.*, [1, 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 $i = 1, \ldots, s$ with $\sum_{i=1}^s \lambda_i = 1$. Therefore, we see that

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

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 (1.5), which will be useful for our study of r-factorizability in subsequent sections.

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

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* (1.5) *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 $UV^\top = R[\widetilde{\mathrm{Diag}}(\sigma(U))]^2P^\top = R[\widetilde{\mathrm{Diag}}(\sigma(V))]^2P^\top$ 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 (1.5) 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

$$\begin{aligned} \nabla h(UV^\top)V + \lambda U &= -R\widetilde{\mathrm{Diag}}(d)P^\top P\widetilde{\mathrm{Diag}}(\sigma(V))Q^\top + \lambda R\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top \\ &= -\,R\widetilde{\mathrm{Diag}}(d)P^\top P\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top + \lambda R\widetilde{\mathrm{Diag}}(\sigma(U))Q^\top \end{aligned}$$

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

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 (1.5), *i.e.*, (3.2) holds. It then follows from [19, Proposition 2] that $U^\top U = V^\top V$. 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. □

We next establish an equivalent characterization of second-order stationary points of F_r in (1.5). We start by introducing a useful way to partition matrices. 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:

$$U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \tag{3.6}$$

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.

Proposition 3.4 (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* (1.5) *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[3]

$$\begin{aligned} &-2\lambda\operatorname{tr}(U_{11}^\top V_{11} + U_{12}V_{12}^\top) - 2\operatorname{tr}(D^\top(U_{21}V_{21}^\top + U_{22}V_{22}^\top)) \\ &\quad + \lambda(\|U\|_F^2 + \|V\|_F^2) + \nabla^2 h(\bar{U}\bar{V}^\top)\left[R\begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0 \end{bmatrix}P^\top\right]^2 \ge 0, \end{aligned} \tag{3.7}$$

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

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

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 Item (i).

We now reformulate the second-order optimality condition: $\nabla^2 F_r(\bar{U},\bar{V})[(U,V),(U,V)] \ge 0$ for all $(U,V) \in \mathbb{R}^{m\times r}\times\mathbb{R}^{n\times r}$. Let $R \in \mathcal{O}^m$, $P \in \mathcal{O}^n$ and $Q \in \mathcal{O}^r$ be orthogonal matrices given in Item (i). Since $(U,V) \mapsto (RUQ^\top, PVQ^\top)$ is a bijective linear map on $\mathbb{R}^{m\times r}\times\mathbb{R}^{n\times r}$, the condition is equivalent to that $\nabla^2 F_r(\bar{U},\bar{V})[(RUQ^\top,PVQ^\top),(RUQ^\top,PVQ^\top)] \ge 0$ for all $(U,V) \in \mathbb{R}^{m\times r}\times\mathbb{R}^{n\times r}$. Denoting $\bar{X} = \bar{U}\bar{V}^\top$ and using [19, Equation (3.14)], we have that

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

The second-order condition is therefore further equivalent to that for all $(U,V) \in \mathbb{R}^{m\times r}\times\mathbb{R}^{n\times r}$,

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

[3] Here, we use the partition (3.6) with respect to $(\bar{U},\bar{V})$; this is well defined because $\sigma(\bar{U}) = \sigma(\bar{V})$ holds in Item (i).

$$+ \nabla^2 h(\bar{X}) \left[R \left(\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^\top + U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \right) P^\top \right]^2 \geq 0 \tag{3.8}$$

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

$$+ \lambda(\|U\|_F^2 + \|V\|_F^2) + \nabla^2 h(\bar{X}) \left[R \begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0 \end{bmatrix} P^\top \right]^2 \geq 0, \tag{3.9}$$

where (a) follows from the decomposition for $-\nabla h(\bar{X})$ in Item (i) and the definition of D and Σ, and (b) from the definition of the blocks in (3.6). This shows that Item (i) and Item (ii) together form an equivalent characterization of the second-order stationary points.

We next prove the second claim under the additional assumption of $s = \text{rank}(\bar{U}) < r$. Note that this implies that U_{22} and V_{22} are not null. Hence, we can take U and V in (3.7) 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 Item (i). □

4 Approximate stationary points of f

Recall that our ultimate goal is to solve problem (1.4) to global optimality. Problem (1.5) is only a surrogate that is computationally more friendly. In practice, it is customary to invoke a first- or second-order optimization algorithm to solve problem (1.5), which often returns a second-order stationary point [17, 30]. A natural question is therefore when the second-order stationary points (U, V) of F_r in (1.5) correspond to the global optima of problem (1.4), through the correspondence $(U, V) \mapsto UV^\top$. In general, a second-order stationary point of problem (1.5) may not even correspond to a stationary point of problem (1.4). To see this, suppose that h in (1.4) is a strongly convex function, which implies that $f(\cdot) = h(\cdot) + \lambda\|\cdot\|_*$ is also strongly convex and hence problem (1.4) has a unique stationary point X^*. Assume that $\text{rank}(X^*) > 1$ and pick $r < \text{rank}(X^*)$. Then, F_r is the sum of the level-bounded function $(U, V) \mapsto \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2)$ and the function $(U, V) \mapsto h(UV^\top)$, which is lower bounded due to the strong convexity of h. Consequently, the objective function F_r is level-bounded, and problem (1.5) must have a global minimizer $(\bar{U}, \bar{V})$. However, $\bar{U}\bar{V}^\top$ is not a stationary point of f, since $\text{rank}(\bar{U}\bar{V}^\top) \leq r < \text{rank}(X^*)$. Although in general second-order stationary points of F_r do not correspond to a stationary point of f in (1.4), we show that they correspond to an approximate stationary point of f. A similar result was obtained in [15, Theorem 1] under the restricted isometry property.

Theorem 4.1 (Approximate 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* (1.5)*, $s = \text{rank}(\bar{U})$ and $d \in \mathbb{R}^m$ be given in Proposition 3.4. Then,*[4] *$d_{s+1} \leq \lambda + L\sigma_r(\bar{U}\bar{V}^\top)$, where*

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

Furthermore, $\inf_{Y \in \partial f(\bar{U}\bar{V}^\top)} \|Y\|_2 \leq L\,\sigma_r(\bar{U}\bar{V}^\top)$.

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

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,

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

where the first equality follows from Remark 3.3(i) and the second inequality follows from Proposition 3.4. Therefore, we have

$$-\frac{1}{\lambda}\nabla h(\bar{U}\bar{V}^\top) = R\widetilde{\mathrm{Diag}}(1,\ldots,1,d_{s+1}/\lambda,\ldots,d_m/\lambda)P^\top \in \partial\|\bar{U}\bar{V}^\top\|_*,$$

where the inclusion follows from Proposition 2.1 and the fact that $0 \le d_m \le \cdots \le d_{s+1} \le \lambda$. This proves $0 \in \partial f(\bar{U}\bar{V}^\top)$, and hence $0 = \inf_{Y\in\partial f(\bar{U}\bar{V}^\top)}\|Y\|_2 \le L\,\sigma_r(\bar{U}\bar{V}^\top) = 0$.

We next consider the case where $\mathrm{rank}(\bar{U}) = r$. If $r = m$, then $d_{m+1} = 0$; moreover, we have $d_1 = \cdots = d_m = \lambda$ from Proposition 3.4, which implies (as in the above display) that $\inf_{Y\in\partial f(\bar{U}\bar{V}^\top)}\|Y\|_2 = 0$. 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 (3.6) are null. Since $(\bar{U},\bar{V})$ is a second-order stationary point of F_r in (1.5), by Proposition 3.4, it satisfies the following inequality for any matrices $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$:

$$\begin{aligned}0 \le& -2\lambda\mathrm{tr}(U_{11}^\top V_{11}) - 2\mathrm{tr}(D^\top U_{21}V_{21}^\top) + \lambda(\|U\|_F^2 + \|V\|_F^2)\\ &+ \nabla^2 h(\bar{U}\bar{V}^\top)\left[R\begin{bmatrix}U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top\\ U_{21}\Sigma & 0_{(m-r)\times(n-r)}\end{bmatrix}P^\top\right]^2.\end{aligned}\tag{4.1}$$

Taking U and V to be the matrices with U_{11} and V_{11} being zero and $U_{21} = [e_r\ 0]^\top = e_1e_r^\top \in \mathbb{R}^{(m-r)\times r}$ and $V_{21} = [e_r\ 0]^\top = e_1e_r^\top \in \mathbb{R}^{(n-r)\times r}$, we have that $\Sigma V_{21}^\top = \sigma_r(\bar{U})e_re_1^\top$ and $U_{21}\Sigma = \sigma_r(\bar{U})e_1e_r^\top$ and that $\mathrm{tr}(D^\top U_{21}V_{21}^\top) = e_1^\top De_1 = d_{r+1}$. Substituting these into (4.1) yields

$$\begin{aligned}0 &\le -2d_{r+1} + 2\lambda + \nabla^2 h(\bar{U}\bar{V}^\top)\left[\sigma_r(\bar{U})R\begin{bmatrix}0_{r\times r} & e_re_1^\top\\ e_1e_r^\top & 0_{(m-r)\times(n-r)}\end{bmatrix}P^\top\right]^2\\ &\le -2d_{r+1} + 2\lambda + 2\sigma_r^2(\bar{U})L = -2d_{r+1} + 2\lambda + 2\sigma_r(\bar{U}\bar{V}^\top)L,\end{aligned}$$

where the second inequality follows from the fact that $\left\|\sigma_r(\bar{U})R\begin{bmatrix}0_{r\times r} & e_re_1^\top\\ e_1e_r^\top & 0_{(m-r)\times(n-r)}\end{bmatrix}P^\top\right\|_F = \sqrt{2}\sigma_r(\bar{U})$ and the definition of L, and the equality follows from Remark 3.3(ii). Hence, $d_{r+1} \le \lambda + L\,\sigma_r(\bar{U}\bar{V}^\top)$. In addition, we can further compute that

$$\begin{aligned}\inf_{Y\in\partial f(\bar{U}\bar{V}^\top)}\|Y\|_2 &= \inf_{S\in\partial\|\bar{U}\bar{V}^\top\|_*}\|\nabla h(\bar{U}\bar{V}^\top) + \lambda S\|_2\\ &= \inf_{\|W\|_2\le 1}\left\|-R\begin{bmatrix}\lambda I_r & 0_{r\times(n-r)}\\ 0_{(m-r)\times r} & \widetilde{\mathrm{Diag}}(d_{r+1},\ldots,d_m)\end{bmatrix}P^\top + \lambda R\begin{bmatrix}I_r & 0_{r\times(n-r)}\\ 0_{(m-r)\times r} & W\end{bmatrix}P^\top\right\|_2\\ &= \inf_{\|W\|_2\le\lambda}\|\widetilde{\mathrm{Diag}}(d_{r+1},\ldots,d_m) - W\|_2 = \max\{d_{r+1}-\lambda, 0\} \le L\sigma_r(\bar{U}\bar{V}^\top),\end{aligned}$$

where the first equality follows from the definition of f, the second follows from Proposition 2.1 and Proposition 3.2, the third holds upon making a simple change of variables, the fourth makes use of the unitary invariance of $\|\cdot\|_2$, [23, Proposition 2.1] and the fact that $d_{r+1} \ge \cdots \ge d_m \ge 0$, and the inequality holds because $d_{r+1} \le \lambda + L\,\sigma_r(\bar{U}\bar{V}^\top)$. This completes the proof. □

In view of Theorem 4.1, for a second-order stationary point $(\bar{U},\bar{V})$ of F_r in (1.5), by setting $\bar{X} = \bar{U}\bar{V}^\top$, we see that the smaller $\sigma_r(\bar{X})$ is, the closer $\bar{X}$ is to being a stationary point of f in (1.4). When $\sigma_r(\bar{X}) = 0$, we see that $\bar{X}$ is a stationary point of f, as proved in [2]; see, also [41, Section 3], and [14, 3, 4] for similar results.

Corollary 4.2. *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* (1.5) *satisfying* $\operatorname{rank}(\bar{U})<r$, $\bar{U}\bar{V}^\top$ *is a stationary point of* f *in* (1.4).

Our next goal is to characterize rank-deficient second-order stationary points of F_r in (1.5) under the convexity of the function h. To do so, we need the following lemma.

Lemma 4.3. *Let* $(\bar{U},\bar{V})\in\mathbb{R}^{m\times r}\times\mathbb{R}^{n\times r}$ *satisfy* $\bar{U}^\top\bar{U}=\bar{V}^\top\bar{V}$ *and that* $\bar{U}\bar{V}^\top$ *is a local (global) minimizer of problem* (1.4)*. Then,* $(\bar{U},\bar{V})$ *is a local (global) minimizer of problem* (1.5).

