Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach

Dohuyng Park Facebook
Anastasios Kyrillidis UT Austin
Constantine Caramanis UT Austin
Sujay Sanghavi UT Austin

Abstract

We consider the *non-square* matrix sensing problem, under restricted isometry property (RIP) assumptions. We focus on the non-convex formulation, where any rank-r matrix $X \in \mathbb{R}^{m \times n}$ is represented as $UV^\top$, where $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$. In this paper, we complement recent findings on the non-convex geometry of the analogous PSD setting [5], and show that matrix factorization does not introduce any spurious local minima, under RIP.

1 Introduction and Problem Formulation

Consider the following matrix sensing problem:

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}} \quad & f(X) := \|\mathcal{A}(X) - b\|_2^2 \\ \text{subject to} \quad & \text{rank}(X) \leq r. \end{aligned} \tag{1}$$

Here, $b \in \mathbb{R}^p$ denotes the set of observations and $\mathcal{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^p$ is the sensing linear map. The motivation behind this task comes from several applications, where we are interested in inferring an unknown matrix $X^\star \in \mathbb{R}^{m \times n}$ from b. Common assumptions are (i) $p \ll m \cdot n$, (ii) $b = \mathcal{A}(X^\star) + w$, *i.e.*, we have a linear measurement system, and (iii) $X^\star$ is rank-r, $r \ll \min\{m, n\}$. Such problems appear in a variety of research fields and include image processing [11, 44], data analytics [13, 11], quantum computing [1, 19, 26], systems [32], and sensor localization [23] problems.

Appearing in Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

There are numerous approaches that solve (1), both in its original non-convex form or through its convex relaxation; see [28, 16] and references therein. However, satisfying the rank constraint (or any nuclear norm constraints in the convex relaxation) per iteration requires SVD computations, which could be prohibitive in practice for large-scale settings. To overcome this obstacle, recent approaches reside on non-convex parametrization of the variable space and encode the low-rankness directly into the objective [25, 22, 2, 43, 49, 14, 4, 48, 42, 50, 24, 35, 46, 37, 36, 47, 34, 29, 33]. In particular, we know that a rank-r matrix $X \in \mathbb{R}^{m \times n}$ can be written as a product $UV^\top$, where $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$. Such a re-parametrization technique has a long history [45, 15, 39], and was popularized by Burer and Monteiro [8, 9] for solving semi-definite programs (SDPs). Using this observation in (1), we obtain the following *non-convex, bilinear* problem:

$$\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}} \quad f(UV^\top) := \|\mathcal{A}(UV^\top) - b\|_2^2. \tag{2}$$

Now, (2) has a different form of non-convexity due to the bilinearity of the variable space, which raises the question whether we introduce spurious local minima by doing this transformation.

Contributions: The goal of this paper is to answer negatively to this question: *We show that, under standard regulatory assumptions on $\mathcal{A}$, $UV^\top$ parametrization does not introduce any spurious local minima.* To do so, we non-trivially generalize recent developments for the square, PSD case [5] to the non-square case for $X^\star$. Our result requires a different (but equivalent) problem re-formulation and analysis, with the introduction of an appropriate regularizer in the objective.

Related work. There are several papers that consider similar questions, but for other objectives. [40] characterizes the non-convex geometry of the *complete* dictionary recovery problem, and proves that all local minima are global; [6] considers the problem of non-convex phase synchronization where the task is modeled as a non-convex least-squares optimization problem, and can be globally solved via a modified version of power method; [41] show that a non-convex fourth-order polynomial objective for phase retrieval has no local minimizers and all global minimizers are equivalent; [3, 7] show that the Burer-Monteiro approach works on smooth semidefinite programs, with applications in synchronization and community detection; [17] consider the PCA problem under streaming settings and use martingale arguments to prove that stochastic gradient descent on the factors reaches to the global solution with non-negligible probability; [20] introduces the notion of *strict saddle points* and shows that noisy stochastic gradient descent can escape saddle points for generic objectives f; [30] proves that gradient descent converges to (local) minimizers almost surely, using arguments drawn from dynamical systems theory.

More related to this paper are the works of [21] and [5]: they show that matrix completion and sensing have no spurious local minima, for the case where $X^\star$ is square and PSD. For both cases, extending these arguments for the more realistic non-square case is a non-trivial task.

1.1 Assumptions and Definitions

We first state the assumptions we make for the matrix sensing setting. We consider the case where the linear operator $\mathcal{A}$ satisfies the *Restricted Isometry Property*, according to the following definition [12]:

Definition 1.1 (Restricted Isometry Property (RIP))**.** *A linear operator $\mathcal{A}: \mathbb{R}^{m\times n} \to \mathbb{R}^p$ satisfies the restricted isometry property on rank-r matrices, with parameter δ_r, if the following set of inequalities hold for all rank-r matrices X:*

$$(1-\delta_r)\cdot \|X\|_F^2 \le \|\mathcal{A}(X)\|_2^2 \le (1+\delta_r)\cdot \|X\|_F^2.$$

Characteristic examples are Gaussian-based linear maps [18, 38], Pauli-based measurement operators, used in quantum state tomography applications [31], Fourier-based measurement operators, which lead to computational gains in practice due to their structure [27, 38], or even permuted and sub-sampled noiselet linear operators, used in image and video compressive sensing applications [44].

In this paper, we consider sensing mechanisms that can be expressed as:

$$(\mathcal{A}(X))_i = \langle A_i, X\rangle, \quad \forall i = 1,\dots,p, \text{ and } A_i \in \mathbb{R}^{m\times n}.$$

E.g., for the case of a Gaussian map $\mathcal{A}$, A_i are independent, identically distributed (i.i.d.) Gaussian matrices; for the case of a Pauli map $\mathcal{A}$, $A_i \in \mathbb{R}^{n\times n}$ are i.i.d. and drawn uniformly at random from a set of scaled Pauli "observables" $(P_1 \otimes P_2 \otimes \cdots \otimes P_d)/\sqrt{n}$, where $n = 2^d$ and P_i is a 2×2 Pauli observable matrix [31].

