Information and Inference: A Journal of the IMA (2018) **00**, 1–46
doi:10.1093/imaiai/iay003

The non-convex geometry of low-rank matrix optimization

Qiuwei Li[†], Zhihui Zhu and Gongguo Tang
Department of Electrical Engineering, Colorado School of Mines, CO, USA
[†]Corresponding author. Email: liqiuweiss@gmail.com

[Received on 3 July 2017; revised on 11 January 2018; accepted on 25 January 2018]

This work considers two popular minimization problems: (i) the minimization of a general convex function $f(X)$ with the domain being positive semi-definite matrices, and (ii) the minimization of a general convex function $f(X)$ regularized by the matrix nuclear norm $\|X\|_*$ with the domain being general matrices. Despite their optimal statistical performance in the literature, these two optimization problems have a high computational complexity even when solved using tailored fast convex solvers. To develop faster and more scalable algorithms, we follow the proposal of Burer and Monteiro to factor the low-rank variable $X = UU^\top$ (for semi-definite matrices) or $X = UV^\top$ (for general matrices) and also replace the nuclear norm $\|X\|_*$ with $\left(\|U\|_F^2 + \|V\|_F^2\right)/2$. In spite of the non-convexity of the resulting factored formulations, we prove that each critical point either corresponds to the global optimum of the original convex problems or is a strict saddle where the Hessian matrix has a strictly negative eigenvalue. Such a nice geometric structure of the factored formulations allows many local-search algorithms to find a global optimizer even with random initializations.

Keywords: Burer–Monteiro; global convergence; low rank; matrix factorization; negative curvature; nuclear norm; strict saddle property; weighted PCA; 1-bit matrix recovery.

1. Introduction

Non-convex reformulations of convex optimization problems have received a surge of renewed interest for efficiency and scalability reasons [4,19,24,25,31,34–36,40,41,48–50,52–54,56]. Compared with the convex formulations, the non-convex ones typically involve many fewer variables, allowing them to scale to scenarios with millions of variables. Besides, simple algorithms [23,33,48] applied to the non-convex formulations have surprisingly good performance in practise. However, a complete understanding of this phenomenon, particularly the geometrical structures of these non-convex optimization problems, is still an active research area. Unlike the simple geometry of convex optimization problems where local minimizers are also global ones, the landscapes of general non-convex functions can become extremely complicated. Fortunately, for a range of convex optimization problems, particularly for matrix completion and sensing problems, the corresponding non-convex reformulations have nice geometric structures that allow local-search algorithms to converge to global optimality [23–25,33,36,48,58].

We extend this line of investigation by working with a general convex function $f(X)$ and considering the following two popular optimization problems:

$$\text{for symmetric case: } \underset{X\in\mathbb{R}^{n\times n}}{\text{minimize}}\ f(X) \ \text{ subject to } \ X \succeq 0 \tag{$\mathcal{P}_0$}$$

$$\text{for non-symmetric case: } \underset{X\in\mathbb{R}^{n\times m}}{\text{minimize}}\ f(X) + \lambda\|X\|_* \ \text{ where } \lambda > 0. \tag{$\mathcal{P}_1$}$$

For these two problems, even fast first-order methods, such as the projected gradient descent algorithm [8], require performing an expensive eigenvalue decomposition or singular value decomposition in each iteration. These expensive operations form the major computational bottleneck and prevent them from scaling to scenarios with millions of variables, a typical situation in a diverse range of applications, including quantum state tomography [27], user preferences prediction [20] and pairwise distances estimation in sensor localization [6].

1.1 *Our approach: Burer–Monteiro-style parameterization*

As we have seen, the extremely large dimension of the optimization variable X and the accordingly expensive eigenvalue or singular value decompositions on X form the major computational bottleneck of the convex optimization algorithms. An immediate question might be "Is there a way to directly reduce the dimension of the optimization variable X and meanwhile avoid performing the expensive eigenvalue or singular value decompositions?"

This question can be answered when the original optimization problems ($\mathcal{P}_0$) and ($\mathcal{P}_1$) admit a low-rank solution $X^\star$ with $\operatorname{rank}(X^\star) = r^\star \ll \min\{n, m\}$. Then we can follow the proposal of Burer and Monteiro [9] to parameterize the low-rank variable as $X = UU^\top$ for ($\mathcal{P}_0$) or $X = UV^\top$ for ($\mathcal{P}_1$), where $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{m\times r}$ with $r \geq r^\star$. Moreover, since $\|X\|_* = \operatorname{minimize}_{X=UV^\top} \left(\|U\|_F^2 + \|V\|_F^2\right)/2$, we obtain the following non-convex re-parameterizations of ($\mathcal{P}_0$) and ($\mathcal{P}_1$):

$$\text{for symmetric case:} \quad \underset{U\in\mathbb{R}^{n\times r}}{\operatorname{minimize}}\, g(U) = f(UU^\top), \tag{$\mathcal{F}_0$}$$

$$\text{for non-symmetric case:} \quad \underset{U\in\mathbb{R}^{n\times r}, V\in\mathbb{R}^{m\times r}}{\operatorname{minimize}}\, g(U,V) = f(UV^\top) + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right). \tag{$\mathcal{F}_1$}$$

Since $r \ll \{p, q\}$, the resulting factored problems ($\mathcal{F}_0$) and ($\mathcal{F}_1$) involve many fewer variables. Moreover, because the positive semi-definite constraint is removed from ($\mathcal{P}_0$) and the nuclear norm $\|X\|_*$ in ($\mathcal{P}_1$) is replaced by $\left(\|U\|_F^2 + \|V\|_F^2\right)/2$, there is no need to perform an eigenvalue (or a singular value) decomposition in solving the factored problems.

The past 2 years have seen renewed interest in the Burer–Monteiro factorization for solving low-rank matrix optimization problems [4,24,25,36,37,53]. With technical innovations in analysing the non-convex landscape of the factored objective function, several recent works have shown that with an exact parameterization (i.e. $r = r^\star$) the resulting factored reformulation has no spurious local minima or degenerate saddle points [24,25,36,58]. An important implication is that local-search algorithms such as gradient descent and its variants can converge to the global optima with even random initialization [23,33,48].

We generalize this line of work by assuming a general objective function $f(X)$ in ($\mathcal{P}_0$) and ($\mathcal{P}_1$), not necessarily coming from a matrix inverse problem. This generality allows us to view the resulting factored problems ($\mathcal{F}_0$) and ($\mathcal{F}_1$) as a way to solve the original convex optimization problems to the global optimum, rather than a new modelling method. This perspective, also taken by Burer and Monteiro in their original work [9], frees us from rederiving the statistical performances of the resulting factored optimization problems. Instead, the statistical performances of the resulting factored optimization problems inherit from that of the original convex optimization problems, whose statistical performance can be analysed using a suite of powerful convex analysis techniques, which have accumulated from several decades of research. For example, the original convex optimization problems ($\mathcal{P}_0$) and ($\mathcal{P}_1$) have information-theoretically optimal sampling complexity [15], achieve minimax

denoising rate [13] and satisfy tight oracle inequalities [14]. Therefore, the statistical performances of the factored optimization problems ($\mathcal{F}_0$) and ($\mathcal{F}_1$) share the same theoretical bounds as those of the original convex optimization problems ($\mathcal{P}_0$) and ($\mathcal{P}_1$), as long as we can show that the two problems are equivalent.

In spite of their optimal statistical performance [13–15,18], the original convex optimization problems cannot be scaled to solve the practical problems that originally motivate their development even with specialized first-order algorithms. This was realized since the advent of this field, where the low-rank factorization method was proposed as an alternative to convex solvers [9]. When coupled with stochastic gradient descent, low-rank factorization leads to state-of-the-art performance in practical matrix recovery problems [24,25,36,53,58]. Therefore, our general analysis technique also sheds light on the connection between the geometries of the original convex programmes and their non-convex reformulations.

Although the Burer–Monteiro parameterization tremendously reduces the number of optimization variables from n^2 to nr (or nm to $(n+m)r$) when r is very small, the intrinsic bilinearity makes the factored objective functions non-convex, and introduces additional critical points that are not global optima of the factored optimization problems. One of our main purposes is to show that these additional critical points will not introduce spurious local minima. More precisely, we want to figure out what properties of the convex function f are required for the factored objective functions g to have no spurious local minima.

1.2 *Enlightening examples*

To gain some intuition about the properties of f such that the factored objective function g has no spurious local minima (which is one of the main goals considered in this paper), let us consider the following two examples: weighted principal component analysis (weighted PCA) and the matrix sensing problem.

Weighted PCA: Consider the symmetric weighted PCA problem in which the lifted objective function is

$$f(X) = \frac{1}{2}\left\|W \odot \left(X - X^\star\right)\right\|_F^2,$$

where $\odot$ is the Hadamard product, $X^\star$ is the global optimum we want to recover and W is the known weighting matrix (which is assumed to have no zero entries for simplicity). After applying the Burer–Monteiro parameterization to $f(X)$, we obtain the factored objective function

$$g(U) = \frac{1}{2}\left\|W \odot \left(UU^\top - X^\star\right)\right\|_F^2.$$

To investigate the conditions under which the bilinearity $\phi(U) = UU^\top$ will (not) introduce additional local minima to the factored optimization problems, consider a simple (but enlightening) two-dimensional example, where $W = \begin{bmatrix} \sqrt{1+a} & 1 \\ 1 & \sqrt{1+a} \end{bmatrix}$ for some $a \geq 0, X^\star = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $U = \begin{bmatrix} x \\ y \end{bmatrix}$ for unknowns x, y. Then the factored objective function becomes

$$g(U) = \frac{1+a}{2}(x^2-1)^2 + \frac{1+a}{2}(y^2-1)^2 + (xy-1)^2. \tag{1.1}$$

In this particular setting, we will see that the value of a in the weighting matrix is the deciding factor for the occurrence of spurious local minima.

CLAIM 1.1 The factored objective function $g(U)$ in (1.1) has no spurious local minima when $a \in [0, 2)$; while for $a > 2$, spurious local minima will appear.

Proof. First of all, we compute the gradient $\nabla g(U)$ and Hessian $\nabla^2 g(U)$:

$$\nabla g(U) = 2\begin{bmatrix}(a+1)(x^2-1)x + y(xy-1)\\ (a+1)(y^2-1)y + x(xy-1)\end{bmatrix},$$

$$\nabla^2 g(U) = 2\begin{bmatrix}y^2+(3x^2-1)(a+1) & 2xy-1\\ 2xy-1 & x^2+(3y^2-1)(a+1)\end{bmatrix}.$$

Now we collect all the critical points by solving $\nabla g(U) = 0$ and list the Hessian of g at these points as follows:[1]

1. $U_1 = (0,0)$, $\nabla^2 g(U_1) = -2\begin{bmatrix}a+1 & 1\\ 1 & a+1\end{bmatrix}$,
2. $U_2 = (1,1)$, $\nabla^2 g(U_2) = 2\begin{bmatrix}2a+3 & 1\\ 1 & 2a+3\end{bmatrix}$,
3. $U_3 = \left(\sqrt{\frac{a}{a+2}}, -\sqrt{\frac{a}{a+2}}\right)$, $\nabla^2 g(U_3) = \begin{bmatrix}4a+\frac{8}{a+2}-6 & \frac{8}{a+2}-6\\ \frac{8}{a+2}-6 & 4a+\frac{8}{a+2}-6\end{bmatrix}$,
4. $U_4 = \left(\frac{\sqrt{\frac{\sqrt{a^2-4}+a}{a}}}{\sqrt{2}}, -\frac{\sqrt{2}}{a\sqrt{\frac{\sqrt{a^2-4}+a}{a}}}\right)$, $\nabla^2 g(U_4) = \begin{bmatrix}a+3\sqrt{a^2-4}+2+\frac{2\sqrt{a^2-4}}{a} & -\frac{2(a+2)}{a}\\ -\frac{2(a+2)}{a} & a-3\sqrt{a^2-4}+2-\frac{2\sqrt{a^2-4}}{a}\end{bmatrix}$.

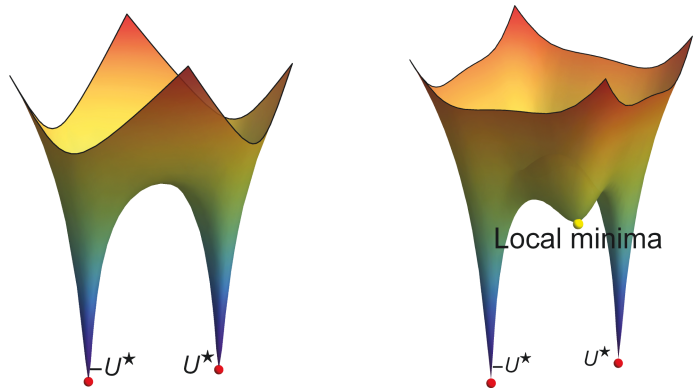
Note that the critical point U_4 exists only for $a \geq 2$. By checking the signs of the two eigenvalues (denoted by λ_1 and λ_2) of these Hessians, we can further classify these critical points as a local minimum, a local maximum or a saddle point:[2]

1. $\lambda_1 = -2(a+2), \lambda_2 = -2a$. So, U_1 is a local maximum for $a > 0$ and a strict saddle for $a = 0$ (see Definition 3).
2. $\lambda_1 = 4(a+1) > 0, \lambda_2 = 4(a+2) > 0$. So, U_2 is a local minimum (also a global minimum as $g(U_2) = 0$).
3. $\lambda_1 = \frac{4(a-2)(a+1)}{a+2}\begin{cases}<0, & a\in[0,2)\\ >0, & a>2\end{cases}$, $\lambda_2 = 4a > 0$. So, U_3 is $\begin{cases}\text{a saddle point,} & a\in[0,2)\\ \text{a } \textit{spurious} \text{ local minimum.} & a>2\end{cases}$
4. From the determinant, we have $\lambda_1 \cdot \lambda_2 = -\frac{8(a-2)(a+1)(a+2)}{a} < 0$ for $a > 2$. So, U_4 is a saddle point for $a > 2$. □

In this example, the value of a controls the dynamic range of the weights as $\max W_{ij}^2 / \min W_{ij}^2 = 1+a$. Therefore, Claim 1.1 can be interpreted as a relationship between the spurious local minima and the

[1] Note that if U is a critical point, so is $-U$, since $\nabla g(-U) = -\nabla g(U)$. Hence, we only list one part of these critical points.

[2] This classification of the critical points using the Hessian information is known as the second derivative test, which says a critical point is a local maximum if the Hessian is negative definite, a local minimum is the Hessian is positive definite and a saddle point if the Hessian matrix has both positive and negative eigenvalues.



(a) Small dynamic range **(b)** Large dynamic range

FIG. 1. Factored function landscapes corresponding to different dynamic ranges of the weights W: (a) a small dynamic range with $\max W_{ij}^2/\min W_{ij}^2 = 1$ and (b) a large dynamic range with $\max W_{ij}^2/\min W_{ij}^2 > 3$.

dynamic range: if the dynamic range $\max W_{ij}^2/\min W_{ij}^2$ is smaller than 3, there will be no spurious local minima; while if the dynamic range is larger than 3, spurious local minima will appear. We also plot the landscapes of the factored objective function $g(U)$ in (1.1) with different dynamic ranges in Fig. 1.

As we have seen, the dynamic range of the weighting matrix serves as a determinant factor for the appearance of the spurious local minima for $g(U)$ in (1.1). To extend the above observations to general objective functions, we now interpret this condition (on the dynamic range of the weighting matrix) by relating it with the condition number of the Hessian matrix $\nabla^2 f(X)$. This can be seen from the following directional-curvature form for $f(X)$:

$$\left[\nabla^2 f(X)\right](D,D) = \left\|W \odot D\right\|_F^2,$$

where $\left[\nabla^2 f(X)\right](D,D)$ is the directional curvature of $f(X)$ along the matrix D of the same dimension as X, defined by $\sum_{i,j,l,k} \frac{\partial^2 f(X)}{\partial X_{ij}\partial X_{lk}} D_{ij}D_{lk}$. This implies that the condition number $\lambda_{\max}\left(\nabla^2 f(X)\right)/\lambda_{\min}\left(\nabla^2 f(X)\right)$ is upper bounded by this dynamic range:

$$\min_{ij}\left|W_{ij}\right|^2 \cdot \left\|D\right\|_F^2 \le \left[\nabla^2 f(X)\right](D,D) \le \max_{ij}\left|W_{ij}\right|^2 \cdot \left\|D\right\|_F^2 \quad \Leftrightarrow \quad \frac{\lambda_{\max}\left(\nabla^2 f(X)\right)}{\lambda_{\min}\left(\nabla^2 f(X)\right)} \le \frac{\max W_{ij}^2}{\min W_{ij}^2}. \tag{1.2}$$

Therefore, we conjecture that the condition number of the general convex function $f(X)$ would be a deciding factor of the behaviour of the landscape of the factored objective function and a large condition number is very likely to introduce spurious local minima to the factored problem.

Matrix Sensing: The above conjecture can be further verified by the matrix sensing problem, where the goal is to recover the low-rank positive semi-definite (PSD) matrix $X^\star \in \mathbb{R}^{n\times n}$ from the linear

measurement $\mathbf{y} = \mathcal{A}(X^\star)$ with $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ being a linear measurement operator. Consider the factored objective function $g(U) = f(UU^\top)$ with $U \in \mathbb{R}^{n\times r}$. In [5,36], the authors showed that the non-convex parametrization $UU^\top$ will not introduce spurious local minima to the factored objective function, provided the linear measurement operator $\mathcal{A}$ satisfies the following restricted isometry property (RIP).

DEFINITION 1.2 (RIP) A linear operator $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ satisfies the r-RIP with constant δ_r if

$$(1-\delta_r)\|D\|_F^2 \le \left\|\mathcal{A}(D)\right\|_2^2 \le (1+\delta_r)\|D\|_F^2 \tag{1.3}$$

holds for all $n \times n$ matrices D with $\operatorname{rank}(D) \le r$.

Note that the required condition (1.3) essentially says that the condition number of Hessian matrix $\nabla^2 f(X)$ should be small at least in the directions of the low-rank matrices D, since the directional curvature form of $f(X)$ is computed as $\left[\nabla^2 f(X)\right](D,D) = \|\mathcal{A}(D)\|_F^2$.

From these two examples, we see that as long as the Hessian matrix of the original convex function $f(X)$ has a small (restricted) condition number the resulting factored objective function has a landscape such that all local minima correspond to the globally optimal solution. Therefore, we believe that such a restricted well-conditioned property might be the key factor that brings us a benign factored landscape, i.e.

$$\alpha\|D\|_F^2 \le \left[\nabla^2 f(X)\right](D,D) \le \beta\|D\|_F^2 \ \text{ with } \ \beta/\alpha \ \text{ being small,}$$

which says that the landscape of $f(X)$ in the lifted space is bowl-shaped, at least in the directions of low-rank matrices.

1.3 *Our results*

Before presenting the main results, we list a few necessary definitions.

DEFINITION 1.3 (Critical points) A point x is a critical point of a function if the gradient of this function vanishes at x.

DEFINITION 1.4 (Strict saddles or ridable saddles [48]) For a twice differentiable function, a strict saddle is one of its critical points whose Hessian matrix has at least one strictly negative eigenvalue.

DEFINITION 1.5 (Strict saddle property [25]) A twice differentiable function satisfies strict saddle property if each critical point either corresponds to the local minima or is a strict saddle.

Heuristically, the strict saddle property describes a geometric structure of the landscape: if a critical point is not a local minimum, then it is a strict saddle, which implies that the Hessian matrix at this point has a strictly negative eigenvalue. Hence, we can continue to decrease the function value at this point along the negative-curvature direction. This nice geometric structure ensures that many local-search algorithms, such as noisy gradient descent [23], vanilla gradient descent with random initialization [33] and the trust region method [48], can escape from all the saddle points along the directions associated with the Hessian's negative eigenvalues, and hence converge to a local minimum.

THEOREM 1.6 (Local convergence for strict saddle property [23,30,32,33,48]) The strict saddle property[3] allows many local-search algorithms to escape all the saddle points and converge to a local minimum.

Our primary interest is to understand how the original convex landscapes are transformed by the factored parameterization $X = UU^\top$ or $X = UV^\top$, particularly how the original global optimum is mapped to the factored space, how other types of critical points are introduced and what are their properties. To answer these questions and conclude from the previous two examples, we require that the function $f(X)$ in ($\mathcal{P}_0$) and ($\mathcal{P}_1$) be restricted well-conditioned:[4]

$$\alpha\|D\|_F^2 \le \left[\nabla^2 f(X)\right](D,D) \le \beta\|D\|_F^2 \text{ with } \beta/\alpha \le 1.5 \text{ whenever } \operatorname{rank}(X) \le 2r \text{ and } \operatorname{rank}(D) \le 4r. \tag{$\mathcal{C}$}$$

We show that as long as the function $f(X)$ in the original convex programmes satisfies the restricted well-conditioned assumption ($\mathcal{C}$), each critical point of the factored programmes either corresponds to the low-rank globally optimal solution of the original convex programmes, or is a strict saddle point where the Hessian matrix $\nabla^2 g$ has a strictly negative eigenvalue. This nice geometric structure coupled with the powerful algorithmic tools provided in Theorem 1.6 thus allows simple iterative algorithms to solve the factored programmes to a global optimum.