Proof. We only prove the statement for local minimizers, as the proof of the statement for global minimizers is the same. Let $(P,Q)\in\mathcal{O}_{\bar{V}}$. By Lemma 3.1, there exists $R\in\mathcal{O}^m$ such that $(R,Q)\in\mathcal{O}_{\bar{U}}$. Lemma 3.1 also asserts that $\sigma(\bar{V})=\sigma(\bar{U})$. Therefore,

$$\|\bar{U}\bar{V}^\top\|_*=\|R\widetilde{\operatorname{Diag}}(\sigma(\bar{U}))Q^\top Q\widetilde{\operatorname{Diag}}(\sigma(\bar{V}))P^\top\|_*=\sigma(\bar{U})^\top\sigma(\bar{V})=\|\sigma(\bar{U})\|_2^2=\frac{1}{2}(\|\bar{U}\|_F^2+\|\bar{V}\|_F^2),$$

which implies that $F_r(\bar{U},\bar{V})=f(\bar{U}\bar{V}^\top)$. Next, since $\bar{U}\bar{V}^\top$ is a local minimizer of f, by definition, there exists an $\epsilon>0$ such $f(X)\geq f(\bar{U}\bar{V}^\top)$ for any X satisfying $\|X-\bar{U}\bar{V}^\top\|\leq\epsilon$. By the continuity of the mapping $(U,V)\mapsto UV^\top$, there exists an $\epsilon'>0$ such that $\|UV^\top-\bar{U}\bar{V}^\top\|_F\leq\epsilon$ whenever $\|(U,V)-(\bar{U},\bar{V})\|_F\leq\epsilon'$. Consider any (U,V) satisfying $\|(U,V)-(\bar{U},\bar{V})\|_F\leq\epsilon'$. Then,

$$\begin{aligned}F_r(U,V)&=h(UV^\top)+\frac{\lambda}{2}(\|U\|_F^2+\|V\|_F^2)\geq h(UV^\top)+\lambda\min_{U'V'^\top=UV^\top}\frac{\|U'\|_F^2+\|V'\|_F^2}{2}\\&=h(UV^\top)+\lambda\|UV^\top\|_*=f(UV^\top)\geq f(\bar{U}\bar{V}^\top)=F_r(\bar{U},\bar{V}),\end{aligned}$$

where the second equality follows from [34, Lemma 1]. Hence, $(\bar{U},\bar{V})$ is a local minimizer of F_r. This completes the proof. □

The theorem below shows that when the function h in problem (1.4) is convex, a rank-deficient second-order stationary point of problem (1.5) does not only induce a stationary point for problem (1.4), but is also a global minimizer of problem (1.5).

Theorem 4.4 (Second-order stationary point with rank deficiency)**.** *Consider problems* (1.4) *and* (1.5) *with convex* h*. For any* $(U,V)\in\mathbb{R}^{m\times r}\times\mathbb{R}^{n\times r}$ *satisfying* $\operatorname{rank}(U)<r$*, the following statements are equivalent:*

(i) (U,V) *is a second-order stationary point of problem* (1.5)*;*

(ii) $UV^\top$ *is a global minimizer of problem* (1.4) *and* $U^\top U=V^\top V$*;*

(iii) (U,V) *is a global minimizer of problem* (1.5)*.*

Proof. It is trivial that (iii) implies (i). We then prove that (i) implies (ii). By Corollary 4.2, $UV^\top$ is a stationary point of problem (1.4). Since h is convex, the objective function f of problem (1.4) is also convex. The global minimality then follows from the stationarity of $UV^\top$, which together with [19, Proposition 2] implies that $U^\top U=V^\top V$. Finally, it follows directly from Lemma 4.3 that (ii) implies (iii). The proof is thus completed. □

5 Characterization of r-factorizability

In view of Theorem 4.4, any rank-deficient second-order stationary point of F_r in (1.5) is a global minimizer of f (1.4) when h is convex. Consequently, if h is convex and all second-order stationary points of (1.5) are *rank-deficient*, then f is r-factorizable. Moreover, the discussion preceding Theorem 4.1 suggests that the rank conditions cannot be dropped in general.

This section approaches the problem from a different perspective. In particular, we aim to derive characterizations of r-factorizability. Our characterizations are based on a set of carefully chosen parameters, as described in the following definition.

Definition 5.1. *Let $L \in (0, \infty)$, $M \in (0, \infty]$, $\mu \geq 0$, $r^* \in [m] \cup \{0\}$, and $0 \leq q \leq m - r^*$. We define $\mathfrak{S}(L, \mu, r^*, M, q)$ to be the set of all $h \in C^2(\mathbb{R}^{m\times n})$ satisfying the following conditions:*

(i) The function $h(\cdot) - \frac{\mu}{2}\|\cdot\|_F^2$ is convex, and ∇h is L-Lipschitz continuous.

(ii) There exists a global minimizer $X^* \in \mathbb{R}^{m\times n}$ of f in (1.4) satisfying $\operatorname{rank}(X^*) = r^*$, $\|X^*\|_2 \leq M$ and $\|\nabla h(X^*)\|_* \leq \lambda(r^* + q)$.

In view of the first-order optimality condition of (1.4) and Proposition 2.1, one can observe that any Lipschitz differentiable convex C^2 function with (1.4) being solvable belongs to $\mathfrak{S}(L, \mu, r^*, M, q)$ for some L, μ, r^*, M and q. We will study conditions on these parameters so that the corresponding $\mathfrak{S}(L, \mu, r^*, M, q)$ consists solely of h whose corresponding f in (1.4) is r-factorizable. We will first consider in Section 5.1 *strongly convex* loss functions (*i.e.*, $\mu > 0$), where our characterizations of the parameters bear similarity to the recent work [44], which also considered strongly convex loss functions. The general case will be studied in Section 5.2.

5.1 Lipschitz differentiable strongly convex C^2 loss functions

We first characterize the r-factorizability of Lipschitz differentiable strongly convex C^2 functions. It turns out that under strong convexity, the characterization can be made independent of the parameters M and q. We thus consider the class $\mathfrak{S}(L, \mu, r^*, \infty, m - r^*)$. One can observe that any Lipschitz differentiable strongly convex C^2 function belongs to $\mathfrak{S}(L, \mu, r^*, \infty, m - r^*)$ for some L, μ and r^*. Our main result is the following theorem, which characterizes when $\mathfrak{S}(L, \mu, r^*, \infty, m - r^*)$ consists solely of functions h whose corresponding f in (1.4) is r-factorizable.

Theorem 5.2. *Let $r^* \in [m] \cup \{0\}$, $r \in [m]$, $\infty > L \geq \mu > 0$ and $\kappa = \frac{L}{\mu} \geq 1$. If r^*, r and κ satisfy any of the following conditions:*

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

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

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

(4) $r \geq r^*$, $\kappa > 3$ and $\frac{r - r^*}{1 - \frac{4}{(\kappa-1)^2}} > \min\{r, m - r^*\}$,

then for all $h \in \mathfrak{S}(L, \mu, r^*, \infty, m - r^*)$, the corresponding f in (1.4) is r-factorizable. Otherwise, there exists a quadratic $h \in \mathfrak{S}(L, \mu, r^*, \infty, m - r^*)$ such that the corresponding f in (1.4) is not r-factorizable.

Proof. This follows from Proposition B.1, Proposition B.2 and Proposition B.4. □

Note that Item (3) in Theorem 5.2 also appeared in [44, Corollary 1.2] to ensure that, in (1.7), all second-order stationary points U of $\widetilde{h}$ satisfy that $UU^\top$ globally minimizes h over $\mathbb{S}^n_+$.

5.2 Lipschitz differentiable convex C^2 loss functions

In this subsection, we characterize r-factorizability of f in (1.4) with a Lipschitz differentiable convex C^2 function h. In view of the first-order optimality condition of (1.4) and Proposition 2.1, one can observe that any such h with (1.4) being solvable belongs to $\mathfrak{S}(L, 0, r^*, M, q)$ for some L, r^*, M and q. The main result here is similar to Theorem 5.2. It characterizes when $\mathfrak{S}(L, 0, r^*, M, q)$ consists solely of functions h whose corresponding f in (1.4) is r-factorizable.

Theorem 5.3. *Assume $r^* \in [m] \cup \{0\}$ and $r \in [m]$. Let $q \in [0, m - r^*]$, L, $M \in (0, \infty)$ and W^* be the optimal value of the optimization problem:*

$$\begin{aligned} &\sup_{d\in\mathbb{N}_0} \ (q-d)(1-\frac{q}{d}) + \frac{(q-d+LM(r^*-r+d)/\lambda)^2}{4(r^*-r+d)} \\ &s.t. \quad \max\{r-r^*+1, 1\} \le d \le \min\{r, m-r^*\}, \ q-d+LM(r^*-r+d)/\lambda > 0. \end{aligned} \tag{5.1}$$

If r, r^*, L, M, q, W^* satisfy any of the following conditions:

(1) $r^* = 0$,

(2) $r^* > 0$, $r \ge r^* + \lfloor q \rfloor$, *and* $W^* < 0$,

then for all $h \in \mathfrak{S}(L, 0, r^*, M, q)$, the corresponding f in (1.4) *is r-factorizable. Otherwise, there exists a quadratic $h \in \mathfrak{S}(L, 0, r^*, M, q)$ such that the corresponding f in* (1.4) *is not r-factorizable.*

Remark 5.4. *In view of the inequality constraint involving q in* (5.1)*, one can see that the optimal value of* (5.1) *is decreasing in λ and increasing in M. In addition, when $r \ge r^* + \lfloor q \rfloor$, the constraint in* (5.1) *implies that $d \ge r - r^* + 1 \ge \lfloor q \rfloor + 1 \ge q$. This further implies that the optimal value of* (5.1) *is increasing in $q \in [0, \min\{d, m - r^*\}]$. Therefore, the W^* in Theorem 5.3 is more likely negative when M and q are small and λ is large.*

Proof for Theorem 5.3. This follows from Proposition C.1, Proposition C.2 and Lemma C.4. □

A Pseudo-stationarity

Our main strategy for characterizing the r-factorizability is to analyze the pseudo-stationary points of f in (1.4), which we define as follows.

Definition A.1 (Pseudo-stationarity). *A matrix $X \in \mathbb{R}^{m\times n}$ is said to be a pseudo-stationary point of f in* (1.4) *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 $s = \mathrm{rank}(X)$.*

Definition A.1 is motivated by Proposition 2.1 and the first-order optimality condition of (1.4). Indeed, any pseudo-stationary point with the d in Definition A.1 satisfying $\lambda \ge \max\{d_{s+1}, \ldots, d_m\}$ is, in fact, a stationary point of F_r in (1.5). Here, we present several lemmas concerning pseudo-stationary points that are useful for proving the main results in Section 5.

Lemma A.2. *Let X_1, X_2 be two pseudo-stationary points of f in* (1.4) *in the sense of Definition A.1, i.e., there exist R_1, $R_2 \in \mathcal{O}^m$ and P_1, $P_2 \in \mathcal{O}^n$ such that*

$$\begin{aligned} X_1 &= R_1 \begin{bmatrix} \Sigma_1 & 0_{m\times(n-m)} \end{bmatrix} P_1^\top, \quad -\nabla h(X_1) = R_1 \begin{bmatrix} D_1 & 0_{m\times(n-m)} \end{bmatrix} P_1^\top, \\ X_2 &= R_2 \begin{bmatrix} \Sigma_2 & 0_{m\times(n-m)} \end{bmatrix} P_2^\top, \quad -\nabla h(X_2) = R_2 \begin{bmatrix} D_2 & 0_{m\times(n-m)} \end{bmatrix} P_2^\top, \end{aligned} \tag{A.1}$$

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* (1.4) *satisfies that* $h(\cdot) - \frac{\mu}{2}\|\cdot\|_F^2$ *is convex for some* $\mu \geq 0$*, and* ∇h *is Lipschitz continuous with modulus* $L \geq \mu$*. Then, we have*

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

Proof. Define $\phi(\cdot) := h(\cdot) - \frac{\mu}{2}\|\cdot\|_F^2$. Then we see that ϕ is convex and $\nabla\phi$ is Lipschitz continuous with modulus $L - \mu$. We can now invoke [28, Theorem 2.1.5, (2.1.11)] on ϕ to deduce that

$$0 \leq (L-\mu)\langle \nabla\phi(X_1) - \nabla\phi(X_2), X_1 - X_2\rangle - \|\nabla\phi(X_1) - \nabla\phi(X_2)\|_F^2. \tag{A.3}$$

For the right hand side of (A.3), a direct computation shows that

$$\begin{aligned}&(L-\mu)\langle \nabla\phi(X_1) - \nabla\phi(X_2), X_1 - X_2\rangle - \|\nabla\phi(X_1) - \nabla\phi(X_2)\|_F^2\\ &= \langle \nabla\phi(X_1) - \nabla\phi(X_2), (L-\mu)X_1 - \nabla\phi(X_1) - ((L-\mu)X_2 - \nabla\phi(X_2))\rangle\\ &= \langle -\nabla\phi(X_1), (L-\mu)X_2 - \nabla\phi(X_2)\rangle + \langle -\nabla\phi(X_2), (L-\mu)X_1 - \nabla\phi(X_1)\rangle\\ &\quad - \langle -\nabla\phi(X_1), (L-\mu)X_1 - \nabla\phi(X_1)\rangle - \langle -\nabla\phi(X_2), (L-\mu)X_2 - \nabla\phi(X_2)\rangle\\ &= \langle \mu X_1 - \nabla h(X_1), LX_2 - \nabla h(X_2)\rangle + \langle \mu X_2 - \nabla h(X_2), LX_1 - \nabla h(X_1)\rangle\\ &\quad - \langle \mu X_1 - \nabla h(X_1), LX_1 - \nabla h(X_1)\rangle - \langle \mu X_2 - \nabla h(X_2), LX_2 - \nabla h(X_2)\rangle\\ &=: S_1 + S_2,\end{aligned} \tag{A.4}$$

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 (A.1): for $i = 1$, 2,

$$-\langle \mu X_i - \nabla h(X_i), LX_i - \nabla h(X_i)\rangle = -\sum_{j=1}^m (\mu\sigma_j(X_i) + d_j^i)(L\sigma_j(X_i) + d_j^i). \tag{A.5}$$

