An equivalence between stationary points for rank constraints versus low-rank factorizations

Wooseok Ha*, Haoyang Liu†, Rina Foygel Barber‡

December 4, 2018

Abstract

Two common approaches in low-rank optimization problems are either working directly with a rank constraint on the matrix variable, or optimizing over a low-rank factorization so that the rank constraint is implicitly ensured. In this paper, we study the natural connection between the rank-constrained and factorized approaches. We show that all second-order stationary points of the factorized objective function correspond to stationary points of projected gradient descent run on the original problem (where the projection step enforces the rank constraint). This result allows us to unify many existing optimization guarantees that have been proved specifically in either the rank-constrained or the factorized setting, and leads to new results for certain settings of the problem.

1 Introduction

We consider the following low rank optimization problem

$$\min_{X \in \mathbb{R}^{m \times n}} \{f(X) : \text{rank}(X) \leq r\},$$

for a differentiable function $f : \mathbb{R}^{m \times n} \to \mathbb{R}$. Due to a wide range of applications, this type of optimization problem has been studied extensively in the past decade.

In some special cases, the unconstrained minimizer of $f(X)$ may already be low-rank, i.e.

$$\hat{X} = \arg \min_{X \in \mathbb{R}^{m \times n}} f(X) = \arg \min_{X \in \mathbb{R}^{m \times n}} \{f(X) : \text{rank}(X) \leq r\}.$$

This setting arises naturally in the study of semidefinite programs (SDP)—a wide class of SDP problems admit low rank solution that are global optimal (e.g. Barvinok).

*Department of Statistics, University of California, Berkeley
†Department of Statistics, University of Chicago
Since SDP problems are convex, they can already be solved by convex optimization algorithms such as interior point methods. A low-rank solution to the unconstrained minimization problem can also arise in matrix inverse problems, including noiseless matrix sensing [Recht et al., 2010] and noiseless matrix completion [Candes and Recht, 2009]. In these settings, while the rank constraint is not active at the global optimum $\hat{X}$, enforcing the constraint over the course of an iterative algorithm may still be useful in speeding up convergence [Oymak et al., 2018].

In most settings, however, the rank constraint $\text{rank}(X) \leq r$ will be active in the solution to the minimization problem (1), meaning that we must necessarily work with the rank constraint in the optimization. In this case, the optimization strategies in the literature can be broadly categorized into two types: either working with the full variable $X \in \mathbb{R}^{m \times n}$ while enforcing $\text{rank}(X) \leq r$ (e.g. by projecting to this constraint after each iteration), or reformulating the problem in terms of a factorization $X = AB^\top$ with $A \in \mathbb{R}^{m \times r}$ and $B \in \mathbb{R}^{n \times r}$, so that the factorization ensures the rank constraint. (There is also extensive literature on relaxing to a convex penalty or constraint, such as the nuclear norm [Recht et al., 2010], but here we will focus on optimization techniques that work with the original rank constraint rather than a relaxation.) Working either with $X$ or with a factorization, we can implement a gradient descent algorithm to attempt to find the solution to (1).

Specifically, working with the full variable $X \in \mathbb{R}^{m \times n}$, we can consider the projected gradient descent method (also known as iterative hard thresholding, see Jain et al. [2014]):

$$X \leftarrow {\mathcal{P}}_r(X - \eta \nabla f(X)), \quad (2)$$

where ${\mathcal{P}}_r(\cdot)$ denotes projection to the rank-$r$ constraint (calculated by taking the top $r$ components of a singular value decomposition). If we work instead in the factorized setting, we would aim to solve

$$\min_{A \in \mathbb{R}^{m \times r}, B \in \mathbb{R}^{n \times r}} f(AB^\top). \quad (3)$$

For instance, we might approach this minimization via alternating gradient descent, which would iterate steps of the form

$$\begin{cases}
    A \leftarrow A - \eta_A \cdot \nabla f(AB^\top)B, \\
    B \leftarrow B - \eta_B \cdot \nabla f(AB^\top)^\top A,
\end{cases} \quad (4)$$

alternating between updates of each of the two factors.

### 1.1 Comparing full-dimensional vs factorized approaches

In this work, we are interested in comparing the output of full-dimensional approaches such as (2) against factorized approaches aiming to solve (3).
Special case: semidefinite programs  This problem has been studied in the context of semidefinite programs of the form

\[
\hat{X} = \arg\min_{X \in \mathbb{R}^{n \times n}, X \succeq 0} \{ f(X) : f_1(X) = a_1, \ldots, f_k(X) = a_k \}
\]

where \( f, f_1, \ldots, f_k \) are all linear functions. The factorized form of this problem is given by writing \( X = AA^\top \), and solving

\[
\min_{A \in \mathbb{R}^{n \times r}} \{ f(AA^\top) : f_1(AA^\top) = a_1, \ldots, f_k(AA^\top) = a_k \}.
\]

If we take \( r = n \), then the global minimizer of this problem coincides with that of the full SDP—and in fact, this holds as long as \( r \geq \text{rank}(\hat{X}) \). On the other hand, the factorized problem is highly nonconvex so finding the global minimum may be challenging. Remarkably, Bhojanapalli et al. [2018] (building on the earlier work of Burer and Monteiro [2003]) show that taking \( r \sim \sqrt{k} \) is sufficient to ensure that any second-order stationary point (SOSP) of the factorized problem is a global minimizer of the full SDP; it is also shown that approximate SOSPs are approximately globally optimal (see also Boumal et al. [2016, 2018]). Of course, for \( A \in \mathbb{R}^{n \times r} \) to achieve the global minimum at \( r \sim \sqrt{k} \), this means that the global minimizer \( \hat{X} \) itself must have rank on the order of \( \sqrt{k} \).

Contributions  In this work we strengthen the connection between the factorized problem (3) and the original problem (2), extending existing results into a much broader setting. The results mentioned above apply in the setting where:

- The optimization problem is a SDP, meaning that the objective function is linear and the factorized form is given by \( X = AA^\top \),

- The unconstrained global minimizer \( \hat{X} \) is rank-deficient (without imposing a rank constraint),

- Results apply to finding the global minimum.

In contrast, in our work, we will allow:

- The objective function is any twice-differentiable function \( f(X) \), and \( X \) is not necessarily symmetric, i.e. the factorized form is given by \( X = AB^\top \),

- The unconstrained global minimizer, \( \arg\min_X f(X) \), may be full rank in general—in the rank-constrained problem, \( \hat{X} = \arg\min_X \{ f(X) : \text{rank}(X) \leq r \} \), the rank constraint may be active,

- Results no longer apply to finding the global minimum (since this is NP-hard), but instead we study stationary points.
In this general setting, our main result is the following: we show that any second-order stationary point (SOSP) of the factorized objective function \((3)\) must also be a stationary point of projected gradient descent on the original objective function \((2)\). In the special case that the problem is a SDP and the SOSP is rank-deficient, i.e. rank strictly less than \(r\), this result reduces to the known global optimality result proved by Bhojanapalli et al. [2018].

### 1.2 Notation

Throughout the paper, \(f : \mathbb{R}^{m \times n} \to \mathbb{R}\) is a twice-differentiable objective function. Its gradient \(\nabla f(X)\) is represented as a matrix in \(\mathbb{R}^{m \times n}\) while its second derivative \(\nabla^2 f(X) : \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} \to \mathbb{R}\) will be written as a quadratic form, i.e. \(\nabla^2 f(X)(X_1, X_2)\).