THEOREM 1.7 (Informal statement of our results) Suppose the objective function $f(X)$ satisfies the restricted well-conditioned assumption (C). Assume $X^\star$ is an optimal solution of ($\mathcal{P}_0$) or ($\mathcal{P}_1$) with $\operatorname{rank}(X^\star) = r^\star$. Set $r \ge r^\star$ for the factored variables U and V. Then any critical point U (or (U, V)) of the factored objective function g in ($\mathcal{F}_0$) and ($\mathcal{F}_1$) either corresponds to the global optimum $X^\star$ such that $X^\star = UU^\top$ for ($\mathcal{P}_0$) (or $X^\star = UV^\top$ for ($\mathcal{P}_1$)) or is a strict saddle point (which includes a local maximum) of g.

First note that our result covers both over-parameterization where $r > r^\star$ and exact parameterization where $r = r^\star$, while most existing results in low-rank matrix optimization problems [24,25,36] mainly consider the exact parameterization case, i.e. $r = r^\star$, due to the hardness of fulfilling the gap between the metric in the factored space and the one in the lifted space for the over-parameterization case. The geometric property established in the theorem ensures that many iterative algorithms [23,33,48] converge to a square-root factor (or a factorization) of $X^\star$, even with random initialization. Therefore,

---

[3] To be precise, Lee *et al.* [32] showed that for any function that has a Lipschitz continuous gradient and obeys the strict saddle property first-order methods with a random initialization almost always escape all the saddle points and converge to a local minimum. The Lipschitz-gradient assumption is commonly adopted for analysing the convergence of local-search algorithms, and we will discuss this issue after Theorem 3.1. To obtain explicit convergence rate, other properties (like the gradient at the points that are away from the critical points is not small) about the objective functions may be required [21,23,30,48]. In this paper, similar to [25], we mostly focus on the properties of the critical points, and we omit the details about the convergence rate. However, we should note that, by utilizing the similar approach in [58], it is possible to extend the strict saddle property so that we can obtain explicit convergence rate for certain algorithms [23,30,48] when applied for solving the factored low-rank problems.

[4] Note that the constant 1.5 for the dynamic range $\frac{\beta}{\alpha}$ in ($\mathcal{C}$) is not optimized, and it is possible to slightly relax this constraint with more sophisticated analysis. However, the example of the weighted PCA in (1.1) implies that the room for improving this constant is rather limited. In particular, Claim 1.1 and (1.2) indicate that, when $\frac{\beta}{\alpha} > 3$, the spurious local minima will occur for the weighted PCA in (1.1). Thus, as a sufficient condition for any general objective function to have no spurious local minima, a universal bound on the condition number should be at least no larger than 3, i.e. $\frac{\beta}{\alpha} \le 3$. Also, aside from the lack of spurious local minima, as stated in Theorem 1.7, the strict saddle property is the other one that needs to be guaranteed.

we can recover the rank-$r^\star$ global minimizer $X^\star$ of $(\mathcal{P}_0)$ and $(\mathcal{P}_1)$ by running local-search algorithms on the factored function $g(U)$ (or $g(U, V)$) if we know an upper bound on the rank $r^\star$. For problems with additional linear constraints, such as those studied in [9], one can combine the original objective function with a least-squares term that penalizes the deviation from the linear constraints. As long as the penalization parameter is large enough, the solution is equivalent to that of the constrained minimization problems, and hence is also covered by our result.

1.4 *Stylized applications*

Our main result only relies on the restricted well-conditioned assumption of $f(X)$. Therefore, in addition to low-rank matrix recovery problems [24,25,36,53,58], it is also applicable to many other low-rank matrix optimization problems with non-quadratic objective functions, including 1-bit matrix recovery, robust PCA [24] and low-rank matrix recovery with non-Gaussian noise [44]. For ease of exposition, we list the following stylized applications regarding the PSD matrices. But we note that the results listed below also hold for the cases where X is general non-symmetric matrices.

1.4.1 *Weighted PCA* We already know that in the two-dimensional case, the landscape for the factored weighted PCA problem is closely related with the dynamic range of the weighting matrix. Now we exploit Theorem 1.7 to derive the result for the high-dimensional case. Consider the *symmetric* weighted PCA problem, where the goal is to recover the ground-truth $X^\star$ from a pointwisely weighted observation $Y = W \odot X^\star$. Here $W \in \mathbb{R}^{n\times n}$ is the known weighting matrix and the desired solution $X^\star \geq 0$ is of rank $r^\star$. A natural approach is to minimize the following squared ℓ_2 loss:

$$\underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}}\ \frac{1}{2}\left\|W \odot (UU^\top - X^\star)\right\|_F^2. \tag{1.4}$$

Unlike the low-rank approximation problem where W is the all-ones matrix, in general there is no analytic solutions for the weighted PCA problem (1.4) [47] and directly solving this traditional ℓ_2 loss (1.4) is known to be NP-hard [26]. We now apply Theorem 1.7 to the weighted PCA problem and show the objective function in (1.4) has nice geometric structures. Towards that end, define $f(X) = \frac{1}{2}\|W \odot (X - X^\star)\|_F^2$ and compute its directional curvature as

$$\left[\nabla^2 f(X)\right](D, D) = \|W \odot D\|_F^2.$$

Since β/α is a restricted condition number (conditioning on directions of low-rank matrices), which must be no larger than the standard condition number $\lambda_{\max}(\nabla^2 f(X))/\lambda_{\min}(\nabla^2 f(X))$. Thus, together with (1.2), we have

$$\frac{\beta}{\alpha} \leq \frac{\lambda_{\max}\left(\nabla^2 f(X)\right)}{\lambda_{\min}\left(\nabla^2 f(X)\right)} \leq \frac{\max W_{ij}^2}{\min W_{ij}^2}.$$

Now we apply Theorem 1.7 to characterize the geometry of the factored problem of (1.4).

COROLLARY 1.8 Suppose the weighting matrix W has a small dynamic range $\frac{\max W_{ij}^2}{\min W_{ij}^2} \leq 1.5$. Then the objective function of (1.4) with $r \geq r^\star$ satisfies the strict saddle property and has no spurious local minima.

1.4.2 *Matrix sensing* We now consider the matrix sensing problem which is presented before in Section 1.2. To apply Theorem 1.7, we first compare the RIP (1.3) with our restricted well-conditioned assumption ($\mathcal{C}$), which is copied below:

$$\alpha\|D\|_F^2 \leq \left[\nabla^2 f(X)\right](D,D) \leq \beta\|D\|_F^2 \text{ with } \beta/\alpha \leq 1.5 \text{ whenever } \operatorname{rank}(X) \leq 2r \text{ and } \operatorname{rank}(D) \leq 4r.$$

Clearly, the restricted well-conditioned assumption ($\mathcal{C}$) would hold if the linear measurement operator $\mathcal{A}$ satisfies the $4r$-RIP with a constant δ_r such that

$$\frac{1+\delta_{4r}}{1-\delta_{4r}} \leq 1.5 \iff \delta_{4r} \in \left[0, \frac{1}{5}\right].$$

Now we can apply Theorem 1.7 to characterize the geometry of the following matrix sensing problem after the factored parameterization:

$$\underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}} \frac{1}{2}\left\|\mathbf{y} - \mathcal{A}(UU^\top)\right\|_2^2. \tag{1.5}$$

COROLLARY 1.9 Suppose the linear map $\mathcal{A}$ satisfies the $4r$-RIP (1.3) with $\delta_{4r} \in [0, 1/5]$. Then the objective function of (1.5) with $r \geq r^\star$ satisfies the strict saddle property and has no spurious local minima.

1.4.3 *1-Bit matrix completion* 1-Bit matrix completion, as its name indicates, is the inverse problem of completing a low-rank matrix from a set of 1-bit quantized measurements

$$Y_{ij} = \text{bit}\left(X_{ij}^\star\right) \quad \text{for } (i,j) \in \Omega.$$

Here, $X^\star \in \mathbb{R}^{n\times n}$ is the low-rank PSD matrix of rank $r^\star$, Ω is a subset of the indices $[n] \times [n]$ and $\text{bit}(\cdot)$ is the 1-bit quantifier which outputs 0 or 1 in a probabilistic manner:

$$\text{bit}(x) = \begin{cases} 1, & \text{with probability } \sigma(x), \\ 0, & \text{with probability } 1-\sigma(x). \end{cases}$$

One typical choice for $\sigma(x)$ is the sigmoid function $\sigma(x) = \frac{e^x}{1+e^x}$. To recover $X^\star$, the authors of [17] propose to minimizing the negative log-likelihood function

$$\underset{X\succeq 0}{\text{minimize}} f(X) := -\sum_{(i,j)\in\Omega}\left[Y_{ij}\log\left(\sigma(X_{ij})\right) + \left(1-Y_{ij}\right)\log\left(1-\sigma(X_{ij})\right)\right] \tag{1.6}$$

and show that if $\|X^\star\|_* \leq cn\sqrt{r^\star}$, $\max_{ij}|X^\star_{ij}| \leq c$ for some small constant c, and Ω follows certain random binomial model, solving the minimization of the negative log-likelihood function with some nuclear norm constraint would be very likely to produce a satisfying approximation to $X^\star$ [**17**, Theorem 1].

However, when $X^\star$ is extremely high-dimensional (which is the typical case in practise), it is not efficient to deal with the nuclear norm constraint, and hence we propose to minimize the factored formulation of (1.6):

$$\underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}}\, g(U) := -\sum_{(i,j)\in\Omega}\left[Y_{ij}\log\left(\sigma\left((UU^\top)_{ij}\right)\right) + (1-Y_{ij})\log\left(1-\sigma\left((UU^\top)_{ij}\right)\right)\right]. \tag{1.7}$$

In order to utilize Theorem 1.7 to understand the landscape of the factored objective function (1.7), we then check the following directional Hessian quadratic form of $f(X)$:

$$\left[\nabla^2 f(X)\right](D,D) = \sum_{(ij)\in\Omega}\sigma'(X_{ij})D^2_{ij}.$$

For simplicity, consider the case where $\Omega = [n]\times[n]$, i.e. observe full quantized measurements. This will not increase the acquisition cost too much, since each measurement is of 1-bit. Under this assumption, we have

$$\min\sigma'(X_{ij})\|D\|_F^2 \leq \left[\nabla^2 f(X)\right](D,D) \leq \max\sigma'(X_{ij})\|D\|_F^2 \quad\Leftrightarrow\quad \frac{\beta}{\alpha} \leq \frac{\max\sigma'(X_{ij})}{\min\sigma'(X_{ij})}.$$

LEMMA 1.10 Let $\Omega = [n]\times[n]$. Assume $\|X\|_\infty := \max|X_{i,j}|$ is bounded by 1.3169. Then the negative log-likelihood function (1.6) $f(X)$ satisfies the restricted well-conditioned property.

Proof. First of all, we claim $\sigma(x)$ is an even, positive function and decreasing when $x \geq 0$. This is because the sigmoid function $\sigma(x)$ is odd, $\sigma'(x) = \sigma(x)(1-\sigma(x)) > 0$ by $\sigma(x)\in(0,1)$ and $\sigma''(x) = -\frac{e^x(e^x-1)}{(e^x+1)^3} < 0$ for $x \geq 0$. Therefore, for any $|X_{ij}| \leq 1.3169$, we have $\frac{\max\sigma'(X_{ij})}{\min\sigma'(X_{ij})} = \frac{\max\sigma'(0)}{\min\sigma'(1.3169)} \leq 1.49995 \leq 1.5$. □

We now use Theorem 1.7 to characterize the landscape of the factored formulation (1.7) in the set $\mathbb{B}_U := \{U\in\mathbb{R}^{n\times r} : \|UU^\top\|_\infty \leq 1.3169\}$.

COROLLARY 1.11 Set $r \geq r^\star$ in (1.7). Then the objective function (1.7) satisfies the strict saddle property and has no spurious local minima in $\mathbb{B}_U$.

We remark that such a constraint on $\|X\|_\infty$ is also required in the seminal work [**17**], while by using the Burer–Monteiro parameterization, our result removes the time-consuming nuclear norm constraint.

1.4.4 *Robust PCA* For the symmetric variant of robust PCA, the observed matrix $Y = X^\star + S$ with S being sparse and $X^\star$ being PSD. Traditionally, we recover $X^\star$ by minimizing $\|Y-X\|_1 = \sum_{ij}|Y_{ij}-X_{ij}|$ subject to a PSD constraint. However, this formulation does not directly fit into our framework due to the non-smoothness of the ℓ_1 norm. An alternative approach is to minimize $\sum_{ij}h_a(Y_{ij}-X_{ij})$, where

$h_a(.)$ is chosen to be a convex smooth approximation to the absolute value function. A possible choice is $h_a(x) = a\log((\exp(x/a) + \exp(-x/a))/2)$, which is shown to be strictly convex and smooth in [50, Lemma A.1].

1.4.5 *Low-rank matrix recovery with non-Gaussian noise* Consider the PCA problem where the underlying noise is non-Gaussian:

$$Y = X^\star + Z,$$

i.e. the noise matrix $Z \in \mathbb{R}^{n\times n}$ may not follow the Gaussian distributions. Here, $X^\star \in \mathbb{R}^{n\times n}$ is a PSD matrix of rank $r^\star$. It is known that when the noise is from normal distribution, the according maximum likelihood estimator (MLE) is given by the minimizer of a squared loss function $\text{minimize}_{X\succeq 0}\ \frac{1}{2}\|Y-X\|_F^2$. However, in practise, the noise is often from other distributions [45], such as Poisson, Bernoulli, Laplacian and Cauchy, just to name a few. In these cases, the resulting MLE, obtained by minimizing the negative log-likelihood function, is not the square-loss one. Such a noise-adaptive estimator is more effective than square-loss minimization. To have a strongly convex and smooth objective function, the noise distribution should be log-strongly-concave, e.g. the Subbotin densities [44, Example 2.13], the Weibull density $f_\beta(x) = \beta x^{\beta-1}\exp(-x^\beta)$ for $\beta \geq 2$ [44, Example 2.14] and the Chernoff's density [3, Conjecture 3.1]. Once the restricted well-conditioned assumption ($\mathcal{C}$) is satisfied, we can then apply Theorem 1.7 to characterize the landscape of the factored formulation. Similar results apply to matrix sensing and weighted PCA when the underlying noise is non-Gaussian.

1.5 *Prior arts and inspirations*

Prior Arts in Non-convex Optimization Problems. The past few years have seen a surge of interest in non-convex reformulations of convex optimization problems for efficiency and scalability reasons. However, fully understanding this phenomenon, mainly the landscapes of these non-convex reformulations could be hard. Even certifying the local optimality of a point might be an NP-hard problem [38]. The existence of spurious local minima that are not global optima is a common issue [22,46]. Also, degenerate saddle points or those surrounded by plateaus of small curvature could also prevent local-search algorithms from converging quickly to local optima [16]. Fortunately, for a range of convex optimization problems, particularly those involving low-rank matrices, the corresponding non-convex reformulations have nice geometric structures that allow local-search algorithms to converge to global optimality. Examples include low-rank matrix factorization, completion and sensing [24,25,36,58], tensor decomposition and completion [2,23], dictionary learning [50], phase retrieval [49] and many more. Based on whether smart initializations are needed, these previous works can be roughly classified into two categories. In one case, the algorithms require a problem-dependent initialization plus local refinement. A good initialization can lead to global convergence if the initial iterate lies in the attraction basin of the global optima [2,4,12,51]. For low-rank matrix recovery problems, such initializations can be obtained using spectral methods [4,51]; for other problems, it is more difficult to find an initial point located in the attraction basin [2]. The second category of works attempts to understand the empirical success of simple algorithms such as gradient descent [33], which converge to global optimality even with random initialization [23–25,33,36,58]. This is achieved by analysing the objective function's landscape and showing that they have no spurious local minima and no degenerate saddle points. Most of the works in the second category are for specific matrix sensing problems with quadratic objective functions. Our work expands this line of geometry-based convergence analysis by considering low-rank matrix optimization problems with general objective functions.

Burer–Monteiro Reformulation for PSD Matrices. In [4], the authors also considered low-rank and PSD matrix optimization problems with general objective functions. They characterized the local landscape around the global optima, and hence their algorithms require proper initializations for global convergence. We instead characterize the global landscape by categorizing all critical points into global optima and strict saddles. This guarantees that several local-search algorithms with random initialization will converge to the global optima. Another closely related work is low-rank and PSD matrix recovery from linear observations by minimizing the factored quadratic objective function [5]. Low-rank matrix recovery from linear measurements is a particular case of our general objective function framework. Furthermore, by relating the first-order optimality condition of the factored problem with the global optimality of the original convex programme, our work provides a more transparent relationship between geometries of these two problems and dramatically simplifies the theoretical argument. More recently, the authors of [7] showed that for general semi-definite programmes with linear objective functions and linear constraints, the factored problems have no spurious local minimizers. In addition to showing non-existence of spurious local minimizers for general objective functions, we also quantify the curvature around the saddle points, and our result covers both over and exact parameterizations.

Burer–Monteiro Reformulation for General Matrices. The most related work is non-symmetric matrix sensing from linear observations, which minimizes the factored quadratic objective function [42]. The ambiguity in the factored parameterization

$$UV^\top = (UR)(VR^{-\top})^\top \text{for all non-singular } R$$

tends to make the factored quadratic objective function badly conditioned, especially when the matrix R or its inverse is close to being singular. To overcome this problem, the regularizer

$$\Theta_E(U,V) = \left\| U^\top U - V^\top V \right\|_F^2 \tag{1.8}$$

is proposed to ensure that U and V have almost equal energy [42,53,57]. In particular, with the regularizer in (1.8), it was shown in [42,57] that $\widetilde{g}(U,V) = f(UV^\top) + \mu\Theta_E(U,V)$ with a properly chosen $\mu > 0$ has similar geometric result as the one provided in Theorem 1.6 for ($\mathcal{P}_1$), i.e. $\widetilde{g}(U,V)$ also obeys the strict saddle property. Compared with [42,53,57], our result shows that it is not necessary to introduce the extra regularization (1.8) if we solve ($\mathcal{P}_1$) with the factorization approach. Indeed, the optimization form $\left\|X\right\|_* = \min_{X=UV^\top}\left(\left\|U\right\|_F^2 + \left\|V\right\|_F^2\right)/2$ of the nuclear norm implicitly requires U and V to have equal energy. On the other hand, we stress that our interest is to analyse the non-convex geometry of the convex problem ($\mathcal{P}_1$) which, as we explained before, has a very nice statistical performance such as it achieves minimax denoising rate [13]. Our geometrical result implies that instead of using convex solvers to solve ($\mathcal{P}_1$), one can turn to apply local-search algorithms to solve its factored problem ($\mathcal{F}_1$) efficiently. In this sense, as a reformulation of the convex programme ($\mathcal{P}_1$), the non-convex optimization problem ($\mathcal{F}_1$) inherits all the statistical performance bounds for ($\mathcal{P}_1$). Cabral *et al.* [10] worked on a similar problem and showed all global optima of ($\mathcal{F}_1$) corresponds to the solution of the convex programme ($\mathcal{P}_1$). The work [28] applied the factorization approach to a more broad class of problems. When specialized to matrix inverse problems, their results show that any local minimizer U and V with zero columns is a global minimum for the over-parameterization case, i.e. $r > \text{rank}(X^\star)$. However, there are no results discussing the existence of spurious local minima or the degenerate saddles in these previous works. We extend these works and further prove that as long as the loss function $f(X)$ is restricted well-conditioned, all local minima are global minima, and there are no degenerate saddles

with no requirement on the dimension of the variables.We finally note that compared with [28], our result (Theorem 1.7) does not depend on the existence of zero columns at the critical points, and hence can provide guarantees for many local-search algorithms.

1.6 *Notations*

Denote $[n]$ as the collection of all positive integers up to n. The symbols $\mathbf{I}$ and $\mathbf{0}$ are reserved for the identity matrix and zero matrix/vector, respectively. A subscript is used to indicate its dimension when this is not clear from context. We call a matrix PSD, denoted by $X \succeq 0$, if it is symmetric and all its eigenvalues are non-negative. The notation $X \succeq Y$ means $X - Y \succeq 0$, i.e. $X - Y$ is PSD. The set of $r \times r$ orthogonal matrices is denoted by $\mathbb{O}_r = \{R \in \mathbb{R}^{r\times r} : RR^\top = \mathbf{I}_r\}$. Matrix norms, such as the spectral, nuclear and Frobenius norms, are denoted by $\|\cdot\|$, $\|\cdot\|_*$ and $\|\cdot\|_F$, respectively.

The gradient of a scalar function $f(Z)$ with a matrix variable $Z \in \mathbb{R}^{m\times n}$ is an $m \times n$ matrix, whose (i,j)th entry is $[\nabla f(Z)]_{i,j} = \frac{\partial f(Z)}{\partial Z_{ij}}$ for $i \in [m], j \in [n]$. Alternatively, we can view the gradient as a linear form $[\nabla f(Z)](G) = \langle \nabla f(Z), G\rangle = \sum_{i,j} \frac{\partial f(Z)}{\partial Z_{ij}} G_{ij}$ for any $G \in \mathbb{R}^{m\times n}$. The Hessian of $f(Z)$ can be viewed as a fourth-order tensor of dimension $m \times n \times m \times n$, whose (i,j,k,l)th entry is $[\nabla^2 f(Z)]_{i,j,k,l} = \frac{\partial^2 f(Z)}{\partial Z_{ij}\partial Z_{k,l}}$ for $i,k \in [m], j,l \in [n]$. Similar to the linear form representation of the gradient, we can view the Hessian as a bilinear form defined via $[\nabla^2 f(Z)](G,H) = \sum_{i,j,k,l} \frac{\partial^2 f(Z)}{\partial Z_{ij}\partial Z_{kl}} G_{ij}H_{kl}$ for any $G, H \in \mathbb{R}^{m\times n}$. Yet another way to represent the Hessian is as an $mn \times mn$ matrix $[\nabla^2 f(Z)]_{i,j} = \frac{\partial^2 f(Z)}{\partial z_i \partial z_j}$ for $i,j \in [mn]$, where z_i is the ith entry of the vectorization of Z. We will use these representations interchangeably whenever the specific form can be inferred from context. For example, in the restricted well-conditioned assumption ($\mathcal{C}$), the Hessian is apparently viewed as an $n^2 \times n^2$ matrix and the identity $\mathbf{I}$ is of dimension $n^2 \times n^2$.