Next, for S_1, notice that

$$\begin{aligned}S_1 &= \langle \mu X_1 - \nabla h(X_1), LX_2 - \nabla h(X_2)\rangle + \langle \mu X_2 - \nabla h(X_2), LX_1 - \nabla h(X_1)\rangle\\ &\overset{\text{(a)}}{=} \left\langle R_1 \begin{bmatrix}\mu\Sigma_1 + D_1 & 0\end{bmatrix} P_1^\top, R_2 \begin{bmatrix}L\Sigma_2 + D_2 & 0\end{bmatrix} P_2^\top\right\rangle\\ &\quad + \left\langle R_1 \begin{bmatrix}L\Sigma_1 + D_1 & 0\end{bmatrix} P_1^\top, R_2 \begin{bmatrix}\mu\Sigma_2 + D_2 & 0\end{bmatrix} P_2^\top\right\rangle\\ &= \left\langle R_2^\top R_1 \begin{bmatrix}\mu\Sigma_1 + D_1 & 0\end{bmatrix} P_1^\top P_2, \begin{bmatrix}L\Sigma_2 + D_2 & 0\end{bmatrix}\right\rangle\\ &\quad + \left\langle R_2^\top R_1 \begin{bmatrix}L\Sigma_1 + D_1 & 0\end{bmatrix} P_1^\top P_2, \begin{bmatrix}\mu\Sigma_2 + D_2 & 0\end{bmatrix}\right\rangle,\end{aligned} \tag{A.6}$$

where in (a) we have used (A.1). Using the above display and Lemma 2.3, we see that

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

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

Lemma A.3. *Let $\tau \in \mathfrak{P}_m$, $L \geq \mu \geq 0$, and define $\tau(y) = \begin{bmatrix} y_{\tau(1)} & \cdots & y_{\tau(m)} \end{bmatrix}^\top \in \mathbb{R}^m$ for all $y = \begin{bmatrix} y_1 & \cdots & y_m \end{bmatrix}^\top \in \mathbb{R}^m$. Let v^1, v^2, d^1, $d^2 \in \mathbb{R}^m_+$ with $v^1 \neq 0$, and assume that v^1 and v^2 are sorted in descending order. Let $r = |\{i \in [m] :\ v^1_i > 0\}|$ and $r^* = |\{i \in [m] :\ v^2_i > 0\}|$ and assume that*

$$\forall i \in [r],\ j \in [r^*],\ d^1_i = \lambda,\ d^2_j = \lambda, \quad \text{and} \quad \forall k \in [m] \setminus [r],\ d^1_k \leq \lambda + L v^1_r. \tag{A.7}$$

Let X_1, X_2, G_1, $G_2 \in \mathbb{R}^{m\times n}$ be defined as

$$X_1 = \widetilde{\mathrm{Diag}}(v^1),\ X_2 = \widetilde{\mathrm{Diag}}(\tau(v^2)),\ G_1 = \widetilde{\mathrm{Diag}}(d^1),\ G_2 = \widetilde{\mathrm{Diag}}(\tau(d^2)), \tag{A.8}$$

and suppose that

$$\begin{aligned} &\sum_{i=1}^m (L\sigma_i(X_1) + d^1_i)(\mu\sigma_{\tau(i)}(X_2) + d^2_{\tau(i)}) + \sum_{i=1}^m (\mu\sigma_i(X_1) + d^1_i)(L\sigma_{\tau(i)}(X_2) + d^2_{\tau(i)}) \\ &- \sum_{i=1}^m (\mu\sigma_i(X_1) + d^1_i)(L\sigma_i(X_1) + d^1_i) - \sum_{i=1}^m (\mu\sigma_i(X_2) + d^2_i)(L\sigma_i(X_2) + d^2_i) \geq 0 \end{aligned} \tag{A.9}$$

If $G_1 + \mu X_1 \neq G_2 + \mu X_2$, we define a quadratic function h as follows:

$$\begin{aligned} h(X) =& \frac{L}{2}\sum_{i=1}^m \sum_{j\neq i}^n X_{ij}^2 + \frac{\mu}{2}\sum_{i=1}^m (X_{ii} - (X_1)_{ii})^2 - \langle G_1, X\rangle \\ &+ \frac{(\langle X - X_1, -G_2 - \mu X_2 + G_1 + \mu X_1\rangle)^2}{2\langle X_2 - X_1, -G_2 - \mu X_2 + G_1 + \mu X_1\rangle}, \end{aligned} \tag{A.10}$$

Otherwise, we set

$$h(X) = \frac{L}{2}\sum_{i=1}^m \sum_{j\neq i}^n X_{ij}^2 + \frac{\mu}{2}\sum_{i=1}^m (X_{ii} - (X_1)_{ii})^2 - \langle G_1, X\rangle. \tag{A.11}$$

Then h is well defined, $h(\cdot) - \frac{\mu}{2}\|\cdot\|_F^2$ is convex, ∇h is Lipschitz continuous with modulus L, $\nabla h(X_1) = -G_1$, and $\nabla h(X_2) = -G_2$. Moreover, if we define F_r as in (1.5) *with the above h and define $(\bar{U}, \bar{V}) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ as*

$$\bar{U} = \widetilde{\mathrm{Diag}}(\sqrt{\sigma_1(X_1)}, \ldots, \sqrt{\sigma_r(X_1)}) \quad \text{and} \quad \bar{V} = \widetilde{\mathrm{Diag}}(\sqrt{\sigma_1(X_1)}, \ldots, \sqrt{\sigma_r(X_1)}), \tag{A.12}$$

then $(\bar{U}, \bar{V})$ is a second-order stationary point of F_r.

Remark A.4. *Notice that the SVDs of X_1 and X_2 are given by*

$$X_1 = I_m \widetilde{\mathrm{Diag}}(\sigma(X_1)) I_n^\top \quad \text{and} \quad X_2 = W \widetilde{\mathrm{Diag}}(\sigma(X_2)) \begin{bmatrix} W & 0 \\ 0 & I_{n-m} \end{bmatrix}^\top,$$

where $W \in \mathcal{P}_m$ corresponds to the permutation $\tau \in \mathfrak{P}_m$. Based on (A.4), (A.5) *and* (A.6)*, one can show that the inequality* (A.9) *is equivalent to*

$$(L - \mu)\langle (G_2 + \mu X_2) - (G_1 + \mu X_1), X_1 - X_2\rangle \geq \|(G_1 + \mu X_1) - (G_2 + \mu X_2)\|_F^2. \tag{A.13}$$

Proof. We first consider the case where $G_1 + \mu X_1 = G_2 + \mu X_2$. In this case, the function h in (A.11) is clearly well defined, and one can verify that the function $h(\cdot) - \frac{\mu}{2}\|\cdot\|_F^2$ is convex, and ∇h is Lipschitz continuous with modulus L. Moreover, $\nabla h(X_1) = -G_1$ and

$$\nabla h(X_2) = \mu(X_2 - X_1) - G_1 = -G_2.$$

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 3.4, we know that $(\bar{U}, \bar{V})$ is a second-order stationary point of F_r if and only if for all $U_{11}, V_{11} \in \mathbb{R}^{r\times r}$, $U_{21} \in \mathbb{R}^{(m-r)\times r}$, $V_{21} \in \mathbb{R}^{(n-r)\times r}$,[5] it holds that

$$\begin{aligned}&-2\lambda \mathrm{tr}(U_{11}^\top V_{11}) - 2\mathrm{tr}(D^\top U_{21} V_{21}^\top) + \lambda(\|U_{11}\|_F^2 + \|V_{11}\|_F^2 + \|U_{21}\|_F^2 + \|V_{21}\|_F^2) \\ &+ \nabla^2 h(X_1)\left[\begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0\end{bmatrix}\right]^2 \geq 0,\end{aligned} \tag{A.14}$$

where

$$\begin{aligned}&\Sigma = \mathrm{diag}(\sqrt{\sigma_1(X_1)}, \ldots, \sqrt{\sigma_r(X_1)}) \overset{(a)}{=} \mathrm{diag}(\sqrt{v_1^1}, \ldots, \sqrt{v_r^1}) \in \mathbb{R}^{r\times r}, \\ &D = \widetilde{\mathrm{Diag}}(d_{r+1}^1, \ldots, d_m^1) \in \mathbb{R}^{(m-r)\times(n-r)},\end{aligned} \tag{A.15}$$

and in (a) we have used the fact that v^1 is a nonnegative vector sorted in descending order. We will verify (A.14).

To this end, we first use the representation of h in (A.11) to deduce that

$$\begin{aligned}\nabla^2 h(X_1)\left[\begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0\end{bmatrix}\right]^2 &\geq L\|\Sigma V_{21}^\top\|_F^2 + L\|U_{21}\Sigma\|_F^2 \\ &\overset{\text{(a)}}{\geq} L v_r^1(\|U_{21}\|_F^2 + \|V_{21}\|_F^2),\end{aligned} \tag{A.16}$$

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

$$\begin{aligned}&-2\lambda \mathrm{tr}(U_{11}^\top V_{11}) - 2\mathrm{tr}(D^\top U_{21} V_{12}^\top) + \lambda(\|U_{11}\|_F^2 + \|V_{11}\|_F^2 + \|U_{21}\|_F^2 + \|V_{21}\|_F^2) \\ &+ \nabla^2 h(X_1)\left[\begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0\end{bmatrix}\right]^2 \\ &\overset{\text{(a)}}{\geq} -2\mathrm{tr}(D^\top U_{21} V_{21}^\top) + \lambda(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) + \nabla^2 h(X_1)\left[\begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0\end{bmatrix}\right]^2 \\ &\overset{\text{(b)}}{\geq} -2\|D^\top\|_2\|U_{21}\|_F\|V_{21}^\top\|_F + \lambda(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) + \nabla^2 h(X_1)\left[\begin{bmatrix} U_{11}\Sigma + \Sigma V_{11}^\top & \Sigma V_{21}^\top \\ U_{21}\Sigma & 0\end{bmatrix}\right]^2 \\ &\overset{\text{(c)}}{\geq} -2(\lambda + L v_r^1)\|U_{21}\|_F\|V_{21}\|_F + \lambda(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) + L v_r^1(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) \\ &\overset{\text{(d)}}{\geq} -(\lambda + L v_r^1)(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) + \lambda(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) + L v_r^1(\|U_{21}\|_F^2 + \|V_{21}\|_F^2) = 0,\end{aligned}$$

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

[5] Here, we use the partition in (3.6), 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 $\mathrm{rank}(\bar{U}) = r$.

(c) we have used (A.7), (A.16) and the definition of D in (A.15), and in (d) we have used the fact $\|A\|_F\|B\|_F \leq \frac{1}{2}(\|A\|_F^2 + \|B\|_F^2)$. This verifies (A.14) and hence $(\bar{U}, \bar{V})$ is a second-order stationary point of F_r.

Next, we consider the case where $G_1 + \mu X_1 \neq G_2 + \mu X_2$. By Remark A.4, we know that (A.9) is equivalent to

$$(L-\mu)\langle (G_2 + \mu X_2) - (G_1 + \mu X_1), X_1 - X_2\rangle \geq \|(G_1 + \mu X_1) - (G_2 + \mu X_2)\|_F^2.$$

Since $G_1+\mu X_1 \neq G_2+\mu X_2$, we see that $(L-\mu)\langle (G_2+\mu X_2)-(G_1+\mu X_1), X_1-X_2\rangle \geq \|(G_1+\mu X_1)-(G_2+\mu X_2)\|_F^2 > 0$. In particular, this implies $L > \mu$ and $\langle (G_2 + \mu X_2) - (G_1 + \mu X_1), X_1 - X_2\rangle > 0$, showing that h in (A.10) is well defined. Furthermore, we have

$$L - \mu \geq \frac{\|(G_1 + \mu X_1) - (G_2 + \mu X_2)\|_F^2}{\langle (G_2 + \mu X_2) - (G_1 + \mu X_1), X_1 - X_2\rangle}. \tag{A.17}$$

Now, it is routine to check that $h(\cdot) - \frac{\mu}{2}\|\cdot\|_F^2$ is convex, $\nabla h(X_1) = -G_1$, and $\nabla h(X_2) = -G_2$. Moreover, the relation in (A.16) and hence the second-order stationarity of $(\bar{U}, \bar{V})$ can be verified similarly to that in the case where $G_1 + \mu X_1 = G_2 + \mu X_2$. 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 (A.10), it holds that

$$\begin{aligned}
\nabla^2 h(X)[Y,Y] &= L\sum_{i=1}^m\sum_{j\neq i}^n Y_{ij}^2 + \mu\sum_{i=1}^n Y_{ii}^2 + \frac{(\langle Y, -G_2 - \mu X_2 + G_1 + \mu X_1\rangle)^2}{\langle X_2 - X_1, -G_2 - \mu X_2 + G_1 + \mu X_1\rangle} \\
&\overset{\text{(a)}}{=} L\sum_{i=1}^m\sum_{j\neq i}^n Y_{ij}^2 + \mu\sum_{i=1}^n Y_{ii}^2 + \frac{(\langle \widetilde{\mathrm{Diag}}(Y_{11},\ldots,Y_{mm}), -G_2 - \mu X_2 + G_1 + \mu X_1\rangle)^2}{\langle X_2 - X_1, -G_2 - \mu X_2 + G_1 + \mu X_1\rangle} \\
&\overset{\text{(b)}}{\leq} L\sum_{i=1}^m\sum_{j\neq i}^n Y_{ij}^2 + \mu\sum_{i=1}^n Y_{ii}^2 + \frac{\|\widetilde{\mathrm{Diag}}(Y_{11},\ldots,Y_{mm})\|_F^2\,\| - G_2 - \mu X_2 + G_1 + \mu X_1\|_F^2}{\langle X_2 - X_1, -G_2 - \mu X_2 + G_1 + \mu X_1\rangle} \overset{\text{(c)}}{\leq} L\|Y\|_F^2,
\end{aligned}$$

where in (a) we have used the fact that X_1, X_2, G_1, G_2 are diagonal (see (A.8)), in (b) we have used the Cauchy-Schwartz inequality, and in (c) we have used (A.17). This proves that ∇h is Lipschitz continuous with modulus L, and completes the proof. □

B Proof of Theorem 5.2

This subsection contains the essential auxiliary results for the proof of Theorem 5.2. Note that Theorem 5.2 is about deriving conditions for $\mathfrak{S}(L, \mu, r^*, \infty, m - r^*)$ to consist solely of h whose corresponding f in (1.4) is r-factorizable. Our first task is to reduce this problem to a simpler one.