A useful property derived from the RIP definition is the following [10]:

Proposition 1.2 (Useful property due to RIP)**.** *For a linear operator $\mathcal{A}: \mathbb{R}^{m\times n} \to \mathbb{R}^p$ that satisfies the restricted isometry property on rank-r matrices, the following inequality holds for any two rank-r matrices $X,\ Y \in \mathbb{R}^{m\times n}$:*

$$\left|\sum_{i=1}^{p} \langle A_i, X\rangle \cdot \langle A_i, Y\rangle - \langle X, Y\rangle\right| \le \delta_{2r}\cdot \|X\|_F \cdot \|Y\|_F.$$

An important issue in optimizing f over the factored space is the existence of non-unique possible factorizations for a given X. Since we are interested in obtaining a low-rank solution in the original space, we need a notion of distance to the low-rank solution $X^\star$ over the factors. Among infinitely many possible decompositions of $X^\star$, we focus on the set of "equally-footed" factorizations [43]:

$$\mathcal{X}_r^\star = \Big\{ (U^\star, V^\star):\ U^\star {V^\star}^\top = X^\star, \\ \sigma_i(U^\star) = \sigma_i(V^\star) = \sigma_i(X^\star)^{1/2}, \forall i \in [r]\Big\}. \quad (3)$$

Given a pair (U, V), we define the distance to $X^\star$ as:

$$\textsc{Dist}\,(U, V; X^\star) = \min_{(U^\star,V^\star)\in\mathcal{X}_r^\star} \left\| \begin{bmatrix} U \\ V\end{bmatrix} - \begin{bmatrix} U^\star \\ V^\star\end{bmatrix}\right\|_F.$$

1.2 Problem Re-formulation

Before we delve into the main results, we need to further reformulate the objective (2) for our analysis. First, we use a well-known transformation to reduce (2) to a semidefinite optimization. Let us define auxiliary variables

$$W = \begin{bmatrix} U \\ V\end{bmatrix} \in \mathbb{R}^{(m+n)\times r}, \quad \tilde{W} = \begin{bmatrix} U \\ -V\end{bmatrix} \in \mathbb{R}^{(m+n)\times r}.$$

Based on the auxiliary variables, we define the linear map $\mathcal{B}: \mathbb{R}^{(m+n)\times(m+n)} \to \mathbb{R}^p$ such that $(\mathcal{B}(WW^\top))_i = \langle B_i, WW^\top\rangle$, and $B_i \in \mathbb{R}^{(m+n)\times(m+n)}$. To make a connection between the variable spaces (U, V) and W, $\mathcal{A}$ and $\mathcal{B}$ are related via matrices A_i and B_i as follows:

$$B_i = \frac{1}{2}\cdot \begin{bmatrix} 0 & A_i \\ A_i^\top & 0\end{bmatrix}.$$

This further implies that:

$$
\begin{aligned}
(\mathcal{B}(WW^\top))_i &= \frac{1}{2} \cdot \langle B_i, WW^\top \rangle \\
&= \frac{1}{2} \cdot \left\langle \begin{bmatrix} 0 & A_i \\ A_i^\top & 0 \end{bmatrix}, \begin{bmatrix} UU^\top & UV^\top \\ VU^\top & VV^\top \end{bmatrix} \right\rangle \\
&= \left\langle A_i, UV^\top \right\rangle .
\end{aligned}
$$

Given the above, we re-define $f : \mathbb{R}^{(m+n)\times r} \to \mathbb{R}$ such that

$$
f(W) := \|\mathcal{B}(WW^\top) - b\|_2^2. \tag{4}
$$

It is important to note that $\mathcal{B}$ operates on $(m+n) \times (m+n)$ matrices, while we assume RIP on $\mathcal{A}$ and $m \times n$ matrices. Making no other assumptions for $\mathcal{B}$, we cannot directly apply [5] on (4), but a rather different analysis is required.

In addition to this redefinition, we also introduce a *regularizer* $g : \mathbb{R}^{(m+n)\times r} \to \mathbb{R}$ such that

$$
g(W) := \lambda \left\| \tilde{W}^\top W \right\|_F^2 = \lambda \left\| U^\top U - V^\top V \right\|_F^2 .
$$

This regularizer was first introduced in [43] to prove convergence of its algorithm for non-square matrix sensing, and it is also used in this paper to analyze local minima of the problem. After setting $\lambda = \frac{1}{4}$, (2) can be *equivalently* written as:

$$
\underset{W \in \mathbb{R}^{(m+n)\times r}}{\text{minimize}} \quad \|\mathcal{B}(WW^\top) - b\|_2^2 + \frac{1}{4} \cdot \left\| \tilde{W}^\top W \right\|_F^2 . \tag{5}
$$

By equivalent, we note that the addition of g in the objective does not change the problem, since for any rank-r matrix X there is a pair of factors (U, V) such that $g(W) = 0$. It merely reduces the set of optimal points from all possible factorizations of $X^\star$ to *balanced* factorizations of $X^\star$ in $\mathcal{X}_r^\star$. $U^\star$ and $V^\star$ have the same set of singular values, which are the square roots of the singular values of $X^\star$. A key property of the balanced factorizations is the following.

Proposition 1.3. *For any factorization of the form* (3)*, it holds that*

$$
\tilde{W}^{\star\top} W^\star = U^{\star\top} U^\star - V^{\star\top} V^\star = 0
$$

Proof. By "balanced factorizations" of $X^\star = U^\star V^{\star\top}$, we mean that factors $U^\star$ and $V^\star$ satisfy

$$
U^\star = A\Sigma^{1/2} R, \quad V^\star = B\Sigma^{1/2} R \tag{6}
$$

where $X^\star = A\Sigma B^\top$ is the SVD, and $R \in \mathbb{R}^{r\times r}$ is an orthonormal matrix. Apply this to $\tilde{W}^{\star\top} W^\star$ to get the result. □

Therefore, we have $g(W^\star) = 0$, and $(U^\star, V^\star)$ is an optimal point of (5).

2 Main Results

This section describes our main results on the function landscape of the non-square matrix sensing problem. The following theorem bounds the distance of any local minima to the global minimum, by the function value at the global minimum.

Theorem 2.1. *Suppose $W^\star$ is any target matrix of the optimization problem* (5)*, under the balanced singular values assumption for $U^\star$ and $V^\star$. If W is a critical point satisfying the first- and the second-order optimality conditions, i.e., $\nabla(f+g)(W) = 0$ and $\nabla^2(f+g)(W) \succeq 0$, then we have*

$$
\begin{aligned}
\tfrac{1 - 5\delta_{2r} - 544\delta_{4r}^2 - 1088\delta_{2r}\delta_{4r}^2}{8(40+68\delta_{2r})(1+\delta_{2r})} &\left\| WW^\top - W^\star W^{\star\top} \right\|_F^2 \\
&\leq \left\| \mathcal{A}(U^\star V^{\star\top}) - b \right\|^2 . \quad (7)
\end{aligned}
$$

Observe that for this bound to make sense, the term $\frac{1 - 5\delta_{2r} - 544\delta_{4r}^2 - 1088\delta_{2r}\delta_{4r}^2}{8(40+68\delta_{2r})(1+\delta_{2r})}$ needs to be positive. We provide some intuition of this result next. Combined with Lemma 5.14 in [43], we can also obtain the distance between (U, V) and $(U^\star, V^\star)$.

