The Nonconvex Geometry of the Non-square Low Rank Matrix Optimization

Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B. Wakin *
Department of Electrical Engineering and Computer Science
Colorado School of Mines

November 15, 2016

Abstract

This paper considers the minimization of a convex objective function $f(\boldsymbol{X})$ over non-square matrices with a low-rank optimal solution $\boldsymbol{X}^\star$. We focus on the nonconvex reformulation where $\boldsymbol{X}$ is factored into the product of two rectangular matrices with much smaller size than the original one. In spite of the nonconvexity, recent studies in matrix sensing and completion problems have shown that the corresponding factored problems have no spurious local minima or degenerate saddle points and that many local search algorithms (such as gradient descent) can efficiently find the global solution. We extend this line of geometry-based convergence analysis by considering a general objective function $f(\boldsymbol{X})$ satisfying certain restricted strict convexity and smoothness conditions. In particular, we show that each critical point of the reformulated objective function either corresponds to the global minimum point $\boldsymbol{X}^\star$ of the original convex program, or is a strict saddle point such that the Hessian has a strictly negative eigenvalue. This property ensures that many local search algorithms for the factored problem can converge to the global optimal solution $\boldsymbol{X}^\star$.

1 Introduction

Consider the minimization of a convex objective function $f(\boldsymbol{X})$ over all $n \times m$ matrices:

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

which admits a low-rank solution $\boldsymbol{X}^\star \in \mathbb{R}^{n\times m}$ with $\text{rank}(\boldsymbol{X}^\star) = r^\star$. This low-rank matrix optimization appears in a broad variety of research fields, including quantum computing [1, 26], collaborative filtering [25, 45, 48, 55], sensor localization [7, 8], low rank matrix recovery from compressive measurements [19, 44], and matrix completion [20, 43].

In order to find a low-rank solution, the nuclear norm is widely exploited in matrix-based inverse problems [19, 20, 43, 44], which have found a variety of applications in signal processing [24], machine learning [29], and control [38]. The performance of nuclear norm based minimization in recovering a low-rank matrix has been extensively studied using convex analysis techniques [20]. For example, it has information-theoretically optimal sampling complexity [21], has a general sharp oracle inequality [19] and achieves nearly minimax optimal rates [18]. In spite of its optimal performance, the computational complexities of nuclear norm minimization is very high, prohibiting it from scaling to practical problems even with specialized first-order algorithms. For example, the singular value thresholding algorithm [13] requires performing an expensive singular value decomposition in each iteration, which is computationally prohibitive for large-scale settings.

For relieving the computational bottleneck and inspired by the pioneering work in [11, 12] for solving a semi-definite program (SDP), recent studies propose to factorize the variable into Burer-Monteiro

*Email: {zzhu,qiuli,gtang,mwakin}@mines.edu. This work was supported by NSF grants CCF-1409261, CCF-1464205.

style decomposition and then turn to solve a factored nonconvex problem instead of the convex one [3, 5, 6, 11, 12, 28, 36, 42, 52, 54, 56]. Specifically, we factorize the variable $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ into a low-rank decomposition, where $\boldsymbol{U} \in \mathbb{R}^{n\times r}, \boldsymbol{V} \in \mathbb{R}^{m\times r}$ with $r \geq r^\star$ (but typically still $r \ll \min\{m, n\}$). By matrix factorization, the low-rankness is automatically incorporated into the new introduced variables $\boldsymbol{U}$ and $\boldsymbol{V}$. Using this matrix factorization, we can recast the original problem (1) into a new Burer-Monteiro-based formulation

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r},\boldsymbol{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \; h(\boldsymbol{U},\boldsymbol{V}) := f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}). \tag{2}$$

The bilinear nature in (2) renders its objective function nonconvex, exhibiting potentially spurious local minima or bad saddle points. With technical innovations in analyzing the nonconvex landscape of the factored function, several recent works have shown that with exact factorization $r = r^\star$, the factored objective function $h(\boldsymbol{U}, \boldsymbol{V})$ in matrix inverse problems has no spurious local minima or degenerate saddle[1] points [6, 28, 42].

We generalize this line of work by focusing on a general objective function $f(\boldsymbol{X})$ in the optimization (1), not necessarily a quadratic loss function coming from a matrix inverse problem. As also illustrated in Burer and Monterio's work [11, 12], the factored problem (2) serves as a way to solve the convex optimization (1) globally, rather than a new modeling method. From this perspective, there is no need to redevelop the statistical performances for the factored optimization (2), whose performance is in fact inherited from that of the convex optimization (1). Exploiting the powerful and elegant language of convex analysis which has been unified for several decades of research, the performance of the convex counterpart (1) can be well developed. For example, the optimal sampling complexity for matrix inverse problems needs not to be rederived once one knows the equivalence between the convex and the factored formulations. In addition, our general analysis technique also sheds light on the connection between the geometries of the convex program (1) and its nonconvex counterpart (2).

1.1 Summary of Results

The purpose of this paper is through matrix factorization approach to understand how the geometric structures of the convex objective function $f(\boldsymbol{X})$ is transformed into its nonconvex counterpart $h(\boldsymbol{U}, \boldsymbol{V})$, how the global minimum is mapped into the factored space, and whether any other type of critical points are introduced (like degenerate saddle or strict saddle). To answer these questions, we focus our major attentions on characterizing the type of each critical point of the nonconvex function $h(\boldsymbol{U}, \boldsymbol{V})$.

Before presenting our main results, we lay out the necessary assumption on the objective function $f(\boldsymbol{X})$. As is known, without any assumptions on the problem, even minimizing traditional quadratic objective functions is challenging [28, 31, 52, 54]. For this purpose, we focus on the model where $f(\boldsymbol{X})$ is $2r$-restricted strongly convex and smooth, i.e., for any $n \times m$ matrix $\boldsymbol{X}$ with $\mathrm{rank}(\boldsymbol{X}) \leq 2r$, the Hessian of $f(\boldsymbol{X})$ satisfies

$$\alpha\mathbf{I} \preceq \nabla^2 f(\boldsymbol{X}) \preceq \beta\mathbf{I} \tag{3}$$

for some positive numbers α and β. Here, $\mathbf{I}$ is an identity matrix of appropriate size and $\boldsymbol{X} \preceq \boldsymbol{Y}$ means $\boldsymbol{X} - \boldsymbol{Y} \preceq \mathbf{0}$, i.e., $\boldsymbol{X} - \boldsymbol{Y}$ is positive semi-definite (PSD). This assumption is standard in matrix inverse problem [19]. With this assumption on $f(\boldsymbol{X})$, we summarize our main results in the following informal theorem:

Theorem 1. (informal) *Suppose the function $f(\boldsymbol{X})$ satisfies $2r$-restricted strong convexity and smoothness condition (3) and admits a low-rank minimizer $\boldsymbol{X}^\star \in \mathbb{R}^{n\times m}$ with $\mathrm{rank}(\boldsymbol{X}^\star) = r^\star \leq r$. Then each critical point of the factored objective function $h(\boldsymbol{U}, \boldsymbol{V})$ (with an additional regularizer, see Theorem 2) in (2) either corresponds to the low-rank global solution of the original convex program (1), or is a strict saddle point such that the Hessian at this point has a strictly negative eigenvalue.*

Remark 1. The strict saddle property implies that we can recover the rank-$r^\star$ global minimizer $\boldsymbol{X}^\star$ of (1) by many iterative algorithms, for example, trust region method [49, 51], stochastic gradient descent [27], and gradient descent with sufficient small stepsize [32], even with random initialization.