Proposition B.1. *Let $\infty > L \geq \mu > 0$, $r \in [m]$ and $r^* \in [m] \cup \{0\}$. The following statements are equivalent.*

(i) There exists an $h \in \mathfrak{S}(L, \mu, r^*, \infty, m - r^*)$ (see Definition 5.1) such that f in (1.4) *is not r-factorizable.*

(ii) There exist $x, g, y, v \in \mathbb{R}^m$ with $\|g\|_\infty > \lambda$ and $\tau \in \mathfrak{P}_m$ such that

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

and

$$\forall i \in [r],\ x_i > 0,\ g_i = \lambda, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = \lambda, \tag{B.1b}$$

$$\forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, \lambda + L \min_{j\in[r]} x_j], \quad \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, \lambda]. \tag{B.1c}$$

Proof. Assume (i). By (i), we can select $h \in \mathfrak{S}(L, \mu, r^*, \infty, m-r^*)$ 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 (1.5) 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

$$\begin{aligned} X_1 &= R_1 \begin{bmatrix}\Sigma_1 & 0\end{bmatrix} P_1^\top, \quad -\nabla h(X_1) = R_1 \begin{bmatrix}D_1 & 0\end{bmatrix} P_1^\top, \\ X_2 &= R_2 \begin{bmatrix}\Sigma_2 & 0\end{bmatrix} P_2^\top, \quad -\nabla h(X_2) = R_2 \begin{bmatrix}D_2 & 0\end{bmatrix} P_2^\top, \end{aligned} \tag{B.2}$$

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 Theorem 4.4 we can conclude that X_1 is a stationary point of f, leading to a contradiction. Applying Lemma A.2, there exists $\bar{\tau} \in \mathfrak{P}_m$ such that

$$\begin{aligned} &\sum_{i=1}^m (L\sigma_i(X_1) + d^1_i)(\mu\sigma_{\bar{\tau}(i)}(X_2) + d^2_{\bar{\tau}(i)}) + \sum_{i=1}^m (\mu\sigma_i(X_1) + d^1_i)(L\sigma_{\bar{\tau}(i)}(X_2) + d^2_{\bar{\tau}(i)}) \\ &- \sum_{i=1}^m (L\sigma_i(X_1) + d^1_i)(\mu\sigma_i(X_1) + d^1_i) - \sum_{i=1}^m (L\sigma_i(X_2) + d^2_i)(\mu\sigma_i(X_2) + d^2_i) \geq 0, \end{aligned} \tag{B.3}$$

where L and μ are defined in Definition 5.1 for the h we selected. Next, applying Theorem 4.1, 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, (B.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 $-\nabla h(X_2) \in \lambda\partial\|X_2\|_*$, (B.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 (B.1a)–(B.1c) and $\|g\|_\infty > \lambda$. Therefore we know that (ii) holds.

Next, assume (ii). Then, we are able to select (x, g, y, v, τ) satisfying (B.1a)–(B.1c). We sort x and y in descending order to get $\bar{x}$ and $\bar{y}$, respectively. Pick $\tau_1, \tau_2 \in \mathfrak{P}_m$ such that

$$\bar{x} = \tau_1(x),\ \bar{y} = \tau_2(y) \text{ and we define } \bar{g} = \tau_1(g),\ \bar{v} = \tau_2(v). \tag{B.4}$$

Since we have in view of (B.1b) and (B.1c) that

$$\begin{aligned} &\forall i \in [r],\ x_i > 0, \quad \forall i \in [m] \setminus [r],\ x_i = 0, \\ &\forall i \in [r^*],\ y_i > 0, \quad \forall i \in [m] \setminus [r^*],\ y_i = 0, \end{aligned} \tag{B.5}$$

we can see from (B.4) that

$$\tau_1([r]) = [r], \quad \tau_1([m] \setminus [r]) = [m] \setminus [r], \quad \tau_2([r^*]) = [r^*], \quad \tau_2([m] \setminus [r^*]) = [m] \setminus [r^*]. \tag{B.6}$$

Moreover, notice that from (B.1b) and (B.1c) we have

$$\begin{aligned} &\forall i \in [r],\ g_i = \lambda, \quad \forall i \in [m] \setminus [r],\ g_i \in [0, \lambda + L \min_{i\in[r]} x_i], \\ &\forall i \in [r^*],\ v_i = \lambda, \quad \forall i \in [m] \setminus [r^*],\ v_i \in [0, \lambda], \end{aligned} \tag{B.7}$$

This together with the definitions of $\bar{g}$ and $\bar{v}$ in (B.4) and the relations (B.5) and (B.6) implies that

$$\begin{aligned} &\forall i \in [r],\ \bar{x}_i > 0,\ \bar{g}_i = \lambda, \quad \forall i \in [m] \setminus [r],\ \bar{x}_i = 0, \bar{g}_i \in [0, \lambda + L \min_{i\in[r]} \bar{x}_i], \\ &\forall i \in [r^*],\ \bar{y}_i > 0,\ \bar{v}_i = \lambda, \quad \forall i \in [m] \setminus [r^*],\ \bar{y}_i = 0,\ \bar{v}_i \in [0, \lambda]. \end{aligned} \tag{B.8}$$

Let $\rho \in \mathfrak{P}_m$ be defined as $\rho := \tau_2^{-1}\tau\tau_1$ and define the following $m \times n$ matrices:

$$X_1 = \widetilde{\mathrm{Diag}}(\bar{x}),\ X_2 = \widetilde{\mathrm{Diag}}(\rho(\bar{y})),\ G_1 = \widetilde{\mathrm{Diag}}(\bar{g}),\ G_2 = \widetilde{\mathrm{Diag}}(\rho(\bar{v})). \tag{B.9}$$

By direct calculation, we have:

$$\begin{aligned} &\sum_{i=1}^m (L\sigma_i(X_1) + \bar{g}_i)(\mu\sigma_{\rho(i)}(X_2) + \bar{v}_{\rho(i)}) + \sum_{i=1}^m (\mu\sigma_i(X_1) + \bar{g}_i)(L\sigma_{\rho(i)}(X_2) + \bar{v}_{\rho(i)}) \\ &- \sum_{i=1}^m (L\sigma_i(X_1) + \bar{g}_i)(\mu\sigma_i(X_1) + \bar{g}_i) - \sum_{i=1}^m (L\sigma_i(X_2) + \bar{v}_i)(\mu\sigma_i(X_2) + \bar{v}_i) \\ &\overset{\text{(a)}}{=} \sum_{i=1}^m (L\bar{x}_i + \bar{g}_i)(\mu\bar{y}_{\rho(i)} + \bar{v}_{\rho(i)}) + \sum_{i=1}^m (\mu\bar{x}_i + \bar{g}_i)(L\bar{y}_{\rho(i)} + \bar{v}_{\rho(i)}) \\ &- \sum_{i=1}^m (L\bar{x}_i + \bar{g}_i)(\mu\bar{x}_i + \bar{g}_i) - \sum_{i=1}^m (L\bar{y}_i + \bar{v}_i)(\mu\bar{y}_i + \bar{v}_i) \\ &\overset{\text{(b)}}{=} \sum_{i=1}^m (Lx_{\tau_1(i)} + g_{\tau_1(i)})(\mu y_{\tau_2(\rho(i))} + v_{\tau_2(\rho(i))}) + \sum_{i=1}^m (\mu x_{\tau_1(i)} + g_{\tau_1(i)})(Ly_{\tau_2(\rho(i))} + v_{\tau_2(\rho(i))}) \\ &- \sum_{i=1}^m (Lx_{\tau_1(i)} + g_{\tau_1(i)})(\mu x_{\tau_1(i)} + g_{\tau_1(i)}) - \sum_{i=1}^m (Ly_{\tau_2(i)} + v_{\tau_2(i)})(\mu y_{\tau_2(i)} + v_{\tau_2(i)}) \\ &\overset{\text{(c)}}{=} \sum_{i=1}^m (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i + g_i)(Ly_{\tau(i)} + v_{\tau(i)}) \\ &- \sum_{i=1}^m (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^m (Ly_i + v_i)(\mu y_i + v_i) \overset{\text{(d)}}{\geq} 0, \end{aligned} \tag{B.10}$$

where in (a) we have used the fact that $\bar{x}$ and $\bar{y}$ are nonnegative vectors sorted in descending order to calculate $\sigma_i(X_1)$ and $\sigma_i(X_2)$ using (B.9), in (b) we have used (B.4), in (c) we have used that $\rho = \tau_2^{-1}\tau\tau_1$, the substitution $i \leftarrow \tau_1^{-1}(i)$ for the first three terms, and the substitution $i \leftarrow \tau_2^{-1}(i)$ for the last term, and in (d) we have used (B.1a). Applying Lemma A.3, and noticing that the prerequisites in Lemma A.3 are satisfied by (B.8), (B.10) and the fact that $\bar{x}$ and $\bar{y}$ are nonnegative vectors sorted in descending order, we know there exists a quadratic $h \in C^2(\mathbb{R}^{m\times n})$, such that $\nabla h(X_i) = -G_i$ for $i = 1, 2$, $h(\cdot) - \frac{\mu}{2}\|\cdot\|_F^2$ is convex, and ∇h is Lipschitz continuous with modulus L. Moreover, $(\bar{U}, \bar{V})$ is a second-order stationary point of F_r in (1.5), where $(\bar{U}, \bar{V}) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ is defined as:

$$\bar{U} = \widetilde{\mathrm{Diag}}(\sqrt{\sigma_1(X_1)}, \ldots, \sqrt{\sigma_r(X_1)}), \quad \bar{V} = \widetilde{\mathrm{Diag}}(\sqrt{\sigma_1(X_1)}, \ldots, \sqrt{\sigma_r(X_1)}). \tag{B.11}$$

However, since $\|\bar{g}\|_\infty > \lambda$, we know $-\nabla h(X_1) = G_1 \notin \lambda\partial\|X_1\|_*$ by Proposition 2.1. This shows that X_1 is not a stationary point of f in (1.4), and hence f is not r-factorizable. Finally, in view of the definitions of X_2 and G_2 in (B.9), the relations in (B.8) and Proposition 2.1, we can deduce that $-\nabla h(X_2) = G_2 \in \lambda\partial\|X_2\|_*$ and $\mathrm{rank}(X_2) = r^*$, from which we conclude further that $h \in \mathfrak{S}(L, \mu, r^*, \infty, m - r^*)$. Consequently, we know (i) holds. □

Let $\infty > L \geq \mu > 0$. Consider the next optimization problem:

$$
\begin{aligned}
\sup_{\substack{x,g,y,v\in\mathbb{R}^m\\ \tau\in\mathfrak{P}_m}} \quad & \sum_{i=1}^m (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i + g_i)(L y_{\tau(i)} + v_{\tau(i)}) \\
& - \sum_{i=1}^m (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^m (Ly_i + v_i)(\mu y_i + v_i) \\
s.t. \quad & \forall i \in [r],\ x_i > 0,\ g_i = \lambda, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = \lambda, \\
& \forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, \lambda + L \min_{j\in[r]} x_j], \\
& \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, \lambda].
\end{aligned}
\tag{B.12}
$$

One can deduce from Proposition B.1 that determining the r-factorizability of f in (1.4) with $h \in \mathfrak{S}(L, \mu, r^*, \infty, m - r^*)$ is equivalent to determining the existence of a feasible solution (x, g, y, v, τ) with $\|g\|_\infty > \lambda$ to (B.12) having nonnegative objective function value. We now turn to (B.12) and observe that the objective can be rewritten as a sum with the ith summand depending only on $(x_i, g_i, y_{\tau(i)}, v_{\tau(i)})$.[6] In addition, from the structure of the constraints in (B.12), 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 $\tau(i) \in [r^*]$. These motivate the definitions of the next four associated index sets, for any fixed $\tau \in \mathfrak{P}_m$:

$$
\mathcal{J}_1^\tau := [r] \cap \tau^{-1}[r^*],\ \mathcal{J}_2^\tau := [r] \setminus \mathcal{J}_1^\tau,\ \mathcal{J}_3^\tau := ([m] \setminus [r]) \cap \tau^{-1}[r^*],\ \mathcal{J}_4^\tau := ([m] \setminus [r]) \setminus \mathcal{J}_3^\tau. \tag{B.13}
$$

Let $d_\tau := |\mathcal{J}_2^\tau|$. Then, by the definition of $\{\mathcal{J}_i^\tau\}_{i\in[4]}$ in (B.13), we have

$$
|\mathcal{J}_1| = r - |\mathcal{J}_2^\tau| = r - d_\tau,\ |\mathcal{J}_3^\tau| = r^* - |\mathcal{J}_1^\tau| = r^* - r + d_\tau,\ |\mathcal{J}_4^\tau| = m - r - |\mathcal{J}_3^\tau| = m - r^* - d_\tau. \tag{B.14}
$$

To solve (B.12), our strategy is to introduce an auxiliary variable $w \in \mathbb{R}$ to transform (B.12) to the next equivalent form:

$$
\begin{aligned}
\sup_{\substack{x,g,y,v\in\mathbb{R}^m\\ \tau\in\mathfrak{P}_m, w\in\mathbb{R}}} \quad & \sum_{i=1}^m (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i + g_i)(L y_{\tau(i)} + v_{\tau(i)}) \\
& - \sum_{i=1}^m (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^m (Ly_i + v_i)(\mu y_i + v_i) \\
s.t. \quad & \forall i \in [r],\ x_i \geq w > 0,\ g_i = \lambda, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = \lambda, \\
& \forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, \lambda + Lw], \\
& \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, \lambda].
\end{aligned}
\tag{B.15}
$$

Problem (B.12) is equivalent to (B.15) in the following sense: for any feasible solution (x, g, y, v, τ) of (B.12), $(x, g, y, v, \tau, \min_{i\in[r]} x_i)$ is a feasible solution of (B.12) having the same objective function value; for any feasible solution (x, g, y, v, τ, w) of (B.15), (x, g, y, v, τ) is a feasible solution of (B.12) having the same objective function value. Consequently, we have the following result.