For a matrix-valued function $\phi : \mathbb{R}^{p\times q} \to \mathbb{R}^{m\times n}$, it is notationally easier to represent its gradient (or Jacobian) and Hessian as multi-linear operators. For example, the gradient, as a linear operator from $\mathbb{R}^{p\times q}$ to $\mathbb{R}^{m\times n}$, is defined via $[\nabla[\phi(U)](G)]_{ij} = \sum_{k\in[p],\, l\in[q]} \frac{\partial[\phi(U)]_{ij}}{\partial U_{kl}} G_{kl}$ for $i \in [m]$, $j \in [n]$ and $G \in \mathbb{R}^{p\times q}$; the Hessian, as a bilinear operator from $\mathbb{R}^{p\times q} \times \mathbb{R}^{p\times q}$ to $\mathbb{R}^{m\times n}$, is defined via $[\nabla^2[\phi(U)](G,H)]_{ij} = \sum_{k_1,k_2\in[p],\, l_1,l_2\in[q]} \frac{\partial^2[\phi(U)]_{ij}}{\partial U_{k_1 l_1}\partial U_{k_2 l_2}} G_{k_1 l_1} H_{k_2 l_2}$ for $i \in [m]$, $j \in [n]$ and $G, H \in \mathbb{R}^{p\times q}$. Using this notation, the Hessian of the scalar function $f(Z)$ of the previous paragraph, which is also the gradient of $\nabla f(Z) : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$, can be viewed as a linear operator from $\mathbb{R}^{m\times m}$ to $\mathbb{R}^{m\times n}$ denoted by $[\nabla^2 f(Z)](G)$ and satisfies $\langle [\nabla^2 f(Z)](G), H\rangle = [\nabla^2 f(Z)](G,H)$ for $G, H \in \mathbb{R}^{m\times n}$.

2. Problem formulation

This work considers two problems: (i) the minimization of a general convex function $f(X)$ with the domain being positive semi-definite matrices, and (ii) the minimization of a general convex function $f(X)$ regularized by the matrix nuclear norm $\|X\|_*$ with the domain being general matrices. Let $X^\star$ be an optimal solution of ($\mathcal{P}_0$) or ($\mathcal{P}_1$) of rank $r^\star$. To develop faster and scalable algorithms, we apply Burer–Monteiro-style parameterization [9] to the low-rank optimization variable X in ($\mathcal{P}_0$) and ($\mathcal{P}_1$):

$$\text{for symmetric case:} \quad X = \phi(U) := UU^\top,$$

$$\text{for non-symmetric case:} \quad X = \psi(U,V) := UV^\top,$$

where $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{m\times r}$ with $r \geq r^\star$. With the optimization variable X being parameterized, the convex programmes are transformed into the factored problems ($\mathcal{F}_0$)–($\mathcal{F}_1$):

$$\text{for symmetric case:} \quad \underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}}\ g(U) = f\big(\phi(U)\big),$$

$$\text{for non-symmetric case:} \quad \underset{U\in\mathbb{R}^{n\times r},\, V\in\mathbb{R}^{m\times r}}{\text{minimize}}\ g(U,V) = f\big(\psi(U,V)\big) + \frac{\lambda}{2}\Big(\|U\|_F^2 + \|V\|_F^2\Big).$$

Inspired by the lifting technique in constructing SDP relaxations, we refer to the variable X as the lifted variable, and the variables U, V as the factored variables. Similar naming conventions apply to the optimization problems, their domains and objective functions.

2.1 *Consequences of the restricted well-conditioned assumption*

First the restricted well-conditioned assumption reduces to (1.3) when the objective function is quadratic. Moreover, the restricted well-conditioned assumption ($\mathcal{C}$) shares a similar spirit with (1.3) in that the operator $\frac{2}{\beta+\alpha}[\nabla^2 f(X)]$ preserves geometric structure for low-rank matrices:

PROPOSITION 2.1 Let $f(X)$ satisfy the restricted well-conditioned assumption ($\mathcal{C}$). Then

$$\left|\frac{2}{\beta+\alpha}\Big[\nabla^2 f(X)\Big](G,H) - \langle G,H\rangle\right| \leq \frac{\beta-\alpha}{\beta+\alpha}\|G\|_F\|H\|_F \leq \frac{1}{5}\|G\|_F\|H\|_F \tag{2.1}$$

for any matrices X, G, H of rank at most $2r$.

Proof. We extend the argument in [**11**] to a general function $f(X)$. If either G or H is zero, (2.1) holds since both sides are zero. For non-zero G and H, we can assume $\|G\|_F = \|H\|_F = 1$ without loss of generality.[5] Then the assumption ($\mathcal{C}$) implies

$$\alpha\,\|G-H\|_F^2 \leq \Big[\nabla^2 f(X)\Big](G-H,G-H) \leq \beta\,\|G-H\|_F^2,$$
$$\alpha\,\|G+H\|_F^2 \leq \Big[\nabla^2 f(X)\Big](G+H,G+H) \leq \beta\,\|G+H\|_F^2.$$

Thus, we have

$$\left|2\Big[\nabla^2 f(X)\Big](G,H) - \big(\beta+\alpha\big)\langle G,H\rangle\right| \leq \frac{\beta-\alpha}{2}\underbrace{\Big(\|G\|_F^2 + \|H\|_F^2\Big)}_{=2} = \beta-\alpha = \big(\beta-\alpha\big)\underbrace{\|G\|_F\|H\|_F}_{=1}.$$

We complete the proof by dividing both sides by $\beta+\alpha$:

$$\left|\frac{2}{\beta+\alpha}\Big[\nabla^2 f(X)\Big](G,H) - \langle G,H\rangle\right| \leq \frac{\beta-\alpha}{\beta+\alpha}\|G\|_F\|H\|_F \leq \frac{\beta/\alpha-1}{\beta/\alpha+1}\|G\|_F\|H\|_F \leq \frac{1}{5}\|G\|_F\|H\|_F,$$

where in the last inequality we use the assumption that $\beta/\alpha \leq 1.5$. □

[5] Otherwise, we can divide both sides of the equation (2.1) by $\|G\|_F\|H\|_F$, and use the homogeneity to get an equivalent version of Proposition 2.1 with $G = G/\|G\|_F$ and $H = H/\|H\|_F$, i.e. $\|G\|_F = \|H\|_F = 1$.

Another immediate consequence of this assumption is that if the original convex programme ($\mathcal{P}_0$) has an optimal solution $X^\star$ with $\text{rank}\left(X^\star\right) \leq r$, then there is no other optimum of ($\mathcal{P}_0$) of rank less than or equal to r:

PROPOSITION 2.2 Suppose the function $f(X)$ satisfies the restricted well-conditioned assumption ($\mathcal{C}$). Let $X^\star$ be an optimum of ($\mathcal{P}_0$) with $\text{rank}(X^\star) \leq r$. Then $X^\star$ is the unique global optimum of ($\mathcal{P}_0$) of rank at most r.

Proof. For the sake of a contradiction, suppose there exists another optimum X of ($\mathcal{P}_0$) with $\text{rank}(X) \leq r$ and $X \neq X^\star$. We begin with the second-order Taylor expansion, which reads

$$f(X) = f\left(X^\star\right) + \left\langle \nabla f\left(X^\star\right), X - X^\star \right\rangle + \frac{1}{2}\left[\nabla^2 f\left(tX^\star + (1-t)X\right)\right]\left(X - X^\star, X - X^\star\right),$$

for some $t \in [0, 1]$. The Karush-Kuhn-Tucker (KKT) conditions for the convex optimization problem ($\mathcal{P}_0$) state that $\nabla f(X^\star) \succeq 0$ and $\nabla f(X^\star)X^\star = \mathbf{0}$, implying that the second term in the above Taylor expansion

$$\left\langle \nabla f(X^\star), X - X^\star \right\rangle = \left\langle \nabla f(X^\star), X \right\rangle \geq 0,$$

since X is feasible, and hence PSD. Further, since $\text{rank}(tX^\star + (1-t)X) \leq \text{rank}(X) + \text{rank}(X^\star) \leq 2r$ and similarly $\text{rank}(X - X^\star) \leq 2r < 4r$, then from the restricted well-conditioned assumption ($\mathcal{C}$) we have

$$\left[\nabla^2 f(\tilde{X})\right]\left(X - X^\star, X - X^\star\right) \geq \alpha \left\|X - X^\star\right\|_F^2.$$

Combining all, we obtain a contradiction when $X \neq X^\star$:

$$f(X) \geq f(X^\star) + \frac{1}{2}\alpha \left\|X - X^\star\right\|_F^2 \geq f(X) + \frac{1}{2}\alpha \left\|X - X^\star\right\|_F^2 > f(X),$$

where the second inequality follows from the optimality of $X^\star$ and the third inequality holds for any $X \neq X^\star$. □

At a high level, the proof essentially depends on the restricted strongly convexity of the objective function of the convex programme ($\mathcal{P}_0$), which is guaranteed by the restricted well-conditioned assumption ($\mathcal{C}$) on $f(X)$. The similar argument holds for ($\mathcal{P}_1$) by noting that the sum of a (restricted) strongly convex function and a standard convex function is still (restricted) strongly convex. However, showing this requires a slightly more complicated argument due to the non-smoothness of $\|X\|_*$ around those non-singular matrices. Mainly, we need to use the concept of subgradient.

PROPOSITION 2.3 Suppose the function $f(X)$ satisfies the restricted well-conditioned assumption ($\mathcal{C}$). Let $X^\star$ be a global optimum of ($\mathcal{P}_1$) with $\text{rank}(X^\star) \leq r$. Then $X^\star$ is the unique global optimum of ($\mathcal{P}_1$) of rank at most r.

Proof. For the sake of contradiction, suppose that there exists another optimum X of ($\mathcal{P}_1$) with $\text{rank}(X) \leq r$ and $X \neq X^\star$. We begin with the second-order Taylor expansion of $f(X)$, which reads

$$f(X) = f\left(X^\star\right) + \left\langle \nabla f\left(X^\star\right), X - X^\star \right\rangle + \frac{1}{2}\left[\nabla^2 f\left(tX^\star + (1-t)X\right)\right]\left(X - X^\star, X - X^\star\right)$$

for some $t \in [0, 1]$. From the convexity of $\|X\|_*$, for any $D \in \partial \|X^\star\|_*$, we also have

$$\|X\|_* \geq \|X^\star\|_* + \langle D, X - X^\star \rangle.$$

Combining both, we obtain

$$\begin{aligned}
f(X) + \lambda\|X\|_* &\overset{①}{\geq} f\left(X^\star\right) + \lambda \left\|X^\star\right\|_* + \left\langle \nabla f\left(X^\star\right) + \lambda D, X - X^\star \right\rangle \\
&\quad + \frac{1}{2}\left[\nabla^2 f\big(tX^\star + (1-t)X\big)\right]\left(X - X^\star, X - X^\star\right) \\
&\overset{②}{\geq} f\left(X^\star\right) + \lambda \left\|X^\star\right\|_* + \frac{1}{2}\left[\nabla^2 f\big(tX^\star + (1-t)X\big)\right]\left(X - X^\star, X - X^\star\right) \\
&\overset{③}{\geq} f\left(X^\star\right) + \lambda \left\|X^\star\right\|_* + \frac{1}{2}\alpha \left\|X - X^\star\right\|_F^2 \\
&\overset{④}{=} f(X) + \lambda\|X\|_* + \frac{1}{2}\alpha \left\|X - X^\star\right\|_F^2 \\
&\overset{⑤}{>} f(X) + \lambda\|X\|_*,
\end{aligned}$$

where ① holds for any $D \in \partial \|X^\star\|_*$. For ②, we use fact that $\partial f_1 + \partial f_2 = \partial (f_1 + f_2)$ for any convex functions f_1, f_2, to obtain that $\nabla f(X^\star) + \lambda \partial \|X^\star\|_* = \partial(f(X^\star) + \lambda\|X^\star\|_*)$, which includes $\mathbf{0}$ since $X^\star$ is a global optimum of ($\mathcal{P}_1$). Therefore, ② follows by choosing $D \in \partial\|X^\star\|_*$ such that $\nabla f(X^\star) + \lambda D = \mathbf{0}$. ③ uses the restricted well-conditioned assumption ($\mathcal{C}$) as $\operatorname{rank}(tX^\star + (1-t)X) \leq 2r$ and $\operatorname{rank}(X - X^\star) \leq 4r$. ④ comes from the assumption that both X and $X^\star$ are global optimal solutions of ($\mathcal{P}_1$). ⑤ uses the assumption that $X \neq X^\star$. □

3. Understanding the factored landscapes for PSD matrices

In the convex programme ($\mathcal{P}_0$), we minimize a convex function $f(X)$ over the PSD cone. Let $X^\star$ be an optimal solution of ($\mathcal{P}_0$) of rank $r^\star$. We re-parameterize the low-rank PSD variable X as

$$X = \phi(U) = UU^\top,$$

where $U \in \mathbb{R}^{n \times r}$ with $r \geq r^\star$ is a rectangular, matrix square root of X. After this parameterization, the convex programme is transformed into the factored problem ($\mathcal{F}_0$) whose objective function is $g(U) = f(\phi(U))$.

3.1 *Transforming the landscape for PSD matrices*

Our primary interest is to understand how the landscape of the lifted objective function $f(X)$ is transformed by the factored parameterization $\phi(U) = UU^\top$, particularly how its global optimum is mapped to the factored space, how other types of critical points are introduced and what their properties are.

We show that if the function $f(X)$ is restricted well-conditioned, then each critical point of the factored objective function $g(U)$ in ($\mathcal{F}_0$) either corresponds to the low-rank global solution of the original

convex programme ($\mathcal{P}_0$) or is a strict saddle where the Hessian $\nabla^2 g(U)$ has a strictly negative eigenvalue. This implies that the factored objective function $g(U)$ satisfies the strict saddle property.

THEOREM 3.1 (Transforming the landscape for PSD matrices) Suppose the function $f(X)$ in ($\mathcal{P}_0$) is twice continuously differentiable and is restricted well-conditioned assumption ($\mathcal{C}$). Assume $X^\star$ is an optimal solution of ($\mathcal{P}_0$) with $\operatorname{rank}(X^\star) = r^\star$. Set $r \geq r^\star$ in ($\mathcal{F}_0$). Let U be any critical point of $g(U)$ satisfying $\nabla g(U) = \mathbf{0}$. Then U either corresponds to a square-root factor of $X^\star$, i.e.

$$X^\star = UU^\top,$$

or is a strict saddle of the factored problem ($\mathcal{F}_0$). More precisely, let $U^\star \in \mathbb{R}^{n\times r}$ such that $X^\star = U^\star U^{\star\top}$ and set $D = U - U^\star R$ with $R = \operatorname{argmin}_{R:R\in\mathbb{O}_r} \|U - U^\star R\|_F^2$, then the curvature of $\nabla^2 g(U)$ along D is strictly negative:

$$\left[\nabla^2 g(U)\right](D, D) \leq \begin{cases} -0.24\alpha \min\left\{\rho(U)^2, \rho(X^\star)\right\} \|D\|_F^2 & \text{when } r > r^\star; \\ -0.19\alpha\rho(X^\star)\|D\|_F^2 & \text{when } r = r^\star; \\ -0.24\alpha\rho(X^\star)\|D\|_F^2 & \text{when } U = \mathbf{0} \end{cases}$$

with $\rho(\cdot)$ denoting the smallest non-zero singular value of its argument. This further implies

$$\lambda_{\min}\left(\nabla^2 g(U)\right) \leq \begin{cases} -0.24\alpha \min\left\{\rho(U)^2, \rho\left(X^\star\right)\right\} & \text{when } r > r^\star; \\ -0.19\alpha\rho(X^\star) & \text{when } r = r^\star; \\ -0.24\alpha\rho(X^\star) & \text{when } U = \mathbf{0}. \end{cases}$$

Several remarks follow. First, the matrix D is the direction from the saddle point U to its closest globally optimal factor $U^\star R$ of the same dimension as U. Secondly, our result covers both over-parameterization where $r > r^\star$ and exact parameterization where $r = r^\star$. Thirdly, we can recover the rank-$r^\star$ global minimizer $X^\star$ of ($\mathcal{P}_0$) by running local-search algorithms on the factored function $g(U)$ if we know an upper bound on the rank $r^\star$. In particular, to apply the results in [32] where the first-order algorithms are proved to escape all the strict saddles, aside from the strict saddle property, one needs $g(U)$ to have a Lipschitz continuous gradient, i.e. $\|\nabla g(U) - \nabla g(V)\|_F \leq L_c\|U - V\|_F$ or $\|\nabla^2 g(U)\| \leq L_c$ for some positive constant L_c (also known as the Lipschitz constant). As indicated by the expression of $\nabla^2 g(U)$ in (3.5), it is possible that one cannot find such a constant L_c for the whole space. Similar to [30] which considers the low-rank matrix factorization problem, suppose the local-search algorithm starts at U_0 and sequentially decreases the objective value (which is true as long as the algorithm obeys certain sufficient decrease property [55]). Then it is adequate to focus on the sublevel set of g

$$\mathcal{L}_{U_0} = \left\{U : g(U) \leq g(U_0)\right\}, \tag{3.1}$$

and show that g has a Lipschitz gradient on $\mathcal{L}_{U_0}$. This is formally established in Proposition 3.2, whose proof is given in Appendix A.

PROPOSITION 3.2 Under the same setting as in Theorem 3.1, for any initial point U_0, $g(U)$ on $\mathcal{L}_{U_0}$ defined in (3.1) has a Lipschitz continuous gradient with the Lipschitz constant

$$L_c = \sqrt{2\beta\sqrt{\frac{2}{\alpha}\left(f\left(U_0U_0^T\right) - f(X^\star)\right)} + 2\left\|\nabla f(X^\star)\right\|_F + 4\beta\left(\|U^\star\|_F + \frac{\sqrt{\frac{2}{\alpha}\left(f\left(U_0U_0^T\right) - f(X^\star)\right)}}{2\left(\sqrt{2}-1\right)\rho(U^\star)}\right)^2},$$

where $\rho(\cdot)$ denotes the smallest non-zero singular value of its argument.

3.2 *Metrics in the lifted and factored spaces*

Before continuing this geometry-based argument, it is essential to have a good understanding of the domain of the factored problem and establish a metric for this domain. Since for any U, $\phi(U) = \phi(UR)$ where $R \in \mathbb{O}_r$, the domain of the factored objective function $g(U)$ is stratified into equivalence classes and can be viewed as a quotient manifold [1]. The matrices in each of these equivalence classes differ by an orthogonal transformation (not necessarily unique when the rank of U is less than r). One implication is that, when working in the factored space, we should consider all factorizations of $X^\star$:

$$\mathcal{A}^\star = \left\{U^\star \in \mathbb{R}^{n\times r} : \phi(U^\star) = X^\star\right\}.$$

A second implication is that when considering the distance between two points U_1 and U_2, one should use the distance between their corresponding equivalence classes:

$$\mathrm{d}(U_1, U_2) = \min_{R_1\in\mathbb{O}_r, R_2\in\mathbb{O}_r} \|U_1R_1 - U_2R_2\|_F = \min_{R\in\mathbb{O}_r} \|U_1 - U_2R\|_F. \tag{3.2}$$

Under this notation, $\mathrm{d}\left(U, U^\star\right) = \min_{R\in\mathbb{O}_r} \left\|U - U^\star R\right\|_F$ represents the distance between the class containing a critical point $U \in \mathbb{R}^{n\times r}$ and the optimal factor class $\mathcal{A}^\star$. The second minimization problem in the definition (3.2) is known as the orthogonal Procrustes problem, where the global optimum R is characterized by the following lemma:

LEMMA 3.3 [29] An optimal solution for the orthogonal Procrustes problem

$$R = \operatorname*{argmin}_{\tilde{R}\in\mathbb{O}_r} \left\|U_1 - U_2\tilde{R}\right\|_F^2 = \operatorname*{argmax}_{\tilde{R}\in\mathbb{O}_r} \left\langle U_1, U_2\tilde{R}\right\rangle$$

is given by $R = LP^\top$, where the orthogonal matrices $L, P \in \mathbb{R}^{r\times r}$ are defined via the singular value decomposition of $U_2^\top U_1 = L\Sigma P^\top$. Moreover, we have $U_1^\top U_2 R = (U_2R)^\top U_1 \succeq 0$ and $\langle U_1, U_2R\rangle = \|U_1^\top U_2\|_*$.

For any two matrices $U_1, U_2 \in \mathbb{R}^{n\times r}$, the following lemma relates the distance $\left\|U_1U_1^\top - U_2U_2^\top\right\|_F$ in the lifted space to the distance $\mathrm{d}(U_1, U_2)$ in the factored space. The proof is deferred to Appendix B.

