The Global Optimization Geometry of Low-Rank Matrix Optimization

Zhihui Zhu, *Member, IEEE*, Qiuwei Li, *Student Member, IEEE*, Gongguo Tang, *Member, IEEE*, and Michael B. Wakin, *Fellow, IEEE*

***Abstract*— This paper considers general rank-constrained optimization problems that minimize a general objective function $f(\boldsymbol{X})$ over the set of rectangular $n \times m$ matrices that have rank at most r. To tackle the rank constraint and also to reduce the computational burden, we factorize $\boldsymbol{X}$ into $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $n \times r$ and $m \times r$ matrices, respectively, and then optimize over the small matrices $\boldsymbol{U}$ and $\boldsymbol{V}$. We characterize the global optimization geometry of the nonconvex factored problem and show that the corresponding objective function satisfies the robust strict saddle property as long as the original objective function f satisfies restricted strong convexity and smoothness properties, ensuring global convergence of many local search algorithms (such as noisy gradient descent) in polynomial time for solving the factored problem. We also provide a comprehensive analysis for the optimization geometry of a matrix factorization problem where we aim to find $n \times r$ and $m \times r$ matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ such that $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ approximates a given matrix $\boldsymbol{X}^\star$. Aside from the robust strict saddle property, we show that the objective function of the matrix factorization problem has no spurious local minima and obeys the strict saddle property not only for the exact-parameterization case where $\operatorname{rank}(\boldsymbol{X}^\star) = r$, but also for the over-parameterization case where $\operatorname{rank}(\boldsymbol{X}^\star) < r$ and the under-parameterization case where $\operatorname{rank}(\boldsymbol{X}^\star) > r$. These geometric properties imply that a number of iterative optimization algorithms (such as gradient descent) converge to a global solution with random initialization.**

***Index Terms*— Low-rank optimization, matrix factorization, matrix sensing, nonconvex optimization, optimization geometry.**

I. Introduction

LOW-RANK matrices arise in a wide variety of applications throughout science and engineering, ranging from quantum tomography [1], signal processing [2], machine learning [3], [4], and so on; see [5] for a comprehensive review. In all of these settings, we often encounter the following rank-constrained optimization problem:

$$\begin{aligned} &\operatorname*{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times m}} f(\boldsymbol{X}),\\ &\text{subject to } \operatorname{rank}(\boldsymbol{X}) \le r, \end{aligned} \tag{1}$$

where the objective function $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ is smooth.

Whether the objective function f is convex or nonconvex, the rank constraint renders low-rank matrix optimizations of the form (1) highly nonconvex and computationally NP-hard in general [6]. Significant efforts have been devoted to transforming (1) into a convex problem by replacing the rank constraint with one involving the nuclear norm. This strategy has been widely utilized in matrix inverse problems [7] arising in signal processing [5], machine learning [8], and control [6]. With convex analysis techniques, nuclear norm minimization has been proved to provide optimal performance in recovering low-rank matrices [9]. However, in spite of the optimal performance, solving nuclear norm minimization is very computationally expensive even with specialized first-order algorithms. For example, the singular value thresholding algorithm [10] requires performing an expensive singular value decomposition (SVD) in each iteration, making it computationally prohibitive in large-scale settings. This prevents nuclear norm minimization from scaling to practical problems.

To relieve the computational bottleneck, recent studies propose to factorize the variable into $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$, and optimize over the $n \times r$ and $m \times r$ matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ rather than the $n \times m$ matrix $\boldsymbol{X}$. The rank constraint in (1) then is automatically satisfied through the factorization. This strategy is usually referred to as the Burer-Monteiro type decomposition after the authors in [11], [12]. Plugging this parameterization of $\boldsymbol{X}$ in (1), we can recast the program into the following one:

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

The bilinear nature of the parameterization renders the objective function of (2) nonconvex. Hence, it can potentially have spurious local minima (i.e., local minimizers that are not global minimizers) or even saddle points. With technical innovations in analyzing the landscape of nonconvex functions, however, several recent works have shown that the factored objective function $h(\boldsymbol{U},\boldsymbol{V})$ in certain matrix inverse problems has no spurious local minima [13]–[15].

Manuscript received January 4, 2018; revised July 26, 2019; accepted August 28, 2020. Date of publication January 5, 2021; date of current version January 21, 2021. This work was supported in part by the NSF under Grant CCF-1409261, Grant CCF-1464205, Grant 2008460, under Grant CAREER CCF-1149225; and in part by the DARPA Lagrange Program under Award N660011824020. *(Zhihui Zhu and Qiuwei Li contributed equally to this work.) (Corresponding author: Zhihui Zhu.)*

Zhihui Zhu is with the Department of Electrical and Computer Engineering, University of Denver, Denver, CO 80208 USA (e-mail: zhihui.zhu@du.edu).

Qiuwei Li is with the Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: qiuweili@math.ucla.edu).

Gongguo Tang and Michael B. Wakin are with the Department of Electrical Engineering, Colorado School of Mines, Golden, CO 80401 USA (e-mail: gtang@mines.edu; mwakin@mines.edu).

Communicated by P. Grohs, Associate Editor for Signal Processing.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3049171.

Digital Object Identifier 10.1109/TIT.2021.3049171

A. Summary of Results and Outline

In this paper, we provide a comprehensive geometric analysis for solving general low-rank optimizations of the form (1) using the factorization approach (2). Our work actually rests on the recent works [16]–[20] ensuring a number of iterative optimization methods (such as gradient descent) converge to a local minimum with random initialization provided the problem satisfies the so-called strict saddle property (see Definition 3 in Section II). If the objective function further obeys the robust strict saddle property [16] (see Definition 4 in Section II) or belongs to the class of so-called $\mathcal{X}$ functions [17], the recent works [16], [17] show that many local search algorithms can converge to a local minimum in polynomial time. The implications of this line of work have had a tremendous impact on a number of nonconvex problems in applied mathematics, signal processing, and machine learning.

We begin this paper in Section II with the notions of strict saddle, strict saddle property, and robust strict saddle property. Considering that many invariant functions are not strongly convex (or even convex) in any neighborhood around a local minimum point, we then provide a revised robust strict saddle property[1] requiring a regularity condition (see Definition 8 in Section II) rather than strong convexity near the local minimum points (which is one of the requirements for the strict saddle property). The stochastic gradient descent algorithm is guaranteed to converge to a local minimum point in polynomial time for problems satisfying the revised robust strict saddle property [16], [20].

In Section III, we consider the geometric analysis for solving general low-rank optimizations of the form (1) using the factorization approach (2). Provided the objective function f satisfies certain restricted strong convexity and smoothness conditions, we show that the low-rank optimization problem with the factorization (2) (with an additional regularizer—see Section III for the details) obeys the revised robust strict saddle property. In Section III-C, we consider a stylized application in matrix sensing where the measurement operator satisfies the restricted isometry property (RIP) [7]. In the case of Gaussian measurements, as guaranteed by this robust strict saddle property, a number of iterative optimizations can find the unknown matrix $\boldsymbol{X}^\star$ of rank r in polynomial time with high probability when the number of measurements exceeds a constant times $(n+m)r^2$.

Our main approach for analyzing the optimization geometry of (2) is based on the geometric analysis for the following non-square low-rank matrix factorization problem: given $\boldsymbol{X}^\star \in \mathbb{R}^{n\times m}$,

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r},\boldsymbol{V}^{m\times r}}{\text{minimize}} \left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star\right\|_F^2. \tag{3}$$

In particular, we show the optimization geometry for the low-rank matrix factorization problem (3) is preserved for the general low-rank optimization (2) under certain restricted strong convexity and smoothness conditions on f. Thus, in Appendix A, we provide a comprehensive geometric analysis for (3), which can be viewed as an important foundation of many popular matrix factorization problems such as the matrix sensing problem and matrix completion. We show that the low-rank matrix factorization problem (3) (with an additional regularizer) has no spurious local minima and obeys the strict saddle property—that is the objective function in (3) has a directional negative curvature at all critical points but local minima—not only for the exact-parameterization case where $\operatorname{rank}(\boldsymbol{X}^\star) = r$, but also for the over-parameterization case where $\operatorname{rank}(\boldsymbol{X}^\star) < r$ and the under-parameterization case where $\operatorname{rank}(\boldsymbol{X}^\star) > r$. The strict saddle property and lack of spurious local minima ensure that a number of local search algorithms applied to the matrix factorization problem (3) converge to global optima which correspond to the best rank-r approximation to $\boldsymbol{X}^\star$. Further, we completely analyze the low-rank matrix factorization problem (3) for the exact-parameterization case and show that it obeys the revised robust strict saddle property.

B. Relation to Existing Work

Unlike the objective functions of convex optimizations that have simple landscapes, such as where all local minimizers are global ones, the objective functions of general nonconvex programs have much more complicated landscapes. In recent years, by exploiting the underlying optimization geometry, a surge of progress has been made in providing theoretical justifications for matrix factorization problems such as (2) using a number of previously heuristic algorithms (such as alternating minimization [21], gradient descent, and the trust region method). Typical examples include phase retrieval [22]–[24], blind deconvolution [25], [26], dictionary learning [27]–[29], phase synchronization [30] and matrix sensing and completion [14], [31]–[36].

These iterative algorithms can be sorted into two categories based on whether a good initialization is required. One set of algorithms consist of two steps: initialization and local refinement. Provided the function satisfies a regularity condition or similar properties, 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 [23], phase synchronization [30] and low-rank matrix recovery problems [31], [32], [37], [38]. As we have mentioned, a regularity condition is also adopted in the revised robust strict saddle property.

Another category of works attempt to analyze the landscape of the objective functions in a larger space rather than the regions near the global optima. We can further separate these approaches into two types based on whether they involve the strict saddle property or the robust strict saddle property. The strict saddle property and lack of spurious local minima are proved for low-rank, positive semidefinite (PSD) matrix recovery [13] and completion [14], PSD matrix optimization

[1] A similar notion of a revised robust strict saddle property has also been utilized in [20], which shows that noisy gradient descent converges to a local minimum in a number iterations that depends only poly-logarithmically on the dimension. In a nutshell, [20] has a different focus than this work: the focus in [20] is on providing convergence analysis of a noisy gradient descent algorithm with a robust strict saddle property, while in the present paper, we establish a robust strict saddle property for the nonsymmetric matrix factorization and more general low-rank optimization (including matrix sensing) problems with the factorization approach.

problems with generic objective functions [39], low-rank non-square matrix estimation from linear observations [15], low-rank nonsquare optimization problems with generic objective functions [40] and generic nuclear norm regularized problems [39]. The strict saddle property along with the lack of spurious local minima ensures a number of iterative algorithms such as gradient descent [16] and the trust region method [41] converge to the global minimum with random initialization [16], [18], [29].

A few other works which are closely related to our work attempt to study the *global geometry* by characterizing the landscapes of the objective functions in the whole space rather than the regions near the global optima or all the critical points. As we discussed before, a number of local search algorithms are guaranteed to find a local optimum (which is also the global optimum if there are no spurious local minima) because of this robust strict saddle property. In [16], the authors proved that tensor decomposition problems satisfy this robust strict saddle property. Sun *et al.* [22] studied the global geometry of the phase retrieval problem. The very recent work in [42] analyzed the global geometry for PSD low-rank matrix factorization of the form (3) and the related matrix sensing problem when the rank is exactly parameterized (i.e., $r = \operatorname{rank}(\boldsymbol{X}^\star)$). The factorization approach for matrix inverse problems with quadratic loss functions is considered in [36]. We extend this line by considering general rank-constrained optimization problems including a set of matrix inverse problems.

Finally, we remark that our work is also closely related to the recent works in low-rank matrix factorization of the form (3) and its variants [13]–[15], [31]–[33], [36], [40], [42]. As we discussed before, most of these works except [36], [42] (but including [15] which also focuses on nonsymmetric matrix sensing) only characterize the geometry either near the global optima or all the critical points. Instead, we characterize the *global*geometry for general (rather than PSD) low-rank matrix factorization and sensing. Because the analysis is different, the proof strategy in the present paper is also very different than that of [15], [40]. The results for PSD matrix sensing in [42] build heavily on the concentration properties of Gaussian measurements, while our results for matrix sensing depend on the RIP of the measurement operator and thus can be applied to other matrix sensing problems whose measurement operator is not necessarily from a Gaussian measurement ensemble. Also, [36] considers matrix inverse problems with quadratic loss functions and its proof strategy is very different than that in the present paper: the proof in [36] is specified to quadratic loss functions, while we consider the rank-constrained optimization problem with general objective functions in (1) and our proof utilizes the fact that the gradient and Hessian of the low-rank matrix sensing are respectively very close to those in low-rank matrix factorization. Furthermore, in terms of the matrix factorization, we show that the objective function in (3) obeys the strict saddle property and has no spurious local minima not only for exact-parameterization ($r = \operatorname{rank}(\boldsymbol{X}^\star)$), but also for over-parameterization ($r > \operatorname{rank}(\boldsymbol{X}^\star)$) and under-parameterization ($r < \operatorname{rank}(\boldsymbol{X}^\star)$). Local (rather than global) geometry results for exact-parameterization and under-parameterization are also covered in [40]. As noted above, the work in [36], [42] for low-rank matrix factorization only focuses on exact-parameterization ($r = \operatorname{rank}(\boldsymbol{X}^\star)$). The under-parameterization implies that we can find the best rank-r approximation to $\boldsymbol{X}^\star$ by many efficient iterative optimization algorithms such as gradient descent.

C. Notation

Before proceeding, we first briefly introduce some notation used throughout the paper. The symbols $\mathbf{I}$ and $\mathbf{0}$ respectively represent the identity and zero matrices with appropriate sizes. Also $\mathbf{I}_n$ is used to denote the $n \times n$ identity matrix. For any natural number n, we let $[n]$ or $1 : n$ denote the set $\{1, 2, ..., n\}$. We use $|\Omega|$ denote the cardinality (i.e., the number of elements) of a set Ω. MATLAB notations are adopted for matrix indexing; that is, for the $n \times m$ matrix $\boldsymbol{A}$, its (i, j)-th element is denoted by $\boldsymbol{A}[i, j]$, its i-th row (or column) is denoted by $\boldsymbol{A}[i, :]$ (or $\boldsymbol{A}[:, i]$), and $\boldsymbol{A}[\Omega_1, \Omega_2]$ refers to a $|\Omega_1| \times |\Omega_2|$ submatrix obtained by taking the elements in rows Ω_1 of columns Ω_2 of matrix $\boldsymbol{A}$. Here $\Omega_1 \subset [n]$ and $\Omega_2 \subset [n]$. We use $a \gtrsim b$ (or $a \lesssim b$) to represent that there is a constant so that $a \geq \text{Const} \cdot b$ (or $a \leq \text{Const} \cdot b$).

We say that a (not necessarily square) matrix $\boldsymbol{A} \in \mathbb{R}^{n \times r}$ is orthonormal if the columns of $\boldsymbol{A}$ are normalized and orthogonal to each other, i.e., $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \mathbf{I}$. The set of $r \times r$ orthonormal matrices is denoted by $\mathcal{O}_r := \{\boldsymbol{R} \in \mathbb{R}^{r \times r} : \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \mathbf{I}\}$. We say that a (not necessarily square) matrix $\boldsymbol{A} \in \mathbb{R}^{n \times r}$ is orthogonal if $\langle \boldsymbol{A}[:, i], \boldsymbol{A}[:, j] \rangle = 0$ for all $i \neq j$; that is the columns of $\boldsymbol{A}$ are orthogonal to each other, but are not necessarily normalized and could even be zero.

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 $[\nabla f(\boldsymbol{Z})][i, j] = \frac{\partial f(\boldsymbol{Z})}{\partial \boldsymbol{Z}[i,j]}$ 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 $[\nabla^2 f(\boldsymbol{Z})][i, j] = \frac{\partial^2 f(\boldsymbol{Z})}{\partial \boldsymbol{z}[i] \partial \boldsymbol{z}[j]}$ 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 $[\nabla^2 f(\boldsymbol{Z})](\boldsymbol{A}, \boldsymbol{B}) = \sum_{i,j,k,l} \frac{\partial^2 f(\boldsymbol{Z})}{\partial \boldsymbol{Z}[i,j] \partial \boldsymbol{Z}[k,\ell]} \boldsymbol{A}[i, j]\boldsymbol{B}[k, \ell]$ 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.

II. Preliminaries

In this section, we provide a number of important definitions in optimization and group theory. To begin, suppose $h(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable.

Definition 1 (Critical Points): A point $\boldsymbol{x}$ is a critical point of $h(\boldsymbol{x})$ if $\nabla h(\boldsymbol{x}) = \mathbf{0}$.

Definition 2 (Strict Saddles; or Ridable Saddles in [29]): A critical point $\boldsymbol{x}$ is a strict saddle if the Hessian matrix evaluated at this point has a strictly negative eigenvalue, i.e., $\lambda_{\min}(\nabla^2 h(\boldsymbol{x})) < 0$.

Definition 3 (Strict Saddle Property [16]): A twice differentiable function satisfies the strict saddle property if each critical point either corresponds to a local minimum or is a strict saddle.

Intuitively, the strict saddle property requires a function to have a directional negative curvature at all of the critical points but local minima. This property allows a number of iterative algorithms such as noisy gradient descent [16] and the trust region method [41] to further decrease the function value at all the strict saddles and thus converge to a local minimum.

In [16], the authors proposed a noisy gradient descent algorithm for the optimization of functions satisfying the robust strict saddle property.

Definition 4 (Robust Strict Saddle Property [16]): Given $\alpha, \gamma, \epsilon, \delta$, a twice differentiable $h(\boldsymbol{x})$ satisfies the $(\alpha, \gamma, \epsilon, \delta)$-robust strict saddle property if for every point $\boldsymbol{x}$ at least one of the following applies:

1) There exists a local minimum point $\boldsymbol{x}^\star$ such that $\|\boldsymbol{x}^\star - \boldsymbol{x}\| \le \delta$, and the function $h(\boldsymbol{x}')$ restricted to a 2δ neighborhood of $\boldsymbol{x}^\star$ (i.e., $\|\boldsymbol{x}^\star - \boldsymbol{x}'\| \le 2\delta$) is α-strongly convex;
2) $\lambda_{\min}\left(\nabla^2 h(\boldsymbol{x})\right) \le -\gamma$;
3) $\|\nabla h(\boldsymbol{x})\| \ge \epsilon$.

In words, the above robust strict saddle property says that for any point whose gradient is small, then either the Hessian matrix evaluated at this point has a strictly negative eigenvalue, or it is close to a local minimum point. Thus the robust strict saddle property not only requires that the function obeys the strict saddle property, but also that it is well-behaved (i.e., strongly convex) near the local minima and has large gradient at the points far way to the critical points.

Intuitively, when the gradient is large, the function value will decrease in one step by gradient descent; when the point is close to a saddle point, the noise introduced in the noisy gradient descent could help the algorithm escape the saddle point and the function value will also decrease; when the point is close to a local minimum point, the algorithm then converges to a local minimum. Ge *et al.* [16] rigorously showed that the noisy gradient descent algorithm (see [16, Algorithm 1]) outputs a local minimum in a polynomial number of steps if the function $h(\boldsymbol{x})$ satisfies the robust strict saddle property.

It is proved in [16] that tensor decomposition problems satisfy this robust strict saddle property. However, requiring the local strong convexity prohibits the potential extension of the analysis in [16] for the noisy gradient descent algorithm to many other problems, for which it is not possible to be strongly convex in any neighborhood around the local minimum points. Typical examples include the matrix factorization problems due to the rotational degrees of freedom for any critical point. This motivates us to weaken the local strong convexity assumption relying on the approach used by [23], [31] and to provide the following revised robust strict saddle property for such problems. To that end, we list some necessary definitions related to groups and invariance of a function under the group action.

Definition 5 (Definition 7.1 [43])): A (closed) binary operation, $\circ$, is a law of composition that produces an element of a set from two elements of the same set. More precisely, let $\mathcal{G}$ be a set and $a_1, a_2 \in \mathcal{G}$ be arbitrary elements. Then $(a_1, a_2) \to a_1 \circ a_2 \in \mathcal{G}$.

Definition 6 (Definition 7.2 [43])): A **group** is a set $\mathcal{G}$ together with a (closed) binary operation $\circ$ such that for any elements $a, a_1, a_2, a_3 \in \mathcal{G}$ the following properties hold:

- Associative property: $a_1 \circ (a_2 \circ a_3) = (a_1 \circ a_2) \circ a_3$.
- There exists an identity element $e \in \mathcal{G}$ such that $e \circ a = a \circ e = a$.
- There is an element $a^{-1} \in \mathcal{G}$ such that $a^{-1} \circ a = a \circ a^{-1} = e$.

With this definition, it is common to denote a group just by $\mathcal{G}$ without saying the binary operation $\circ$ when it is clear from the context.

Definition 7: Given a function $h(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ and a group $\mathcal{G}$ of operators on $\mathbb{R}^n$, we say h is invariant under the group action (or under an element a of the group) if

$$h(a(\boldsymbol{x})) = h(\boldsymbol{x})$$

for all $\boldsymbol{x} \in \mathbb{R}^n$ and $a \in \mathcal{G}$.

Suppose the group action also preserves the energy of $\boldsymbol{x}$, i.e., $\|a(\boldsymbol{x})\| = \|\boldsymbol{x}\|$ for all $a \in \mathcal{G}$. Since for any $\boldsymbol{x} \in \mathbb{R}^n$, $h(a(\boldsymbol{x})) = h(\boldsymbol{x})$ for all $a \in \mathcal{G}$, it is straightforward to stratify the domain of $h(\boldsymbol{x})$ into equivalent classes. The vectors in each of these equivalent classes differ by a group action. One implication is that when considering the distance of two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, it would be helpful to use the distance between their corresponding classes:

$$\begin{aligned}\operatorname{dist}(\boldsymbol{x}_1, \boldsymbol{x}_2) :&= \min_{a_1 \in \mathcal{G}, a_2 \in \mathcal{G}} \|a_1(\boldsymbol{x}_1) - a_2(\boldsymbol{x}_2)\| \\ &= \min_{a \in \mathcal{G}} \|\boldsymbol{x}_1 - a(\boldsymbol{x}_2)\|, \end{aligned} \tag{4}$$

where the second equality follows because $\|a_1(\boldsymbol{x}_1) - a_2(\boldsymbol{x}_2)\| = \|a_1(\boldsymbol{x}_1 - a_1^{-1} \circ a_2(\boldsymbol{x}_2))\| = \|\boldsymbol{x}_1 - a_1^{-1} \circ a_2(\boldsymbol{x}_2)\|$ and $a_1^{-1} \circ a_2 \in \mathcal{G}$. Another implication is that the function $h(\boldsymbol{x})$ cannot possibly be strongly convex (or even convex) in any neighborhood around its local minimum points because of the existence of the equivalent classes. Before presenting the revised robust strict saddle property for invariant functions, we list two examples to illuminate these concepts.

Example 1: As one example, consider the phase retrieval problem of recovering an n-dimensional complex vector $\boldsymbol{x}^\star$ from $\left\{y_i = \left|\boldsymbol{b}_i^{\mathrm{H}} \boldsymbol{x}^\star\right|, i = 1, \ldots, p\right\}$, the magnitude of its projection onto a collection of known complex vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_p$ [22], [23]. The unknown $\boldsymbol{x}^\star$ can be estimated by solving the following natural least-squares formulation [22], [23]

$$\operatorname*{minimize}_{\boldsymbol{x} \in \mathbb{C}^n} h(\boldsymbol{x}) = \frac{1}{2p} \sum_{i=1}^{p} \left(y_i^2 - \left|\boldsymbol{b}_i^{\mathrm{H}} \boldsymbol{x}\right|^2\right)^2,$$

where we note that here the domain of $\boldsymbol{x}$ is $\mathbb{C}^n$. For this case, we denote the corresponding

$$\mathcal{G} = \{e^{j\theta} : \theta \in [0, 1)\}$$

and the group action as $a(\boldsymbol{x}) = e^{j\theta}\boldsymbol{x}$, where $a = e^{j\theta}$ is an element in $\mathcal{G}$. It is clear that $h(a(\boldsymbol{x})) = h(\boldsymbol{x})$ for all $a \in \mathcal{G}$. Due to this invariance of $h(\boldsymbol{x})$, it is impossible to recover the global phase factor of the unknown $\boldsymbol{x}^\star$ and the function $h(\boldsymbol{x})$ is not strongly convex in any neighborhood of $\boldsymbol{x}^\star$.

Example 2: As another example, we revisit the general factored low-rank optimization problem (2):

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

We recast the two variables $\boldsymbol{U},\boldsymbol{V}$ into $\boldsymbol{W}$ as $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$. For this example, we denote the corresponding $\mathcal{G} = \mathcal{O}_r$ and the group action on $\boldsymbol{W}$ as $a(\boldsymbol{W}) = \begin{bmatrix}\boldsymbol{U}\boldsymbol{R}\\ \boldsymbol{V}\boldsymbol{R}\end{bmatrix}$ where $a = \boldsymbol{R}\in\mathcal{G}$. We have that $h(a(\boldsymbol{W})) = h(\boldsymbol{W})$ for all $a\in\mathcal{G}$ since $\boldsymbol{U}\boldsymbol{R}(\boldsymbol{V}\boldsymbol{R})^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ for any $\boldsymbol{R}\in\mathcal{O}_r$. Because of this invariance, in general $h(\boldsymbol{W})$ is not strongly convex in any neighborhood around its local minimum points even though $f(\boldsymbol{X})$ is a strongly convex function; see [42] for the symmetric low-rank factorization problem and Theorem 2 in Appendix A for the nonsymmetric low-rank factorization problem.