[1] Degenerate saddle point is usually referred to the critical point that is neither a local minimum nor a local maximum and the Hessian matrix evaluated at this point is singular. At this point, one needs higher order information instead of its Hessian information to find a direction to reduce the function value.

Remark 2. Our main result only relies on the strong convexity and smoothness property (3). Therefore, apart from nonconvex phase synchronization [9] and low-rank matrix recovery problem [19], it can also be applied to many other low-rank matrix optimization problems which do not necessarily involve quadratic loss functions, including 1-bit matrix completion [14, 23] with negative log-likelihood function, Possion principal component analysis (PCA) [46], robust PCA [10,16], and other low-rank models with generazied loss functions [53].

1.2 Related Works

In recent years, a surge of convex optimizations are reformulated into the corresponding nonconvex form to get rid of the computational bottleneck. For example, one can naturally reformulate several convex optimizations in machine learning, signal processing, and statistical problems into nonconvex forms [5, 6, 15, 30, 35, 40, 51, 52]. Compared with the convex optimization, the nonconvex form typically involves many fewer variables (or variables with much smaller size) and thus can be efficiently solved by simple but powerful methods (such as gradient descent [27, 32], trust region method [49], alternating methods [30]) for large-scale settings.

Although it is known that these nonconvex formulations work surprisingly well in practice, more work is still needed to fully understand the underlying theoretical foundation for this phenomenon, especially the geometric structures of the nonconvex optimizations. Unlike the objective functions of convex optimizations that have simple landscapes such that local minimizers are always global ones, the objective functions of general nonconvex functions have much more complicated landscapes. For example, it is an NP-hard problem to even certify the local optimality of a point for general function [39]. It is also very common that a general nonconvex function has spurious local minima that are not global optima [47].

For many convex optimizations—typically those involving structured matrices—the landscapes of the corresponding nonconvex reformulations fortunately have nice geometric structures that allow simple local search algorithms to find the globally optimal solution [50]. Typical examples are dictionary learning [2, 4, 49], matrix sensing and completion [28, 30, 31, 52, 54], phase retrieval [17, 22, 51] and blind deconvolution [33, 34, 37]. Based on whether a good initialization is utilized, these algorithms can be separated into two categories. One set of algorithms consist of two steps: initialization and local refinement. A good guess lying in the attraction basin of the global optimum can lead to global convergence of the following iterative step. We can obtain such initializations by spectral methods for phase retrieval [15] and low-rank matrix recovery problems [5, 52, 56]. Another category of works attempt to analyze the landscape of the objective function and show that they have no spurious local minima and no degenerate saddle points. If this particular property holds, then simple algorithms such as gradient descent and trust region method are guaranteed to converge to global optimality with random initialization [27,32,49]. Our work follows this approach of geometry-based convergence analysis for low-rank matrix optimization with generic objective functions.

Our work is most closely related to the recent works in low-rank matrix optimization [6, 28, 36]. Bhojanapalli et al. [6] showed that the low-rank, PSD matrix recovery problem from linear observations has no spurious local minimum or degenerate saddle point. A similar argument for low-rank, PSD matrix completion was obtained by Ge et al. [28]. In [36], the authors expand this line by considering low-rank, PSD matrix optimization problems with generic objective functions. Our work extends this line of analysis to the general low-rank matrix (not necessary PSD or even square) optimization problems. Another closely related work is low-rank, non-square matrix estimation from linear observations by minimizing the factored quadratic objective function in [42]. We note that the low-rank sensing problem is a special case of our general objective function framework. Furthermore, our result covers both over-parameterization where $r > r^\star$ and exact parameterization where $r = r^\star$. In addition, Wang et al. [54] also considered the factored low-rank matrix minimization problem with a general objective function which satisfies restricted the strong convexity and smoothness conditions. Their algorithms require good initializations for global convergence since they only characterized the local landscapes around the global optima. By categorizing all critical points into global optima and strict saddle points, our work differs from [54] in that we instead characterize the global landscapes of the factored objective function.

Before proceeding, we first briefly introduce some notations used throughout the paper. The symbols $\mathbf{I}$ and $\mathbf{0}$ respectively represent the identity matrix and zero matrix with appropriate sizes. The set of $r \times r$ orthonormal matrices is denoted by $\mathcal{O}_r := \{\boldsymbol{R} \in \mathbb{R}^{r\times r} : \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \mathbf{I}\}$. If a function $h(\boldsymbol{U}, \boldsymbol{V})$ has two

arguments, $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{m\times r}$, we occasionally use the notation $h(\boldsymbol{W})$ when we put these two arguments into a new one as $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$. For a scalar function $f(\boldsymbol{Z})$ with a matrix variable $\boldsymbol{Z} \in \mathbb{R}^{n\times m}$, its gradient is an $n \times m$ matrix whose (i,j)-th entry is $[\nabla f(\boldsymbol{Z})]_{ij} = \frac{\partial f(\boldsymbol{Z})}{\partial \boldsymbol{Z}_{ij}}$ for all $i \in \{1,2,\ldots,n\}, j \in \{1,2,\ldots,m\}$. The Hessian of $f(\boldsymbol{Z})$ can be viewed as an $nm \times nm$ matrix $[\nabla^2 f(\boldsymbol{Z})]_{ij} = \frac{\partial^2 f(\boldsymbol{Z})}{\partial \boldsymbol{z}_i \partial \boldsymbol{z}_j}$ for all $i,j \in \{1,\ldots,nm\}$, where $\boldsymbol{z}_i$ is the i-th entry of the vectorization of $\boldsymbol{Z}$. An alternative way to represent the Hessian is by a bilinear form defined via $[\nabla^2 f(\boldsymbol{Z})](\boldsymbol{A},\boldsymbol{B}) = \sum_{i,j,k,l} \frac{\partial^2 f(\boldsymbol{Z})}{\partial \boldsymbol{Z}_{ij}\partial \boldsymbol{Z}_{kl}} \boldsymbol{A}_{ij}\boldsymbol{B}_{kl}$ for any $\boldsymbol{A},\boldsymbol{B} \in \mathbb{R}^{n\times m}$. These two notations will be used interchangeably whenever the specific form can be inferred from context.