Corollary 2.2. *For $W = \begin{bmatrix} U \\ V \end{bmatrix}$ and given the assumptions of Theorem 2.1, we have*

$$
\begin{aligned}
\sigma_r(X^\star) \cdot \tfrac{1 - 5\delta_{2r} - 544\delta_{4r}^2 - 1088\delta_{2r}\delta_{4r}^2}{10(40+68\delta_{2r})(1+\delta_{2r})} &\cdot \textsc{Dist}\left(U, V; X^\star\right)^2 \\
&\leq \left\| \mathcal{A}(U^\star V^{\star\top}) - b \right\|^2 . \quad (8)
\end{aligned}
$$

Implications of these results are described next, where we consider specific settings.

Remark 1 (Noiseless matrix sensing)**.** Suppose that $W^\star = \begin{bmatrix} U^\star \\ V^\star \end{bmatrix}$ is the underlying unknown true matrix, *i.e.*, $X^\star = U^\star V^{\star\top}$ is rank-r and $b = \mathcal{A}(U^\star V^{\star\top})$. We assume the noiseless setting, $w = 0$. If $0 \leq \delta_{2r} \leq \delta_{4r} \lesssim 0.0363$, then $\frac{1 - 5\delta_{2r} - 544\delta_{4r}^2 - 1088\delta_{2r}\delta_{4r}^2}{10(40+68\delta_{2r})(1+\delta_{2r})} > 0$ in Corollary 2.2. Since the RHS of (8) is zero, this further implies that $\textsc{Dist}\left(U, V; X^\star\right) = 0$, *i.e.*, any critical point W that satisfies first- and second-order optimality conditions is global minimum.

Remark 2 (Noisy matrix sensing)**.** Suppose that $W^\star$ is the underlying true matrix, such that $X^\star = U^\star V^{\star\top}$ and is rank-r, and $b = \mathcal{A}(U^\star V^{\star\top}) + w$, for some noise term w. If $0 \leq \delta_{2r} \leq \delta_{4r} < 0.02$, then it follows from (7) that for any local minima W the distance to $U^\star V^{\star\top}$ is bounded by

$$
\frac{1}{500} \left\| WW^\top - W^\star W^{\star\top} \right\|_F \leq \|w\| .
$$

Remark 3 (High-rank matrix sensing)**.** Suppose that $X^\star$ is of arbitrary rank and let $X_r^\star$ denote its best rank-r approximation. Let $b = \mathcal{A}(X^\star) + w$ where w is some noise and let $(U^\star, V^\star)$ be a balanced factorization of $X_r^\star$. If $0 \leq \delta_{2r} \leq \delta_{4r} < 0.005$, then it follows from (8) that for any local minima (U, V) the distance to $(U^\star, V^\star)$ is bounded by

$$\textsc{Dist}\left(U, V; X^\star\right) \leq \tfrac{1250}{3\sigma_r(X^\star)} \cdot \|\mathcal{A}(X^\star - X_r^\star) + w\| .$$

In plain words, the above remarks state that, given sensing mechanism $\mathcal{A}$ with small RIP constants, any critical point of the non-square matrix sensing objective—with low rank optimum and no noise—is a global minimum. As we describe in Section 3, due to this fact, gradient descent over the factors can converge, with high probability, to (or very close to) the global minimum.

3 What About Saddle Points?

Our discussion so far concentrates on whether $UV^\top$ parametrization introduces spurious local minima. Our main results show that any point (U, V) that satisfies both first- and second-order optimality conditions[1] should be (or lie close to) the global optimum. However, we have not discussed what happens with saddle points, *i.e.*, points (U, V) where the Hessian matrix contains both positive and negative eigenvalues.[2] This is important for practical reasons: first-order methods rely on gradient information and, thus, can easily get stuck to saddle points that may be far away from the global optimum.

[20] studied conditions that guarantee that stochastic gradient descent—randomly initialized—converges to a local minimum; *i.e.*, we can avoid getting stuck to non-degenerate saddle points. These conditions include $f + g$ being bounded and smooth, having Lipschitz Hessian, being locally strongly convex, and satisfying the strict saddle property, according to the following definition.

Definition 3.1. *[20] A twice differentiable function $f+g$ is strict saddle, if all its stationary points, that are not local minima, satisfy $\lambda_{\min}(\nabla^2(f+g)(\cdot)) < 0$.*

[30] relax some of these conditions and prove the following theorem (for standard gradient descent).

[1] Note here that the second-order optimality condition includes positive *semi*-definite second-order information; *i.e.*, Theorem 2.1 also handles saddle points due to the semi-definiteness of the Hessian at these points.

[2] Here, we do not consider the harder case where saddle points have Hessian with positive, negative and zero eigenvalues.

Theorem 3.2 ([30] - Informal)**.** *If the objective is twice differentiable and satisfies the strict saddle property, then gradient descent, randomly initialized and with sufficiently small step size, converges to a local minimum almost surely.*

In this section, based on the analysis in [5], we show that $f + g$ satisfy the strict saddle property, which implies that gradient descent can avoid saddle points and converge to the global minimum, with high probability.

Theorem 3.3. *Consider noiseless measurements $b = \mathcal{A}(X^\star)$, with $\mathcal{A}$ satisfying RIP with constant $\delta_{4r} \leq \frac{1}{100}$. Assume that rank$(X^\star) = r$. Let (U, V) be a pair of factors that satisfies the first order optimality condition $\nabla f(W) = 0$, for $W = \begin{bmatrix} U \\ V \end{bmatrix}$, and $UV^\top \neq X^\star$. Then,*

$$\lambda_{\min}\left(\nabla^2(f+g)(W)\right) \leq -\frac{1}{7} \cdot \sigma_r(X^\star).$$