Proposition B.2. *There exists a feasible solution (x, g, y, v, τ) with $\|g\|_\infty > \lambda$ to* (B.12) *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* (B.15) *having nonnegative objective function value.*

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

Next, we plan to fix τ and w to analyze the optimal value and the optimal solution of (B.15). 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 B.3. *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* (B.13)*. Consider the following optimization problem:*

$$\begin{aligned}\sup_{x,g,y,v\in\mathbb{R}^m} \quad & \sum_{i=1}^m (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i + g_i)(L y_{\tau(i)} + v_{\tau(i)}) \\ & - \sum_{i=1}^m (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^m (L y_i + v_i)(\mu y_i + v_i) \\ s.t. \quad & \forall i \in [r],\ x_i \geq w,\ g_i = 1, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = 1, \\ & \forall i \in [m]\setminus[r],\ x_i = 0,\ g_i \in [0, 1 + Lw], \quad \forall i \in [m]\setminus[r^*],\ y_i = 0,\ v_i \in [0,1].\end{aligned} \tag{B.16}$$

Then, the optimization problem (B.16) *has optimal solutions, and the optimal value is*

$$\left(-L\mu|\mathcal{J}_2^\tau| + |\mathcal{J}_3^\tau|\frac{L(L-\mu)^2}{4\mu}\right) w^2. \tag{B.17}$$

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* (B.16)*, we have $\|\bar{g}\|_\infty > 1$.*
- *There exists one optimal solution $(\bar{x}, \bar{g}, \bar{y}, \bar{v})$ of* (B.16) *such that $\|\bar{g}\|_\infty > 1$.*

Proof. Using the definition of permutation, we notice that

$$\sum_{i=1}^m (L y_i + v_i)(\mu y_i + v_i) = \sum_{i=1}^m (L y_{\tau(i)} + v_{\tau(i)})(\mu y_{\tau(i)} + v_{\tau(i)}). \tag{B.18}$$

Substituting (B.18) into (B.16), we get that

$$\begin{aligned}\sup_{x,g,y,v\in\mathbb{R}^m} \quad & \sum_{i=1}^m (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i + g_i)(L y_{\tau(i)} + v_{\tau(i)}) \\ & - \sum_{i=1}^m (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^m (L y_{\tau(i)} + v_{\tau(i)})(\mu y_{\tau(i)} + v_{\tau(i)}) \\ s.t. \quad & \forall i \in [r],\ x_i \geq w,\ g_i = 1, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = 1, \\ & \forall i \in [m]\setminus[r],\ x_i = 0,\ g_i \in [0, 1 + Lw], \quad \forall i \in [m]\setminus[r^*],\ y_i = 0,\ v_i \in [0,1].\end{aligned} \tag{B.19}$$

Observe that (B.19) can be decomposed into the next m subproblems for each $i \in [m]$:

$$\begin{aligned}\sup_{x_i,g_i,y_i,v_i\in\mathbb{R}} \quad & (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + (\mu x_i + g_i)(L y_{\tau(i)} + v_{\tau(i)}) \\ & - (Lx_i + g_i)(\mu x_i + g_i) - (L y_{\tau(i)} + v_{\tau(i)})(\mu y_{\tau(i)} + v_{\tau(i)}) \\ s.t. \quad & \begin{cases} x_i \geq w,\ g_i = 1,\ y_{\tau(i)} > 0,\ v_{\tau(i)} = 1, & \text{if } i \in \mathcal{J}_1^\tau, \\ x_i \geq w,\ g_i = 1,\ y_{\tau(i)} = 0,\ v_{\tau(i)} \in [0,1], & \text{if } i \in \mathcal{J}_2^\tau, \\ x_i = 0,\ g_i \in [0, 1+Lw],\ y_{\tau(i)} > 0,\ v_{\tau(i)} = 1, & \text{if } i \in \mathcal{J}_3^\tau, \\ x_i = 0,\ g_i \in [0, 1+Lw],\ y_{\tau(i)} = 0,\ v_{\tau(i)} \in [0,1], & \text{if } i \in \mathcal{J}_4^\tau, \end{cases}\end{aligned} \tag{B.20}$$

where we recall the definition of $\{\mathcal{J}_i^\tau\}_{i\in[4]}$ in (B.13). We now consider the solution and optimal value of each subproblem (B.20) for fixed i:

1. $i \in \mathcal{J}_1^\tau$. Then (B.20) takes the following form:

$$\begin{aligned} \sup_{x_i, y_{\tau(i)}} \ & (Lx_i+1)(\mu y_{\tau(i)}+1) + (\mu x_i+1)(Ly_{\tau(i)}+1) \\ & - (Lx_i+1)(\mu x_i+1) - (Ly_{\tau(i)}+1)(\mu y_{\tau(i)}+1) \\ s.t. \ & x_i \geq w, \ y_{\tau(i)} > 0, \end{aligned} \tag{B.21}$$

   where we used the fact that g_i and $v_{\tau(i)}$ are 1. Denote the objective function of (B.21) 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 (B.21) is 0, and it is achieved if and only if

$$x_i = y_{\tau(i)} \geq w. \tag{B.22}$$

2. $i \in \mathcal{J}_2^\tau$. Then (B.20) takes the following form:

$$\begin{aligned} \sup_{x_i, v_{\tau(i)}} \ & (Lx_i+1)v_{\tau(i)} + (\mu x_i+1)v_{\tau(i)} - (Lx_i+1)(\mu x_i+1) - v_{\tau(i)}^2 \\ s.t. \ & x_i \geq w, \ v_{\tau(i)} \in [0,1], \end{aligned} \tag{B.23}$$

   where we used the fact that g_i and $y_{\tau(i)}$ are 1 and 0, respectively. Denote the objective function of (B.23) 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

$$\sup_{x_i \geq w} S_2(x_i, v_{\tau(i)}) = -L\mu w^2 + (L+\mu)w(v_{\tau(i)}-1) - (v_{\tau(i)}-1)^2, \tag{B.24}$$

   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 (B.24). 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 (B.23) is $-L\mu w^2$, and it is achieved if and only if

$$x_i = w, \ v_{\tau(i)} = 1. \tag{B.25}$$

3. $i \in \mathcal{J}_3^\tau$. Then (B.20) has the following form:

$$\begin{aligned} \sup_{g_i, y_{\tau(i)}} \ & g_i(\mu y_{\tau(i)}+1) + g_i(Ly_{\tau(i)}+1) - g_i^2 - (Ly_{\tau(i)}+1)(\mu y_{\tau(i)}+1) \\ s.t. \ & g_i \in [0, 1+Lw], \ y_{\tau(i)} > 0, \end{aligned} \tag{B.26}$$

   where we used the fact that x_i and $v_{\tau(i)}$ are 0 and 1, respectively. Denote the objective function of (B.26) by S_3. By direct calculation we can rewrite S_3 as follows

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

$$= \frac{L(L-\mu)^2w^2}{4\mu} + \frac{(L-\mu)^2}{4L\mu}(g_i - (1+Lw))(Lw + g_i - 1)$$
$$- L\mu\left(y_{\tau(i)} - \frac{(L+\mu)(g_i-1)}{2L\mu}\right)^2. \tag{B.27}$$

Notice that $g_i - (1+Lw) \le 0$ 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 $\frac{L(L-\mu)^2w^2}{4\mu}$; moreover, when $L > \mu$, the optimal value is achieved if and only if

$$g_i = 1 + Lw,\ y_{\tau(i)} = \frac{(L+\mu)w}{2\mu}, \tag{B.28}$$

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

$$g_i \in (1, 1+Lw],\ y_{\tau(i)} = \frac{(L+\mu)(g_i-1)}{2L\mu}. \tag{B.29}$$

On the other hand, when $g_i \le 1$, 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 (B.27) that

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

Consequently, the optimal value of (B.26) is $\frac{L(L-\mu)^2w^2}{4\mu}$, and is achieved as described in (B.28) and (B.29).

4. $i \in \mathcal{J}_4^\tau$. Then (B.20) has the following form:

$$\begin{aligned} \sup_{g_i, v_{\tau(i)}} \quad & 2g_iv_{\tau(i)} - g_i^2 - v_{\tau(i)}^2 \\ s.t. \quad & g_i \in [0, 1+Lw],\ v_{\tau(i)} \in [0,1], \end{aligned} \tag{B.30}$$

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 (B.30) is 0, and is achieved if and only if

$$g_i = v_{\tau(i)} \in [0,1]. \tag{B.31}$$

Consequently, by the solution sets given in (B.22), (B.25), (B.28), (B.29) and (B.31), we know the solution set of (B.16) 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 (B.16) satisfies $\|\bar{g}\|_\infty > 1$ if and only if $|\mathcal{J}_3^\tau| > 0$, according to the structure of $\bar{g}$ given in (B.28), (B.29) and (B.31). □

Proposition B.4. *Let $r^* \in [m] \cup \{0\}, r \in [m]$, and $\infty > L \ge \mu > 0$. Let G be the objective function of* (B.15) *and let $\kappa := \frac{L}{\mu} \ge 1$. 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* (B.15) *satisfying $\|\bar{g}\|_\infty > \lambda$ and $G(\bar{x}, \bar{g}, \bar{y}, \bar{v}, \bar{\tau}, \bar{w}) \ge 0$.*

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

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

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

4. $r \geq r^*$*,* $\kappa > 3$ *and* $\frac{r-r^*}{1-\frac{4}{(\kappa-1)^2}} > \min\{r, m-r^*\}$.

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 (B.15) can be reduced to the following problem:

$$\begin{aligned}\sup_{\substack{x,g,y,v\in\mathbb{R}^m\\ \tau\in\mathfrak{P}_m, w\in\mathbb{R}}} \lambda^2 \Bigg[& \sum_{i=1}^m (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i + g_i)(L y_{\tau(i)} + v_{\tau(i)}) \\ & - \sum_{i=1}^m (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^m (Ly_i + v_i)(\mu y_i + v_i) \Bigg] \\ \text{s.t.}\ \ & \forall i \in [r],\ x_i \geq w > 0,\ g_i = 1, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = 1, \\ & \forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, 1 + Lw], \\ & \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, 1].\end{aligned} \tag{B.32}$$

Since dropping the constant λ^2 won't affect the sign of the function value and our claim only concerns the feasible set of (B.15) and the *sign* of its objective value, we shall consider the following optimization problem instead:

$$\begin{aligned}\sup_{\substack{x,g,y,v\in\mathbb{R}^m\\ \tau\in\mathfrak{P}_m, w\in\mathbb{R}}} & \sum_{i=1}^m (Lx_i + g_i)(\mu y_{\tau(i)} + v_{\tau(i)}) + \sum_{i=1}^m (\mu x_i + g_i)(L y_{\tau(i)} + v_{\tau(i)}) \\ & - \sum_{i=1}^m (Lx_i + g_i)(\mu x_i + g_i) - \sum_{i=1}^m (Ly_i + v_i)(\mu y_i + v_i) \\ \text{s.t.}\ \ & \forall i \in [r],\ x_i \geq w > 0,\ g_i = 1, \quad \forall i \in [r^*],\ y_i > 0,\ v_i = 1, \\ & \forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, 1 + Lw], \\ & \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, 1].\end{aligned} \tag{B.33}$$

Notice that when τ and w are fixed, (B.33) becomes (B.16). Applying Lemma B.3, setting $d_\tau = |\mathcal{J}_2^\tau|$ and recalling the definition of $\mathcal{J}_3^\tau$ in (B.14), we see that the solution set $\Omega_{w,\tau}$ of (B.16) 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 $H : \mathbb{N}_0 \times \mathbb{R}_+ \to \mathbb{R}$:

$$H(d, w) := \left(-L\mu d + (r^* - r + d)\frac{L(L-\mu)^2}{4\mu} \right) w^2. \tag{B.34}$$

Then in view of (B.14) and (B.17), the optimal value of (B.16) is given by $H(|\mathcal{J}_2^\tau|, w)$. Moreover, we see that (B.33) is equivalent to the following problem

$$\sup_{w\in\mathbb{R},\ d\in\mathbb{N}_0} H(d, w) \quad \text{s.t. } w > 0,\ r - r^* \leq d \leq \min\{r, m - r^*\}, \tag{B.35}$$

where the bound for d comes from the requirement that $|\mathcal{J}_i^\tau| \geq 0$ for $i \in [4]$ in (B.14).