2 Problem Formulation and Main Results

This paper considers the problem (1) of minimizing a convex function $f(\boldsymbol{X})$ which admits a low-rank solution $\boldsymbol{X}^\star$ with $\operatorname{rank}(\boldsymbol{X}^\star) = r^\star \le r$. We factorize the variable $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ with $\boldsymbol{U} \in \mathbb{R}^{n\times r}, \boldsymbol{V} \in \mathbb{R}^{m\times r}$ and transform the convex program into its factored counterpart (2) whose objective function is

$$h(\boldsymbol{W}) = h(\boldsymbol{U},\boldsymbol{V}) := f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}),$$

where $\boldsymbol{W} := \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$. Although the new variable $\boldsymbol{W}$ has much smaller size than $\boldsymbol{X}$ when $r \ll \min\{n,m\}$, the factored problem (2) is no longer convex due to the bilinear form about $\boldsymbol{U}$ and $\boldsymbol{V}$. The nonconvexity in the reformulated objective function $h(\boldsymbol{U},\boldsymbol{V})$ may introduce spurious local minima or degenerate saddle points. Our goal is to provide a positive answer to this issue by showing that the critical points either corresponds to $\boldsymbol{X}^\star$ or are strict saddle points where Hessian has a strictly negative eigenvalue.

Let $\boldsymbol{X}^\star = \boldsymbol{Q}_{\boldsymbol{U}^\star}\boldsymbol{\Sigma}^\star\boldsymbol{Q}_{\boldsymbol{V}^\star}^{\mathrm{T}}$ denote an SVD of $\boldsymbol{X}^\star$, where $\boldsymbol{Q}_{\boldsymbol{U}^\star} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{Q}_{\boldsymbol{V}^\star} \in \mathbb{R}^{m\times r}$ are orthonormal matrices of appropriate sizes, and $\boldsymbol{\Sigma}^\star \in \mathbb{R}^{r\times r}$ is a diagonal matrix with non-negative diagonals. Also let $\boldsymbol{U}^\star = \boldsymbol{Q}_{\boldsymbol{U}^\star}\boldsymbol{\Sigma}^{\star 1/2}$ and $\boldsymbol{V}^\star = \boldsymbol{Q}_{\boldsymbol{V}^\star}\boldsymbol{\Sigma}^{\star 1/2}$, where $\boldsymbol{X}^\star = \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}}$ forms a balanced factorization of $\boldsymbol{X}^\star$ since $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$ have the same set of non-zero singular values. Through the paper, we utilize the following two ways to stack $\boldsymbol{U}$ and $\boldsymbol{V}$ together:

$$\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}, \quad \widehat{\boldsymbol{W}} = \begin{bmatrix}\boldsymbol{U}\\ -\boldsymbol{V}\end{bmatrix}, \boldsymbol{W}^\star = \begin{bmatrix}\boldsymbol{U}^\star\\ \boldsymbol{V}^\star\end{bmatrix}, \quad \widehat{\boldsymbol{W}}^\star = \begin{bmatrix}\boldsymbol{U}^\star\\ -\boldsymbol{V}^\star\end{bmatrix}.$$

One useful consequence of this notation is as follows

$$\widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}} = \boldsymbol{0}. \tag{4}$$

Before moving on, we note that for any solution $(\boldsymbol{U},\boldsymbol{V})$ to (2), $(\boldsymbol{U}\boldsymbol{\Psi},\boldsymbol{V}\boldsymbol{\Phi})$ is also a solution to (2) for any $\boldsymbol{\Psi},\boldsymbol{\Phi} \in \mathbb{R}^{r\times r}$ such that $\boldsymbol{U}\boldsymbol{\Psi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$. In order to address this ambiguity (i.e., to reduce the search space of $\boldsymbol{W}$ for (2)), we utilize the trick in [42, 52, 54] by introducing a regularizer g and solving the following problem

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r},\boldsymbol{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \ \rho(\boldsymbol{W}) = \rho(\boldsymbol{U},\boldsymbol{V}) := f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}) + g(\boldsymbol{U},\boldsymbol{V}) \tag{5}$$

where

$$g(\boldsymbol{U},\boldsymbol{V}) = \frac{\mu}{4}\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right\|_F^2.$$

We remark that $\boldsymbol{W}^\star$ is still a global minimizer to the factored problem (5) since both $f(\boldsymbol{X})$ and $g(\boldsymbol{W})$ achieve their global minimum at $\boldsymbol{W}^\star$. The regularizer $g(\boldsymbol{W})$ is applied to force the difference between the two Gram matrices of $\boldsymbol{U}$ and $\boldsymbol{V}$ as small as possible. The global minimum of $g(\boldsymbol{W})$ is 0, which is achieved when $\boldsymbol{U}$ and $\boldsymbol{V}$ have the same set of singular values and right singular vectors, i.e., $\boldsymbol{W}$ belongs to

$$\mathcal{E} := \left\{\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} : \boldsymbol{U} = \boldsymbol{Q}_{\boldsymbol{U}}\boldsymbol{\Sigma}\boldsymbol{R}^{\mathrm{T}}, \quad \boldsymbol{V} = \boldsymbol{Q}_{\boldsymbol{V}}\boldsymbol{\Sigma}\boldsymbol{R}^{\mathrm{T}}, \quad \boldsymbol{Q}_{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{Q}_{\boldsymbol{U}} = \boldsymbol{Q}_{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{Q}_{\boldsymbol{V}} = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \mathbf{I}\right\},$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with non-negative diagonals. Informally, we can view (5) as finding a point from $\mathcal{E}$ that also minimizes $f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})$.

Our main argument is that any critical point $\boldsymbol{W}$ of ρ satisfying $\nabla\rho(\boldsymbol{W}) = \boldsymbol{0}$ is either the global solution of the original convex problem (1) or a strict saddle point of the factored problem (5) such that the Hessian matrix $\nabla^2\rho(\boldsymbol{W})$ evaluated at this point has a strictly negative eigenvalue. We formally establish this in the following theorem, whose proof is given in the next section.

Theorem 2. *Suppose the function $f(\boldsymbol{X})$ satisfies the strong convexity and smoothness condition (3) with positive α and β satisfying $\frac{\beta}{\alpha} \leq 1.12$. Set $\mu = \frac{\alpha+\beta}{4}$ for the factored problem (5). Then any critical point $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$ of ρ, i.e., any $\boldsymbol{W}$ such that $\nabla\rho(\boldsymbol{W}) = \boldsymbol{0}$, is either the global solution of the original convex problem (1), i.e.,*

$$\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{X}^\star,$$

or a strict saddle point of the factored problem (5) with

$$\lambda_{\min}\left(\nabla^2\left(\rho(\boldsymbol{W})\right)\right) \leq -0.069\alpha\left(\sigma_{r'}^2(\boldsymbol{W}) + 2\sigma_{r'}(\boldsymbol{X}^\star) + \sqrt{2}\sigma_{r'}(\boldsymbol{W})\sqrt{\sigma_{r'}(\boldsymbol{X}^\star)}\right), \tag{6}$$

where $r' = \max\{r^\star, r^c\}$, $r^c \leq r$ is the rank of $\boldsymbol{W}$, $\lambda_{\min}(\cdot)$ represents the smallest eigenvalue and $\sigma_\ell(\cdot)$ denotes the ℓ-th largest singular value.