We will work also with \(g(A, B) = f(AB^\top)\), the function defining the factorized problem. Writing \(g : \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} \to \mathbb{R}\), we will work with first derivative \(\nabla g(A, B) = (\nabla_A g(A, B), \nabla_B g(A, B)) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}\), while the second derivative \(\nabla^2 g(A, B)\) will denote the quadratic form mapping from \((\mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}) \times (\mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r})\) to \(\mathbb{R}\).

For a matrix \(X\), we write \(\|X\|_F\) and \(\|X\|\) to denote the Frobenius norm and the spectral norm, respectively.

### 2 Main result

Our main result concerns stationary points of the rank-constrained minimization problem \((1)\). We will consider projected gradient descent (PGD) algorithms, which have update steps of the form

\[X_{t+1} \leftarrow P_r(X_t - \eta \nabla f(X_t)),\]

where \(\eta > 0\) is the step size, while \(P_r\) denotes (possibly non-unique) projection to the rank constraint, i.e. \(P_r(X) = \arg \min_{\text{rank}(X') \leq r} \|X' - X\|_F\).

A matrix \(X \in \mathbb{R}^{m \times n}\) is therefore a stationary point of PGD at step size \(\eta > 0\) if it satisfies

\[X = P_r(X - \eta \nabla f(X)).\]

By examining this condition, we can easily determine that \(X\) is a stationary point if and only if

\[\nabla f(X)^\top U_X = 0 \text{ and } \nabla f(X)V_X = 0 \text{ and } \eta \|\nabla f(X)\| \leq \sigma_r,\]

\[\text{(5)}\]

\(^1\)If the projection step is not unique, we need to be more precise with our definition. We say that \(X\) is a stationary point of PGD at step size \(\eta\) if \(X\) is equal to a (possibly non-unique) solution of the projection step, i.e. \(X \in \arg \min_{\text{rank}(X') \leq r} \|X' - (X - \eta \nabla f(X))\|_F\).
where $X = U X \cdot \text{diag}\{\sigma_1, \ldots, \sigma_r\} \cdot V_X^T$ is a (possibly non-unique) singular value decomposition of $X$, with $\sigma_1 \geq \cdots \geq \sigma_r$.

We will also consider factorized gradient descent (FGD) algorithms of the form \[ g(A, B) = f(A B^\top), \] where we write $X = A B^\top$ for $A \in \mathbb{R}^{m \times r}$ and $B \in \mathbb{R}^{n \times r}$. For this type of algorithm, defining

\begin{align*}
    g(A, B) &= f(A B^\top), \\
    \begin{cases}
        A &\leftarrow A - \eta_A \cdot \nabla_A g(A, B), \\
        B &\leftarrow B - \eta_B \cdot \nabla_B g(A, B).
    \end{cases}
\end{align*}

Given any step sizes $\eta_A > 0, \eta_B > 0$, a pair $(A, B)$ is a stationary point of factorized gradient descent if and only if $\nabla g(A, B) = 0$. In other words, the stationary points of factorized gradient descent are simply the first-order stationary points (FOSPs) of the function $g$. By definition of $g$, we can calculate

\[ \nabla_A g(A, B) = \nabla f(A B^\top) B \quad \text{and} \quad \nabla_B g(A, B) = \nabla f(A B^\top)^\top A, \]

and so the FOSPs of the factorized minimization problem can equivalently be characterized by

\[ \nabla g(A, B) = 0 \iff f(A B^\top)^\top A = 0 \quad \text{and} \quad \nabla f(A B^\top) B = 0. \]  

This calculation leads to the following well-known result, which follows directly from calculations of the gradients, and requires no proof:

**Lemma 1.** If $X \in \mathbb{R}^{m \times n}$ is a stationary point of the projected gradient descent algorithm at any step size $\eta > 0$, then for any factorization $X = A B^\top$ with $(A, B) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$, the pair $(A, B)$ is a first-order stationary point of the factorized objective function $g(A, B)$.

In other words, stationary points of PGD are first-order stationary points of the factorized problem.

However, we cannot hope for the converse to be true, since FOSPs of $g$ can exhibit some counterintuitive behavior that does not arise in the full-dimensional problem. A well-known example is the pair $(A, B) = (0_{m \times r}, 0_{n \times r})$. This point is always a FOSP of the factorized problem, but in general $X = 0_{m \times n}$ does not correspond to a stationary point of projected gradient descent (and indeed, will be far from optimal). From this trivial example, we see that considering only the first-order conditions of $g$ is not sufficient to understand the correspondence between the projected and the factorized forms of the problem. We will therefore consider second-order stationary points (SOSPs) of the factorized problem, which are characterized by the conditions

\[ \nabla g(A, B) = 0 \quad \text{and} \quad \nabla^2 g(A, B) \succeq 0. \]  

\[ (7) \]
2.1 Characterization of SOSP for factorized problem

Our main theoretical result establishes a partial converse to Lemma 1, proving that any second-order stationary point (SOSP) of the factorized objective function \( g(A, B) \) must also be a stationary point of projected gradient descent on the original function \( f(X) \). We need one additional piece of notation before we can state our result, to allow us to quantify the smoothness of \( f \) on the space of low-rank matrices.

When running projected gradient descent algorithms with a constant step size \( \eta \), typically the step size is chosen with respect to the curvature of \( f \). Specifically, at the current point \( X \), one step of projected gradient descent with step size \( \eta = 1/\beta \) can be interpreted as minimizing the function \( Y \mapsto f(X) + \langle \nabla f(X), Y - X \rangle + \frac{\beta}{2} \| X - Y \|_F^2 \) under the constraint \( \text{rank}(Y) \leq r \). If this function majorizes \( f \), i.e. \( f \) is bounded by this function, then our update step will make progress towards minimizing \( f \). With this in mind, we define the local curvature of \( f \) at \( X \) as

\[
\beta_f(X) = \lim_{\epsilon \to 0} \left\{ \sup_{\substack{0 < \| Y - X \|_F \leq \epsilon \\
\text{rank}(Y) \leq r}} \frac{f(Y) - f(X) - \langle \nabla f(X), Y - X \rangle}{\frac{1}{2} \| X - Y \|_F^2} \right\}.
\]

Note that, if we were to remove the rank constraint, \( \text{rank}(Y) \leq r \), from this definition, then we are simply calculating the operator norm of the Hessian of \( f \), i.e. \( \| \nabla^2 f(X) \| \). In particular, this proves that \( \beta_f(X) \leq \| \nabla^2 f(X) \| \), and thus is always finite as long as \( f \) is twice differentiable.

With this definition in place, we can now state our main result.

**Theorem 1.** Consider any pair \( (A, B) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} \) that is a SOSP of the factorized objective function \( g(A, B) \). Then \( X = AB^\top \) is a stationary point of the projected gradient descent algorithm at any step size \( \eta \leq 1/\beta_f(X) \). Furthermore, if \( \text{rank}(X) < r \), then \( \nabla f(X) = 0 \), and so \( X \) is a stationary point of the projected gradient descent algorithm at any step size \( \eta \).