LEMMA 3.4 Assume that $U_1, U_2 \in \mathbb{R}^{n\times r}$. Then

$$\left\| U_1U_1^\top - U_2U_2^\top \right\|_F \geq \min\left\{\rho(U_1), \rho(U_2)\right\}\mathrm{d}(U_1, U_2).$$

In particular, when one matrix is of full rank, we have a similar but tighter result to relate these two distances.

LEMMA 3.5 [53, Lemma 5.4] Assume that $U_1, U_2 \in \mathbb{R}^{n\times r}$ and $\mathrm{rank}(U_1) = r$. Then

$$\left\| U_1U_1^\top - U_2U_2^\top \right\|_F \geq 2(\sqrt{2}-1)\rho(U_1)\mathrm{d}(U_1, U_2).$$

3.3 *Proof idea: connecting the optimality conditions*

The proof is inspired by connecting the optimality conditions for the two programmes ($\mathcal{P}_0$) and ($\mathcal{F}_0$). First of all, as the critical points of the convex optimization problem ($\mathcal{P}_0$), they are global optima and are characterized by the necessary and sufficient KKT conditions [8]

$$\nabla f\big(X^\star\big) \succeq 0, \nabla f\big(X^\star\big)X^\star = \mathbf{0}, X^\star \succeq 0. \tag{3.3}$$

The factored optimization problem ($\mathcal{F}_0$) is unconstrained, with the critical points being specified by the zero gradient condition

$$\nabla g(U) = 2\nabla f\big(\phi(U)\big)U = \mathbf{0}. \tag{3.4}$$

To classify the critical points of ($\mathcal{F}_0$), we compute the Hessian quadratic form $[\nabla^2 g(U)](D, D)$ as

$$\left[\nabla^2 g(U)\right](D, D) = 2\left\langle \nabla f\big(\phi(U)\big), DD^\top \right\rangle + \left[\nabla^2 f\big(\phi(U)\big)\right]\big(DU^\top + UD^\top, DU^\top + UD^\top\big). \tag{3.5}$$

Roughly speaking, the Hessian quadratic form has two terms—the first term involves the gradient of $f(X)$ and the Hessian of $\phi(U)$, while the second term involves the Hessian of $f(X)$ and the gradient of $\phi(U)$. Since $\phi(U + D) = \phi(U) + UD^\top + DU^\top + DD^\top$, the gradient of ϕ is the linear operator $[\nabla\phi(U)](D) = UD^\top + DU^\top$ and the Hessian bilinear operator applies as $\frac{1}{2}[\nabla^2\phi(U)](D, D) = DD^\top$. Note in (3.5) the second quadratic form is always non-negative since $\nabla^2 f \succeq 0$ due to the convexity of f.

For any critical point U of $g(U)$, the corresponding lifted variable $X := UU^\top$ is PSD and satisfies $\nabla f(X)X = \mathbf{0}$. On one hand, if X further satisfies $\nabla f(X) \succeq 0$, then in view of the KKT conditions (3.3) and noting $\mathrm{rank}(X) = \mathrm{rank}(U) \leq r$, we must have $X = X^\star$, the global optimum of ($\mathcal{P}_0$). On the other hand, if $X \neq X^\star$, implying $\nabla f(X) \not\succeq 0$ due to the necessity of (3.3), then additional critical points can be introduced into the factored space. Fortunately, $\nabla f(X) \not\succeq 0$ also implies that the first quadratic form in (3.5) might be negative for a properly chosen direction D. To sum up, the critical points of $g(U)$ can be classified into two categories: the global optima in the optimal factor set $\mathcal{A}^\star$ with $\nabla f(UU^\top) \succeq 0$ and those with $\nabla f(UU^\top) \not\succeq 0$. For the latter case, by choosing a proper direction D, we will argue that the Hessian quadratic form (3.5) has a strictly negative eigenvalue, and hence moving in the direction of D

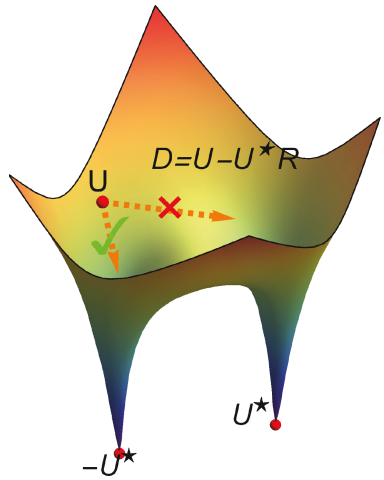


FIG. 2. The matrix $D = U - U^\star R$ is the direction from the critical point U to its nearest optimal factor $U^\star R$, whose norm $\|U - U^\star R\|_F$ defines the distance $\mathrm{d}(U, U^\star)$. Here, U is closer to $-U^\star$ than $U^\star$, and the direction from U to $-U^\star$ has more negative curvature compared to the direction from U to $U^\star$.

in a short distance will decrease the value of $g(U)$, implying that they are strict saddles and are not local minima.

We argue that a good choice of D is the direction from the current U to its closest point in the optimal factor set $\mathcal{A}^\star$. Formally, $D = U - U^\star R$ where $R = \operatorname{argmin}_{R:R\in\mathbb{O}_r} \|U - U^\star R\|_F$ is the optimal rotation for the orthogonal Procrustes problem. As illustrated in Fig. 2 where we have two global solutions $U^\star$ and $-U^\star$ and U is closer to $-U^\star$, the direction from U to $-U^\star$ has more negative curvature compared to the direction from U to $U^\star$.

Plugging this choice of D into the first term of (3.5), we simplify it as

$$\begin{aligned}\left\langle \nabla f(UU^\top), DD^\top \right\rangle &= \left\langle \nabla f(UU^\top), U^\star U^{\star\top} - U^\star R U^\top - U(U^\star R)^\top + UU^\top \right\rangle \\ &= \left\langle \nabla f(UU^\top), U^\star U^{\star\top} \right\rangle \\ &= \left\langle \nabla f(UU^\top), U^\star U^{\star\top} - UU^\top \right\rangle, \end{aligned} \tag{3.6}$$

where both the second line and last line follow from the critical point property $\nabla f(UU^\top)U = \mathbf{0}$. To gain some intuition on why (3.6) is negative while the second term in (3.5) remains small, we consider a simple example: the matrix PCA problem.

Matrix PCA Problem. Consider the PCA problem for symmetric PSD matrices

$$\underset{X\in\mathbb{R}^{n\times n}}{\text{minimize}}\ f_{\text{PCA}}(X) := \frac{1}{2}\left\|X - X^\star\right\|_F^2 \ \text{subject to}\ X \succeq 0, \tag{3.7}$$

where $X^\star$ is a symmetric PSD matrix of rank $r^\star$. Trivially, the optimal solution is $X = X^\star$. Now consider the factored problem

$$\underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}}\, g(U) := f_{\text{PCA}}(UU^\top) = \frac{1}{2}\left\|UU^\top - U^\star U^{\star\top}\right\|_F^2,$$

where $U^\star \in \mathbb{R}^{n\times r}$ satisfies $\phi(U^\star) = X^\star$. Our goal is to show that any critical point U such that $X := UU^\top \neq X^\star$ is a strict saddle.

Controlling the first term. Since $\nabla f_{\text{PCA}}(X) = X - X^\star$, by (3.6), the first term of $[\nabla^2 g(U)](D, D)$ in (3.5) becomes

$$2\left\langle \nabla f_{\text{PCA}}(X), DD^\top\right\rangle = 2\left\langle \nabla f_{\text{PCA}}(X), X^\star - X\right\rangle = 2\left\langle X - X^\star, X^\star - X\right\rangle = -2\left\|X - X^\star\right\|_F^2, \tag{3.8}$$

which is strictly negative when $X \neq X^\star$.

Controlling the second term. We show that the second term $[\nabla^2 f(\phi(U))](DU^\top + UD^\top, DU^\top + UD^\top)$ vanishes by showing that $DU^\top = \mathbf{0}$ (hence, $UD^\top = \mathbf{0}$). For this purpose, let $X^\star = Q\text{diag}(\boldsymbol{\lambda})Q^\top = \sum_{i=1}^{r^\star}\lambda_i \mathbf{q}_i\mathbf{q}_i^\top$ be the eigenvalue decomposition of $X^\star$, where $Q = \begin{bmatrix}\mathbf{q}_1 & \cdots & \mathbf{q}_{r^\star}\end{bmatrix} \in \mathbb{R}^{n\times r^\star}$ has orthonormal columns and $\boldsymbol{\lambda} \in \mathbb{R}^{r^\star}$ is composed of positive entries. Similarly, let $\phi(U) = V\text{diag}(\boldsymbol{\mu})V^\top = \sum_{i=1}^{r'}\mu_i\mathbf{v}_i\mathbf{v}_i^\top$ be the eigenvalue decomposition of $\phi(U)$, where $r' = \text{rank}(U)$. The critical point U satisfies $-\nabla g(U) = 2\big(X^\star - \phi(U)\big)U = \mathbf{0}$, implying that

$$\mathbf{0} = \left(X^\star - \sum_{i=1}^{r'}\mu_i\mathbf{v}_i\mathbf{v}_i^\top\right)\mathbf{v}_j = X^\star\mathbf{v}_j - \mu_j\mathbf{v}_j, j = 1, \ldots, r'.$$

This means $(\mu_j, \mathbf{v}_j)$ forms an eigenvalue–eigenvector pair of $X^\star$ for each $j = 1, \ldots, r'$. Consequently,

$$\mu_j = \lambda_{i_j} \text{ and } \mathbf{v}_j = \mathbf{q}_{i_j}, j = 1, \ldots, r'.$$

Hence, $\phi(U) = \sum_{j=1}^{r'}\lambda_{i_j}\mathbf{q}_{i_j}\mathbf{q}_{i_j}^\top = \sum_{j=1}^{r^\star}\lambda_j s_j \mathbf{q}_j\mathbf{q}_j^\top$. Here s_j is equal to either 0 or 1, indicating which of the eigenvalue–eigenvector pair $\big(\lambda_j, \mathbf{q}_j\big)$ appears in the decomposition of $\phi(U)$. Without loss of generality, we can choose $U^\star = Q\begin{bmatrix}\text{diag}(\sqrt{\boldsymbol{\lambda}}) & \mathbf{0}\end{bmatrix}$. Then $U = Q\begin{bmatrix}\text{diag}(\sqrt{\boldsymbol{\lambda}}\odot\mathbf{s}) & \mathbf{0}\end{bmatrix}V^\top$ for some orthonormal matrix $V \in \mathbb{R}^{r\times r}$ and $\mathbf{s} = \begin{bmatrix}s_1 & \cdots & s_{r^\star}\end{bmatrix}$, where the symbol $\odot$ means pointwise multiplication. By Lemma 3.3, we obtain $R = V^\top$. Plugging these into $DU^\top = UU^\top - U^\star RU^\top$ gives $DU^\top = \mathbf{0}$.

Combining the two. Hence, $[\nabla^2 g(U)](D, D)$ is simply determined by its first term

$$\begin{aligned}\left[\nabla^2 g(U)\right](D, D) &= -2\left\|UU^\top - U^\star U^{\star\top}\right\|_F^2\\ &\leq -2\min\left\{\rho(U)^2, \rho\big(U^\star\big)^2\right\}\|D\|_F^2\\ &= -2\min\left\{\rho(\phi(U)), \rho\big(X^\star\big)\right\}\|D\|_F^2\\ &= -2\rho(X^\star)\|D\|_F^2,\end{aligned}$$

where the second line follows from Lemma 3.4 and the last line follows from the fact that all the eigenvalues of $UU^\top$ come from those of $X^\star$. Finally, we obtain the desired strict saddle property of $g(U)$:

$$\lambda_{\min}\left(\nabla^2 g(U)\right) \leq -2\rho(X^\star).$$

This simple example is ideal in several ways, particularly the gradient $\nabla f(\phi(U)) = \phi(U) - \phi(U^\star)$, which directly establishes the negativity of the first term in (3.5), and by choosing $D = U - U^\star R$ and using $DU^\top = \mathbf{0}$, the second term vanishes. Neither of these simplifications hold for general objective functions $f(X)$. However, the example does suggest that the direction $D = U - U^\star R$ is a good choice to show $[\nabla^2 g(U)](D, D) \leq -\tau\|D\|_F^2$ for some $\tau > 0$. For a formal proof, we will also use the direction $D = U - U^\star R$ to show that those critical points U not corresponding to $X^\star$ have a negative directional curvature for the general factored objective function $g(U)$.

3.4 *A formal proof of Theorem* 3

Proof Outline. We present a formal proof of Theorem 3.4 in this section. The main argument involves showing each critical point U of $g(U)$ either corresponds to the optimal solution $X^\star$ or its Hessian matrix $\nabla^2 g(U)$ has at least one strictly negative eigenvalue. Inspired by the discussions in Section 3.3, we will use the direction $D = U - U^\star R$ and show that the Hessian $\nabla^2 g(U)$ has a strictly negative directional curvature in the direction of D, i.e. $[\nabla^2 g(U)](D, D) \leq -\tau\|D\|_F^2$, for some $\tau > 0$.

Supporting Lemmas. We first list two lemmas. The first lemma separates $\left\|(U - Z)U^\top\right\|_F^2$ into two terms: $\left\|UU^\top - ZZ^\top\right\|_F^2$ and $\left\|(UU^\top - ZZ^\top)QQ^\top\right\|_F^2$ with $QQ^\top$ being the projection matrix onto Range(U). It is crucial for the first term $\left\|UU^\top - ZZ^\top\right\|_F^2$ to have a small coefficient. In the second lemma, we will further control the second term as a consequence of U being a critical point. The proof of Lemma 3.6 is given in Appendix C.

LEMMA 3.6 Let U and Z be any two matrices in $\mathbb{R}^{n\times r}$ such that $U^\top Z = Z^\top U$ is PSD. Assume that Q is an orthogonal matrix whose columns span Range(U). Then

$$\left\|(U - Z)U^\top\right\|_F^2 \leq \frac{1}{8}\left\|UU^\top - ZZ^\top\right\|_F^2 + \left(3 + \frac{1}{2\sqrt{2} - 2}\right)\left\|(UU^\top - ZZ^\top)QQ^\top\right\|_F^2.$$

We remark that Lemma 3.6 is a strengthened version of [**5**, Lemma 4.4]. While the result there requires (i) U to be a critical point of the factored objective function $g(U)$, and (ii) Z to be an optimal factor in $\mathcal{A}^\star$ that is closest to U, i.e. $Z = U^\star R$ with $U^\star \in \mathcal{A}^\star$ and $R = \operatorname{argmin}_{R:RR^\top = \mathbf{I}_r}\|W - W^\star R\|_F$. Lemma 3.6 removes these assumptions and requires only $U^\top Z = Z^\top U$ being PSD.

Next, we control the distance between $UU^\top$ and the global solution $X^\star$ when U is a critical point of the factored objective function $g(U)$, i.e. $\nabla g(U) = \mathbf{0}$. The proof, given in Appendix D, relies on writing $\nabla f(X) = \nabla f(X^\star) + \int_0^1 [\nabla^2 f(tX + (1 - t)X^\star)](X - X^\star)\,\mathrm{d}t$ and applying Proposition 2.1.

LEMMA 3.7 (Upper Bound on $\left\|(UU^\top - U^\star U^{\star\top})QQ^\top\right\|_F^2$) Suppose the objective function $f(X)$ in ($\mathcal{P}_0$) is twice continuously differentiable and satisfies the restricted well-conditioned assumption ($\mathcal{C}$). Further, let U be any critical point of ($\mathcal{F}_0$) and Q be the orthonormal basis spanning Range(U). Then

$$\left\|(UU^\top - U^\star U^{\star\top})QQ^\top\right\|_F \leq \frac{\beta - \alpha}{\beta + \alpha}\left\|UU^\top - U^\star U^{\star\top}\right\|_F.$$

Proof of Theorem 3.1 Along the same lines as in the matrix PCA example, it suffices to find a direction D to produce a strictly negative curvature for each critical point U not corresponding to $X^\star$. We choose $D = U - U^\star R$ where $R = \operatorname{argmin}_{R:RR^\top = \mathbf{I}_r} \|W - W^\star R\|_F$. Then

$$\begin{aligned}
&\left[\nabla^2 g(U)\right](D, D) \\
&= 2\left\langle \nabla f(X), DD^\top \right\rangle + \left[\nabla^2 f(X)\right]\left(DU^\top + UD^\top, DU^\top + UD^\top\right) && \text{By Eq. (3.5)} \\
&= 2\left\langle \nabla f(X), X^\star - X \right\rangle + \left[\nabla^2 f(X)\right]\left(DU^\top + UD^\top, DU^\top + UD^\top\right) && \text{By Eq. (3.4)} \\
&\le \underbrace{2\left\langle \nabla f(X) - \nabla f(X^\star), X^\star - X \right\rangle}_{\Pi_1} + \underbrace{\left[\nabla^2 f(X)\right]\left(DU^\top + UD^\top, DU^\top + UD^\top\right)}_{\Pi_2}. && \text{By Eq. (3.3)}
\end{aligned}$$

In the following, we will bound Π_1 and Π_2, respectively.

Bounding Π_1.

$$\begin{aligned}
\Pi_1 = -2\left\langle \nabla f(X^\star) - \nabla f(X), X^\star - X \right\rangle &\overset{①}{=} -2\left\langle \int_0^1 \left[\nabla^2 f\big(tX + (1-t)X^\star\big)\right](X^\star - X)\,\mathrm{d}t,\, X^\star - X \right\rangle \\
&= -2\int_0^1 \left[\nabla^2 f\big(tX + (1-t)X^\star\big)\right](X^\star - X,\, X^\star - X)\,\mathrm{d}t \\
&\overset{②}{\le} -2\alpha\|X^\star - X\|_F^2,
\end{aligned}$$

where ① follows from the Taylor's Theorem for vector-valued functions [**39**, Eq. (2.5) in Theorem 2.1], and ② follows from the restricted strong convexity assumption ($\mathcal{C}$) since the PSD matrix $tX + (1-t)X^\star$ has rank of at most $2r$ and $\operatorname{rank}(X^\star - X) \le 4r$.

Bounding Π_2.

$$\begin{aligned}
\Pi_2 &= \left[\nabla^2 f(X)\right]\left(DU^\top + UD^\top, DU^\top + UD^\top\right) \\
&\le \beta \left\| DU^\top + UD^\top \right\|_F^2 && \text{By } (\mathcal{C}) \\
&\le 4\beta \left\| DU^\top \right\|_F^2 \\
&\le 4\beta \left[\frac{1}{8}\|X - X^\star\|_F^2 + \left(3 + \frac{1}{2\sqrt{2}-2}\right) \left\| (X - X^\star)QQ^\top \right\|_F^2 \right]. && \text{By Lemma 3.6} \\
&\le 4\beta \left[\frac{1}{8} + \left(3 + \frac{1}{2\sqrt{2}-2}\right) \frac{(\beta-\alpha)^2}{(\beta+\alpha)^2} \right] \|X - X^\star\|_F^2 && \text{By Lemma 3.7} \\
&\le 1.76\alpha \left\| X^\star - X \right\|_F^2. && \text{By } \beta/\alpha \le 1.5
\end{aligned}$$

Combining the two. Hence,

$$\Pi_1 + \Pi_2 \leq -0.24\alpha \left\|X^\star - X\right\|_F^2.$$

Then, we relate the lifted distance $\|X^\star - X\|_F^2$ with the factored distance $\left\|U - U^\star R\right\|_F^2$ using Lemma 3.4 when $r > r^\star$, and Lemma 3.5 when $r = r^\star$, respectively:

$$\begin{aligned}\text{When } r > r^\star: \left[\nabla^2 g(U)\right](D, D) &\leq -0.24\alpha \min\left\{\rho(U)^2, \rho\left(U^\star\right)^2\right\}\|D\|_F^2 && \text{By Lemma 3.4}\\ &= -0.24\alpha \min\left\{\rho(U)^2, \rho\left(X^\star\right)\right\}\|D\|_F^2.\end{aligned}$$

$$\begin{aligned}\text{When } r = r^\star: \left[\nabla^2 g(U)\right](D, D) &\leq -0.19\alpha\rho\left(U^\star\right)^2\|D\|_F^2 && \text{By Lemma 3.5}\\ &= -0.19\alpha\rho\left(X^\star\right)\|D\|_F^2.\end{aligned}$$

For the special case where $U = \mathbf{0}$, we have

$$\begin{aligned}\left[\nabla^2 g(U)\right](D, D) &\leq -0.24\alpha \left\|\mathbf{0} - X^\star\right\|_F^2\\ &= -0.24\alpha\left\|U^\star U^{\star\top}\right\|_F^2\\ &\leq -0.24\alpha\rho(U^\star)^2\left\|U^\star\right\|_F^2\\ &= -0.24\alpha\rho(X^\star)\|D\|_F^2,\end{aligned}$$

where the last second line follows from

$$\left\|U^\star U^{\star\top}\right\|_F^2 = \sum_i \sigma_i^4(U^\star) = \sum_{i:\sigma_i(U^\star)\neq 0} \sigma_i^4(U^\star) \geq \min_{i:\sigma_i(U^\star)\neq 0} \sigma_i^2(U^\star)\left(\sum_{j:\sigma_j(U^\star)\neq 0} \sigma_j^2\left(U^\star\right)\right) = \rho^2(U^\star)\left\|U^\star\right\|_F^2,$$

and the last line follows from $D = \mathbf{0} - U^\star R = -U^\star R$ when $U = \mathbf{0}$. Here $\sigma_i(\cdot)$ denotes the ith largest singular value of its argument. □

4. Understanding the factored landscapes for general non-squared matrices

In this section, we will study the second convex programme ($\mathcal{P}_1$): the minimization of a general convex function $f(X)$ regularized by the matrix nuclear norm $\|X\|_*$ with the domain being general matrices. Since the matrix nuclear norm $\|X\|_*$ appears in the objective function, the standard convex solvers or even faster tailored ones require performing singular value decomposition in each iteration, which severely limits the efficiency and scalability of the convex programme. Motivated by this, we will instead solve its Burer–Monteiro re-parameterized counterpart.