Remark 3. The right hand side of (6) can be simplified by using the relative values of r^c and $r^\star$:

$$\sigma_{r'}^2(\boldsymbol{W}) + 2\sigma_{r'}(\boldsymbol{X}^\star) + \sqrt{2}\sigma_{r'}(\boldsymbol{W})\sqrt{\sigma_{r'}(\boldsymbol{X}^\star)} = \begin{cases} \sigma_{r^c}^2(\boldsymbol{W}), & r^c > r^\star; \\ \sigma_{r^\star}^2(\boldsymbol{W}) + 2\sigma_{r^\star}(\boldsymbol{X}^\star) + \sqrt{2}\sigma_{r^\star}(\boldsymbol{W})\sqrt{\sigma_{r^\star}(\boldsymbol{X}^\star)}, & r^c = r^\star; \\ 2\sigma_{r^\star}(\boldsymbol{X}^\star), & r^c < r^\star. \end{cases}$$

For all these cases, the right hand side of (6) is strictly negative, implying $\boldsymbol{W}$ is a strict saddle point. We finally note that Theorem 2 not only covers exact parameterization where $r = r^\star$, but also includes over-parameterization where $r > r^\star$.

3 Proof of Theorem 2

In this section, we provide a formal proof of Theorem 2. The main argument involves showing that each critical point of $\rho(\boldsymbol{W})$ either corresponds to the global solution of (1) or is a strict saddle point whose Hessian $\nabla^2\rho(\boldsymbol{W})$ has a strictly negative eigenvalue. Following the same proof techniques of [6, 36, 42], we show that $\boldsymbol{W}$ is a strict saddle point by arguing that the Hessian $\nabla^2\rho(\boldsymbol{W})$ has a strictly negative curvature along $\boldsymbol{\Delta} := \boldsymbol{W} - \boldsymbol{W}^\star\boldsymbol{R}$, i.e., $[\nabla^2\rho(\boldsymbol{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) \leq -\tau\|\boldsymbol{\Delta}\|_F^2$ for some $\tau > 0$. Here $\boldsymbol{R}$ is an $r \times r$ orthonormal matrix such that the distance between $\boldsymbol{W}$ and $\boldsymbol{W}^\star$ rotated through $\boldsymbol{R}$ as small as possible.

3.1 Supporting Results

We first present some useful results. As a consequence of the strong convexity and smoothness condition (3), the following result establishes that if (1) has an optimal solution $\boldsymbol{X}^\star$ with $\mathrm{rank}(\boldsymbol{X}^\star) \leq r$, then this is the unique global minimum point of rank at most r.

Proposition 1. *Suppose the function $f(\boldsymbol{X})$ satisfies the $2r$-restricted strong convexity and smoothness condition (3) with positive α and β. Assume $\boldsymbol{X}^\star$ is a global minimum point of $f(\boldsymbol{X})$ with $\mathrm{rank}(\boldsymbol{X}^\star) = r^\star \leq r$. Then there is no other optimum of (1) that has rank less than or equal to r.*

Proof of Proposition 1. Suppose there exists another global minimum point $\boldsymbol{X}' \neq \boldsymbol{X}^\star$ with $\mathrm{rank}(\boldsymbol{X}') \leq r$. The convexity of $f(\boldsymbol{X})$ indicates that $f(\boldsymbol{X}^\star) = f(\boldsymbol{X}')$, i.e.,

$$f(\boldsymbol{X}^\star) - f(\boldsymbol{X}') = 0.$$

On the other hand, the second order Taylor expansion gives

$$f(\boldsymbol{X}') = f(\boldsymbol{X}^\star) + \langle\nabla f(\boldsymbol{X}^\star), \boldsymbol{X}' - \boldsymbol{X}^\star\rangle + \frac{1}{2}[\nabla^2 f(\widetilde{\boldsymbol{X}})](\boldsymbol{X}' - \boldsymbol{X}^\star, \boldsymbol{X}' - \boldsymbol{X}^\star),$$

where $\widetilde{\boldsymbol{X}} = t\boldsymbol{X}^\star + (1-t)\boldsymbol{X}'$ for some $t \in [0,1]$ and $[\nabla^2 f(\widetilde{\boldsymbol{X}})](\boldsymbol{X}' - \boldsymbol{X}^\star, \boldsymbol{X}' - \boldsymbol{X}^\star)$ evaluates the Hessian bilinear form at the direction $\boldsymbol{X}' - \boldsymbol{X}^\star$. Combining the above two equations together with $f(\boldsymbol{X}^\star) = \boldsymbol{0}$ gives

$$[\nabla^2 f(\widetilde{\boldsymbol{X}})](\boldsymbol{X}' - \boldsymbol{X}^\star, \boldsymbol{X}' - \boldsymbol{X}^\star) = 0,$$

which contradicts (3) since $\boldsymbol{X}' - \boldsymbol{X}^\star \neq \boldsymbol{0}$ and $\operatorname{rank}(\widetilde{\boldsymbol{X}}) \leq \operatorname{rank}(\boldsymbol{X}') + \operatorname{rank}(\boldsymbol{X}^\star) \leq 2r$. □

The restricted strong convexity and smoothness assumption (3) also implies the following isometry property.

Proposition 2. *Suppose the function $f(\boldsymbol{X})$ satisfies the $2r$-restricted strong convexity and smoothness condition (3) with positive α and β. Then for any $n \times m$ matrix $\boldsymbol{Z}$ of rank at most $2r$, we have*

$$\left| \left[\frac{2}{\alpha+\beta} \nabla^2 f(\boldsymbol{Z}) \right] (\boldsymbol{G}, \boldsymbol{H}) - \langle \boldsymbol{G}, \boldsymbol{H} \rangle \right| \leq \frac{\beta - \alpha}{\beta + \alpha} \|\boldsymbol{G}\|_F \|\boldsymbol{H}\|_F$$

for any $\boldsymbol{G}, \boldsymbol{H} \in \mathbb{R}^{n \times m}$.

Proof of Proposition 2. First note that the $2r$-restricted strong convexity and smoothness condition (3) implies

$$\left\| \frac{2}{\alpha+\beta} \nabla^2 f(\boldsymbol{Z}) - \mathbf{I} \right\| \leq \frac{\beta - \alpha}{\beta + \alpha},$$

where $\nabla^2 f(\boldsymbol{Z})$ is viewed as an $nm \times nm$ matrix. Thus, we obtain

$$\left| \left[\frac{2}{\alpha+\beta} \nabla^2 f(\boldsymbol{Z}) \right] (\boldsymbol{G}, \boldsymbol{H}) - \langle \boldsymbol{G}, \boldsymbol{H} \rangle \right| = \left| \operatorname{vec}(\boldsymbol{G})^{\mathrm{T}} \left(\frac{2}{\alpha+\beta} \nabla^2 f(\boldsymbol{Z}) - \mathbf{I} \right) \operatorname{vec}(\boldsymbol{H}) \right| \leq \frac{\beta - \alpha}{\beta + \alpha} \|\boldsymbol{G}\|_F \|\boldsymbol{H}\|_F .$$

□

The following result provides an upper bound on the energy of the difference $\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star \boldsymbol{W}^{\star \mathrm{T}}$ after projection onto the column space of $\boldsymbol{W}$. Its proof is given in Appendix A.