In the examples illustrated above, due to the invariance, the function is not strongly convex (or even convex) in any neighborhood around its local minimum point and thus it is prohibitive to apply the standard approach in optimization to show the convergence in a small neighborhood around the local minimum point. To overcome this issue, Candès *et al.* [23] utilized the so-called regularity condition as a sufficient condition for local convergence of gradient descent applied for the phase retrieval problem. This approach has also been applied for the matrix sensing problem [31] and semi-definite optimization [37].

Definition 8 (Regularity Condition [23], [31]): Suppose $h(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ is invariant under the group action of the given group $\mathcal{G}$. Let $\boldsymbol{x}^\star \in \mathbb{R}^n$ be a local minimum point of $h(\boldsymbol{x})$. Define the set $B(\delta,\boldsymbol{x}^\star)$ as

$$B(\delta,\boldsymbol{x}^\star) := \{\boldsymbol{x}\in\mathbb{R}^n : \operatorname{dist}(\boldsymbol{x},\boldsymbol{x}^\star)\le\delta\},$$

where the distance $\operatorname{dist}(\boldsymbol{x},\boldsymbol{x}^\star)$ is defined in (4). Then we say the function $h(\boldsymbol{x})$ satisfies the (α,β,δ)-regularity condition if for all $\boldsymbol{x}\in B(\delta,\boldsymbol{x}^\star)$, we have

$$\langle\nabla h(\boldsymbol{x}), \boldsymbol{x} - a(\boldsymbol{x}^\star)\rangle \ge \alpha\operatorname{dist}(\boldsymbol{x},\boldsymbol{x}^\star)^2 + \beta\|\nabla h(\boldsymbol{x})\|^2, \quad (5)$$

where $a = \arg\min_{a'\in\mathcal{G}}\|\boldsymbol{x} - a'(\boldsymbol{x}^\star)\|$.

We remark that (α,β) in the regularity condition (8) must satisfy $\alpha\beta\le\frac{1}{4}$ since by applying Cauchy-Schwarz

$$\langle\nabla h(\boldsymbol{x}), \boldsymbol{x} - a(\boldsymbol{x}^\star)\rangle \le \|\nabla h(\boldsymbol{x})\|\operatorname{dist}(\boldsymbol{x},\boldsymbol{x}^\star)$$

and the inequality of arithmetic and geometric means

$$\alpha\operatorname{dist}^2(\boldsymbol{x},\boldsymbol{x}^\star) + \beta\|\nabla h(\boldsymbol{x})\|^2 \ge 2\sqrt{\alpha\beta}\operatorname{dist}(\boldsymbol{x},\boldsymbol{x}^\star)\|\nabla h(\boldsymbol{x})\|^2.$$

Lemma 1: [23], [31] If the function $h(\boldsymbol{x})$ restricted to a δ neighborhood of $\boldsymbol{x}^\star$ satisfies the (α,β,δ)-regularity condition, then as long as gradient descent starts from a point $\boldsymbol{x}_0\in B(\delta,\boldsymbol{x}^\star)$, the gradient descent update

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \nu\nabla h(\boldsymbol{x}_t)$$

with step size $0<\nu\le 2\beta$ obeys $\boldsymbol{x}_t\in B(\delta,\boldsymbol{x}^\star)$ and

$$\operatorname{dist}^2(\boldsymbol{x}_t,\boldsymbol{x}^\star) \le (1-2\nu\alpha)^t\operatorname{dist}^2(\boldsymbol{x}_0,\boldsymbol{x}^\star)$$

for all $t\ge 0$.

The proof is given in [23]. To keep the paper self-contained, we also provide the proof of Lemma 1 in Appendix B. We remark that the decreasing rate $1-2\nu\alpha\in[0,1)$ since we choose $\nu\le 2\beta$ and $\alpha\beta\le\frac{1}{4}$.

Now we establish the following revised robust strict saddle property for invariant functions by replacing the strong convexity condition in Definition 4 with the regularity condition.

Definition 9 (Revised Robust Strict Saddle Property for Invariant Functions): Given a twice differentiable $h(\boldsymbol{x}) : \mathbb{R}^n\to\mathbb{R}$ and a group $\mathcal{G}$, suppose $h(\boldsymbol{x})$ is invariant under the group action and the energy of $\boldsymbol{x}$ is also preserved under the group action, i.e., $h(a(\boldsymbol{x})) = h(\boldsymbol{x})$ and $\|a(\boldsymbol{x})\|_2 = \|\boldsymbol{x}\|_2$ for all $a\in\mathcal{G}$. Given $\alpha,\beta,\gamma,\epsilon,\delta$, $h(\boldsymbol{x})$ satisfies the $(\alpha,\beta,\gamma,\epsilon,\delta)$-robust strict saddle property if for any point $\boldsymbol{x}$ at least one of the following applies:

1) There exists a local minimum point $\boldsymbol{x}^\star$ such that $\operatorname{dist}(\boldsymbol{x},\boldsymbol{x}^\star)\le\delta$, and the function $h(\boldsymbol{x}')$ restricted to 2δ a neighborhood of $\boldsymbol{x}^\star$ (i.e., $\operatorname{dist}(\boldsymbol{x}',\boldsymbol{x}^\star)\le 2\delta$) satisfies the $(\alpha,\beta,2\delta)$-regularity condition defined in Definition 8;
2) $\lambda_{\min}\left(\nabla^2 h(\boldsymbol{x})\right)\le-\gamma$;
3) $\|\nabla h(\boldsymbol{x})\|\ge\epsilon$.

Compared with Definition 4, the revised robust strict saddle property requires the local descent condition instead of strict convexity in a small neighborhood around any local minimum point. With the convergence guarantee in Lemma 1, the convergence analysis of the stochastic gradient descent algorithm in [16] for the robust strict saddle functions can also be applied for the revised robust strict saddle functions defined in Definition 9 with the same convergence rate.[2] We omit the details here and refer the reader to [20] for more details on this. In the rest of the paper, the robust strict saddle property refers to the one in Definition 9.

III. Low-Rank Matrix Optimization With the Factorization Approach

In this section, we consider the minimization of general rank-constrained optimization problems of the form (1) using the factorization approach (2) (which we repeat as follows):

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r},\boldsymbol{V}\in\mathbb{R}^{m\times r}}{\operatorname{minimize}} h(\boldsymbol{U},\boldsymbol{V}) = f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}),$$

where the rank constraint in (1) is automatically satisfied by the factorization approach. With necessary assumptions on f in Section III-A, we provide geometric analysis of the factored problem in Section III-B. We then present a stylized application in matrix sensing in Section III-C.

A. Assumptions and Regularizer

Before presenting our main results, we lay out the necessary assumptions on the objective function $f(\boldsymbol{X})$. As is known, without any assumptions on the problem, even minimizing traditional quadratic objective functions is challenging. For this reason, we focus on problems satisfying the following two assumptions.

[2] As mentioned previously, a similar notion of a revised robust strict saddle property has also recently been utilized in [20].

Assumption 1: $f(\boldsymbol{X})$ has a critical point $\boldsymbol{X}^\star \in \mathbb{R}^{n\times m}$ which has rank r.

Assumption 2: $f(\boldsymbol{X})$ is $(2r,4r)$-restricted strongly convex and smooth, i.e., for any $n\times m$ matrices $\boldsymbol{X},\boldsymbol{D}$ with $\operatorname{rank}(\boldsymbol{X})\le 2r$ and $\operatorname{rank}(\boldsymbol{D})\le 4r$, the Hessian of $f(\boldsymbol{X})$ satisfies

$$a\left\|\boldsymbol{D}\right\|_F^2 \le [\nabla^2 f(\boldsymbol{X})](\boldsymbol{D},\boldsymbol{D}) \le b\left\|\boldsymbol{D}\right\|_F^2 \tag{6}$$

for some positive a and b.

Assumption 1 is equivalent to the existence of a rank r $\boldsymbol{X}^\star$ such that $\nabla f(\boldsymbol{X}^\star)=\boldsymbol{0}$, which is very mild and holds in many matrix inverse problems including matrix sensing [7], matrix completion [9] and 1-bit matrix completion [44], where the unknown matrix to be recovered is a critical point of f.

Assumption 2 is also utilized in [32, Conditions 5.3 and 5.4] and [40], where weighted low-rank matrix factorization and a set of matrix inverse problems are proved to satisfy the $(2r,4r)$-restricted strong convexity and smoothness condition (6). We discuss matrix sensing as a typical example satisfying this assumption in Section III-C.

Combining Assumption 1 and Assumption 2, we have that $\boldsymbol{X}^\star$ is the unique global minimum of (1).

Proposition 1: Suppose $f(\boldsymbol{X})$ satisfies the $(2r,4r)$-restricted strong convexity and smoothness condition (6) with positive a and b. Assume $\boldsymbol{X}^\star$ is a critical point of $f(\boldsymbol{X})$ with $\operatorname{rank}(\boldsymbol{X}^\star)=r$. Then $\boldsymbol{X}^\star$ is the global minimum of (1), i.e.,

$$f(\boldsymbol{X}^\star)\le f(\boldsymbol{X}),\ \forall\ \boldsymbol{X}\in\mathbb{R}^{n\times m}, \operatorname{rank}(\boldsymbol{X})\le r$$

and the equality holds only at $\boldsymbol{X}=\boldsymbol{X}^\star$.

The proof of Proposition 1 is given in Appendix C. We note that Proposition 1 guarantees that $\boldsymbol{X}^\star$ is the unique global minimum of (1) and it is expected that solving the factorized problem (9) also gives $\boldsymbol{X}^\star$. Proposition 1 differs from [40] in that it only requires $\boldsymbol{X}^\star$ as a critical point, while [40] needs $\boldsymbol{X}^\star$ as a global minimum of f.

Before presenting the main result, we note that if f satisfies (6) with positive a and b and we rescale f as $f'=\frac{2}{a+b}f$, then f' satisfies

$$\frac{2a}{a+b}\left\|\boldsymbol{D}\right\|_F^2 \le [\nabla^2 f'(\boldsymbol{X})](\boldsymbol{D},\boldsymbol{D}) \le \frac{2b}{a+b}\left\|\boldsymbol{D}\right\|_F^2.$$

It is clear that f and f' have the same optimization geometry (despite the scaling difference). Let $a'=\frac{2a}{a+b}=1-c$ and $b'=\frac{2a}{a+b}=1+c$ with $c=\frac{b-a}{a+b}$. We have $0<a'\le 1\le b'$ and $a'+b'=2$. Thus, throughout the paper and without the generality, we assume

$$a=1-c,\ b=1+c,\ c\in[0,1). \tag{7}$$

Now let $\boldsymbol{X}^\star=\boldsymbol{\Phi\Sigma\Psi}^{\mathrm{T}}=\sum_{i=1}^r\sigma_i\boldsymbol{\phi}_i\boldsymbol{\psi}_i^{\mathrm{T}}$ be a reduced SVD of $\boldsymbol{X}^\star$, where $\boldsymbol{\Sigma}$ is a diagonal matrix with $\sigma_1\ge\cdots\ge\sigma_r$ along its diagonal. Denote

$$\boldsymbol{U}^\star=\boldsymbol{\Phi\Sigma}^{1/2}\boldsymbol{R}, \boldsymbol{V}^\star=\boldsymbol{\Psi\Sigma}^{1/2}\boldsymbol{R} \tag{8}$$

for any $\boldsymbol{R}\in\mathcal{O}_r$. We first introduce the following ways to stack $\boldsymbol{U}$ and $\boldsymbol{V}$ together that are widely used through the paper:

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

Before moving on, we note that for any solution $(\boldsymbol{U},\boldsymbol{V})$ to (2), $(\boldsymbol{U}\boldsymbol{R}_1,\boldsymbol{V}\boldsymbol{R}_2)$ is also a solution to (2) for any $\boldsymbol{R}_1,\boldsymbol{R}_2\in\mathbb{R}^{r\times r}$ such that $\boldsymbol{U}\boldsymbol{R}_1\boldsymbol{R}_2^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}}=\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$. As an extreme example, $\boldsymbol{R}_1=c\mathbf{I}$ and $\boldsymbol{R}_2=\frac{1}{c}\mathbf{I}$ where c can be arbitrarily large. In order to address this ambiguity (i.e., to reduce the search space of $\boldsymbol{W}$ for (3)), we utilize the trick in [15], [31], [32], [40] by introducing a regularizer ρ and turn to solve the following problem

$$\operatorname*{minimize}_{\boldsymbol{U}\in\mathbb{R}^{n\times r},\boldsymbol{V}\in\mathbb{R}^{m\times r}} G(\boldsymbol{W}):=h(\boldsymbol{W})+\rho(\boldsymbol{W}), \tag{9}$$

where

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

We remark that $\boldsymbol{W}^\star$ is still a global minimizer of the factored problem (29) since both the first term and $\rho(\boldsymbol{W})$ achieve their global minimum at $\boldsymbol{W}^\star$. The regularizer $\rho(\boldsymbol{W})$ is applied to force the difference between the Gram matrices of $\boldsymbol{U}$ and $\boldsymbol{V}$ as small as possible. The global minimum of $\rho(\boldsymbol{W})$ is 0, which is achieved when $\boldsymbol{U}$ and $\boldsymbol{V}$ have the same Gram matrices, i.e., when $\boldsymbol{W}$ belongs to

$$\mathcal{E}:=\left\{\boldsymbol{W}=\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}:\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}-\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{0}\right\}. \tag{10}$$

Informally, we can view (9) as finding a point from $\mathcal{E}$ that also minimizes the first term in (9). This is rigorously established in the following result which reveals that any critical point $\boldsymbol{W}$ of $g(\boldsymbol{W})$ belongs to $\mathcal{E}$ (that is $\boldsymbol{U}$ and $\boldsymbol{V}$ are balanced factors of their product $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$) for any $\mu>0$.

Lemma 2 ([40, Theorem 3]): Suppose $G(\boldsymbol{W})$ is defined as in (9) with $\mu>0$. Then any critical point $\boldsymbol{W}$ of $G(\boldsymbol{W})$ belongs to $\mathcal{E}$, i.e.,

$$\nabla G(\boldsymbol{W})=\boldsymbol{0}\quad\Rightarrow\quad\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}. \tag{11}$$

For completeness, we include the proof of Lemma 2 in Appendix D.

B. Global Geometry for General Low-Rank Optimization

We now characterize the global optimization geometry of the factored problem (9). As explained in Section II that $G(\boldsymbol{W})$ is invariant under the matrices $\boldsymbol{R}\in\mathcal{O}_r$, we first recall the discussions in Section II about the revised robust strict saddle property for the invariant functions. To that end, we follow the notion of the distance between equivalent classes for invariant functions defined in (4) and define the distance between $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ as follows

$$\begin{aligned}\operatorname{dist}(\boldsymbol{W}_1,\boldsymbol{W}_2):&=\min_{\boldsymbol{R}_1\in\mathcal{O}_r,\boldsymbol{R}_2\in\mathcal{O}_r}\left\|\boldsymbol{W}_1\boldsymbol{R}_1-\boldsymbol{W}_2\boldsymbol{R}_2\right\|_F\\ &=\min_{\boldsymbol{R}\in\mathcal{O}_r}\left\|\boldsymbol{W}_1-\boldsymbol{W}_2\boldsymbol{R}\right\|_F.\end{aligned} \tag{12}$$

For convenience, we also denote the best rotation matrix $\boldsymbol{R}$ so that $\left\|\boldsymbol{W}_1-\boldsymbol{W}_2\boldsymbol{R}\right\|_F$ achieves its minimum by $\boldsymbol{R}(\boldsymbol{W}_1,\boldsymbol{W}_2)$, i.e.,

$$\boldsymbol{R}(\boldsymbol{W}_1,\boldsymbol{W}_2):=\arg\min_{\boldsymbol{R}'\in\mathcal{O}_r}\left\|\boldsymbol{W}_1-\boldsymbol{W}_2\boldsymbol{R}'\right\|_F, \tag{13}$$

which is also known as the orthogonal Procrustes problem [45]. The solution to the above minimization problem is characterized by the following lemma.

Lemma 3 ([45]): Let $\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_1 = \boldsymbol{LSP}^{\mathrm{T}}$ be an SVD of $\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{W}_1$. An optimal solution for the orthogonal Procrustes problem (13) is given by

$$\boldsymbol{R}(\boldsymbol{W}_1, \boldsymbol{W}_2) = \boldsymbol{LP}^{\mathrm{T}}.$$

Moreover, we have

$$\begin{aligned}\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{W}_2\boldsymbol{R}(\boldsymbol{W}_1, \boldsymbol{W}_2) &= (\boldsymbol{W}_2\boldsymbol{R}(\boldsymbol{W}_1, \boldsymbol{W}_2))^{\mathrm{T}}\boldsymbol{W}_1\\ &= \boldsymbol{PSP}^{\mathrm{T}} \succeq \boldsymbol{0}.\end{aligned}$$

To ease the notation, we drop $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ in $\boldsymbol{R}(\boldsymbol{W}_1, \boldsymbol{W}_2)$ and rewrite $\boldsymbol{R}$ instead of $\boldsymbol{R}(\boldsymbol{W}_1, \boldsymbol{W}_2)$ when they ($\boldsymbol{W}_1$ and $\boldsymbol{W}_2$) are clear from the context. Now we are well equipped to present the robust strict saddle property for $G(\boldsymbol{W})$ in the following result.

Theorem 1: Define the following regions

$$\begin{aligned}\mathcal{R}_1 &:= \left\{\boldsymbol{W} : \operatorname{dist}(\boldsymbol{W}, \boldsymbol{W}^\star) \le \sigma_r^{1/2}(\boldsymbol{X}^\star)\right\},\\ \mathcal{R}_2 &:= \Bigg\{\boldsymbol{W} : \sigma_r(\boldsymbol{W}) \le \sqrt{\frac{1}{2}}\sigma_r^{1/2}(\boldsymbol{X}^\star),\\ &\qquad \|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F \le \frac{20}{19}\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F\Bigg\},\\ \mathcal{R}_3' &:= \Bigg\{\boldsymbol{W} : \operatorname{dist}(\boldsymbol{W}, \boldsymbol{W}^\star) > \sigma_r^{1/2}(\boldsymbol{X}^\star), \|\boldsymbol{W}\| \le \frac{20}{19}\|\boldsymbol{W}^\star\|,\\ &\qquad \sigma_r(\boldsymbol{W}) > \sqrt{\frac{1}{2}}\sigma_r^{1/2}(\boldsymbol{X}^\star), \|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F \le \frac{20}{19}\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F\Bigg\},\\ \mathcal{R}_3'' &:= \Bigg\{\boldsymbol{W} : \|\boldsymbol{W}\| > \frac{20}{19}\|\boldsymbol{W}^\star\| = \frac{20}{19}\sqrt{2}\|\boldsymbol{X}^\star\|^{1/2},\\ &\qquad \|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F \le \frac{10}{9}\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F\Bigg\},\\ \mathcal{R}_3''' &:= \left\{\boldsymbol{W} : \|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F > \frac{10}{9}\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F = \frac{20}{9}\|\boldsymbol{X}^\star\|_F\right\}.\end{aligned}$$

Let $G(\boldsymbol{W})$ be defined as in (9) with $\mu = \frac{1}{2}$. Suppose $f(\boldsymbol{X})$ has a critical point $\boldsymbol{X}^\star \in \mathbb{R}^{n\times m}$ of rank r and satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (6) with positive constants $a = 1 - c, b = 1 + c$ and

$$c \le \frac{1}{100}\frac{\sigma_r^{3/2}(\boldsymbol{X}^\star)}{\|\boldsymbol{X}^\star\|_F\|\boldsymbol{X}^\star\|^{1/2}}. \tag{14}$$

Then $G(\boldsymbol{W})$ has the following robust strict saddle property:

1) For any $\boldsymbol{W} \in \mathcal{R}_1$, $G(\boldsymbol{W})$ satisfies the local regularity condition:

$$\begin{aligned}&\langle\nabla G(\boldsymbol{W}), \boldsymbol{W} - \boldsymbol{W}^\star\rangle\\ &\ge \frac{1}{16}\sigma_r(\boldsymbol{X}^\star)\operatorname{dist}^2(\boldsymbol{W}, \boldsymbol{W}^\star) + \frac{1}{260}\frac{1}{\|\boldsymbol{X}^\star\|}\|\nabla G(\boldsymbol{W})\|_F^2.\end{aligned} \tag{15}$$

where $\operatorname{dist}(\boldsymbol{W}, \boldsymbol{W}^\star)$ and $\boldsymbol{R}$ are defined in (12) and (13), respectively.

2) For any $\boldsymbol{W} \in \mathcal{R}_2$, $G(\boldsymbol{W})$ has a directional negative curvature, i.e.,

$$\lambda_{\min}\left(\nabla^2 G(\boldsymbol{W})\right) \le -\frac{1}{6}\sigma_r(\boldsymbol{X}^\star). \tag{16}$$

3) For any $\boldsymbol{W} \in \mathcal{R}_3 = \mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$, $G(\boldsymbol{W})$ has large gradient:

$$\|\nabla G(\boldsymbol{W})\|_F \ge \frac{1}{27}\sigma_r^{3/2}(\boldsymbol{X}^\star), \quad \forall\, \boldsymbol{W} \in \mathcal{R}_3'; \tag{17}$$

$$\|\nabla G(\boldsymbol{W})\|_F \ge \frac{1}{50}\|\boldsymbol{W}\|^3, \quad \forall\, \boldsymbol{W} \in \mathcal{R}_3''; \tag{18}$$

$$\|\nabla G(\boldsymbol{W})\|_F \ge \frac{1}{45}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\right\|_F^{3/2}, \quad \forall\, \boldsymbol{W} \in \mathcal{R}_3'''. \tag{19}$$

The proof of this result is given in Appendix L. The main proof strategy is to utilize Assumption 1 and Assumption 2 about the function f to control the deviation between the gradient (and the Hessian) of the general low-rank optimization (9) and the counterpart of the matrix factorization problem so that the landscape of the general low-rank optimization (9) has a similar geometry property. To that end, in Appendix A, we provide a comprehensive geometric analysis for the matrix factorization problem (3). The reason for choosing $\mu = \frac{1}{2}$ is also discussed in Appendix A-F. We note that the results in Appendix A are also of independent interest, as we show that the objective function in (3) obeys the strict saddle property and has no spurious local minima not only for exact-parameterization ($r = \operatorname{rank}(\boldsymbol{X}^\star)$), but also for over-parameterization ($r > \operatorname{rank}(\boldsymbol{X}^\star)$) and under-parameterization ($r < \operatorname{rank}(\boldsymbol{X}^\star)$). Several remarks follow.

Remark 1: Note that

$$\begin{aligned}\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3' \supseteq \Bigg\{\boldsymbol{W} : \|\boldsymbol{W}\| &\le \frac{20}{19}\|\boldsymbol{W}^\star\|_F,\\ \|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F &\le \frac{10}{9}\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F\Bigg\},\end{aligned}$$

which further implies

$$\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3' \cup \mathcal{R}_3'' \supseteq \{\boldsymbol{W} : \|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F \le \frac{10}{9}\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F\}.$$

Thus, we conclude that $\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3''' = \mathbb{R}^{(n+m)\times r}$. Now the convergence analysis of the stochastic gradient descent algorithm in [16], [20] for the robust strict saddle functions also holds for $G(\boldsymbol{W})$.

Remark 2: Theorem 1 states that the objective function for the general low-rank optimization (9) also satisfies the robust strict saddle property when (14) holds. The requirement for c in (14) can be weakened to ensure the properties of $g(\boldsymbol{W})$ are preserved for $G(\boldsymbol{W})$ in some regions. For example, the local regularity condition (15) holds when

$$c \le \frac{1}{50}$$

which is independent of $\boldsymbol{X}^\star$. With the analysis of the global geometric structure in $G(\boldsymbol{W})$, Theorem 1 ensures that many local search algorithms can converge to $\boldsymbol{X}^\star$ (which is the the global minimum of (1) as guaranteed by Proposition 1) with random initialization. In particular, stochastic gradient descent when applied to the matrix sensing problem (22) is guaranteed to find the global minimum $\boldsymbol{X}^\star$ in polynomial time.

Remark 3: Local (rather than global) geometry results for the general low-rank optimization (9) are also covered in [40], which only characterizes the geometry at all the critical points. Instead, Theorem 1 characterizes the global geometry for general low-rank optimization (9). Because the analysis is different, the proof strategy for Theorem 1 is also very different than that of [40]. Since [40] only considers local geometry, the result in [40] requires $c \le 0.2$, which is slightly less restrictive than the one in (14).

Remark 4: To explain the necessity of the requirement on the constants a and b in (14), we utilize the symmetric weighted PCA problem (so that we can visualize the landscape of the factored problem in Figure 1) as an example where the objective function is

$$f(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{\Omega} \odot (\boldsymbol{X} - \boldsymbol{X}^\star)\|_F^2, \tag{20}$$

where $\boldsymbol{\Omega} \in \mathbb{R}^{n\times n}$ contains positive entries. The Hessian quadratic form for $f(\boldsymbol{X})$ is given by $[\nabla^2 f(\boldsymbol{X})](\boldsymbol{D}, \boldsymbol{D}) = \|\boldsymbol{\Omega} \odot \boldsymbol{D}\|_F^2$ for any $\boldsymbol{D} \in \mathbb{R}^{n\times n}$. Thus, we have

$$\min_{ij} |\boldsymbol{\Omega}[i,j]|^2 \le \frac{[\nabla^2 f(\boldsymbol{X})](\boldsymbol{D},\boldsymbol{D})}{\|\boldsymbol{D}\|_F^2} \le \max_{ij} |\boldsymbol{\Omega}[i,j]|^2.$$

Comparing with (6), we see that f satisfies the restricted strong convexity and smoothness conditions with the constants $a = \min_{ij} |\boldsymbol{\Omega}[i,j]|^2$ and $b = \max_{ij} |\boldsymbol{\Omega}[i,j]|^2$. In this case, we also note that if each entry W_{ij} is nonzero (i.e., $\min_{ij} |\boldsymbol{\Omega}[i,j]|^2 > 0$), the function $f(\boldsymbol{X})$ is strongly convex, rather than only restrictively strongly convex, implying that (20) has a unique optimal solution $\boldsymbol{X}^\star$. By applying the factorization approach, we get the factored objective function

$$h(\boldsymbol{U}) = \frac{1}{2}\|\boldsymbol{\Omega} \odot (\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} - \boldsymbol{X}^\star)\|_F^2. \tag{21}$$

To illustrate the necessity of the requirement on the constants a and b as in (14) so that the factored problem (21) has no spurious local minima and obeys the robust strict saddle property, we set $\boldsymbol{X}^\star = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$ which is a rank-1 matrix and can be factorized as $\boldsymbol{X}^\star = \boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}}$ with $\boldsymbol{U}^\star = \begin{bmatrix}1\\ 1\end{bmatrix}$. We then plot the landscapes of the factored objective function $h(\boldsymbol{U})$ with $\boldsymbol{\Omega} = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$ and $\begin{bmatrix}8 & 1\\ 1 & 8\end{bmatrix}$ in Figure 1. We observe from Figure 1 that as long as the elements in $\boldsymbol{\Omega}$ have a small dynamic range (which corresponds to a small b/a), $h(\boldsymbol{U})$ has no spurious local minima, but if the elements in $\boldsymbol{\Omega}$ have a large dynamic range (which corresponds to a large b/a), spurious local minima can appear in $h(\boldsymbol{U})$.