4.1 *Burer–Monteiro reformulation of the nuclear norm regularization*

Recall the second problem is the nuclear norm regularization ($\mathcal{P}_1$):

$$\underset{X\in\mathbb{R}^{n\times m}}{\text{minimize}}\ f(X)+\lambda\|X\|_* \text{ where } \lambda>0. \tag{$\mathcal{P}_1$}$$

This convex programme has an equivalent SDP formulation [**43**, p. 8]:

$$\underset{X\in\mathbb{R}^{n\times m},\,\Phi\in\mathbb{R}^{n\times n},\,\Psi\in\mathbb{R}^{m\times m}}{\text{minimize}} f(X)+\frac{\lambda}{2}\left(\text{trace}(\Phi)+\text{trace}(\Psi)\right)\ \text{ subject to }\ \begin{bmatrix}\Phi & X\\ X^\top & \Psi\end{bmatrix}\succeq 0. \tag{4.1}$$

When the PSD constraint is implicitly enforced as the following equality constraint

$$\begin{bmatrix}\Phi & X\\ X^\top & \Psi\end{bmatrix}=\begin{bmatrix}U\\ V\end{bmatrix}\begin{bmatrix}U\\ V\end{bmatrix}^\top \Rightarrow X=UV^\top, \Phi=UU^\top, \Psi=VV^\top, \tag{4.2}$$

we obtain the Burer–Monteiro factored reformulation ($\mathcal{F}_1$):

$$\underset{U\in\mathbb{R}^{n\times r},\,V\in\mathbb{R}^{m\times r}}{\text{minimize}}\ g(U,V)=f(UV^\top)+\frac{\lambda}{2}\left(\|U\|_F^2+\|V\|_F^2\right). \tag{$\mathcal{F}_1$}$$

The factored formulation ($\mathcal{F}_1$) can potentially solve the computational issue of ($\mathcal{P}_1$) in two major respects: (i) avoiding expensive SVDs by replacing the nuclear norm $\|X\|_*$ with the squared term $(\|U\|_F^2+\|V\|_F^2)/2$; and (ii) a substantial reduction in the number of the optimization variables from nm to $(n+m)r$.

4.2 *Transforming the landscape for general non-square matrices*

Our primary interest is to understand how the landscape of the lifted objective function $f(X)+\lambda\|X\|_*$ is transformed by the factored parameterization $\psi(U,V)=UV^\top$. The main contribution of this part is establishing that under the restricted well-conditioned assumption of the convex loss function $f(X)$, the factored formulation ($\mathcal{F}_1$) has no spurious local minima and satisfies the strict saddle property.

THEOREM 4.1 (Transforming the landscape for general non-square matrices) Suppose the function $f(X)$ satisfies the restricted well-conditioned property ($\mathcal{C}$). Assume that $X^\star$ of rank $r^\star$ is an optimal solution of ($\mathcal{P}_1$) where $\lambda>0$. Set $r\geq r^\star$ in the factored programme ($\mathcal{F}_1$). Let (U,V) be any critical point of $g(U,V)$ satisfying $\nabla g(U,V)=\mathbf{0}$. Then (U,V) either corresponds to a factorization of $X^\star$, i.e.

$$X^\star=UV^\top,$$

or is a strict saddle of the factored problem

$$\lambda_{\min}\left(\nabla^2 g(U,V)\right) \leq \begin{cases} -0.12\alpha \min\left\{0.5\rho^2(W), \rho(X^\star)\right\} & \text{when } r > r^\star; \\ -0.099\alpha\rho(X^\star) & \text{when } r = r^\star; \\ -0.12\alpha\rho(X^\star) & \text{when } W = \mathbf{0}, \end{cases}$$

where $W := \begin{bmatrix} U^\top & V^\top \end{bmatrix}^\top$ and $\rho(W)$ is the smallest non-zero singular value of W.

Theorem 4.1 ensures that many local-search algorithms[6] when applied for solving the factored programme ($\mathcal{F}_1$) can escape from all the saddle points and converge to a global solution that corresponds to $X^\star$. Several remarks follow.

The Non-triviality of Extending the PSD Case to the Non-symmetric Case. Although the generalization from the PSD case might not seem technically challenging at first sight, we must overcome several technical difficulties to prove this main theorem. We make a few other technical contributions in the process. In fact, the non-triviality of extending to the non-symmetric case is also highlighted in [**36**,**42**,**53**]. The major technique difficulty to complete such an extension is the ambiguity issue existed in the non-symmetric case: $UV^\top = (tU)(1/tV)^\top$ for any non-zero t. This tends to make the factored quadratic objective function badly conditioned, especially when t is very large or small. To prevent this from happening, a popular strategy utilized to adapt the result for the symmetric case to the non-symmetric case is to introduce an additional balancing regularization to ensure that U and V have equal energy [**36**,**42**,**53**]. Sometimes these additional regularizations are quite complicated (see Eq. (13)–(15) in [**51**]). Instead, we find for nuclear norm regularized problems, the critical points are automatically balanced even without these additional complex balancing regularizations (see Section 4.4 for details). In addition, by connecting the optimality conditions of the convex programme ($\mathcal{P}_1$) and the factored programme ($\mathcal{F}_1$), we dramatically simplify the proof argument, making the relationship between the original convex problem and the factored programme more transparent.

Proof Sketch of Theorem 4.1. We try to understand how the parameterization $X = \psi(U, V)$ transforms the geometric structures of the convex objective function $f(X)$ by categorizing the critical points of the non-convex factored function $g(U, V)$. In particular, we will illustrate how the globally optimal solution of the convex programme is transformed in the domain of $g(U, V)$. Furthermore, we will explore the properties of the additional critical points introduced by the parameterization and find a way of utilizing these properties to prove the strict saddle property. For those purposes, the optimality conditions for the two programmes ($\mathcal{P}_1$) and ($\mathcal{F}_1$) will be compared. □

4.3 *Optimality condition for the convex programme*

As an unconstrained convex optimization, all critical points of ($\mathcal{P}_1$) are global optima, and are characterized by the necessary and sufficient KKT condition [**8**]:

$$\nabla f(X^\star) \in -\lambda \partial \|X^\star\|_*, \tag{4.3}$$

[6] The Lipschitz gradient of g at any of its sublevel set can be obtained with similar approach for Proposition 3.2.

where $\partial\|X^\star\|_*$ denotes the subdifferential (the set of subgradient) of the nuclear norm $\|X\|_*$ evaluated at $X^\star$. The subdifferential of the matrix nuclear norm is defined by

$$\partial\|X\|_* = \left\{D \in \mathbb{R}^{n\times m} : \|Y\|_* \geq \|X\|_* + \langle Y - X, D\rangle, \text{ all } Y \in \mathbb{R}^{n\times m}\right\}.$$

We have a more explicit characterization of the subdifferential of the nuclear norm using the singular value decomposition. More specifically, suppose $X = P\Sigma Q^\top$ is the (compact) singular value decomposition of $X \in \mathbb{R}^{n\times m}$ with $P \in \mathbb{R}^{n\times r}, Q \in \mathbb{R}^{m\times r}$ and Σ being an $r \times r$ diagonal matrix. Then the subdifferential of the matrix nuclear norm at X is given by [**43**, Equation (2.9)]

$$\partial\|X\|_* = \left\{PQ^\top + E : P^\top E = \mathbf{0}, EQ = \mathbf{0}, \|E\| \leq 1\right\}.$$

Combining this representation of the subdifferential and the KKT condition (4.3) yields an equivalent expression for the optimality condition

$$\begin{aligned} \nabla f(X^\star)Q^\star &= -\lambda P^\star, \\ \nabla f(X^\star)^\top P^\star &= -\lambda Q^\star, \\ \left\|\nabla f(X^\star)\right\| &\leq \lambda, \end{aligned} \tag{4.4}$$

where we assume the compact SVD of $X^\star$ is given by

$$X^\star = P^\star\Sigma^\star Q^{\star\top} \text{ with } P^\star \in \mathbb{R}^{n\times r^\star}, Q^\star \in \mathbb{R}^{m\times r^\star}, \Sigma^\star \in \mathbb{R}^{r^\star\times r^\star}.$$

Since $r \geq r^\star$ in the factored problem ($\mathcal{F}_1$), to match the dimensions, we define the optimal factors $U^\star \in \mathbb{R}^{n\times r}$, $V^\star \in \mathbb{R}^{m\times r}$ for any $R \in \mathbb{O}_r$ as

$$\begin{aligned} U^\star &= P^\star\left[\sqrt{\Sigma^\star}\,\mathbf{0}_{r^\star\times(r-r^\star)}\right]R, \\ V^\star &= Q^\star\left[\sqrt{\Sigma^\star}\,\mathbf{0}_{r^\star\times(r-r^\star)}\right]R. \end{aligned} \tag{4.5}$$

Consequently, with the optimal factors $U^\star, V^\star$ defined in (4.5), we can rewrite the optimal condition (4.4) as

$$\begin{aligned} \nabla f(X^\star)V^\star &= -\lambda U^\star, \\ \nabla f(X^\star)^\top U^\star &= -\lambda V^\star, \\ \left\|\nabla f(X^\star)\right\| &\leq \lambda. \end{aligned} \tag{4.6}$$

Stacking $U^\star, V^\star$ as $W^\star = \begin{bmatrix} U^\star \\ V^\star \end{bmatrix}$ and defining

$$\Xi(X) := \begin{bmatrix} \lambda\mathbf{I} & \nabla f(X) \\ \nabla f(X)^\top & \lambda\mathbf{I} \end{bmatrix} \quad \text{for all } X \tag{4.7}$$

yield a more concise form of the optimality condition:

$$\begin{aligned} \Xi(X^\star)W^\star &= \mathbf{0}, \\ \left\|\nabla f(X^\star)\right\| &\leq \lambda. \end{aligned} \tag{4.8}$$

4.4 *Characterizing the critical points of the factored programme*

To begin with, the gradient of $g(U, V)$ can be computed and rearranged as

$$\begin{aligned} \nabla g(U,V) &= \begin{bmatrix} \nabla_U g(U,V) \\ \nabla_V g(U,V) \end{bmatrix} \\ &= \begin{bmatrix} \nabla f(UV^\top)V + \lambda U \\ \nabla f(UV^\top)^\top U + \lambda V \end{bmatrix} \\ &= \begin{bmatrix} \lambda \mathbf{I} & \nabla f(UV^\top) \\ \nabla f(UV^\top)^\top & \lambda \mathbf{I} \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix} \\ &= \Xi(UV^\top) \begin{bmatrix} U \\ V \end{bmatrix}, \end{aligned} \tag{4.9}$$

where the last equality follows from the definition (4.7) of $\Xi(\cdot)$. Therefore, all critical points of $g(U, V)$ can be characterized by the following set:

$$\mathcal{X} := \left\{ (U,V) : \Xi(UV^\top) \begin{bmatrix} U \\ V \end{bmatrix} = \mathbf{0} \right\}.$$

We will see that any critical point $(U, V) \in \mathcal{X}$ forms a balanced pair, which is defined as follows:

DEFINITION 4.2 (Balanced pairs) We call (U, V) is a balanced pair if the Gram matrices of U and V are the same: $U^\top U - V^\top V = \mathbf{0}$. All the balanced pairs form the balanced set, denoted by $\mathcal{E} := \{(U, V) : U^\top U - V^\top V = \mathbf{0}\}$.

By Definition 4.2, to show that each critical point forms a balanced pair, we rely on the following fact:

$$W = \begin{bmatrix} U \\ V \end{bmatrix}, \widehat{W} = \begin{bmatrix} U \\ -V \end{bmatrix} \text{ with } (U,V) \in \mathcal{E} \Leftrightarrow \widehat{W}^\top W = W^\top \widehat{W} = U^\top U - V^\top V = \mathbf{0}. \tag{4.10}$$

Now we are ready to relate the critical points and balanced pairs; the proof of which is given in Appendix E.

PROPOSITION 4.3 Any critical point $(U, V) \in \mathcal{X}$ forms a balanced pair in $\mathcal{E}$.

4.4.1 *The properties of the balanced set* In this part, we introduce some important properties of the balanced set $\mathcal{E}$. These properties basically compare the on-diagonal-block energy and the off-diagonal-block energy for a certain block matrix. Hence, it is necessary to introduce two operators defined on block matrices:

$$\begin{aligned}\mathcal{P}_{\text{on}}\left(\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\right) &:= \begin{bmatrix}A_{11} & \mathbf{0}\\ \mathbf{0} & A_{22}\end{bmatrix},\\ \mathcal{P}_{\text{off}}\left(\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\right) &:= \begin{bmatrix}\mathbf{0} & A_{12}\\ A_{21} & \mathbf{0}\end{bmatrix},\end{aligned} \tag{4.11}$$

for any matrices $A_{11} \in \mathbb{R}^{n\times n}, A_{12} \in \mathbb{R}^{n\times m}, A_{21} \in \mathbb{R}^{m\times n}, A_{22} \in \mathbb{R}^{m\times m}$.

According to the definitions of $\mathcal{P}_{\text{on}}$ and $\mathcal{P}_{\text{off}}$ in (4.11), when $\mathcal{P}_{\text{on}}$ and $\mathcal{P}_{\text{off}}$ are acting on the product of two block matrices $W_1W_2^\top$,

$$\begin{aligned}\mathcal{P}_{\text{on}}\left(W_1W_2^\top\right) &= \mathcal{P}_{\text{on}}\left(\begin{bmatrix}U_1U_2^\top & U_1V_2^\top\\ V_1U_2^\top & V_1V_2^\top\end{bmatrix}\right) = \begin{bmatrix}U_1U_2^\top & \mathbf{0}\\ \mathbf{0} & V_1V_2^\top\end{bmatrix} = \frac{W_1W_2^\top + \widehat{W}_1\widehat{W}_2^\top}{2},\\ \mathcal{P}_{\text{off}}\left(W_1W_2^\top\right) &= \mathcal{P}_{\text{on}}\left(\begin{bmatrix}U_1U_2^\top & U_1V_2^\top\\ V_1U_2^\top & V_1V_2^\top\end{bmatrix}\right) = \begin{bmatrix}\mathbf{0} & U_1V_2^\top\\ V_1U_2^\top & \mathbf{0}\end{bmatrix} = \frac{W_1W_2^\top - \widehat{W}_1\widehat{W}_2^\top}{2}.\end{aligned} \tag{4.12}$$

Here, to simplify the notations, for any $U_1, U_2 \in \mathbb{R}^{n\times r}$ and $V_1, V_2 \in \mathbb{R}^{m\times r}$, we define

$$W_1 = \begin{bmatrix}U_1\\ V_1\end{bmatrix}, \qquad \widehat{W}_1 = \begin{bmatrix}U_1\\ -V_1\end{bmatrix}, \qquad W_2 = \begin{bmatrix}U_2\\ V_2\end{bmatrix}, \qquad \widehat{W}_2 = \begin{bmatrix}U_2\\ -V_2\end{bmatrix}.$$

Now, we are ready to present the properties regarding the set $\mathcal{E}$ in Lemma 4.4 and Lemma 4.5, whose proofs are given in Appendix F and Appendix G, respectively.

LEMMA 4.4 Let $W = \begin{bmatrix}U^\top & V^\top\end{bmatrix}^\top$ with $(U, V) \in \mathcal{E}$. Then for every $D = \begin{bmatrix}D_U^\top & D_V^\top\end{bmatrix}^\top$ of proper dimension, we have

$$\left\|\mathcal{P}_{\text{on}}(DW^\top)\right\|_F^2 = \left\|\mathcal{P}_{\text{off}}(DW^\top)\right\|_F^2.$$

LEMMA 4.5 Let $W_1 = \begin{bmatrix}U_1^\top & V_1^\top\end{bmatrix}^\top$, $W_2 = \begin{bmatrix}U_2^\top & V_2^\top\end{bmatrix}^\top$ with $(U_1, V_1), (U_2, V_2) \in \mathcal{E}$. Then

$$\left\|\mathcal{P}_{\text{on}}\left(W_1W_1^\top - W_2W_2^\top\right)\right\|_F^2 \le \left\|\mathcal{P}_{\text{off}}\left(W_1W_1^\top - W_2W_2^\top\right)\right\|_F^2.$$

4.5 *Proof idea: connecting the optimality conditions*

First observe that each $(U^\star, V^\star)$ in (4.5) is a global optimum for the factored programme (we prove this in Appendix H):

PROPOSITION 4.6 Any $(U^\star, V^\star)$ in (4.5) is a global optimum of the factored programme ($\mathcal{F}_1$):

$$g(U^\star, V^\star) \le g(U, V), \text{for all } U \in \mathbb{R}^{n\times r}, V \in \mathbb{R}^{m\times r}.$$

However, due to non-convexity, only characterizing the global optima is not enough for the factored programme to achieve the global convergence by many local-search algorithms. One should also eliminate the possibility of the existence of spurious local minima or degenerate saddles. For this purpose, we focus on the critical point set $\mathcal{X}$ and observe that any critical point $(U, V) \in \mathcal{X}$ of the factored problem satisfies the first part of the optimality condition (4.8):

$$\Xi(X)W = \mathbf{0}$$

by constructing $W = [U^\top \ V^\top]^\top$ and $X = UV^\top$. If the critical point (U, V) additionally satisfies $\|\nabla f(UV^\top)\| \leq \lambda$, then it corresponds to the global optimum $X^\star = UV^\top$.

Therefore, it remains to study the additional critical points (which are introduced by the parameterization $X = \psi(U, V)$) that violate $\|\nabla f(UV^\top)\| \leq \lambda$. In fact, we intend to show the following: for any critical point (U, V), if $X^\star \neq UV^\top$, we can find a direction D, in which the Hessian $\nabla^2 g(U, V)$ has a strictly negative curvature $[\nabla^2 g(U, V)](D, D) < -\tau \|D\|_F^2$ for some $\tau > 0$. Hence, every critical point (U, V) either corresponds to the global optimum $X^\star$ or is a strict saddle point.

To gain more intuition, we take a closer look at the directional curvature of $g(U, V)$ in some direction $D = \left[D_U^\top \ D_V^\top\right]^\top$:

$$\left[\nabla^2 g(U, V)\right](D, D) = \left\langle \Xi(X), DD^\top \right\rangle + \left[\nabla^2 f(X)\right]\left(D_U V^\top + UD_V^\top, D_U V^\top + UD_V^\top\right), \tag{4.13}$$

where the second term is always non-negative by the convexity of f. The sign of the first term $\langle \Xi(X), DD^\top \rangle$ depends on the positive semi-definiteness of $\Xi(X)$, which is related to the boundedness condition $\|\nabla f(X)\| \leq \lambda$ through the Schur complement theorem [**8**, A.5.5]:

$$\Xi(X) \succeq 0 \Leftrightarrow \lambda \mathbf{I} - \frac{1}{\lambda} \nabla f(X)^\top \nabla f(X) \succeq 0 \Leftrightarrow \|\nabla f(X)\| \leq \lambda.$$

Equivalently, whenever $\|\nabla f(X)\| > \lambda$, we have $\Xi(X) \not\succeq 0$. Therefore, for those non-globally optimal critical points (U, V), it is possible to find a direction D such that the first term $\langle \Xi(X), DD^\top \rangle$ is strictly negative. Inspired by the weighted PCA example, we choose D as the direction from the critical point $W = \left[U^\top \ V^\top\right]^\top$ to the nearest globally optimal factor $W^\star R$ with $W^\star = \left[U^{\star\top} \ V^{\star\top}\right]^\top$, i.e.

$$D = W - W^\star R,$$

where $R = \operatorname{argmin}_{R:RR^\top = \mathbf{I}_r} \left\|W - W^\star R\right\|_F$. We will see that with this particular D, the first term of (4.13) will be strictly negative while the second term retains small.

4.6 *A formal proof of Theorem* 4.1

The main argument involves choosing D as the direction from $W = \left[U^\top \ V^\top\right]^\top$ to its nearest optimal factor: $D = W - W^\star R$ with $R = \operatorname{argmin}_{R:RR^\top = \mathbf{I}_r} \|W - W^\star R\|_F$, and showing that the Hessian $\nabla^2 g(U, V)$ has a strictly negative curvature in the direction of D whenever $W \neq W^\star$. To that end, we first introduce the following lemma (with its proof in Appendix I) connecting the distance

$\left\|UV^\top - X^\star\right\|_F$ and the distance $\left\|(WW^\top - W^\star W^{\star\top})QQ^\top\right\|_F$ (where $QQ^\top$ is an orthogonal projector onto the Range(W)).