Lemma 1. *Suppose $f(\boldsymbol{X})$ satisfies $2r$-restricted strong convexity and smoothness condition (3). Let $\mu = \frac{\alpha+\beta}{4}$ in the factored problem (5). For any critical point $\boldsymbol{W}$ of (5), let $\boldsymbol{P}_{\boldsymbol{W}} \in \mathbb{R}^{(m+n)\times(m+n)}$ be the orthogonal projector onto the column space of $\boldsymbol{W}$. Then*

$$\left\| \left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star \boldsymbol{W}^{\star \mathrm{T}} \right) \boldsymbol{P}_{\boldsymbol{W}} \right\|_F \leq 2 \frac{\beta - \alpha}{\beta + \alpha} \left\| \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star \boldsymbol{W}^{\star \mathrm{T}} \right\|_F .$$

We remark that Lemma 1 is a variant of [42, Lemma 3.2]. While the result there requires a $4r$-RIP condition of the objective function, our result depends only on the $2r$-restricted strong convexity and smoothness condition.

In addition, for any matrix $\boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{n \times r}$, the following result relates the distance between $\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}$ and $\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}$ to the distance between $\boldsymbol{C}$ and $\boldsymbol{D}$.

Lemma 2. *[36, Lemma 2] For any matrix $\boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{n \times r}$ with rank r_1 and r_2, respectively, let $\boldsymbol{R} = \arg\min_{\widetilde{\boldsymbol{R}} \in \mathcal{O}_r} \|\boldsymbol{C} - \boldsymbol{D}\boldsymbol{R}\|_F$. Then*

$$\left\| \boldsymbol{C}\boldsymbol{C}^{\mathrm{T}} - \boldsymbol{D}\boldsymbol{D}^{\mathrm{T}} \right\|_F \geq \left(\sigma_{\max\{r_1, r_2\}}(\boldsymbol{C}) + \sigma_{\max\{r_2, r_2\}}(\boldsymbol{D}) \right) \|\boldsymbol{C} - \boldsymbol{D}\boldsymbol{R}\|_F .$$

We present one more useful result in the following lemma.

Lemma 3. *[36, Lemma 3] For any matrix $\boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{n \times r}$, let $\boldsymbol{P}_{\boldsymbol{C}}$ be the orthogonal projector onto the range of $\boldsymbol{C}$. Let $\boldsymbol{R} = \arg\min_{\widetilde{\boldsymbol{R}} \in \mathcal{O}_r} \|\boldsymbol{C} - \boldsymbol{D}\boldsymbol{R}\|_F$. Then*

$$\left\| \boldsymbol{C} \left(\boldsymbol{C} - \boldsymbol{D}\boldsymbol{R} \right)^{\mathrm{T}} \right\|_F^2 \leq \frac{1}{8} \left\| \boldsymbol{C}\boldsymbol{C}^{\mathrm{T}} - \boldsymbol{D}\boldsymbol{D}^{\mathrm{T}} \right\|_F^2 + \left(3 + \frac{1}{2(\sqrt{2}-1)} \right) \left\| \left(\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}} - \boldsymbol{D}\boldsymbol{D}^{\mathrm{T}} \right) \boldsymbol{P}_{\boldsymbol{C}} \right\|_F^2 .$$

Finally, we provide the gradient and Hessian expression for ρ. The gradient of $\rho(\boldsymbol{W})$ is given by

$$\begin{aligned} \nabla_{\boldsymbol{U}} \rho(\boldsymbol{U}, \boldsymbol{V}) &= \nabla f(\boldsymbol{X}) \boldsymbol{V} + \mu \boldsymbol{U} (\boldsymbol{U}^{\mathrm{T}} \boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}} \boldsymbol{V}), \\ \nabla_{\boldsymbol{V}} \rho(\boldsymbol{U}, \boldsymbol{V}) &= \nabla f(\boldsymbol{X})^{\mathrm{T}} \boldsymbol{U} - \mu \boldsymbol{V} (\boldsymbol{U}^{\mathrm{T}} \boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}} \boldsymbol{V}), \end{aligned}$$

equivalently,

$$\nabla\rho(\boldsymbol{W}) = \begin{bmatrix} \nabla f(\boldsymbol{X})\boldsymbol{V} \\ \nabla f(\boldsymbol{X})^{\mathrm{T}}\boldsymbol{U} \end{bmatrix} + \mu\widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W}.$$

Standard computations gives the the Hessian quadrature form $[\nabla^2\rho(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta})$ for any $\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{U}} \\ \boldsymbol{\Delta}_{\boldsymbol{V}} \end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$ as

$$\begin{aligned}[\nabla^2\rho(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta}) = {} & [\nabla^2 f(\boldsymbol{X})](\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}, \boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}) \\ & + 2\left\langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}} \right\rangle + [\nabla^2 g(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta}),\end{aligned}$$

where

$$[\nabla^2 g(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta}) = \mu\left(\left\langle \widehat{\boldsymbol{\Delta}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{W}^{\mathrm{T}} \right\rangle + \left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Delta}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{W}^{\mathrm{T}} \right\rangle + \left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}} \right\rangle\right).$$

3.2 The Formal Proof

Proof of Theorem 2. The convexity of $f(\boldsymbol{X})$ implies that $\boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}} = \boldsymbol{X}^\star$ is the globally optimal solution of the original convex program (1) if and only if

$$\nabla f(\boldsymbol{X}^\star) = \boldsymbol{0}. \tag{7}$$

Also guaranteed by the uniqueness of Proposition 1, $\boldsymbol{X}^\star$ is the unique $n \times m$ matrix with rank at most r such that the gradient of $f(\boldsymbol{X})$ is zero at this point. Hence it is then sufficient to show the following set of critical points of $\rho(\boldsymbol{W})$:

$$\{\boldsymbol{W} \in \mathbb{R}^{(n+m)\times r} : \nabla\rho(\boldsymbol{W}) = \boldsymbol{0}\} \cap \left\{\boldsymbol{W} \in \mathbb{R}^{(n+m)\times r} : \nabla f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}) \neq \boldsymbol{0}\right\}$$

is the set of strict saddle points, i.e., there is a direction $\boldsymbol{\Delta}$ along which the Hessian has a strictly negative curvature for these points.

Following the same approach as in [6, 36, 42], we construct $\boldsymbol{\Delta} = \boldsymbol{W} - \boldsymbol{W}^\star\boldsymbol{R}$, the difference from $\boldsymbol{W}$ to its nearest global factor $\boldsymbol{W}^\star$, where

$$\boldsymbol{R} = \underset{\widetilde{\boldsymbol{R}}\in\mathcal{O}_r}{\arg\min} \left\|\boldsymbol{W} - \boldsymbol{W}^\star\widetilde{\boldsymbol{R}}\right\|_F.$$

Such $\boldsymbol{\Delta}$ satisfies $\boldsymbol{\Delta} \neq \boldsymbol{0}$ since $\nabla f(\boldsymbol{X}) \neq \boldsymbol{0}$ implying $\boldsymbol{X} \neq \boldsymbol{X}^\star$ by the convexity of $f(\boldsymbol{X})$. Then we have

$$\begin{aligned}[\nabla^2\rho(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta}) = {} & \underbrace{[\nabla^2 f(\boldsymbol{X})](\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}, \boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}})}_{\Pi_1} \\ & + 2\underbrace{\left\langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}} \right\rangle}_{\Pi_2} \\ & + \mu\left(\underbrace{\left\langle \widehat{\boldsymbol{\Delta}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{W}^{\mathrm{T}} \right\rangle}_{\Pi_3} + \underbrace{\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Delta}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{W}^{\mathrm{T}} \right\rangle}_{\Pi_4} + \underbrace{\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}} \right\rangle}_{\Pi_5}\right).\end{aligned}$$

The remaining part is to show the term Π_2 plus Π_5 is strictly negative, while the sum of the remaining terms are relatively small, though they maybe nonnegative.