Remark 5: The global geometry of low-rank matrix recovery but with analysis customized to linear measurements and quadratic loss functions is also covered in [36], [42]. Since Theorem 1 only requires the $(2r, 4r)$-restricted strong convexity and smoothness property (6), aside from low-rank matrix recovery [46], it can also be applied to many other low-rank matrix optimization problems [47] which do not necessarily involve quadratic loss functions. Typical examples include 1-bit matrix completion [44], [48] and Poisson principal component analysis (PCA) [49]. We refer to [40] for more discussion on this issue. In next section, we consider a stylized application of Theorem 1 in matrix sensing and compare it with the result in [42].

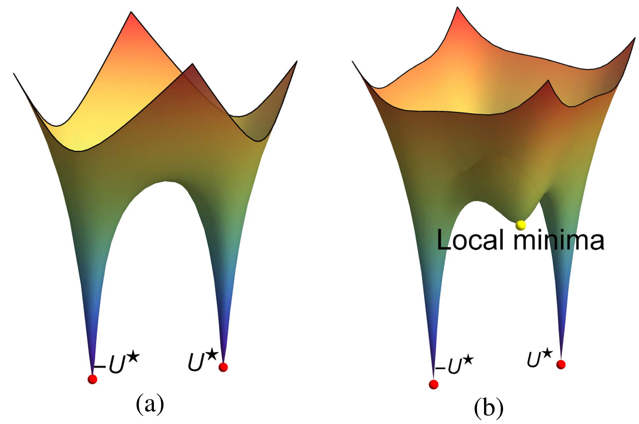

Fig. 1. Landscapes of $h(\boldsymbol{U})$ in (21) with $\boldsymbol{X}^\star = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$ and (a) $\boldsymbol{\Omega} = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$; (b) $\boldsymbol{\Omega} = \begin{bmatrix}8 & 1\\ 1 & 8\end{bmatrix}$.

C. Stylized Application: Matrix Sensing

In this section, we extend the previous geometric analysis to the matrix sensing problem

$$\operatorname*{minimize}_{\boldsymbol{U}\in\mathbb{R}^{n\times r}, \boldsymbol{V}\in\mathbb{R}^{m\times r}} G(\boldsymbol{W}) := \frac{1}{2}\left\|\mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star\right)\right\|_2^2 + \rho(\boldsymbol{W}), \tag{22}$$

where $\mathcal{A} : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ is a known linear measurement operator and $\boldsymbol{X}^\star$ is the unknown rank r matrix to be recovered. In this case, we have

$$f(\boldsymbol{X}) = \frac{1}{2}\|\mathcal{A}(\boldsymbol{X} - \boldsymbol{X}^\star)\|_2^2.$$

The derivative of $f(\boldsymbol{X})$ at $\boldsymbol{X}^\star$ is

$$\nabla f(\boldsymbol{X}^\star) = \mathcal{A}^*\mathcal{A}(\boldsymbol{X}^\star - \boldsymbol{X}^\star) = \boldsymbol{0},$$

which implies that $f(\boldsymbol{X})$ satisfies Assumption 1. The Hessian quadratic form $\nabla^2 f(\boldsymbol{X})[\boldsymbol{D}, \boldsymbol{D}]$ for any $n \times m$ matrices $\boldsymbol{X}$ and $\boldsymbol{D}$ is given by

$$\nabla^2 f(\boldsymbol{X})[\boldsymbol{D},\boldsymbol{D}] = \|\mathcal{A}(\boldsymbol{D})\|^2.$$

The following matrix Restricted Isometry Property (RIP) serves as a way to link the low-rank matrix factorization problem (29) with the matrix sensing problem (22) and certifies $f(\boldsymbol{X})$ satisfying Assumption 2.

Definition 10 (Restricted Isometry Property (RIP) [7], [50]): The map $\mathcal{A} : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ satisfies the r-RIP with constant δ_r if [3]

$$(1-\delta_r)\|\boldsymbol{D}\|_F^2 \le \|\mathcal{A}(\boldsymbol{D})\|^2 \le (1+\delta_r)\|\boldsymbol{D}\|_F^2 \tag{23}$$

holds for any $n \times m$ matrix $\boldsymbol{D}$ with $\operatorname{rank}(\boldsymbol{D}) \le r$.

If $\mathcal{A}$ satisfies the $4r$-restricted isometry property with constant δ_{4r}, then $f(\boldsymbol{X})$ satisfies the $(2r, 4r)$-restricted strong

[3]By abuse of notation, we adopt the conventional notation δ_r for the RIP constant. The subscript r can be used to distinguish the RIP constant δ_r from δ which is used as a small constant in Section II.

convexity and smoothness condition (6) with constants $a = 1-\delta_{4r}$ and $b = 1-\delta_{4r}$ since

$$(1-\delta_{4r})\|\boldsymbol{D}\|_F^2 \le \nabla^2 f(\boldsymbol{X})[\boldsymbol{D},\boldsymbol{D}] = \|\mathcal{A}(\boldsymbol{D})\|^2 \le (1+\delta_{4r})\|\boldsymbol{D}\|_F^2 \tag{24}$$

for any rank-$4r$ matrix $\boldsymbol{D}$. Comparing (24) with (6), we note that the RIP is stronger than the restricted strong convexity and smoothness property (6) as the RIP gives that (24) holds for all $n \times m$ matrices $\boldsymbol{X}$, while Assumption 2 only requires that (6) holds for all rank-$2r$ matrices.

Now, applying Theorem 1, we obtain a similar geometric guarantee to Theorem 1 for the matrix sensing problem (22) when $\mathcal{A}$ satisfies the RIP.

Corollary 1: Let $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3', \mathcal{R}_3'', \mathcal{R}_3'''$ be the regions as defined in Theorem 5. Let $G(\boldsymbol{W})$ be defined as in (22) with $\mu = \frac{1}{2}$ and $\mathcal{A}$ satisfying the $4r$-RIP with

$$\delta_{4r} \le \frac{1}{100}\frac{\sigma_r^{3/2}(\boldsymbol{X}^\star)}{\|\boldsymbol{X}^\star\|_F\|\boldsymbol{X}^\star\|^{1/2}}. \tag{25}$$

Then $G(\boldsymbol{W})$ has the following robust strict saddle property:

1) For any $\boldsymbol{W} \in \mathcal{R}_1$, $G(\boldsymbol{W})$ satisfies the local regularity condition:

$$\langle \nabla G(\boldsymbol{W}), \boldsymbol{W}-\boldsymbol{W}^\star\rangle \ge \frac{1}{16}\sigma_r(\boldsymbol{X}^\star)\,\mathrm{dist}^2(\boldsymbol{W},\boldsymbol{W}^\star) + \frac{1}{260}\frac{1}{\|\boldsymbol{X}^\star\|}\|\nabla G(\boldsymbol{W})\|_F^2. \tag{26}$$

where $\mathrm{dist}(\boldsymbol{W},\boldsymbol{W}^\star)$ and $\boldsymbol{R}$ are defined in (12) and (13), respectively.

2) For any $\boldsymbol{W} \in \mathcal{R}_2$, $G(\boldsymbol{W})$ has a directional negative curvature, i.e.,

$$\lambda_{\min}\left(\nabla^2 G(\boldsymbol{W})\right) \le -\frac{1}{6}\sigma_r(\boldsymbol{X}^\star).$$

3) For any $\boldsymbol{W} \in \mathcal{R}_3 = \mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$, $G(\boldsymbol{W})$ has large gradient:

$$\|\nabla G(\boldsymbol{W})\|_F \ge \frac{1}{27}\sigma_r^{3/2}(\boldsymbol{X}^\star), \quad \forall\, \boldsymbol{W} \in \mathcal{R}_3';$$
$$\|\nabla G(\boldsymbol{W})\|_F \ge \frac{1}{50}\|\boldsymbol{W}\|^3, \quad \forall\, \boldsymbol{W} \in \mathcal{R}_3'';$$
$$\|\nabla G(\boldsymbol{W})\|_F \ge \frac{1}{45}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\right\|_F^{3/2}, \quad \forall\, \boldsymbol{W} \in \mathcal{R}_3'''.$$

Remark 6: Similar to (14), the requirement for δ_{4r} in (25) can be weakened to ensure the properties of $g(\boldsymbol{W})$ are preserved for $G(\boldsymbol{W})$ in some regions. For example, the local regularity condition (26) holds when

$$\delta_{4r} \le \frac{1}{50}$$

which is independent of $\boldsymbol{X}^\star$. Note that Tu *et al.* [31, Section 5.4, (5.15)] provided a similar regularity condition. However, the result there requires $\delta_{6r} \le \frac{1}{25}$ and $\mathrm{dist}(\boldsymbol{W},\boldsymbol{W}^\star) \le \frac{1}{2\sqrt{2}}\sigma_r(\boldsymbol{X}^\star)$ which defines a smaller region than $\mathcal{R}_1$. Based on this local regularity condition, Tu *et al.* [31] showed that gradient descent with a good initialization (which is close enough to $\boldsymbol{W}^\star$) converges to the unknown matrix $\boldsymbol{W}^\star$ (and hence $\boldsymbol{X}^\star$). With the analysis of the global geometric structure in $G(\boldsymbol{W})$, Theorem 1 ensures that many local search algorithms can find the unknown matrix $\boldsymbol{X}^\star$ in polynomial time.

Remark 7: A Gaussian $\mathcal{A}$ will have the RIP with high probability when the number of measurements p is comparable to the number of degrees of freedom in an $n \times m$ matrix with rank r. By Gaussian $\mathcal{A}$ we mean the ℓ-th element in $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X})$, y_ℓ, is given by

$$y_\ell = \langle \boldsymbol{X}, \boldsymbol{A}_\ell\rangle = \sum_{i=1}^n\sum_{j=1}^m \boldsymbol{X}[i,j]\boldsymbol{A}_\ell[i,j],$$

where the entries of each $n \times m$ matrix $\boldsymbol{A}_\ell$ are independent and identically distributed normal random variables with zero mean and variance $\frac{1}{p}$. Specifically, a Gaussian $\mathcal{A}$ satisfies (23) with high probability when [5], [7], [46]

$$p \gtrsim r(n+m)\frac{1}{\delta_r^2}.$$

Now utilizing the inequality $\|\boldsymbol{X}^\star\|_F \le \sqrt{r}\|\boldsymbol{X}^\star\|$ for (14), we conclude that in the case of Gaussian measurements, the robust strict saddle property is preserved for the matrix sensing problem with high probability when the number of measurements exceeds a constant times $(n+m)r^2\kappa(\boldsymbol{X}^\star)^3$ where $\kappa(\boldsymbol{X}^\star) = \frac{\sigma_1(\boldsymbol{X}^\star)}{\sigma_r(\boldsymbol{X}^\star)}$. This further implies that, when applying the stochastic gradient descent algorithm to the matrix sensing problem (22) with Gaussian measurements, we are guaranteed to find the unknown matrix $\boldsymbol{X}^\star$ in polynomial time with high probability when

$$p \gtrsim (n+m)r^2\kappa(\boldsymbol{X}^\star)^3. \tag{27}$$

When $\boldsymbol{X}^\star$ is an $n \times n$ PSD matrix, Li *et al.* [42] showed that the corresponding matrix sensing problem with Gaussian measurements has similar global geometry to the low-rank PSD matrix factorization problem when the number of measurements

$$p \gtrsim nr^2\frac{\sigma_1^4(\boldsymbol{X}^\star)}{\sigma_r^2(\boldsymbol{X}^\star)}. \tag{28}$$

Comparing (27) with (28), we find both results for the number of measurements needed depend similarly on the rank r, but slightly differently on the spectrum of $\boldsymbol{X}^\star$. We finally remark that the sampling complexity in (27) is $O((n+m)r^2)$, which is slightly larger than the information theoretically optimal bound $O((n+m)r)$ for matrix sensing. This is because Theorem 1 is a direct consequence of Theorem 1 in which we directly characterize the landscapes of the objective functions in the whole space by combining the results for matrix factorization in Appendix A and the restricted strong convexity and smoothness condition. We believe this mismatch is an artifact of our proof strategy and could be mitigated by a different approach, like utilizing the properties of quadratic loss functions [36]. If one desires only to characterize the geometry for critical points, then $O((n+m)r)$ measurements are enough to ensure the strict saddle property and lack of spurious local minima for matrix sensing [15], [40]. We finally note that for matrix completion where the RIP is not satisfied, [36] proves the robust strict saddle property for the factorization approach by utilizing an additional regularizer which promotes incoherence of $\boldsymbol{W}$.

Appendix A
The Optimization Geometry of Low-Rank Matrix Factorization

In this appendix, we consider the low-rank matrix factorization problem

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r},\boldsymbol{V}\in\mathbb{R}^{m\times r}}{\text{minimize}}\ g(\boldsymbol{W}) := \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^{\star}\right\|_F^2 + \rho(\boldsymbol{W}) \quad (29)$$

where $\rho(\boldsymbol{W})$ is the regularizer used in (9) and repeated here:

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

We provide a comprehensive geometric analysis for the matrix factorization problem (29). In particular, we show that the objective function in (29) obeys the strict saddle property and has no spurious local minima not only for exact-parameterization ($r = \operatorname{rank}(\boldsymbol{X}^{\star})$), but also for over-parameterization ($r > \operatorname{rank}(\boldsymbol{X}^{\star})$) and under-parameterization ($r < \operatorname{rank}(\boldsymbol{X}^{\star})$). For the exact-parameterization case, we further show that the objective function satisfies the robust strict saddle property, ensuring global convergence of many local search algorithms in polynomial time. As we believe these results are also of independent interest and to make it easy to follow, we only present the main results in this appendix and defer the proofs to other appendices.

A. Relationship to PSD Low-Rank Matrix Factorization

Similar to (8), let $\boldsymbol{X}^{\star} = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{T}} = \sum_{i=1}^{r}\sigma_i\boldsymbol{\phi}_i\boldsymbol{\psi}_i^{\mathrm{T}}$ be a reduced SVD of $\boldsymbol{X}^{\star}$, where $\boldsymbol{\Sigma}$ is a diagonal matrix with $\sigma_1 \geq \cdots \geq \sigma_r$ along its diagonal, and denote $\boldsymbol{U}^{\star} = \boldsymbol{\Phi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{R}, \boldsymbol{V}^{\star} = \boldsymbol{\Psi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{R}$ for any $\boldsymbol{R}\in\mathcal{O}_r$. The following result to some degree characterizes the relationship between the nonsymmetric low-rank matrix factorization problem (29) and the following PSD low-rank matrix factorization problem [42]:

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r}}{\text{minimize}}\left\|\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} - \boldsymbol{M}\right\|_F^2, \quad (30)$$

where $\boldsymbol{M}\in\mathbb{R}^{n\times n}$ is a rank-r PSD matrix.

Lemma 4: Suppose $g(\boldsymbol{W})$ is defined as in (29) with $\mu > 0$. Then we have

$$g(\boldsymbol{W}) \geq \min\{\frac{\mu}{4}, \frac{1}{8}\}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^{\star}\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2.$$

In particular, if we choose $\mu = \frac{1}{2}$, then we have

$$g(\boldsymbol{W}) = \frac{1}{8}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^{\star}\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2 + \frac{1}{4}\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}^{\star} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}^{\star}\right\|_F^2.$$

The proof of Lemma 4 is given in Appendix E. Informally, Lemma 4 indicates that minimizing $g(\boldsymbol{W})$ also results in minimizing $\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^{\star}\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2$ (which is the same form as the objective function in (30)) and hence the distance between $\boldsymbol{W}$ and $\boldsymbol{W}^{\star}$ (though $\boldsymbol{W}^{\star}$ is unavailable at priori). The global geometry for the PSD low-rank matrix factorization problem (30) is recently analyzed by Li et al. in [42].

B. Characterization of Critical Points

We first provide the gradient and Hessian expression for $g(\boldsymbol{W})$. The gradient of $g(\boldsymbol{W})$ is given by

$$\nabla_{\boldsymbol{U}} g(\boldsymbol{U},\boldsymbol{V}) = (\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^{\star})\boldsymbol{V} + \mu\boldsymbol{U}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}),$$
$$\nabla_{\boldsymbol{V}} g(\boldsymbol{U},\boldsymbol{V}) = (\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^{\star})^{\mathrm{T}}\boldsymbol{U} - \mu\boldsymbol{V}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}),$$

which can be rewritten as

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

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

$$\begin{aligned}&[\nabla^2 g(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta})\\ &= \left\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2 + 2\left\langle\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^{\star}, \boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\rangle\\ &\quad + [\nabla^2\rho(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta}),\end{aligned} \quad (31)$$

where

$$\begin{aligned}&[\nabla^2\rho(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta})\\ &= \mu\left\langle\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W}, \widehat{\boldsymbol{\Delta}}^{\mathrm{T}}\boldsymbol{\Delta}\right\rangle + \mu\left\langle\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Delta}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{W}^{\mathrm{T}}\right\rangle\\ &\quad + \mu\left\langle\widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\rangle.\end{aligned} \quad (32)$$

By Lemma 2, we can simplify the equations for critical points as follows

$$\nabla_{\boldsymbol{U}}\rho(\boldsymbol{U},\boldsymbol{V}) = \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{X}^{\star}\boldsymbol{V} = \boldsymbol{0}, \quad (33)$$
$$\nabla_{\boldsymbol{V}}\rho(\boldsymbol{U},\boldsymbol{V}) = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} - \boldsymbol{X}^{\star\mathrm{T}}\boldsymbol{U} = \boldsymbol{0}. \quad (34)$$

Now suppose $\boldsymbol{W}$ is a critical point of $g(\boldsymbol{W})$. We can apply the Gram-Schmidt process to orthonormalize the columns of $\boldsymbol{U}$ such that $\widetilde{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{R}$, where $\boldsymbol{R}\in\mathcal{O}_r = \left\{\boldsymbol{R}\in\mathbb{R}^{r\times r}, \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \mathbf{I}\right\}$ and $\widetilde{\boldsymbol{U}}$ is orthogonal.[4] Also let $\widetilde{\boldsymbol{V}} = \boldsymbol{V}\boldsymbol{R}$. Since $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}$, we have $\widetilde{\boldsymbol{U}}^{\mathrm{T}}\widetilde{\boldsymbol{U}} = \widetilde{\boldsymbol{V}}^{\mathrm{T}}\widetilde{\boldsymbol{V}}$. Thus $\widetilde{\boldsymbol{V}}$ is also orthogonal. Noting that $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{V}}^{\mathrm{T}}$, we conclude that $g(\boldsymbol{W}) = g(\widetilde{\boldsymbol{W}})$ and $\widetilde{\boldsymbol{W}}$ is also a critical point of $g(\boldsymbol{W})$ since $\nabla_{\widetilde{\boldsymbol{U}}} g(\widetilde{\boldsymbol{W}}) = \nabla_{\boldsymbol{U}} g(\boldsymbol{W})\boldsymbol{R} = \boldsymbol{0}$ and $\nabla_{\widetilde{\boldsymbol{V}}} g(\widetilde{\boldsymbol{W}}) = \nabla_{\boldsymbol{V}} g(\boldsymbol{W})\boldsymbol{R} = \boldsymbol{0}$. Also for any $\boldsymbol{\Delta}\in\mathbb{R}^{(n+m)\times r}$, we have $[\nabla^2 g(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta}) = [\nabla^2 g(\widetilde{\boldsymbol{W}})](\boldsymbol{\Delta}\boldsymbol{R},\boldsymbol{\Delta}\boldsymbol{R})$, indicating that $\boldsymbol{W}$ and $\widetilde{\boldsymbol{W}}$ have the same Hessian information. Thus, without loss of generality, we assume $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal (including the possibility that they have zero columns). With this, we use $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ to denote the i-th columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. It follows from $\nabla g(\boldsymbol{W}) = \boldsymbol{0}$ that

$$\|\boldsymbol{u}_i\|^2\boldsymbol{u}_i = \boldsymbol{X}^{\star}\boldsymbol{v}_i,$$
$$\|\boldsymbol{v}_i\|^2\boldsymbol{v}_i = \boldsymbol{X}^{\star\mathrm{T}}\boldsymbol{u}_i,$$

[4]As defined in Section I-C, by orthogonal we mean that $\langle\widetilde{\boldsymbol{U}}[:,i], \widetilde{\boldsymbol{U}}[:,j]\rangle = 0$ for all $i\neq j$. The columns of $\widetilde{\boldsymbol{U}}$ are not required to be normalized, and could even be zero. Also, another way to find $\boldsymbol{R}$ is via the SVD. Let $\boldsymbol{U} = \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{R}^{\mathrm{T}}$ be a reduced SVD of $\boldsymbol{U}$, where $\boldsymbol{L}$ is an $n\times r$ orthonormal matrix, $\boldsymbol{\Sigma}$ is an $r\times r$ diagonal matrix with non-negative diagonals, and $\boldsymbol{R}\in\mathcal{O}_r$. Then $\widetilde{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{R} = \boldsymbol{L}\boldsymbol{\Sigma}$ is orthogonal, with possible zero columns.

which indicates that

$$(\boldsymbol{u}_i, \boldsymbol{v}_i) \in \left\{(\sqrt{\lambda_1}\boldsymbol{p}_1, \sqrt{\lambda_1}\boldsymbol{q}_1), \dots, (\sqrt{\lambda_r}\boldsymbol{p}_r, \sqrt{\lambda_r}\boldsymbol{q}_r), (\mathbf{0}, \mathbf{0})\right\}.$$

Now we identify all the critical points of $g(\boldsymbol{W})$ in the following lemma, which is formally proved with an algebraic approach in Appendix F.

Lemma 5: Let $\boldsymbol{X}^\star = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{T}} = \sum_{i=1}^r \sigma_i \boldsymbol{\phi}_i \boldsymbol{\psi}_i^{\mathrm{T}}$ be a reduced SVD of $\boldsymbol{X}^\star$ and $g(\boldsymbol{W})$ be defined as in (29) with $\mu > 0$. Any $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$ is a critical point of $g(\boldsymbol{W})$ if and only if $\boldsymbol{W} \in \mathcal{C}$ with

$$\mathcal{C} := \left\{\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} : \boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{R} \in \mathcal{O}_r, \right. \\ \left. \boldsymbol{\Lambda} \text{ is diagonal}, \boldsymbol{\Lambda} \geq \mathbf{0}, (\boldsymbol{\Sigma} - \boldsymbol{\Lambda})\boldsymbol{\Sigma} = \mathbf{0}\right\}. \quad (35)$$

Intuitively, (35) means that a critical point $\boldsymbol{W}$ of $g(\boldsymbol{W})$ is one such that $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ is a rank-ℓ approximation to $\boldsymbol{X}^\star$ with $\ell \leq r$ and $\boldsymbol{U}$ and $\boldsymbol{V}$ are equal factors of this rank-ℓ approximation. Let $\lambda_1, \lambda_2, \dots, \lambda_r$ denote the diagonals of $\boldsymbol{\Lambda}$. Unlike $\boldsymbol{\Sigma}$, we note that these diagonals $\lambda_1, \lambda_2, \dots, \lambda_r$ are not necessarily placed in decreasing or increasing order. Actually, this equation $(\boldsymbol{\Sigma} - \boldsymbol{\Lambda})\boldsymbol{\Sigma} = \mathbf{0}$ is equivalent to

$$\lambda_i \in \{\sigma_i, 0\}$$

for all $i \in \{1, 2, \dots, r\}$. Further, we introduce the set of optimal solutions:

$$\mathcal{X} := \left\{\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} : \boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{R}, \boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{R}, \boldsymbol{R} \in \mathcal{O}_r\right\}. \quad (36)$$

It is clear that the set $\mathcal{X}$ containing all the optimal solutions, the set $\mathcal{C}$ containing all the critical points and the set $\mathcal{E}$ containing all the points with balanced factors have the nesting relationship: $\mathcal{X} \subset \mathcal{C} \subset \mathcal{E}$. Before moving to the next section, we provide one more result regarding $\boldsymbol{W} \in \mathcal{E}$. The proof of the following result is given in Appendix G.