To summarize, our main result (combined with the known result in Lemma 1) shows that:

\[
\{ \text{SOSPs of } g(A, B) \} \subseteq \{ \text{Stationary points of PGD on } f(X) \} \subseteq \{ \text{FOSPs of } g(A, B) \}.
\]

### 2.1.1 Proof of Theorem 1

By definition of \( g \), some simple calculations show that \( \nabla^2 g(A, B) \) maps \((A_1, B_1) \times (A_2, B_2)\) to

\[
\langle \nabla f(X), A_1 B_2^\top + A_2 B_1^\top \rangle + \nabla^2 f(X) \left( A B_1^\top + A_1 B^\top, A_2 B_2^\top + A_2 B^\top \right).
\]
Since we assume that $\nabla^2 g(A, B) \succeq 0$ by definition of a SOSP, this means that the above quantity is nonnegative at any $(A_1, B_1) = (A_2, B_2)$, i.e.

$$2\langle \nabla f(X), A_1 B_1^\top \rangle + \nabla^2 f(X) \left( AB_1^\top + A_1 B_1^\top, AB_1^\top + A_1 B_1^\top \right) \geq 0 \text{ for all } (A_1, B_1). \quad (9)$$

By first-order optimality conditions at $(A, B)$ we additionally know that

$$\nabla f(X)^\top A = 0 \text{ and } \nabla f(X) B = 0. \quad (10)$$

Next, let $X = U_X \cdot \text{diag}\{\sigma_1, \ldots, \sigma_r\} \cdot V_X^\top$ be a singular value decomposition of $X$, with $\sigma_1 \geq \cdots \geq \sigma_r$. Let $u_* \in \mathbb{R}^m$ and $v_* \in \mathbb{R}^n$ be the top singular vectors of the gradient $\nabla f(X) \in \mathbb{R}^{m \times n}$, so that $\|\nabla f(X)\| = u_*^\top \nabla f(X) v_*$. We will now split into two cases, $\text{rank}(X) = r$ and $\text{rank}(X) < r$.

**Case 1: full rank** First suppose $\text{rank}(X) = r$. Let $u_r$ and $v_r$ be the last left and right singular vectors of $X$, respectively. Since $X = AB^\top$ has rank $r$, this means that $U_X$ and $A$ span the same column space, and similarly $V_X$ and $B$ span the same column space. Together with the first-order optimality conditions in $(10)$, this implies that $\nabla f(X)V_X = 0$ and $\nabla f(X)^\top U_X = 0$. By our earlier characterization of the stationary points of PGD, we therefore only need to check that $\eta \|\nabla f(X)\| \leq \sigma_r(X)$ in order to verify that $X$ is a stationary point of PGD at step size $\eta$.

Next, if $\nabla f(X) = 0$ then $X$ is obviously a stationary point, so from this point on we will consider the case that $\nabla f(X) \neq 0$. Since we know that $\nabla f(X)^\top U_X = 0$ while $u_*$ is the first left singular vector of $\nabla f(X)$, this implies that $u_*^\top u_* = 0$. Similarly $v_*^\top v_* = 0$. We will consider the curvature of the factorized objective function $g(A, B)$ in the direction given by $(A_1, B_1) = (-u_* u_r^\top A, v_* v_r^\top B)$. Plugging this choice into our earlier calculation $(9)$ we see that

$$\nabla^2 f(X) \left( AB_1^\top + A_1 B_1^\top, AB_1^\top + A_1 B_1^\top \right) \geq -2\langle \nabla f(X), A_1 B_1^\top \rangle$$

$$= 2\langle \nabla f(X), u_* u_r^\top AB_1^\top v_r v_*^\top \rangle = 2\sigma_r \|\nabla f(X)\|,$$

where the last step holds since $u_r, v_r$ are the $r$th singular vectors of $X = AB^\top$.

Next, we will use the following lemma (proved in Appendix $\text{A}$):

**Lemma 2.** Let $f: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ be twice-differentiable at $X = AB^\top$, where $A \in \mathbb{R}^{m \times r}$ and $B \in \mathbb{R}^{n \times r}$. Then, for any matrices $A_1 \in \mathbb{R}^{m \times r}, B_1 \in \mathbb{R}^{n \times r}$,

$$\nabla^2 f(X) \left( AB_1^\top + A_1 B_1^\top, AB_1^\top + A_1 B_1^\top \right) \leq \beta_f(X) \cdot \|AB_1^\top + A_1 B_1^\top\|_F^2.$$

Now fix any step size $\eta > 0$ with $\eta \leq 1/\beta_f(X)$. Then by Lemma $2$ along with the definitions of $A_1$ and $B_1$, we can bound

$$\nabla^2 f(X) \left( AB_1^\top + A_1 B_1^\top, AB_1^\top + A_1 B_1^\top \right) \leq \eta^{-1} \cdot \|AB_1^\top + A_1 B_1^\top\|_F^2$$

$$= \eta^{-1} \cdot \|AB_1^\top v_r v_*^\top - u_* u_r^\top AB_1^\top\|_F^2 = \eta^{-1} \cdot \sigma_r(X)^2 \|u_r v_* - u_* v_r^\top\|_F^2 = 2\eta^{-1} \sigma_r^2,$$
where the next-to-last step holds since \( u_r, v_r \) are the \( r \)th singular vectors of \( X = AB^\top \), while the last step holds since \( u_r, u_* \) and \( v_r, v_* \) are pairs of orthogonal unit vectors. Combining everything, and using the fact that \( \sigma_r > 0 \) since \( \text{rank}(X) = r \), we have proved that

\[
\eta \|\nabla f(X)\| \leq \sigma_r.
\]

Applying (5), this verifies that \( X \) is a stationary point of PGD with step size \( \eta \), which completes the proof for the rank-\( r \) case.

**Case 2: rank deficient.** For the case that \( \text{rank}(X) < r \), our proof closely follows that of Bhojanapalli et al. [2018, Lemma 1], extending their result to the asymmetric case (their work assumes \( X \succeq 0 \) and works with the symmetric factorization \( X = AA^\top \)).

First, since \( A \in \mathbb{R}^{m \times r} \) and \( B \in \mathbb{R}^{n \times r} \), if the product \( X = AB^\top \) has rank \( < r \) then it cannot be the case that both \( A \) and \( B \) are full rank. Without loss of generality suppose \( \text{rank}(A) < r \). This means that there is some unit vector \( w \in \mathbb{R}^r \) with \( Aw = 0 \). Now consider \( (A_1, B_1) = (-u_*w^\top, c \cdot v_*w^\top) \) for any \( c > 0 \). Since \( (A, B) \) is a SOSP of the factorized problem, our earlier calculation (9) yields

\[
2\langle \nabla f(X), -c \cdot u_*w^\top wv_*^\top \rangle + \nabla^2 f(X) \left( c \cdot Awv_*^\top + u_*w^\top B^\top, c \cdot Awv_*^\top + u_*w^\top B^\top \right) \geq 0.
\]

Since \( \|w\|_2 = 1 \) while \( u_*^\top \nabla f(X)v_* = \|\nabla f(X)\| \), and \( Aw = 0 \) by definition of \( w \), we can simplify this to

\[
\nabla^2 f(X) \left( u_*w^\top B^\top, u_*w^\top B^\top \right) \geq 2c\|\nabla f(X)\|.
\]

Now, \( c > 0 \) is arbitrary, and so this holds for any \( c > 0 \). On the other hand, since \( f \) is twice-differentiable, the left-hand side must be finite. This implies that \( \|\nabla f(X)\| = 0 \), i.e. \( \nabla f(X) = 0 \). Therefore clearly \( X \) is a stationary point of projected gradient descent at any step size \( \eta \).