Bounding terms Π_1, Π_3 and Π_4

By the smoothness condition (3), we have

$$\begin{aligned}\Pi_1 &= [\nabla^2 f(\boldsymbol{X})](\boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}}, \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}}) \\ &\le \beta\|\boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}}\|_F^2 \\ &\le 2\beta\left(\|\boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}}\|_F^2\right).\end{aligned}$$

Also the terms Π_3 and Π_4 can be rewritten as

$$\begin{aligned}\Pi_3 &= \left\langle \widehat{\boldsymbol{\Delta}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{W}^{\mathrm{T}}\right\rangle \\ &= \|\boldsymbol{\Delta_U}\boldsymbol{U}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{\Delta_V}\boldsymbol{V}^{\mathrm{T}}\|_F^2 - \|\boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}}\|_F^2 - \|\boldsymbol{\Delta_V}\boldsymbol{U}^{\mathrm{T}}\|_F^2,\end{aligned}$$

$$\begin{aligned}\Pi_4 &= \left\langle \boldsymbol{U}\boldsymbol{\Delta_U}^{\mathrm{T}}, \boldsymbol{\Delta_U}\boldsymbol{U}^{\mathrm{T}}\right\rangle + \left\langle \boldsymbol{V}\boldsymbol{\Delta_V}^{\mathrm{T}}, \boldsymbol{\Delta_V}\boldsymbol{V}^{\mathrm{T}}\right\rangle - 2\left\langle \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}}, \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}}\right\rangle \\ &\le \|\boldsymbol{U}\boldsymbol{\Delta_U}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{V}\boldsymbol{\Delta_V}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{\Delta_V}\boldsymbol{U}^{\mathrm{T}}\|_F^2,\end{aligned}$$

which implies

$$\Pi_1 + \mu\Pi_3 + \mu\Pi_4 \le 2\max\{\beta, \mu\} \cdot \|\boldsymbol{W}\boldsymbol{\Delta}^{\mathrm{T}}\|_F^2. \tag{8}$$

Bounding terms Π_2 and Π_5

To obtain an upper bound for the term Π_2, we utilize the fact that $\boldsymbol{\Delta_U} = \boldsymbol{U} - \boldsymbol{U}^\star\boldsymbol{R}$ and $\boldsymbol{\Delta_V} = \boldsymbol{V} - \boldsymbol{V}^\star\boldsymbol{R}$, which yields

$$\begin{aligned}\Pi_2 &= \left\langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta_U}\boldsymbol{\Delta_V}^{\mathrm{T}}\right\rangle \\ &= \left\langle \nabla f(\boldsymbol{X}), (\boldsymbol{U} - \boldsymbol{U}^\star\boldsymbol{R})(\boldsymbol{V} - \boldsymbol{V}^\star\boldsymbol{R})^{\mathrm{T}}\right\rangle \\ &= \left\langle \nabla f(\boldsymbol{X}), \boldsymbol{X} + \boldsymbol{X}^\star - \boldsymbol{U}^\star\boldsymbol{R}^{\mathrm{T}}\boldsymbol{V}^T - \boldsymbol{U}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{V}^{\star\mathrm{T}}\right\rangle \\ &\overset{(i)}{=} -\langle \nabla f(\boldsymbol{X}), \boldsymbol{X} - \boldsymbol{X}^\star\rangle - \mu\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, (\boldsymbol{W} - \boldsymbol{W}^\star\boldsymbol{R})\boldsymbol{W}^{\mathrm{T}}\right\rangle \\ &\overset{(ii)}{=} -\langle \nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{X}^\star), \boldsymbol{X} - \boldsymbol{X}^\star\rangle - \mu\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, (\boldsymbol{W} - \boldsymbol{W}^\star\boldsymbol{R})\boldsymbol{W}^{\mathrm{T}}\right\rangle \\ &\overset{(iii)}{\le} -\alpha\|\boldsymbol{X} - \boldsymbol{X}^\star\|_F^2 - \mu\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, (\boldsymbol{W} - \boldsymbol{W}^\star\boldsymbol{R})\boldsymbol{W}^{\mathrm{T}}\right\rangle,\end{aligned}$$

where (i) follows because $\boldsymbol{W}$ is the critical point satisfying

$$\begin{bmatrix}\nabla f(\boldsymbol{X})\boldsymbol{V} \\ \nabla f(\boldsymbol{X})^{\mathrm{T}}\boldsymbol{U}\end{bmatrix} = -\mu\widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W},$$

(ii) utilizes $\nabla f(\boldsymbol{X}^\star) = \boldsymbol{0}$, and (iii) follows by using the strict convexity property (3):

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

where the first line follows from the integral form of the mean value theorem for vector-valued functions (see [41, Eq. (A.57)]), and the second line uses the fact that $t\boldsymbol{X} + (1-t)\boldsymbol{X}^\star$ has rank at most $2r$ and

the restricted strong convexity of Hessian $\nabla^2 f(\cdot)$. Note that

$$\begin{aligned}
2\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, (\boldsymbol{W}-\boldsymbol{W}^\star\boldsymbol{R})\boldsymbol{W}^{\mathrm{T}}\right\rangle - \Pi_5 &= \left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, (\boldsymbol{W}-\boldsymbol{W}^\star\boldsymbol{R})(\boldsymbol{W}+\boldsymbol{W}^\star\boldsymbol{R})^{\mathrm{T}}\right\rangle \\
&= \left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\rangle \\
&\geq \left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\rangle - \left\langle \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\rangle \\
&= \left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}} - \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\rangle,
\end{aligned}$$

where the inequality follows because $\left\langle \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\right\rangle \geq 0$ and $\left\langle \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\rangle = 0$ by (4). We then have

$$\begin{aligned}
2\Pi_2 + \mu\Pi_5 &\leq -2\alpha\|\boldsymbol{X}-\boldsymbol{X}^\star\|_F^2 - 2\mu\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, (\boldsymbol{W}-\boldsymbol{W}^\star\boldsymbol{R})\boldsymbol{W}^{\mathrm{T}}\right\rangle + \mu\Pi_5 \\
&\leq -2\alpha\|\boldsymbol{X}-\boldsymbol{X}^\star\|_F^2 - \mu\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}} - \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\rangle \qquad (9) \\
&\leq -\min\{\alpha-\mu,\mu\}\cdot\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2.
\end{aligned}$$