Lemma 6: For any $\boldsymbol{\Delta} = \begin{bmatrix}\boldsymbol{\Delta}_{\boldsymbol{U}}\\ \boldsymbol{\Delta}_{\boldsymbol{V}}\end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$ and $\boldsymbol{W} \in \mathcal{E}$ where $\mathcal{E}$ is defined in (10), we have

$$\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{U}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{\Delta}_{\boldsymbol{V}}\boldsymbol{V}^{\mathrm{T}}\|_F^2 = \|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{\Delta}_{\boldsymbol{V}}\boldsymbol{U}^{\mathrm{T}}\|_F^2, \quad (37)$$

and

$$\nabla^2 \rho(\boldsymbol{W}) \succeq \mathbf{0}. \quad (38)$$

C. Strict Saddle Property

Lemma 6 implies that the Hessian of $\rho(\boldsymbol{W})$ evaluated at any critical point $\boldsymbol{W}$ is PSD, i.e., $\nabla^2\rho(\boldsymbol{W}) \succeq \mathbf{0}$ for all $\boldsymbol{W} \in \mathcal{C}$. Despite this fact, the following result establishes the strict saddle property for $g(\boldsymbol{W})$.

Theorem 2: Let $g(\boldsymbol{W})$ be defined as in (29) with $\mu > 0$ and $\operatorname{rank}(\boldsymbol{X}^\star) = r$. Let $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$ be any critical point satisfying $\nabla g(\boldsymbol{W}) = \mathbf{0}$, i.e., $\boldsymbol{W} \in \mathcal{C}$. Any $\boldsymbol{W} \in \mathcal{C} \setminus \mathcal{X}$ is a strict saddle of $g(\boldsymbol{W})$ satisfying

$$\lambda_{\min}(\nabla^2 g(\boldsymbol{W})) \leq -\frac{1}{2}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\| \leq -\sigma_r(\boldsymbol{X}^\star). \quad (39)$$

Furthermore, $g(\boldsymbol{W})$ is not strongly convex at any global minimum point $\boldsymbol{W} \in \mathcal{X}$.

The proof of Theorem 2 is given in Appendix H. We note that this strict saddle property is also covered in [40, Theorem 3], but with much looser bounds (in particular, directly applying [40, Theorem 3] gives $\lambda_{\min}(\nabla^2 g(\boldsymbol{W})) \leq -0.1\sigma_r(\boldsymbol{X}^\star)$ rather than $\lambda_{\min}(\nabla^2 g(\boldsymbol{W})) \leq -\sigma_r(\boldsymbol{X}^\star)$ in (39)). Theorem 2 actually implies that $g(\boldsymbol{W})$ has no spurious local minima (since all local minima belong to $\mathcal{X}$) and obeys the strict saddle property. With the strict saddle property and lack of spurious local minima for $g(\boldsymbol{W})$, the recent results [18], [19] ensure that gradient descent converges to a global minimizer almost surely with random initialization. We also note that Theorem 2 states that $g(\boldsymbol{W})$ is not strongly convex at any global minimum point $\boldsymbol{W} \in \mathcal{X}$ because of the invariance property of $g(\boldsymbol{W})$. This is the reason we introduce the distance in (12) and also the robust strict saddle property in Definition 9.

D. Extension to Over-Parameterized Case: $\operatorname{rank}(\boldsymbol{X}^\star) < r$

In this section, we briefly discuss the over-parameterized scenario where the low-rank matrix $\boldsymbol{X}^\star$ has rank smaller than r. Similar to Theorem 2, the following result shows that the strict saddle property also holds in this case.

Theorem 3: Let $\boldsymbol{X}^\star = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{T}} = \sum_{i=1}^{r'} \sigma_i \boldsymbol{\phi}_i \boldsymbol{\psi}_i^{\mathrm{T}}$ be a reduced SVD of $\boldsymbol{X}^\star$ with $r' \leq r$, and let $g(\boldsymbol{W})$ be defined as in (29) with $\mu > 0$. Any $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$ is a critical point of $g(\boldsymbol{W})$ if and only if $\boldsymbol{W} \in \mathcal{C}$ with

$$\mathcal{C} := \left\{\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} : \boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{R}\boldsymbol{R}^{\mathrm{T}} = \mathbf{I}_{r'}, \right. \\ \left. \boldsymbol{\Lambda} \text{ is diagonal}, \boldsymbol{\Lambda} \geq \mathbf{0}, (\boldsymbol{\Sigma} - \boldsymbol{\Lambda})\boldsymbol{\Sigma} = \mathbf{0}\right\}.$$

Further, all the local minima (which are also global) belong to the following set

$$\mathcal{X} = \left\{\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} : \boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{R}, \boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{R}, \right. \\ \left. \boldsymbol{R}\boldsymbol{R}^{\mathrm{T}} = \mathbf{I}_{r'}\right\}.$$

Finally, any $\boldsymbol{W} \in \mathcal{C} \setminus \mathcal{X}$ is a strict saddle of $g(\boldsymbol{W})$ satisfying

$$\lambda_{\min}(\nabla^2 g(\boldsymbol{W})) \leq -\frac{1}{2}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\| \leq -\sigma_{r'}(\boldsymbol{X}^\star).$$

The proof of Theorem 3 is given in Appendix I. We note that this strict saddle property is also covered in [40, Theorem 3], but with much looser bounds (in particular, directly applying [40, Theorem 3] gives $\lambda_{\min}(\nabla^2 g(\boldsymbol{W})) \leq -0.1\sigma_{r'}(\boldsymbol{X}^\star)$ rather than $\lambda_{\min}(\nabla^2 g(\boldsymbol{W})) \leq -\sigma_{r'}(\boldsymbol{X}^\star)$ in Theorem 3).

E. Extension to Under-Parameterized Case: $\operatorname{rank}(\boldsymbol{X}^\star) > r$

We further discuss the under-parameterized case where $\operatorname{rank}(\boldsymbol{X}^\star) > r$. In this case, (3) is also known as the low-rank approximation problem as the product $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ forms a rank-r approximation to $\boldsymbol{X}^\star$. Similar to Theorem 2, the following result shows that the strict saddle property also holds for $g(\boldsymbol{W})$ in this scenario.

Theorem 4: Let $\boldsymbol{X}^\star = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{T}} = \sum_{i=1}^{r'} \sigma_i \boldsymbol{\phi}_i \boldsymbol{\psi}_i^{\mathrm{T}}$ be a reduced SVD of $\boldsymbol{X}^\star$ with $r' > r$ and $\sigma_r(\boldsymbol{X}^\star) > \sigma_{r+1}(\boldsymbol{X}^\star)$.[5] Also let $g(\boldsymbol{W})$ be defined as in (29) with $\mu > 0$. Any $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$ is a critical point of $g(\boldsymbol{W})$ if and only if $\boldsymbol{W} \in \mathcal{C}$ with

$$\mathcal{C} := \Bigg\{ \boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} : \boldsymbol{U} = \boldsymbol{\Phi}[:,\Omega]\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{V} = \boldsymbol{\Psi}[:,\Omega]\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{\Lambda} = \boldsymbol{\Sigma}[\Omega,\Omega], \boldsymbol{R}\boldsymbol{R}^{\mathrm{T}} = \mathbf{I}_\ell, \Omega \subset \{1,2,\ldots,r'\}, |\Omega| = \ell \le r \Bigg\}$$

where we recall that $\boldsymbol{\Phi}[:,\Omega]$ is a submatrix of $\boldsymbol{\Phi}$ obtained by keeping the columns indexed by Ω and $\boldsymbol{\Sigma}[\Omega,\Omega]$ is an $\ell \times \ell$ matrix obtained by taking the elements of $\boldsymbol{\Sigma}$ in rows and columns indexed by Ω.

Further, all local minima belong to the following set

$$\mathcal{X} = \Bigg\{ \boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} : \boldsymbol{\Lambda} = \boldsymbol{\Sigma}[1:r,1:r], \boldsymbol{R} \in \mathcal{O}_r, \boldsymbol{U} = \boldsymbol{\Phi}[:,1:r]\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{V} = \boldsymbol{\Psi}[:,1:r]\boldsymbol{\Lambda}^{1/2}\boldsymbol{R} \Bigg\}.$$

Finally, any $\boldsymbol{W} \in \mathcal{C}\setminus\mathcal{X}$ is a strict saddle of $g(\boldsymbol{W})$ satisfying

$$\lambda_{\min}(\nabla^2 g(\boldsymbol{W})) \le -(\sigma_r(\boldsymbol{X}^\star) - \sigma_{r+1}(\boldsymbol{X}^\star)).$$

The proof of Theorem 4 is given in Appendix J. It follows from Eckart-Young-Mirsky theorem [51] that for any $\boldsymbol{W} \in \mathcal{X}$, $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ is the best rank-r approximation to $\boldsymbol{X}^\star$. Thus, this strict saddle property ensures that the local search algorithms applied to the factored problem (29) converge to global optimum which corresponds to the best rank-r approximation to $\boldsymbol{X}^\star$. Note that Theorems 2–4 require $\mu > 0$. Based on these results, it has been recently proved in [52] that the strict saddle property also holds for $g(\boldsymbol{W})$ even when $\mu = 0$, but without an explicit bound on $\lambda_{\min}(\nabla^2 g(\boldsymbol{W}))$ as in Theorems 2–4.

F. Robust Strict Saddle Property

We now consider the revised robust strict saddle property defined in Definition 9 for the low-rank matrix factorization problem (29). As guaranteed by Theorem 2, $g(\boldsymbol{W})$ satisfies the strict saddle property for any $\mu > 0$. However, too small a μ would make analyzing the robust strict saddle property difficult. To see this, we denote

$$f(\boldsymbol{W}) = \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star\right\|_F^2$$

for convenience. Thus we can rewrite $g(\boldsymbol{W})$ as the sum of $f(\boldsymbol{W})$ and $\rho(\boldsymbol{W})$. Note that for any $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} \in \mathcal{C}$ where $\mathcal{C}$ is the set of critical points defined in (35), $\widetilde{\boldsymbol{W}} = \begin{bmatrix}\boldsymbol{U}\boldsymbol{M}\\ \boldsymbol{V}\boldsymbol{M}^{-1}\end{bmatrix}$ is a critical point of $f(\boldsymbol{W})$ for any invertible $\boldsymbol{M} \in \mathbb{R}^{r\times r}$. This further implies that the gradient at $\widetilde{\boldsymbol{W}}$ reduces to

$$\nabla g(\widetilde{\boldsymbol{W}}) = \nabla \rho(\widetilde{\boldsymbol{W}}),$$

which could be very small if μ is very small since $\rho(\boldsymbol{W}) = \frac{\mu}{4}\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right\|_F^2$. On the other hand, $\widetilde{\boldsymbol{W}}$ could be far away from any point in $\mathcal{X}$ for some $\boldsymbol{M}$ that is not well-conditioned. Therefore, we choose a proper μ controlling the importance of the regularization term such that for any $\boldsymbol{W}$ that is not close to the critical points $\mathcal{X}$, $g(\boldsymbol{W})$ has large gradient. Motivated by Lemma 4, we choose $\mu = \frac{1}{2}$.

The following result establishes the robust strict saddle property for $g(\boldsymbol{W})$.

Theorem 5: Let $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3', \mathcal{R}_3'', \mathcal{R}_3'''$ be the regions as defined in Theorem 1. Let $g(\boldsymbol{W})$ be defined as in (29) with $\mu = \frac{1}{2}$. Then $g(\boldsymbol{W})$ has the following robust strict saddle property:

1) For any $\boldsymbol{W} \in \mathcal{R}_1$, $g(\boldsymbol{W})$ satisfies local regularity condition:

$$\langle \nabla g(\boldsymbol{W}), \boldsymbol{W} - \boldsymbol{W}^\star\boldsymbol{R}\rangle \ge \frac{1}{32}\sigma_r(\boldsymbol{X}^\star)\operatorname{dist}^2(\boldsymbol{W},\boldsymbol{W}^\star) + \frac{1}{48\|\boldsymbol{X}^\star\|}\left\|\nabla g(\boldsymbol{W})\right\|_F^2, \tag{40}$$

where $\operatorname{dist}(\boldsymbol{W},\boldsymbol{W}^\star)$ and $\boldsymbol{R}$ are defined in (12) and (13), respectively.

2) For any $\boldsymbol{W} \in \mathcal{R}_2$, $g(\boldsymbol{W})$ has a directional negative curvature:

$$\lambda_{\min}\left(\nabla^2 g(\boldsymbol{W})\right) \le -\frac{1}{4}\sigma_r(\boldsymbol{X}^\star). \tag{41}$$

3) For any $\boldsymbol{W} \in \mathcal{R}_3 = \mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$, $g(\boldsymbol{W})$ has large gradient:

$$\|\nabla g(\boldsymbol{W})\|_F \ge \frac{1}{10}\sigma_r^{3/2}(\boldsymbol{X}^\star), \quad \forall\, \boldsymbol{W} \in \mathcal{R}_3'; \tag{42}$$

$$\|\nabla g(\boldsymbol{W})\|_F > \frac{39}{800}\|\boldsymbol{W}\|^3, \quad \forall\, \boldsymbol{W} \in \mathcal{R}_3''; \tag{43}$$

$$\langle \nabla g(\boldsymbol{W}), \boldsymbol{W}\rangle > \frac{1}{20}\left\|\boldsymbol{W}\boldsymbol{W}^{\top}\right\|_F^2, \quad \forall\, \boldsymbol{W} \in \mathcal{R}_3'''. \tag{44}$$

The proof is given in Appendix K.

Remark 8: Recall that all the strict saddles of $g(\boldsymbol{W})$ are actually rank deficient (see Theorem 2). Thus the region $\mathcal{R}_2$ attempts to characterize all the neighbors of the saddle saddles by including all rank deficient points. Actually, (41) holds not only for $\boldsymbol{W} \in \mathcal{R}_2$, but for all $\boldsymbol{W}$ such that $\sigma_r(\boldsymbol{W}) \le \sqrt{\frac{1}{2}}\sigma_r^{1/2}(\boldsymbol{X}^\star)$. The reason we add another constraint controlling the term $\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F$ is to ensure this negative curvature property in the region $\mathcal{R}_2$ also holds for the matrix sensing problem discussed in next section. This is the same reason we add two more constraints $\|\boldsymbol{W}\| \le \frac{20}{19}\|\boldsymbol{W}^\star\|_F$ and $\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F \le \frac{10}{9}\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F$ for the region $\mathcal{R}_3'$.

[5] If $\sigma_{r_1} = \cdots = \sigma_r = \cdots = \sigma_{r_2}$ with $r_1 \le r \le r_2$, then the optimal rank-r approximation to $\boldsymbol{X}^\star$ is not unique. For this case, the optimal solution set $\mathcal{X}$ for the factorized problem needs to be changed correspondingly, but the main arguments still hold.

Appendix B
Proof of Lemma 1

Denote $a_{\boldsymbol{x},\boldsymbol{x}^\star} = \arg\min_{a'\in\mathcal{G}} \|\boldsymbol{x} - a'(\boldsymbol{x}^\star)\|$. Utilizing the definition of distance in (4), the regularity condition (5) and the assumption that $\mu \le 2\beta$, we have

$$\begin{aligned}
&\operatorname{dist}^2(\boldsymbol{x}_{t+1},\boldsymbol{x}^\star)\\
&= \left\|\boldsymbol{x}_{t+1} - a_{\boldsymbol{x}_{t+1},\boldsymbol{x}^\star}(\boldsymbol{x}^\star)\right\|^2\\
&\le \|\boldsymbol{x}_t - \nu\nabla h(\boldsymbol{x}_t) - a_{\boldsymbol{x}_t,\boldsymbol{x}^\star}(\boldsymbol{x}^\star)\|^2\\
&= \|\boldsymbol{x}_t - a_{\boldsymbol{x}_t,\boldsymbol{x}^\star}(\boldsymbol{x}^\star)\|^2 + \nu^2\|\nabla h(\boldsymbol{x}_t)\|^2\\
&\quad - 2\nu\left\langle \boldsymbol{x}_t - a_{\boldsymbol{x}_t,\boldsymbol{x}^\star}(\boldsymbol{x}^\star), \nabla h(\boldsymbol{x}_t)\right\rangle\\
&\le (1-2\nu\alpha)\operatorname{dist}^2(\boldsymbol{x}_t,\boldsymbol{x}^\star) - \nu(2\beta-\nu)\|\nabla h(\boldsymbol{x}_t)\|^2\\
&\le (1-2\nu\alpha)\operatorname{dist}^2(\boldsymbol{x}_t,\boldsymbol{x}^\star)
\end{aligned}$$

where the fourth line uses the regularity condition (5) and the last line holds because $\nu \le 2\beta$. Thus we conclude $\boldsymbol{x}_t \in B(\delta)$ for all $t\in\mathbb{N}$ if $\boldsymbol{x}_0\in B(\delta)$ by noting that $0 \le 1-2\nu\alpha < 1$ since $\alpha\beta \le \frac{1}{4}$ and $\nu \le 2\beta$.

Appendix C
Proof of Proposition 1

First note that if $\boldsymbol{X}^\star$ is a critical point of f, then

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

Now for any $\boldsymbol{X}\in\mathbb{R}^{n\times m}$ with $\operatorname{rank}(\boldsymbol{X}) \le r$, the second order Taylor expansion gives

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

where $\widetilde{\boldsymbol{X}} = t\boldsymbol{X}^\star + (1-t)\boldsymbol{X}$ for some $t\in[0,1]$. This Taylor expansion together with (6) (both $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{X}'-\boldsymbol{X}^\star$ have rank at most $2r$) gives

$$f(\boldsymbol{X}) - f(\boldsymbol{X}^\star) \ge \frac{a}{2}\|\boldsymbol{X}-\boldsymbol{X}^\star\|_F^2.$$

Appendix D
Proof of Lemma 2

Any critical point (see Definition 1) $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$ satisfies $\nabla G(\boldsymbol{W}) = \boldsymbol{0}$, i.e.,

$$\nabla f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{V} + \mu\boldsymbol{U}\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right) = \boldsymbol{0}, \tag{45}$$

$$(\nabla f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}))^{\mathrm{T}}\boldsymbol{U} - \mu\boldsymbol{V}\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right) = \boldsymbol{0}. \tag{46}$$

By (46), we obtain

$$(\nabla f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}))^{\mathrm{T}}\boldsymbol{U} = \mu\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right)\boldsymbol{V}^{\mathrm{T}}.$$

Multiplying (45) by $\boldsymbol{U}^{\mathrm{T}}$ and plugging it in the expression for $\boldsymbol{U}^{\mathrm{T}}\nabla f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})$ from the above equation gives

$$\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right)\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right) = \boldsymbol{0},$$

which further implies

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}.$$

In order to show (11), note that $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$ and $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}$ are the principal square roots (i.e., PSD square roots) of $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$ and $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}$, respectively. Utilizing the result that a PSD matrix $\boldsymbol{A}$ has a unique PSD matrix $\boldsymbol{B}$ such that $\boldsymbol{B}^k = \boldsymbol{A}$ for any $k\ge 1$ [51, Theorem 7.2.6], we obtain

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}$$

for any critical point $\boldsymbol{W}$.

Appendix E
Proof of Lemma 4

We first rewrite the objective function $g(\boldsymbol{W})$:

$$\begin{aligned}
g(\boldsymbol{W}) &= \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}}\right\|_F^2 + \frac{\mu}{4}\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right\|_F^2\\
&\ge \min\{\mu,\frac{1}{2}\}\left(\left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}}\right\|_F^2 + \frac{1}{4}\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right\|_F^2\right)\\
&= \min\{\mu,\frac{1}{2}\}\left(\frac{1}{4}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2 + g'(\boldsymbol{W})\right),
\end{aligned}$$

where the second line attains the equality when $\mu = \frac{1}{2}$, and $g'(\boldsymbol{W})$ in the last line is defined as

$$\begin{aligned}
g'(\boldsymbol{W}) :=& \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}}\right\|_F^2 - \frac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} - \boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}}\right\|_F^2\\
&- \frac{1}{4}\left\|\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{V}^\star\boldsymbol{V}^{\star\mathrm{T}}\right\|_F^2 + \frac{1}{4}\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right\|_F^2.
\end{aligned}$$

We further show $g'(\boldsymbol{W})$ is always nonnegative:

$$\begin{aligned}
g'(\boldsymbol{W}) =& \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}}\right\|_F^2 - \frac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} - \boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}}\right\|_F^2\\
&- \frac{1}{4}\left\|\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{V}^\star\boldsymbol{V}^{\star\mathrm{T}}\right\|_F^2 + \frac{1}{4}\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right\|_F^2.\\
=& \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}}\right\|_F^2 + \frac{1}{2}\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}^\star\right\|_F^2 + \frac{1}{2}\left\|\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}^\star\right\|_F^2\\
&- \frac{1}{2}\operatorname{trace}\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right) - \frac{1}{4}\left\|\boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}}\right\|_F^2 - \frac{1}{4}\left\|\boldsymbol{V}^\star\boldsymbol{V}^{\star\mathrm{T}}\right\|_F^2\\
=& \frac{1}{2}\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}^\star - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}^\star\right\|_F^2 + \frac{1}{2}\left\|\boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}}\right\|_F^2\\
&- \frac{1}{4}\left\|\boldsymbol{U}^\star\boldsymbol{U}^{\star\mathrm{T}}\right\|_F^2 - \frac{1}{4}\left\|\boldsymbol{V}^\star\boldsymbol{V}^{\star\mathrm{T}}\right\|_F^2\\
=& \frac{1}{2}\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}^\star - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}^\star\right\|_F^2 \ge 0,
\end{aligned}$$

where the last line follows because $\boldsymbol{U}^{\star\mathrm{T}}\boldsymbol{U}^\star = \boldsymbol{V}^{\star\mathrm{T}}\boldsymbol{V}^\star$. Thus, we have

$$g(\boldsymbol{W}) \ge \min\{\frac{\mu}{4},\frac{1}{8}\}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2,$$

and

$$g(\boldsymbol{W}) = \frac{1}{8}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2 + \frac{1}{4}\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}^\star - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}^\star\right\|_F^2$$

if $\mu = \frac{1}{2}$.

APPENDIX F
PROOF OF LEMMA 5

We first repeat that $\boldsymbol{X}^\star = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{T}} =$ is a reduced SVD of $\boldsymbol{X}^\star$. We separate $\boldsymbol{U}$ into two parts—the projections onto the column space of $\boldsymbol{\Phi}$ and its orthogonal complement—by denoting $\boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^{1/2}\boldsymbol{R}_1 + \boldsymbol{E}_1$ with $\boldsymbol{R}_1 \in \mathcal{O}_r$, $\boldsymbol{E}_1^{\mathrm{T}}\boldsymbol{\Phi} = \boldsymbol{0}$ and $\boldsymbol{\Lambda}_1$ being a $r \times r$ diagonal matrix with non-negative elements along its diagonal. Similarly, denote $\boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{R}_2 + \boldsymbol{E}_2$, where $\boldsymbol{R}_2 \in \mathcal{O}_r$, $\boldsymbol{E}_2^{\mathrm{T}}\boldsymbol{\Psi} = \boldsymbol{0}$, $\boldsymbol{\Lambda}_2$ is a $r \times r$ diagonal matrix with non-negative elements along its diagonal. Recall that any critical point $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$ satisfies

$$\nabla_{\boldsymbol{U}}\rho(\boldsymbol{U},\boldsymbol{V}) = \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{X}^\star\boldsymbol{V} = \boldsymbol{0},$$
$$\nabla_{\boldsymbol{V}}\rho(\boldsymbol{U},\boldsymbol{V}) = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} - \boldsymbol{X}^{\star\mathrm{T}}\boldsymbol{U} = \boldsymbol{0}.$$

Plugging $\boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^{1/2}\boldsymbol{R}_1 + \boldsymbol{E}_1$ and $\boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{R}_2 + \boldsymbol{E}_2$ into the above equations gives

$$\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^{3/2}\boldsymbol{R}_1 + \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^{1/2}\boldsymbol{R}_1\boldsymbol{E}_1^{\mathrm{T}}\boldsymbol{E}_1 + \boldsymbol{E}_1\boldsymbol{R}_1^{\mathrm{T}}\boldsymbol{\Lambda}_1\boldsymbol{R}_1 + \boldsymbol{E}_1\boldsymbol{E}_1^{\mathrm{T}}\boldsymbol{E}_1 - \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{R}_2 = \boldsymbol{0}, \tag{47}$$
$$\boldsymbol{\Psi}\boldsymbol{\Lambda}_2^{3/2}\boldsymbol{R}_2 + \boldsymbol{\Psi}\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{R}_2\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}_2 + \boldsymbol{E}_2\boldsymbol{R}_2^{\mathrm{T}}\boldsymbol{\Lambda}_2\boldsymbol{R}_2 + \boldsymbol{E}_2\boldsymbol{E}_2^{\mathrm{T}}\boldsymbol{E}_2 - \boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Lambda}_1^{1/2}\boldsymbol{R}_1 = \boldsymbol{0}. \tag{48}$$

Since $\boldsymbol{E}_1$ is orthogonal to $\boldsymbol{\Phi}$, (47) further implies that

$$\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^{3/2}\boldsymbol{R}_1 + \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^{1/2}\boldsymbol{R}_1\boldsymbol{E}_1^{\mathrm{T}}\boldsymbol{E}_1 - \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{R}_2 = \boldsymbol{0}, \tag{49}$$
$$\boldsymbol{E}_1\boldsymbol{R}_1^{\mathrm{T}}\boldsymbol{\Lambda}_1\boldsymbol{R}_1 + \boldsymbol{E}_1\boldsymbol{E}_1^{\mathrm{T}}\boldsymbol{E}_1 = \boldsymbol{0}. \tag{50}$$

From (50), we have

$$\left\langle \boldsymbol{E}_1\boldsymbol{R}_1^{\mathrm{T}}\boldsymbol{\Lambda}_1\boldsymbol{R}_1 + \boldsymbol{E}_1\boldsymbol{E}_1^{\mathrm{T}}\boldsymbol{E}_1, \boldsymbol{E}_1\right\rangle = \left\langle \boldsymbol{R}_1^{\mathrm{T}}\boldsymbol{\Lambda}_1\boldsymbol{R}_1, \boldsymbol{E}_1^{\mathrm{T}}\boldsymbol{E}_1\right\rangle + \|\boldsymbol{E}_1\|_F^2 = 0,$$

which further implies $\|\boldsymbol{E}_1\|_F^2 = 0$ by noting that $\left\langle \boldsymbol{R}_1^{\mathrm{T}}\boldsymbol{\Lambda}_1\boldsymbol{R}_1, \boldsymbol{E}_1^{\mathrm{T}}\boldsymbol{E}_1\right\rangle \geq 0$ since it is the inner product between two PSD matrices. Thus $\boldsymbol{E}_1 = \boldsymbol{0}$. With a similar argument we also have $\boldsymbol{E}_2 = \boldsymbol{0}$.