LEMMA 4.7 Suppose the function $f(X)$ in ($\mathcal{P}_1$) is restricted well-conditioned ($\mathcal{C}$). Let $W = \begin{bmatrix} U^\top & V^\top \end{bmatrix}^\top$ with $(U, V) \in \mathcal{X}$, $W^\star = \begin{bmatrix} U^{\star\top} & V^{\star\top} \end{bmatrix}^\top$ correspond to the global optimum of ($\mathcal{P}_1$) and $QQ^\top$ be the orthogonal projector onto Range(W). Then

$$\left\|\left(WW^\top - W^\star W^{\star\top}\right)QQ^\top\right\|_F \le 2\frac{\beta-\alpha}{\beta+\alpha}\|UV^\top - X^\star\|_F.$$

Proof of Theorem 4.1 Let $D = W - W^\star R$ with $R = \operatorname{argmin}_{R:RR^\top=\mathbf{I}_r} \|W - W^\star R\|_F$. Then

$$\begin{aligned}
&\left[\nabla^2 g(U,V)\right](D,D)\\
&= \left\langle \Xi(X), DD^\top\right\rangle + \left[\nabla^2 f(X)\right]\left(D_U V^\top + UD_V^\top, D_U V^\top + UD_V^\top\right)\\
&\overset{①}{=} \left\langle \Xi(X), W^\star W^{\star\top} - WW^\top\right\rangle + \left[\nabla^2 f(X)\right]\left(D_U V^\top + UD_V^\top, D_U V^\top + UD_V^\top\right)\\
&\overset{②}{\le} \left\langle \Xi(X) - \Xi\left(X^\star\right), W^\star W^{\star\top} - WW^\top\right\rangle + \left[\nabla^2 f(X)\right]\left(D_U V^\top + UD_V^\top, D_U V^\top + UD_V^\top\right)\\
&= \left\langle \begin{bmatrix} \lambda\mathbf{I} & \nabla f(X)\\ \nabla f(X)^\top & \lambda\mathbf{I}\end{bmatrix} - \begin{bmatrix} \lambda\mathbf{I} & \nabla f(X^\star)\\ \nabla f\left(X^\star\right)^\top & \lambda\mathbf{I}\end{bmatrix}, W^\star W^{\star\top} - WW^\top\right\rangle\\
&\quad + \left[\nabla^2 f(X)\right]\left(D_U V^\top + UD_V^\top, D_U V^\top + UD_V^\top\right)\\
&\overset{③}{=} \left\langle \begin{bmatrix} \mathbf{0} & \int_0^1 \left[\nabla^2 f\left(X^\star + t(X-X^\star)\right)\right]\left(X - X^\star\right)\mathrm{d}t\\ * & \mathbf{0}\end{bmatrix}, W^\star W^{\star\top} - WW^\top\right\rangle\\
&\quad + \left[\nabla^2 f(X)\right]\left(D_U V^\top + UD_V^\top, D_U V^\top + UD_V^\top\right)\\
&= -2\int_0^1 \left[\nabla^2 f\left(X^\star + t(X-X^\star)\right)\right]\left(X - X^\star, X - X^\star\right)\mathrm{d}t + \left[\nabla^2 f(X)\right]\left(D_U V^\top + UD_V^\top, D_U V^\top + UD_V^\top\right),
\end{aligned}$$

where ① follows from $\nabla g(U,V) = \Xi(X)W = \mathbf{0}$ and (4.9). For ②, we note that $\langle \Xi(X^\star), W^\star W^{\star\top} - WW^\top\rangle \le 0$ since $\Xi(X^\star)W^\star = \mathbf{0}$ in (4.8) and $\Xi(X^\star) \succeq 0$ by the optimality condition. For ③, we first use $* = \left(\int_0^1 [\nabla^2 f(X^\star + t(X-X^\star))](X-X^\star)\,\mathrm{d}t\right)^\top$ for convenience and then it follows from the Taylor's Theorem for vector-valued functions [**39**, Eq. (2.5) in Theorem 2.1]:

$$\nabla f(X) - \nabla f\left(X^\star\right) = \int_0^1 \left[\nabla^2 f\left(X^\star + t(X-X^\star)\right)\right]\left(X - X^\star\right)\,\mathrm{d}t.$$

Now, we continue the argument:

$$
\begin{aligned}
&\left[\nabla^2 g(U,V)\right](D,D) \\
&\leq -2\int_0^1 \left[\nabla^2 f\left(X^\star + t\left(X - X^\star\right)\right)\right]\left(X - X^\star, X - X^\star\right) \mathrm{d}t \\
&\quad + \left[\nabla^2 f(X)\right]\left(D_U V^\top + U D_V^\top, D_U V^\top + U D_V^\top\right) \\
&\overset{④}{\leq} -2\alpha \left\|X^\star - X\right\|_F^2 + \beta \left\|D_U V^\top + U D_V^\top\right\|_F^2, \\
&\overset{⑤}{\leq} -0.5\alpha \left\|WW^\top - W^\star W^{\star\top}\right\|_F^2 + 2\beta \left(\left\|D_U V^\top\right\|_F^2 + \left\|U D_V^\top\right\|_F^2\right) \\
&\overset{⑥}{=} -0.5\alpha \left\|WW^\top - W^\star W^{\star\top}\right\|_F^2 + \beta \left\|DW^\top\right\|_F^2 \\
&\overset{⑦}{\leq} \left[-0.5\alpha + \beta/8 + 4.208\beta\left(\frac{\beta-\alpha}{\beta+\alpha}\right)^2\right]\left\|WW^\top - W^\star W^{\star\top}\right\|_F^2 \\
&\overset{⑧}{\leq} -0.06\alpha \left\|WW^\top - W^\star W^{\star\top}\right\|_F^2 \\
&\overset{⑨}{\leq} \begin{cases} -0.06\alpha \min\left\{\rho^2(W), \rho^2\left(W^\star\right)\right\}\|D\|_F^2, & \text{By Lemma 3.4 when } r > r^\star \\ -0.0495\alpha\rho^2\left(W^\star\right)\|D\|_F^2, & \text{By Lemma 3.5 when } r = r^\star \\ -0.06\alpha\rho^2\left(W^\star\right)\|D\|_F^2, & \text{when } W = \mathbf{0}, \end{cases}
\end{aligned}
$$

where ④ uses the restricted well-conditioned assumption ($\mathcal{C}$) since $\operatorname{rank}(X^\star + t(X - X^\star)) \leq 2r$, $\operatorname{rank}(X - X^\star) \leq 4r$ and $\operatorname{rank}\left(D_U V^\top + U D_V^\top\right) \leq 4r$. ⑤ comes from Lemma 4.5 and the fact $\left\|x + y\right\|_F^2 \leq 2\left(\left\|x\right\|_F^2 + \left\|y\right\|_F^2\right)$. ⑥ follows from Lemma 4.4. ⑦ first uses Lemma 3.6 to bound $\left\|DW^\top\right\|_F^2 = \left\|(W - W^\star R)W^\top\right\|_F^2$ since $W^\top W^\star \succeq 0$ and then uses Lemma 4.7 to further bound $\left\|(W^\star - W)QQ^\top\right\|_F^2$. ⑧ holds when $\beta/\alpha \leq 1.5$. ⑨ uses the similar argument as in the proof of Theorem 3.1 to relate the lifted distance and factored distance. Particularly, three possible cases are considered: (i) $r > r^\star$, (ii) $r = r^\star$ and (iii) $W = \mathbf{0}$. We apply Lemma 3.4 to Case (i) and Lemma 3.5 to Case (ii). For the third case that $W = \mathbf{0}$, we obtain from ⑧ that

$$
\left[\nabla^2 g(U,V)\right](D,D) \leq -0.06\alpha \left\|W^\star W^{\star\top}\right\|_F^2 \leq -0.06\alpha\rho\left(W^\star\right)^2 \|W^\star\|_F^2 = -0.06\alpha\rho\left(W^\star\right)^2 \|D\|_F^2,
$$

where the last equality follows from $D = \mathbf{0} - W^\star R = -W^\star R$ because $W = \mathbf{0}$.

The final result follows from the definition of $U^\star, V^\star$ in (4.5):

$$W^\star = \begin{bmatrix} P^\star\sqrt{\Sigma^\star}R \\ Q^\star\sqrt{\Sigma^\star}R \end{bmatrix} = \begin{bmatrix} P^\star/\sqrt{2} \\ Q^\star/\sqrt{2} \end{bmatrix} \left(\sqrt{2\Sigma^\star}\right) R,$$

which implies $\sigma_\ell\left(W^\star\right) = \sqrt{2\sigma_\ell\left(X^\star\right)}$. □

5. Conclusion

In this work, we considered two popular minimization problems: the minimization of a general convex function $f(X)$ with the domain being positive semi-definite matrices and the minimization of a general convex function $f(X)$ regularized by the matrix nuclear norm $\|X\|_*$, with the domain being general matrices. To improve the computational efficiency, we applied the Burer–Monteiro re-parameterization and showed that, as long as the convex function $f(X)$ is (restricted) well-conditioned, the resulting factored problems have the following properties: each critical point either corresponds to a global optimum of the original convex programmes or is a strict saddle, where the Hessian matrix has a strictly negative eigenvalue. Such a benign landscape then allows many iterative optimization methods to escape from all the saddle points and converge to a global optimum with even random initializations.

Funding

National Science Foundation (CCF-1704204 to G.T. and Q.L., CCF-1409261 to Z.Z.).

References

1. Absil, P.-A., Mahony, R. & Sepulchre, R. (2009) *Optimization Algorithms on Matrix Manifolds*. Princeton, New Jersey: Princeton University Press.
2. Anandkumar, A., Ge, R. & Janzamin, M. (2014) Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. arXiv preprint arXiv:1402.5180.
3. Balabdaoui, F. & Wellner, J. A. (2014) Chernoff's density is log-concave. *Bernoulli*, **20**, 231.
4. Bhojanapalli, S., Kyrillidis, A. & Sanghavi, S. (2016) Dropping convexity for faster semi-definite optimization. *29th Annual Conference on Learning Theory*. pp. 530–582.
5. Bhojanapalli, S., Neyshabur, B. & Srebro, N. (2016) Global optimality of local search for low rank matrix recovery. *Advances in Neural Information Processing Systems*. pp. 3873–3881.
6. Biswas, P. & Ye, Y. (2004) Semidefinite programming for ad hoc wireless sensor network localization. *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks*. Berkeley, CA, USA: Association for Computing Machinery. pp. 46–54.
7. Boumal, N., Voroninski, V. & Bandeira, A. (2016) The non-convex Burer–Monteiro approach works on smooth semidefinite programs. *Advances in Neural Information Processing Systems*. pp. 2757–2765.
8. Boyd, S. & Vandenberghe, L. (2004) *Convex Optimization*. Cambridge, England: Cambridge University Press.
9. Burer, S. & Monteiro, R. D. (2003) A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Math. Program.*, **95**, 329–357.
10. Cabral, R., De la Torre, F., Costeira, J. P. & Bernardino, A. (2013) Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2488–2495.
11. Candes, E. J. (2008) The restricted isometry property and its implications for compressed sensing. *C. R. Math.*, **346**, 589–592.

12. Candes, E. J., Eldar, Y. C., Strohmer, T. & Voroninski, V. (2015) Phase retrieval via matrix completion. *SIAM Rev.*, **57**, 225–251.

13. Candes, E. J. & Plan, Y. (2010) Matrix completion with noise. *Proc. IEEE*, **98**, 925–936.

14. Candes, E. J. & Plan, Y. (2011) Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theory*, **57**, 2342–2359.

15. Candès, E. J. & Tao, T. (2010) The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inf. Theory*, **56**, 2053–2080.

16. Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S. & Bengio, Y. (2014) Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in Neural Information Processing Systems*. pp. 2933–2941.

17. Davenport, M. A., Plan, Y., van den Berg, E. & Wootters, M. (2014) 1-bit matrix completion. *Inf. Inference*, **3**, 189–223.

18. Davenport, M. A. & Romberg, J. (2016) An overview of low-rank matrix recovery from incomplete observations. *IEEE J. Sel. Top. Signal Process.*, **10**, 608–622.

19. De Sa, C., Re, C. & Olukotun, K. (2015) Global convergence of stochastic gradient descent for some non-convex matrix problems. *International Conference on Machine Learning*. pp. 2332–2341.

20. DeCoste, D. (2006) Collaborative prediction using ensembles of maximum margin matrix factorizations. *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery. pp. 249–256.

21. Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A. & Poczos, B. (2017) Gradient descent can take exponential time to escape saddle points. *Advances in Neural Information Processing Systems*. pp. 1067–1077.

22. Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

23. Ge, R., Huang, F., Jin, C. & Yuan, Y. (2015) Escaping from saddle pointsłonline stochastic gradient for tensor decomposition. *Proceedings of the 28th Conference on Learning Theory*. pp. 797–842.

24. Ge, R., Jin, C. & Zheng, Y. (2017) No spurious local minima in nonconvex low rank problems: a unified geometric analysis. *Proceedings of the 34th International Conference on Machine Learning* (D. Precup & Y. W. Teh, eds). vol. 70 of *Proceedings of Machine Learning Research*. pp. 1233–1242, International Convention Centre, Sydney, Australia. PMLR.

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

26. Gillis, N. & Glineur, F. (2011) Low-rank matrix approximation with weights or missing data is NP-hard. *SIAM J. Matrix Anal. Appl.*, **32**, 1149–1165.

27. Gross, D., Liu, Y.-K., Flammia, S. T., Becker, S. & Eisert, J. (2010) Quantum state tomography via compressed sensing. *Physical Rev. Lett.*, **105**, 150401.

28. Haeffele, B. D. & Vidal, R. (2015) Global optimality in tensor factorization, deep learning, and beyond. arXiv preprint arXiv:1506.07540.

29. Higham, N. & Papadimitriou, P. (1995) *Matrix procrustes problems. Rapport Technique*. UK: University of Manchester.

30. Jin, C., Ge, R., Netrapalli, P., Kakade, S. M. & Jordan, M. I. (2017) How to escape saddle points efficiently. arXiv preprint arXiv:1703.00887.

31. Kyrillidis, A., Kalev, A., Park, D., Bhojanapalli, S., Caramanis, C. & Sanghavi, S. (2017) Provable quantum state tomography via non-convex methods. arXiv preprint arXiv:1711.02524.

32. Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I. & Recht, B. (2017) First-order methods almost always avoid saddle points. arXiv preprint arXiv:1710.07406.

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

34. Li, Q., Prater, A., Shen, L. & Tang, G. (2016) Overcomplete tensor decomposition via convex optimization. arXiv preprint arXiv:1602.08614.

35. Li, Q. & Tang, G. (2017) Convex and nonconvex geometries of symmetric tensor factorization. *IEEE 2017 Asilomar Conference on Signals, Systems and Computers*.

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

37. Li, Y., Sun, Y. & Chi, Y. (2017) Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements. *IEEE Trans. Signal Process.*, **65**, 397–408.

38. Murty, K. G. & Kabadi, S. N. (1987) Some NP-complete problems in quadratic and nonlinear programming. *Math. Program.*, **39**, 117–129.

39. Nocedal, J. & Wright, S. (2006) *Numerical Optimization*, 2nd edn. New York: Springer Science & Business Media.

40. Park, D., Kyrillidis, A., Bhojanapalli, S., Caramanis, C. & Sanghavi, S. (2016) Provable Burer-Monteiro factorization for a class of norm-constrained matrix problems. *Stat*., **1050**, 1.

41. Park, D., Kyrillidis, A., Caramanis, C. & Sanghavi, S. (2016) Finding low-rank solutions via non-convex matrix factorization, efficiently and provably. arXiv preprint arXiv:1606.03168.

42. Park, D., Kyrillidis, A., Carmanis, C. & Sanghavi, S. (2017) Non-square matrix sensing without spurious local minima via the Burer–Monteiro approach. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. FL, USA, pp. 65–74.

43. Recht, B., Fazel, M. & Parrilo, P. A. (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, **52**, 471–501.

44. Saumard, A. & Wellner, J. (2014) Log-concavity and strong log-concavity: a review. *Stat. Surv.*, **8**, 45.

45. Sciacchitano, F. (2017) Image reconstruction under non-Gaussian noise. *Ph.D. Thesis*, Denmark: Technical University of Denmark (DTU).

46. Sontag, E. D. & Sussmann, H. J. (1989) Backpropagation can give rise to spurious local minima even for networks without hidden layers. *Complex Syst.*, **3**, 91–106.

47. Srebro, N. & Jaakkola, T. (2003) Weighted low-rank approximations. *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (T. Fawcett & N. Mishra eds.). California: AAAI Press, pp. 720–727.

48. Sun, J. (2016) When are nonconvex optimization problems not scary? *Ph.D.Thesis*, NY, USA: Columbia University.

49. Sun, J., Qu, Q. & Wright, J. (2016) A geometric analysis of phase retrieval. *2016 IEEE International Symposium on Information Theory (ISIT)*. pp. 2379–2383.

50. Sun, J., Qu, Q. & Wright, J. (2017) Complete dictionary recovery over the sphere II: recovery by Riemannian trust-region method. *IEEE Trans. Inf. Theory*, **63**, 885–914.

51. Sun, R. & Luo, Z.-Q. (2015) Guaranteed matrix completion via nonconvex factorization. *2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*. pp. 270–289.

52. Tran-Dinh, Q. & Zhang, Z. (2016) Extended Gauss–Newton and Gauss–Newton-ADMM algorithms for low-rank matrix optimization. arXiv preprint arXiv:1606.03358.

53. Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M. & Recht, B. (2016) Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pp. 964–973.

54. Wang, L., Zhang, X. & Gu, Q. (2017) A unified computational and statistical framework for nonconvex low-rank matrix estimation. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. FL, USA, pp. 981–990.

55. Wolfe, P. (1969) Convergence conditions for ascent methods. *SIAM Rev.*, **11**, 226–235.

56. Zhao, T., Wang, Z. & Liu, H. (2015) Nonconvex low rank matrix factorization via inexact first order oracle. *Advances in Neural Information Processing Systems*.

57. Zhu, Z., Li, Q., Tang, G. & Wakin, M. B. (2017a) Global optimality in low-rank matrix optimization. arXiv preprint arXiv:1702.07945.

58. Zhu, Z., Li, Q., Tang, G. & Wakin, M. B. (2017b) The global optimization geometry of low-rank matrix optimization. arXiv preprint arXiv:1703.01256.

Appendix A. Proof of Proposition 3.2

To that end, we first show that for any $U \in \mathcal{L}_{U_0}$, $\|U\|_F$ is upper bounded. Let $X = UU^\top$ and consider the following second-order Taylor expansion of $f(X)$:

$$\begin{aligned} f(X) &= f\left(X^\star\right) + \left\langle \nabla f\left(X^\star\right), X - X^\star \right\rangle + \frac{1}{2}\int_0^1 \left[\nabla^2 f\left(tX^\star + (1-t)X\right)\right]\left(X - X^\star, X - X^\star\right)\, \mathrm{d}t \\ &\geq f\left(X^\star\right) + \frac{1}{2}\int_0^1 \left[\nabla^2 f\left(tX^\star + (1-t)X\right)\right]\left(X - X^\star, X - X^\star\right)\, \mathrm{d}t \\ &\geq f\left(X^\star\right) + \frac{\alpha}{2}\left\|X - X^\star\right\|_F^2, \end{aligned}$$

which implies that

$$\left\|UU^\top - X^\star\right\|_F^2 \leq \frac{2}{\alpha}\left(f\left(UU^\top\right) - f\left(X^\star\right)\right) \leq \frac{2}{\alpha}\left(f\left(U_0U_0^\top\right) - f\left(X^\star\right)\right) \tag{A.1}$$

with the second inequality following from the assumption $U \in \mathcal{L}_{U_0}$. Thus, we have

$$\|U\|_F \leq \left\|U^\star\right\|_F + d\left(U, U^\star\right) \leq \left\|U^\star\right\|_F + \frac{\left\|UU^\top - X^\star\right\|_F}{2\left(\sqrt{2}-1\right)\rho\left(U^\star\right)} \leq \left\|U^\star\right\|_F + \frac{\sqrt{\frac{2}{\alpha}\left(f\left(U_0U_0^\top\right) - f\left(X^\star\right)\right)}}{2\left(\sqrt{2}-1\right)\rho\left(U^\star\right)}. \tag{A.2}$$

Now we are ready to show the Lipschitz gradient for g at $\mathcal{L}_{U_0}$:

$$\begin{aligned} \left\|\nabla^2 g(U)\right\|^2 &= \max_{\|D\|_F=1}\left|\left[\nabla^2 g(U)\right](D,D)\right| \\ &= \max_{\|D\|_F=1}\left|2\left\langle \nabla f\left(UU^\top\right), DD^\top\right\rangle + \left[\nabla^2 f\left(UU^\top\right)\right]\left(DU^\top + UD^\top, DU^\top + UD^\top\right)\right| \\ &\leq 2\max_{\|D\|_F=1}\left|\left\langle \nabla f\left(UU^\top\right), DD^\top\right\rangle\right| + \max_{\|D\|_F=1}\left|\left[\nabla^2 f\left(UU^\top\right)\right]\left(DU^\top + UD^\top, DU^\top + UD^\top\right)\right| \\ &\leq 2\max_{\|D\|_F=1}\left|\left\langle \nabla f\left(UU^\top\right) - \nabla f\left(X^\star\right), DD^\top\right\rangle\right| + 2\left\|\nabla f\left(X^\star\right)\right\|_F + \beta\left\|DU^\top + UD^\top\right\|_F^2 \\ &\leq 2\beta\left\|UU^\top - X^\star\right\|_F + 2\left\|\nabla f\left(X^\star\right)\right\|_F + 4\beta\|U\|_F^2 \\ &\leq 2\beta\sqrt{\frac{2}{\alpha}\left(f\left(U_0U_0^\top\right) - f\left(X^\star\right)\right)} + 2\|\nabla f(X^\star)\|_F + 4\beta\left(\|U^\star\|_F + \frac{\sqrt{\frac{2}{\alpha}\left(f\left(U_0U_0^T\right) - f(X^\star)\right)}}{2\left(\sqrt{2}-1\right)\rho\left(U^\star\right)}\right)^2 \\ &:= L_c^2. \end{aligned}$$

The last inequality follows from (A.1) and (A.2). This concludes the proof of Proposition 3.2.