Merging together

Putting (8) and (9) together gives

$$\begin{aligned}
[\nabla^2\rho(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta}) &= \Pi_1 + 2\Pi_2 + \mu\Pi_3 + \mu\Pi_4 + 2\Pi_2 + \mu\Pi_5 \\
&\leq 2\max\{\beta,\mu\}\cdot\|\boldsymbol{W}\boldsymbol{\Delta}^{\mathrm{T}}\|_F^2 - \min\{\alpha-\mu,\mu\}\cdot\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2 \\
&\overset{(i)}{\leq} \left(-\frac{3}{4}\alpha + \frac{1}{2}\beta + 8\beta\left(3 + \frac{1}{2(\sqrt{2}-1)}\right)\left(\frac{\beta-\alpha}{\beta+\alpha}\right)^2\right)\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2 \\
&\overset{(ii)}{\leq} -0.069\alpha\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2 \\
&\overset{(iii)}{\leq} -0.069\alpha\left(\sigma_{r'}^2(\boldsymbol{W}) + 2\sigma_{r'}(\boldsymbol{W})\sigma_{r'}(\boldsymbol{W}^\star) + \sigma_{r'}^2(\boldsymbol{W}^\star)\right)\|\boldsymbol{\Delta}\|_F^2 \\
&\overset{(iv)}{\leq} -0.069\alpha\left(\sigma_{r'}^2(\boldsymbol{W}) + \sqrt{2}\sigma_{r'}(\boldsymbol{W})\sqrt{\sigma_{r'}(\boldsymbol{X}^\star)} + 2\sigma_{r'}(\boldsymbol{X}^\star)\right)\|\boldsymbol{\Delta}\|_F^2,
\end{aligned}$$

where $r' = \max\{r^\star, r^c\}$ and (i) utilizes $\mu = \frac{\alpha+\beta}{4}$, Lemma 1 and Lemma 3, (ii) holds for $\frac{\beta}{\alpha} \leq 1.12$, (iii) follows from Lemma 2, and (iv) follows because $\sigma_\ell(\boldsymbol{W}^\star) = \sqrt{2\sigma_\ell(\boldsymbol{X}^\star)}$ by noting that

$$\boldsymbol{W}^\star = \begin{bmatrix}\boldsymbol{Q}_{\boldsymbol{U}^\star}\boldsymbol{\Sigma}^{\star 1/2} \\ \boldsymbol{Q}_{\boldsymbol{V}^\star}\boldsymbol{\Sigma}^{\star 1/2}\end{bmatrix} = \begin{bmatrix}\boldsymbol{Q}_{\boldsymbol{U}^\star}/\sqrt{2} \\ \boldsymbol{Q}_{\boldsymbol{V}^\star}/\sqrt{2}\end{bmatrix}\left(\sqrt{2}\boldsymbol{\Sigma}^{\star 1/2}\right)\mathbf{I}$$

is an SVD of $\boldsymbol{W}^\star$.

□

4 Conclusion

This paper considers low-rank matrix optimization with general objective functions. To reduce the computational complexity, we apply a matrix factorization technique to reformulate the convex program. Although the resulting optimization problem is not convex, we show the reformulated objection function has simple landscapes: each critical point of the factored objective function either corresponds to the global optimal solution of the original convex program or is a strict saddle point such that the Hessian at this point has a strictly negative eigenvalue. This property guarantees that many local search algorithms (such as gradient descent and trust region method) can converge to the global optimum from a random initialization.

References

[1] Scott Aaronson. The learnability of quantum states. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 463, pages 3089–3114. The Royal Society, 2007.

[2] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *COLT*, pages 123–137, 2014.

[3] Anima Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. *arXiv preprint arXiv:1602.05908*, 2016.

[4] Anima Anandkumar, Rong Ge, and Majid Janzamin. Analyzing tensor power method dynamics: Applications to learning overcomplete latent variable models. *CoRR abs/1411.1488*, 17, 2014.

[5] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. *arXiv preprint*, 2015.

[6] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.

[7] Pratik Biswas, Tzu-Chen Liang, Kim-Chuan Toh, Yinyu Ye, and Ta-Chung Wang. Semidefinite programming approaches for sensor network localization with noisy distance measurements. *IEEE transactions on automation science and engineering*, 3(4):360, 2006.

[8] Pratik Biswas and Yinyu Ye. Semidefinite programming for ad hoc wireless sensor network localization. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 46–54. ACM, 2004.

[9] Nicolas Boumal. Nonconvex phase synchronization. *arXiv preprint arXiv:1601.06114*, 2016.

[10] Thierry Bouwmans, Necdet Serhat Aybat, and El-hadi Zahzah. *Handbook of Robust Low-Rank and Sparse Matrix Decomposition: Applications in Image and Video Processing*. CRC Press, 2016.

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

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

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

[14] Tony Cai and Wen-Xin Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *Journal of Machine Learning Research*, 14(1):3619–3647, 2013.

[15] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

[16] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[17] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[18] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[19] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[20] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[21] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[22] Yuxin Chen and Emmanuel Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747, 2015.

[23] Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

[24] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[25] Dennis DeCoste. Collaborative prediction using ensembles of maximum margin matrix factorizations. In *Proceedings of the 23rd international conference on Machine learning*, pages 249–256. ACM, 2006.

[26] Steven T Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012.

[27] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

[28] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.

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

[30] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[31] Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *arXiv preprint arXiv:1605.08370*, 2016.

[32] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *University of California, Berkeley*, 1050:16, 2016.

[33] Kiryung Lee and Marius Junge. Rip-like properties in subsampled blind deconvolution. *arXiv preprint arXiv:1511.06146*, 2015.

[34] Kiryung Lee, Ning Tian, and Justin Romberg. Fast and guaranteed blind multichannel deconvolution under a bilinear system model. *arXiv preprint arXiv:1610.06469*, 2016.

[35] Qiuwei Li, Ashley Prater, Lixin Shen, and Gongguo Tang. Overcomplete tensor decomposition via convex optimization. *arXiv preprint arXiv:1602.08614*, 2016.

[36] Qiuwei Li and Gongguo Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. *arXiv:1611.03060*, 2016.

[37] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv preprint arXiv:1606.04933*, 2016.

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

[39] Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

[40] Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

[41] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[42] Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. *arXiv preprint arXiv:1609.03240*, 2016.

[43] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

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

[45] Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.

[46] Joseph Salmon, Zachary Harmany, Charles-Alban Deledalle, and Rebecca Willett. Poisson noise reduction with non-local pca. *Journal of mathematical imaging and vision*, 48(2):279–294, 2014.

[47] Eduardo D Sontag and Héctor J Sussmann. Backpropagation can give rise to spurious local minima even for networks without hidden layers. *Complex Systems*, 3(1):91–106, 1989.

[48] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2004.

[49] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *arXiv preprint arXiv:1511.04777*, 2015.

[50] Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

[51] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*, 2016.

[52] Stephen Tu, Ross Boczar, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

[53] Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *arXiv preprint arXiv:1410.0342*, 2014.

[54] Lingxiao Wang, Xiao Zhang, and Quanquan Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. *arXiv preprint arXiv:1610.05275*, 2016.

[55] Markus Weimer, Alexandros Karatzoglou, Quoc Viet Le, and Alex Smola. Maximum margin matrix factorization for collaborative ranking. *Advances in neural information processing systems*, pages 1–8, 2007.

[56] Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.