With $\boldsymbol{E}_1 = \boldsymbol{E}_2 = \boldsymbol{0}$, (49) reduces to

$$\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^{3/2}\boldsymbol{R}_1 - \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{R}_2 = \boldsymbol{0}.$$

Since $\boldsymbol{\Phi}$ is orthonormal and $\boldsymbol{R}_1 \in \mathcal{O}_r$, the above equation implies that

$$\boldsymbol{\Lambda}_1^{3/2} = \boldsymbol{\Sigma}\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{R}_2\boldsymbol{R}_1^{\mathrm{T}}.$$

Let Ω denote the set of locations of the non-zero diagonals in $\boldsymbol{\Lambda}_2$, i.e., $\boldsymbol{\Lambda}_2[i,i] > 0$ for all $i \in \Omega$. Then $[\boldsymbol{R}_1^{\mathrm{T}}]_\Omega = [\boldsymbol{R}_2^{\mathrm{T}}]_\Omega$ since otherwise $\boldsymbol{\Sigma}\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{R}_2\boldsymbol{R}_1^{\mathrm{T}}$ is not a diagonal matrix anymore. Then we have

$$\boldsymbol{\Lambda}_1^{3/2} = \boldsymbol{\Sigma}\boldsymbol{\Lambda}_2^{1/2} \tag{51}$$

implying that the set of the locations of non-zero diagonals in $\boldsymbol{\Lambda}_1$ is identical to Ω. A similar argument applied to (48) gives

$$\boldsymbol{\Lambda}_2^{3/2} = \boldsymbol{\Sigma}\boldsymbol{\Lambda}_1^{1/2}. \tag{52}$$

Noting that (51) implies $\boldsymbol{\Lambda}_1^{3/2}[i,i] = \boldsymbol{\Sigma}[i,i]\boldsymbol{\Lambda}_2^{1/2}[i,i]$ and (52) implies $\boldsymbol{\Lambda}_2^{3/2}[i,i] = \boldsymbol{\Sigma}[i,i]\boldsymbol{\Lambda}_1^{1/2}[i,i]$, for all $i \in \Omega$ we have $\boldsymbol{\Lambda}_1[i,i] = \boldsymbol{\Lambda}_2[i,i] = \boldsymbol{\Sigma}[i,i]$. For $i \notin \Omega$, we have $\boldsymbol{\Lambda}_1[i,i] = \boldsymbol{\Lambda}_2[i,i] = 0$. Thus $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$. For convenience, denote $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$ with $\boldsymbol{\Lambda}[i,i] = \lambda_i$.

Finally, we note that $\boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}_1 = \sum_{i\in\Omega}\lambda_i\boldsymbol{\phi}_i\boldsymbol{R}_1[i,:]$ and $\boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}_2 = \sum_{i\in\Omega}\lambda_i\boldsymbol{\psi}_i\boldsymbol{R}_2[i,:]$ implying that only $[\boldsymbol{R}_1^{\mathrm{T}}]_\Omega$ and $[\boldsymbol{R}_2^{\mathrm{T}}]_\Omega$ play a role in $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Thus one can set $\boldsymbol{R}_1 = \boldsymbol{R}_2$ since we already proved $[\boldsymbol{R}_1^{\mathrm{T}}]_\Omega = [\boldsymbol{R}_2^{\mathrm{T}}]_\Omega$.

APPENDIX G
PROOF OF LEMMA 6

Utilizing the result that any point $\boldsymbol{W} \in \mathcal{E}$ satisfies $\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{0}$, we directly obtain

$$\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{U}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{\Delta}_{\boldsymbol{V}}\boldsymbol{V}^{\mathrm{T}}\|_F^2 = \|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{\Delta}_{\boldsymbol{V}}\boldsymbol{U}^{\mathrm{T}}\|_F^2$$

since $\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{U}^{\mathrm{T}}\|_F^2 = \operatorname{trace}\left(\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{U}}\right) = \operatorname{trace}\left(\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Delta}_{\boldsymbol{U}}\right) = \|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}}\|_F^2$ (and similarly for the other two terms).

We then rewrite the last two terms in (32) as

$$\begin{aligned}
&\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Delta}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{W}^{\mathrm{T}}\right\rangle + \left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\rangle \\
&= \left\langle \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta}, \boldsymbol{\Delta}^{\mathrm{T}}\widehat{\boldsymbol{W}}\right\rangle + \left\langle \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta}, \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta}\right\rangle \\
&= \left\langle \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta}, \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta} + \boldsymbol{\Delta}^{\mathrm{T}}\widehat{\boldsymbol{W}}\right\rangle \\
&= \frac{1}{2}\left\langle \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta} + \boldsymbol{\Delta}^{\mathrm{T}}\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta} + \boldsymbol{\Delta}^{\mathrm{T}}\widehat{\boldsymbol{W}}\right\rangle \\
&\quad + \frac{1}{2}\left\langle \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta} - \boldsymbol{\Delta}^{\mathrm{T}}\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta} + \boldsymbol{\Delta}^{\mathrm{T}}\widehat{\boldsymbol{W}}\right\rangle \\
&= \frac{1}{2}\left\|\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta} + \boldsymbol{\Delta}^{\mathrm{T}}\widehat{\boldsymbol{W}}\right\|_F^2,
\end{aligned}$$

where the last line holds because $\left\langle \boldsymbol{A} - \boldsymbol{A}^{\mathrm{T}}, \boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}}\right\rangle = 0$. Plugging these with the factor $\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{0}$ into the Hessian quadratic form $[\nabla^2\rho(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta})$ defined in (32) gives

$$[\nabla^2\rho(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta}) \geq \frac{\mu}{2}\left\|\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta} + \boldsymbol{\Delta}^{\mathrm{T}}\widehat{\boldsymbol{W}}\right\|_F^2 \geq 0.$$

This implies that the Hessian of ρ evaluated at any $\boldsymbol{W} \in \mathcal{E}$ is PSD, i.e., $\nabla^2\rho(\boldsymbol{W}) \succeq \boldsymbol{0}$.[6]

APPENDIX H
PROOF OF THEOREM 2 (STRICT SADDLE PROPERTY FOR (29))

We begin the proof of Theorem 2 by characterizing any $\boldsymbol{W} \in \mathcal{C} \setminus \mathcal{X}$. For this purpose, let $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$, where $\boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{R} \in \mathcal{O}_r, \boldsymbol{\Lambda}$ is diagonal, $\boldsymbol{\Lambda} \geq \boldsymbol{0}, (\boldsymbol{\Sigma} - \boldsymbol{\Lambda})\boldsymbol{\Sigma} = \boldsymbol{0}$, and $\operatorname{rank}(\boldsymbol{\Lambda}) < r$. Denote the corresponding optimal solution $\boldsymbol{W}^\star = \begin{bmatrix}\boldsymbol{U}^\star\\ \boldsymbol{V}^\star\end{bmatrix}$, where $\boldsymbol{U}^\star = \boldsymbol{\Phi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{R}, \boldsymbol{V}^\star = \boldsymbol{\Psi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{R}$. Let

$$k = \arg\max_i \sigma_i - \lambda_i$$

[6]This can also be observed since any critical point $\boldsymbol{W}$ is a global minimum of $\rho(\boldsymbol{W})$, which directly indicates that $\nabla^2\rho(\boldsymbol{W}) \succeq \boldsymbol{0}$.

denote the location of the first zero diagonal element in $\mathbf{\Lambda}$. Noting that $\lambda_i \in \{\sigma_i, 0\}$, we conclude that

$$\lambda_k = 0, \quad \boldsymbol{\phi}_k^{\mathrm{T}} \boldsymbol{U} = \mathbf{0}, \quad \boldsymbol{\psi}_k^{\mathrm{T}} \boldsymbol{V} = \mathbf{0}. \tag{53}$$

In words, $\boldsymbol{\phi}_k$ and $\boldsymbol{\psi}_k$ are orthogonal to $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Let $\boldsymbol{\alpha} \in \mathbb{R}^r$ be the eigenvector associated with the smallest eigenvalue of $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}$. Such $\boldsymbol{\alpha}$ simultaneously lives in the null spaces of $\boldsymbol{U}$ and $\boldsymbol{V}$ since $\boldsymbol{W}$ is rank deficient indicating

$$0 = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\alpha} + \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\alpha},$$

which further implies

$$\begin{cases}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\alpha} = 0,\\ \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\alpha} = 0.\end{cases} \tag{54}$$

With this property, we construct $\mathbf{\Delta}$ by setting $\mathbf{\Delta}_{\boldsymbol{U}} = \boldsymbol{\phi}_k \boldsymbol{\alpha}^{\mathrm{T}}$ and $\mathbf{\Delta}_{\boldsymbol{V}} = \boldsymbol{\psi}_k \boldsymbol{\alpha}^{\mathrm{T}}$. Now we show that $\boldsymbol{W}$ is a strict saddle by arguing that $g(\boldsymbol{W})$ has a strictly negative curvature along the constructed direction $\mathbf{\Delta}$, i.e., $[\nabla^2 g(\boldsymbol{W})](\mathbf{\Delta}, \mathbf{\Delta}) < 0$. To that end, we compute the five terms in (31) as follows

$$\begin{aligned}\left\|\mathbf{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\mathbf{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2 &= 0 \quad (\text{since} \quad (54)),\\ \left\langle \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star, \mathbf{\Delta}_{\boldsymbol{U}}\mathbf{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\rangle &= \lambda_k - \sigma_k = -\sigma_k \quad (\text{since} \quad (53)),\\ \left\langle \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W}, \widehat{\mathbf{\Delta}}^{\mathrm{T}}\mathbf{\Delta}\right\rangle &= 0 \quad (\text{since} \quad \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W} = \mathbf{0}),\\ \left\langle \widehat{\boldsymbol{W}}\widehat{\mathbf{\Delta}}^{\mathrm{T}}, \mathbf{\Delta}\boldsymbol{W}^{\mathrm{T}}\right\rangle &= \operatorname{trace}\left(\widehat{\mathbf{\Delta}}^{\mathrm{T}}\boldsymbol{W}\mathbf{\Delta}^T\widehat{\boldsymbol{W}}\right) = 0,\\ \left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \mathbf{\Delta}\mathbf{\Delta}^{\mathrm{T}}\right\rangle &= \operatorname{trace}\left(\widehat{\boldsymbol{W}}^{\mathrm{T}}\mathbf{\Delta}\mathbf{\Delta}^{\mathrm{T}}\widehat{\boldsymbol{W}}\right) = 0,\end{aligned}$$

where $\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W} = \mathbf{0}$ since $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \mathbf{0}$, the last two lines utilize $\widehat{\mathbf{\Delta}}^{\mathrm{T}}\boldsymbol{W} = \mathbf{0}$ (or $\widehat{\boldsymbol{W}}^{\mathrm{T}}\mathbf{\Delta} = \mathbf{0}$) because $\widehat{\mathbf{\Delta}}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{\alpha}\boldsymbol{\phi}_k^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{\alpha}\boldsymbol{\psi}_k^{\mathrm{T}}\boldsymbol{V} = \mathbf{0}$ (see (53)). Plugging these terms into (31) gives

$$\begin{aligned}&[\nabla^2 g(\boldsymbol{W})](\mathbf{\Delta}, \mathbf{\Delta})\\ &= \left\|\mathbf{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\mathbf{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2 + 2\left\langle \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star, \mathbf{\Delta}_{\boldsymbol{U}}\mathbf{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\rangle\\ &\quad + \mu\left\langle \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W}, \widehat{\mathbf{\Delta}}^{\mathrm{T}}\mathbf{\Delta}\right\rangle + \mu\left\langle \widehat{\boldsymbol{W}}\widehat{\mathbf{\Delta}}^{\mathrm{T}}, \mathbf{\Delta}\boldsymbol{W}^{\mathrm{T}}\right\rangle\\ &\quad + \mu\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \mathbf{\Delta}\mathbf{\Delta}^{\mathrm{T}}\right\rangle\\ &= -2\sigma_k.\end{aligned}$$

The proof of the strict saddle property is completed by noting that

$$\|\mathbf{\Delta}\|_F^2 = \|\mathbf{\Delta}_{\boldsymbol{U}}\|_F^2 + \|\mathbf{\Delta}_{\boldsymbol{V}}\|_F^2 = \left\|\boldsymbol{\phi}_k\boldsymbol{\alpha}^{\mathrm{T}}\right\|_F^2 + \left\|\boldsymbol{\psi}_k\boldsymbol{\alpha}^{\mathrm{T}}\right\|_F^2 = 2,$$

which further implies

$$\begin{aligned}\lambda_{\min}\left(\nabla^2 g(\boldsymbol{W})\right) &\le \frac{[\nabla^2 g(\boldsymbol{W})](\mathbf{\Delta}, \mathbf{\Delta})}{\|\mathbf{\Delta}\|_F^2} \le -\frac{2\sigma_k}{2}\\ &= -\|\mathbf{\Lambda} - \mathbf{\Sigma}\| = -\frac{1}{2}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|,\end{aligned}$$

where the first equality holds because

$$\|\mathbf{\Lambda} - \mathbf{\Sigma}\| = \max_i \sigma_i - \lambda_i = \sigma_k,$$

and the second equality follows since

$$\begin{aligned}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}} &= \frac{1}{2}\boldsymbol{Q}\left(\mathbf{\Lambda} - \mathbf{\Sigma}\right)\boldsymbol{Q}^{\mathrm{T}},\\ \boldsymbol{Q} = \begin{bmatrix}\mathbf{\Phi}/\sqrt{2}\\ \mathbf{\Psi}/\sqrt{2}\end{bmatrix}, &\quad \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \mathbf{I}.\end{aligned}$$

We finish the proof of (39) by noting that

$$\sigma_k = \sigma_k(\boldsymbol{X}^\star) \ge \sigma_r(\boldsymbol{X}^\star).$$

Now suppose $\boldsymbol{W}^\star \in \mathcal{X}$. Applying (38), which states that the Hessian of ρ evaluated at any critical point $\boldsymbol{W}$ is PSD, we have

$$\begin{aligned}&[\nabla^2 g(\boldsymbol{W}^\star)][\mathbf{\Delta}, \mathbf{\Delta}]\\ &= \left\|\mathbf{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\star\mathrm{T}} + \boldsymbol{U}^\star\mathbf{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2 + 2\left\langle \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}} - \boldsymbol{X}^\star, \mathbf{\Delta}_{\boldsymbol{U}}\mathbf{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\rangle\\ &\quad + [\nabla^2 \rho(\boldsymbol{W}^\star)][\mathbf{\Delta}, \mathbf{\Delta}]\\ &\ge \left\|\mathbf{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\star\mathrm{T}} + \boldsymbol{U}^\star\mathbf{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2 + 2\left\langle \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}} - \boldsymbol{X}^\star, \mathbf{\Delta}_{\boldsymbol{U}}\mathbf{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\rangle\\ &\ge 0\end{aligned}$$

since $\boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}} - \boldsymbol{X}^\star = \mathbf{0}$. We show g is not strongly convex at $\boldsymbol{W}^\star$ by arguing that $\lambda_{\min}(\nabla^2 g(\boldsymbol{W}^\star)) = 0$. For this purpose, we first recall that $\boldsymbol{U}^\star = \mathbf{\Phi}\mathbf{\Sigma}^{1/2}, \boldsymbol{V}^\star = \mathbf{\Psi}\mathbf{\Sigma}^{1/2}$, where we assume $\boldsymbol{R} = \mathbf{I}$ without loss of generality. Let $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_r\}$ be the standard orthobasis for $\mathbb{R}^r$, i.e., $\boldsymbol{e}_\ell$ is the ℓ-th column of the $r \times r$ identity matrix. Construct $\mathbf{\Delta}_{(i,j)} = \begin{bmatrix}\mathbf{\Delta}_{\boldsymbol{U}}^{(i,j)}\\ \mathbf{\Delta}_{\boldsymbol{V}}^{(i,j)}\end{bmatrix}$, where

$$\mathbf{\Delta}_{\boldsymbol{U}}^{(i,j)} = \boldsymbol{U}^\star\boldsymbol{e}_j\boldsymbol{e}_i^{\mathrm{T}} - \boldsymbol{U}^\star\boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}}, \quad \mathbf{\Delta}_{\boldsymbol{V}}^{(i,j)} = \boldsymbol{V}^\star\boldsymbol{e}_j\boldsymbol{e}_i^{\mathrm{T}} - \boldsymbol{U}^\star\boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}},$$

for any $1 \le i < j \le r$. That is, the ℓ-th columns of the matrices $\mathbf{\Delta}_{\boldsymbol{U}}^{(i,j)}$ and $\mathbf{\Delta}_{\boldsymbol{V}}^{(i,j)}$ are respectively given by

$$\mathbf{\Delta}_{\boldsymbol{U}}^{(i,j)}[:, \ell] = \begin{cases}\sigma_j^{1/2}\boldsymbol{\phi}_j, & \ell = i,\\ -\sigma_i^{1/2}\boldsymbol{\phi}_i, & \ell = j,\\ \mathbf{0}, & \text{otherwise},\end{cases},$$

$$\mathbf{\Delta}_{\boldsymbol{V}}^{(i,j)}[:, \ell] = \begin{cases}\sigma_j^{1/2}\boldsymbol{\psi}_j, & \ell = i,\\ -\sigma_i^{1/2}\boldsymbol{\psi}_i, & \ell = j,\\ \mathbf{0}, & \text{otherwise},\end{cases}$$

for any $1 \le i < j \le r$. We then compute the five terms in (31) as follows

$$\begin{aligned}&\left\|\mathbf{\Delta}_{\boldsymbol{U}}^{(i,j)}\boldsymbol{V}^{\star\mathrm{T}} + \boldsymbol{U}^\star(\mathbf{\Delta}_{\boldsymbol{V}}^{(i,j)})^{\mathrm{T}}\right\|_F^2\\ &= \left\|\sigma_i^{1/2}\sigma_j^{1/2}\left(\boldsymbol{\phi}_j\boldsymbol{\psi}_i^{\mathrm{T}} - \boldsymbol{\phi}_i\boldsymbol{\psi}_j^{\mathrm{T}} + \boldsymbol{\phi}_i\boldsymbol{\psi}_j^{\mathrm{T}} - \boldsymbol{\phi}_j\boldsymbol{\psi}_i^{\mathrm{T}}\right)\right\|_F^2 = 0,\\ &\langle \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}} - \boldsymbol{X}^\star, \mathbf{\Delta}_{\boldsymbol{U}}^{(i,j)}(\mathbf{\Delta}_{\boldsymbol{V}}^{(i,j)})^{\mathrm{T}}\rangle = 0 \ (\text{as } \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}} - \boldsymbol{X}^\star = \mathbf{0}),\\ &\langle \widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}^\star, \widehat{\mathbf{\Delta}}_{(i,j)}^{\mathrm{T}}\mathbf{\Delta}_{(i,j)}\rangle = 0 \quad (\text{as } \widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}^\star = \mathbf{0}),\\ &\langle \widehat{\boldsymbol{W}}^\star\widehat{\mathbf{\Delta}}_{(i,j)}^{\mathrm{T}}, \mathbf{\Delta}_{(i,j)}\boldsymbol{W}^{\star\mathrm{T}}\rangle = \operatorname{trace}(\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\mathbf{\Delta}_{(i,j)}\boldsymbol{W}^{\star\mathrm{T}}\widehat{\mathbf{\Delta}}_{(i,j)}) = 0,\\ &\langle \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \mathbf{\Delta}_{(i,j)}\mathbf{\Delta}_{(i,j)}^{\mathrm{T}}\rangle = \operatorname{trace}(\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\mathbf{\Delta}_{(i,j)}\mathbf{\Delta}_{(i,j)}^{\mathrm{T}}\widehat{\boldsymbol{W}}^\star) = 0,\end{aligned}$$

where the last two lines hold because

$$\begin{aligned}\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\mathbf{\Delta}_{(i,j)} &= \boldsymbol{U}^{\star\mathrm{T}}\boldsymbol{U}^\star(\boldsymbol{e}_j\boldsymbol{e}_i^{\mathrm{T}} - \boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}}) - \boldsymbol{V}^{\star\mathrm{T}}\boldsymbol{V}^\star(\boldsymbol{e}_j\boldsymbol{e}_i^{\mathrm{T}} - \boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}})\\ &= \mathbf{0}\end{aligned}$$

since $\boldsymbol{U}^{\star\mathrm{T}}\boldsymbol{U}^\star = \boldsymbol{V}^{\star\mathrm{T}}\boldsymbol{V}^\star$.

Thus, we obtain the Hessian evaluated at the optimal solution point $\boldsymbol{W}^\star$ along the direction $\boldsymbol{\Delta}^{(i,j)}$:

$$\left[\nabla^2 g(\boldsymbol{W}^\star)\right]\left(\boldsymbol{\Delta}^{(i,j)}, \boldsymbol{\Delta}^{(i,j)}\right) = 0$$

for all $1 \le i < j \le r$. This proves that $g(\boldsymbol{W})$ is not strongly convex at a global minimum point $\boldsymbol{W}^\star \in \mathcal{X}$.

Appendix I
Proof of Theorem 3 (Strict Saddle Property of $g(\boldsymbol{W})$ When Over-Parameterized)

Let $\boldsymbol{X}^\star = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{T}} = \sum_{i=1}^{r'} \sigma_i \boldsymbol{\phi}_i \boldsymbol{\psi}_i^{\mathrm{T}}$ be a reduced SVD of $\boldsymbol{X}^\star$ with $r' \le r$. Using an approach similar to that in Appendix F for proving Lemma 5, we can show that any $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$ is a critical point of $g(\boldsymbol{W})$ if and only if $\boldsymbol{W} \in \mathcal{C}$ with

$$\mathcal{C} = \left\{\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} : \boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{R}\boldsymbol{R}^{\mathrm{T}} = \mathbf{I}_{r'}, \right.$$
$$\left. \boldsymbol{\Lambda} \text{ is diagonal}, \boldsymbol{\Lambda} \ge \mathbf{0}, (\boldsymbol{\Sigma} - \boldsymbol{\Lambda})\boldsymbol{\Sigma} = \mathbf{0}\right\}.$$

Recall that

$$\mathcal{X} = \left\{\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} : \boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{R}, \boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{R}, \right.$$
$$\left. \boldsymbol{R}\boldsymbol{R}^{\mathrm{T}} = \mathbf{I}_{r'}\right\}.$$

It is clear that $\mathcal{X}$ is the set of optimal solutions since for any $\boldsymbol{W} \in \mathcal{X}$, $g(\boldsymbol{W})$ achieves its global minimum, i.e., $g(\boldsymbol{W}) = 0$.

Using an approach similar to that in Appendix H for proving Theorem 2, we can show that any $\boldsymbol{W} \in \mathcal{C}\setminus\mathcal{X}$ is a strict saddle satisfying

$$\lambda_{\min}\left(\nabla^2 g(\boldsymbol{W})\right) \le -\sigma_{r'}(\boldsymbol{X}^\star).$$

Appendix J
Proof of Theorem 4 (Strict Saddle Property of $g(\boldsymbol{W})$ When Under-Parameterized)

Let $\boldsymbol{X}^\star = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{T}} = \sum_{i=1}^{r'} \sigma_i \boldsymbol{\phi}_i \boldsymbol{\psi}_i^{\mathrm{T}}$ be a reduced SVD of $\boldsymbol{X}^\star$ with $r' > r$ and $\sigma_r(\boldsymbol{X}^\star) > \sigma_{r+1}(\boldsymbol{X}^\star)$. Using an approach similar to that in Appendix F for proving Lemma 5, we can show that any $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$ is a critical point of $g(\boldsymbol{W})$ if and only if $\boldsymbol{W} \in \mathcal{C}$ with

$$\mathcal{C} = \left\{\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} : \boldsymbol{U} = \boldsymbol{\Phi}[:,\Omega]\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{V} = \boldsymbol{\Psi}[:,\Omega]\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \right.$$
$$\left. \boldsymbol{\Lambda} = \boldsymbol{\Sigma}[\Omega,\Omega], \boldsymbol{R}\boldsymbol{R}^{\mathrm{T}} = \mathbf{I}_\ell, \Omega \subset \{1,\dots,r'\}, |\Omega| = \ell \le r\right\}.$$

Intuitively, a critical point is one such that $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ is a rank-ℓ approximation to $\boldsymbol{X}^\star$ with $\ell \le r$ and $\boldsymbol{U}$ and $\boldsymbol{V}$ are equal factors of their product $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$.

