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

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

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 fixed 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. We demonstrate application of our results to several concrete low-rank optimization problems arising in matrix inverse problems.

1 Introduction

We consider the following low rank optimization problem

$$\min_{X\in\mathbb{R}^{m\times n}} \big\{\mathsf{f}(X) : \mathrm{rank}(X)\le r\big\}, \tag{1}$$

for a differentiable function $\mathsf{f}:\mathbb{R}^{m\times n}\to\mathbb{R}$. 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 $\mathsf{f}(X)$ may already be low-rank, i.e.

$$\widehat{X}\in\operatorname*{arg\,min}_{X\in\mathbb{R}^{m\times n}} \big\{\mathsf{f}(X) : \mathrm{rank}(X)\le r\big\} \subseteq \operatorname*{arg\,min}_{X\in\mathbb{R}^{m\times n}} \mathsf{f}(X).$$

This setting arises naturally in the matrix inverse problems, such as matrix sensing [Recht et al., 2010] and matrix completion [Candès and Recht, 2009], where the low-rank solution typically represents a matrix signal to recover from a fewer number of measurements. In these settings, while there may exist many full rank minimizers due to the

*Department of Statistics, University of California, Berkeley
†Department of Statistics, University of Chicago

nature of under-determined system, enforcing the constraint over the course of an iterative algorithm allows to accurately find the one with low rank [Oymak et al., 2018]. A low-rank solution to the unconstrained minimization problem can also arise in the study of semidefinite programs (SDP)—a wide class of SDP problems[1] admit low rank solution that are global optimal (e.g., Bhojanapalli et al. [2018]). While SDP problems are convex and can be solved by convex optimization algorithms, restricting the search space via rank constraint may still be useful in speeding up the algorithm [Burer and Monteiro, 2003].

In other settings, the rank constraint $\text{rank}(X) \leq r$ will be active in the solution to the minimization problem (1), meaning that the unconstrained minimizer will no longer be low rank and we must necessarily work with the rank constraint in the optimization. In this case, two of the most common optimization strategies in the literature are: 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. (Riemannian optimization [Absil et al., 2009, Vandereycken, 2013, Mishra et al., 2013] is another well-studied approach to optimizaiton under rank constraints which we do not consider in this work. There is also extensive literature on relaxing rank constraint 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 various optimization methods to attempt to find the solution to (1). When working with the full variable, a standard approach is to treat the rank-constrained set as a subset of the Euclidean space $\mathbb{R}^{m\times n}$, and apply constrained optimization algorithms. As our central example of this work, we consider the projected gradient descent method (also known as iterative hard thresholding, see Jain et al. [2014]):

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

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

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

For instance, we might apply any unconstrained optimization techniques to this minimization, which attempt to update each of the two factors A, B. In contrast to the full-dimensional approach, these methods implicitly explore the space of low rank matrix manifold embedded in $\mathbb{R}^{m\times n}$.

Comparing these options naturally raises the following question: is there a connection between the output of full-dimensional approaches such as PGD (2) versus factorized

[1] While canonical forms of SDPs involve linear constraints and do not fall within the framework of (1), here we mainly focus on the penalized formulation of SDPs, as proposed in [Bhojanapalli et al., 2018], i.e., the linear constraints are replaced by a quadratic penalty in the objective function—see Section 2.4.

approaches aiming to solve (3)? Our work is intended to partially answer this question, and further highlighting the implication of this result to a range of low rank estimation problems.

1.1 Comparing full-dimensional vs factorized approaches

In this work we strengthen the connection between solving the rank-constrained optimization problem via its factorized representation (3) versus projecting directly to the constraint (2). Our key finding is that these two approaches, treated more or less separately in the literature, can in fact be considered to be equivalent for a wide class of low-rank optimization problems, and thus lead to the same guarantees in a range of settings. Specifically we can state our main result as follows:

> Any second-order stationary point (SOSP) of the factorized objective function $\mathsf{g}(A, B) = \mathsf{f}(AB^\top)$, must also be a fixed point of projected gradient descent on the original objective function $\mathsf{f}(X)$.

Based on this finding, we further verify the following results:

- In Section 3, under conditions of restricted strong convexity/smoothness on f, we give a range of different optimality guarantees for SOSPs of the factorized objective function. Here the strength of the guarantee (e.g., global or local optimality) varies depending on the strength of our assumptions on problem.
- In Section 4, we specialize these optimality guarantees to several concrete matrix inverse problems arising in low-rank signal recovery, such as matrix sensing, matrix completion, and robust PCA.

As we will see, these results directly follow from our main equivalence result, in combination with some properties of fixed points of PGD (2). It is not the aim of this work to provide novel guarantees for estimation and convergence of these various problems—and indeed, some of these guarantees are already known in the literature, although in other cases new guarantees arise as a byproduct of our main results. Rather, we aim to bring in a new perspective and broaden our understanding on the landscape of nonconvex low-rank minimization problems through our equivalence result.

1.2 Notation

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

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

For a matrix X, we write, respectively, $\|X\|_{\mathrm{F}}$ and $\|X\|$ to denote the Frobenius norm and the spectral norm, while $\|X\|_{2,\infty}$ will be denoted as the largest ℓ_2 norm of any row. The ℓ_0 norm, $\|\cdot\|_0$, will denote the number of nonzero entries in a vector. If $\mathrm{rank}(X) \le r$, we will write $X = U_X \cdot \mathrm{diag}\{\sigma_1, \dots, \sigma_r\} \cdot V_X^\top$ to denote a (possibly non-unique) singular value decomposition of X, with $\sigma_1 \ge \cdots \ge \sigma_r$.

2 Main result

We now turn to our main result, relating critical points of factorized optimization of $\mathsf{g}(A, B) = \mathsf{f}(AB^\top)$ to the fixed points of PGD on the full-dimensional problem $\mathsf{f}(X)$. Before proceeding, we need one additional piece of notation that allow us to quantify the smoothness of f on the space of low-rank matrices:

$$\beta_{\mathsf{local}}(X) = \lim_{\epsilon \to 0} \left\{ \sup_{\substack{0 < \|Y - X\|_{\mathrm{F}} \le \epsilon \\ \mathrm{rank}(Y) \le r}} \frac{\mathsf{f}(Y) - \mathsf{f}(X) - \langle \nabla \mathsf{f}(X), Y - X \rangle}{\frac{1}{2}\|X - Y\|_{\mathrm{F}}^2} \right\}. \tag{4}$$

Note that, if f is twice differentiable, then $\beta_{\mathsf{local}}(X) \le \|\nabla^2 \mathsf{f}(X)\|$.

This local curvature measure will relate to the step size of PGD, since the step size for PGD is typically chosen with respect to the curvature of f—in particular, if the second derivative of f is globally bounded by some β, then a constant step size $\eta \le 1/\beta$ ensures that each step of PGD will make progress towards minimizing f.

2.1 Preliminaries: characterizing critical points

We begin by characterizing a critical point (CP) or fixed point for each of the relevant representations and algorithms.

2.1.1 Critical points, fixed points, and local minima of rank-constrained minimization

First consider the rank-constrained minimization problem (1) over the full-dimensional matrix variable X. For a matrix X with $\mathrm{rank}(X) \le r$,

$$X \text{ is a CP of (1) iff } \begin{cases} \mathrm{rank}(X) = r,\ \nabla \mathsf{f}(X)^\top U_X = 0, \text{ and } \nabla \mathsf{f}(X) V_X = 0, \text{ or} \\ \mathrm{rank}(X) < r \text{ and } \nabla \mathsf{f}(X) = 0. \end{cases} \tag{5}$$

These conditions are necessary for local optimality Rockafellar and Wets [2009, Theorem 6.12]—that is, any local minimum X for the function $\mathsf{f}(X)$ must satisfy (5)—but in general are not sufficient. In particular, we can verify the following stronger property necessary for local optimality:

Lemma 1. *Suppose that X is a local minimum of the rank-constrained optimization problem (1). Then, in addition to the first-order conditions (5), the gradient $\nabla \mathsf{f}(X)$ satisfies*

$$\|\nabla \mathsf{f}(X)\| \leq \beta_{\mathsf{local}}(X) \cdot \sigma_r, \tag{6}$$

where σ_r is the r-th singular value of X.

Next we turn to the PGD algorithm in particular, and characterize its fixed points. Recall that the PGD algorithm has update steps of the form

$$X_{t+1} \leftarrow \mathcal{P}_r\big(X_t - \eta \nabla \mathsf{f}(X_t)\big),$$

where $\eta > 0$ is the step size, while $\mathcal{P}_r$ denotes (possibly non-unique) projection to the rank constraint, i.e., $\mathcal{P}_r(X) \in \arg\min_{\mathrm{rank}(X') \leq r} \|X' - X\|_{\mathrm{F}}$.

A matrix $X \in \mathbb{R}^{m \times n}$ is therefore a *fixed point* of PGD at step size $\eta > 0$ if it satisfies[2]

$$X = \mathcal{P}_r\big(X - \eta \nabla \mathsf{f}(X)\big).$$

By examining this condition, we can easily determine that X is a fixed point of PGD if and only if

$$\nabla \mathsf{f}(X)^\top U_X = 0 \text{ and } \nabla \mathsf{f}(X) V_X = 0 \text{ and } \eta \|\nabla \mathsf{f}(X)\| \leq \sigma_r. \tag{7}$$

Comparing to the result of Lemma 1 and the critical point conditions (5), we see that this implies

$$\left\{ \begin{array}{c} \text{Local minima} \\ \text{of } \min\limits_{\mathrm{rank}(X) \leq r} \mathsf{f}(X) \end{array} \right\} \subseteq \left\{ \begin{array}{c} \text{Fixed pts. of PGD} \\ \text{on } \min\limits_{\mathrm{rank}(X) \leq r} \mathsf{f}(X) \\ \text{with } \eta \leq 1/\beta_{\mathsf{local}} \end{array} \right\} \subseteq \left\{ \begin{array}{c} \text{Critical pts.} \\ \text{of } \min\limits_{\mathrm{rank}(X) \leq r} \mathsf{f}(X) \end{array} \right\}.$$

2.1.2 Critical points of factorized minimization

Next, we will consider the critical points of the factorized objective function $\mathsf{g}(A, B)$, defined over the variables $A \in \mathbb{R}^{m \times r}$ and $B \in \mathbb{R}^{n \times r}$ (with no constraints on these variables). A first-order stationary point (FOSP), or critical point, of g is any pair (A, B) with $\nabla \mathsf{g}(A, B) = (\nabla_A \mathsf{g}(A, B), \nabla_B \mathsf{g}(A, B)) = 0$. By definition of g, we can calculate

$$\nabla_A \mathsf{g}(A, B) = \nabla \mathsf{f}(AB^\top) B \text{ and } \nabla_B \mathsf{g}(A, B) = \nabla \mathsf{f}(AB^\top)^\top A,$$

and therefore,

$$\begin{aligned} &\text{The pair } (A, B) \text{ is a FOSP of } \mathsf{g} \text{ iff } \nabla \mathsf{g}(A, B) = 0, \\ &\qquad \text{or equivalently, } \nabla \mathsf{f}(AB^\top)^\top A = 0 \text{ and } \nabla \mathsf{f}(AB^\top) B = 0. \end{aligned} \tag{8}$$

Comparing to the first-order optimality conditions for the original (full-dimensional) objective function $\mathsf{f}(X)$, given in (5), we obtain the following result (which requires no proof):

[2] If the projection step is not unique, we need to be more precise with our definition. We say that X is a fixed point of PGD at step size η if X is equal to a (possibly non-unique) solution of the projection step, i.e., $X \in \arg\min_{\mathrm{rank}(X') \leq r} \|X' - \big(X - \eta \nabla \mathsf{f}(X)\big)\|_{\mathrm{F}}$.