We consider the following scenarios:

(S1) The optimal value of (B.35) is nonpositive, and (B.35) has no feasible solution (d, w) satisfying $H(d, w) \geq 0$ and $r^* - r + d > 0$.

In this scenario, we claim that (B.15) 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 (B.15). A short proof is provided below.

Suppose such a feasible solution $(\bar{x}, \bar{g}, \bar{y}, \bar{v}, \bar{\tau}, \bar{w})$ of (B.15) exists, then either $(\bar{x}, \bar{g}, \bar{y}, \bar{v})/\lambda$ is optimal for (B.16) 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 (B.35) must be positive. In the former case, we see from Lemma B.3 that $|\mathcal{J}_3^{\bar{\tau}}| > 0$, and hence (B.35) 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 (B.14)). Both cases yield a contradiction.

(S2) There exists a feasible solution $(d, \tilde{w})$ of (B.35) satisfying $H(d, \tilde{w}) \geq 0$ and $r^* - r + d > 0$.

In this scenario, (B.15) 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 (B.16) 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 (B.35) is positive and there is no feasible solution (d, w) of (B.35) 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 (x, g, y, v, τ, w) of (B.12) satisfies $\|g\|_\infty = \lambda$; moreover, every feasible point (d, w) to (B.35) must satisfy $d = r - r^*$ and $H(d, w) = -L\mu d w^2 \leq 0$. Then (S1) holds. If $r^* = 0$, we see that every feasible point (d, w) to (B.35) must satisfy $d = r$. Then, in view of (B.34), we can rewrite (B.35) as:

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

This means that every feasible solution of (B.35) 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 (B.35) is 0, and for all feasible solution (d, w) of (B.35) 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, (B.35) 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 (B.35) 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 (B.35) 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 (B.34) 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 (B.35) as:

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

If $\kappa < 3$ and $r \geq r^*$, then $\alpha \leq 0$, and we see that the optimal value of (B.35) is 0. Moreover, for all feasible solution (d, w) of (B.35) 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 (B.35) is 0, and for all (d, w) that is feasible to (B.35), 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 (B.35) 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 $\alpha > \min\{r, m - r^*\}$.

□

C Proof of Theorem 5.3

The development here is similar to Appendix B. First, we would like to transform the r-factorizability of f to a more concrete problem.

Proposition C.1. *Let $r^* \in [m] \cup \{0\}, r \in [m]$, $q \in [0, m - r^*]$, and $L, M \in (0, \infty)$. Then the following statements are equivalent.*

(i) There exists an $h \in \mathfrak{S}(L, 0, r^*, M, q)$ (see Definition 5.1) such that f in (1.4) *is not r-factorizable.*

(ii) There exist $x, g, y, v \in \mathbb{R}^m$ with $\|g\|_\infty > \lambda$ and $\tau \in \mathfrak{P}_m$ such that

$$\sum_{i=1}^{m}(Lx_i + g_i)v_{\tau(i)} + \sum_{i=1}^{m} g_i(Ly_{\tau(i)} + v_{\tau(i)}) - \sum_{i=1}^{m}(Lx_i + g_i)g_i - \sum_{i=1}^{m}(Ly_i + v_i)v_i \geq 0, \tag{C.1a}$$

and

$$\forall i \in [r],\ x_i > 0,\ g_i = \lambda, \quad \forall i \in [r^*],\ 0 < y_i \leq M,\ v_i = \lambda,\ \sum_{j=r^*+1}^{m} v_j \leq \lambda q, \tag{C.1b}$$

$$\forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, \lambda + L \min_{j \in [r]} x_j], \quad \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, \lambda]. \tag{C.1c}$$

Proof. Assume (i), we can argue as in the proof of Proposition B.1 to get (B.3) with $\mu = 0$, and $(x, g, y, v, \tau) = (\operatorname{diag}(\Sigma_1), \operatorname{diag}(D_1), \operatorname{diag}(\Sigma_2), \operatorname{diag}(D_2), \bar{\tau})$ as defined in the proof of Proposition B.1. The additional constraints $y_i \leq M$ and $\sum_{j=r^*+1}^{m} v_j \leq \lambda q$ come from the conditions $\|X_2\|_2 \leq M$ and $\|\nabla h(X^*)\|_* \leq \lambda(r^* + q)$ (see Definition 5.1(ii)) and the observation that (see (B.2))

$$\|\nabla h(X^*)\|_* = \operatorname{tr}(D_2) = \lambda r^* + \sum_{j=r^*+1}^{m} v_j.$$

This proves (ii).

Assume (ii), we can construct h as in the proof of Proposition B.1, and we note that (B.10) holds with $\mu = 0$, which corresponds to (C.1a). □

Similarly, we would introduce a new variable w to decouple x and g as in Appendix B.

$$\begin{array}{ll}\sup\limits_{\substack{x,g,y,v\in\mathbb{R}^m \\ \tau\in\mathfrak{P}_m, w\in\mathbb{R}}} & \sum_{i=1}^{m}(Lx_i + g_i)v_{\tau(i)} + \sum_{i=1}^{m} g_i(Ly_{\tau(i)} + v_{\tau(i)}) \\ & -\sum_{i=1}^{m}(Lx_i + g_i)g_i - \sum_{i=1}^{m}(Ly_i + v_i)v_i \\ s.t. & \forall i \in [r],\ x_i \geq w > 0,\ g_i = \lambda, \\ & \forall i \in [r^*],\ 0 < y_i \leq M,\ v_i = \lambda,\ \sum_{j=r^*+1}^{m} v_j \leq \lambda q, \\ & \forall i \in [m] \setminus [r],\ x_i = 0,\ g_i \in [0, \lambda + Lw], \\ & \forall i \in [m] \setminus [r^*],\ y_i = 0,\ v_i \in [0, \lambda]. \end{array} \tag{C.2}$$

Moreover, we similarly have the next result.

Proposition C.2. *There exists a* (x,g,y,v,τ) *with* $\|g\|_\infty > \lambda$ *satisfying* (C.1a)–(C.1c) *if and only if there exists a feasible solution* (x,g,y,v,τ,w) *with* $\|g\|_\infty > \lambda$ *to* (C.2) *having nonnegative objective function value.*

The next lemma considers how to solve (C.2) when w and τ are fixed. We note that here we also set $\lambda = 1$ to simplify the calculation.

Lemma C.3. *Let* $r^* \in [m] \cup \{0\}, r \in [m]$, $q \in [0, m - r^*]$ *and* $L, \hat{M} \in (0,\infty)$. *Let* $\tau \in \mathfrak{P}_m$ *and* $w > 0$. *Let* $\{\mathcal{J}_i^\tau\}_{i\in[4]}$ *be defined in* (B.13). *Consider the following optimization problem:*

$$\begin{array}{lll} \sup\limits_{x,g,y,v\in\mathbb{R}^m} & \sum\limits_{i=1}^m (Lx_i + g_i)v_{\tau(i)} + \sum\limits_{i=1}^m g_i(Ly_{\tau(i)} + v_{\tau(i)}) - \sum\limits_{i=1}^m (Lx_i + g_i)g_i - \sum\limits_{i=1}^m (Ly_i + v_i)v_i & \\ s.t. & \forall i \in [r],\ x_i \ge w,\ g_i = 1, \quad \forall i \in [r^*],\ 0 < y_i \le \hat{M},\ v_i = 1, \ \sum\limits_{j=r^*+1}^m v_j \le q, & \\ & \forall i \in [m]\setminus[r],\ x_i = 0,\ g_i \in [0, 1 + Lw], \quad \forall i \in [m]\setminus[r^*],\ y_i = 0,\ v_i \in [0,1]. & \end{array} \tag{C.3}$$

Then, the optimization problem (C.3) *has optimal solutions, and the optimal value is*

$$\begin{cases} (q - |\mathcal{J}_2^\tau|)(Lw + 1 - \frac{q}{|\mathcal{J}_2^\tau|}) + |\mathcal{J}_3^\tau| L^2(\hat{M} - \min\{w, \hat{M}/2\})\min\{w,\hat{M}/2\} & \text{if } q < |\mathcal{J}_2^\tau|, \\ |\mathcal{J}_3^\tau| L^2(\hat{M} - \min\{w, \hat{M}/2\})\min\{w,\hat{M}/2\} & \text{otherwise.} \end{cases} \tag{C.4}$$

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* (C.3), *we have* $\|\bar{g}\|_\infty > 1$.
- *There exists one optimal solution* $(\bar{x},\bar{g},\bar{y},\bar{v})$ *of* (C.3) *such that* $\|\bar{g}\|_\infty > 1$.

Proof. Recalling the definition of $\{\mathcal{J}_i^\tau\}_{i\in[4]}$ in (B.13), we can decompose (C.3) into two separate problems:

$$\begin{array}{ll} \sup\limits_{\{x_i, g_i, y_{\tau(i)}, v_{\tau(i)}\}_{i\in\mathcal{J}_2^\tau\cup\mathcal{J}_4^\tau}} & \sum\limits_{i\in\mathcal{J}_2^\tau\cup\mathcal{J}_4^\tau} \big[(Lx_i + g_i)v_{\tau(i)} + g_i(Ly_{\tau(i)} + v_{\tau(i)}) \\ & \qquad -(Lx_i + g_i)g_i - (Ly_{\tau(i)} + v_{\tau(i)})v_{\tau(i)}\big] \\ s.t. & \forall i \in \mathcal{J}_2^\tau,\ x_i \ge w,\ g_i = 1,\ y_{\tau(i)} = 0,\ v_{\tau(i)} \in [0,1], \\ & \forall i \in \mathcal{J}_4^\tau,\ x_i = 0,\ g_i \in [0, 1+Lw], y_{\tau(i)} = 0,\ v_{\tau(i)} \in [0,1], \\ & \sum\limits_{j\in\mathcal{J}_2^\tau\cup\mathcal{J}_4^\tau} v_{\tau(j)} \le q. \end{array} \tag{C.5}$$

$$\begin{array}{ll} \sup\limits_{\{x_i, g_i, y_{\tau(i)}, v_{\tau(i)}\}_{i\in\mathcal{J}_1^\tau\cup\mathcal{J}_3^\tau}} & \sum\limits_{i\in\mathcal{J}_1^\tau\cup\mathcal{J}_3^\tau} \big[(Lx_i + g_i)v_{\tau(i)} + g_i(Ly_{\tau(i)} + v_{\tau(i)}) \\ & \qquad -(Lx_i + g_i)g_i - (Ly_{\tau(i)} + v_{\tau(i)})v_{\tau(i)}\big] \\ s.t. & \forall i \in \mathcal{J}_1^\tau,\ x_i \ge w,\ g_i = 1,\ 0 < y_{\tau(i)} \le \hat{M},\ v_{\tau(i)} = 1, \\ & \forall i \in \mathcal{J}_3^\tau,\ x_i = 0,\ g_i \in [0, 1+Lw], 0 < y_{\tau(i)} \le \hat{M},\ v_{\tau(i)} = 1. \end{array} \tag{C.6}$$

where in (C.5) we have used $\tau(\mathcal{J}_2^\tau \cup \mathcal{J}_4^\tau) = [m]\setminus[r^*]$ by (B.13) to transform the constraint $\sum_{j=r^*+1}^m v_j \le q$ to the constraint $\sum_{j\in\mathcal{J}_2^\tau\cup\mathcal{J}_4^\tau} v_{\tau(j)} \le q$.