It follows from the Eckart-Young-Mirsky theorem [51] that the set of optimal solutions is given by

$$\mathcal{X} = \left\{\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} : \boldsymbol{U} = \boldsymbol{\Phi}[:,1:r]\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \right.$$
$$\left. \boldsymbol{V} = \boldsymbol{\Psi}[:,1:r]\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{\Lambda} = \boldsymbol{\Sigma}[1:r,1:r], \boldsymbol{R} \in \mathcal{O}_r\right\}.$$

Now we characterize any $\boldsymbol{W} \in \mathcal{C}\setminus\mathcal{X}$ by letting $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$, where

$$\boldsymbol{U} = \boldsymbol{\Phi}[:,\Omega]\boldsymbol{\Lambda}^{1/2}\boldsymbol{R}, \boldsymbol{V} = \boldsymbol{\Psi}[:,\Omega]\boldsymbol{\Lambda}^{1/2}\boldsymbol{R},$$
$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}[\Omega,\Omega], \boldsymbol{R} \in \mathbb{R}^{\ell\times r}, \boldsymbol{R}\boldsymbol{R}^{\mathrm{T}} = \mathbf{I}_\ell,$$
$$\Omega \subset \{1,2,\dots,r'\}, |\Omega| = \ell \le r, \Omega \ne \{1,2,\dots,r\}.$$

Let $\boldsymbol{\alpha} \in \mathbb{R}^r$ be the eigenvector associated with the smallest eigenvalue of $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$ (or $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}$). By the typical structures in $\boldsymbol{U}$ and $\boldsymbol{V}$ (see the above equation), we have

$$\|\boldsymbol{V}\boldsymbol{\alpha}\|_F^2 = \|\boldsymbol{U}\boldsymbol{\alpha}\|_F^2 = \sigma_r^2(\boldsymbol{U}) = \begin{cases}\sigma_j(\boldsymbol{X}^\star), & |\Omega| = r \text{ and } j = \max\Omega\\ 0, & |\Omega| < r,\end{cases} \tag{55}$$

where $j > r$ because $\Omega \ne \{1,2,\dots,r\}$. Note that there always exists an index

$$i \in \{1,2,\dots,r\}, i \ne \Omega$$

since $\Omega \ne \{1,2,\dots,r\}$ and $|\Omega| \le r$. We construct $\boldsymbol{\Delta}$ by setting

$$\boldsymbol{\Delta}_{\boldsymbol{U}} = \boldsymbol{\phi}_i\boldsymbol{\alpha}^{\mathrm{T}}, \quad \boldsymbol{\Delta}_{\boldsymbol{V}} = \boldsymbol{\psi}_i\boldsymbol{\alpha}^{\mathrm{T}}.$$

Since $i \notin \Omega$, we have

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Delta}_{\boldsymbol{U}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\phi}_i\boldsymbol{\alpha}^{\mathrm{T}} = \mathbf{0},$$
$$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Delta}_{\boldsymbol{V}} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\psi}_i\boldsymbol{\alpha}^{\mathrm{T}} = \mathbf{0}. \tag{56}$$

We compute the five terms in (31) as follows

$$\begin{aligned}
&\left\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2\\
&= \left\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}}\right\|_F^2 + \left\|\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2 + 2\,\mathrm{trace}\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Delta}_{\boldsymbol{V}}\right)\\
&= 2\sigma_r^2(\boldsymbol{U}),
\end{aligned}$$

$$\begin{aligned}
\left\langle\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star, \boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\rangle &= \left\langle\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star, \boldsymbol{\phi}_i\boldsymbol{\psi}_i^{\mathrm{T}}\right\rangle\\
&= -\left\langle\boldsymbol{X}^\star, \boldsymbol{\phi}_i\boldsymbol{\psi}_i^{\mathrm{T}}\right\rangle = -\sigma_i(\boldsymbol{X}^\star),\\
\left\langle\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W}, \widehat{\boldsymbol{\Delta}}^{\mathrm{T}}\boldsymbol{\Delta}\right\rangle &= 0 \quad (\text{since } \widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W} = \mathbf{0}),\\
\left\langle\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Delta}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{W}^{\mathrm{T}}\right\rangle &= \mathrm{trace}\left(\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{W}^{\mathrm{T}}\widehat{\boldsymbol{\Delta}}\right) = 0,\\
\left\langle\widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\rangle &= \mathrm{trace}\left(\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\widehat{\boldsymbol{W}}\right) = 0,
\end{aligned}$$

where the last equality in the first line holds because $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Delta}_{\boldsymbol{U}} = \mathbf{0}$ (see (56)) and $\left\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}}\right\|_F^2 = \left\|\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2 = \sigma_r^2(\boldsymbol{U})$ (see (55)), $\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W} = \mathbf{0}$ in the third line holds since $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \mathbf{0}$, and $\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta} = \mathbf{0}$ in the fourth and last lines holds because

$$\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Delta}_{\boldsymbol{U}} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Delta}_{\boldsymbol{V}} = \mathbf{0}.$$

Now plugging these terms into (31) yields

$$\begin{aligned}
&[\nabla^2 g(\boldsymbol{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta})\\
&= \|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\|_F^2 + 2\langle\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star, \boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\rangle\\
&\quad + \mu(\langle\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W}, \widehat{\boldsymbol{\Delta}}^{\mathrm{T}}\boldsymbol{\Delta}\rangle + \langle\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Delta}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{W}^{\mathrm{T}}\rangle + \langle\widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\rangle)\\
&= -2(\sigma_i(\boldsymbol{X}^\star) - \sigma_r^2(\boldsymbol{U})).
\end{aligned}$$

The proof of the strict saddle property is completed by noting that

$$\|\boldsymbol{\Delta}\|_F^2 = \|\boldsymbol{\Delta}_{\boldsymbol{U}}\|_F^2 + \|\boldsymbol{\Delta}_{\boldsymbol{V}}\|_F^2 = 2,$$

which further implies

$$\begin{aligned}\lambda_{\min}\left(\nabla^2 g(\boldsymbol{W})\right) &\leq -2\frac{\sigma_i(\boldsymbol{X}^\star) - \sigma_r^2(\boldsymbol{U})}{\|\boldsymbol{\Delta}\|_F^2} \\ &\leq -\left(\sigma_r(\boldsymbol{X}^\star) - \sigma_{r+1}(\boldsymbol{X}^\star)\right),\end{aligned}$$

where the last inequality holds because of (55) and because $i \leq r$.

Appendix K
Proof of Theorem 5 (Robust Strict Saddle for $g(\boldsymbol{W})$)

We first establish the following useful results.

Lemma 7: For any two PSD matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$, we have

$$\sigma_n(\boldsymbol{A})\operatorname{trace}(\boldsymbol{B}) \leq \operatorname{trace}(\boldsymbol{A}\boldsymbol{B}) \leq \|\boldsymbol{A}\|\operatorname{trace}(\boldsymbol{B}).$$

Proof of Lemma 7: Let $\boldsymbol{A} = \boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1^{\mathrm{T}}$ and $\boldsymbol{B} = \boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2^{\mathrm{T}}$ be the eigendecompositions of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. Here $\boldsymbol{\Lambda}_1$ ($\boldsymbol{\Lambda}_2$) is a diagonal matrix with the eigenvalues of $\boldsymbol{A}$ ($\boldsymbol{B}$) along its diagonal. We first rewrite $\operatorname{trace}(\boldsymbol{A}\boldsymbol{B})$ as

$$\operatorname{trace}(\boldsymbol{A}\boldsymbol{B}) = \operatorname{trace}\left(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1^{\mathrm{T}}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2^{\mathrm{T}}\boldsymbol{\Phi}_1\right).$$

Noting that $\boldsymbol{\Lambda}_1$ is a diagonal matrix, we have

$$\begin{aligned}&\operatorname{trace}\left(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1^{\mathrm{T}}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2^{\mathrm{T}}\boldsymbol{\Phi}_1\right) \\ &\geq \min_i \boldsymbol{\Lambda}_1[i,i]\cdot\operatorname{trace}\left(\boldsymbol{\Phi}_1^{\mathrm{T}}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2^{\mathrm{T}}\boldsymbol{\Phi}_1\right) \\ &= \sigma_n(\boldsymbol{A})\operatorname{trace}(\boldsymbol{B}).\end{aligned}$$

The other direction follows similarly. □

Corollary 2: For any two matrices $\boldsymbol{A} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{B} \in \mathbb{R}^{r\times r}$, we have

$$\sigma_r(\boldsymbol{B})\|\boldsymbol{A}\|_F \leq \|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{B}\|\|\boldsymbol{A}\|_F.$$

We provide one more result before proceeding to prove the main theorem.

Lemma 8: Suppose $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times r}$ such that $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{A} \succeq \boldsymbol{0}$ is PSD. If $\|\boldsymbol{A}-\boldsymbol{B}\| \leq \frac{\sqrt{2}}{2}\sigma_r(\boldsymbol{B})$, we have

$$\begin{aligned}&\underbrace{\left\langle\left(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} - \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right)\boldsymbol{A}, \boldsymbol{A}-\boldsymbol{B}\right\rangle}_{(\aleph_1)} \\ &\geq \frac{1}{16}(\underbrace{\operatorname{trace}((\boldsymbol{A}-\boldsymbol{B})^{\mathrm{T}}(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})}_{(\aleph_2)} + \underbrace{\|\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}-\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\|_F^2}_{(\aleph_3)}).\end{aligned} \tag{57}$$

Proof: Denote $\boldsymbol{E} = \boldsymbol{A} - \boldsymbol{B}$. We first rewrite the terms $(\aleph_1)$, $(\aleph_2)$ and $(\aleph_3)$ as follows

$$\begin{aligned}(\aleph_1) &= \operatorname{trace}\left(\left(\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)^2 + 3\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{B} + \left(\boldsymbol{E}^{\mathrm{T}}\boldsymbol{B}\right)^2 \right. \\ &\qquad \left. + \boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\right), \\ (\aleph_2) &= \operatorname{trace}\left(\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\right), \\ (\aleph_3) &= \operatorname{trace}\left(\left(\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)^2 + 4\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{B} + 2\left(\boldsymbol{E}^{\mathrm{T}}\boldsymbol{B}\right)^2 \right. \\ &\qquad \left. + 2\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\right),\end{aligned}$$

where $\boldsymbol{E}^{\mathrm{T}}\boldsymbol{B} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{B} - \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{E}$. Now we have

$$\begin{aligned}&(\aleph_1) - \frac{1}{16}(\aleph_2) - \frac{1}{16}(\aleph_3) \\ &= \operatorname{trace}\left(\frac{15}{16}\left(\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)^2 + \frac{11}{4}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{B} + \frac{7}{8}\left(\boldsymbol{E}^{\mathrm{T}}\boldsymbol{B}\right)^2 \right. \\ &\qquad \left. + \frac{13}{16}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\right) \\ &= \left\|\sqrt{\frac{121}{56}}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E} + \sqrt{\frac{7}{8}}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{B}\right\|_F^2 \\ &\quad + \operatorname{trace}\left(\frac{13}{16}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} - \frac{137}{112}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right) \\ &\geq \operatorname{trace}\left(\frac{13}{16}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\sigma_r^2(\boldsymbol{B}) - \frac{137}{112}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\|\boldsymbol{E}\|^2\right) \\ &\geq \operatorname{trace}\left(\left(\frac{13}{16} - \frac{137}{112}\frac{1}{2}\right)\sigma_r^2(\boldsymbol{B})\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right) \\ &\geq 0,\end{aligned}$$

where the third line follows from Lemma 7 and the fourth line holds because by assumption $\|\boldsymbol{E}\| \leq \frac{\sqrt{2}}{2}\sigma_r(\boldsymbol{B})$. □

Now we turn to prove the main results. Recall that $\mu = \frac{1}{2}$ throughout the proof.

A. Regularity Condition for the Region $\mathcal{R}_1$

It follows from Lemma 3 that $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}^\star\boldsymbol{R} = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{W}^{\star\mathrm{T}}\boldsymbol{W}$ is PSD, where $\boldsymbol{R} = \arg\min_{R'\in\mathbb{O}_r}\|\boldsymbol{W} - \boldsymbol{W}^\star\boldsymbol{R}'\|_F^2$. We first perform the change of variable $\boldsymbol{W}^\star\boldsymbol{R} \to \boldsymbol{W}^\star$ to avoid $\boldsymbol{R}$ in the following equations. With this change of variable we have instead $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}^\star = \boldsymbol{W}^{\star\mathrm{T}}\boldsymbol{W}$ is PSD. We now rewrite the gradient $\nabla g(\boldsymbol{W})$ as follows:

$$\begin{aligned}\nabla g(\boldsymbol{W}) &= \begin{bmatrix}\boldsymbol{0} & \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}} \\ \boldsymbol{V}\boldsymbol{U}^{\mathrm{T}} - \boldsymbol{V}^\star\boldsymbol{U}^{\star\mathrm{T}} & \boldsymbol{0}\end{bmatrix}\boldsymbol{W} \\ &\quad + \mu\widehat{\boldsymbol{W}}(\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W}) \\ &= \frac{1}{2}\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{W} + \frac{1}{2}\widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W} \\ &\quad + (\mu - \frac{1}{2})\widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W} \\ &= \frac{1}{2}\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{W} + \frac{1}{2}\widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}.\end{aligned} \tag{58}$$

Plugging this into the left hand side of (40) gives

$$\begin{aligned}&\langle\nabla g(\boldsymbol{W}), \boldsymbol{W} - \boldsymbol{W}^\star\rangle \\ &= \frac{1}{2}\left\langle\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{W}, \boldsymbol{W} - \boldsymbol{W}^\star\right\rangle \\ &\quad + \frac{1}{2}\left\langle\widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}, \boldsymbol{W} - \boldsymbol{W}^\star\right\rangle \\ &= \frac{1}{2}\left\langle\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{W}, \boldsymbol{W} - \boldsymbol{W}^\star\right\rangle \\ &\quad + \frac{1}{2}\left\langle\widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\right\rangle\end{aligned} \tag{59}$$

where the last line follows from the fact that $\boldsymbol{W}^{\star \mathrm{T}}\widehat{\boldsymbol{W}}^{\star}=\mathbf{0}$. We first show the first term in the right hand side of the above equation is sufficiently large

$$\begin{aligned}
&\left\langle\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}-\boldsymbol{W}^{\star}\boldsymbol{W}^{\star \mathrm{T}}\right)\boldsymbol{W},\boldsymbol{W}-\boldsymbol{W}^{\star}\right\rangle\\
&\geq \frac{1}{16}\operatorname{trace}\left((\boldsymbol{W}-\boldsymbol{W}^{\star})^{\mathrm{T}}(\boldsymbol{W}-\boldsymbol{W}^{\star})\boldsymbol{W}^{\star \mathrm{T}}\boldsymbol{W}^{\star}\right)\\
&+\frac{1}{16}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}-\boldsymbol{W}^{\star}\boldsymbol{W}^{\star \mathrm{T}}\right\|_F^2\\
&\geq \frac{1}{16}\sigma_r(\boldsymbol{W}^{\star \mathrm{T}}\boldsymbol{W}^{\star})\left\|\boldsymbol{W}-\boldsymbol{W}^{\star}\right\|_F^2\\
&\quad+\frac{1}{16}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}-\boldsymbol{W}^{\star}\boldsymbol{W}^{\star \mathrm{T}}\right\|_F^2\\
&=\frac{1}{8}\sigma_r(\boldsymbol{X}^{\star})\left\|\boldsymbol{W}-\boldsymbol{W}^{\star}\right\|_F^2+\frac{1}{16}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}-\boldsymbol{W}^{\star}\boldsymbol{W}^{\star \mathrm{T}}\right\|_F^2,
\end{aligned} \tag{60}$$

where the first inequality follows from Lemma 8 since $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}^{\star}=\boldsymbol{W}^{\star \mathrm{T}}\boldsymbol{W}$ is PSD and $\|\boldsymbol{W}-\boldsymbol{W}^{\star}\|\leq \sigma_r^{1/2}(\boldsymbol{X}^{\star})=\frac{\sqrt{2}}{2}\sigma_r(\boldsymbol{W}^{\star})$, the second inequality follows from Lemma 7, and the last line holds because $\sigma_r\left(\widehat{\boldsymbol{W}}^{\star \mathrm{T}}\widehat{\boldsymbol{W}}^{\star}\right)=\sigma_r\left(\widehat{\boldsymbol{U}}^{\star \mathrm{T}}\widehat{\boldsymbol{U}}^{\star}+\widehat{\boldsymbol{V}}^{\star \mathrm{T}}\widehat{\boldsymbol{V}}^{\star}\right)=2\sigma_r\left(\boldsymbol{\Sigma}\right)=2\sigma_r\left(\boldsymbol{X}^{\star}\right)$. We then show the second term in the right hand side of (59) is lower bounded by

$$\begin{aligned}
&\left\langle\widehat{\boldsymbol{W}}^{\star}\widehat{\boldsymbol{W}}^{\star \mathrm{T}},\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\right\rangle\\
&=\frac{1}{2\left\|\boldsymbol{X}^{\star}\right\|}\left\|\widehat{\boldsymbol{W}}^{\star \mathrm{T}}\widehat{\boldsymbol{W}}^{\star}\right\|\operatorname{trace}\left(\widehat{\boldsymbol{W}}^{\star \mathrm{T}}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\widehat{\boldsymbol{W}}^{\star}\right)\\
&\geq\frac{1}{2\left\|\boldsymbol{X}^{\star}\right\|}\operatorname{trace}\left(\widehat{\boldsymbol{W}}^{\star \mathrm{T}}\widehat{\boldsymbol{W}}^{\star}\widehat{\boldsymbol{W}}^{\star \mathrm{T}}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\widehat{\boldsymbol{W}}^{\star}\right)\\
&=\frac{1}{2\left\|\boldsymbol{X}^{\star}\right\|}\left\|\widehat{\boldsymbol{W}}^{\star}\widehat{\boldsymbol{W}}^{\star \mathrm{T}}\boldsymbol{W}\right\|_F^2
\end{aligned} \tag{61}$$

where the first line holds because $\left\|\widehat{\boldsymbol{W}}^{\star \mathrm{T}}\widehat{\boldsymbol{W}}^{\star}\right\|=\left\|\widehat{\boldsymbol{U}}^{\star \mathrm{T}}\widehat{\boldsymbol{U}}^{\star}+\widehat{\boldsymbol{V}}^{\star \mathrm{T}}\widehat{\boldsymbol{V}}^{\star}\right\|=2\left\|\boldsymbol{\Sigma}\right\|=2\left\|\boldsymbol{X}^{\star}\right\|$, and the inequality follows from Lemma 7.

On the other hand, we attempt to control the gradient of $g(\boldsymbol{W})$. To that end, it follows from (58) that

$$\begin{aligned}
&\|\nabla g(\boldsymbol{W})\|_F^2\\
&=\frac{1}{4}\left\|\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}-\boldsymbol{W}^{\star}\boldsymbol{W}^{\star \mathrm{T}}\right)\boldsymbol{W}+\widehat{\boldsymbol{W}}^{\star}\widehat{\boldsymbol{W}}^{\star \mathrm{T}}\boldsymbol{W}\right\|_F^2\\
&\leq\frac{12}{47}\left\|\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}-\boldsymbol{W}^{\star}\boldsymbol{W}^{\star \mathrm{T}}\right)\boldsymbol{W}\right\|_F^2+12\left\|\widehat{\boldsymbol{W}}^{\star}\widehat{\boldsymbol{W}}^{\star \mathrm{T}}\boldsymbol{W}\right\|_F^2\\
&\leq\frac{12}{47}\|\boldsymbol{W}\|^2\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}-\boldsymbol{W}^{\star}\boldsymbol{W}^{\star \mathrm{T}}\right\|_F^2+12\left\|\widehat{\boldsymbol{W}}^{\star}\widehat{\boldsymbol{W}}^{\star \mathrm{T}}\boldsymbol{W}\right\|_F^2,
\end{aligned} \tag{62}$$

where the first inequality holds since $(a+b)^2\leq\frac{1+\epsilon}{\epsilon}a^2+(1+\epsilon)b^2$ for any $\epsilon>0$.

Combining (59)-(62), we can conclude the proof of (40) as long as we can show the following inequality:

$$\begin{aligned}
&\frac{1}{8}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}-\boldsymbol{W}^{\star}\boldsymbol{W}^{\star \mathrm{T}}\right\|_F^2\\
&\geq\frac{1}{47}\frac{\|\boldsymbol{W}\|^2}{\|\boldsymbol{X}^{\star}\|}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}-\boldsymbol{W}^{\star}\boldsymbol{W}^{\star \mathrm{T}}\right\|_F^2.
\end{aligned}$$

To that end, we upper bound $\|\boldsymbol{W}\|$ as follows:

$$\begin{aligned}
\|\boldsymbol{W}\|&\leq\|\boldsymbol{W}^{\star}\|+\|\boldsymbol{W}-\boldsymbol{W}^{\star}\|\\
&\leq\sqrt{2}\sigma_1^{1/2}(\boldsymbol{X}^{\star})+\|\boldsymbol{W}-\boldsymbol{W}^{\star}\|_F\\
&\leq(\sqrt{2}+1)\sigma_1^{1/2}(\boldsymbol{X}^{\star})
\end{aligned}$$

since $\|\boldsymbol{W}^{\star}\|=\sqrt{2}\sigma_1^{(1/2)}(\boldsymbol{X}^{\star})$ and $\operatorname{dist}(\boldsymbol{W},\boldsymbol{W}^{\star})\leq\sigma_r^{(1/2)}(\boldsymbol{X}^{\star})$. This completes the proof of (40).

B. Negative Curvature for the Region $\mathcal{R}_2$

To show (41), we utilize a strategy similar to that used in Appendix H for proving the strict saddle property of $g(\boldsymbol{W})$ by constructing a direction $\boldsymbol{\Delta}$ such that the Hessian evaluated at $\boldsymbol{W}$ along this direction is negative. For this purpose, denote

$$\boldsymbol{Q}=\begin{bmatrix}\boldsymbol{\Phi}/\sqrt{2}\\ \boldsymbol{\Psi}/\sqrt{2}\end{bmatrix}, \tag{63}$$

where we recall that $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ consist of the left and right singular vectors of $\boldsymbol{X}^{\star}$, respectively. The optimal solution $\boldsymbol{W}^{\star}$ has a compact SVD $\boldsymbol{W}^{\star}=\boldsymbol{Q}(\sqrt{2}\boldsymbol{\Sigma}^{1/2})\boldsymbol{R}$. For notational convenience, we denote $\overline{\boldsymbol{\Sigma}}=2\boldsymbol{\Sigma}$, where $\overline{\boldsymbol{\Sigma}}$ is a diagonal matrix whose diagonal entries in the upper left corner are $\overline{\sigma}_1,\ldots,\overline{\sigma}_r$.

For any $\boldsymbol{W}$, we can always divide it into two parts, the projections onto the column spaces of $\boldsymbol{Q}$ and its orthogonal complement, respectively. Equivalently, we can write

$$\boldsymbol{W}=\boldsymbol{Q}\overline{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{R}+\boldsymbol{E}, \tag{64}$$

where $\boldsymbol{Q}\overline{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{R}$ is a compact SVD form representing the projection of $\boldsymbol{W}$ onto the column space of $\boldsymbol{Q}$, and $\boldsymbol{E}^{\mathrm{T}}\boldsymbol{Q}=\mathbf{0}$ (i.e., $\boldsymbol{E}$ is orthogonal to $\boldsymbol{Q}$). Here $\boldsymbol{R}\in\mathcal{O}_r$ and $\overline{\boldsymbol{\Lambda}}$ is a diagonal matrix whose diagonal entries in the upper left corner are $\overline{\lambda}_1,\ldots,\overline{\lambda}_r$, but the diagonal entries are not necessarily placed either in decreasing or increasing order. In order to characterize the neighborhood near all strict saddles $\mathcal{C}\setminus\mathcal{X}$, we consider $\boldsymbol{W}$ such that $\sigma_r(\boldsymbol{W})\leq\sqrt{\frac{3}{8}}\sigma_r^{1/2}(\boldsymbol{X}^{\star})$. Let $k:=\arg\min_i\overline{\lambda}_i$ denote the location of the smallest diagonal entry in $\overline{\boldsymbol{\Lambda}}$. It is clear that

$$\overline{\lambda}_k\leq\sigma_r^2(\boldsymbol{W})\leq\frac{3}{8}\sigma_r(\boldsymbol{X}^{\star}). \tag{65}$$

Let $\boldsymbol{\alpha}\in\mathbb{R}^r$ be the eigenvector associated with the smallest eigenvalue of $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}$.

Recall that $\mu=\frac{1}{2}$. We show that the function $g(\boldsymbol{W})$ at $\boldsymbol{W}$ has directional negative curvature along the direction

$$\boldsymbol{\Delta}=\boldsymbol{q}_k\boldsymbol{\alpha}^{\mathrm{T}}. \tag{66}$$

We repeat the Hessian evaluated at $\boldsymbol{W}$ for $\boldsymbol{\Delta}$ as follows

$$\begin{aligned}
&[\nabla^2 g(\boldsymbol{W})](\boldsymbol{\Delta},\boldsymbol{\Delta})\\
&=\underbrace{\left\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}}+\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2}_{\Pi_1}+2\underbrace{\left\langle\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}-\boldsymbol{X}^{\star},\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\rangle}_{\Pi_2}\\
&\quad+\frac{1}{2}\underbrace{\left\langle\widehat{\boldsymbol{\Delta}}\widehat{\boldsymbol{W}}^{\mathrm{T}},\boldsymbol{\Delta}\boldsymbol{W}^{\mathrm{T}}\right\rangle}_{\Pi_3}+\frac{1}{2}\underbrace{\left\langle\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Delta}}^{\mathrm{T}},\boldsymbol{\Delta}\boldsymbol{W}^{\mathrm{T}}\right\rangle}_{\Pi_4}\\
&\quad+\frac{1}{2}\underbrace{\left\langle\widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}},\boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\rangle}_{\Pi_5}.
\end{aligned}$$

The remaining part is to bound the five terms.