A Proof of Lemma 1

Proof of Lemma 1. First recall the notation $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$, $\boldsymbol{X}^\star = \boldsymbol{U}^\star\boldsymbol{V}^\star$, and

$$\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}, \quad \widehat{\boldsymbol{W}} = \begin{bmatrix}\boldsymbol{U}\\ -\boldsymbol{V}\end{bmatrix}, \boldsymbol{W}^\star = \begin{bmatrix}\boldsymbol{U}^\star\\ \boldsymbol{V}^\star\end{bmatrix}, \quad \widehat{\boldsymbol{W}}^\star = \begin{bmatrix}\boldsymbol{U}^\star\\ -\boldsymbol{V}^\star\end{bmatrix}.$$

Now note that any critical point $\boldsymbol{W}$ satisfies

$$\begin{bmatrix}\boldsymbol{0} & \nabla f(\boldsymbol{X})\\ \nabla f(\boldsymbol{X})^{\mathrm{T}} & \boldsymbol{0}\end{bmatrix}\boldsymbol{W} = -\mu\widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W},$$

which gives

$$\begin{aligned}
-\mu\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{Z}\boldsymbol{W}^{\mathrm{T}}\right\rangle &= \left\langle \begin{bmatrix}\boldsymbol{0} & \nabla f(\boldsymbol{X})\\ \nabla f(\boldsymbol{X})^{\mathrm{T}} & \boldsymbol{0}\end{bmatrix}, \boldsymbol{Z}\boldsymbol{W}^{\mathrm{T}}\right\rangle\\
&= \left\langle \begin{bmatrix}\boldsymbol{0} & \nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{X}^\star)\\ \nabla f(\boldsymbol{X})^{\mathrm{T}} - \nabla f(\boldsymbol{X}^\star)^{\mathrm{T}} & \boldsymbol{0}\end{bmatrix}, \boldsymbol{Z}\boldsymbol{W}^{\mathrm{T}}\right\rangle\\
&= \underbrace{\left\langle \nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{X}^\star) - \frac{\alpha+\beta}{2}(\boldsymbol{X}-\boldsymbol{X}^\star), \boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Z}_{\boldsymbol{V}}^{\mathrm{T}}\right\rangle}_{\urcorner_1} \qquad (10)\\
&\quad + \frac{\alpha+\beta}{2}\underbrace{\left\langle \boldsymbol{X}-\boldsymbol{X}^\star, \boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Z}_{\boldsymbol{V}}^{\mathrm{T}}\right\rangle}_{\urcorner_2}
\end{aligned}$$

for any $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z_U} \\ \boldsymbol{Z_V} \end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$. Here the second line utilizes the fact $\nabla f(\boldsymbol{X}^\star) = \boldsymbol{0}$. We bound $\urcorner_1$ by first using integral form of the mean value theorem for $\nabla f(\boldsymbol{X})$:

$$\urcorner_1 = \int_0^1 \left[\nabla^2 f(t\boldsymbol{X} + (1-t)\boldsymbol{X}^\star)\right](\boldsymbol{X} - \boldsymbol{X}^\star, \boldsymbol{Z_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Z_V}^{\mathrm{T}})\,\mathrm{d}\,t - \frac{\alpha+\beta}{2}\left\langle \boldsymbol{X} - \boldsymbol{X}^\star, \boldsymbol{Z_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Z_V}^{\mathrm{T}}\right\rangle.$$

Noting that $\mathrm{rank}(t\boldsymbol{X} + (1-t)\boldsymbol{X}^\star) \le 2r$, it follows from Proposition 2 that

$$\begin{aligned}
|\urcorner_1| &\le \int_0^1 \left|\left[\nabla^2 f(t\boldsymbol{X} + (1-t)\boldsymbol{X}^\star)\right](\boldsymbol{X} - \boldsymbol{X}^\star, \boldsymbol{Z_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Z_V}^{\mathrm{T}}) - \frac{\alpha+\beta}{2}\left\langle \boldsymbol{X} - \boldsymbol{X}^\star, \boldsymbol{Z_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Z_V}^{\mathrm{T}}\right\rangle\right| \mathrm{d}\,t \\
&\le \frac{\beta-\alpha}{2}\left\|\boldsymbol{X} - \boldsymbol{X}^\star\right\|_F \left\|\boldsymbol{Z_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Z_V}^{\mathrm{T}}\right\|_F,
\end{aligned}$$

which when plugged into (10) gives

$$\mu\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{Z}\boldsymbol{W}^{\mathrm{T}}\right\rangle + \frac{\alpha+\beta}{2}\urcorner_2 = -\urcorner_1 \le \frac{\beta-\alpha}{2}\left\|\boldsymbol{X} - \boldsymbol{X}^\star\right\|_F \left\|\boldsymbol{Z_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Z_V}^{\mathrm{T}}\right\|_F. \tag{11}$$

Now let $\boldsymbol{Z} = (\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}})\boldsymbol{W}^{T\dagger}$, which gives $\boldsymbol{Z}\boldsymbol{W}^{\mathrm{T}} = (\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}})\boldsymbol{P_W}$. Here $\dagger$ denotes the pseudoinverse of a matrix and $\boldsymbol{P_W}$ is the orthogonal projector onto the range of $\boldsymbol{W}$. Utilizing the fact $\mu = \frac{\alpha+\beta}{4}$, we further connect the left hand side of (11) with $\left\|\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{P_W}\right\|_F^2$ by

$$\begin{aligned}
&\frac{\alpha+\beta}{2}\urcorner_2 + \mu\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{Z}\boldsymbol{W}^{\mathrm{T}}\right\rangle \\
&= \frac{\alpha+\beta}{4}\left\langle \begin{bmatrix} \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} & \boldsymbol{X} - 2\boldsymbol{X}^\star \\ \boldsymbol{X}^{\mathrm{T}} - 2\boldsymbol{X}^{\star\mathrm{T}} & \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} \end{bmatrix}, \left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{P_W}\right\rangle \\
&= \frac{\alpha+\beta}{4}\left\langle \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}, \left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{P_W}\right\rangle \\
&\quad + \frac{\alpha+\beta}{4}\left\langle \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{P_W}\right\rangle \\
&\ge \frac{\alpha+\beta}{4}\left\langle \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}, \left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{P_W}\right\rangle \\
&= \frac{\alpha+\beta}{4}\left\|\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{P_W}\right\|_F^2,
\end{aligned} \tag{12}$$

where the inequality follows because $\left\langle \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\boldsymbol{P_W}\right\rangle = 0$ by (4) and $\left\langle \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{P_W}\right\rangle = \left\langle \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\right\rangle \ge 0$ since it is the inner product between two PSD matrices.

On the other hand, we give an upper bound on the right hand side of (11)

$$\begin{aligned}
\left\|\boldsymbol{X} - \boldsymbol{X}^\star\right\|_F \left\|\boldsymbol{Z_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Z_V}^{\mathrm{T}}\right\|_F &\le \frac{\sqrt{2}}{2}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F \sqrt{2\left\|\boldsymbol{Z_U}\boldsymbol{V}^{\mathrm{T}}\right\|_F^2 + 2\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{Z_V}^{\mathrm{T}}\right\|_F^2} \\
&\le \left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F \left\|\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{P_W}\right\|_F,
\end{aligned}$$

which together with (11) and (12) completes the proof. □