We first try to solve (C.5). Denote the objective function in (C.5) by S_{24}. Using the constraints in (C.5), we see that for all $(\{x_i, g_i, y_{\tau(i)}, v_{\tau(i)}\}_{i\in\mathcal{J}_2^\tau\cup\mathcal{J}_4^\tau})$ feasible to (C.5), we have

$$\begin{aligned}
&S_{24}(\{x_i, g_i, y_{\tau(i)}, v_{\tau(i)}\}_{i\in\mathcal{J}_2^\tau\cup\mathcal{J}_4^\tau}) \\
&\overset{\text{(a)}}{=} \sum_{i\in\mathcal{J}_2^\tau}\left[(Lx_i+1)v_{\tau(i)} + v_{\tau(i)} - (Lx_i+1) - v_{\tau(i)}^2\right] + \sum_{i\in\mathcal{J}_4^\tau}\left[g_i v_{\tau(i)} + g_i v_{\tau(i)} - g_i^2 - v_{\tau(i)}^2\right] \\
&= \sum_{i\in\mathcal{J}_2^\tau}\left[Lx_i(v_{\tau(i)}-1) - (v_{\tau(i)}-1)^2\right] - \sum_{i\in\mathcal{J}_4^\tau}(g_i - v_{\tau(i)})^2
\end{aligned} \tag{C.7}$$

where in (a) we have used the following constraints in (C.5):

$$\forall i\in\mathcal{J}_2^\tau,\quad g_i = 1,\ y_{\tau(i)} = 0;\quad \forall i\in\mathcal{J}_4^\tau,\quad x_i = 0,\ y_{\tau(i)} = 0. \tag{C.8}$$

By performing maximization with respect to $\{g_i\}_{i\in\mathcal{J}_4^\tau}$ over the constraint set of (C.5), we see that

$$\begin{aligned}
G_{24}(\{x_i, v_{\tau(i)}\}_{i\in\mathcal{J}_2^\tau}, \{v_{\tau(i)}\}_{i\in\mathcal{J}_4^\tau}) &:= \sup_{\forall i\in\mathcal{J}_4^\tau,\ g_i\in[0,1+Lw]} S_{24}(\{x_i, g_i, y_{\tau(i)}, v_{\tau(i)}\}_{i\in\mathcal{J}_2^\tau\cup\mathcal{J}_4^\tau}) \\
&\overset{\text{(a)}}{=} \sum_{i\in\mathcal{J}_2^\tau}\left[Lx_i(v_{\tau(i)}-1) - (v_{\tau(i)}-1)^2\right],
\end{aligned} \tag{C.9}$$

where in (a) we have used the fact that $v_{\tau(i)}\in[0,1]$ for all $i\in\mathcal{J}_4^\tau$ to show that the optimal value is achievable; moreover, this optimal value is achieved if and only if:

$$\forall i\in\mathcal{J}_4^\tau,\quad g_i = v_{\tau(i)}. \tag{C.10}$$

Consequently, (C.5) is equivalent to that

$$\begin{aligned}
\sup_{\{x_i, v_{\tau(i)}\}_{i\in\mathcal{J}_2^\tau}, \{v_{\tau(i)}\}_{i\in\mathcal{J}_4^\tau}} \quad & \sum_{i\in\mathcal{J}_2^\tau}\left[Lx_i(v_{\tau(i)}-1) - (v_{\tau(i)}-1)^2\right] \\
s.t. \quad & \forall i\in\mathcal{J}_2^\tau,\ x_i\geq w,\ v_{\tau(i)}\in[0,1]. \\
& \forall i\in\mathcal{J}_4^\tau,\ v_{\tau(i)}\in[0,1],\ \sum_{j\in\mathcal{J}_2^\tau\cup\mathcal{J}_4^\tau} v_{\tau(j)}\leq q.
\end{aligned} \tag{C.11}$$

Notice that the objective function in (C.11) is irrelevant to $\{v_{\tau(i)}\}_{i\in\mathcal{J}_4^\tau}$. Performing maximization with respect to $\{v_{\tau(i)}\}_{i\in\mathcal{J}_4^\tau}$, we see that (C.11) is equivalent to that

$$\begin{aligned}
\sup_{\{x_i, v_{\tau(i)}\}_{i\in\mathcal{J}_2^\tau}} \quad & \sum_{i\in\mathcal{J}_2^\tau}\left[Lx_i(v_{\tau(i)}-1) - (v_{\tau(i)}-1)^2\right] \\
s.t. \quad & \forall i\in\mathcal{J}_2^\tau,\ x_i\geq w,\ v_{\tau(i)}\in[0,1],\ \sum_{j\in\mathcal{J}_2^\tau} v_{\tau(j)}\leq q.
\end{aligned} \tag{C.12}$$

Moreover, this is achieved if and only if:

$$\forall i\in\mathcal{J}_4^\tau,\ v_{\tau(i)}\in[0,1];\quad \sum_{j\in\mathcal{J}_2^\tau\cup\mathcal{J}_4^\tau} v_{\tau(j)}\leq q. \tag{C.13}$$

Next, performing maximization with respect to $\{x_i\}_{i\in\mathcal{J}_2^\tau}$ in (C.12), we see that (C.12) is equivalent to that

$$\begin{aligned}
\sup_{\{v_{\tau(i)}\}_{i\in\mathcal{J}_2^\tau}} \quad & \sum_{i\in\mathcal{J}_2^\tau}\left[Lw(v_{\tau(i)}-1) - (v_{\tau(i)}-1)^2\right] \\
s.t. \quad & \forall i\in\mathcal{J}_2^\tau,\ v_{\tau(i)}\in[0,1],\ \sum_{j\in\mathcal{J}_2^\tau} v_{\tau(j)}\leq q.
\end{aligned} \tag{C.14}$$

This is achieved if and only if:

$$\forall i \in \mathcal{J}_2^\tau, \ \begin{cases} x_i = w & \text{if } v_{\tau(i)} < 1, \\ x_i \in [w, \infty) & \text{if } v_{\tau(i)} = 1. \end{cases} \tag{C.15}$$

Denote the objective function in (C.14) by V_2. Notice that when $\mathcal{J}_2^\tau \neq \emptyset$, V_2 can be rewritten as:

$$\begin{aligned} V_2(\{v_{\tau(i)}\}_{i \in \mathcal{J}_2^\tau}) &\overset{\text{(a)}}{=} |\mathcal{J}_2^\tau| Lw(\bar{v} - 1) - \sum_{i \in \mathcal{J}_2^\tau} (v_{\tau(i)} - \bar{v} + \bar{v} - 1)^2 \\ &= |\mathcal{J}_2^\tau| Lw(\bar{v} - 1) - |\mathcal{J}_2^\tau| (\bar{v} - 1)^2 - \sum_{i \in \mathcal{J}_2^\tau} (v_{\tau(i)} - \bar{v})^2 - 2(\bar{v} - 1) \sum_{i \in \mathcal{J}_2^\tau} (v_{\tau(i)} - \bar{v}) \\ &= |\mathcal{J}_2^\tau| Lw(\bar{v} - 1) - |\mathcal{J}_2^\tau| (\bar{v} - 1)^2 - \sum_{i \in \mathcal{J}_2^\tau} (v_{\tau(i)} - \bar{v})^2, \end{aligned}$$

where in (a) we set $\bar{v} = \frac{1}{|\mathcal{J}_2^\tau|} \sum_{i \in \mathcal{J}_2^\tau} v_{\tau(i)}$. This motivates us to introduce an auxiliary variable $\bar{v} \in \mathbb{R}$, to transform (C.14) to the following problem:

$$\begin{aligned} \sup_{\bar{v} \in \mathbb{R}} \ \sup_{\{v_{\tau(i)}\}_{i \in \mathcal{J}_2^\tau}} \quad & |\mathcal{J}_2^\tau| Lw(\bar{v} - 1) - |\mathcal{J}_2^\tau| (\bar{v} - 1)^2 - \sum_{i \in \mathcal{J}_2^\tau} (v_{\tau(i)} - \bar{v})^2, \\ s.t. \quad & \forall i \in \mathcal{J}_2^\tau, \ v_{\tau(i)} \in [0, 1], \ \sum_{i \in \mathcal{J}_2^\tau} v_{\tau(i)} = |\mathcal{J}_2^\tau| \bar{v}, \ \ |\mathcal{J}_2^\tau| \bar{v} \leq q. \end{aligned} \tag{C.16}$$

By solving the inner problem of (C.16), we see that (C.16) is equivalent to the following problem:

$$\begin{aligned} \sup_{\bar{v} \in \mathbb{R}} \quad & |\mathcal{J}_2^\tau| Lw(\bar{v} - 1) - |\mathcal{J}_2^\tau| (\bar{v} - 1)^2, \\ s.t. \quad & \bar{v} \in [0, 1], \ |\mathcal{J}_2^\tau| \bar{v} \leq q. \end{aligned} \tag{C.17}$$

Moreover, the optimal value of the inner problem is achieved if and only if

$$\forall i \in \mathcal{J}_2^\tau, \ \ v_{\tau(i)} = \bar{v}. \tag{C.18}$$

The optimization problem in (C.17) is an elementary univariate quadratic optimization problem. We can directly see that the optimal value V_2^* of (C.17) is

$$V_2^* := \begin{cases} (q - |\mathcal{J}_2^\tau|)(Lw + 1 - \frac{q}{|\mathcal{J}_2^\tau|}) & \text{if } q < |\mathcal{J}_2^\tau|, \\ 0 & \text{if } q \geq |\mathcal{J}_2^\tau|, \end{cases} \tag{C.19}$$

and it is achieved if and only if

$$\bar{v} = \begin{cases} \frac{q}{|\mathcal{J}_2^\tau|} & \text{if } q < |\mathcal{J}_2^\tau|, \\ 1 & \text{if } q \geq |\mathcal{J}_2^\tau|. \end{cases} \tag{C.20}$$

Combining (C.10), (C.13), (C.15), (C.18) and (C.20), we see the optimal solution $(\{\hat{x}_i, \hat{g}_i, \hat{y}_{\tau(i)}, \hat{v}_{\tau(i)}\}_{i \in \mathcal{J}_2^\tau \cup \mathcal{J}_4^\tau})$ to (C.5) is given by

$$\begin{aligned} &\forall i \in \mathcal{J}_2^\tau, \ \hat{x}_i \in \begin{cases} \{w\} & \text{if } q < |\mathcal{J}_2^\tau|, \\ [w, \infty) & \text{if } q \geq |\mathcal{J}_2^\tau|, \end{cases} \ \hat{g}_i = 1, \ \hat{y}_{\tau(i)} = 0, \ \hat{v}_{\tau(i)} = \begin{cases} \frac{q}{|\mathcal{J}_2^\tau|} & \text{if } q < |\mathcal{J}_2^\tau|, \\ 1 & \text{if } q \geq |\mathcal{J}_2^\tau|, \end{cases} \\ &\forall i \in \mathcal{J}_4^\tau, \ \hat{x}_i = 0, \ \hat{g}_i = \hat{v}_{\tau(i)}, \ \hat{y}_{\tau(i)} = 0, \ \hat{v}_{\tau(i)} \in [0, 1] \quad s.t. \sum_{j \in \mathcal{J}_2^\tau \cup \mathcal{J}_4^\tau} v_{\tau(j)} \leq q. \end{aligned} \tag{C.21}$$

Next, our goal is to solve (C.6). Since (C.6) is separable, we would solve it for each fixed i.

1. $i \in \mathcal{J}_1^\tau$. Then (C.6) has the following form:

$$\begin{aligned} \sup_{x_i, y_{\tau(i)}} \quad & (Lx_i + 1) + (Ly_{\tau(i)} + 1) - (Lx_i + 1) - (Ly_{\tau(i)} + 1) \\ s.t. \quad & x_i \geq w,\ 0 < y_{\tau(i)} \leq \hat{M}, \end{aligned} \tag{C.22}$$

   Denote the objective function in (C.22) by S_1. We see that

$$S_1(x_i, y_{\tau(i)}) \equiv 0, \tag{C.23}$$

   and every feasible solution to (C.22) is optimal.

2. $i \in \mathcal{J}_3^\tau$. Then (C.6) has the following form:

$$\begin{aligned} \sup_{g_i, y_{\tau(i)}} \quad & g_i + g_i(Ly_{\tau(i)} + 1) - g_i^2 - (Ly_{\tau(i)} + 1) \\ s.t. \quad & g_i \in [0, 1 + Lw],\ 0 < y_{\tau(i)} \leq \hat{M}, \end{aligned} \tag{C.24}$$

   Denote the objective function in (C.24) by S_3. By direct calculation we can rewrite S_3 as

$$S_3(x_i, y_{\tau(i)}) = (g_i - 1)Ly_{\tau(i)} - (g_i - 1)^2.$$

   When $g_i \leq 1$, we see that $S_3(x_i, y_{\tau(i)}) \leq 0$, and when $g_i > 1$, we see that S_3 is increasing as a function of $y_{\tau(i)}$ when fixing g_i, this means when $g_i > 1$ it holds that

$$\sup_{y_{\tau(i)} \in (0, \hat{M}]} S_3(x_i, y_{\tau(i)}) = (g_i - 1)L\hat{M} - (g_i - 1)^2.$$

   Moreover, this is achieved if and only if $y_{\tau(i)} = \hat{M}$. Maximizing with respect to $g_i \in (1, 1+Lw]$, we see that the quadratic function on the right hand side of the above display is maximized at $g_i = 1 + L\min\{\hat{M}/2, w\}$. Consequently, the optimal value (C.24) is

$$L^2(\hat{M} - \min\{w, \hat{M}/2\}) \min\{w, \hat{M}/2\}. \tag{C.25}$$

   and is attained at

$$y_{\tau(i)} = \hat{M}, \quad g_i = 1 + L\min\{\hat{M}/2, w\}. \tag{C.26}$$

The optimal value of (C.3) is obtained by summing the optimal values given in (C.19), (C.23) and (C.25), and the characterization of the optimal solution follows from the optimal solutions given in (C.21), (C.23) and (C.26). Moreover, every solution $(\bar{x}, \bar{g}, \bar{y}, \bar{v})$ of (C.3) satisfies $\|\bar{g}\|_\infty > 1$ if and only if $|\mathcal{J}_3^\tau| > 0$, according to the structure of $\bar{g}$ given in (C.21), (C.23) and (C.26). □

Lemma C.4. *Let $r^* \in [m] \cup \{0\}, r \in [m]$, $0 \leq q \leq m - r^*$ and $L, M \in (0, \infty)$. Let the objective function of* (C.2) *be G. Suppose that W^* is the optimal value of the next optimization problem:*

$$\begin{aligned} \sup_{d \in \mathbb{N}_0} \quad & (q - d)(1 - \frac{q}{d}) + \frac{(q - d + LM(r^* - r + d)/\lambda)^2}{4(r^* - r + d)} \\ s.t. \quad & \max\{r - r^* + 1, 1\} \leq d \leq \min\{r, m - r^*\},\ q - d + LM(r^* - r + d)/\lambda > 0. \end{aligned}$$

If r, r^*, L, M, q, W^* satisfy any of the following conditions, then (C.2) *has no feasible solution $(\bar{x}, \bar{g}, \bar{y}, \bar{v}, \bar{\tau}, \bar{w})$ with $\|\bar{g}\|_\infty > \lambda$ and $G(\bar{x}, \bar{g}, \bar{y}, \bar{v}, \bar{\tau}, \bar{w}) \geq 0$.*

(1) $r^* = 0$.