### 3 Convergence guarantees

In this section, we investigate the implications of our main result Theorem 1 on the landscape of the factorized problem (3). We are interested in determining settings where factorized optimization methods can be expected to achieve optimality guarantees. Depending on the structure of the objective function \( f \) and other assumptions in the problem, we will see wide variation in the types of guarantees that can be obtained for the output \( \hat{X} \) of a particular algorithm. From strongest to weakest, the three main styles of guarantees that appear in the literature are:

- **Global optimality:** the algorithm converges to a global minimizer.
- Local optimality: if initialized near a global minimizer, then the algorithm converges to that global minimizer.

- Restricted optimality: the algorithm converges to a matrix $X$ that satisfies $f(X) \leq f(X')$ for any rank-$r'$ matrix $X'$, where $r' < r$ is a strictly lower rank constraint.

To simplify our comparison of these three styles of guarantees, we will consider the setting where the original objective function $f$ satisfies $\alpha$-restricted strong convexity (abbreviated as $\alpha$-RSC) with respect to the rank constraint $r$, meaning that for all $X, Y \in \mathbb{R}^{m \times n}$ with rank($X$), rank($Y$) $\leq r$,

$$f(Y) \geq f(X) + \langle \nabla f(X), Y - X \rangle + \frac{\alpha}{2} \|X - Y\|_F^2. \quad (11)$$

Similarly, we assume that $f$ satisfies $\beta$-restricted smoothness with parameter $\beta$ (abbreviated as $\beta$-RSM) with respect to the rank constraint $r$, meaning that for all $X, Y$ with rank($X$), rank($Y$) $\leq r$,

$$f(Y) \leq f(X) + \langle \nabla f(X), Y - X \rangle + \frac{\beta}{2} \|X - Y\|_F^2. \quad (12)$$

Throughout this section, we will always write $\kappa = \beta/\alpha$ to denote the rank-restricted condition number of $f$. Note that $\kappa \geq 1$ always.

We will consider two different regimes for the condition number $\kappa$:

Near-isometry ($\kappa \approx 1$) vs. Arbitrary conditioning ($\kappa \gg 1$).

We can expect to see $\kappa \approx 1$ in certain well-behaved problems, for instance the matrix sensing problem, where $f(X)$ represents matching $X$ with random linear measurements of the form $\langle A_i, X \rangle$, where e.g. the measurement matrices $A_i$ have i.i.d. entries. In general, however, most problems do not have $\kappa \approx 1$.

We also need to consider a second important distinction between different classes of problems. In many statistical settings, we may have an objective function $f(X)$ that comes from a data likelihood, where $\mathbb{E}[f(X)]$ is minimized at some true low-rank parameter matrix $X_*$. When this is the case, it is common to see $\|\nabla f(X_*)\| \approx 0$. In other settings, though, there might not be any natural underlying low-rank structure, and the gradient $\nabla f(X)$ is large at any low-rank $X$. We will therefore distinguish between two scenarios:

Vanishing gradient ($\min_{\text{rank}(X) \leq r} \|\nabla f(X)\| \approx 0$) vs. Arbitrary gradient ($\min_{\text{rank}(X) \leq r} \|\nabla f(X)\| \gg 0$).

### 3.1 Existing results

We now summarize the existing results as well as our own findings, for the different types of assumptions and different styles of guarantees outlined above:
• Near-isometry + Vanishing gradient ⇒ Global optimality.
  For the most well-behaved problems, where the objective function $f(X)$ exhibits both near-isometry and a vanishing gradient, it is possible to prove convergence to an (approximate) globally optimal estimate $\hat{X}$. For full-dimensional projected gradient descent algorithm, this has been established in the case of a least squares objective [Oymak et al., 2018]; for factorized algorithms, an analogous result (no spurious local minima) has been established for certain least squares objectives [Bhojanapalli et al., 2016b, Ge et al., 2016, 2017, Park et al., 2016] and more generally for functions $f$ with a near-isometry property [Zhu et al., 2017]. (We will show in the present work that under near-isometry + vanishing gradient, both full-dimensional and factorized approaches contain no spurious local minima.)

• Arbitrary conditioning + Vanishing gradient ⇒ Local optimality.
  With a non-ideal condition number $\kappa > 1$, assuming a vanishing gradient condition is sufficient to prove a local optimality result, both for full-dimensional PGD [Barber and Ha, 2018] and for factorized approaches [Chen and Wainwright, 2015]; in the stronger setting of a near-isometry and a vanishing gradient, the local optimality result for factorized approaches has been also established by many works, including [Candes et al., 2015], Zheng and Lafferty [2015], Tu et al. [2015], Bhojanapalli et al. [2016a], Jain et al. [2013]. Note that all of the previous local optimality results for factorized problems are built upon identifying local region of attraction for globally optimal solution $\hat{X}$ in the factorized space $(A, B)$. (We will give in the present work the local region of attraction in the full-dimensional representations $X = AB^\top$.)

• Arbitrary conditioning + Arbitrary gradient ⇒ Restricted optimality.
  In the most challenging setting, where we allow both arbitrary condition number $\kappa$ and an arbitrarily large gradient, restricted optimality guarantees can still be obtained. This is established for the full-dimensional PGD algorithm [Jain et al., 2014, Liu and Barber, 2018], as well as its variants, such as approximate low-rank projection [Becker et al., 2013, Soltani and Hegde, 2017], and projection with debiasing step [Yuan et al., 2018]; for sparse problems specifically, the analogous restricted optimality result has been established [Shen and Li, 2017]. On the other hand, there is no known result for restricted optimality guarantees within the factorized approach. (We will show in the present work that it holds also for the factorized approach.)

This extensive literature has enabled us to understand the landscape of the non-convex low-rank optimization problem, but the various results have been proved somewhat disjointly, using very different techniques for analyzing full-dimensional PGD type algorithms versus factorized algorithms. It is natural to ask whether this collection of results can be unified into a single framework. Our main result,
Theorem 1 allows us to connect established results between PGD algorithms and factorized algorithms, allowing us to establish simpler proofs of some existing results, and provide new results in certain settings. Overall, it is the goal of this section to provide a broader view of the landscape of results known for low-rank optimization problems through the lens of the equivalence between PGD and factorized algorithms established in Theorem 1.

3.2 Results for global and local optimality

In the special case of least squares objective, i.e. \( f(X) = \frac{1}{2} \| A(X) - b \|^2_F \) for a linear operator \( A : \mathbb{R}^{m \times n} \to \mathbb{R}^p \), Oymak et al. [2018] show that, in the near-isometry setting (\( \kappa \approx 1 \)), projected gradient descent offers a global convergence guarantee starting from any initialization point. Here we extend some of their technical tools to general functions \( f(X) \).

**Lemma 3.** Suppose that \( f : \mathbb{R}^{m \times n} \to \mathbb{R} \) satisfies \( \alpha \)-RSC (11) with respect to the rank constraint \( r \). If \( X_0, X_1 \) are both stationary points of projected gradient descent run with rank constraint \( r \) and step size \( \eta \), then either \( X_0 = X_1 \) or

\[
\min \left\{ \frac{1}{\eta} \frac{\| \nabla f(X_0) \|}{\sigma_r(X_0)} \right\} + \min \left\{ \frac{1}{\eta} \frac{\| \nabla f(X_1) \|}{\sigma_r(X_1)} \right\} \geq 2\alpha.
\]

The proof of this lemma is given in Appendix A. We also verify a simple result:

**Lemma 4.** Suppose that \( f : \mathbb{R}^{m \times n} \to \mathbb{R} \) satisfies \( \beta \)-RSM (12) with respect to the rank constraint \( r \). If \( \hat{X} \) is a global minimizer, i.e. \( f(\hat{X}) = \min_{\text{rank}(X) \leq r} f(X) \), then \( \hat{X} \) is a stationary point of projected gradient descent run with rank constraint \( r \) and any step size \( \eta \leq 1/\beta \).

These lemmas will allow us to easily prove global optimality and local optimality results under the appropriate assumptions. We now turn to the question of obtaining global and local optimality results for PGD and factorized algorithms. While results of this flavor are already known in the literature (see Section 3.1 for some references), our goal here is to give extremely short and clean proofs that illuminate the connection between the full-dimensional and factorized representations of the optimization problem, and thereby also highlight the utility of our main result, Theorem 1. In some cases, our work also establishes guarantees in a broader setting than previous results.

3.2.1 Global optimality

In the setting where \( f(X) \) satisfies the near-isometry property, with condition number \( \kappa < 2 \), we can obtain global optimality guarantees for both PGD and factorized methods whenever \( \| \nabla f(X) \| \) is sufficiently small, i.e. the vanishing gradient condition. (See Section 3.1 for related existing results in the literature.)
Theorem 2. Assume that \( f(X) \) satisfies \( \alpha\)-RSC \((11) \) and \( \beta\)-RSM \((12) \) with respect to rank \( r \), and that \( \beta < 2\alpha \). If \( \hat{X} \) is a global minimizer, i.e. \( f(\hat{X}) = \min_{\text{rank}(X) \leq r} f(X) \), and \( \hat{X} \) satisfies

\[
\| \nabla f(\hat{X}) \| < (2\alpha - \beta) \cdot \sigma_r(\hat{X}),
\]

then

- \( \hat{X} \) is the unique stationary point of PGD for any step size \( \eta \) satisfying

\[
\frac{1}{2\alpha - \frac{\| \nabla f(\hat{X}) \|}{\sigma_r(X)}} < \eta \leq \frac{1}{\beta}.
\]

- \( \hat{X} \) is the unique second-order stationary point of factorized gradient descent.

This result proves that global optimality guarantees can be achieved as long as \( \kappa < 2 \), i.e. the map \( f \) is a near-isometry. This type of assumption on \( \kappa \) is crucial to achieving global optimality guarantees. For instance,\(^\text{[1]}\) Zhang et al. \([2018]\) Example 3 construct an example of objective function \( f(X) \) with \( \beta = 3\alpha \), i.e. \( \kappa = 3 \), where there exists a stationary point \( X \) that is not globally optimal. This proves that \( \kappa < 3 \) is necessary for achieving a global optimality guarantee, while our work shows \( \kappa < 2 \) is sufficient. While it is not the goal of the present work, an interesting open question is to close the gap between these necessary and sufficient conditions to identify an exact correspondence between condition number and the global optimality guarantee.

We now compare this result with some recent works in the literature. The first part of Theorem \([2]\) i.e. the result for stationary points of PGD on \( X \in \mathbb{R}^{m \times n} \), is an extension of global optimality results established in Oymak et al. \([2018]\) —their work is specific to a least-squares objective function, i.e. \( f \) is quadratic.\(^\text{[2]}\) On the other hand, the second part of the theorem, i.e. the result on SOSPs of the factorized problem, is already known for various types of problems, such as the matrix sensing and the matrix completion problems \([Bhojanapalli et al., 2016a, Ge et al., 2016, 2017]\). Similarly, Zhu et al. \([2017]\) also establish “no spurious local minima” under conditions similar to Theorem \([2]\) i.e. when \( f(X) \) satisfies \( \alpha\)-RSC and \( \beta\)-RSM with \( \alpha \approx \beta \). While these results typically require more involved analysis than our framework presented here, they further prove strict saddle property (see, for instance, Jin et al. \([2017]\), Assumption A2)) of the factorized problems under which polynomial time convergence is ensured for finding approximate SOSPs (hence approximate globally optimal solution). Such guarantee on the rate of convergence is not provided in Theorem \([2]\) and we leave the study of approximate SOSPs in the future work.

\(^2\)In Oymak et al. \([2018]\), the authors mention that their results are more broadly applicable than least squares objective functions, but we are not aware of any such results that have appeared in the follow-up papers.
Proof of Theorem 2. First consider PGD with step size $\eta$ lying in the specified interval. By Lemma 4 we know that $\hat{X}$ is a stationary point. Next suppose that $X$ is another stationary point, with $X \neq \hat{X}$. Applying Lemma 3 with $X_0 = \hat{X}$ and $X_1 = X$ yields
\[
\frac{\|\nabla f(\hat{X})\|}{\sigma_r(\hat{X})} + \frac{1}{\eta} \geq 2\alpha,
\]
but this inequality cannot hold by our assumption on $\eta$, and so we have reached a contradiction.

Next we turn to factorized gradient descent. Let $X$ be any rank-$r$ SOSP. Comparing the definition of $\beta$-RSM with that of the local smoothness parameter $\beta_f(X)$ defined in (8), we can see that $\beta_f(X) \leq \beta$ by definition, and therefore $\eta \leq 1/\beta_f(X)$. Therefore, applying our main result, Theorem 1, we see that $X$ must be a stationary point of PGD at step size $\eta = 1/\beta$, which proves that $X = \hat{X}$ by our work above.

3.2.2 Local optimality

Next we turn to the local optimality guarantees that can be obtained when $f$ exhibits a vanishing gradient, but may have an arbitrarily large condition number $\kappa$. (See Section 3.1 for related existing results in the literature.)

**Theorem 3.** Assume that $f(X)$ satisfies $\alpha$-RSC (11) and $\beta$-RSM (12) with respect to rank $r$. If $\hat{X}$ is a global minimizer, i.e. $f(\hat{X}) = \min_{\text{rank}(X) \leq r} f(X)$, and $\hat{X}$ satisfies
\[
\|\nabla f(\hat{X})\| < \alpha \cdot \sigma_r(\hat{X}),
\]
Define
\[
\mathcal{N} = \left\{ X \in \mathbb{R}^{m \times n} : \text{rank}(X) \leq r \text{ and } \frac{\|\nabla f(\hat{X})\|}{\sigma_r(\hat{X})} + \frac{\|\nabla f(X)\|}{\sigma_r(X)} < 2\alpha \right\}.
\]
(Note that $\mathcal{N}$ must contain some neighborhood of $\hat{X}$, since $X \mapsto \|\nabla f(X)\|/\sigma_r(X)$ is continuous.) Then

- For any stationary point $X$ of PGD with step size $\eta \leq 1/\beta$, if $X \in \mathcal{N}$ then $X = \hat{X}$.

- For any second-order stationary point $X$ of factorized gradient descent, if $X \in \mathcal{N}$ then $X = \hat{X}$.

In this setting where $\kappa$ may be arbitrarily large, global optimality does not hold in general (as shown by Zhang et al. [2018]'s counterexample, discussed in Section 3.2.1 above). Nonetheless, the results in Theorem 3 still assure the existence of regions of
attraction $\mathcal{N}$ within which the global minimum $\hat{X}$ will be discovered, for both the full-dimensional and factorized methods.

To compare with the existing results, the first part of Theorem 3 (for stationary points of PGD) is an immediate result given the work in [Barber and Ha, 2018]. Next, turning to the second part of the result, on the SOSPs of the factorized approach, some related results in the existing literature have shown that certain rank-constrained problems exhibit local region of attraction near the global minimum $\hat{X}$ [Candes et al., 2015; Zheng and Lafferty, 2015; Tu et al., 2015; Bhojanapalli et al., 2016a; Jain et al., 2013]. While these problems satisfy the near-isometry property with $κ \approx 1$, our result in Theorem 3 extends to a broader setting with an arbitrarily large condition number $κ$. Chen and Wainwright [2015] have also established local convergence guarantees under conditions similar to restricted strong convexity and smoothness, but the difference is that they work with RSC and RSM type conditions defined directly on the factorized variable pair $(A, B)$. In addition, many of these works address the positive semidefinite setting, $X = AA^\top$, rather than the generic setting $X = AB^\top$ considered here.

**Proof of Theorem 3.** By Lemma 4, we know that $\hat{X}$ is a stationary point for PGD with any step size $η \leq 1/β$. Next suppose that $X$ is another stationary point, with $X \neq \hat{X}$. Applying Lemma 3 with $X_0 = \hat{X}$ and $X_1 = X$ yields

$$\frac{\|\nabla f(\hat{X})\|}{σ_r(\hat{X})} + \frac{\|\nabla f(X)\|}{σ_r(X)} \geq 2α,$$

which implies that $X \notin \mathcal{N}$ by definition of $\mathcal{N}$.

Next we turn to factorized gradient descent. As in the proof of Theorem 2, any rank-$r$ SOSP $X$ must be a stationary point of PGD at step size $η = 1/β$. If also $X \in \mathcal{N}$ then this proves that $X = \hat{X}$ by our work above.

### 3.3 A restricted optimality guarantee

In this last setting, we will make no assumptions on either the gradient or the condition number, i.e. it may be possible that $\|\nabla f(\hat{X})\|$ is large and the condition $κ$ is large as well. (See Section 3.1 for related existing results in the literature.)

Under such assumptions, to the best of our knowledge, there is no guaranteed result to solve the low-rank minimization problem either locally or globally—identifying a region of attraction in a deterministic way is a nontrivial task. Therefore, we may wish to instead establish a weaker *restricted optimality* guarantee, which entails proving that the algorithm converges to some matrix $X$ satisfying

$$f(X) \leq \min_{\text{rank}(Y) \leq r'} f(Y),$$
where the rank $r' < r$ proves a more restrictive constraint. In a statistical setting where we are aiming to recover some true low-rank parameter, we might think of $r'$ as the true underlying rank, while $r \geq r'$ is a relaxed rank constraint that we place on our optimization scheme. More generally, we are simply aiming to show that optimizing over rank $r$, while not ensuring the best rank-$r$ solution, is competitive with the best lower-rank solution.

Under these conditions, Liu and Barber [2018] prove that any stationary point $X$ of PGD with step size $\eta = 1/\beta$ satisfies restricted optimality with respect to any rank $r' < r/\kappa^2$. Based on our main result, Theorem 1, the same guarantee also holds for any SOSP of the factorized problem. For completeness, we restate their result along with the new extension to the factorized problem:

**Theorem 4.** Assume that $f(X)$ satisfies $\alpha$-RSC (11) and $\beta$-RSM with respect to rank $r$. Then:

- [Liu and Barber, 2018] For any stationary point $X$ of PGD with step size $\eta = 1/\beta$,
  \[ f(X) \leq \min_{\text{rank}(Y) < r/\kappa^2} f(Y), \]
  i.e. $X$ satisfies restricted optimality with respect to any rank $r' < r/\kappa^2$.

- For any second-order stationary point $X = AB^\top$ of the factorized problem,
  \[ f(AB^\top) \leq \min_{\text{rank}(Y) < r/\kappa^2} f(Y), \]
  i.e. $X = AB^\top$ satisfies restricted optimality with respect to any rank $r' < r/\kappa^2$.

**Proof of Theorem 4.** The first claim is proved by Liu and Barber [2018], while the second claim follows immediately by combining the first claim with Theorem 1 as in the proof of Theorem 2.

Conversely, Liu and Barber [2018] also establish that this factor of $\kappa^2$ is sharp, i.e. restricted optimality cannot be guaranteed relative to rank $r' > r/\kappa^2$. Here we establish the analogous result for the factorized problem. For completeness, we state the two results together.

**Theorem 5.** For any parameters $\beta \geq \alpha > 0$ and any rank $r' > r/\kappa^2$, there exists a function $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ satisfying $\alpha$-RSC (11) and $\beta$-RSM (12) with respect to the rank constraint $r$, such that:

- [Liu and Barber, 2018] There exists a stationary point $X$ of PGD with step size $\eta = 1/\beta$, such that
  \[ f(X) > \min_{\text{rank}(Y) \leq r'} f(Y). \]
- There exists a second-order stationary point \((A, B)\) of the factorized problem, such that

\[
f(AB^\top) > \min_{\text{rank}(Y) \leq r'} f(Y).
\]

This result is proved in Appendix A. Unlike the restricted optimality guarantee above (Theorem 1), this converse result does not follow directly from Liu and Barber [2018]'s work, and instead requires a new construction.

4 Discussion

In this paper, we establish a deep connection between the full-dimensional PGD and the factorized approaches for solving nonconvex low-rank optimization problems. Our main result shows that any SOSP of the factorized problem must also be a stationary point of projected gradient descent algorithms on the original function, connecting naturally the optimization landscape of the unconstrained factorized approaches with the full-dimensional rank-constrained approaches. In particular, this allows us to obtain various types of established optimality results for PGD algorithms and factorized algorithms in a single framework. Overall, our result provides a new perspective on understanding the optimization landscape of the factorized approaches.

While the present work only considers exact stationary points of PGD and exact SOSPs of the factorized problems, finding such points is practically challenging. Standard optimization techniques such as stochastic or perturbed gradient descent are known to converge to an approximate SOSP [Ge et al., 2015; Jin et al., 2017]. Characterizing equivalence between approximate stationary points for full-dimensional PGD versus factorized approaches is therefore of practical interest. Another interesting direction would be to establish similar results under additional constraints on the full matrix \(X = AB^\top\) or on the factorized matrices \(A\) and \(B\).

Acknowledgements

W.H. is supported by the NSF via the TRIPODS program and by Berkeley Institute for Data Science. R.F.B. was partially supported by the NSF via grant DMS–1654076 and by an Alfred P. Sloan fellowship.

References

Rina Foygel Barber and Wooseok Ha. Gradient descent with non-convex constraints: local concavity determines convergence. Information and Inference: A Journal of the IMA, 2018.
Alexander I. Barvinok. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13(2):189–202, 1995.

Stephen Becker, Volkan Cevher, and Anastasios Kyrillidis. Randomized low-memory singular value projection. *arXiv preprint arXiv:1303.0167*, 2013.

Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582, 2016a.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016b.

Srinadh Bhojanapalli, Nicolas Boumal, Prateek Jain, and Praneeth Netrapalli. Smoothed analysis for low-rank solutions to semidefinite programs in quadratic penalty form. *arXiv preprint arXiv:1803.00186*, 2018.

Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016.

Nicolas Boumal, Vladislav Voroninski, and Afonso S Bandeira. Deterministic guarantees for burer-monteiro factorizations of smooth semidefinite programs. *arXiv preprint arXiv:1804.02008*, 2018.

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

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

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

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.
Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In Advances in Neural Information Processing Systems, pages 2973–2981, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. arXiv preprint arXiv:1704.00708, 2017.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In Acm Symposium on Theory of Computing, pages 665–674, 2013.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In Advances in Neural Information Processing Systems, pages 685–693, 2014.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. arXiv preprint arXiv:1703.00887, 2017.

Haoyang Liu and Rina Foygel Barber. Between hard and soft thresholding: optimal iterative thresholding algorithms. arXiv preprint arXiv:1804.08841, 2018.

Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Sharp time–data trade-offs for linear inverse problems. IEEE Transactions on Information Theory, 64(6):4129–4158, 2018.

Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. arXiv preprint arXiv:1609.03240, 2016.

Gábor Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. Mathematics of operations research, 23(2):339–358, 1998.

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

Jie Shen and Ping Li. A tight bound of hard thresholding. The Journal of Machine Learning Research, 18(1):7650–7691, 2017.

Mohammadreza Soltani and Chinmay Hegde. Fast low-rank matrix estimation without the condition number. arXiv preprint arXiv:1712.03281, 2017.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. arXiv preprint arXiv:1507.03566, 2015.
Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18(166):1–43, 2018.

Richard Y Zhang, Cédric Josz, Somayeh Sojoudi, and Javad Lavaei. How much restricted isometry is needed in nonconvex matrix recovery? *arXiv preprint arXiv:1805.10251*, 2018.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015.

Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. The global optimization geometry of low-rank matrix optimization. *arXiv preprint arXiv:1703.01256*, 2017.

### A Additional proofs

**Proof of Lemma 2** Define $Y_t = (A + tA_1)(B + tB_1)^\top$ for $t > 0$. Note that $\|X - Y_t\|_F \to 0$ as $t \to 0$. By definition of $\beta_t(X)$,

$$
\limsup_{t \to 0} \frac{f(Y_t) - f(X) - \langle \nabla f(X), Y_t - X \rangle}{\frac{1}{2} \|X - Y_t\|_F^2} \leq \beta_t(X).
$$

Since $f$ is twice-differentiable at $X$, we can also take a Taylor expansion to see that

$$
\liminf_{t \to 0} \frac{f(Y_t) - f(X) - \langle \nabla f(X), Y_t - X \rangle - \frac{1}{2} \nabla^2 f(X)(Y_t - X, Y_t - X)}{\frac{1}{2} \|X - Y_t\|_F^2} = 0.
$$

Combining these two, we see that

$$
\limsup_{t \to 0} \frac{\nabla^2 f(X)(Y_t - X, Y_t - X)}{\|X - Y_t\|_F^2} \leq \beta_t(X).
$$

Now we calculate this fraction. Since $Y_t - X = t \cdot (AB_1^\top + A_1B^\top) + t^2 \cdot A_1B_1^\top$, we have

$$
\|X - Y_t\|_F^2 = t^2 \|AB_1^\top + A_1B^\top\|_F^2 + O(t^3)
$$

and

$$
\nabla^2 f(X)(Y_t - X, Y_t - X) = t^2 \nabla^2 f(X)(AB_1^\top + A_1B^\top, AB_1^\top + A_1B^\top) + O(t^3),
$$

and therefore,

$$
\limsup_{t \to 0} \frac{\nabla^2 f(X)(Y_t - X, Y_t - X)}{\|X - Y_t\|_F^2} = \frac{\nabla^2 f(X)(AB_1^\top + A_1B^\top, AB_1^\top + A_1B^\top)}{\|AB_1^\top + A_1B^\top\|_F^2}.
$$
(as long as we are not in the degenerate case that \( \| AB_1^T + A_1 B_1^T \|_F = 0 \)--but if this were the case, then the result would hold trivially). Combining everything, we have proved the desired bound.

Proof of Lemma 3. By assumption, \( X_0 \) is a stationary point of PGD for some step size \( \eta > 0 \), meaning that

\[
X_0 = \mathcal{P}_r(X_0 - \eta \nabla f(X_0)).
\]

Since \( \text{rank}(X_1) \leq r \), by definition of the projection operator this implies that

\[
\| X_0 - (X_0 - \eta \nabla f(X_0)) \|_F^2 \leq \| X_1 - (X_0 - \eta \nabla f(X_0)) \|_F^2.
\]

Rearranging terms, we obtain

\[
\langle X_1 - X_0, \nabla f(X_0) \rangle \geq -\frac{1}{2\eta} \| X_0 - X_1 \|_F^2.
\]

Additionally, noting that \( X_0 \) is a solution to the quadratic problem with rank constraint (again by definition of projection), i.e.

\[
X_0 = \arg \min_{\text{rank}(X) \leq r} \| X_0 - \eta \nabla f(X_0) - X \|_F^2,
\]

we can obtain another lower bound on the same quantity \( \langle X_1 - X_0, \nabla f(X_0) \rangle \) by applying [Barber and Ha, 2018, Lemma 7], which proves a first-order optimality condition for rank-constrained optimization:

\[
\langle X_1 - X_0, \nabla f(X_0) \rangle \geq -\frac{1}{2\sigma_r(X_0)} \| \nabla f(X_0) \| \| X_0 - X_1 \|_F^2.
\]

Taking the best of these two lower bounds yields

\[
\langle X_1 - X_0, \nabla f(X_0) \rangle \geq -\frac{1}{2\min \left\{ \frac{1}{\eta}, \frac{\| \nabla f(X_0) \|}{\sigma_r(X_0)} \right\}} \| X_0 - X_1 \|_F^2.
\]

Combined with the \( \alpha \)-RSC assumption, we see that

\[
f(X_1) \geq f(X_0) + \langle X_1 - X_0, \nabla f(X_0) \rangle + \frac{\alpha}{2} \| X_0 - X_1 \|_F^2
\]

\[
\geq f(X_0) + \frac{1}{2} \left( \alpha - \min \left\{ \frac{1}{\eta}, \frac{\| \nabla f(X_0) \|}{\sigma_r(X_0)} \right\} \right) \| X_0 - X_1 \|_F^2. \tag{14}
\]

Applying the same arguments with the roles of \( X_0 \) and \( X_1 \) reversed yields

\[
f(X_0) \geq f(X_1) + \frac{1}{2} \left( \alpha - \min \left\{ \frac{1}{\eta}, \frac{\| \nabla f(X_1) \|}{\sigma_r(X_1)} \right\} \right) \| X_0 - X_1 \|_F^2. \tag{15}
\]
Adding the two inequalities (14) and (15) yields

\[ 0 \geq \frac{1}{2} \left( 2\alpha - \min \left\{ \frac{1}{\eta}, \frac{\|\nabla f(X_0)\|}{\sigma_r(X_0)} \right\} - \min \left\{ \frac{1}{\eta}, \frac{\|\nabla f(X_1)\|}{\sigma_r(X_1)} \right\} \right) \|X_0 - X_1\|_F^2. \]

This implies that either \(X_0 = X_1\), or

\[ \min \left\{ \frac{1}{\eta}, \frac{\|\nabla f(X_0)\|}{\sigma_r(X_0)} \right\} + \min \left\{ \frac{1}{\eta}, \frac{\|\nabla f(X_1)\|}{\sigma_r(X_1)} \right\} \geq 2\alpha, \]

as desired.

\[ \Box \]

**Proof of Lemma 4** Suppose that \(X_\eta\) is the matrix that we obtain after running one step of projected gradient descent on \(\hat{X}\) with step size \(\eta\), i.e.

\[ X_\eta = P_r(\hat{X} - \eta \nabla f(\hat{X})). \]

Since \(\operatorname{rank}(X_\eta), \operatorname{rank}(\hat{X}) \leq r\), invoking \(\beta\)-RSM (12) of \(f\) on \((X_\eta, \hat{X})\), we have

\[ f(X_\eta) \leq f(\hat{X}) + \langle \nabla f(\hat{X}), X_\eta - \hat{X} \rangle + \frac{\beta}{2}\|X_\eta - \hat{X}\|_F^2. \tag{16} \]

Now we first assume \(\eta < 1/\beta\). By definition of projection, we know that

\[ \|\hat{X} - \eta \nabla f(\hat{X}) - X\|_F^2 \leq \|\hat{X} - \eta \nabla f(\hat{X}) - \hat{X}\|_F^2 = \|\eta \nabla f(\hat{X})\|_F^2, \]

hence implying

\[ \langle \nabla f(\hat{X}), X_\eta - \hat{X} \rangle \leq -\frac{1}{2\eta}\|X_\eta - \hat{X}\|_F^2. \]

Plugging into the inequality (16), and using \(\eta < 1/\beta\) along with global optimality of \(\hat{X}\), we obtain

\[ 0 \geq f(\hat{X}) - f(X_\eta) \geq -\langle \nabla f(\hat{X}), X_\eta - \hat{X} \rangle - \frac{\beta}{2}\|\hat{X} - X_\eta\|_F^2 \geq \left( \frac{1}{2\eta} - \frac{\beta}{2} \right) \|X_\eta - \hat{X}\|_F^2. \]

For any \(\eta < 1/\beta\), then, this implies that \(\|X_\eta - \hat{X}\|_F^2 = 0\), i.e. \(X_\eta = \hat{X}\). In other words, \(\hat{X}\) is a stationary point at any step size \(\eta < 1/\beta\).

Next we consider \(\eta = 1/\beta\). Let \(X_\eta\) be defined as before. Now, for any step size \(\eta' < 1/\beta\), our work above proves that \(\hat{X} = P_r(\hat{X} - \eta' \nabla f(\hat{X}))\), which by definition means that

\[ \|\hat{X} - (\hat{X} - \eta' \nabla f(\hat{X}))\|_F \leq \|X_\eta - (\hat{X} - \eta' \nabla f(\hat{X}))\|_F. \]

Now, taking a limit as \(\eta' \to \eta = 1/\beta\), this proves that

\[ \|\hat{X} - (\hat{X} - \eta \nabla f(\hat{X}))\|_F \leq \|X_\eta - (\hat{X} - \eta \nabla f(\hat{X}))\|_F = \min_{\operatorname{rank}(X) \leq r} \|X - (\hat{X} - \eta \nabla f(\hat{X}))\|_F. \]

This proves that \(\hat{X}\) is a (potentially non-unique) solution to the projection step, at step size \(\eta = 1/\beta\). In other words, \(\hat{X}\) is stationary at step size \(\eta = 1/\beta\). \(\Box\)
Proof of Theorem 5: Without loss of generality, take $m \leq n$. Define the matrices

$$X_0 = \sum_{i=1}^{r} e_i e_i^\top$$

and

$$X_1 = \sum_{i=r+1}^{r+r'} e_i e_i^\top,$$

and

$$M = \begin{pmatrix} 0_{r\times r} & I_{r\times (n-r)} \\ 1_{(m-r)\times r} & 0_{(m-r)\times (n-r)} \end{pmatrix}.$$ 

Writing $\circ$ to denote the elementwise product, we will consider the objective function

$$f(X) = -\beta \cdot \langle X_1, X - X_0 \rangle + \frac{\alpha}{2} \cdot \|X - X_0\|_F^2 + \frac{\beta - \alpha}{2} \cdot \|M \circ (X - X_0)\|_F^2,$$

which clearly is $\alpha$-strongly convex and $\beta$-smooth (and therefore trivially satisfies $\alpha$-RSC and $\beta$-RSM). Define

$$A_0 = \begin{pmatrix} I_r \\ 0_{(m-r)\times r} \end{pmatrix} \text{ and } B_0 = \begin{pmatrix} I_r \\ 0_{(m-r)\times r} \end{pmatrix}.$$ 

Then $A_0 B_0^\top = X_0$, and a trivial calculation verifies that

$$f(A_0 B_0^\top) = f(X_0) = 0 > f(\kappa \cdot X_1) \geq \min_{\text{rank}(Y) \leq r'} f(Y).$$ 

Therefore, $A_0 B_0^\top$ does not satisfy restricted optimality relative to the rank $r'$. Now it remains to be shown that the pair $(A_0, B_0)$ is a second-order stationary point of the factorized objective function $g(A, B)$. We can trivially see that

$$\nabla f(X_0) = \beta X_1,$$

and so $\nabla f(X_0)^\top A_0 = 0$ and $\nabla f(X_0) B_0 = 0$, verifying that $(A_0, B_0)$ satisfies the first-order conditions. Now we examine the second-order conditions. We need to prove that, for any pair of matrices $(A_1, B_1)$, the operator $\nabla^2 g(A_0, B_0)$ maps $(A_1, B_1) \times (A_1, B_1)$ to a nonnegative value. Using our earlier calculation (9) to derive $\nabla^2 g(A, B)$, we can calculate

$$\nabla^2 g(A_0, B_0) \left( (A_1, B_1), (A_1, B_1) \right)$$

$$= 2 \langle \nabla f(X_0), A_1 B_1^\top \rangle + \nabla^2 f(X_0) \left( A_0 B_1^\top + A_1 B_0^\top, A_0 B_0^\top + A_0 B_1^\top \right)$$

$$= 2 \beta \cdot \langle X_1, A_1 B_1^\top \rangle + \alpha \cdot \|A_0 B_1^\top + A_1 B_0^\top\|_F^2 + (\beta - \alpha) \cdot \|M \circ (A_0 B_1^\top + A_1 B_0^\top)\|_F^2,$$

where the last step holds by definition of $f$. Now we split the matrices $A_1$ and $B_1$ into block form, writing

$$A_1 = \begin{pmatrix} A_1' \\ A_1'' \end{pmatrix}, \quad B_1 = \begin{pmatrix} B_1' \\ B_1'' \end{pmatrix},$$

and...
where $A'_1$ and $B'_1$ contain the first $r$ rows of $A_1$ and of $B_1$, respectively. Then, plugging in the definitions of $X_1$, $A_0$, $B_0$, and $M$, the expression in (17) can be rewritten as

$$2\beta \cdot \text{trace} \left( A''_1 B''_1^\top \right) + \alpha \cdot \left\| \begin{pmatrix} A'_1 + B'_1^\top & B''_1^\top \\ A''_1 & 0_{(m-r)\times(n-r)} \end{pmatrix} \right\|_F^2 + (\beta - \alpha) \cdot \left\| \begin{pmatrix} 0_{r\times r} & B''_1^\top \\ 0_{(m-r)\times(n-r)} & A''_1 \end{pmatrix} \right\|_F^2.$$

This is trivially lower-bounded by

$$2\beta \cdot \text{trace} \left( A''_1 B''_1^\top \right) + \alpha \cdot \left\| A''_1 \right\|_F^2 + \beta \cdot \left\| B''_1 \right\|_F^2.$$

Using the fact that $|\text{trace}(YZ)| \leq \|Y\|_F \|Z\|_F$ for all matrices $Y$, $Z$, this expression is clearly nonnegative. We have therefore proved that $\nabla^2 g(A_0, B_0) \succeq 0$, thus verifying that $(A_0, B_0)$ is a SOSP and proving the desired result. $\Box$