Appendix B. Proof of Lemma 3.4

Let $X_1 = U_1U_1^\top$, $X_2 = U_2U_2^\top$ and their full eigenvalue decompositions be

$$X_1 = \sum_{j=1}^n \lambda_j \mathbf{p}_j\mathbf{p}_j^\top, \qquad X_2 = \sum_{j=1}^n \eta_j \mathbf{q}_j\mathbf{q}_j^\top,$$

where $\{\lambda_j\}$ and $\{\eta_j\}$ are the eigenvalues in decreasing order. Since $\operatorname{rank}(U_1) = r_1$ and $\operatorname{rank}(U_2) = r_2$, we have $\lambda_j = 0$ for $j > r_1$ and $\eta_j = 0$ for $j > r_2$. We compute $\|X_1 - X_2\|_F^2$ as follows

$$\begin{aligned}
\|X_1 - X_2\|_F^2 &= \|X_1\|_F^2 + \|X_2\|_F^2 - 2\langle X_1, X_2\rangle \\
&= \sum_{i=1}^n \lambda_i^2 + \sum_{j=1}^n \eta_j^2 - \sum_{i=1}^n \sum_{j=1}^n 2\lambda_i \eta_j \langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 \\
&\overset{①}{=} \sum_{i=1}^n \lambda_i^2 \sum_{j=1}^n \langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 + \sum_{j=1}^n \eta_j^2 \sum_{i=1}^n \langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 - \sum_{i=1}^n \sum_{j=1}^n 2\lambda_i \eta_j \langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 \\
&\overset{②}{=} \sum_{i=1}^n \sum_{j=1}^n (\lambda_i - \eta_j)^2 \langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n \left(\sqrt{\lambda_i} - \sqrt{\eta_j}\right)^2 \left(\sqrt{\lambda_i} + \sqrt{\eta_j}\right)^2 \langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 \\
&\overset{③}{\geq} \min\left\{\sqrt{\lambda_{r_1}}, \sqrt{\eta_{r_2}}\right\}^2 \sum_{i=1}^n \sum_{j=1}^n \left(\sqrt{\lambda_i} - \sqrt{\eta_j}\right)^2 \langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 \\
&\overset{④}{=} \min\left\{\lambda_{r_1}, \eta_{r_2}\right\} \left\|\sqrt{X_1} - \sqrt{X_2}\right\|_F^2,
\end{aligned}$$

where ① uses the fact $\sum_{j=1}^n \langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 = \|\mathbf{p}_i\|_2^2 = 1$, with $\{\mathbf{q}_j\}$ being an orthonormal basis and similarly $\sum_{i=1}^n \langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 = \|\mathbf{q}_j\|_2^2 = 1$. ② is by first an exchange of the summations, secondly the fact that $\lambda_j = 0$ for $j > r_1$ and $\eta_j = 0$ for $j > r_2$ and thirdly completing squares. ③ is because $\{\lambda_j\}$ and $\{\eta_j\}$ are sorted in decreasing order. ④ follows from ② and that $\{\sqrt{\lambda_j}\}$ and $\{\sqrt{\eta_j}\}$ are eigenvalues of $\sqrt{X_1}$ and $\sqrt{X_2}$, the matrix square root of X_1 and X_2, respectively.

Finally, we can conclude the proof as long as we can show the following inequality:

$$\left\|\sqrt{X_1} - \sqrt{X_2}\right\|_F^2 \geq \min_{R:RR^\top = \mathbf{I}_r} \|U_1 - U_2 R\|_F^2. \tag{B.1}$$

By expanding $\|\cdot\|_F^2$ in (B.1) and noting that $\langle\sqrt{X_1}, \sqrt{X_1}\rangle = \operatorname{trace}(X_1) = \operatorname{trace}\left(U_1 U_1^\top\right)$ and $\langle\sqrt{X_2}, \sqrt{X_2}\rangle = \operatorname{trace}(X_2) = \operatorname{trace}\left(U_2 U_2^\top\right)$, (B.1) reduces to

$$\left\langle\sqrt{X_1}, \sqrt{X_2}\right\rangle \leq \max_{R:RR^\top = \mathbf{I}_r} \langle U_1, U_2 R\rangle. \tag{B.2}$$

To show (B.2), we write the SVDs of U_1, U_2 as $U_1 = P_1 \Sigma_1 Q_1^\top$ and $U_2 = P_2 \Sigma_2 Q_2^\top$, respectively, with $P_1, P_2 \in \mathbb{R}^{n\times r}$, $\Sigma_1, \Sigma_2 \in \mathbb{R}^{r\times r}$ and $Q_1, Q_2 \in \mathbb{R}^{r\times r}$. Then we have $\sqrt{X_1} = P_1 \Sigma_1 P_1^\top$, $\sqrt{X_2} = P_2 \Sigma_2 P_2^\top$.

On one hand,

$$\begin{aligned}
\text{Right-hand side of (B.2)} &= \max_{R:RR^\top=\mathbf{I}_r} \left\langle P_1\Sigma_1 Q_1^\top, P_2\Sigma_2 Q_2^\top R \right\rangle \\
&= \max_{R:RR^\top=\mathbf{I}_r} \left\langle P_1\Sigma_1, P_2\Sigma_2 Q_2^\top R Q_1 \right\rangle \\
&= \max_{R:RR^\top=\mathbf{I}_r} \left\langle P_1\Sigma_1, P_2\Sigma_2 R \right\rangle && \text{By } R \leftarrow Q_2^\top R Q_1 \\
&= \left\| \left(P_2\Sigma_2\right)^\top P_1\Sigma_1 \right\|_*. && \text{By Lemma 2}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\text{Left-hand side of (B.2)} &= \left\langle P_1\Sigma_1 P_1^\top, P_2\Sigma_2 P_2^\top \right\rangle \\
&= \left\langle \left(P_2\Sigma_2\right)^\top P_1\Sigma_1, P_2^\top P_1 \right\rangle \\
&\le \left\| \left(P_2\Sigma_2\right)^\top P_1\Sigma_1 \right\|_* \left\| P_2^\top P_1 \right\| && \text{By Hölder's Inequality} \\
&\le \left\| \left(P_2\Sigma_2\right)^\top P_1\Sigma_1 \right\|_*. && \text{Since } \left\| P_2^\top P_1 \right\| \le \|P_2\|\|P_1\| \le 1
\end{aligned}$$

This proves (B.2), and hence completes the proof of Lemma 3.4.

Appendix C. Proof of Lemma 3.6

The proof relies on the following lemma.

LEMMA 10 [5, Lemma E.1] Let U and Z be any two matrices in $\mathbb{R}^{n\times r}$ such that $U^\top Z = Z^\top U$ is PSD. Then

$$\left\| \left(U - Z\right)U^\top \right\|_F^2 \le \frac{1}{2\sqrt{2}-2} \left\| UU^\top - ZZ^\top \right\|_F^2.$$

Proof of Lemma 5. Define two orthogonal projectors

$$\mathcal{Q} = QQ^\top \qquad \text{and} \qquad \mathcal{Q}_\perp = Q_\perp Q_\perp^\top,$$

so $\mathcal{Q}$ is the orthogonal projector onto Range(U) and $\mathcal{Q}_\perp$ is the orthogonal projector onto the orthogonal complement of Range(U). Then

$$\begin{aligned}
\left\|(U-Z)U^\top\right\|_F^2 &\overset{①}{=} \left\|(U-\mathcal{Q}Z)U^\top\right\|_F^2 + \left\|\mathcal{Q}_\perp ZU^\top\right\|_F^2 \\
&\overset{②}{=} \left\|(U-\mathcal{Q}Z)U^\top\right\|_F^2 + \left\langle Z^\top\mathcal{Q}_\perp Z, U^\top U\right\rangle \\
&\overset{③}{\le} \frac{1}{2\sqrt{2}-2}\left\|UU^\top - (\mathcal{Q}Z)(\mathcal{Q}Z)^\top\right\|_F^2 + \left\langle Z^\top\mathcal{Q}_\perp Z, U^\top U - Z^\top\mathcal{Q}Z\right\rangle + \left\langle Z^\top\mathcal{Q}_\perp Z, Z^\top\mathcal{Q}Z\right\rangle \\
&\overset{④}{\le} \frac{1}{2\sqrt{2}-2}\left\|UU^\top - \mathcal{Q}ZZ^\top\right\|_F^2 + \left\langle Z^\top\mathcal{Q}_\perp Z, U^\top U - Z^\top\mathcal{Q}Z\right\rangle + \left\langle Z^\top\mathcal{Q}_\perp Z, Z^\top\mathcal{Q}Z\right\rangle \\
&\overset{⑤}{\le} \frac{1}{2\sqrt{2}-2}\left\|UU^\top - \mathcal{Q}ZZ^\top\right\|_F^2 + \frac{1}{8}\left\|Z^\top\mathcal{Q}_\perp Z\right\|_F^2 + 2\left\|U^\top U - Z^\top\mathcal{Q}Z\right\|_F^2 \\
&\quad + \left\langle Z^\top\mathcal{Q}_\perp Z, Z^\top\mathcal{Q}Z\right\rangle, \qquad \text{(C.1)}
\end{aligned}$$

where ① is by expressing $\left(U-Z\right)U^\top$ as the sum of two orthogonal factors $\left(U-\mathcal{Q}Z\right)U^\top$ and $-\mathcal{Q}_\perp ZU^\top$. ② is because $\left\|\mathcal{Q}_\perp ZU^\top\right\|_F^2 = \left\langle \mathcal{Q}_\perp ZU^\top, \mathcal{Q}_\perp ZU^\top\right\rangle = \left\langle \mathcal{Q}_\perp ZU^\top, ZU^\top\right\rangle = \left\langle Z^\top\mathcal{Q}_\perp Z, U^\top U\right\rangle$. ③ uses Lemma 10 by noting that $U^\top\mathcal{Q}Z = \left(\mathcal{Q}U\right)^\top Z = U^\top Z \succeq 0$ satisfying the assumptions of Lemma 10. ④ uses the fact that $\left\|UU^\top - (\mathcal{Q}Z)(\mathcal{Q}Z)^\top\right\|_F^2 = \left\|UU^\top - \mathcal{Q}ZZ^\top\mathcal{Q}\right\|_F^2 \le \left\|UU^\top - \mathcal{Q}ZZ^\top\mathcal{Q}\right\|_F^2 + \left\|\mathcal{Q}ZZ^\top\mathcal{Q}_\perp\right\|_F^2 = \left\|UU^\top - \mathcal{Q}ZZ^\top\mathcal{Q} - \mathcal{Q}ZZ^\top\mathcal{Q}_\perp\right\|_F^2 = \left\|UU^\top - \mathcal{Q}ZZ^\top\right\|_F^2$. ⑤ uses the following basic inequality that

$$\frac{1}{8}\|A\|_F^2 + 2\|B\|_F^2 \ge 2\sqrt{\frac{2}{8}\|A\|_F^2\|B\|_F^2} = \|A\|_F\|B\|_F \ge \langle A, B\rangle,$$

where $A = Z^\top\mathcal{Q}_\perp Z$ and $B = U^\top U - Z^\top\mathcal{Q}Z$.

The Remaining Steps. The remaining steps involve showing the following bounds:

$$\left\|Z^\top\mathcal{Q}_\perp Z\right\|_F^2 \le \left\|UU^\top - ZZ^\top\right\|_F^2, \qquad \text{(C.2)}$$

$$\left\langle Z^\top\mathcal{Q}_\perp Z, Z^\top\mathcal{Q}Z\right\rangle \le \left\|UU^\top - \mathcal{Q}ZZ^\top\right\|_F^2, \qquad \text{(C.3)}$$

$$\left\|U^\top U - Z^\top\mathcal{Q}Z\right\|_F^2 \le \left\|UU^\top - \mathcal{Q}ZZ^\top\right\|_F^2. \qquad \text{(C.4)}$$

This is because when plugging these bounds (C.2)–(C.4) into (C.1), we can obtain the desired result:

$$\left\|(U-Z)U^\top\right\|_F^2 \le \frac{1}{8}\left\|UU^\top - ZZ^\top\right\|_F^2 + \left(3 + \frac{1}{2\sqrt{2}-2}\right)\left\|\left(UU^\top - ZZ^\top\right)QQ^\top\right\|_F^2.$$

Showing (C.2).

$$\begin{aligned}\left\|Z^\top \mathcal{Q}_\perp Z\right\|_F^2 &= \left\langle ZZ^\top \mathcal{Q}_\perp, \mathcal{Q}_\perp ZZ^\top \right\rangle \\ &\overset{①}{=} \left\langle \mathcal{Q}_\perp ZZ^\top \mathcal{Q}_\perp, \mathcal{Q}_\perp ZZ^\top \mathcal{Q}_\perp \right\rangle \\ &= \left\| \mathcal{Q}_\perp ZZ^\top \mathcal{Q}_\perp \right\|_F^2 \\ &\overset{②}{=} \left\| \mathcal{Q}_\perp (ZZ^\top - UU^\top) \mathcal{Q}_\perp \right\|_F^2 \\ &\overset{③}{\leq} \left\| ZZ^\top - UU^\top \right\|_F^2,\end{aligned}$$

where ① follows from the idempotence property that $\mathcal{Q}_\perp = \mathcal{Q}_\perp \mathcal{Q}_\perp$. ② follows from $\mathcal{Q}_\perp U = \mathbf{0}$. ③ follows from the non-expansiveness of projection operator:

$$\left\| \mathcal{Q}_\perp \left(ZZ^\top - UU^\top \right) \mathcal{Q}_\perp \right\|_F \leq \left\| \left(ZZ^\top - UU^\top \right) \mathcal{Q}_\perp \right\|_F \leq \left\| ZZ^\top - UU^\top \right\|_F.$$

Showing (C.3). The argument here is pretty similar to that for (C.2):

$$\begin{aligned}\left\langle Z^\top \mathcal{Q}_\perp Z, Z^\top \mathcal{Q} Z \right\rangle &= \left\langle \mathcal{Q} ZZ^\top, ZZ^\top \mathcal{Q}_\perp \right\rangle \\ &= \left\langle \mathcal{Q} ZZ^\top \mathcal{Q}_\perp, \mathcal{Q} ZZ^\top \mathcal{Q}_\perp \right\rangle \\ &= \left\| \mathcal{Q} ZZ^\top \mathcal{Q}_\perp \right\|_F^2 \\ &\overset{①}{=} \left\| \mathcal{Q} \left(ZZ^\top - UU^\top \right) \mathcal{Q}_\perp \right\|_F^2 \\ &\overset{②}{\leq} \left\| \mathcal{Q} ZZ^\top - UU^\top \right\|_F^2,\end{aligned}$$

where ① is by $\mathcal{Q}_\perp U = \mathbf{0}$. ② uses the non-expansiveness of projection operator and $\mathcal{Q} UU^\top = UU^\top$.

Showing (C.4). First by expanding $\|\cdot\|_F^2$ using inner products, (C.4) is equivalent to the following inequality

$$\left\| U^\top U \right\|_F^2 + \left\| U^\top U - Z^\top \mathcal{Q} Z \right\|_F^2 - 2\left\langle U^\top U, Z^\top \mathcal{Q} Z \right\rangle \leq \left\| UU^\top \right\|_F^2 + \left\| \mathcal{Q} ZZ^\top \right\|_F^2 - 2\left\langle UU^\top, \mathcal{Q} ZZ^\top \right\rangle. \tag{C.5}$$

First of all, we recognize that

$$\left\| U^\top U \right\|_F^2 = \sum_i \sigma_i(U)^2 = \left\| UU^\top \right\|_F^2,$$

$$\left\| Z^\top \mathcal{Q} Z \right\|_F^2 = \left\langle Z^\top \mathcal{Q} Z, Z^\top \mathcal{Q} Z \right\rangle = \left\langle \mathcal{Q} ZZ^\top, ZZ^\top \mathcal{Q} \right\rangle = \left\langle \mathcal{Q} ZZ^\top \mathcal{Q}, \mathcal{Q} ZZ^\top \mathcal{Q} \right\rangle = \left\| \mathcal{Q} ZZ^\top \mathcal{Q} \right\|_F^2 \leq \left\| ZZ^\top \mathcal{Q} \right\|_F^2,$$

where we use the idempotence and non-expansiveness property of the projection matrix $\mathcal{Q}$ in the second line. Plugging these to (C.5), we find (C.5) reduces to

$$\left\langle U^\top U, Z^\top \mathcal{Q} Z \right\rangle \geq \left\langle UU^\top, \mathcal{Q} ZZ^\top \right\rangle = \left\langle UU^\top, ZZ^\top \right\rangle = \left\| U^\top Z \right\|_F^2. \tag{C.6}$$

To show (C.6), let $Q\Sigma P^\top$ be the SVD of U with $\Sigma \in \mathbb{R}^{r' \times r'}$ and $P \in \mathbb{R}^{r \times r'}$ where r' is the rank of U. Then

$$U^\top U = P\Sigma^2 P^\top, \qquad Q = UP\Sigma^{-1} \quad \text{and} \quad \mathcal{Q} = QQ^\top = UP\Sigma^{-2}P^\top U^\top. \tag{C.7}$$

Now

$$\begin{aligned}
\text{Left-hand side of (C.6)} &= \left\langle U^\top U, Z^\top \mathcal{Q} Z \right\rangle \\
&\overset{①}{=} \left\langle P\Sigma^2 P^\top, Z^\top UP\Sigma^{-2}P^\top U^\top Z \right\rangle \\
&\overset{②}{=} \left\langle \Sigma^2, P^\top \left(U^\top Z\right) P\Sigma^{-2} P^\top \left(U^\top Z\right) P \right\rangle \\
&\overset{③}{=} \left\langle \Sigma^2, G\Sigma^{-2} G \right\rangle \\
&= \left\| \Sigma G \Sigma^{-1} \right\|_F^2 \\
&\overset{④}{\geq} \|G\|_F^2 \\
&\overset{⑤}{=} \left\| U^\top Z \right\|_F^2,
\end{aligned}$$

where ① is by (C.7) and ② uses the assumption that $Z^\top U = U^\top Z \succeq 0$. In ③, we define $G := P^\top \left(U^\top Z\right) P$. ⑤ is because $\|G\|_F^2 = \left\| P^\top \left(U^\top Z\right) P \right\|_F^2 = \left\| U^\top Z \right\|_F^2$ due to the rotational invariance of $\|\cdot\|_F$. ④ is because

$$\begin{aligned}
\left\| \Sigma G \Sigma^{-1} \right\|_F^2 &= \sum_{i,j} \frac{\sigma_i^2}{\sigma_j^2} G_{ij}^2 \\
&= \sum_{i=j} G_{ii}^2 + \sum_{i>j} \left(\frac{\sigma_i^2}{\sigma_j^2} + \frac{\sigma_j^2}{\sigma_i^2} \right) G_{ij}^2 \\
&\geq \sum_{i=j} G_{ii}^2 + \sum_{i>j} 2 \left(\frac{\sigma_i}{\sigma_j} \right) \left(\frac{\sigma_j}{\sigma_i} \right) G_{ij}^2 \\
&= \sum_{i,j} G_{ij}^2 \\
&= \|G\|_F^2,
\end{aligned}$$

where the second line follows from the symmetric property of G since $G = P^\top \left(U^\top Z\right) P \succeq 0$ and $U^\top Z \succeq 0$. □