(2) $r^* > 0$, $r \geq r^* + \lfloor q \rfloor$, and $W^* < 0$.

Otherwise, such a feasible solution exists.

Proof. The whole proof is similar to Proposition B.4, but since the optimization steps are different, we here present all the details. First, we do the variable change $(x, g, y, v, \tau, w) \leftarrow (x/\lambda, g/\lambda, y/\lambda, v/\lambda, \tau, w/\lambda)$ for (C.2), and then get the next optimization problem:

$$\begin{aligned} \sup_{\substack{x,g,y,v\in\mathbb{R}^m \\ \tau\in\mathfrak{P}_m, w\in\mathbb{R}}} \; & \lambda^2\left(\sum_{i=1}^{m}(Lx_i+g_i)v_{\tau(i)}+\sum_{i=1}^{m}g_i(Ly_{\tau(i)}+v_{\tau(i)})-\sum_{i=1}^{m}(Lx_i+g_i)g_i-\sum_{i=1}^{m}(Ly_i+v_i)v_i\right) \\ s.t. \quad & \forall i\in[r],\ x_i\geq w>0,\ g_i=1, \quad \forall i\in[r^*],\ 0<y_i\leq \hat{M},\ v_i=1,\ \sum_{j=r^*+1}^{m} v_j\leq q, \\ & \forall i\in[m]\setminus[r],\ x_i=0,\ g_i\in[0,1+Lw], \quad \forall i\in[m]\setminus[r^*],\ y_i=0,\ v_i\in[0,1], \end{aligned} \tag{C.27}$$

where $\hat{M} = \frac{M}{\lambda}$. Next, define the following function $H : \mathbb{N}_0 \times \mathbb{R}_+ \to \mathbb{R}$:

$$H(d,w) := \begin{cases} (q-d)(Lw+1-\frac{q}{d}) + H_0(d,w) & \text{if } q < d, \\ H_0(d,w) & \text{otherwise.} \end{cases} \tag{C.28}$$

where $H_0(d,w) := (r^* - r + d)L^2(\hat{M} - \min\{w, \hat{M}/2\})\min\{w, \hat{M}/2\}$. Then in view of (B.14) and (C.4), the optimal value of (C.3) is given by $H(|\mathcal{J}_2^\tau|, w)$.

Using Lemma C.3 and the definition of H in (C.28), upon dropping λ^2 in the objective function since it would not affect the sign of the function value, we can consider (C.29) instead because our claim only concerns the feasible set of (C.2) and the *sign* of its objective value:

$$\sup_{w\in\mathbb{R},\ d\in\mathbb{N}_0} H(d,w) \qquad s.t.\ w>0,\ r-r^*\leq d\leq \min\{r, m-r^*\}, \tag{C.29}$$

where the constraint on d comes from the requirement that $|\mathcal{J}_i^\tau| \geq 0$ for all $i \in [4]$ (see (B.14)).

We consider the following scenarios:

(S1) The optimal value of (C.29) is nonpositive, and (C.29) has no feasible solution (d, w) satisfying $H(d, w) \geq 0$ and $r^* - r + d > 0$.

In this scenario, we claim that (C.2) has no feasible solution $(\bar{x}, \bar{g}, \bar{y}, \bar{v}, \bar{\tau}, \bar{w})$ with $\|\bar{g}\|_\infty > \lambda$ and $G(\bar{x}, \bar{g}, \bar{y}, \bar{v}, \bar{\tau}, \bar{w}) \geq 0$. A short proof is provided below.

Suppose such a feasible solution $(\bar{x}, \bar{g}, \bar{y}, \bar{v}, \bar{\tau}, \bar{w})$ of (C.2) exists, then either $(\bar{x}, \bar{g}, \bar{y}, \bar{v})/\lambda$ is optimal for (C.3) with $\tau = \bar{\tau}$ and $w = \bar{w}/\lambda$, or $(\bar{x}, \bar{g}, \bar{y}, \bar{v})/\lambda$ is not optimal. In the latter case, the optimal value of (C.29) must be positive. In the former case, we see from Lemma C.3 that $|\mathcal{J}_3^\tau| > 0$, and hence (C.29) 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 (B.14)). Both cases yield a contradiction.

(S2) There exists a feasible solution $(\tilde{d}, \tilde{w})$ of (C.29) satisfying $H(\tilde{d}, \tilde{w}) \geq 0$ and $r^* - r + \tilde{d} > 0$. In this scenario, (C.2) 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 a $\bar{\tau} \in \mathfrak{P}_m$ with $|\mathcal{J}_2^{\bar{\tau}}| = \tilde{d}$, and then take $(\tilde{x}, \tilde{g}, \tilde{y}, \tilde{v})$ to be the optimal solution of (C.3) with $\tau = \bar{\tau}$ and $w = \tilde{w}$, and set $(\bar{x}, \bar{g}, \bar{y}, \bar{v}, \bar{w}) = \lambda(\tilde{x}, \tilde{g}, \tilde{y}, \tilde{v}, \tilde{w})$.

Consider the following cases:

Case 1: $r^* = 0$. In this case, we see that every feasible solution (d, w) of (C.29) satisfies $d = r$ and hence $r^* - r + d = 0$, and the optimal value of (C.29) is nonpositive. Then (S1) holds.

Case 2: $r^* > 0$ and $r \le r^* + \lfloor q \rfloor - 1$. In this case, we can set $d = r - r^* + 1$ and $w = \hat{M}/2$, which is feasible due to that $r - r^* + 1 \le r$ and $r - r^* + 1 \le \lfloor q \rfloor \le m - r^*$. Then $r^* - r + d = 1$ and $d \le \lfloor q \rfloor \le q$. Thus,

$$H(d, w) = H_0(d, w) = (r^* - r + d)L^2(\hat{M} - \min\{w, \hat{M}/2\}) \min\{w, \hat{M}/2\} = \frac{L^2\hat{M}^2}{4}.$$

Hence (S2) holds.

Case 3: $r^* > 0$ and $r \ge r^* + \lfloor q \rfloor$. Consider the following problem:

$$\begin{aligned} \sup_{w\in\mathbb{R}, d\in\mathbb{N}_0} \quad & (q-d)(1+wL-\frac{q}{d})+L^2(r^*-r+d)(\hat{M}-\min\{w,\hat{M}/2\})\min\{w,\hat{M}/2\} \\ s.t. \quad & r - r^* + 1 \le d \le \min\{r, m - r^*\},\ w > 0; \end{aligned} \tag{C.30}$$

notice that $d \ge r - r^* + 1 \ge \lfloor q \rfloor + 1 > q$ for all (w, d) feasible for (C.30), and hence the objective of (C.30) equals the $H(d, w)$ defined in (C.28). Observe that (C.30) is obtained by excluding $d = r - r^*$ from (C.29), and we have by direct computation that $H(r - r^*, w) \le H_0(r - r^*, w) = 0$. Thus, if (C.30) has a feasible solution (d, w) with $H(d, w) \ge 0$, then (S2) holds; otherwise (S1) holds.

Note the objective in (C.30) is nonincreasing on $[\frac{\hat{M}}{2}, \infty)$ as a function of w when d is fixed. Hence, it holds that (C.30) has a feasible solution (d, w) with $H(d, w) \ge 0$ if and only if the following optimization problem has a feasible solution (d, w) with nonnegative objective value:

$$\begin{aligned} \sup_{w\in\mathbb{R}, d\in\mathbb{N}_0} \quad & H_1(d, w) := (q - d)(1 + wL - \frac{q}{d}) + L^2(r^* - r + d)(\hat{M}w - w^2) \\ s.t. \quad & r - r^* + 1 \le d \le \min\{r, m - r^*\},\ w \in (0, \hat{M}/2]. \end{aligned} \tag{C.31}$$

Let $\alpha = \frac{q-d+L(r^*-r+d)\hat{M}}{2L(r^*-r+d)}$, and we note that $\alpha < \hat{M}/2$ since $d > q$. We rewrite H_1 as follows:

$$\begin{aligned} H_1(d, w) &= (q - d)(1 - \frac{q}{d}) + L(L(r^* - r + d)\hat{M} + (q - d))w - L^2(r^* - r + d)w^2 \\ &= (q - d)(1 - \frac{q}{d}) - L^2(r^* - r + d)[(w - \alpha)^2 - \alpha^2]. \end{aligned}$$

We can now rewrite (C.31) as:

$$\begin{aligned} \sup_{w\in\mathbb{R}, d\in\mathbb{N}_0} \quad & H_1(d, w) = (q - d)(1 - \frac{q}{d}) - L^2(r^* - r + d)[(w - \alpha)^2 - \alpha^2] \\ s.t. \quad & r - r^* + 1 \le d \le \min\{r, m - r^*\},\ w \in (0, \hat{M}/2], \\ & \alpha = \frac{q - d + L(r^* - r + d)\hat{M}}{2L(r^* - r + d)}. \end{aligned} \tag{C.32}$$

If the choice of d makes $\alpha \le 0$, then we have $(w - \alpha)^2 - \alpha^2 = w^2 - 2\alpha w > 0$, which together with $r^* - r + d \ge 1 > 0$ implies that

$$H_1(d, w) < (q - d)(1 - \frac{q}{d}) \overset{\text{(a)}}{<} 0,$$

where in (a) we have used $d > q$. Therefore, (C.32) has a feasible solution with nonnegative function value if and only if it does so when $\alpha > 0$. This leads us to consider the following problem:

$$\begin{aligned}\sup_{w\in\mathbb{R},d\in\mathbb{N}_0} \quad & (q-d)(1-\frac{q}{d}) - L^2(r^*-r+d)[(w-\alpha)^2-\alpha^2] \\ s.t. \quad & r-r^*+1 \le d \le \min\{r, m-r^*\},\ w\in(0,\hat{M}/2], \\ & \alpha = \frac{q-d+L(r^*-r+d)\hat{M}}{2L(r^*-r+d)} > 0.\end{aligned} \tag{C.33}$$

Maximizing with respect to w and noticing that $\alpha < \frac{\hat{M}}{2}$, we see that (C.33) and the following problem must have or do not have a feasible solution with nonnegative function value simultaneously:

$$\begin{aligned}\sup_{d\in\mathbb{N}_0} \quad & (q-d)(1-\frac{q}{d}) + L^2(r^*-r+d)\alpha^2 \\ s.t. \quad & r-r^*+1 \le d \le \min\{r, m-r^*\},\ \alpha = \frac{q-d+L(r^*-r+d)\hat{M}}{2L(r^*-r+d)} > 0.\end{aligned} \tag{C.34}$$

Simplifying this problem, we obtain

$$\begin{aligned}\sup_{d\in\mathbb{N}_0} \quad & (q-d)(1-\frac{q}{d}) + \frac{(q-d+L\hat{M}(r^*-r+d))^2}{4(r^*-r+d)} \\ s.t. \quad & r-r^*+1 \le d \le \min\{r, m-r^*\},\ q-d+L\hat{M}(r^*-r+d) > 0.\end{aligned} \tag{C.35}$$

The problem (C.35) is compactly discrete, and hence optimal value can be always achieved as long as it is not negative infinity. Therefore, (C.35) has a feasible point with nonnegative function value if and only if the optimal value of (C.35) is nonnegative, in which case (S2) holds, and otherwise (S1) holds.

□

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

References

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

[2] 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/`.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] Maryam Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

[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] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176:311–337, 2019.

[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] Wenjing Li, Wei Bian, and Kim-Chuan Toh. On solving a rank regularized minimization problem via equivalent factorized column-sparse regularized models. To appear in *Mathematical Programming*, 2024.

[21] Tianxiang Liu, Ivan Markovsky, Ting Kei Pong, and Akiko Takeda. A hybrid penalty method for a class of optimization problems with multiple rank constraints. *SIAM Journal on Matrix Analysis and Applications*, 41(3):1260–1283, 2020.

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

[23] Zhaosong Lu and Yong Zhang. Penalty decomposition methods for rank minimization. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NeurIPS'11, page 46–54, Red Hook, NY, USA, 2011. Curran Associates Inc.

[24] Andrew D McRae. Low solution rank of the matrix LASSO under RIP with consequences for rank-constrained algorithms. *arXiv preprint arXiv:2404.12828*, 2024.

[25] 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.

[26] 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.

[27] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

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

[29] Liam O'Carroll, Vaidehi Srinivas, and Aravindan Vijayaraghavan. The Burer-Monteiro SDP method can fail even above the Barvinok-Pataki bound. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NeurIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[30] Ioannis Panageas, Georgios Piliouras, and Xiao Wang. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, NeurIPS '19, pages 6474–6483. Curran Associates, Inc., 2019.

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

[32] 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.

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

[34] Jasson D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 713–719, New York, NY, USA, 2005. Association for Computing Machinery.

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

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

[37] Jonathan Weed. Approximately certifying the restricted isometry property is hard. *IEEE Transactions on Information Theory*, 64(8):5488–5497, 2017.

[38] 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.

[39] Man-Chung Yue and Anthony Man-Cho So. A perturbation inequality for concave functions of singular values and its applications in low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 40(2):396–416, 2016.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.