Lemma 2. *Let $(A, B) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$. If $X = AB^\top$ is a critical point of $\min_{\mathrm{rank}(X)\le r} \mathsf{f}(X)$, then the pair (A, B) is a FOSP of the factorized objective function $\mathsf{g}(A, B)$.*

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 $(A, B) = (\mathbf{0}_{m\times r}, \mathbf{0}_{n\times r})$. This point is always a FOSP of the factorized problem, but in general $X = \mathbf{0}_{m\times n}$ does not correspond to a critical point of $\mathsf{f}(X)$ (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 full-dimensional and the factorized forms of the problem. We will therefore next consider second-order stationary points (SOSPs), or critical points, of the factorized problem, which are characterized by the conditions

$$\nabla \mathsf{g}(A, B) = 0 \text{ and } \nabla^2 \mathsf{g}(A, B) \succeq 0. \tag{9}$$

2.2 Characterization of SOSP for factorized problem

From the discussion above, we see clearly that any fixed point of the PGD is first-order stationary point (FOSP) of the factorized objective function. Our main theoretical result establishes a partial converse to this, proving that any second-order stationary point (SOSP) of the factorized objective function $\mathsf{g}(A, B)$ must also be a fixed point of projected gradient descent on the original function $\mathsf{f}(X)$.

Theorem 1. *Assume that f is twice differentiable, and let $(A, B) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$.*

(a) If (A, B) is a SOSP of the factorized objective function $\mathsf{g}(A, B)$, then $X = AB^\top$ is a fixed point of the projected gradient descent algorithm on $\min_{\mathrm{rank}(X)\le r} \mathsf{f}(X)$ with any step size $\eta \le 1/\beta_{\mathsf{local}}(X)$.

(b) Conversely, if (A, B) is not a SOSP of g, then $X = AB^\top$ is not a local minimum of $\min_{\mathrm{rank}(X)\le r} \mathsf{f}(X)$.

To summarize, our main result (combined with the discussion of Section 2.1) shows that, for the case of a twice-differentiable function f, we have:

$$\left\{\begin{matrix}\text{Local minima}\\ \text{of } \min\limits_{\mathrm{rank}(X)\le r} \mathsf{f}(X)\end{matrix}\right\} \subseteq \left\{\begin{matrix}AB^\top \text{ for SOSPs}\\ (A,B) \text{ of } \mathsf{g}(A,B)\end{matrix}\right\} \subseteq \left\{\begin{matrix}\text{Fixed pts. of PGD}\\ \text{on } \min\limits_{\mathrm{rank}(X)\le r} \mathsf{f}(X)\\ \text{with } \eta \le 1/\beta_{\mathsf{local}}\end{matrix}\right\} \subseteq \left\{\begin{matrix}\text{Critical pts.}\\ \text{of } \min\limits_{\mathrm{rank}(X)\le r} \mathsf{f}(X)\end{matrix}\right\} \subseteq \left\{\begin{matrix}AB^\top \text{ for FOSPs}\\ (A,B) \text{ of } \mathsf{g}(A,B)\end{matrix}\right\}.$$

2.2.1 Regularized factored optimization

The factors A and B are not identifiable in the factored optimization problem—in particular, $\mathsf{g}(A, B) = \mathsf{g}(AC, BC^{-1})$ for any invertible $C \in \mathbb{R}^{r\times r}$. While the product $X = AB^\top$ is in principle not affected by the nonidentifiability of the individual factors, it is known that this issue may lead to instability and numerical issues when solving the factorized minimization problem (3). To alleviate this, it is common to add a regularizer on A and

B to align the two factors on the same scale (e.g., Tu et al. [2015], Zheng and Lafferty [2016], Zhu et al. [2018]). The regularized objective function is

$$\mathsf{g}_{\mathsf{reg}}(A, B) = \mathsf{g}(A, B) + \frac{\lambda}{2}\|A^\top A - B^\top B\|_{\mathrm{F}}^2, \tag{10}$$

for a regularization parameter $\lambda > 0$. In fact, we can verify that our main result, Theorem 1, applies in this setting as well.

Lemma 3. *For any $\lambda > 0$, the result of Theorem 1(a) holds with $\mathsf{g}_{\mathsf{reg}}$ in place of g. Furthermore, a modification of Theorem 1(b) holds: if X is a local minimum of $\min_{\mathrm{rank}(X)\le r} \mathsf{f}(X)$, then there exists a factorization $X = AB^\top$ such that (A, B) is a SOSP of $\mathsf{g}_{\mathsf{reg}}$.*

2.3 Proof of Theorem 1

By definition of g, some simple calculations show that $\nabla^2\mathsf{g}(A, B)$ maps $(A_1, B_1) \times (A_1, B_1)$ to

$$2\langle \nabla \mathsf{f}(X), A_1 B_1 \rangle + \nabla^2 \mathsf{f}(X)\Big(AB_1^\top + A_1 B^\top, AB_1^\top + A_1 B^\top\Big). \tag{11}$$

2.3.1 Claim (a): a SOSP is a fixed point of PGD

Since we assume that $\nabla^2\mathsf{g}(A, B) \succeq 0$ by definition of a SOSP, the calculation in (11) implies that

$$2\langle \nabla \mathsf{f}(X), A_1 B_1^\top \rangle + \nabla^2 \mathsf{f}(X)\Big(AB_1^\top + A_1 B^\top, AB_1^\top + A_1 B^\top\Big) \ge 0 \text{ for all } (A_1, B_1). \tag{12}$$

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

$$\nabla \mathsf{f}(X)^\top A = 0 \text{ and } \nabla \mathsf{f}(X) B = 0. \tag{13}$$

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

Case 1: full rank First suppose $\mathrm{rank}(X) = r$. 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 (13), this implies that $\nabla \mathsf{f}(X) V_X = 0$ and $\nabla \mathsf{f}(X)^\top U_X = 0$. By our earlier characterization (7) of the fixed points of PGD, we therefore only need to check that $\eta \|\nabla \mathsf{f}(X)\| \le \sigma_r(X)$ in order to verify that X is a fixed point of PGD at step size η.

Next, if $\nabla \mathsf{f}(X) = 0$ then X is obviously a fixed point, so from this point on we will consider the case that $\nabla \mathsf{f}(X) \neq 0$. Since we know that $\nabla \mathsf{f}(X)^\top U_X = 0$ while $u_\star$ is the first

left singular vector of $\nabla \mathsf{f}(X)$, this implies that $u_r^\top u_\star = 0$. Similarly $v_r^\top v_\star = 0$. We will consider the curvature of the factorized objective function $\mathsf{g}(A, B)$ in the direction given by $(A_1, B_1) = \big(-u_\star u_r^\top A, v_\star v_r^\top B\big)$. Plugging this choice into our earlier calculation (12) we see that

$$\begin{aligned}\nabla^2 \mathsf{f}(X)\Big(AB_1^\top + A_1 B^\top, AB_1^\top + A_1 B^\top\Big) &\geq -2\langle \nabla \mathsf{f}(X), A_1 B_1^\top \rangle \\ &= 2\langle \nabla \mathsf{f}(X), u_\star u_r^\top A B^\top v_r v_\star^\top \rangle = 2\sigma_r \|\nabla \mathsf{f}(X)\|,\end{aligned}$$

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

Next, we will use the following lemma (proved in Appendix A):

Lemma 4. *Let $\mathsf{f} : \mathbb{R}^{m\times n} \to \mathbb{R}$ 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 $A_1 \in \mathbb{R}^{m\times r}, B_1 \in \mathbb{R}^{n\times r}$,*

$$\nabla^2 \mathsf{f}(X)\big(AB_1^\top + A_1 B^\top, AB_1^\top + A_1 B^\top\big) \leq \beta_{\mathsf{local}}(X) \cdot \|AB_1^\top + A_1 B^\top\|_{\mathrm{F}}^2.$$

Now fix any step size $\eta > 0$ with $\eta \leq 1/\beta_{\mathsf{local}}(X)$. Then by Lemma 4, along with the definitions of A_1 and B_1, we can bound

$$\begin{aligned}\nabla^2 \mathsf{f}(X)\big(AB_1^\top + A_1 B^\top, AB_1^\top + A_1 B^\top\big) &\leq \eta^{-1} \cdot \|AB_1^\top + A_1 B^\top\|_{\mathrm{F}}^2 \\ &= \eta^{-1} \cdot \|AB^\top v_r v_\star^\top - u_\star u_r^\top A B^\top\|_{\mathrm{F}}^2 = \eta^{-1} \cdot \sigma_r(X)^2 \|u_r v_\star^\top - u_\star v_r^\top\|_{\mathrm{F}}^2 = 2\eta^{-1}\sigma_r^2,\end{aligned}$$

where the next-to-last step holds since u_r, v_r are the rth singular vectors of $X = AB^\top$, while the last step holds since $u_r, u_\star$ and $v_r, v_\star$ are pairs of orthogonal unit vectors. Combining everything, and using the fact that $\sigma_r > 0$ since $\mathrm{rank}(X) = r$, we have proved that

$$\eta \|\nabla \mathsf{f}(X)\| \leq \sigma_r.$$

Applying (7), this verifies that X is a fixed point of PGD with step size η, which completes the proof for the rank-r case.

Case 2: rank deficient For the case that $\mathrm{rank}(X) < r$, our proof closely follows that of Bhojanapalli et al. [2018, Lemma1], 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 $\mathrm{rank}(A) < r$. This means that there is some unit vector $w \in \mathbb{R}^r$ with $Aw = 0$. Now consider $(A_1, B_1) = (-u_\star w^\top, c \cdot v_\star w^\top)$ for any $c > 0$. Since (A, B) is a SOSP of the factorized problem, our earlier calculation (12) yields

$$2\langle \nabla \mathsf{f}(X), -c \cdot u_\star w^\top w v_\star^\top \rangle + \nabla^2 \mathsf{f}(X)\Big(c \cdot Awv_\star^\top + u_\star w^\top B^\top, c \cdot Awv_\star^\top + u_\star w^\top B^\top\Big) \geq 0.$$

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

$$\nabla^2 \mathsf{f}(X)\Big(u_\star w^\top B^\top, u_\star w^\top B^\top\Big) \geq 2c\|\nabla \mathsf{f}(X)\|.$$

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 $\|\nabla \mathsf{f}(X)\| = 0$, i.e., $\nabla \mathsf{f}(X) = 0$. Therefore clearly X is a fixed point of projected gradient descent at any step size η.

2.3.2 Claim (b): a local minimum is a SOSP

The second claim follows from a simple Taylor series argument. Suppose that $X = AB^\top$ is a local minimum of $\min_{\mathrm{rank}(X) \le r} \mathsf{f}(X)$. The work in Section 2.1 implies that (A, B) is therefore a FOSP of $\mathsf{g}(A, B)$, that is, $\nabla \mathsf{g}(A, B) = 0$. We therefore only need to verify that $\nabla^2 \mathsf{g}(A, B) \succeq 0$ to prove that (A, B) is a SOSP. Fix any $(A_1, B_1) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$ and let $\delta > 0$. Define

$$X_\delta = (A + \delta A_1)(B + \delta B_1)^\top.$$

Since X is a local minimum, this implies that $\mathsf{f}(X_\delta) \ge \mathsf{f}(X)$ for all sufficiently small $\delta > 0$. Next, taking a Taylor expansion,

$$\begin{aligned}
0 \le \frac{\mathsf{f}(X_\delta) - \mathsf{f}(X)}{\delta^2} &= \frac{\langle \nabla \mathsf{f}(X), X_\delta - X \rangle}{\delta^2} + \frac{1}{2\delta^2} \nabla^2 \mathsf{f}(X)\big(X_\delta - X, X_\delta - X\big) + \mathcal{O}(\delta) \\
&= \langle \nabla \mathsf{f}(X), \frac{AB_1^\top + A_1 B^\top}{\delta} + A_1 B_1^\top \rangle + \frac{1}{2} \nabla^2 \mathsf{f}(X)\big(AB_1^\top + A_1 B^\top, AB_1^\top + A_1 B^\top\big) + \mathcal{O}(\delta) \\
&= \langle \nabla \mathsf{f}(X), A_1 B_1^\top \rangle + \frac{1}{2} \nabla^2 \mathsf{f}(X)\big(AB_1^\top + A_1 B^\top, AB_1^\top + A_1 B^\top\big) + \mathcal{O}(\delta) \\
&= \nabla^2 \mathsf{g}(A, B)\Big((A_1, B_1), (A_1, B_1)\Big) + \mathcal{O}(\delta),
\end{aligned}$$

where the last step applies (11), while the next-to-last step holds since $\nabla \mathsf{f}(X)^\top A = 0$ and $\nabla \mathsf{f}(X) B = 0$ due to the fact that (A, B) is a FOSP (8). Since this bound holds for all sufficiently small $\delta > 0$, taking a limit we see that (A, B) is a SOSP.