Appendix D. Proof of Lemma 3.7

Let $X = UU^\top$ and $X^\star = U^\star U^{\star\top}$. We start with the critical point condition $\nabla f(X)U = \mathbf{0}$ which implies

$$\nabla f(X)UU^\dagger = \nabla f(X)QQ^\top = \mathbf{0},$$

where † denotes the pseudoinverse. Then for all $Z \in \mathbb{R}^{n\times n}$, we have

$$\begin{aligned}
&\Rightarrow \left\langle \nabla f(X), ZQQ^\top \right\rangle = 0\\
&\overset{①}{\Rightarrow} \left\langle \nabla f(X^\star) + \int_0^1 \left[\nabla^2 f\left(tX + (1-t)X^\star\right)\right](X - X^\star)\ \mathrm{d}t, ZQQ^\top \right\rangle = 0\\
&\Rightarrow \left\langle \nabla f\left(X^\star\right), ZQQ^\top \right\rangle + \left[\int_0^1 \nabla^2 f\left(tX + (1-t)X^\star\right)\ \mathrm{d}t\right]\left(X - X^\star, ZQQ^\top\right) = 0\\
&\overset{②}{\Rightarrow} \left| -\frac{2}{\beta+\alpha}\left\langle \nabla f(X^\star), ZQQ^\top \right\rangle - \left\langle X - X^\star, ZQQ^\top \right\rangle \right| \le \frac{\beta-\alpha}{\beta+\alpha}\left\|X - X^\star\right\|_F \left\|ZQQ^\top\right\|_F\\
&\Rightarrow \left| \frac{2}{\beta+\alpha}\left\langle \nabla f(X^\star), ZQQ^\top \right\rangle + \left\langle X - X^\star, ZQQ^\top \right\rangle \right| \le \frac{\beta-\alpha}{\beta+\alpha}\left\|X - X^\star\right\|_F \left\|ZQQ^\top\right\|_F\\
&\overset{③}{\Rightarrow} \left| \frac{2}{\beta+\alpha}\left\langle \nabla f\left(X^\star\right), \left(X - X^\star\right)QQ^\top \right\rangle + \left\|\left(X - X^\star\right)QQ^\top\right\|_F^2 \right| \le \frac{\beta-\alpha}{\beta+\alpha}\left\|X - X^\star\right\|_F \left\|\left(X - X^\star\right)QQ^\top\right\|_F\\
&\overset{④}{\Rightarrow} \frac{2}{\beta+\alpha}\left\langle \nabla f\left(X^\star\right), \left(X - X^\star\right)QQ^\top \right\rangle + \left\|\left(X - X^\star\right)QQ^\top\right\|_F^2 \le \frac{\beta-\alpha}{\beta+\alpha}\left\|X - X^\star\right\|_F \left\|\left(X - X^\star\right)QQ^\top\right\|_F\\
&\Rightarrow \left\|\left(X - X^\star\right)QQ^\top\right\|_F \le \delta\left\|X - X^\star\right\|_F,
\end{aligned}$$

where ① uses the Taylor's Theorem for vector-valued functions [39, Eq. (2.5) in Theorem 2.1]. ② uses Proposition 1 by noting that the PSD matrix $[tX^\star + (1-t)X]$ has rank at most $2r$ for all $t \in [0, 1]$ and $\operatorname{rank}\left(X - X^\star\right) \le 4r, \operatorname{rank}\left(ZQQ^\top\right) \le 4r$. ③ is by choosing $Z = X - X^\star$. ④ follows from $\left\langle \nabla f\left(X^\star\right), \left(X - X^\star\right)QQ^\top \right\rangle \ge 0$ since

$$\left\langle \nabla f\left(X^\star\right), \left(X - X^\star\right)QQ^\top \right\rangle \overset{\text{(i)}}{=} \left\langle \nabla f\left(X^\star\right), X - X^\star QQ^\top \right\rangle \overset{\text{(ii)}}{=} \left\langle \nabla f\left(X^\star\right), X \right\rangle \overset{\text{(iii)}}{\ge} 0,$$

where (i) follows from $XQQ^\top = UU^\top QQ^\top = UU^\top$ since $QQ^\top$ is the orthogonal projector onto Range(U), (ii) uses the fact that

$$\nabla f\left(X^\star\right)X^\star = \mathbf{0} = X^\star \nabla f\left(X^\star\right)$$

and (iii) is because $\nabla f\left(X^\star\right) \succeq 0, X \succeq 0$. □

Appendix E. Proof of Proposition 4.3

For any critical point (U, V), we have

$$\nabla g(U, V) = \Xi\left(UV^\top\right)W = \mathbf{0},$$

where $W = \begin{bmatrix} U^\top & V^\top \end{bmatrix}^\top$. Further denote $\widehat{W} = \begin{bmatrix} U^\top & -V^\top \end{bmatrix}^\top$. Then

$$\overset{①}{\Rightarrow} \widehat{W}^\top \nabla g(U,V) + \nabla g(U,V)^\top \widehat{W} = \mathbf{0}$$

$$\overset{②}{\Rightarrow} \widehat{W}^\top \Xi\left(UV^\top\right) W + W^\top \Xi\left(UV^\top\right) \widehat{W} = \mathbf{0}$$

$$\overset{③}{\Rightarrow} \begin{bmatrix} U^\top & -V^\top \end{bmatrix} \begin{bmatrix} \lambda \mathbf{I} & \nabla f\left(UV^\top\right) \\ \nabla f\left(UV^\top\right)^\top & \lambda \mathbf{I} \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix} + \begin{bmatrix} U^\top V^\top \end{bmatrix} \begin{bmatrix} \lambda \mathbf{I} & \nabla f\left(UV^\top\right) \\ \nabla f\left(UV^\top\right)^\top & \lambda \mathbf{I} \end{bmatrix} \begin{bmatrix} U \\ -V \end{bmatrix} = \mathbf{0}$$

$$\overset{④}{\Rightarrow} \lambda\left(2U^\top U - 2V^\top V\right) + \underbrace{U^\top\left(\nabla f\left(UV^\top\right) - \nabla f\left(UV^\top\right)\right) V}_{=\mathbf{0}} + \underbrace{V^\top\left(\nabla f\left(UV^\top\right)^\top - \nabla f\left(UV^\top\right)^\top\right) U}_{=\mathbf{0}} = \mathbf{0}$$

$$\Rightarrow 2\lambda\left(U^\top U - V^\top V\right) = \mathbf{0}$$

$$\overset{⑤}{\Rightarrow} U^\top U - V^\top V = \mathbf{0},$$

where ① follows from $\nabla g(U, V) = \mathbf{0}$ and ② follows from $\nabla g(U,V) = \Xi\left(UV^\top\right) W$. ③ follows by plugging the definitions of $W, \widehat{W}$ and $\Xi(\cdot)$ into the second line. ④ follows from direct computations. ⑤ holds since $\lambda > 0$.

Appendix F. Proof of Lemma 4.4

First recall

$$W = \begin{bmatrix} U \\ V \end{bmatrix}, \qquad \widehat{W} = \begin{bmatrix} U \\ -V \end{bmatrix}, \qquad D = \begin{bmatrix} D_U \\ D_V \end{bmatrix}, \qquad \widehat{D} = \begin{bmatrix} D_U \\ -D_V \end{bmatrix}.$$

By performing the following change of variables

$$W_1 \leftarrow D, \qquad \widehat{W}_1 \leftarrow \widehat{D}, \qquad W_2 \leftarrow W, \qquad \widehat{W}_2 \leftarrow \widehat{W}$$

in (4.12), we have

$$\left\|\mathcal{P}_{\mathrm{on}}(DW^\top)\right\|_F^2 = \frac{1}{4}\left\|DW^\top + \widehat{D}\widehat{W}^\top\right\|_F^2 = \frac{1}{4}\left\langle DW^\top + \widehat{D}\widehat{W}^\top, DW^\top + \widehat{D}\widehat{W}^\top\right\rangle,$$

$$\left\|\mathcal{P}_{\mathrm{off}}(DW^\top)\right\|_F^2 = \frac{1}{4}\left\|DW^\top - \widehat{D}\widehat{W}^\top\right\|_F^2 = \frac{1}{4}\left\langle DW^\top - \widehat{D}\widehat{W}^\top, DW^\top - \widehat{D}\widehat{W}^\top\right\rangle.$$

Then it implies that

$$\begin{aligned}\left\|\mathcal{P}_{\mathrm{on}}(DW^\top)\right\|_F^2 - \left\|\mathcal{P}_{\mathrm{off}}\left(DW^\top\right)\right\|_F^2 &= \frac{1}{4}\left\langle DW^\top + \widehat{D}\widehat{W}^\top, DW^\top + \widehat{D}\widehat{W}^\top\right\rangle \\ &\quad - \frac{1}{4}\left\langle DW^\top - \widehat{D}\widehat{W}^\top, DW^\top - \widehat{D}\widehat{W}^\top\right\rangle \\ &= \left\langle DW^\top, \widehat{D}\widehat{W}^\top\right\rangle = \left\langle \widehat{D}^\top D, \widehat{W}^\top W\right\rangle = 0,\end{aligned}$$

since $\widehat{W}^\top W = \mathbf{0}$ from (4.10).

Appendix G. Proof of Lemma 4.5

To begin with, we define $\widehat{W}_1 = \begin{bmatrix} U_1 \\ -V_1 \end{bmatrix}$, $\widehat{W}_2 = \begin{bmatrix} U_2 \\ -V_2 \end{bmatrix}$. Then

$$
\begin{aligned}
&\left\|\mathcal{P}_{\text{on}}\left(W_1W_1^\top - W_2W_2^\top\right)\right\|_F^2 - \left\|\mathcal{P}_{\text{off}}\left(W_1W_1^\top - W_2W_2^\top\right)\right\|_F^2 \\
&\overset{①}{=} \left\|\mathcal{P}_{\text{on}}\left(W_1W_1^\top\right) - \mathcal{P}_{\text{on}}\left(W_2W_2^\top\right)\right\|_F^2 - \left\|\mathcal{P}_{\text{off}}(W_1W_1^\top) - \mathcal{P}_{\text{off}}\left(W_2W_2^\top\right)\right\|_F^2 \\
&\overset{②}{=} \left\|\frac{W_1W_1^\top + \widehat{W}_1\widehat{W}_1^\top}{2} - \frac{W_2W_2^\top + \widehat{W}_2\widehat{W}_2^\top}{2}\right\|_F^2 - \left\|\frac{W_1W_1^\top - \widehat{W}_1\widehat{W}_1^\top}{2} - \frac{W_2W_2^\top - \widehat{W}_2\widehat{W}_2^\top}{2}\right\|_F^2 \\
&= \left\|\frac{W_1W_1^\top - W_2W_2^\top}{2} + \frac{\widehat{W}_1\widehat{W}_1^\top - \widehat{W}_2\widehat{W}_2^\top}{2}\right\|_F^2 - \left\|\frac{W_1W_1^\top - W_2W_2^\top}{2} - \frac{\widehat{W}_1\widehat{W}_1^\top - \widehat{W}_2\widehat{W}_2^\top}{2}\right\|_F^2 \\
&\overset{③}{=} \left\langle W_1W_1^\top - W_2W_2^\top, \widehat{W}_1\widehat{W}_1^\top - \widehat{W}_2\widehat{W}_2^\top\right\rangle \\
&= \left\langle W_1W_1^\top, \widehat{W}_1\widehat{W}_1^\top\right\rangle + \left\langle W_2W_2^\top, \widehat{W}_2\widehat{W}_2^\top\right\rangle - \left\langle W_1W_1^\top, \widehat{W}_2\widehat{W}_2^\top\right\rangle - \left\langle \widehat{W}_1\widehat{W}_1^\top, W_2W_2^\top\right\rangle \\
&\overset{④}{=} -\left\langle W_1W_1^\top, \widehat{W}_2\widehat{W}_2^\top\right\rangle - \left\langle \widehat{W}_1\widehat{W}_1^\top, W_2W_2^\top\right\rangle \\
&\overset{⑤}{\leq} 0,
\end{aligned}
$$

where ① is due to the linearity of $\mathcal{P}_{\text{on}}$ and $\mathcal{P}_{\text{off}}$. ② follows from (4.12). ③ is by expanding $\|\cdot\|_F^2$. ④ comes from (4.10) that

$$\widehat{W}_i^\top W_i = W_i^\top \widehat{W}_i = \mathbf{0}, \qquad \text{for } i = 1, 2.$$

⑤ uses the fact that

$$W_1W_1^\top \succeq 0, \qquad \widehat{W}_1\widehat{W}_1^\top \succeq 0, \qquad W_2W_2^\top \succeq 0, \qquad \widehat{W}_2\widehat{W}_2^\top \succeq 0.$$

Appendix H. Proof of Proposition 4.6

From (4.5), we have

$$
\begin{aligned}
\frac{1}{2}\left(\left\|U^\star\right\|_F^2 + \left\|V^\star\right\|_F^2\right) &\overset{①}{=} \frac{1}{2}\left(\left\|P^\star\left[\sqrt{\Sigma^\star}\,\mathbf{0}_{r^\star\times(r-r^\star)}\right]R\right\|_F^2 + \left\|Q^\star\left[\sqrt{\Sigma^\star}\,\mathbf{0}_{r^\star\times(r-r^\star)}\right]R\right\|_F^2\right) \\
&\overset{②}{=} \frac{1}{2}\left(\left\|\sqrt{\Sigma^\star}\right\|_F^2 + \left\|\sqrt{\Sigma^\star}\right\|_F^2\right) \\
&= \left\|\sqrt{\Sigma^\star}\right\|_F^2 \\
&\overset{③}{=} \left\|X^\star\right\|_*,
\end{aligned}
$$

where ① uses the definitions of $U^\star$ and $V^\star$ in (4.5). ② uses the rotational invariance of $\|\cdot\|_F$. ③ is because $\left\|\sqrt{\Sigma^\star}\right\|_F^2 = \sum_j \sigma_k\left(X^\star\right) = \|X^\star\|_*$.

Therefore,

$$\begin{aligned} f\left(U^\star V^{\star\top}\right)+\lambda\left(\|U^\star\|_F^2+\|V^\star\|_F^2\right)/2 &\overset{①}{=} f\left(X^\star\right)+\lambda\left\|X^\star\right\|_* \\ &\le f(X)+\lambda\|X\|_* \\ &\overset{②}{=} f\left(UV^\top\right)+\lambda\left\|UV^\top\right\|_* \\ &\overset{③}{\le} f\left(UV^\top\right)+\lambda\left(\|U\|_F^2+\|V\|_F^2\right)/2, \end{aligned}$$

where ① comes from the optimality of $X^\star$ for $(\mathcal{P}_1)$. ② is by choosing $X = UV^\top$. ③ is because $\left\|UV^\top\right\|_* \le \left(\|U\|_F^2+\|V\|_F^2\right)/2$ by the optimization formulation of the matrix nuclear norm [43, Lemma 5.1] that

$$\|X\|_* = \min_{X=UV^\top} \frac{1}{2}\left(\|U\|_F^2+\|V\|_F^2\right).$$

□

Appendix I. Proof of Lemma 4.7

Let $Z = \begin{bmatrix} Z_U \\ Z_V \end{bmatrix}$ with arbitrary $Z_U \in \mathbb{R}^{n\times r}$ and $Z_V \in \mathbb{R}^{m\times r}$. Then

$$\begin{aligned} &\Rightarrow \langle \Xi(X)W, Z\rangle = \langle \mathbf{0}, Z\rangle = 0 \\ &\Rightarrow \left\langle \Xi(X) - \Xi(X^\star) + \Xi(X^\star), ZW^\top\right\rangle = 0 \\ &\Rightarrow \left\langle \begin{bmatrix} \lambda\mathbf{I} & \nabla f(X) \\ \nabla f(X)^\top & \lambda\mathbf{I}\end{bmatrix} - \begin{bmatrix} \lambda\mathbf{I} & \nabla f(X^\star) \\ \nabla f(X^\star)^\top & \lambda\mathbf{I}\end{bmatrix} + \Xi(X^\star), ZW^\top\right\rangle = 0 \\ &\Rightarrow \left\langle \begin{bmatrix} \mathbf{0} & \nabla f(X) - \nabla f(X^\star) \\ \nabla f(X)^\top - \nabla f(X^\star)^\top & \mathbf{0}\end{bmatrix} + \Xi(X^\star), ZW^\top\right\rangle = 0 \\ &\Rightarrow \left\langle \begin{bmatrix} \mathbf{0} & \int_0^1 [\nabla^2 f(X^\star + t(X-X^\star))](X-X^\star)\,\mathrm{d}t \\ * & \mathbf{0}\end{bmatrix} + \Xi(X^\star), ZW^\top\right\rangle = 0 \\ &\Rightarrow \left\langle \begin{bmatrix} \mathbf{0} & \int_0^1 \left[\nabla^2 f\left(X^\star + t\left(X-X^\star\right)\right)\right](X-X^\star)\,\mathrm{d}t \\ * & \mathbf{0}\end{bmatrix}, \begin{bmatrix} Z_U U^\top & Z_U V^\top \\ Z_V U^\top & Z_V V^\top\end{bmatrix}\right\rangle + \left\langle \Xi\left(X^\star\right), ZW^\top\right\rangle = 0 \\ &\Rightarrow \int_0^1 \left[\nabla^2 f\left(X^\star + t\left(X-X^\star\right)\right)\right]\left(X-X^\star, Z_U V^\top + U Z_V^\top\right)\,\mathrm{d}t + \left\langle \Xi(X^\star), ZW^\top\right\rangle = 0, \end{aligned}$$

where the fifth line follows from the Taylor's Theorem for vector-valued functions [39, Eq. (2.5) in Theorem 2.1] and for convenience $* = \left(\int_0^1 \left[\nabla^2 f\left(X^\star + t\left(X-X^\star\right)\right)\right](X-X^\star)\,\mathrm{d}t\right)^\top$ in the fifth and sixth lines. Then, from Proposition 1 and Eq. (4.12), we have

$$\left|\frac{2}{\beta+\alpha}\underbrace{\left\langle \Xi\left(X^\star\right), ZW^\top\right\rangle}_{\Pi_1(Z)} + \underbrace{\left\langle \mathcal{P}_{\mathrm{off}}\left(WW^\top - W^\star W^{\star\top}\right), ZW^\top\right\rangle}_{\Pi_2(Z)}\right| \le \frac{\beta-\alpha}{\beta+\alpha}\left\|X-X^\star\right\|_F \underbrace{\left\|\mathcal{P}_{\mathrm{off}}\left(ZW^\top\right)\right\|_F}_{\Pi_3(Z)}. \tag{I.1}$$

The Remaining Steps. The remaining steps are choosing $Z = \left(WW^\top - W^\star W^{\star\top}\right) W^{\top\dagger}$ and showing the following

$$\Pi_1(Z) \geq 0, \tag{I.2}$$

$$\Pi_2(Z) \geq \frac{1}{2}\left\|\left(WW^\top - W^\star W^{\star\top}\right) QQ^\top\right\|_F^2, \tag{I.3}$$

$$\Pi_3(Z) \leq \left\|\left(WW^\top - W^\star W^{\star\top}\right) QQ^\top\right\|_F. \tag{I.4}$$

Then plugging (I.2)–(I.4) into (I.1) yields the desired result:

$$\frac{1}{2}\left\|\left(WW^\top - W^\star W^{\star\top}\right) QQ^\top\right\|_F^2 \leq \frac{\beta-\alpha}{\beta+\alpha}\left\|X - X^\star\right\|_F \left\|\left(WW^\top - W^\star W^{\star\top}\right) QQ^\top\right\|_F,$$

or equivalently,

$$\left\|\left(WW^\top - W^\star W^{\star\top}\right) QQ^\top\right\|_F \leq 2\frac{\beta-\alpha}{\beta+\alpha}\left\|X - X^\star\right\|_F.$$

Showing (I.2). Choosing $Z = \left(WW^\top - W^\star W^{\star\top}\right) W^{\top\dagger}$ and noting that $QQ^\top = W^T W^{\top\dagger}$, we have $ZW^\top = \left(WW^\top - W^\star W^{\star\top}\right) W^{\top\dagger} W^\top = \left(WW^\top - W^\star W^{\star\top}\right) QQ^\top$. Then

$$\Pi_1(Z) = \left\langle \Xi(X^\star), \left(WW^\top - W^\star W^{\star\top}\right) QQ^\top \right\rangle = \left\langle \Xi\left(X^\star\right), WW^\top \right\rangle \geq 0,$$

where the second equality holds since $WW^\top QQ^\top = WW^\top$ and $\Xi(X^\star)W^\star = \mathbf{0}$ by (4.8). The inequality is due to $\Xi(X^\star) \succeq 0$.

Showing (I.3). First recognize that $\mathcal{P}_{\text{off}}\left(WW^\top - W^\star W^{\star\top}\right) = \frac{1}{2}\left(WW^\top - W^\star W^{\star\top} - \widehat{W}\widehat{W}^\top + \widehat{W}^\star \widehat{W}^{\star\top}\right)$. Then

$$\begin{aligned}
\Pi_2(Z) &= \left\langle \mathcal{P}_{\text{off}}\left(WW^\top - W^\star W^{\star\top}\right), ZW^\top \right\rangle \\
&= \frac{1}{2}\left\langle WW^\top - W^\star W^{\star\top}, \left(WW^\top - W^\star W^{\star\top}\right) QQ^\top \right\rangle \\
&\quad - \frac{1}{2}\left\langle \widehat{W}\widehat{W}^\top - \widehat{W}^\star \widehat{W}^{\star\top}, \left(WW^\top - W^\star W^{\star\top}\right) QQ^\top \right\rangle.
\end{aligned}$$

Therefore, (I.3) follows from

$$\left\langle \widehat{W}\widehat{W}^\top - \widehat{W}^\star \widehat{W}^{\star\top}, \left(WW^\top - W^\star W^{\star\top}\right) QQ^\top \right\rangle = \left\langle \widehat{W}\widehat{W}^\top, -W^\star W^{\star\top} \right\rangle + \left\langle -\widehat{W}^\star \widehat{W}^{\star\top}, WW^\top \right\rangle \leq 0,$$

where the first equality uses (4.10) and the inequality is because

$$\widehat{W}\widehat{W}^\top \succeq 0, \qquad W^\star W^{\star\top} \succeq 0, \qquad \widehat{W}^\star \widehat{W}^{\star\top} \succeq 0, \qquad WW^\top \succeq 0.$$

Showing (I.4). Plugging $Z = \left(WW^\top - W^\star W^{\star\top}\right) W^{\top\dagger}$ gives

$$\Pi_3(Z) = \left\|\mathcal{P}_{\text{off}}\left(\left(WW^\top - W^\star W^{\star\top}\right) QQ^\top\right)\right\|_F,$$

which is obviously no larger than $\left\|\left(WW^\top - W^\star W^{\star\top}\right) QQ^\top\right\|_F$ by the definition of the operation $\mathcal{P}_{\text{off}}$.