Bounding terms Π_1, Π_3 and Π_4: We first rewrite these three terms:

$$\Pi_1 = \|\boldsymbol{\Delta_U} \boldsymbol{V}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\|_F^2 + 2\left\langle \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}, \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}}\right\rangle,$$

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

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

which implies

$$\begin{aligned}&\Pi_1 + \frac{1}{2}\Pi_3 + \frac{1}{2}\Pi_4\\ &\leq \|\boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{\Delta_U}\boldsymbol{U}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{\Delta_V}\boldsymbol{V}^{\mathrm{T}}\|_F^2\\ &\quad - \frac{1}{2}\|\boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}}\|_F^2 - \frac{1}{2}\|\boldsymbol{\Delta_V}\boldsymbol{U}^{\mathrm{T}}\|_F^2 + \left\langle \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}, \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}}\right\rangle\\ &= \|\boldsymbol{W}\boldsymbol{\Delta}^{\mathrm{T}}\|_F^2 - \frac{1}{2}\left\|\boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2\\ &\leq \|\boldsymbol{W}\boldsymbol{\Delta}^{\mathrm{T}}\|_F^2. \end{aligned}\tag{67}$$

Noting that $\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\Delta} = \boldsymbol{\alpha}\boldsymbol{q}_k^{\mathrm{T}}\boldsymbol{q}_k\boldsymbol{\alpha}^{\mathrm{T}} = \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}}$, we now compute $\|\boldsymbol{W}\boldsymbol{\Delta}^{\mathrm{T}}\|_F^2$ as

$$\begin{aligned}\|\boldsymbol{W}\boldsymbol{\Delta}^{\mathrm{T}}\|_F^2 &= \operatorname{trace}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\Delta}\right) = \operatorname{trace}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}}\right)\\ &= \sigma_r^2(\boldsymbol{W}).\end{aligned}$$

Plugging this into (67) gives

$$\Pi_1 + \frac{1}{2}\Pi_3 + \frac{1}{2}\Pi_4 \leq \sigma_r^2(\boldsymbol{W}). \tag{68}$$

Bounding terms Π_2 and Π_5: To obtain an upper bound for the term Π_2, we first rewrite it as follows

$$\begin{aligned}\Pi_2 &= \left\langle \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star, \boldsymbol{\Delta_U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\rangle\\ &= \frac{1}{2}\left\langle \begin{bmatrix} \boldsymbol{0} & \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}}\\ \boldsymbol{V}\boldsymbol{U}^{\mathrm{T}} - \boldsymbol{V}^\star\boldsymbol{U}^{\star\mathrm{T}} & \boldsymbol{0}\end{bmatrix}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\rangle\\ &= \frac{1}{4}\left\langle \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\rangle - \frac{1}{4}\left\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\rangle\\ &\quad + \frac{1}{4}\left\langle \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\rangle.\end{aligned}$$

We then have

$$\begin{aligned}2\Pi_2 + \frac{1}{2}\Pi_5 &= \frac{1}{2}\left\langle \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\rangle\\ &\quad + \frac{1}{2}\left\langle \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\rangle.\end{aligned}\tag{69}$$

To bound these two terms in the above equation, we note that

$$\boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}} = \sum_{i=1}^{r}\alpha_i^2\boldsymbol{q}_k\boldsymbol{q}_k^{\mathrm{T}} = \boldsymbol{q}_k\boldsymbol{q}_k^{\mathrm{T}} = \frac{1}{2}\begin{bmatrix}\boldsymbol{\phi}_k\boldsymbol{\phi}_k^{\mathrm{T}} & \boldsymbol{\phi}_k\boldsymbol{\psi}_k^{\mathrm{T}}\\ \boldsymbol{\psi}_k\boldsymbol{\phi}_k^{\mathrm{T}} & \boldsymbol{\psi}_k\boldsymbol{\psi}_k^{\mathrm{T}}\end{bmatrix}.$$

Then we have

$$\begin{aligned}&\left\langle \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\rangle\\ &= \frac{1}{2}\left\langle \begin{bmatrix}\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}} & -\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{T}}\\ -\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}} & \boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{T}}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\phi}_k\boldsymbol{\phi}_k^{\mathrm{T}} & \boldsymbol{\phi}_k\boldsymbol{\psi}_k^{\mathrm{T}}\\ \boldsymbol{\psi}_k\boldsymbol{\phi}_k^{\mathrm{T}} & \boldsymbol{\psi}_k\boldsymbol{\psi}_k^{\mathrm{T}}\end{bmatrix}\right\rangle = 0,\end{aligned}$$

and

$$\begin{aligned}&\left\langle \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}, \boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}}\right\rangle\\ &= \left\langle \boldsymbol{Q}\overline{\boldsymbol{\Lambda}}\boldsymbol{Q}^{\mathrm{T}} - 2\boldsymbol{Q}\overline{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{R}\boldsymbol{E}^{\mathrm{T}} + \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}} - \boldsymbol{Q}\overline{\boldsymbol{\Sigma}}\boldsymbol{Q}^{\mathrm{T}}, \boldsymbol{q}_k\boldsymbol{q}_k^{\mathrm{T}}\right\rangle\\ &= \overline{\lambda}_k - \overline{\sigma}_k\end{aligned}$$

where the last utilizes the fact that $\boldsymbol{E}^{\mathrm{T}}\boldsymbol{q}_k = \boldsymbol{0}$ since $\boldsymbol{E}$ is orthogonal to $\boldsymbol{Q}$.

Plugging these into (69) gives

$$2\Pi_2 + \frac{1}{2}\Pi_5 = \frac{1}{2}(\overline{\lambda}_k - \overline{\sigma}_k). \tag{70}$$

Merging together: Putting (68) and (70) together yields

$$\begin{aligned}[\nabla^2 g(\boldsymbol{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) &= \Pi_1 + \frac{1}{2}\Pi_3 + \frac{1}{2}\Pi_4 + 2\Pi_2 + \frac{1}{2}\Pi_5\\ &\leq \sigma_r^2(\boldsymbol{W}) + \frac{1}{2}(\overline{\lambda}_k - \overline{\sigma}_k)\\ &\leq \frac{1}{2}\sigma_r(\boldsymbol{X}^\star) + \frac{1}{2}(\frac{1}{2}\sigma_r(\boldsymbol{X}^\star) - 2\sigma_r(\boldsymbol{X}^\star))\\ &\leq -\frac{1}{4}\sigma_r(\boldsymbol{X}^\star),\end{aligned}$$

where the third line follows because by assumption $\sigma_r(\boldsymbol{W}) \leq \sqrt{\frac{1}{2}\sigma_r^{1/2}(\boldsymbol{X}^\star)}$, by construction $\overline{\lambda}_k \leq \frac{1}{2}\sigma_r(\boldsymbol{X}^\star)$ (see (65)), and $\overline{\sigma}_k \geq \overline{\sigma}_r = 2\sigma_r(\boldsymbol{X}^\star)$. This completes the proof of (41).

C. Large Gradient for the Region $\mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$

In order to show that $g(\boldsymbol{W})$ has a large gradient in the three regions $\mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$, we first provide a lower bound for the gradient. By (58), we have

$$\begin{aligned}&\|\nabla g(\boldsymbol{W})\|_F^2\\ &= \frac{1}{4}\left\|\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{W} + \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}\right\|_F^2\\ &= \frac{1}{4}\left(\left\|\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{W}\right\|_F^2 + \left\|\widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}\right\|_F^2\right)\\ &\quad + \frac{1}{2}\left\langle\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{W}, \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}\right\rangle\\ &= \frac{1}{4}\left(\left\|\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{W}\right\|_F^2 + \left\|\widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}\right\|_F^2\right)\\ &\quad + \frac{1}{2}\left\langle\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}, \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\right\rangle\\ &\geq \frac{1}{4}\left\|\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{W}\right\|_F^2,\end{aligned}\tag{71}$$

where the third equality follows because $\boldsymbol{W}^{\star\mathrm{T}}\widehat{\boldsymbol{W}}^\star = \boldsymbol{U}^{\star\mathrm{T}}\boldsymbol{U}^\star - \boldsymbol{V}^{\star\mathrm{T}}\boldsymbol{V}^\star = \boldsymbol{0}$ and the last line utilizes the fact that the inner product between two PSD matrices is nonnegative.

1) Large Gradient for the Region $\mathcal{R}_3'$: To show $\|\nabla g(\boldsymbol{W})\|_F^2$ is large for any $\boldsymbol{W} \in \mathcal{R}_3'$, again, for any $\boldsymbol{W} \in \mathbb{R}^{(n+m)\times r}$, we utilize (64) to write $\boldsymbol{W} = \boldsymbol{Q}\overline{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{R} + \boldsymbol{E}$, where $\boldsymbol{Q}$ is defined in (63), $\boldsymbol{Q}\overline{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{R}$ is a compact SVD form representing the projection of $\boldsymbol{W}$ onto the column space of $\boldsymbol{Q}$, and $\boldsymbol{E}^{\mathrm{T}}\boldsymbol{Q} = \boldsymbol{0}$ (i.e., $\boldsymbol{E}$ is orthogonal to $\boldsymbol{Q}$). Plugging this form of $\boldsymbol{W}$ into the last term of (71) gives

$$\begin{aligned}&\left\|\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{W}\right\|_F^2 =\\ &\left\|\boldsymbol{Q}\overline{\boldsymbol{\Lambda}}^{1/2}(\overline{\boldsymbol{\Lambda}} - \overline{\boldsymbol{\Sigma}})\boldsymbol{R} + \boldsymbol{Q}\overline{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{R}\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}} + \boldsymbol{E}\boldsymbol{R}^{\mathrm{T}}\overline{\boldsymbol{\Lambda}}\boldsymbol{R} + \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right\|_F^2\end{aligned}$$

$$= \left\| Q\overline{\Lambda}^{1/2}(\overline{\Lambda} - \overline{\Sigma})R + Q\overline{\Lambda}^{1/2}REE^{\mathrm{T}} \right\|_F^2 + \left\| ER^{\mathrm{T}}\overline{\Lambda}R + EE^{\mathrm{T}}E \right\|_F^2 \tag{72}$$

since Q is orthogonal to E. The remaining part is to show at least one of the two terms is large for any $W \in \mathcal{R}_3'$ by considering the following two cases.

Case I: $\|E\|_F^2 \geq \frac{4}{25}\sigma_r(X^\star)$. As E is large, we bound the second term in (72):

$$\begin{aligned} \left\| ER^{\mathrm{T}}\overline{\Lambda}R + EE^{\mathrm{T}}E \right\|_F^2 &\geq \sigma_r^2\left(R^{\mathrm{T}}\overline{\Lambda}R + E^{\mathrm{T}}E\right)\|E\|_F^2 \\ &= \sigma_r^4(W)\|E\|_F^2 \\ &\geq (\frac{1}{2})^2\frac{4}{25}\sigma_r^3(X^\star) = \frac{1}{25}\sigma_r^3(X^\star), \end{aligned} \tag{73}$$

where the first inequality follows from Corollary 2, the first equality follows from the fact $W^{\mathrm{T}}W = R^{\mathrm{T}}\overline{\Lambda}R + E^{\mathrm{T}}E$, and the last inequality holds because by assumption that $\sigma_r^2(W) \geq \frac{1}{2}\sigma_r(X^\star)$ and $\|E\|_F^2 \geq \frac{4}{25}\sigma_r(X^\star)$.

Case II: $\|E\|_F^2 \leq \frac{4}{25}\sigma_r(X^\star)$. In this case, we start by bounding the diagonal entries in $\overline{\Lambda}$. First, utilizing Weyl's inequality for perturbation of singular values [51, Theorem 3.3.16] gives

$$\left|\sigma_r(W) - \min_i \overline{\lambda}_i^{1/2}\right| \leq \|E\|_2,$$

which implies

$$\min_i \overline{\lambda}_i^{1/2} \geq \sigma_r(W) - \|E\|_2 \geq \sqrt{\frac{1}{2}}\sigma_r^{1/2}(X^\star) - \frac{2}{5}\sigma_r^{1/2}(X^\star), \tag{74}$$

where we utilize $\|E\|_2 \leq \|E\|_F \leq \frac{2}{5}\sigma_r^{1/2}(X^\star)$. On the other hand,

$$\begin{aligned} \operatorname{dist}(W, W^\star) &\leq \left\| Q(\overline{\Lambda}^{1/2} - \overline{\Sigma}^{1/2})R + E \right\|_F \\ &\leq \left\| Q(\overline{\Lambda}^{1/2} - \overline{\Sigma}^{1/2})R \right\|_F + \|E\|_F, \end{aligned}$$

which together with the assumption that $\operatorname{dist}(W, W^\star) \geq \sigma_r^{1/2}(X^\star)$ gives

$$\left\|\overline{\Lambda}^{1/2} - \overline{\Sigma}^{1/2}\right\|_F \geq \sigma_r^{1/2}(X^\star) - \frac{2}{5}\sigma_r^{1/2}(X^\star) = \frac{3}{5}\sigma_r^{1/2}(X^\star).$$

We now bound the first term in (72):

$$\begin{aligned} &\left\| Q\overline{\Lambda}^{1/2}(\overline{\Lambda} - \overline{\Sigma})R + Q\overline{\Lambda}^{1/2}REE^{\mathrm{T}} \right\|_F \\ &\geq \min_i \overline{\lambda}_i^{1/2}\left\| (\overline{\Lambda} - \overline{\Sigma})R + REE^{\mathrm{T}} \right\|_F \\ &\geq \min_i \overline{\lambda}_i^{1/2}\left(\|(\overline{\Lambda} - \overline{\Sigma})R\|_F - \left\|REE^{\mathrm{T}}\right\|_F \right) \\ &\geq (\sqrt{\frac{1}{2}} - \frac{2}{5})\left(\left(\sqrt{2} + \sqrt{\frac{1}{2} - \frac{2}{5}}\right)\frac{3}{5} - \frac{4}{25}\right)\sigma_r^{3/2}(X^\star) \end{aligned} \tag{75}$$

where the third line holds because $\|EE^{\mathrm{T}}\|_F \leq \|E\|_F^2 \leq \frac{4}{25}\sigma_r(X^\star)$, $\min_i \overline{\lambda}_i^{1/2} \geq \left(\sqrt{\frac{1}{2}} - \frac{2}{5}\right)\sigma_r^{1/2}(X^\star)$ by (74), and

$$\begin{aligned} \|\overline{\Lambda} - \overline{\Sigma}\|_F &= \sqrt{\sum_{i=1}^r (\overline{\sigma}_i - \overline{\lambda}_i)^2} \\ &= \sqrt{\sum_{i=1}^r \left(\overline{\sigma}_i^{1/2} - \overline{\lambda}_i^{1/2}\right)^2\left(\overline{\sigma}_i^{1/2} + \overline{\lambda}_i^{1/2}\right)^2} \\ &\geq \left(\overline{\sigma}_r^{1/2} + \min_i \overline{\lambda}_i^{1/2}\right)\sqrt{\sum_{i=1}^r \left(\overline{\sigma}_i^{1/2} - \overline{\lambda}_i^{1/2}\right)^2} \\ &= \left(\overline{\sigma}_r^{1/2} + \min_i \overline{\lambda}_i^{1/2}\right)\left\|\overline{\Lambda}^{1/2} - \overline{\Sigma}^{1/2}\right\|_F \\ &\geq \left(\sqrt{2} + \sqrt{\frac{1}{2} - \frac{2}{5}}\right)\frac{3}{5}\sigma_r(X^\star). \end{aligned}$$

Combining (71) with (72), (73) and (75) gives

$$\|\nabla g(W)\|_F \geq \frac{1}{10}\sigma_r^{3/2}(X^\star).$$

This completes the proof of (42).

2) Large Gradient for the Region $\mathcal{R}_3''$: By (71), we have

$$\|\nabla g(W)\|_F \geq \frac{1}{2}\left\|\left(WW^{\mathrm{T}} - W^\star W^{\star\mathrm{T}}\right)W\right\|_F^2.$$

Now (43) follows directly from the fact $\|W\| > \frac{20}{19}\|W^\star\|$ and the following result.

Lemma 9: For any $A, B \in \mathbb{R}^{n\times r}$ with $\|A\| \geq \alpha\|B\|$ and $\alpha > 1$, we have

$$\left\|\left(AA^{\mathrm{T}} - BB^{\mathrm{T}}\right)A\right\|_F \geq (1 - \frac{1}{\alpha^2})\|A\|^3.$$

Proof: Let $A = \Phi_1\Lambda_1R_1^{\mathrm{T}}$ and $B = \Phi_2\Lambda_2R_2^{\mathrm{T}}$ be the SVDs of A and B, respectively. Then

$$\begin{aligned} \left\|\left(AA^{\mathrm{T}} - BB^{\mathrm{T}}\right)A\right\|_F &= \left\|\Phi_1\Lambda_1^3 - \Phi_2\Lambda_2^2\Phi_2^{\mathrm{T}}\Phi_1\Lambda_1\right\|_F \\ &\geq \left\|\Lambda_1^3 - \Phi_1^{\mathrm{T}}\Phi_2\Lambda_2^2\Phi_2^{\mathrm{T}}\Phi_1\Lambda_1\right\|_F \\ &\geq \left\|\Lambda_1^3 - \Lambda_2^2\Lambda_1\right\|_F \\ &\geq (1 - \frac{1}{\alpha^2})\|A\|^3. \end{aligned}$$

□

3) Large Gradient for the Region $\mathcal{R}_3'''$: By (58), we have

$$\begin{aligned} &\langle\nabla g(W), W\rangle \\ &= \left\langle \frac{1}{2}\left(WW^{\mathrm{T}} - W^\star W^{\star\mathrm{T}}\right)W + \frac{1}{2}\widehat{W}^\star\widehat{W}^{\star\mathrm{T}}W, W \right\rangle \\ &\geq \frac{1}{2}\left\langle\left(WW^{\mathrm{T}} - W^\star W^{\star\mathrm{T}}\right)W, W\right\rangle \\ &\geq \frac{1}{2}\left(\left\|WW^{\mathrm{T}}\right\|_F^2 - \left\|WW^{\mathrm{T}}\right\|_F\left\|W^\star W^{\star\mathrm{T}}\right\|_F\right) \\ &> \frac{1}{20}\left\|WW^{\mathrm{T}}\right\|_F^2 \end{aligned} \tag{76}$$

where the last line holds because $\|W^\star W^{\star\mathrm{T}}\|_F < \frac{9}{10}\|WW^{\mathrm{T}}\|_F$.

APPENDIX L
PROOF OF THEOREM 1 (ROBUST STRICT SADDLE FOR $G(W)$)

Throughout the proofs, we always utilize $X = UV^{\mathrm{T}}$ unless stated otherwise. To give a sense that the geometric result in

Theorem 5 for $g(\boldsymbol{W})$ is also possibly preserved for $G(\boldsymbol{W})$, we first compute the derivative of $G(\boldsymbol{W})$ as

$$\nabla G(\boldsymbol{W}) = \begin{bmatrix} \nabla f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{V} \\ (\nabla f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}))^{\mathrm{T}}\boldsymbol{U} \end{bmatrix} + \mu \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{W}. \tag{77}$$

For any $\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\Delta_U} \\ \boldsymbol{\Delta_V} \end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$, algebraic calculation gives the Hessian quadratic form $[\nabla^2 G(\boldsymbol{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta})$ as

$$\begin{aligned} &[\nabla^2 G(\boldsymbol{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) \\ &= [\nabla^2 f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})](\boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}}, \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}}) \\ &\quad + 2\langle \nabla f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}), \boldsymbol{\Delta_U}\boldsymbol{\Delta_V}^{\mathrm{T}} \rangle + [\nabla^2 \rho(\boldsymbol{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) \end{aligned} \tag{78}$$

where $[\nabla^2 \rho(\boldsymbol{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta})$ is defined in (32). Thus, it is expected that $G(\boldsymbol{W})$, $\nabla G(\boldsymbol{W})$, and $\nabla^2 G(\boldsymbol{W})$ are close to their counterparts (i.e., $g(\boldsymbol{W})$, $\nabla g(\boldsymbol{W})$ and $\nabla^2 g(\boldsymbol{W})$) for the matrix factorization problem when $f(\boldsymbol{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (6).

Before moving to the main proofs, we provide several useful results regarding the deviations of the gradient and Hessian. We start with a useful characterization of the restricted strong convexity and smoothness condition.

Lemma 10: Suppose f satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (6) with positive constants $a = 1 - c$ and $b = 1 + c, c \in [0, 1)$. Then any $n \times m$ matrices $\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{H}$ with $\operatorname{rank}(\boldsymbol{C}), \operatorname{rank}(\boldsymbol{D}) \le r$ and $\operatorname{rank}(\boldsymbol{H}) \le 2r$, we have

$$|\langle \nabla f(\boldsymbol{C}) - \nabla f(\boldsymbol{D}) - (\boldsymbol{C} - \boldsymbol{D}), \boldsymbol{H} \rangle| \le c \|\boldsymbol{C} - \boldsymbol{D}\|_F \|\boldsymbol{H}\|_F.$$

Proof of Lemma 10: We first invoke [40, Proposition 2] which states that under Assumption 2 for any $n \times m$ matrices $\boldsymbol{Z}, \boldsymbol{D}, \boldsymbol{H}$ of rank at most $2r$, we have

$$\left| [\nabla^2 f(\boldsymbol{Z})](\boldsymbol{D}, \boldsymbol{H}) - \langle \boldsymbol{D}, \boldsymbol{H} \rangle \right| \le c \|\boldsymbol{D}\|_F \|\boldsymbol{H}\|_F. \tag{79}$$

Now using integral form of the mean value theorem for ∇f, we have

$$\begin{aligned} &|\langle \nabla f(\boldsymbol{C}) - \nabla f(\boldsymbol{D}) - (\boldsymbol{C} - \boldsymbol{D}), \boldsymbol{H} \rangle| \\ &= \left| \int_0^1 [\nabla^2 f(t\boldsymbol{C} + (1-t)\boldsymbol{D})] (\boldsymbol{C} - \boldsymbol{D}, \boldsymbol{H}) - \langle \boldsymbol{C} - \boldsymbol{D}, \boldsymbol{H} \rangle dt \right| \\ &\le \int_0^1 \left| [\nabla^2 f(t\boldsymbol{C} + (1-t)\boldsymbol{D})] (\boldsymbol{C} - \boldsymbol{D}, \boldsymbol{H}) - \langle \boldsymbol{C} - \boldsymbol{D}, \boldsymbol{H} \rangle \right| dt \\ &\le \int_0^1 c \|\boldsymbol{C} - \boldsymbol{D}\|_F \|\boldsymbol{H}\|_F dt = c \|\boldsymbol{C} - \boldsymbol{D}\|_F \|\boldsymbol{H}\|_F. \end{aligned}$$

where the second inequality follows from (79) since $t\boldsymbol{C} + (1-t)\boldsymbol{D}$, $\boldsymbol{C} - \boldsymbol{D}$, and $\boldsymbol{H}$ all are rank at most $2r$. □

The following result controls the deviation of the gradient between the general low-rank optimization (9) and the matrix factorization problem by utilizing the $(2r, 4r)$-restricted strong convexity and smoothness condition (6).

Lemma 11: Suppose $f(\boldsymbol{X})$ has a critical point $\boldsymbol{X}^\star \in \mathbb{R}^{n \times m}$ of rank r and satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (6) with positive constants $a = 1 - c$ and $b = 1 + c, c \in [0, 1)$. Then, we have

$$\|\nabla G(\boldsymbol{W}) - \nabla g(\boldsymbol{W})\|_F \le c \left\| \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star \boldsymbol{W}^{\star \mathrm{T}} \right\|_F \|\boldsymbol{W}\|.$$

Proof of Lemma 11: We bound the deviation directly:

$$\begin{aligned} &\|\nabla G(\boldsymbol{W}) - \nabla g(\boldsymbol{W})\|_F = \max_{\|\boldsymbol{\Delta}\|_F = 1} \langle \nabla G(\boldsymbol{W}) - \nabla g(\boldsymbol{W}), \boldsymbol{\Delta} \rangle \\ &= \max_{\|\boldsymbol{\Delta}\|_F = 1} \left\langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} \right\rangle - \left\langle \boldsymbol{X} - \boldsymbol{X}^\star, \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} \right\rangle \\ &\quad + \left\langle \nabla f(\boldsymbol{X}), \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\rangle - \left\langle \boldsymbol{X} - \boldsymbol{X}^\star, \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\rangle \\ &= \max_{\|\boldsymbol{\Delta}\|_F = 1} \left\langle \nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{X}^\star) - (\boldsymbol{X} - \boldsymbol{X}^\star), \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} \right\rangle \\ &\quad + \left\langle \nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{X}^\star) - (\boldsymbol{X} - \boldsymbol{X}^\star), \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\rangle \\ &\le \max_{\|\boldsymbol{\Delta}\|_F = 1} c \|\boldsymbol{X} - \boldsymbol{X}^\star\|_F \left(\left\| \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} \right\|_F + \left\| \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\|_F \right) \\ &\le c \|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star\|_F (\|\boldsymbol{V}\| + \|\boldsymbol{U}\|) \\ &\le c \|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star \boldsymbol{W}^{\star \mathrm{T}}\|_F \|\boldsymbol{W}\|, \end{aligned}$$

where the last equality follows from Assumption 1 that $\nabla f(\boldsymbol{X}^\star) = \boldsymbol{0}$ and and the first inequality utilizes Lemma 10. □

Similarly, the next result controls the deviation of the Hessian between the matrix sensing problem and the matrix factorization problem.