Proof. Let $Z \in \mathbb{R}^{(m+n)\times r}$. Then, by (10), the proof of Theorem 2.1 and the fact that $b = \mathcal{A}(X^\star)$ (noiseless), $\nabla^2(f+g)(W)$ satisfies the following:

$$\begin{aligned}
&\texttt{vec}(Z)^\top \cdot \nabla^2(f+g)(W) \cdot \texttt{vec}(Z) \\
&\overset{(13),(12)}{\leq} \frac{1+2\delta_{2r}}{2} \sum_{j=1}^{r} \left\| W e_j e_j^\top (W - W^\star R) \right\|_F^2 \\
&\quad - \frac{3-8\delta_{2r}}{16} \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F^2 \\
&\overset{(14),(15)}{\leq} \left(\tfrac{1+2\delta_{2r}}{16} \cdot (1 + 34 \cdot 16\delta_{4r}^2) - \tfrac{3-8\delta_{2r}}{16} \right) \\
&\qquad \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F^2 \\
&\leq \tfrac{-1+5\delta_{4r}+272\delta_{4r}^2+544\delta_{4r}^3}{8} \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F^2 \\
&\leq -\tfrac{1}{10} \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F
\end{aligned} \tag{9}$$

where the last inequality is due to the requirement $\delta_{4r} \leq \frac{1}{100}$. For the LHS of (9), we can lower bound as follows:

$$\begin{aligned}
&\texttt{vec}(Z)^\top \cdot \nabla^2(f+g)(W) \cdot \texttt{vec}(Z) \\
&\qquad \geq \|Z\|_F^2 \cdot \lambda_{\min}\left(\nabla^2(f+g)(W)\right) \\
&\qquad = \|W - W^\star R\|_F^2 \cdot \lambda_{\min}\left(\nabla^2(f+g)(W)\right)
\end{aligned}$$

where the last equality is by setting $Z = W - W^\star R$. Combining this expression with (9), we obtain:

$$\begin{aligned}
&\lambda_{\min}\left(\nabla^2(f+g)(W)\right) \\
&\leq -\frac{1/10}{\|W - W^\star R\|_F^2} \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F \\
&\overset{(a)}{\leq} -\frac{1/10}{\|W - W^\star R\|_F^2} \cdot 2(\sqrt{2}-1) \cdot \sigma_r(X^\star) \cdot \|W - W^\star R\|_F^2 \\
&\leq -\frac{1}{7} \cdot \sigma_r(X^\star),
\end{aligned}$$

where (a) is due to Lemma 5.4, [43]. This completes the proof. $\square$

4 Proof of Main Results

We first describe the first- and second-order optimality conditions for $f+g$ objective with W variable. Then, we provide a detailed proof of the main results: by carefully analyzing the conditions, we study how a local optimum is related to the global optimum.

4.1 Gradient and Hessian of f and g

The gradients of f and g w.r.t. W are given by:

$$\nabla f(W) = \sum_{i=1}^{p} (\left\langle B_i, WW^\top \right\rangle - b_i) \cdot B_i \cdot W$$

$$\nabla g(W) = \frac{1}{4}\tilde{W}\tilde{W}^\top W \quad \left(\equiv \frac{1}{4} \cdot \begin{bmatrix} U \\ -V \end{bmatrix} \cdot (U^\top U - V^\top V) \right)$$

Regarding Hessian information, we are interested in the positive semi-definiteness of $\nabla^2 (f+g)$; for this case, it is easier to write the second-order Hessian information with respect to to some matrix direction $Z \in \mathbb{R}^{(m+n)\times r}$, as follows:

$$\begin{aligned}
&\texttt{vec}(Z)^\top \cdot \nabla^2 f(W) \cdot \texttt{vec}(Z) \\
&\quad = \left\langle \lim_{t\to 0} \left[\tfrac{\nabla f(W+tZ) - \nabla f(W)}{t} \right], Z \right\rangle \\
&\quad = \sum_{i=1}^{p} \langle B_i, ZW^\top + WZ^\top \rangle \cdot \langle B_i, ZW^\top \rangle \\
&\qquad + \sum_{i=1}^{p} \left(\langle B_i, WW^\top \rangle - b_i \right) \cdot \langle B_i, ZZ^\top \rangle
\end{aligned}$$

and

$$\begin{aligned}
&\texttt{vec}(Z)^\top \cdot \nabla^2 g(W) \cdot \texttt{vec}(Z) \\
&\quad = \left\langle \lim_{t\to 0} \left[\tfrac{\nabla g(W+tZ) - \nabla g(W)}{t} \right], Z \right\rangle \\
&\quad = \frac{1}{4} \left\langle \tilde{Z}\tilde{W}^\top, ZW^\top \right\rangle + \frac{1}{4} \left\langle \tilde{W}\tilde{Z}^\top, ZW^\top \right\rangle \\
&\qquad + \frac{1}{4} \left\langle \tilde{W}\tilde{W}^\top, ZZ^\top \right\rangle .
\end{aligned}$$

4.2 Optimality conditions

Given the expressions above, we now describe first- and second-order optimality conditions on the composite objective $f+g$.

First-order optimality condition. By the first-order optimality condition of a pair $(U,\ V)$ such that $W = \begin{bmatrix} U \\ V \end{bmatrix}$, we have $\nabla (f+g)(W) = 0$. This further implies:

$$\nabla (f+g)(W) = 0 \quad \Rightarrow$$
$$\sum_{i=1}^{p} \left(\langle B_i, WW^\top \rangle - b_i \right) \cdot B_i \cdot W + \frac{1}{4}\tilde{W}\tilde{W}^\top W = 0 \tag{11}$$

Second-order optimality condition. For a point W that satisfies the second-order optimality condition $\nabla^2 (f+g)(W) \succeq 0$, (10) holds for any $Z \in \mathbb{R}^{(m+n)\times r}$.

4.3 Proof of Theorem 2.1

Suppose that W is a critical point satisfying the optimality conditions (11) and (10). As in [5], we sum up the above condition for $Z_1 \triangleq (W - W^\star R) e_1 e_1^\top, \dots, Z_r \triangleq (W - W^\star R) e_r e_r^\top$. For simplicity, we first assume $Z = W - W^\star R$.

Bounding terms (A), (C) and (D) in (10)**.** The following bounds work for any Z.

$$\begin{aligned}
(A) &= \sum_{i=1}^{p} \left\langle B_i, ZW^\top \right\rangle^2 + \sum_{i=1}^{p} \left\langle B_i, ZW^\top \right\rangle \cdot \left\langle B_i, WZ^\top \right\rangle \\
&\overset{(a)}{=} 2 \cdot \sum_{i=1}^{p} \left\langle B_i, ZW^\top \right\rangle^2 \\
&= \frac{1}{2} \sum_{i=1}^{p} \left(\left\langle A_i, Z_U V^\top \right\rangle + \left\langle A_i, U Z_V^\top \right\rangle \right)^2 \\
&\overset{(b)}{\leq} \frac{1+\delta_{2r}}{2} \left\| Z_U V^\top \right\|_F^2 + \frac{1+\delta_{2r}}{2} \left\| U Z_V^\top \right\|_F^2 \\
&\quad + \left\langle Z_U V^\top, U Z_V^\top \right\rangle + \delta_{2r} \cdot \left\| Z_U V^\top \right\|_F \cdot \left\| U Z_V^\top \right\|_F \\
&\overset{(c)}{\leq} \underbrace{\frac{1+2\delta_{2r}}{2} \left\| Z_U V^\top \right\|_F^2 + \frac{1+2\delta_{2r}}{2} \left\| U Z_V^\top \right\|_F^2}_{(A1)} \\
&\quad + \underbrace{\left\langle Z_U V^\top, U Z_V^\top \right\rangle}_{(A2)}
\end{aligned}$$

where (a) follows from that every B_i is symmetric, (b) follows from Proposition 1.2, and (c) follows from the AM-GM inequality. We also have

$$\begin{aligned}
(C) = \left\langle \tilde{Z}\tilde{W}^\top, ZW^\top \right\rangle &= \left\| Z_U U^\top \right\|_F^2 \\
&+ \left\| Z_V V^\top \right\|_F^2 \\
&- \left\| Z_U V^\top \right\|_F^2 - \left\| Z_V U^\top \right\|_F^2 ,
\end{aligned}$$

$$(A1) + \frac{1}{4}(C) \leq \frac{1+4\delta_{2r}}{4} \left\| ZW^\top \right\|_F^2 ,$$

$$
\begin{aligned}
&\mathtt{vec}(Z)^\top \cdot \nabla^2 (f+g)(W) \cdot \mathtt{vec}(Z) \geq 0 \\
&\qquad \underbrace{\sum_{i=1}^p \langle B_i, ZW^\top + WZ^\top \rangle \cdot \left\langle B_i, ZW^\top \right\rangle}_{(A)} + \underbrace{\sum_{i=1}^p \left(\left\langle B_i, WW^\top \right\rangle - b_i \right) \cdot \left\langle B_i, ZZ^\top \right\rangle}_{(B)} \\
&\qquad\qquad + \frac{1}{4} \underbrace{\left\langle \tilde{Z}\tilde{W}^\top, ZW^\top \right\rangle}_{(C)} + \frac{1}{4} \underbrace{\left\langle \tilde{W}\tilde{Z}^\top, ZW^\top \right\rangle}_{(D)} + \frac{1}{4} \underbrace{\left\langle \tilde{W}\tilde{W}^\top, ZZ^\top \right\rangle}_{(E)} \geq 0, \quad \forall Z = \begin{bmatrix} Z_U \\ Z_V \end{bmatrix} \in \mathbb{R}^{(m+n)\times r}. \quad (10)
\end{aligned}
$$

$$
\begin{aligned}
(D) &= \left\langle \tilde{W}\tilde{Z}^\top, ZW^\top \right\rangle = \left\langle UZ_U^\top, Z_U U^\top \right\rangle \\
&\quad + \left\langle VZ_V^\top, Z_V V^\top \right\rangle - \left\langle UZ_V^\top, Z_U V^\top \right\rangle \\
&\quad - \left\langle VZ_U^\top, Z_V U^\top \right\rangle, \\
(A2) + \frac{1}{4}(D) &= \frac{1}{4} \left\langle WZ^\top, ZW^\top \right\rangle, \\
(A) + \frac{1}{4}(C) + \frac{1}{4}(D) &\leq \frac{1}{8} \left\| WZ^\top + ZW^\top \right\|_F^2 \\
&\quad + \delta_{2r} \left\| ZW^\top \right\|_F^2 .
\end{aligned}
$$

Bounding terms (B) and (E). We have

$$
\begin{aligned}
(B) &= \sum_{i=1}^p \left(\left\langle B_i, WW^\top \right\rangle - y_i \right) \cdot \left\langle B_i, ZZ^\top \right\rangle \\
&\overset{(a)}{=} - \sum_{i=1}^p \left(\left\langle B_i, WW^\top \right\rangle - y_i \right) \cdot \left\langle B_i, WW^\top - W^\star W^{\star\top} \right\rangle \\
&\quad - \frac{1}{2} \left\langle \tilde{W}\tilde{W}^\top, (W - W^\star R)W^\top \right\rangle \\
&= - \sum_{i=1}^p \left\langle B_i, WW^\top \right\rangle^2 - \frac{1}{2} \left\langle \tilde{W}\tilde{W}^\top, (W - W^\star R)W^\top \right\rangle \\
&\quad - \sum_{i=1}^p \left(\left\langle B_i, W^\star W^{\star\top} \right\rangle - y_i \right) \cdot \left\langle B_i, WW^\top - W^\star W^{\star\top} \right\rangle \\
&\overset{(b)}{\leq} - \underbrace{(1-\delta_{2r}) \left\| UV^\top - U^\star V^{\star\top} \right\|_F^2}_{(B1)} - \frac{1}{4} \cdot \underbrace{\left\langle \tilde{W}\tilde{W}^\top, 2ZW^\top \right\rangle}_{(B2)} \\
&\quad \underbrace{- \sum_{i=1}^p \left(\left\langle B_i, W^\star W^{\star\top} \right\rangle - y_i \right) \cdot \left\langle B_i, WW^\top - W^\star W^{\star\top} \right\rangle}_{(B3)}
\end{aligned}
$$

where at (a) we add the first-order optimality equation

$$
\begin{aligned}
&\left\langle \sum_{i=1}^p \left(\left\langle B_i, WW^\top \right\rangle - y_i \right) \cdot B_i \cdot W, 2W - 2W^\star R \right\rangle \\
&= -\frac{1}{2} \left\langle \tilde{W}\tilde{W}^\top W, W - W^\star R \right\rangle,
\end{aligned}
$$

and (b) follows from Proposition 1.2. Then we have

$$
\begin{aligned}
&(B2) - (E) \\
&= \left\langle \tilde{W}\tilde{W}^\top, 2ZW^\top - ZZ^\top \right\rangle \\
&\overset{(a)}{=} \left\langle \tilde{W}\tilde{W}^\top,\ 2WW^\top - W^\star RW^\top - WR^\top W^{\star\top} \right. \\
&\qquad \left. - (W - W^\star R)(W - W^\star R)^\top \right\rangle \\
&= \left\langle \tilde{W}\tilde{W}^\top, WW^\top - W^\star W^{\star\top} \right\rangle \\
&\overset{(b)}{=} \left\langle \tilde{W}\tilde{W}^\top, WW^\top - W^\star W^{\star\top} \right\rangle + \left\langle \tilde{W}^\star\tilde{W}^{\star\top}, W^\star W^{\star\top} \right\rangle \\
&\overset{(c)}{\geq} \left\langle \tilde{W}\tilde{W}^\top, WW^\top - W^\star W^{\star\top} \right\rangle + \left\langle \tilde{W}^\star\tilde{W}^{\star\top}, W^\star W^{\star\top} \right\rangle \\
&\qquad - \left\langle \tilde{W}^\star\tilde{W}^{\star\top}, WW^\top \right\rangle \\
&= \left\langle \tilde{W}\tilde{W}^\top - \tilde{W}^\star\tilde{W}^{\star\top}, WW^\top - W^\star W^{\star\top} \right\rangle
\end{aligned}
$$

where (a) follows from that $\tilde{W}\tilde{W}^\top$ is symmetric, (b) follows from Proposition 1.3, (c) follows from that the inner product of two PSD matrices is non-negative. We then have,

$$
\begin{aligned}
&(B1) + \frac{1}{4}(B2) - \frac{1}{4}(E) \\
&\geq (1-\delta_{2r}) \left\| UV^\top - U^\star V^{\star\top} \right\|_F^2 \\
&\quad + \frac{1}{4} \left\langle \tilde{W}\tilde{W}^\top - \tilde{W}^\star\tilde{W}^{\star\top}, WW^\top - W^\star W^{\star\top} \right\rangle \\
&= \left(1 - \delta_{2r} - \frac{1}{2} \right) \left\| UV^\top - U^\star V^{\star\top} \right\|_F^2 \\
&\quad + \frac{1}{4} \left\| UU^\top - U^\star U^{\star\top} \right\|_F^2 + \frac{1}{4} \left\| VV^\top - V^\star V^{\star\top} \right\|_F^2 \\
&\geq \frac{1 - 2\delta_{2r}}{4} \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F^2
\end{aligned}
$$

For (B3), we have

$$
\begin{aligned}
&-(B3) \\
&= \sum_{i=1}^p \left(\left\langle B_i, W^\star W^{\star\top} \right\rangle - b_i \right) \cdot \left\langle B_i, WW^\top - W^\star W^{\star\top} \right\rangle \\
&\overset{(a)}{\leq} \left\| \mathcal{A}(U^\star V^{\star\top}) - b \right\| \cdot \left(\sum_{i=1}^p \left\langle B_i, WW^\top - W^\star W^{\star\top} \right\rangle^2 \right)^{\frac{1}{2}} \\
&\overset{(b)}{\leq} \sqrt{1+\delta_{2r}} \cdot \left\| \mathcal{A}(U^\star V^{\star\top}) - b \right\| \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F
\end{aligned}
$$

where (a) follows from the Cauchy-Schwarz inequality, and (b) follows from Proposition 1.2. We get

$$\begin{aligned}
&(B) + \frac{1}{4}(E) \\
&\le -\frac{1-2\delta_{2r}}{4} \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F^2 \\
&\quad + \sqrt{1+\delta_{2r}} \cdot \left\| \mathcal{A}(U^\star V^{\star\top}) - b \right\| \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F \\
&\le -\frac{3-8\delta_{2r}}{16} \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F^2 \\
&\quad + 16(1+\delta_{2r}) \cdot \left\| \mathcal{A}(U^\star V^{\star\top}) - b \right\|^2 \qquad (12)
\end{aligned}$$

where the last inequality follows from the AM-GM inequality.

Summing up the inequalities for $Z_1, \ldots, Z_r$**.** Now we apply $Z_j = Z e_j e_j^\top$. Since $ZZ^\top = \sum_{j=1}^r Z_j Z_j^\top$ in (10), the analysis does not change for (B) and (E). For (A), (C), and (D), we obtain

$$\begin{aligned}
&(A) + \frac{1}{4}(C) + \frac{1}{4}(D) \\
&\quad \le \sum_{j=1}^r \left\{ \frac{1}{8} \left\| WZ_j^\top + Z_j W^\top \right\|_F^2 + \delta_{2r} \left\| Z_j W^\top \right\|_F^2 \right\}
\end{aligned}$$

We have

$$\begin{aligned}
&\sum_{j=1}^r \left\| WZ_j^\top + Z_j W^\top \right\|_F^2 \\
&= 2 \cdot \sum_{j=1}^r \left\| W e_j e_j^\top Z^\top \right\|_F^2 + 2 \cdot \sum_{i=1}^r \left\langle W e_j e_j^\top Z^\top, Z e_j e_j^\top W^\top \right\rangle \\
&= 2 \cdot \sum_{j=1}^r \left\| W e_j e_j^\top Z^\top \right\|_F^2 + 2 \cdot \sum_{i=1}^r (e_j^\top Z^\top W e_j)^2 \\
&\le 2 \cdot \sum_{j=1}^r \left\| W e_j e_j^\top Z^\top \right\|_F^2 + 2 \cdot \sum_{i=1}^r \|Z e_j\|^2 \cdot \|W e_j\|^2 \\
&= 4 \cdot \sum_{j=1}^r \left\| W e_j e_j^\top Z^\top \right\|_F^2
\end{aligned}$$

where the inequality follows from the Cauchy-Schwarz inequality. Applying this bound, we get

$$\begin{aligned}
&(A) + \frac{1}{4}(C) + \frac{1}{4}(D) \\
&\quad \le \frac{1+2\delta_{2r}}{2} \sum_{j=1}^r \left\| W e_j e_j^\top (W - W^\star R) \right\|_F^2. \qquad (13)
\end{aligned}$$

Next, we re-state [5, Lemma 4.4]:

Lemma 4.1. *Let W and $W^\star$ be two matrices, and Q is an orthonormal matrix that spans the column space of W. Then, there exists an orthonormal matrix R such that, for any stationary point W of $g(W)$ that satisfies first and second order condition, the following holds:*

$$\begin{aligned}
&\sum_{j=1}^r \|W e_j e_j^\top (W - W^\star R)\|_F^2 \\
&\quad \le \tfrac{1}{8} \cdot \|WW^\top - W^\star W^{\star\top}\|_F^2 \\
&\quad\quad + \tfrac{34}{8} \cdot \|(WW^\top - W^\star W^{\star\top}) QQ^\top\|_F^2 \qquad (14)
\end{aligned}$$

And we have the following variant of [5, Lemma 4.2].

Lemma 4.2. *For any pair of points (U, V) that satisfies the first-order optimality condition, and $\mathcal{A}$ be a linear operator satisfying the RIP condition with parameter δ_{4r}, the following inequality holds:*

$$\begin{aligned}
&\frac{1}{4} \cdot \left\| (WW^\top - W^\star W^{\star\top}) QQ^\top \right\|_F \\
&\quad \le \delta_{4r} \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F \\
&\quad\quad + \sqrt{\frac{1+\delta_{2r}}{2}} \cdot \left\| \mathcal{A}(U^\star V^{\star\top}) - b \right\| \qquad (15)
\end{aligned}$$

Applying the above two lemmas, we can get

$$\begin{aligned}
&(A) + \frac{1}{4}(C) + \frac{1}{4}(D) \\
&\quad \le \tfrac{(1+2\delta_{2r}) \cdot (1+1088\delta_{4r}^2)}{16} \left\| WW^\top - W^\star W^{\star\top} \right\|_F^2 \\
&\quad\quad + 34(1+2\delta_{2r})(1+\delta_{2r}) \left\| \mathcal{A}(U^\star V^{\star\top}) - b \right\|^2. \qquad (16)
\end{aligned}$$

Final inequality. Plugging (16) and (12) to (10), we get

$$\begin{aligned}
&\tfrac{1-5\delta_{2r}-544\delta_{4r}^2-1088\delta_{2r}\delta_{4r}^2}{8(40+68\delta_{2r})(1+\delta_{2r})} \left\| WW^\top - W^\star W^{\star\top} \right\|_F^2 \\
&\quad \le \left\| \mathcal{A}(U^\star V^{\star\top}) - b \right\|^2,
\end{aligned}$$

which completes the proof.

5 Appendix: Proof of Lemma 4.2

The first-order optimality condition can be written as

$$\begin{aligned}
0 &= \langle \nabla (f+g)(W), Z \rangle \\
&= \sum_{i=1}^p \left(\left\langle B_i, WW^\top \right\rangle - b_i \right) \cdot \langle B_i W, Z \rangle + \frac{1}{4} \left\langle \tilde{W}\tilde{W}^\top W, Z \right\rangle \\
&= \sum_{i=1}^p \left\langle B_i, WW^\top - W^\star W^{\star\top} \right\rangle \left\langle B_i, ZW^\top \right\rangle \\
&\quad + \sum_{i=1}^p \left(\left\langle B_i, W^\star W^{\star\top} \right\rangle - b_i \right) \cdot \left\langle B_i, ZW^\top \right\rangle \\
&\quad + \frac{1}{4} \left\langle \tilde{W}\tilde{W}^\top, ZW^\top \right\rangle
\end{aligned}$$

$$= \frac{1}{2} \cdot \sum_{i=1}^{p} \left\langle A_i, UV^\top - U^\star V^{\star\top} \right\rangle \left\langle A_i, Z_U V^\top + U Z_V^\top \right\rangle$$
$$+ \frac{1}{2} \cdot \sum_{i=1}^{p} \left(\left\langle A_i, U^\star V^{\star\top} \right\rangle - b_i \right) \cdot \left\langle A_i, Z_U V^\top + U Z_V^\top \right\rangle$$
$$+ \frac{1}{4} \left\langle \tilde{W}\tilde{W}^\top, ZW^\top \right\rangle,$$

$\forall Z = \begin{bmatrix} Z_U \\ Z_V \end{bmatrix} \in \mathbb{R}^{(m+n)\times r}$. Applying Proposition 1.2 and the Cauchy-Schwarz inequality to the condition, we obtain

$$\frac{1}{2} \underbrace{\left\langle UV^\top - U^\star V^{\star\top}, Z_U V^\top + U Z_V^\top \right\rangle}_{(A)} + \frac{1}{4} \underbrace{\left\langle \tilde{W}\tilde{W}^\top, ZW^\top \right\rangle}_{(B)}$$
$$\leq \delta_{4r} \underbrace{\left\| UV^\top - U^\star V^{\star\top} \right\|_F \left\| Z_U V^\top + U Z_V^\top \right\|_F}_{(C)}$$
$$+ \frac{\sqrt{1+\delta_{2r}}}{2} \underbrace{\left\| \mathcal{A}(U^\star V^{\star\top}) - b \right\| \left\| Z_U V^\top + U Z_V^\top \right\|_F}_{(D)} \tag{17}$$

Let $Z = (WW^\top - W^\star W^{\star\top}) Q R^{-1\top}$ where $W = QR$ is the QR decomposition. Then we obtain

$$ZW^\top = (WW^\top - W^\star W^{\star\top}) QQ^\top.$$

We have

$$2(A) = 2 \left\langle \begin{bmatrix} 0 & UV^\top - U^\star V^{\star\top} \\ VU^\top - V^\star U^{\star\top} & 0 \end{bmatrix}, ZW^\top \right\rangle$$
$$= \Big\langle (WW^\top - \tilde{W}\tilde{W}^\top) - (W^\star W^{\star\top} - \tilde{W}^\star \tilde{W}^{\star\top}),$$
$$(WW^\top - W^\star W^{\star\top}) QQ^\top \Big\rangle$$

and

$$(B) = \left\langle \tilde{W}\tilde{W}^\top, (WW^\top - W^\star W^{\star\top}) QQ^\top \right\rangle$$
$$\stackrel{(a)}{=} \left\langle \tilde{W}\tilde{W}^\top, (WW^\top - W^\star W^{\star\top}) QQ^\top \right\rangle$$
$$+ \left\langle \tilde{W}^\star \tilde{W}^{\star\top}, W^\star W^{\star\top} QQ^\top \right\rangle$$
$$\stackrel{(b)}{\geq} \left\langle \tilde{W}\tilde{W}^\top, (WW^\top - W^\star W^{\star\top}) QQ^\top \right\rangle$$
$$- \left\langle \tilde{W}^\star \tilde{W}^{\star\top}, (WW^\top - W^\star W^{\star\top}) QQ^\top \right\rangle$$
$$= \left\langle \tilde{W}\tilde{W}^\top - \tilde{W}^\star \tilde{W}^{\star\top}, (WW^\top - W^\star W^{\star\top}) QQ^\top \right\rangle$$

where (a) follows from Proposition 1.3, and (b) follows from that the inner product of two PSD matrices is non-negative. Then we obtain

$$2(A) + (B)$$
$$\geq \left\langle WW^\top - W^\star W^{\star\top}, (WW^\top - W^\star W^{\star\top}) QQ^\top \right\rangle$$
$$= \left\| (WW^\top - W^\star W^{\star\top}) Q \right\|_F^2$$
$$= \left\| (WW^\top - W^\star W^{\star\top}) QQ^\top \right\|_F^2$$

For (C), we have

$$(C) = \left\| UV^\top - U^\star V^{\star\top} \right\|_F \cdot \left\| Z_U V^\top + U Z_V^\top \right\|_F$$
$$\leq \frac{1}{\sqrt{2}} \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F$$
$$\cdot \sqrt{2 \left\| Z_U V^\top \right\|_F^2 + 2 \left\| U Z_V^\top \right\|_F^2}$$
$$\leq \left\| WW^\top - W^\star W^{\star\top} \right\|_F \cdot \sqrt{\left\| ZW^\top \right\|_F^2}$$
$$= \left\| WW^\top - W^\star W^{\star\top} \right\|_F$$
$$\cdot \left\| (WW - W^\star W^{\star\top}) QQ^\top \right\|_F$$

Plugging the above bounds into (17), we get

$$\frac{1}{4} \left\| (WW^\top - W^\star W^{\star\top}) QQ^\top \right\|_F^2 \leq$$
$$\delta_{4r} \left\| WW^\top - W^\star W^{\star\top} \right\|_F \left\| (WW^\top - W^\star W^{\star\top}) QQ^\top \right\|_F$$
$$+ \sqrt{\tfrac{1+\delta_{2r}}{2}} \left\| \mathcal{A}(U^\star V^{\star\top}) - b \right\| \left\| (WW^\top - W^\star W^{\star\top}) QQ^\top \right\|_F$$

In either case of $\left\| (WW^\top - W^\star W^{\star\top}) QQ^\top \right\|_F$ being zero or positive, we can obtain

$$\frac{1}{4} \cdot \left\| (WW^\top - W^\star W^{\star\top}) QQ^\top \right\|_F$$
$$\leq \delta_{4r} \cdot \left\| WW^\top - W^\star W^{\star\top} \right\|_F$$
$$+ \sqrt{\tfrac{1+\delta_{2r}}{2}} \cdot \left\| \mathcal{A}(U^\star V^{\star\top}) - b \right\|$$

This completes the proof.

References

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

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

[3] A. Bandeira, N. Boumal, and V. Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. *arXiv preprint arXiv:1602.04426*, 2016.

[4] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi. Dropping convexity for faster semi-definite optimization. *arXiv preprint arXiv:1509.03917*, 2015.

[5] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.

[6] N. Boumal. Nonconvex phase synchronization. *arXiv preprint arXiv:1601.06114*, 2016.

[7] N. Boumal, V. Voroninski, and A. Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. *arXiv preprint arXiv:1606.04970*, 2016.

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

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

[10] E. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

[11] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

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

[13] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Sparse and low-rank matrix decompositions. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pages 962–967. IEEE, 2009.

[14] Y. Chen and M. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[15] A. Christoffersson. *The one component model with incomplete data*. Uppsala., 1970.

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

[17] C. De Sa, K. Olukotun, and C. Re. Global convergence of stochastic gradient descent for some non-convex matrix problems. *arXiv preprint arXiv:1411.1134*, 2014.

[18] M. Fazel, E. Candes, B. Recht, and P. Parrilo. Compressed sensing and robust recovery of low rank matrices. In *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, pages 1043–1047. IEEE, 2008.

[19] S. Flammia, D. Gross, Y.-K. Liu, and J. Eisert. Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012.

[20] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

[21] R. Ge, J. Lee, and T. Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.

[22] P. Jain, C. Jin, S. Kakade, and P. Netrapalli. Computing matrix squareroot via non convex local search. *arXiv preprint arXiv:1507.05854*, 2015.

[23] A. Javanmard and A. Montanari. Localization from incomplete noisy distance measurements. *Foundations of Computational Mathematics*, 13(3):297–345, 2013.

[24] C. Jin, S. Kakade, and P. Netrapalli. Provable efficient online matrix completion via nonconvex stochastic gradient descent. *arXiv preprint arXiv:1605.08370*, 2016.

[25] M. Journée, F. Bach, P-A Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

[26] A. Kalev, R. Kosut, and I. Deutsch. Quantum tomography protocols with positivity are compressed sensing protocols. *Nature partner journals (npj) Quantum Information*, 1:15018, 2015.

[27] F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

[28] A. Kyrillidis and V. Cevher. Matrix recipes for hard thresholding methods. *Journal of mathematical imaging and vision*, 48(2):235–265, 2014.

[29] L. Le and M. White. Global optimization of factor models using alternating minimization. *arXiv preprint arXiv:1604.04942*, 2016.

[30] J. Lee, M. Simchowitz, M. Jordan, and B. Recht. Gradient descent converges to minimizers. In *Proceedings of The 29th Conference on Learning Theory*, 2016.

[31] Y.-K. Liu. Universal low-rank matrix recovery from Pauli measurements. In *Advances in Neural Information Processing Systems*, pages 1638–1646, 2011.

[32] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.

[33] F. Mirzazadeh, Y. Guo, and D. Schuurmans. Convex co-embedding. In *AAAI*, pages 1989–1996, 2014.

[34] F. Mirzazadeh, M. White, A. György, and D. Schuurmans. Scalable metric learning for co-embedding. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 625–642. Springer, 2015.

[35] D. Park, A. Kyrillidis, S. Bhojanapalli, C. Caramanis, and S. Sanghavi. Provable non-convex projected gradient descent for a class of constrained matrix optimization problems. *arXiv preprint arXiv:1606.01316*, 2016.

[36] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Finding low-rank solutions to convex smooth problems via the Burer-Monteiro approach. In *54th Annual Allerton Conference on Communication, Control, and Computing*, 2016.

[37] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Finding low-rank solutions to matrix problems, efficiently and provably. *arXiv preprint arXiv:1606.03168*, 2016.

[38] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[39] A. Ruhe. *Numerical computation of principal components when several observations are missing*. Univ., 1974.

[40] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *arXiv preprint arXiv:1511.03607*, 2015.

[41] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*, 2016.

[42] R Sun and Z.-Q. Luo. Guaranteed matrix completion via nonconvex factorization. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015*, pages 270–289, 2015.

[43] S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. *arXiv preprint arXiv:1507.03566v2*, 2016.

[44] A. Waters, A. Sankaranarayanan, and R. Baraniuk. SpaRCS: Recovering low-rank and sparse matrices from compressive measurements. In *Advances in neural information processing systems*, pages 1089–1097, 2011.

[45] H. Wold and E. Lyttkens. Nonlinear iterative partial least squares (NIPALS) estimation procedures. *Bulletin of the International Statistical Institute*, 43(1), 1969.

[46] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. *arXiv preprint arXiv:1605.07784*, 2016.

[47] X. Zhang, D. Schuurmans, and Y. Yu. Accelerated training for matrix-norm regularization: A boosting approach. In *Advances in Neural Information Processing Systems*, pages 2906–2914, 2012.

[48] T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems 28*, pages 559–567. 2015.

[49] Q. Zheng and J. 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.

[50] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.