2.4 Comparison to related work: penalized SDPs

The comparison of full-dimensional approaches versus factorized approaches has been studied in the context of semidefinite programs. The existing work closest to the results of our paper is the "penalized" form of SDPs [Bhojanapalli et al., 2018]

$$\widehat{X} \in \operatorname*{arg\,min}_{X \in \mathbb{R}^{n \times n}, X \succeq 0} \big\{ \mathsf{f}_0(X) + \mu \sum_{i=1}^{k} (\mathsf{f}_i(X) - a_i)^2 \big\},$$

where $\mathsf{f}_0, \mathsf{f}_1, \ldots, \mathsf{f}_k$ are all *linear* functions. The factorized form of this problem is given by writing $X = AA^\top$, and solving

$$\min_{A \in \mathbb{R}^{n \times r}} \big\{ f_0(AA^\top) + \mu \sum_{i=1}^{k} (\mathsf{f}_i(AA^\top) - a_i)^2 \big\}.$$

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 $r \ge \mathrm{rank}(\widehat{X})$. On the other hand, the factorized

problem is 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 (if one exists); it is also shown that approximate SOSPs are approximately globally optimal. 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 $\widehat{X}$ itself must have rank on the order of $\sqrt{k}$.

Summarizing, the results mentioned above apply in the setting where:

- The optimization problem is a (penalized) SDP, meaning that the functions $\mathsf{f}_0, \mathsf{f}_1, \ldots, \mathsf{f}_k$ are linear and the factorized form is given by $X = AA^\top$,
- The unconstrained global minimizer $\widehat{X}$ 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 $\mathsf{f}(X)$, and X is not necessarily symmetric, i.e., the factorized form is given by $X = AB^\top$,
- The unconstrained global minimizer, $\arg\min_X \mathsf{f}(X)$, may be full rank in general—in the rank-constrained problem, $\widehat{X} \in \arg\min_X \{\mathsf{f}(X) : \mathrm{rank}(X) \leq r\}$, the rank constraint may be active,
- Results no longer apply to finding the global minimum (since this is NP-hard), but instead we study fixed points.

In particular, if a fixed point of the rank-constrained approach is itself a global minimizer, our main result can be made globally, i.e., any SOSP of (3) is also global optimal. 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].[3]

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 $\widehat{X}$ of a particular algorithm. From strongest to weakest, the three main styles of guarantees that appear in the literature are:

[3] More precisely, the global solution to the full SDP may not exist in general and [Bhojanapalli et al., 2018] proves the result when it exists. To ensure existence of the solution, additional conditions on the linear functions $\mathsf{f}_0, \mathsf{f}_1, \ldots, \mathsf{f}_k$ are required; see [Bhojanapalli et al., 2018] for further details.

- Global optimality: the algorithm converges to a global minimizer.
- Local optimality, or basin of attraction: 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 $\mathsf{f}(X) \leq \mathsf{f}(X')$ 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 *α-restricted strong convexity* (abbreviated as α-RSC) with respect to the rank constraint r (Negahban et al. [2012], Agarwal et al. [2010]), meaning that for all $X, Y \in \mathbb{R}^{m \times n}$ with $\mathrm{rank}(X), \mathrm{rank}(Y) \leq r$,

$$\mathsf{f}(Y) \geq \mathsf{f}(X) + \langle \nabla \mathsf{f}(X), Y - X \rangle + \frac{\alpha}{2} \|X - Y\|_{\mathrm{F}}^2. \tag{14}$$

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

$$\mathsf{f}(Y) \leq \mathsf{f}(X) + \langle \nabla \mathsf{f}(X), Y - X \rangle + \frac{\beta}{2} \|X - Y\|_{\mathrm{F}}^2. \tag{15}$$

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 κ:

$$\text{Near-isometry } (\kappa \approx 1) \quad \text{vs.} \quad \text{Arbitrary conditioning } (\kappa \gg 1).$$

We can expect to see $\kappa \approx 1$ in certain well-behaved problems, for instance the matrix sensing problem, where $\mathsf{f}(X)$ 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 $\kappa \approx 1$.

In some cases, the restricted strong convexity and/or restricted smoothness conditions might not be satisfied globally (i.e., for all rank-r matrices), but is satisfied for a more restricted subset of matrices X, Y; in these settings we may write, for instance, that f satisfies α-RSC over a particular subset.

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

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

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 $\mathsf{f}(X)$ exhibits both near-isometry and a vanishing gradient, it is possible to prove convergence to an (approximate) globally optimal estimate $\widehat{X}$. 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., 2018]. (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, or the existence of basin of attraction, 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, 2016], 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 $\widehat{X}$ 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 κ 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 nonconvex 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., $\mathsf{f}(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_{\mathrm{F}}^2$ for a linear operator $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$, 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 $\mathsf{f}(X)$. We will write $\mathbb{R}^{m\times n}_{\mathrm{rank}(r)}$ to denote the set of $m \times n$ matrices with rank $\leq r$.

Lemma 5. *Suppose that $\mathsf{f} : \mathbb{R}^{m\times n} \to \mathbb{R}$ satisfies α-RSC (14) over a subset $\mathcal{X} \subseteq \mathbb{R}^{m\times n}_{\mathrm{rank}(r)}$, where $\alpha > 0$. If $X_0, X_1 \in \mathcal{X}$ are both fixed points of PGD run with step size $\eta_0 > 0$ or $\eta_1 > 0$, respectively, then one of the following must hold:*

- $X_0 = X_1$, *or*
- $\mathrm{rank}(X_0) = r$ *and* $\mathrm{rank}(X_1) < r$ *and* $\frac{\|\nabla \mathsf{f}(X_0)\|}{\sigma_r(X_0)} \geq 2\alpha$, *or*
- $\mathrm{rank}(X_1) = r$ *and* $\mathrm{rank}(X_0) < r$ *and* $\frac{\|\nabla \mathsf{f}(X_1)\|}{\sigma_r(X_1)} \geq 2\alpha$, *or*
- $\mathrm{rank}(X_0) = \mathrm{rank}(X_1) = r$ *and*

$$\frac{\|\nabla \mathsf{f}(X_0)\|}{\sigma_r(X_0)} + \frac{\|\nabla \mathsf{f}(X_1)\|}{\sigma_r(X_1)} \geq 2\alpha.$$

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

Lemma 6. *Suppose that $\mathsf{f} : \mathbb{R}^{m\times n} \to \mathbb{R}$ satisfies β-RSM (15) over an open subset $\mathcal{X} \subseteq \mathbb{R}^{m\times n}_{\mathrm{rank}(r)}$. If $\widehat{X}$ is a global minimizer (i.e., $\mathsf{f}(\widehat{X}) = \min_{\mathrm{rank}(X)\leq r} \mathsf{f}(X)$) and $\widehat{X} \in \mathcal{X}$, then $\widehat{X}$ is a fixed 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 $\mathsf{f}(X)$ satisfies the near-isometry property, with condition number $\kappa < 2$, we can obtain global optimality guarantees for both PGD and factorized methods whenever $\|\nabla \mathsf{f}(X)\|$ is sufficiently small, i.e., the vanishing gradient condition. (See Section 3.1 for related existing results in the literature.)

Theorem 2. *Assume that $\mathsf{f}(X)$ satisfies α-RSC (14) and β-RSM (15) over an open subset $\mathcal{X} \subseteq \mathbb{R}^{m\times n}_{\mathrm{rank}(r)}$, and that $\beta < 2\alpha$. Suppose $\widehat{X}$ is a global minimizer, i.e., $\mathsf{f}(\widehat{X}) = \min_{\mathrm{rank}(X)\le r} \mathsf{f}(X)$. If $\widehat{X} \in \mathcal{X}$ and $\widehat{X}$ satisfies*

$$\textit{Either } \mathrm{rank}(\widehat{X}) < r, \textit{ or } \mathrm{rank}(\widehat{X}) = r \textit{ and } \|\nabla \mathsf{f}(\widehat{X})\| < (2\alpha - \beta)\cdot \sigma_r(\widehat{X}) \ ,$$

then

- *$\widehat{X}$ is the unique fixed point of PGD in $\mathcal{X}$ for any step size $1/(2\alpha) < \eta \le 1/\beta$ in the case that $\mathrm{rank}(\widehat{X}) < r$, or in the case $\mathrm{rank}(\widehat{X}) = r$, for any step size satisfying*

$$\frac{1}{2\alpha - \frac{\|\nabla \mathsf{f}(\widehat{X})\|}{\sigma_r(\widehat{X})}} < \eta \le \frac{1}{\beta}. \tag{16}$$

- *If $X = AB^\top \in \mathcal{X}$ where (A, B) is a SOSP of $\mathsf{g}(A, B)$, then $X = \widehat{X}$.*

Note that, in the case that $\mathrm{rank}(\widehat{X}) = r$, due to the condition $\|\nabla \mathsf{f}(\widehat{X})\| < (2\alpha-\beta)\cdot\sigma_r(\widehat{X})$ the interval (16) given for step size η is always non-empty.

Remark 1. *In some examples, such as the matrix sensing problem discussed later in Section 4.1, the RSC/RSM conditions will hold globally, i.e., for $\mathcal{X} = \mathbb{R}^{m\times n}_{\mathrm{rank}(r)}$; this is why we use the term "global optimality" to describe this result. In other settings, the RSC/RSM conditions may not hold universally over all rank-r matrices but hold for a subset of matrices, e.g., all matrices satisfying an incoherence condition such as in the robust PCA problem (Section 4.2); the above theorem is formulated to cover this type of scenario as well even though the term "global optimality" may no longer apply.*

Theorem 2 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 κ is crucial to achieving global optimality guarantees. For instance, Zhang et al. [2018, Example 3] construct an example of objective function $\mathsf{f}(X)$ with $\beta = 3\alpha$, i.e., $\kappa = 3$, where there exists a fixed 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; see also Zhang et al. [2019] for the sufficient and necessary conditions when rank $r = 1$.

We now compare this result with some recent works in the literature. The first part of Theorem 2, i.e., the result for fixed 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.[4] 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., 2016b, Ge et al., 2016, 2017]. Similarly, Zhu et al. [2018] also establish "no spurious local minima" under conditions similar to Theorem 2, i.e., when $\mathsf{f}(X)$ satisfies α-RSC and β-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.

Proof of Theorem 2. First we consider PGD. By Lemma 6, we know that $\widehat{X}$ is a fixed point for any $\eta \le 1/\beta$.

We first consider the case that $\operatorname{rank}(\widehat{X}) < r$. Let $X \in \mathcal{X}$ be another fixed point of PGD for any step size $1/2\alpha < \eta \le 1/\beta$. Suppose that $X \neq \widehat{X}$. Then applying Lemma 5, we must have $\operatorname{rank}(X) = r$ with $\frac{\|\nabla \mathsf{f}(X)\|}{\sigma_r(X)} \ge 2\alpha$. However, by (7), we have $\|\nabla \mathsf{f}(X)\| \le \eta^{-1}\sigma_r(X) < 2\alpha\sigma_r(X)$, which is a contradiction.

Next, consider the case that $\operatorname{rank}(\widehat{X}) = r$, and let $X \in \mathcal{X}$ be another fixed point of PGD for any step size η satisfying (16). Suppose $X \neq \widehat{X}$. By Lemma 5, we either have $\operatorname{rank}(X) < r$ and $\frac{\|\nabla \mathsf{f}(\widehat{X})\|}{\sigma_r(\widehat{X})} \ge 2\alpha$, or alternatively $\operatorname{rank}(X) = r$ and

$$2\alpha \le \frac{\|\nabla \mathsf{f}(\widehat{X})\|}{\sigma_r(\widehat{X})} + \frac{\|\nabla \mathsf{f}(X)\|}{\sigma_r(X)} \le \frac{\|\nabla \mathsf{f}(\widehat{X})\|}{\sigma_r(\widehat{X})} + \frac{1}{\eta}$$

by applying (7). In either case, this contradicts our assumption (16) on η.

Next we turn to the factorized setting. Let $X = AB^\top \in \mathcal{X}$ where (A, B) is a SOSP of $\mathsf{g}(A, B)$. Comparing the definition of β-RSM over $\mathcal{X}$ with that of the local smoothness parameter $\beta_{\mathsf{local}}(X)$ defined in (4), we can see that since $\mathcal{X} \subseteq \mathbb{R}^{m\times n}_{\operatorname{rank}(r)}$ is an open subset, $\beta_{\mathsf{local}}(X) \le \beta$ by definition, and therefore $\eta \le 1/\beta_{\mathsf{local}}(X)$. Therefore, applying our main result, Theorem 1, we see that X must be a fixed point of PGD at step size $\eta = 1/\beta$, which proves that $X = \widehat{X}$ by our work above. ∎

3.2.2 Local optimality

Next we turn to the local optimality guarantees, i.e., the existence of local region of attraction, that can be obtained when f exhibits a vanishing gradient, but may have an arbitrarily large condition number κ. (See Section 3.1 for related existing results in the literature.)

[4] 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.

Theorem 3. *Assume that* $\mathsf{f}(X)$ *satisfies* α*-RSC* (14) *over a subset* $\mathcal{X} \subseteq \mathbb{R}^{m\times n}_{\mathrm{rank}(r)}$. *Assume that* $\widehat{X}$ *is a global minimizer, i.e.,* $\mathsf{f}(\widehat{X}) = \min_{\mathrm{rank}(X)\leq r} \mathsf{f}(X)$, *that* $\widehat{X} \in \mathcal{X}$, *and that* $\widehat{X}$ *satisfies*

$$\textit{Either } \mathrm{rank}(\widehat{X}) < r, \textit{ or } \mathrm{rank}(\widehat{X}) = r \textit{ and } \|\nabla \mathsf{f}(\widehat{X})\| < \alpha \cdot \sigma_r(\widehat{X}) \ .$$

Let

$$\mathcal{N} = \left\{ X \in \mathcal{X} : \mathrm{rank}(X) < r \textit{ or } \frac{\|\nabla \mathsf{f}(X)\|}{\sigma_r(X)} < 2\alpha \right\}$$

in the case that $\mathrm{rank}(\widehat{X}) < r$, *or*

$$\mathcal{N} = \left\{ X \in \mathcal{X} : \mathrm{rank}(X) < r \textit{ or } \frac{\|\nabla \mathsf{f}(\widehat{X})\|}{\sigma_r(\widehat{X})} + \frac{\|\nabla \mathsf{f}(X)\|}{\sigma_r(X)} < 2\alpha \right\}$$

in the case that $\mathrm{rank}(\widehat{X}) = r$. *Then:*

- *For any fixed point* X *of PGD with any step size* $\eta > 0$, *if* $X \in \mathcal{N}$, *then* $X = \widehat{X}$.
- *If* $X = AB^\top \in \mathcal{N}$ *where* (A, B) *is a SOSP of* $\mathsf{g}(A, B)$, *then* $X = \widehat{X}$.

We note that $\widehat{X} \in \mathcal{N}$ by the assumptions of the theorem. In this setting where κ 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 ensure the existence of regions of attraction $\mathcal{N}$ within which the global minimum $\widehat{X}$ will be discovered, for both the full-dimensional and factorized methods.

To compare with the existing results, the first part of Theorem 3 (for fixed 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 $\widehat{X}$ [Candes et al., 2015, Zheng and Lafferty, 2015, 2016, Tu et al., 2015, Bhojanapalli et al., 2016a, Jain et al., 2013]. While these problems satisfy the near-isometry property with $\kappa \approx 1$, 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 1, we know that $\widehat{X}$ is a fixed point for PGD with step size $\eta \leq 1/\beta_{\mathsf{local}}(\widehat{X})$. (Note that β_{local} may be arbitrarily large, but must be finite since f is twice differentiable.) Suppose that $X \in \mathcal{N}$ is another fixed point at some (potentially different) step size $\eta > 0$, with $X \neq \widehat{X}$. Applying Lemma 5 with $X_0 = \widehat{X}$ and $X_1 = X$ yields a contradiction to the definition of $\mathcal{N}$, either for $\mathrm{rank}(\widehat{X}) < r$ or $\mathrm{rank}(\widehat{X}) = r$, unless $X = \widehat{X}$.

Next we turn to factorized setting. By Theorem 1 we know that any $X = AB^\top \in \mathcal{X}$ for a SOSP (A, B) must be a fixed point of PGD at any step size $\eta \leq 1/\beta_{\mathsf{local}}(X)$ (again $\beta_{\mathsf{local}}(X)$ can be extremely large). If also $X \in \mathcal{N}$ then this proves that $X = \widehat{X}$ 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 $\|\nabla \mathsf{f}(\widehat{X})\|$ 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

$$\mathsf{f}(X) \leq \min_{\mathrm{rank}(Y) \leq r'} \mathsf{f}(Y),$$

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 fixed 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 $\mathsf{f}(X)$ satisfies α-RSC (14) and β-RSM (15) over an open subset $\mathcal{X} \subseteq \mathbb{R}^{m\times n}_{\mathrm{rank}(r)}$. Let $\kappa = \beta/\alpha$. Then:*

- *[Liu and Barber, 2018] For any fixed point $X \in \mathcal{X}$ of PGD with step size $\eta = 1/\beta$,*

$$\mathsf{f}(X) \leq \min_{\mathrm{rank}(Y) < r/\kappa^2, Y \in \mathcal{X}} \mathsf{f}(Y), \tag{17}$$

  i.e., X satisfies restricted optimality with respect to any rank $r' < r/\kappa^2$ within $\mathcal{X}$.

- *For any $X = AB^\top \in \mathcal{X}$ where (A, B) is a SOSP of the factorized problem $\mathsf{g}(A, B)$,*

$$\mathsf{f}(AB^\top) \leq \min_{\mathrm{rank}(Y) < r/\kappa^2, Y \in \mathcal{X}} \mathsf{f}(Y),$$

  i.e., $X = AB^\top$ satisfies restricted optimality with respect to any rank $r' < r/\kappa^2$ within $\mathcal{X}$.

Proof of Theorem 4. The first claim follows by the definition of α-RSC used on the two points X, Y together with Lemma 1 in Liu and Barber [2018], which bounds the restricted concavity of hard thresholding. 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 κ^2 is sharp in general (on all low-rank matrices), 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 $\mathsf{f} : \mathbb{R}^{m\times n} \to \mathbb{R}$ satisfying α-RSC (14) and β-RSM (15) over $\mathbb{R}^{m\times n}_{\mathrm{rank}(r)}$, such that:*

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

This result is proved in Appendix A. Unlike the restricted optimality guarantee above (Theorem 4), this converse result does not follow directly from Liu and Barber [2018]'s work, and instead requires a new construction.

4 Applications

In this section we apply our framework developed in Section 2 and/or Section 3 to several concrete low-rank optimization problems, including matrix sensing, matrix completion, and robust PCA. These problems typically involve an unknown ground truth matrix $X_\star \in \mathbb{R}^{m\times n}$ that is low-rank and the goal is to accurately recover it from a few or sparse or corrupted measurements. In many cases, $X_\star$ itself becomes a global minimizer of the rank-constrained minimization problem (1) in which case we also denote by $X_\star$ (instead of $\widehat{X}$) to represent a global minimizer.

4.1 Matrix sensing

In the matrix sensing problem [Recht et al., 2010], we aim to recover a low rank matrix $X_\star \in \mathbb{R}^{m\times n}$ given k linear observations $b_1 = \langle L_1, X_\star\rangle, \ldots, b_k = \langle L_k, X_\star\rangle$. Therefore, the least square objective function $\mathsf{f} : \mathbb{R}^{m\times n} \to \mathbb{R}$ for the matrix sensing problem takes the form

$$\mathsf{f}(X) = \frac{1}{2k}\sum_{i=1}^{k} \left(\langle L_i, X\rangle - b_i\right)^2. \tag{18}$$

The corresponding rank-r factorized objective function $\mathsf{g} : \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} \to \mathbb{R}$ is defined as

$$\mathsf{g}(A, B) = \mathsf{f}(AB^\top)$$

To conform with our notion of condition number, we define the following set of sensing matrices:

$$\mathcal{L}(\alpha, \beta, r) = \{(L_1, \ldots, L_k) : \text{the map } X \mapsto \frac{\sum_{i=1}^k \langle L_i, X\rangle^2}{2k} \text{ is } \alpha\text{-RSC and } \beta\text{-RSM} \\ \text{with respect to the rank constraint } r\}.$$

Then direct application of our results in Section 3 gives the following lemma (without proof).

Lemma 7. *Consider a matrix sensing model with a rank-r matrix $X_\star$. Define the objective function $\mathsf{f}(X)$ as in equation (18). Then*

- *If the sensing matrices satisfy $(L_1, \ldots, L_k) \in \mathcal{L}(\alpha, \beta, r)$ with $\beta < 2\alpha$, then for any second-order stationary point (A, B) of the factorized objective g, $AB^\top = X_\star$.*
- *Define the following neighborhood of $X_\star$,*

  $$\mathcal{N}(X_\star) = \{X \in \mathbb{R}^{m\times n}_{\mathrm{rank}(r)} : \mathrm{rank}(X) < r \text{ or } \|\nabla \mathsf{f}(X)\| < 2\alpha \cdot \sigma_r(X)\}.$$

  If the sensing matrices satisfy $(L_1, \ldots, L_k) \in \mathcal{L}(\alpha, \beta, r)$, then for any second-order stationary point (A, B) of the factorized objective g, if $X = AB^\top \in \mathcal{N}(X_\star)$ then $X = X_\star$.
- *If the sensing matrices satisfy $(L_1, \ldots, L_k) \in \mathcal{L}(\alpha', \beta', r')$ with $r' > (\frac{\beta'}{\alpha'})^2 r$, then for any second-order stationary point (A, B) of the rank-r' factorized objective function $\mathsf{g}_{r'}$, $\mathsf{g}_{r'}(A, B) = 0$.[5] Since f satisfies α'-RSC over $\mathbb{R}^{m\times n}_{\mathrm{rank}(r')}$ and $\mathsf{f}(X_\star) = 0$ while $\nabla \mathsf{f}(X_\star) = 0$, this further implies that $AB^\top = X_\star$.*

In the simplest setting, the sensing matrices $(L_1, \ldots, L_k)$ are drawn i.i.d. from standard normal distribution $\mathcal{N}(0, 1)$, then, with high probability, $(L_1, \ldots, L_k) \in \mathcal{L}(\alpha, \beta, r)$ with $\beta < 2\alpha$ (e.g., Recht et al. [2010]). In this case global optimality of SOSPs of g follows from Lemma 7 in a straightforward manner. In the more general setting where the sensing matrices $(L_1, \ldots, L_k)$ are drawn i.i.d. from normal distribution $\mathcal{N}(0, \Sigma)$ with covariance matrix $\Sigma \in \mathbb{R}^{mn\times mn}$, Agarwal et al. [2010, Lemma 7] proves that with high probability $(L_1, \ldots, L_k) \in \mathcal{L}(\alpha, \beta, r)$ with $\alpha = c_1\lambda_{\mathsf{min}}(\Sigma)$ and $\beta = c_2\lambda_{\mathsf{max}}(\Sigma)$ for some $c_1, c_2 > 0$. In this case, Lemma 7 provides respectively local and restricted optimality guarantees for SOSPs of the factorized problem. In particular, we observe that any SOSPs (A, B) of the factorized problem still achieve the global minimizer, i.e. $AB^\top = X_\star$, if we over-parametrize f with rank $r' \approx \lambda^2_{\mathsf{max}}(\Sigma)/\lambda^2_{\mathsf{min}}(\Sigma) \cdot r$. This result has not been known previously in the literature of matrix sensing but somewhat follows directly from our main correspondence result Theorem 1.

[5] The rank-r' objective $\mathsf{g}_{r'}$ is defined as the function $\mathsf{g}_{r'}(A, B) = \mathsf{f}(AB^\top)$ where the factorization $X = AB^\top$ is over-parametrized by rank $r' > r$, i.e. $A \in \mathbb{R}^{m\times r'}$ and $B \in \mathbb{R}^{n\times r'}$.

4.2 Robust PCA

Robust PCA [Candès et al., 2011, Chandrasekaran et al., 2011] refers to the decomposition of the data matrix $D_\star$ into a low-rank component $X_\star$ and a sparse component $S_\star$, so that the sum of two components recover the original matrix. One can view the sparse component to be some outliers which we wish to separate from the low-rank signal. Concretely, suppose that we are given $D_\star = X_\star + S_\star \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(X_\star) \le r$ and where $S_\star$ is s-sparse in each column, then we consider the following minimization problem

$$\min_X \left\{ \mathsf{f}(X) = \frac{1}{2} \min_{S\in\mathcal{S}} \|D_\star - (X+S)\|_{\mathrm{F}}^2 : \mathrm{rank}(X) \le r \right\}. \tag{19}$$

Here to specify the sparsity of the sparse component S, we set

$$\mathcal{S} = \{S \in \mathbb{R}^{m\times n} : \|S_j\|_0 \le \|S_{\star j}\|_0 = s \text{ for } j = 1,\dots,n\},$$

where S_j and $S_{\star j}$ denote j-th columns of S and $S_\star$ respectively.

Before we apply our results to the above problem, note that the factorized objective function $\mathsf{g}(A,B) = \mathsf{f}(AB^\top)$ in (19) is not twice-differentiable with respect to (A,B). To extend our discussion on this setting, at any point X with rank $\le r$, we consider the following majorization function of f:

$$\mathsf{f}_X(\widetilde{X}) = \frac{1}{2}\|D_\star - (\widetilde{X} + S(X))\|_{\mathrm{F}}^2,$$

where we define $S(X) \in \arg\min_{S\in\mathcal{S}} \|D_\star - (X+S)\|_{\mathrm{F}}^2$. Then it is easy to see that $\mathsf{f}_X(\widetilde{X})$ majorizes the original function f while matches at X up to the first-order term, i.e. $\mathsf{f}_X(\widetilde{X}) \ge \mathsf{f}(\widetilde{X})$ and $\mathsf{f}_X(X) = \mathsf{f}(X)$ and $\nabla\mathsf{f}_X(X) = \nabla\mathsf{f}(X)$. Denoting $\mathsf{g}_X(\widetilde{A},\widetilde{B}) = \mathsf{f}_X(\widetilde{A}\widetilde{B}^\top)$ to be the factorized form of f_X, we utilize the following version of second-order stationary points of g, which is defined as (see also Ge et al. [2017, Definition 5])

$$\nabla\mathsf{g}(A,B) = 0 \text{ and } \nabla^2\mathsf{g}_X(A,B) \succeq 0. \tag{20}$$

That is, we say (A,B) is a second-order stationary point of g if it satisfies the conditions (20) with $X = AB^\top$.

Next suppose that $X_\star = A_\star B_\star^\top$, where $A_\star = U_\star\sqrt{\Sigma_\star}$ and $B_\star = V_\star\sqrt{\Sigma_\star}$, and $X_\star = U_\star\Sigma_\star V_\star^\top$ be the rank-r SVD of $X_\star$. It is well-known that the robust PCA model suffers from non-identifiability issue and in particular we cannot recover the true matrix $X_\star$ if $X_\star$ is itself both low-rank and *sparse*. To prevent this, we assume that $X_\star$ is incoherent relative to the sparse matrices [Candès et al., 2011], i.e.,

$$\|A_\star\|_{2,\infty} \le \sqrt{\sigma_1(X_\star)\frac{\mu r}{m}}, \quad \|B_\star\|_{2,\infty} \le \sqrt{\sigma_1(X_\star)\frac{\mu r}{n}}. \tag{21}$$

With this definition in place, now suppose that (A,B) is a SOSP of g with $X = AB^\top$. Since $\mathsf{f}_X(\cdot)$ trivially satisfies RSC (14) and RSM (15) with $\alpha = \beta = 1$ (over the entire space

of matrices $\mathbb{R}^{m\times n}$), and at a global minimum $\widetilde{X}_{\mathsf{global}}$, i.e., $\mathsf{f}(\widetilde{X}_{\mathsf{global}}) = \min_{\mathrm{rank}(\widetilde{X})\le r} \mathsf{f}_X(\widetilde{X})$,

$$\|\nabla \mathsf{f}_X(\widetilde{X}_{\mathsf{global}})\| = \sigma_{r+1}(D_\star - S(X)) < \sigma_r(D_\star - S(X)) = (2\alpha-\beta)\sigma_r(\widetilde{X}_{\mathsf{global}}),$$[6]

our result Theorem 2 applies to show that $AB^\top = \widetilde{X}_{\mathsf{global}}$ is the unique fixed point of PGD run on $\mathsf{f}_X(\cdot)$ at step size $\eta = 1/\beta = 1$. In other words, $X = \mathcal{P}_r\big(D_\star - S(X)\big)$.

Furthermore, by definition of $S(X)$, this means that $(X, S(X))$ is together a joint fixed point of PGD with step sizes $\eta_X = \eta_S = 1$ run on the joint problem

$$\min_{X,S\in\mathbb{R}^{m\times n}} \{\mathsf{f}_{\mathsf{joint}}(X,S) = \frac{1}{2}\|D_\star - (X+S)\|_{\mathrm{F}}^2 : \mathrm{rank}(X)\le r,\ S\in\mathcal{S}\}. \tag{22}$$

Under the assumption that $X_\star$ is μ-incoherent and each column of $S_\star$ is s-sparse, we can then prove that the joint objective function $\mathsf{f}_{\mathsf{joint}}$ exhibits *joint* restricted strong convexity and restricted smoothness over the pairs $\big((X,S),(X_\star,S_\star)\big)$, with $2\alpha > \beta$, whenever $(X,S)\in\mathbb{R}^{m\times n}_{\mathrm{rank}(r)}\times\mathcal{S}$ *and* X is incoherent. This allows simple extension of (first part of) Theorem 2 to the joint minimization problem (22) to guarantee that $(X, S(X))$ is the unique fixed point of PGD for joint minimization, as long as X is incoherent relative to sparse matrices. Since $(X_\star, S_\star)$ is trivially a fixed point, this means that $(X, S(X)) = (X_\star, S_\star)$, and in particular we get $AB^\top = X_\star$. We state this result in the following lemma, whose proof is given in Appendix A.

Lemma 8. *Consider the robust PCA problem (19). Suppose that the rank-r matrix $X_\star$ is μ-incoherent (21), and the sparse matrix $S_\star$ is s-sparse in each column. Let $\kappa(X_\star) = \sigma_1(X_\star)/\sigma_r(X_\star)$. Then there exists a constant $c_1 > 0$ such that the following holds: for any second-order stationary point (A,B) of the factorized objective g, as in equation (20), if (A,B) satisfies the following conditions,*

$$A^\top A = B^\top B^{7} \text{ and } \|A\|_{2,\infty} \le c_2\sqrt{\sigma_1(X_\star)\frac{\mu r}{m}} \text{ and } \|B\|_{2,\infty} \le c_2\sqrt{\sigma_1(X_\star)\frac{\mu r}{n}}, \tag{23}$$

for some constant $c_2 > 0$ such that $4c_2\sqrt{\frac{\kappa(X_\star)\mu r s}{\min\{m,n\}}} \le c_1$, then $AB^\top = X_\star$. In other words, $X_\star = A_\star B_\star^\top$ is the unique "incoherent" second-order stationary point of g.

The result of the lemma requires the stationary point (A,B) to be μ-incoherent (23). To enforce the incoherence on the factored matrices (A,B), some works are focused on putting explicit incoherence penalty/constraint at each iterate (e.g., Chen and Wainwright [2015], Zheng and Lafferty [2016], Ge et al. [2017]); while other works have proved that each update of the factorized function g stays incoherent near the true matrix $X_\star$ *without*

[6] It can be shown that the pathological case, $\sigma_r(D_\star - S(X)) = \sigma_{r+1}(D_\star - S(X))$, does not occur under an additional assumption on the magnitude of the true sparse matrix $S_\star$, see Appendix B for further details; see also Ge et al. [2017]. To simplify the presentation, here we assume this is the case.

[7] Note that this condition can be easily imposed on the factors A and B if we add a regularizer to the factorized objective function. In particular, by Zhu et al. [2018, Theorem 3], any critical point (A,B) of $\mathsf{g}_{\mathsf{reg}}$ satisfies $A^\top A = B^\top B$ and the correspondence result still holds by Lemma 3. See Section 2.2.1.

explicit incoherence regularization (e.g., Ma et al. [2017], Chen et al. [2019]). In practice simple algorithms such as gradient descent are observed to work well without incoherence regularizaiton, even globally. Examining the incoherence property of any second-order stationary point of g, or a fixed point of PGD in the full-dimensional space, is therefore an interesting direction which we leave for future study.

4.3 Matrix completion

Next consider the matrix completion minimization problem ([Candès and Recht, 2009, Negahban and Wainwright, 2012]) where we are given an unknown low-rank matrix $X_\star \in \mathbb{R}^{m\times n}$ while only a subset $\Omega \subset [m]\times[n]$ of entries are observed. Here we assume each entry $(i,j)\in\Omega$ of $X_\star$ is observed independently with probability p. Writing $\mathcal{P}_\Omega(X)$ to denote the matrix whose entries are set to 0 on Ω^c, i.e., $(\mathcal{P}_\Omega(X))_{ij} = X_{ij}\cdot \mathbf{1}_{(i,j)\in\Omega}$, we solve the following minimization problem

$$\min_X \left\{ \mathsf{f}(X) = \frac{1}{2p}\|\mathcal{P}_\Omega(X - X_\star)\|_{\mathrm{F}}^2 : \mathrm{rank}(X)\le r \right\}. \tag{24}$$

As in the case of robust PCA, the matrix completion problem is ill-posed without any incoherence type of conditions on the true matrix—indeed, if $X_\star$ is sparse, $\mathcal{P}_\Omega(X_\star)$ is likely to be a zero matrix and the optimization probem (24) owns a trivial solution which will be far from $X_\star$. To prevent this pathological case and allow reliable estimation, we therefore focus on recovering the incoherent matrix, as defined in (21) [Candès and Recht, 2009].

Recent results have verified that the matrix completion objective function is locally well-behaved near the true matrix if it satisfies the incoherence condition. Specifically, if the matrix is initialized within $\mathcal{O}(\sigma_r(X_\star))$-neighborhood of $X_\star$, then with high probability the factorized objective $\mathsf{g}(A,B) = \mathsf{f}(AB^\top)$ satisfies restricted strong convexity and restricted smoothness type of conditions on the space of factored matrices (A,B) [Chen and Wainwright, 2015, Zheng and Lafferty, 2016, Ge et al., 2017, Ma et al., 2017] (here randomness arises from the sampling operator Ω). Adapting this result to the original function f, and combining with Theorem 3, we can characterize the basin of attraction for matrix completion model.

Lemma 9. *Consider the matrix completion problem (24). Suppose that the rank-r matrix $X_\star$ is μ-incoherent (21). Let $\kappa(X_\star) = \sigma_1(X_\star)/\sigma_r(X_\star)$, and suppose the sampling probability $p \ge c_1 \frac{\mu^2 r^2 \kappa^4(X_\star)(m+n)\log(m+n)}{mn}$ for $c_1 > 0$. Define the local region around $X_\star$ given by*

$$\mathcal{N}(X_\star) = \Big\{ X\in\mathbb{R}^{m\times n}_{\mathrm{rank}(r)} : \|X - X_\star\|_{\mathrm{F}} \le 0.1\kappa^{-1}(X_\star)\sigma_r(X_\star),\ \text{and}\ X = AB^\top \text{ where } (A,B) \text{ satisfies (23)} \Big\}.$$

Then the following holds with probability larger than $1-\mathcal{O}(\min\{m,n\}^{-3})$: for any second-order stationary point (A,B) of the factorized objective g, if $X = AB^\top \in \mathcal{N}(X_\star)$, then $X = X_\star$. In other words, $X_\star = A_\star B_\star^\top$ is the unique "incoherent" second-order stationary point of g in the region $\mathcal{N}(X_\star)$.

The proof is given in Appendix A.

The initialization condition, i.e. the condition $\|X - X_\star\|_{\mathrm{F}} \leq 0.1\kappa^{-1}(X_\star)\sigma_r(X_\star)$ in $\mathcal{N}(X_\star)$, can typically be achieved via spectral initialization, or relaxing the rank constraint to a convex constraint (such as nuclear-norm penalty) and solving the corresponding convex problem. Similarly to the case of robust PCA, the incoherence of the factored matrices (A, B) can be achieved via explicit constraint/penalty, or in certain settings the incoherence is implicitly imposed via an iterative algorithm such as gradient descent on the factorized space.

5 Discussion

In this paper, we establish a connection between the full-dimensional approach and the factorized approach for solving nonconvex low-rank optimization problems. Our main result shows that any SOSP of the factorized problem must also be a fixed 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. We also illustrate applications of our framework to certain low-rank estimation problems arising in matrix signal recovery, such as matrix sensing, matrix completion, and robust PCA. Overall, our result provides a new perspective on understanding the optimization landscape of the factorized approaches.

While the present work only considers exact fixed 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 fixed 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. was 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

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

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.

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.

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

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.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2): 572–596, 2011.

Ji Chen and Xiaodong Li. Memory-efficient kernel PCA via partial matrix sampling and nonconvex optimization: a model-free analysis of local minima. *arXiv preprint arXiv:1711.01742*, 2017.

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.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *arXiv preprint arXiv:1902.07698*, 2019.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online 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 *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 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.

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

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

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

Bamdev Mishra, Gilles Meyer, Francis Bach, and Rodolphe Sepulchre. Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4):2124–2149, 2013.

Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May):1665–1697, 2012.

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

Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Sharp time–data tradeoffs 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.

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.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

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.

Bart Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18(166):1–43, 2018.

Richard Zhang, Cédric Josz, Somayeh Sojoudi, and Javad Lavaei. How much restricted isometry is needed in nonconvex matrix recovery? In *Advances in neural information processing systems*, pages 5586–5597, 2018.

Richard Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *arXiv preprint arXiv:1901.01631*, 2019.

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.

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B. Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.

A Additional proofs

Proof of Theorem 5. Without loss of generality, take $m \leq n$. Define the matrices

$$X_0 = \sum_{i=1}^{r} \mathbf{e}_i \mathbf{e}_i^\top \text{ and } X_1 = \sum_{i=r+1}^{r+r'} \mathbf{e}_i \mathbf{e}_i^\top,$$

and

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

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

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

which clearly is α-strongly convex and β-smooth (and therefore trivially satisfies α-RSC and β-RSM). Define

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

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

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

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 $\mathsf{g}(A, B)$. We can trivially see that $\nabla \mathsf{f}(X_0) = \beta X_1$, and so $\nabla \mathsf{f}(X_0)^\top A_0 = 0$ and $\nabla \mathsf{f}(X_0) B_0 = 0$, 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 $\nabla^2 \mathsf{g}(A_0, B_0)$ maps $(A_1, B_1) \times (A_1, B_1)$ to a nonnegative value. Using our earlier calculation (12) to derive $\nabla^2 \mathsf{g}(A, B)$, we can calculate

$$\begin{aligned}
&\nabla^2 \mathsf{g}(A_0, B_0)\Big((A_1, B_1), (A_1, B_1)\Big) \\
&= 2\langle \nabla \mathsf{f}(X_0), A_1 B_1^\top \rangle + \nabla^2 \mathsf{f}(X_0)\Big(A_0 B_1^\top + A_1 B_0^\top, A_0 B_1^\top + A_1 B_0^\top\Big) \\
&= 2\beta \cdot \langle X_1, A_1 B_1^\top \rangle + \alpha \cdot \|A_0 B_1^\top + A_1 B_0^\top\|_{\mathrm{F}}^2 + (\beta - \alpha) \cdot \|M \circ (A_0 B_1^\top + A_1 B_0^\top)\|_{\mathrm{F}}^2, \quad (25)
\end{aligned}$$

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

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

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 (25) can be rewritten as

$$2\beta \cdot \mathrm{trace}\left(A_1'' B_1''^\top\right) + \alpha \cdot \left\| \begin{pmatrix} A_1' + B_1'^\top & B_1''^\top \\ A_1'' & \mathbf{0}_{(m-r)\times(n-r)} \end{pmatrix} \right\|_{\mathrm{F}}^2 + (\beta - \alpha) \cdot \left\| \begin{pmatrix} \mathbf{0}_{r\times r} & B_1''^\top \\ A_1'' & \mathbf{0}_{(m-r)\times(n-r)} \end{pmatrix} \right\|_{\mathrm{F}}^2.$$

This is trivially lower-bounded by

$$2\beta \cdot \mathrm{trace}\left(A_1'' B_1''^\top\right) + \beta \cdot \|A_1''\|_{\mathrm{F}}^2 + \beta \cdot \|B_1''\|_{\mathrm{F}}^2.$$

Using the fact that $\left|\mathrm{trace}(YZ)\right| \le \|Y\|_{\mathrm{F}} \|Z\|_{\mathrm{F}}$ for all matrices Y, Z, this expression is clearly nonnegative. We have therefore proved that $\nabla^2 \mathsf{g}(A_0, B_0) \succeq 0$, thus verifying that (A_0, B_0) is a SOSP and proving the desired result. □

Proof of Lemma 1. First, we must have $\nabla \mathsf{f}(X)^\top U_X = 0$ and $\nabla \mathsf{f}(X) V_X = 0$ since X is a critical point of the rank-constrained minimization problem. Next, let $X = U_X \cdot \mathrm{diag}\{\sigma_1, \ldots, \sigma_r\} \cdot V_X^\top$ be a SVD of X and $u_\star \in \mathbb{R}^m$ and $v_\star \in \mathbb{R}^n$ be the top singular vectors of the gradient $\nabla \mathsf{f}(X)$. For $t \in [0, 1]$, define

$$X_t = \sum_{i=1}^{r-1} \sigma_i u_{X,i} v_{X,i}^\top + \sigma_r \left[\sqrt{1-t^2} \cdot u_{X,r} + t u_\star\right] \left[\sqrt{1-t^2} \cdot v_{X,r} - t v_\star\right]^\top.$$

Since $\nabla \mathsf{f}(X)^\top U_X = 0$ and $\nabla \mathsf{f}(X) V_X = 0$, some calculations yield

$$\langle \nabla \mathsf{f}(X), X_t - X \rangle = \langle \nabla \mathsf{f}(X), -t^2 \sigma_r u_\star v_\star \rangle = -t^2 \sigma_r \|\nabla \mathsf{f}(X)\| = -\frac{1}{2\sigma_r} \|\nabla \mathsf{f}(X)\| \|X_t - X\|_{\mathrm{F}}^2.$$

Now, if $\|\nabla \mathsf{f}(X)\| > \beta_{\mathsf{loc}} \cdot \sigma_r$, then we can find some small $\delta > 0$ such that

$$\langle \nabla \mathsf{f}(X), X_t - X \rangle < -\frac{\beta_{\mathsf{loc}} + \delta}{2} \|X_t - X\|_{\mathrm{F}}^2$$

for all $t \in (0, 1]$ (note that this step uses the fact that $\|X_t - X\|_{\mathrm{F}} > 0$ for all $t \neq 0$).

On the other hand, by definition of β_{loc} (4), for sufficiently small $t_0 > 0$ we have

$$\mathsf{f}(X_t) \le \mathsf{f}(X) + \langle \nabla \mathsf{f}(X), X_t - X \rangle + \frac{\beta_{\mathsf{loc}}(X) + \delta}{2} \|X_t - X\|_{\mathrm{F}}^2$$

for all $0 \le t \le t_0$. Combining these calculations, for all $t \in (0, t_0]$ we have

$$\mathsf{f}(X_t) < \mathsf{f}(X),$$

which contradicts the assumption that X is a local minimum. □

Proof of Lemma 3. In order for the proof of Theorem 1 to hold for this new setting, we need to verify that the equations (13) and (12) both hold.

By Zhu et al. [2018, Theorem 3, (16)–(18), and Remark 8], for any pair (A, B) for which $A^\top A = B^\top B$, the first derivative satisfies

$$\nabla \mathsf{g}_{\mathsf{reg}}(A, B) = \nabla \mathsf{g}(A, B) \tag{26}$$

while the second derivative $\nabla^2 \mathsf{g}_{\mathsf{reg}}(A, B)$ maps $(A_1, B_1) \times (A_1, B_1)$ to

$$\nabla^2 \mathsf{g}(A, B)\Big((A_1, B_1), (A_1, B_1)\Big) + 4\lambda \|A^\top A_1 + A_1^\top A - B^\top B_1 - B_1^\top B\|_{\mathrm{F}}^2, \tag{27}$$

and furthermore $A^\top A = B^\top B$ holds for any critical point (A, B) of $\mathsf{g}_{\mathsf{reg}}$. Comparing to the proof of Theorem 1, the first-derivative property therefore verify that (13) holds, while the second-derivative property verifies that (12) holds for any (A_1, B_1) with $A^\top A_1 = 0$ and $B^\top B_1 = 0$. Now, following the proof of Theorem 1, for both the full-rank case and the rank-deficient case, we set $(A_1, B_1) = (-u_\star z^\top, v_\star z'^\top)$ for some vectors z, z', where $u_\star, v_\star$ are the top singular vectors of $\nabla \mathsf{f}(AB^\top)$ and, therefore, satisfy $u_\star \perp A$ and $v_\star \perp B$ by (13). This means that we indeed have $A^\top A_1 = 0$ and $B^\top B_1 = 0$, and so (12) holds for the relevant choice of (A_1, B_1). This is sufficient for the proof of Theorem 1(a) to yield the desired result for the regularized setting. To verify that Theorem 1(b) holds in this setting, consider any X that is a local minimum of $\min_{\mathrm{rank}(X) \le r} \mathsf{f}(X)$. Define $A = U_X \cdot \mathrm{diag}\{\sigma_1, \dots, \sigma_r\}^{1/2}$ and $B = V_X \cdot \mathrm{diag}\{\sigma_1, \dots, \sigma_r\}^{1/2}$. Then clearly $A^\top A = B^\top B$. Since Theorem 1(a) implies that (A, B) is a SOSP of g, we know that $\nabla \mathsf{g}(A, B) = 0$ and $\nabla^2 \mathsf{g}(A, B) \succeq 0$. Combined with (26) and (27), this proves that (A, B) is a SOSP of $\mathsf{g}_{\mathsf{reg}}$. □

Proof of Lemma 4. Define $Y_t = (A + tA_1)(B + tB_1)^\top$ for $t > 0$. Note that $\|X - Y_t\|_{\mathrm{F}} \to 0$ as $t \to 0$. By definition of $\beta_{\mathsf{local}}(X)$,

$$\limsup_{t \to 0} \frac{\mathsf{f}(Y_t) - \mathsf{f}(X) - \langle \nabla \mathsf{f}(X), Y_t - X \rangle}{\frac{1}{2}\|X - Y_t\|_{\mathrm{F}}^2} \le \beta_{\mathsf{local}}(X).$$

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

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

Combining these two, we see that

$$\limsup_{t \to 0} \frac{\nabla^2 \mathsf{f}(X)\big(Y_t - X, Y_t - X\big)}{\|X - Y_t\|_{\mathrm{F}}^2} \le \beta_{\mathsf{local}}(X).$$

Now we calculate this fraction. Since $Y_t - X = t \cdot (AB_1^\top + A_1 B^\top) + t^2 \cdot A_1 B_1^\top$, we have

$$\|X - Y_t\|_{\mathrm{F}}^2 = t^2 \|AB_1^\top + A_1 B^\top\|_{\mathrm{F}}^2 + \mathcal{O}(t^3)$$

and

$$\nabla^2 \mathsf{f}(X)\big(Y_t - X, Y_t - X\big) = t^2 \nabla^2 \mathsf{f}(X)\big(AB_1^\top + A_1 B^\top, AB_1^\top + A_1 B^\top\big) + \mathcal{O}(t^3),$$

and therefore,

$$\limsup_{t \to 0} \frac{\nabla^2 \mathsf{f}(X)\big(Y_t - X, Y_t - X\big)}{\|X - Y_t\|_{\mathrm{F}}^2} = \frac{\nabla^2 \mathsf{f}(X)\big(AB_1^\top + A_1 B^\top, AB_1^\top + A_1 B^\top\big)}{\|AB_1^\top + A_1 B^\top\|_{\mathrm{F}}^2},$$

(as long as we are not in the degenerate case that $\|AB_1^\top + A_1 B^\top\|_{\mathrm{F}} = 0$—but if this were the case, then the result would hold trivially). Combining everything, we have proved the desired bound. □

Proof of Lemma 5. By assumption, X_0 is a fixed point of PGD for some step size $\eta_0 > 0$, meaning that

$$X_0 = \mathcal{P}_r\big(X_0 - \eta_0 \nabla \mathsf{f}(X_0)\big).$$

For the case $\mathrm{rank}(X_0) = r$, then X_0 is a solution to the quadratic problem with rank constraint (by definition of projection), i.e.

$$X_0 = \operatorname*{arg\,min}_{\mathrm{rank}(X) \le r} \|X_0 - \eta_0 \nabla \mathsf{f}(X_0) - X\|_{\mathrm{F}}^2,$$

Then, in the case that $\mathrm{rank}(X_0) = r$, [Barber and Ha, 2018, Lemma 7] proves a first-order optimality condition for rank-constrained optimization:

$$\langle X_1 - X_0, \nabla \mathsf{f}(X_0) \rangle \ge -\frac{1}{2\sigma_r(X_0)} \|\nabla \mathsf{f}(X_0)\| \|X_0 - X_1\|_{\mathrm{F}}^2.$$

Combined with the α-RSC assumption over $\mathcal{X}$, we see that

$$\begin{aligned} \mathsf{f}(X_1) &\ge \mathsf{f}(X_0) + \langle X_1 - X_0, \nabla \mathsf{f}(X_0) \rangle + \frac{\alpha}{2} \|X_0 - X_1\|_{\mathrm{F}}^2 \\ &\ge \mathsf{f}(X_0) + \frac{1}{2} \left(\alpha - \frac{\|\nabla \mathsf{f}(X_0)\|}{\sigma_r(X_0)} \right) \|X_0 - X_1\|_{\mathrm{F}}^2. \end{aligned} \tag{28}$$

If instead $\mathrm{rank}(X_0) < r$ then $\nabla \mathsf{f}(X_0) = 0$ by the conditions of a fixed point (7), and so

$$\mathsf{f}(X_1) \ge \mathsf{f}(X_0) + \langle X_1 - X_0, \nabla \mathsf{f}(X_0) \rangle + \frac{\alpha}{2} \|X_0 - X_1\|_{\mathrm{F}}^2 = \mathsf{f}(X_0) + \frac{1}{2} \alpha \|X_0 - X_1\|_{\mathrm{F}}^2. \tag{29}$$

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

$$\mathsf{f}(X_0) \ge \mathsf{f}(X_1) + \frac{1}{2} \left(\alpha - \frac{\|\nabla \mathsf{f}(X_1)\|}{\sigma_r(X_1)} \right) \|X_0 - X_1\|_{\mathrm{F}}^2 \tag{30}$$

if $\mathrm{rank}(X_1) = r$, or

$$\mathsf{f}(X_0) \ge \mathsf{f}(X_1) + \frac{1}{2} \alpha \|X_0 - X_1\|_{\mathrm{F}}^2 \tag{31}$$

if $\operatorname{rank}(X_1) < r$.

Now suppose $\operatorname{rank}(X_0) = \operatorname{rank}(X_1) = r$. Adding the two inequalities (28) and (30) yields

$$0 \geq \frac{1}{2}\left(2\alpha - \frac{\|\nabla \mathsf{f}(X_0)\|}{\sigma_r(X_0)} - \frac{\|\nabla \mathsf{f}(X_1)\|}{\sigma_r(X_1)}\right)\|X_0 - X_1\|_{\mathrm{F}}^2.$$

This implies that either $X_0 = X_1$, or

$$\frac{\|\nabla \mathsf{f}(X_0)\|}{\sigma_r(X_0)} + \frac{\|\nabla \mathsf{f}(X_1)\|}{\sigma_r(X_1)} \geq 2\alpha,$$

as desired. If instead $\operatorname{rank}(X_0) = r$ and $\operatorname{rank}(X_1) < r$, then adding (28) and (31) yields

$$0 \geq \frac{1}{2}\left(2\alpha - \frac{\|\nabla \mathsf{f}(X_0)\|}{\sigma_r(X_0)}\right)\|X_0 - X_1\|_{\mathrm{F}}^2$$

and therefore since $X_0 \neq X_1$ we have

$$\frac{\|\nabla \mathsf{f}(X_0)\|}{\sigma_r(X_0)} \geq 2\alpha.$$

If instead $\operatorname{rank}(X_0) < r$ and $\operatorname{rank}(X_1) = r$, this case is symmetric to the one above. Finally if $\operatorname{rank}(X_0) < r$ and $\operatorname{rank}(X_1) < r$, then adding (29) and (31) proves that we must have $X_0 = X_1$ since $\alpha > 0$. □

Proof of Lemma 6. This lemma is an easy consequence of Lemma 1. First, since $\widehat{X}$ is a global minimum, it is also a local minimum and so $\widehat{X}$ satisfies the conditions (7) with $\eta = 1/\beta_{\mathsf{local}}(\widehat{X})$. Since the set $\mathcal{X}$ is open relative to the set of low-rank matrices, it contains an intersection of a neighborhood of $\widehat{X}$ and the set of low-rank matrices. Then comparing the definition of β-RSM with that of $\beta_{\mathsf{local}}(\widehat{X})$ (4), it follows that $\beta_{\mathsf{local}}(\widehat{X}) \leq \beta$, and in particular, $\widehat{X}$ also satisfies the conditions (7) with $\eta = 1/\beta$. This proves that $\widehat{X}$ is a fixed point of PGD at step size $\eta \leq 1/\beta$. □

Proof of Lemma 8. First we prove that the joint objective function $\mathsf{f}_{\mathsf{joint}}(X, S)$, defined in equation (22), satisfies joint α-RSC/β-RSM, that is

$$\frac{1}{2}\|D_\star - (X + S)\|_{\mathrm{F}}^2 \geq \frac{\alpha}{2}\|X - X_\star\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|S - S_\star\|_{\mathrm{F}}^2,$$

and similarly for β-RSM, over the set

$$\{(X, S) \in \mathbb{R}^{m\times n}_{\operatorname{rank}(r)} \times \mathcal{S} : \; X = AB^\top \text{ where } (A, B) \text{ satisfies (23)}\}. \tag{32}$$

Given the data matrix $D_\star = X_\star + S_\star$, we have

$$\frac{1}{2}\|D_\star - (X + S)\|_{\mathrm{F}}^2 = \frac{1}{2}\|X - X_\star\|_{\mathrm{F}}^2 + \frac{1}{2}\|S - S_\star\|_{\mathrm{F}}^2 + 2\langle X - X_\star, S - S_\star\rangle. \tag{33}$$

To bound the term $\langle X - X_\star, S - S_\star\rangle$, we closely follow Chen and Wainwright [2015, Corollary 6] and extend their result to the asymmetric and global case (their work assumes

$X = AA^\top$ and verifies RSC locally on the factorized space). Note that since $X = AB^\top$ for some (A, B) satisfying $A^\top A = B^\top B$ (23), it follows that $A = U\sqrt{\Sigma}R$ and $B = V\sqrt{\Sigma}R$ for some R, where $X = U\Sigma V^\top$ is a SVD of X and R is a rotation matrix, i.e. $R^\top R = \mathbf{I}_r$. Without loss of generality, we assume R is chosen to be the best transformation to $(A_\star, B_\star)$, i.e.

$$R \in \mathop{\arg\min}_{\widetilde{R}\in\mathbb{R}^{r\times r}} \{\|(\widetilde{A}^\top, \widetilde{B}^\top)^\top \widetilde{R} - (A_\star^\top, B_\star^\top)^\top\|_{\mathrm{F}} : \widetilde{A} = U\sqrt{\Sigma}, \widetilde{B} = V\sqrt{\Sigma}, \widetilde{R}^\top\widetilde{R} = \mathbf{I}_r\}.$$

Writing $X - X_\star = A(B - B_\star)^\top + (A - A_\star)B_\star^\top$, then:

$$\begin{aligned}
&|\langle X - X_\star, S - S_\star\rangle| = |\langle B^\top - B_\star^\top, A^\top(S - S_\star)\rangle| + |\langle A^\top - A_\star^\top, B_\star^\top(S - S_\star)\rangle| \\
&\le \|B - B_\star\|_{\mathrm{F}}\|A^\top(S - S_\star)\|_{\mathrm{F}} + \|A - A_\star\|_{\mathrm{F}}\|B_\star^\top(S - S_\star)\|_{\mathrm{F}} \\
&\le \|B - B_\star\|_{\mathrm{F}}\sqrt{\sum_{j=1}^n \|A^\top(S - S_\star)e_j\|_2^2} + \|A - A_\star\|_{\mathrm{F}}\sqrt{\sum_{j=1}^n \|B_\star^\top(S - S_\star)e_j\|_2^2} \\
&\le \|B - B_\star\|_{\mathrm{F}}\sqrt{\sum_{j=1}^n \|A\|_{2,\infty}^2\|(S - S_\star)e_j\|_1^2} + \|A - A_\star\|_{\mathrm{F}}\sqrt{\sum_{j=1}^n \|B_\star\|_{2,\infty}^2\|(S - S_\star)e_j\|_1^2},
\end{aligned}$$

where e_j denotes the j-th standard basis vector. Since X and $X_\star$ are both μ-incoherent (23), we further have $\|A\|_{2,\infty} \le c_2\sqrt{\frac{\sigma_1(X_\star)\mu r}{m}}$ and $\|B_\star\|_{2,\infty} \le c_2\sqrt{\frac{\sigma_1(X_\star)\mu r}{n}}$, while for each column of S, we have $\|Se_j\|_0 \le \|S_\star e_j\|_0 = s$ and thus $\|(S - S_\star)e_j\|_1 \le \sqrt{2s}\|(S - S_\star)e_j\|_2$. Putting these bounds together, we have

$$|\langle X - X_\star, S - S_\star\rangle| \le c_2\sqrt{\frac{2\sigma_1(X_\star)\mu r s}{\min\{m, n\}}}\|S - S_\star\|_{\mathrm{F}}(\|A - A_\star\|_{\mathrm{F}} + \|B - B_\star\|_{\mathrm{F}}).$$

From Tu et al. [2015, Lemma 5.4] and Zheng and Lafferty [2016, Lemma 4], we know that $\sqrt{\sigma_r(X_\star)}(\|A - A_\star\|_{\mathrm{F}} + \|B - B_\star\|_{\mathrm{F}}) \le 2\|X - X_\star\|_{\mathrm{F}}$. Plugging into the inequality above,

$$\begin{aligned}
|\langle X - X_\star, S - S_\star\rangle| &\le 2c_2\sqrt{\frac{2\kappa(X_\star)\mu r s}{\min\{m, n\}}}\|S - S_\star\|_{\mathrm{F}}\|X - X_\star\|_{\mathrm{F}} \\
&\le \frac{c_1}{2}\|X - X_\star\|_{\mathrm{F}}^2 + \frac{c_1}{2}\|S - S_\star\|_{\mathrm{F}}^2,
\end{aligned}$$

where the second step uses the assumption $2c_2\sqrt{\frac{2\kappa(X_\star)\mu r s}{\min\{m,n\}}} \le c_1$, together with the identity $ab \le \frac{a^2+b^2}{2}$. Combining with (33), we have proved the joint restricted strong convexity and restricted smoothness conditions over the set (32), with $\alpha = 1 - c_1, \beta = 1 + c_1$.

Next we give a brief outline on extending the result of first part of Theorem 2 to ensure the uniqueness of the fixed point $(X, S(X))$ of PGD on the joint problem (22). Note that, by definition of the fixed point, we have (with step sizes $\eta = 1$)

$$\begin{cases}
X = \mathcal{P}_r\big(X - \nabla_X \mathsf{f}_{\mathsf{joint}}(X, S(X))\big), \\
S(X) = \mathcal{P}_{\mathcal{S}}\big(S(X) - \nabla_S \mathsf{f}_{\mathsf{joint}}(X, S(X))\big).
\end{cases}$$

Since $\|S_{\star j}\|_0 \leq s$ for each column of $S_\star$, by definition of projection operator this implies that

$$\|S(X)_j - (S(X)_j - \nabla_{S_j} \mathsf{f}_{\mathsf{joint}}(X, S(X)))\|_{\mathrm{F}}^2 \leq \|S_{\star j} - (S(X)_j - \nabla_{S_j} \mathsf{f}_{\mathsf{joint}}(X, S(X)))\|_{\mathrm{F}}^2.$$

Rearranging terms, and combining across columns $j = 1, \ldots, n$, we obtain

$$\langle S_\star - S(X), \nabla_S \mathsf{f}_{\mathsf{joint}}(X, S(X)) \rangle \geq -\frac{1}{2}\|S_\star - S(X)\|_{\mathrm{F}}^2.$$

Turning to X, in the case that $\mathrm{rank}(X) = r$, by Barber and Ha [2018, Lemma 7], we know that

$$\langle X_\star - X, \nabla_X \mathsf{f}_{\mathsf{joint}}(X, S(X)) \rangle \geq -\frac{1}{2\sigma_r(X)}\|\nabla_X \mathsf{f}_{\mathsf{joint}}(X, S(X))\|\|X_\star - X\|_{\mathrm{F}}^2.$$

Instead, if $\mathrm{rank}(X) < r$, by condition (7), $\nabla_X \mathsf{f}_{\mathsf{joint}}(X, S(X)) = 0$. Following the same arguement as in the proof of Lemma 5, then (with $\alpha = 1 - c_1$ and $\nabla \mathsf{f}_{\mathsf{joint}}(X_\star, S_\star) = 0$),

$$0 \geq \frac{1}{2}\left(2(1-c_1) - \frac{\|\nabla_X \mathsf{f}_{\mathsf{joint}}(X, S(X))\|}{\sigma_r(X)}\right)\|X - X_\star\|_{\mathrm{F}}^2 + \left(\frac{1-2c_1}{2}\right)\|S(X) - S_\star\|_{\mathrm{F}}^2, \quad (34)$$

if $\mathrm{rank}(X) = r$, or

$$0 \geq (1-c_1)\|X - X_\star\|_{\mathrm{F}}^2 + \left(\frac{1-2c_1}{2}\right)\|S(X) - S_\star\|_{\mathrm{F}}^2, \quad (35)$$

if $\mathrm{rank}(X) < r$.

Now suppose that $\mathrm{rank}(X) = r$. For $c_1 > 0$ sufficiently small, the second term on the right-hand side of (34) is non-negative. This implies that either $X = X_\star$ and $S(X) = S_\star$, or

$$2(1-c_1) \leq \frac{\|\nabla_X \mathsf{f}_{\mathsf{joint}}(X, S(X))\|}{\sigma_r(X)} \leq \frac{1}{\eta} = 1,$$

where the last step is by condition (7)— but this cannot hold for $c_1 > 0$ sufficiently small. If instead $\mathrm{rank}(X) < r$, then since c_1 is small the inequality (35) yields $X = X_\star$ and $S(X) = S_\star$ which is a contradiction since $X_\star$ is full-rank. Therefore it follows that $\mathrm{rank}(X) = r$ and $X = X_\star$ and $S(X) = S_\star$, proving the lemma. □

Proof of Lemma 9. Let (A, B) be a SOSP of g with $X = AB^\top \in \mathcal{N}(X_\star)$ and let $X_\star = A_\star B_\star^\top$ with $A_\star = U_\star\sqrt{\Sigma_\star}$ and $B_\star = V_\star\sqrt{\Sigma_\star}$ where $X_\star = U_\star\Sigma_\star V_\star^\top$ is a SVD of $X_\star$. For (A, B), let $R \in \mathbb{R}^{r\times r}$ be the best orthogonal rotation matrix to $(A_\star, B_\star)$, i.e. R is the solution to $\min_{R^\top R = \mathbf{I}_r} \|(A^\top, B^\top)^\top R - (A_\star^\top, B_\star^\top)^\top\|_{\mathrm{F}}$.

Let $\mathcal{X} = \{X, X_\star\}$. Now to apply our Theorem 3 to this setting, we need to verify that f satisfies α-RSC over $\mathcal{X}$, and that $\|\nabla \mathsf{f}(X)\| < 2\alpha \cdot \sigma_r(X)$ (note $\nabla \mathsf{f}(X_\star) = 0$ in our case; X cannot be rank-deficient since $X_\star$ is full-rank and $\|X - X_\star\|_{\mathrm{F}} \leq 0.1\kappa^{-1}(X_\star)\sigma_r(X_\star)$, see equation (37) below). First, writing $\Delta_A = AR - A_\star$ and $\Delta_B = BR - B_\star$, we have the decomposition $X - X_\star = A_\star\Delta_B^\top + \Delta_A B_\star^\top + \Delta_A\Delta_B^\top$. Then, by the work of Ge et al. [2017,

Proof of Lemma 21], if $\|\Delta_A\|_{\mathrm{F}}^2 + \|\Delta_B\|_{\mathrm{F}}^2 \le \sigma_r(X_\star)/40$, then by our choice of p, we have with high probability

$$\begin{aligned}\frac{1}{2p}\|\mathcal{P}_\Omega\left(X - X_\star\right)\|_{\mathrm{F}}^2 &= \frac{1}{2p}\|\mathcal{P}_\Omega\left(A_\star\Delta_B^\top + \Delta_A B_\star^\top + \Delta_A\Delta_B^\top\right)\|_{\mathrm{F}}^2 \\ &\ge \frac{1}{4}\sigma_r(X_\star)(\|\Delta_A\|_{\mathrm{F}}^2 + \|\Delta_B\|_{\mathrm{F}}^2).\end{aligned}$$

Similarly, we can prove that with high probability

$$\frac{1}{2p}\|\mathcal{P}_\Omega\left(X - X_\star\right)\|_{\mathrm{F}}^2 \le \frac{3}{2}\sigma_1(X_\star)(\|\Delta_A\|_{\mathrm{F}}^2 + \|\Delta_B\|_{\mathrm{F}}^2).$$

Furthermore, by Tu et al. [2015, Lemma 5.3], we can deduce that $0.35\sigma_1^{-1}(X_\star)\|X - X_\star\|_{\mathrm{F}}^2 \le \|\Delta_A\|_{\mathrm{F}}^2 + \|\Delta_B\|_{\mathrm{F}}^2$ while by Tu et al. [2015, Lemma 5.4] and Zheng and Lafferty [2016, Lemma 4], together with (23) that $A^\top A = B^\top B$, we have $\|\Delta_A\|_{\mathrm{F}}^2 + \|\Delta_B\|_{\mathrm{F}}^2 \le 2.5\sigma_r^{-1}(X_\star)\|X - X_\star\|_{\mathrm{F}}^2$. Putting everything together, it follows that

$$0.08\kappa^{-1}(X_\star)\|X - X_\star\|_{\mathrm{F}}^2 \le \frac{1}{2p}\|\mathcal{P}_\Omega\left(X - X_\star\right)\|_{\mathrm{F}}^2 \le 3.75\kappa(X_\star)\|X - X_\star\|_{\mathrm{F}}^2,$$

whenever $\|X - X_\star\|_{\mathrm{F}}^2 \le 0.01\sigma_r^2(X_\star)$. In particular this proves restricted strong convexity over $\mathcal{X} = \{X, X_\star\}$, with $\alpha = 0.08\kappa^{-1}(X_\star)$.

Next, to show $\|\nabla \mathsf{f}(X)\| < 2\alpha \cdot \sigma_r(X)$, it suffices to show

$$\frac{1}{p}\|\mathcal{P}_\Omega\left(X - X_\star\right)\| < \frac{0.16\sigma_r(X)}{\kappa(X_\star)}. \tag{36}$$

We denote $\widetilde{\mathcal{P}}_\Omega\left(L\right) = \frac{1}{p}\mathcal{P}_\Omega\left(L\right) - L$. Then we split the left hand side of equation (36) into two terms

$$\frac{1}{p}\|\mathcal{P}_\Omega\left(X - X_\star\right)\| \le \|X - X_\star\| + \|\widetilde{\mathcal{P}}_\Omega\left(X - X_\star\right)\|.$$

The first term is bounded by our assumption as $\|X - X_\star\| \le 0.1\kappa^{-1}(X_\star)\sigma_r(X_\star)$. The second term is upper bounded by

$$\begin{aligned}\|\widetilde{\mathcal{P}}_\Omega\left(X - X_\star\right)\| &\le \|\widetilde{\mathcal{P}}_\Omega\left(\Delta_A(BR)^\top\right)\| + \|\widetilde{\mathcal{P}}_\Omega\left(A_\star\Delta_B^\top\right)\| \\ &\le \|\widetilde{\mathcal{P}}_\Omega\left(11^\top\right)\|\|\Delta_A\|_{2,\infty}\|BR\|_{2,\infty} + \|\widetilde{\mathcal{P}}_\Omega\left(11^\top\right)\|\|A_\star\|_{2,\infty}\|\Delta_B\|_{2,\infty} \\ &\lesssim \frac{\sqrt{m+n}}{\sqrt{p}}\|\Delta_A\|_{2,\infty}\|BR\|_{2,\infty} + \frac{\sqrt{m+n}}{\sqrt{p}}\|A_\star\|_{2,\infty}\|\Delta_B\|_{2,\infty} \\ &\le c_3\frac{\sqrt{m+n}}{\sqrt{p}}\frac{\sigma_1(X_\star)\mu r}{\sqrt{mn}} \le 0.04\kappa^{-1}(X_\star)\sigma_r(X_\star),\end{aligned}$$

with high probability. Here the first step uses the identity $X - X_\star = \Delta_A(BR)^\top + A_\star\Delta_B^\top$ together with triangle inequality; the second and the third steps are respectively due to Chen and Li [2017, Lemma 4.5] and Keshavan et al. [2010, Lemma 3.2]; the fourth step

uses the fact that $\|BR\|_{2,\infty} = \|B\|_{2,\infty}$ for orthogonal matrix R, as well as the incoherence assumption (23); and the last step holds since $pmn \geq \mathcal{O}\left(\mu^2 r^2 \kappa^4(X_\star)(m+n)\log(m+n)\right)$. The right hand side of (36) is lower bounded as follows:

$$\sigma_r(X) \geq \sigma_r(X_\star) - \|X - X_\star\| \geq 0.9\sigma_r(X_\star). \tag{37}$$

Combining the above bounds, it is straightforward to see that (36) holds. Finally, applying Theorem 3 proves the desired result.

□

B Nondegenerate case for robust PCA

Consider the robust PCA problem given in Section 4.2, and we additionally assume that the sparse component $S_\star$ has bounded entries, i.e., $\|S_\star\|_\infty \leq c\frac{\mu r \sigma_1(X_\star)}{\sqrt{mn}}$ for some constant $c > 0$. As pointed out by Ge et al. [2017], this requirement is not without loss of generality, because any μ-incoherent matrix $X_\star$ has maximum entries bounded by $\frac{\mu r \sigma_1(X_\star)}{\sqrt{mn}}$. Here we consider the following sparsity constraint set,

$$\bar{\mathcal{S}} = \{S \in \mathbb{R}^{m\times n} : \|S_j\|_0 \leq \|S_{\star j}\|_0 = s \text{ and } \|S\|_\infty \leq \frac{c\mu r \sigma_1(X_\star)}{\sqrt{mn}}\}.$$

Then the following statement holds: if $m \geq \mathcal{O}(s \cdot \mu^2 r^2 \kappa^2)$,

$$\text{For any } S \in \bar{\mathcal{S}},\ \sigma_{r+1}(D_\star - S) < \sigma_r(D_\star - S).$$

To see why, we first bound $\|S - S_\star\|$. Writing $\Delta_S = S - S_\star$, we can calculate

$$\begin{aligned}\|\Delta_S\| = \sup_{\|x\|_2=\|y\|_2=1} x^\top \Delta_S y = \sup_{\|x\|_2=\|y\|_2=1} \sum_{j=1}^n y_j \cdot \Delta_{S,j}^\top x \leq \sup_{\|x\|_2=\|y\|_2=1} \sum_{j=1}^n |y_j| \|\Delta_{S,j}\|_2 \|x\|_2 \\ \leq \sup_{\|y\|_2=1} \|y\|_2 \sqrt{\sum_{j=1}^n \|\Delta_{S,j}\|_2^2} \leq \sqrt{\frac{s}{m}} \cdot c_2 \mu r \sigma_1(X_\star),\end{aligned}$$

where $\Delta_{S,j}$ denotes the j-th column of Δ_S, and the last step holds since $\|\Delta_{S,j}\|_2 \leq \sqrt{\frac{s \cdot c_2^2 \mu^2 r^2 \sigma_1^2(X_\star)}{mn}}$ for some $c_2 > 0$. Then if $m \gtrsim s \cdot \mu^2 r^2 \kappa^2$, we can find $c_3 > 0$ such that $\|\Delta_S\| \leq c_3 \sigma_r(X_\star)$. Therefore, given the data matrix $D_\star = X_\star + S_\star$, we get

$$\sigma_r(D_\star - S) \geq \sigma_r(X_\star) - \|S - S_\star\| \geq (1 - c_3)\sigma_r(X_\star) \text{ and } \sigma_{r+1}(D_\star - S) \leq c_3 \sigma(X_\star).$$

Taking $c_3 < 1/2$ then proves the desired result.