Lemma 12: Suppose $f(\boldsymbol{X})$ has a critical point $\boldsymbol{X}^\star \in \mathbb{R}^{n \times m}$ of rank r and satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (6) with positive constants $a = 1 - c$ and $b = 1 + c, c \in [0, 1)$. Then, for any $\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\Delta_U} \\ \boldsymbol{\Delta_V} \end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$ the following holds:

$$\begin{aligned} &\left| \nabla^2 G(\boldsymbol{W})[\boldsymbol{\Delta}, \boldsymbol{\Delta}] - \nabla^2 g(\boldsymbol{W})[\boldsymbol{\Delta}, \boldsymbol{\Delta}] \right| \\ &\le 2c \left\| \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star \right\|_F \left\| \boldsymbol{\Delta_U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\|_F + c \left\| \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\|_F^2. \end{aligned}$$

Proof of Lemma 12: First note that

$$\begin{aligned} &\nabla^2 G(\boldsymbol{W})[\boldsymbol{\Delta}, \boldsymbol{\Delta}] - \nabla^2 g(\boldsymbol{W})[\boldsymbol{\Delta}, \boldsymbol{\Delta}] \\ &= 2 \left\langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta_U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\rangle - 2 \left\langle \boldsymbol{X} - \boldsymbol{X}^\star, \boldsymbol{\Delta_U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\rangle \\ &\quad + [\nabla^2 f(\boldsymbol{X})](\boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}}) - \left\| \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\|_F^2. \end{aligned}$$

Now utilizing Lemma 10 and (6), we have

$$\begin{aligned} &\left| \nabla^2 G(\boldsymbol{W})[\boldsymbol{\Delta}, \boldsymbol{\Delta}] - \nabla^2 g(\boldsymbol{W})[\boldsymbol{\Delta}, \boldsymbol{\Delta}] \right| \\ &\le 2 \left| \left\langle \nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{X}^\star), \boldsymbol{\Delta_U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\rangle - \langle \boldsymbol{X} - \boldsymbol{X}^\star, \boldsymbol{\Delta_U}\boldsymbol{\Delta_V}^{\mathrm{T}} \rangle \right| \\ &\quad + \left| [\nabla^2 f(\boldsymbol{X})](\boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}}) - \left\| \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\|_F^2 \right| \\ &\le 2c \left\| \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star \right\|_F \left\| \boldsymbol{\Delta_U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\|_F \\ &\quad + c \left\| \boldsymbol{\Delta_U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta_V}^{\mathrm{T}} \right\|_F^2. \end{aligned}$$

□

We provide one more result before proceeding to prove the main theorem.

Lemma 13 ([13, Lemma E.1]): Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two $n \times r$ matrices such that $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}$ is PSD. Then

$$\left\| (\boldsymbol{A} - \boldsymbol{B}) \boldsymbol{A}^{\mathrm{T}} \right\|_F^2 \le \frac{1}{2(\sqrt{2} - 1)} \left\| \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} - \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} \right\|_F^2.$$

A. Local Descent Condition for the Region $\mathcal{R}_1$

Similar to what used in Appendix K-A, we perform the change of variable $\boldsymbol{W}^\star\boldsymbol{R} \to \boldsymbol{W}^\star$ to avoid $\boldsymbol{R}$ in the following equations. With this change of variable we have instead $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}^\star = \boldsymbol{W}^{\star\mathrm{T}}\boldsymbol{W}$ is PSD.

We first control $|\langle\nabla G(\boldsymbol{W}) - \nabla g(\boldsymbol{W}), \boldsymbol{W} - \boldsymbol{W}^\star\rangle|$ as follows:

$$\begin{aligned}
&|\langle\nabla G(\boldsymbol{W}) - \nabla g(\boldsymbol{W}), \boldsymbol{W} - \boldsymbol{W}^\star\rangle| \\
&\le \left|\langle\nabla f(\boldsymbol{X}), (\boldsymbol{U}-\boldsymbol{U}^\star)\boldsymbol{V}^{\mathrm{T}}\rangle - \langle\boldsymbol{X}-\boldsymbol{X}^\star, (\boldsymbol{U}-\boldsymbol{U}^\star)\boldsymbol{V}^{\mathrm{T}}\rangle\right| \\
&\quad + \left|\langle\nabla f(\boldsymbol{X}), \boldsymbol{U}(\boldsymbol{V}-\boldsymbol{V}^\star)^{\mathrm{T}}\rangle - \langle\boldsymbol{X}-\boldsymbol{X}^\star, \boldsymbol{U}(\boldsymbol{V}-\boldsymbol{V}^\star)^{\mathrm{T}}\rangle\right| \\
&\le c\,\|\boldsymbol{X}-\boldsymbol{X}^\star\|_F\left(\|(\boldsymbol{U}-\boldsymbol{U}^\star)\boldsymbol{V}^{\mathrm{T}}\|_F + \|\boldsymbol{U}(\boldsymbol{V}-\boldsymbol{V}^\star)^{\mathrm{T}}\|_F\right) \\
&\le c\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F\,\|\boldsymbol{W}(\boldsymbol{W}-\boldsymbol{W}^\star)^{\mathrm{T}}\|_F \\
&\le \frac{c}{2(\sqrt{2}-1)}\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F^2
\end{aligned}$$

where the second inequality utilizes $\nabla f(\boldsymbol{X}^\star) = \boldsymbol{0}$ and Lemma 10, and the last inequality follows from Lemma 13. The above result along with (59)-(60) gives

$$\begin{aligned}
&\langle\nabla G(\boldsymbol{W}), \boldsymbol{W} - \boldsymbol{W}^\star\rangle \\
&\ge \langle\nabla g(\boldsymbol{W}), \boldsymbol{W}-\boldsymbol{W}^\star\rangle - |\langle\nabla G(\boldsymbol{W}) - \nabla g(\boldsymbol{W}), \boldsymbol{W} - \boldsymbol{W}^\star\rangle| \\
&\ge \langle\nabla g(\boldsymbol{W}), \boldsymbol{W}-\boldsymbol{W}^\star\rangle - \frac{c}{2(\sqrt{2}-1)}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2 \\
&\ge \frac{1}{16}\sigma_r(\boldsymbol{X}^\star)\operatorname{dist}^2(\boldsymbol{W},\boldsymbol{W}^\star) + \frac{1}{32}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2 \\
&\quad + \frac{1}{4\|\boldsymbol{X}^\star\|}\left\|\widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}\right\|_F^2 \\
&\quad - \frac{c}{2(\sqrt{2}-1)}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2 \\
&\ge \frac{1}{16}\sigma_r(\boldsymbol{X}^\star)\operatorname{dist}^2(\boldsymbol{W},\boldsymbol{W}^\star) + \frac{1}{160}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2 \\
&\quad + \frac{1}{4\|\boldsymbol{X}^\star\|}\left\|\widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}\right\|_F^2 \qquad (80)
\end{aligned}$$

where we utilize $c \le \frac{1}{50}$.

On the other hand, we control $\|\nabla G(\boldsymbol{W})\|_F$ with Lemma 11 controlling the deviation between $\nabla G(\boldsymbol{W})$ and $\nabla g(\boldsymbol{W})$ as follows:

$$\begin{aligned}
&\|\nabla G(\boldsymbol{W})\|_F^2 = \|\nabla g(\boldsymbol{W}) + \nabla G(\boldsymbol{W}) - \nabla g(\boldsymbol{W})\|_F^2 \\
&\le \frac{20}{19}\|\nabla g(\boldsymbol{W})\|_F^2 + 20\,\|\nabla g(\boldsymbol{W}) - \nabla G(\boldsymbol{W})\|_F^2 \\
&\le \frac{20}{19}\|\nabla g(\boldsymbol{W})\|_F^2 + 20c^2\|\boldsymbol{W}\|^2\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2 \\
&= \frac{5}{19}\left\|\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{W} + \widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}\right\|_F^2 \\
&\quad + 20c^2\|\boldsymbol{W}\|^2\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F^2 \\
&\le \left(\frac{5}{19}\frac{100}{99} + 20c^2\right)\left\|\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right)\boldsymbol{W}\right\|_F^2 \\
&\quad + 25\left\|\widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}\right\|_F^2 \\
&\le (\frac{5}{19}\frac{100}{99} + 50c^2)(\sqrt{2}+1)^2\|\boldsymbol{X}^\star\|\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F^2 \\
&\quad + 25\|\widehat{\boldsymbol{W}}^\star\widehat{\boldsymbol{W}}^{\star\mathrm{T}}\boldsymbol{W}\|_F^2, \qquad (81)
\end{aligned}$$

where the first inequality holds since $(a+b)^2 \le \frac{1+\epsilon}{\epsilon}a^2 + (1+\epsilon)b^2$ for any $\epsilon > 0$, and the fourth line follows from (58).

Now combining (80)-(81) and assuming $c \le \frac{1}{50}$ gives

$$\begin{aligned}
&\langle\nabla G(\boldsymbol{W}), \boldsymbol{W} - \boldsymbol{W}^\star\rangle \\
&\ge \frac{1}{16}\sigma_r(\boldsymbol{X}^\star)\operatorname{dist}^2(\boldsymbol{W},\boldsymbol{W}^\star) + \frac{1}{260\|\boldsymbol{X}^\star\|}\|\nabla G(\boldsymbol{W})\|_F^2.
\end{aligned}$$

This completes the proof of (15).

B. Negative Curvature for the Region $\mathcal{R}_2$

Let $\boldsymbol{\Delta} = \boldsymbol{q}_k\boldsymbol{\alpha}^{\mathrm{T}}$ be defined as in (66). First note that

$$\begin{aligned}
\left\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2 &\le 2\left\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}}\right\|_F^2 + 2\left\|\boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2 \\
&\le 2\left\|\boldsymbol{W}\boldsymbol{\Delta}^{\mathrm{T}}\right\|_F^2 = 2\sigma_r^2(\boldsymbol{W}) \le \sigma_r(\boldsymbol{X}^\star),
\end{aligned}$$

where the last equality holds because $\sigma_r(\boldsymbol{W}) \le \sqrt{\frac{1}{2}}\sigma_r^{1/2}(\boldsymbol{X}^\star)$. Also utilizing the particular structure in $\boldsymbol{\Delta}$ yields

$$\left\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F = \frac{1}{2}\left\|\boldsymbol{\phi}_k\boldsymbol{\psi}_k^{\mathrm{T}}\right\|_F = \frac{1}{2}.$$

Due to the assumption $\frac{20}{19}\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F \ge \|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F$, we have

$$\begin{aligned}
&\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star\|_F \le \frac{\sqrt{2}}{2}\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F \\
&\le \frac{\sqrt{2}}{2}(\frac{20}{19}\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F + \|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F) = \frac{39\sqrt{2}}{19}\|\boldsymbol{X}^\star\|_F.
\end{aligned}$$

Now combining the above results with Lemma 12, we have

$$\begin{aligned}
&\nabla^2 G(\boldsymbol{W})[\boldsymbol{\Delta},\boldsymbol{\Delta}] \\
&\le \nabla^2 g(\boldsymbol{W})[\boldsymbol{\Delta},\boldsymbol{\Delta}] + \left|\nabla^2 G(\boldsymbol{W})[\boldsymbol{\Delta},\boldsymbol{\Delta}] - \nabla^2 g(\boldsymbol{W})[\boldsymbol{\Delta},\boldsymbol{\Delta}]\right| \\
&\le -\frac{1}{4}\sigma_r(\boldsymbol{X}^\star) + 2c\left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star\right\|_F\left\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F \\
&\quad + c\left\|\boldsymbol{\Delta}_{\boldsymbol{U}}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{\Delta}_{\boldsymbol{V}}^{\mathrm{T}}\right\|_F^2 \\
&\le -\frac{1}{4}\sigma_r(\boldsymbol{X}^\star) + \frac{39}{19}\sqrt{2}c\|\boldsymbol{X}^\star\|_F + c\sigma_r(\boldsymbol{X}^\star) \\
&\le -\frac{1}{6}\sigma_r(\boldsymbol{X}^\star),
\end{aligned}$$

where the last line holds when $c \le \frac{\sigma_r(\boldsymbol{X}^\star)}{50\|\boldsymbol{X}^\star\|_F}$. This completes the proof of (16).

C. Large Gradient for the Region $\mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$

To show that $G(\boldsymbol{W})$ has large gradient in these three regions, we mainly utilize Lemma 11 to guarantee that $\nabla G(\boldsymbol{W})$ is close to $\nabla g(\boldsymbol{W})$.

1) Large Gradient for the Region $\mathcal{R}_3'$: Utilizing Lemma 11, we have

$$\begin{aligned}
&\|\nabla G(\boldsymbol{W})\|_F \\
&\ge \|\nabla g(\boldsymbol{W})\|_F - \|\nabla G(\boldsymbol{W}) - \nabla g(\boldsymbol{W})\|_F \\
&\ge \|\nabla g(\boldsymbol{W})\|_F - c\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F\|\boldsymbol{W}\| \\
&\ge \|\nabla g(\boldsymbol{W})\|_F - c(\frac{10}{9}\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F + \|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\|_F)\,\|\boldsymbol{W}\|
\end{aligned}$$

$$\geq \frac{1}{10}\sigma_r^{3/2}(\boldsymbol{X}^\star) - c\frac{19}{9}2\|\boldsymbol{X}^\star\|_F\frac{20}{19}\sqrt{2}\|\boldsymbol{X}^\star\|^{1/2}$$
$$\geq \frac{1}{27}\sigma_r^{3/2}(\boldsymbol{X}^\star),$$

where the fourth line follows because $\left\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F = 2\|\boldsymbol{X}^\star\|_F$ and $\|\boldsymbol{W}\| \leq \frac{20}{19}\sqrt{2}\|\boldsymbol{X}^\star\|^{1/2}$, and the last line holds if $c \leq \frac{1}{100}\frac{\sigma_r^{3/2}(\boldsymbol{X}^\star)}{\|\boldsymbol{X}^\star\|_F\|\boldsymbol{X}^\star\|^{1/2}}$. This completes the proof of (17).

2) Large Gradient for the Region $\mathcal{R}_3''$: Utilizing Lemma 11 again, we have

$$\|\nabla G(\boldsymbol{W})\|_F$$
$$\geq \|\nabla g(\boldsymbol{W})\|_F - c\left(\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\right\|_F + \left\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F\right)\|\boldsymbol{W}\|$$
$$\geq \frac{39}{800}\|\boldsymbol{W}\|^3 - c\left(\frac{10}{9}\left\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F + \left\|\boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F\right)\|\boldsymbol{W}\|$$
$$\geq \frac{39}{800}\|\boldsymbol{W}\|^3 - c\frac{19}{9}2\left\|\boldsymbol{X}^\star\right\|_F\|\boldsymbol{W}\|$$
$$\geq \frac{39}{800}\|\boldsymbol{W}\|^3 - \frac{19}{450}\left\|\boldsymbol{X}^\star\right\|\|\boldsymbol{W}\|$$
$$\geq \frac{1}{50}\|\boldsymbol{W}\|^3,$$

where the fourth line holds if $c \leq \frac{1}{100}\frac{\sigma_r^{3/2}(\boldsymbol{X}^\star)}{\|\boldsymbol{X}^\star\|_F\|\boldsymbol{X}^\star\|^{1/2}}$ and the last follows from the fact that

$$\|\boldsymbol{W}\| > \frac{20}{19}\|\boldsymbol{W}^\star\| \geq \frac{20}{19}\sqrt{2}\|\boldsymbol{X}^\star\|^{1/2}.$$

This completes the proof of (18).

3) Large Gradient for the Region $\mathcal{R}_3'''$: To show (19), we first control $|\langle\nabla G(\boldsymbol{W}) - \nabla g(\boldsymbol{W}), \boldsymbol{W}\rangle|$ as follows:

$$|\langle\nabla G(\boldsymbol{W}) - \nabla g(\boldsymbol{W}), \boldsymbol{W}\rangle|$$
$$= 2\left|\left\langle\nabla f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}), \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\right\rangle - \left\langle\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star, \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\right\rangle\right|$$
$$\leq 2c\left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star\right\|_F\left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\right\|_F$$
$$\leq 2c\frac{19}{20}\sqrt{2}\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F\frac{1}{2}\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F = \frac{19}{20}\sqrt{2}c\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F^2,$$

where the first inequality utilizes the fact $\nabla f(\boldsymbol{X}^\star) = \boldsymbol{0}$ and Lemma 10, and the last inequality holds because

$$\left\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X}^\star\right\|_F \leq \frac{\sqrt{2}}{2}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{W}^\star\boldsymbol{W}^{\star\mathrm{T}}\right\|_F$$
$$\leq \frac{\sqrt{2}}{2}\left(\frac{9}{10}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\right\|_F + \left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\right\|_F\right)$$
$$= \frac{19\sqrt{2}}{20}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\right\|_F$$

and

$$\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F^2 = \|\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\|_F^2 + 2\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\|_F^2 \geq 4\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\|_F^2$$

by noting that

$$\|\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\|_F^2 + \|\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\|_F^2 - 2\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\|_F^2 = \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\|_F^2 \geq 0.$$

Now utilizing (76) to provide a lower bound for $\langle\nabla g(\boldsymbol{W}), \boldsymbol{W}\rangle$, we have

$$|\langle\nabla G(\boldsymbol{W}), \boldsymbol{W}\rangle|$$
$$\geq \langle\nabla g(\boldsymbol{W}), \boldsymbol{W}\rangle - |\langle\nabla G(\boldsymbol{W}) - \nabla g(\boldsymbol{W}), \boldsymbol{W}\rangle|$$
$$> \frac{1}{20}\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F^2 - \frac{19}{20}\sqrt{2}c\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F^2$$
$$\geq \frac{1}{45}\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F^2,$$

where the last line holds when $c \leq \frac{1}{50}$. Thus,

$$\|\nabla G(\boldsymbol{W})\|_F \geq \frac{1}{\|\boldsymbol{W}\|}|\langle\nabla G(\boldsymbol{W}), \boldsymbol{W}\rangle| > \frac{1}{45}\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F^{3/2},$$

where we utilize $\|\boldsymbol{W}\| \leq \left(\|\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\|_F\right)^{1/2}$. This completes the proof of (19).

References

[1] S. Aaronson, "The learnability of quantum states," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 463, no. 2088, pp. 3089–3114, Dec. 2007.

[2] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 3, pp. 1235–1256, Jan. 2010.

[3] N. Srebro, J. Rennie, and T. S. Jaakkola, "Maximum-margin matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 1329–1336.

[4] L. Xu and M. Davenport, "Dynamic matrix recovery from incomplete observations under an exact low-rank constraint," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3585–3593.

[5] M. A. Davenport and J. Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 608–622, Jun. 2016.

[6] M. Fazel, H. Hindi, and S. Boyd, "Rank minimization and applications in system theory," in *Proc. Amer. Control Conf.*, vol. 4, 2004, pp. 3273–3278.

[7] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, Jan. 2010.

[8] Z. Harchaoui, M. Douze, M. Paulin, M. Dudik, and J. Malick, "Large-scale image classification with trace-norm regularization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3386–3393.

[9] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Dec. 2009.

[10] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, Jan. 2010.

[11] S. Burer and R. D. C. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Math. Program.*, vol. 95, no. 2, pp. 329–357, Feb. 2003.

[12] S. Burer and R. D. C. Monteiro, "Local minima and convergence in low-rank semidefinite programming," *Math. Program.*, vol. 103, no. 3, pp. 427–444, Jul. 2005.

[13] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3873–3881.

[14] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2973–2981.

[15] D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi, "Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach," in *Proc. Artif. Intell. Statist.*, 2017, pp. 65–74.

[16] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—Online stochastic gradient for tensor decomposition," in *Proc. 28th Conf. Learn. Theory*, 2015, pp. 797–842.

[17] J. Sun, Q. Qu, and J. Wright, "When are nonconvex problems not scary?" 2015, *arXiv:1510.06096*. [Online]. Available: http://arxiv.org/abs/1510.06096

[18] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent converges to minimizers," in *Proc. Conf. Learn. Theory*, 2016, pp. 1246–1257.

[19] I. Panageas and G. Piliouras, "Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions," in *Proc. 8th Innov. Theor. Comput. Sci. Conf. (ITCS)*, 2017, pp. 1–11.

[20] C. Jin, M. Jordan, R. Ge, P. Netrapalli, and S. Kakade, "How to escape saddle points efficiently," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2727–2752.

[21] Q. Li, Z. Zhu, and G. Tang, "Alternating minimizations converge to second-order optimal solutions," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3935–3943.

[22] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *Found. Comput. Math.*, vol. 18, no. 5, pp. 1131–1198, Oct. 2018.

[23] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[24] Y. Chen and E. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 739–747.

[25] K. Lee, N. Tian, and J. Romberg, "Fast and guaranteed blind multichannel deconvolution under a bilinear system model," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 4792–4818, Jul. 2018.

[26] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Appl. Comput. Harmon. Anal.*, vol. 47, no. 3, pp. 893–934, Nov. 2019.

[27] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning sparsely used overcomplete dictionaries," in *Proc. Conf. Learn. Theory (COLT)*, 2014, pp. 123–137.

[28] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere I: Overview and the geometric picture," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 853–884, Feb. 2017.

[29] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 885–914, Feb. 2017.

[30] H. Liu, M.-C. Yue, and A. M.-C. So, "On the estimation performance and convergence rate of the generalized power method for phase synchronization," *SIAM J. Optim.*, vol. 27, no. 4, pp. 2426–2446, Jan. 2017.

[31] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 964–973.

[32] L. Wang, X. Zhang, and Q. Gu, "A unified computational and statistical framework for nonconvex low-rank matrix estimation," in *Proc. Artif. Intell. Statist.*, 2017, pp. 981–990.

[33] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6535–6579, Nov. 2016.

[34] C. Jin, S. M. Kakade, and P. Netrapalli, "Provable efficient online matrix completion via non-convex stochastic gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4520–4528.

[35] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. 45th Annu. ACM Symp. Symp. Theory Comput. (STOC)*, 2013, pp. 665–674.

[36] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 1233–1242.

[37] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi, "Dropping convexity for faster semi-definite optimization," in *Proc. Conf. Learn. Theory*, 2016, pp. 530–582.

[38] T. Zhao, Z. Wang, and H. Liu, "A nonconvex optimization framework for low rank matrix estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 559–567.

[39] Q. Li, Z. Zhu, and G. Tang, "The non-convex geometry of low-rank matrix optimization," *Inf. Inference, J. IMA*, vol. 8, no. 1, pp. 51–96, Mar. 2019.

[40] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "Global optimality in low-rank matrix optimization," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3614–3628, Jul. 2018.

[41] A. R. Conn, N. I. Gould, and P. L. Toint, *Trust Region Methods*. Philadelphia, PA, USA: SIAM, 2000.

[42] X. Li *et al.*, "Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3489–3514, Jun. 2019.

[43] G. S. Chirikjian and A. B. Kyatkin, *Harmonic Analysis for Engineers and Applied Scientists: Updated and Expanded Edition*. New York, NY, USA: Dover, 2016.

[44] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Inf. Inference*, vol. 3, no. 3, pp. 189–223, Sep. 2014.

[45] P. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1996.

[46] E. J. Candès and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2342–2359, Apr. 2011.

[47] M. Udell, C. Horn, R. Zadeh, and S. Boyd, "Generalized low rank models," *Found. Trends Mach. Learn.*, vol. 9, no. 1, pp. 1–118, 2016.

[48] T. Cai and W.-X. Zhou, "A max-norm constrained minimization approach to 1-bit matrix completion," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3619–3647, 2013.

[49] J. Salmon, Z. Harmany, C.-A. Deledalle, and R. Willett, "Poisson noise reduction with non-local PCA," *J. Math. Imag. Vis.*, vol. 48, no. 2, pp. 279–294, Feb. 2014.

[50] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[51] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[52] Z. Zhu, Q. Li, X. Yang, G. Tang, and M. B. Wakin, "Distributed low-rank matrix factorization with exact consensus," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8422–8432.

Zhihui Zhu (Member, IEEE) received the Ph.D. degree in electrical engineering from the Colorado School of Mines, Golden, CO, USA, in 2017. He was a Post-Doctoral Fellow with the Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, MD, USA, from 2018 to 2019. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Denver, Denver, CO, USA. His research interests include exploiting inherent low-dimensional structures and applying optimization methods with guaranteed performance for signal processing and machine learning.

Qiuwei Li (Student Member, IEEE) received the Ph.D. degree in electrical engineering from the Colorado School of Mines, Golden, CO, USA, in 2019. He was a Post-Doctoral Fellow with the Department of Electrical Engineering, Colorado School of Mines, from 2019 to 2020. He is currently a Hedrick Assistant Adjunct Professor with the Department of Mathematics, University of California at Los Angeles, Los Angeles, CA, USA. His research interests include using convex and non-convex optimization methods to model and solve inverse problems in machine learning and signal processing.

Gongguo Tang (Member, IEEE) received the Ph.D. degree in electrical engineering from Washington University in St. Louis in 2011. He was a Post-Doctoral Research Associate with the Department of Electrical and Computer Engineering, University of Wisconsin–Madison, from 2011 to 2013, and a Visiting Scholar with the Simons Institute, University of California at Berkeley, in 2013. He is currently an Associate Professor with the Colorado School of Mines. His research interests include revolves around modeling and optimization in order to make sense of data through computation. He is also especially interested in the design of learning models, optimization formulations, and numerical procedures that come with theoretical performance guarantees and are scalable to large datasets.

Michael B. Wakin (Fellow, IEEE) received the B.S. degree in electrical engineering and the B.A. degree *(summa cum laude)* in mathematics, the M.S. degree in electrical engineering, and the Ph.D. degree in electrical engineering from Rice University, Houston, TX, USA, in 2000, 2002, and 2007, respectively.

He was an NSF Mathematical Sciences Post-Doctoral Research Fellow with Caltech from 2006 to 2007 and an Assistant Professor with the University of Michigan from 2007 to 2008. He is currently a Professor with the Department of Electrical Engineering, Colorado School of Mines. His research interests include sparse, geometric, and manifold-based models for signal processing and compressive sensing.

Dr. Wakin shared the Hershel M. Rich Invention Award from Rice University for the design of a single-pixel camera based on compressive sensing in 2007. He received the DARPA Young Faculty Award for his research in compressive multisignal processing for environments, such as sensor and camera networks in 2008, the NSF CAREER Award for research into dimensionality reduction techniques for structured data sets in 2012, and the Excellence in Research Award for his research as a Junior Faculty Member at CSM in 2014. He was a recipient of the Best Paper Award from the IEEE Signal Processing Society. He has served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS. He is currently an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING.