Global Optimality in Low-rank Matrix Optimization
Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B. Wakin

Abstract—This paper considers the minimization of a general objective function \( f(X) \) over the set of rectangular \( n \times m \) matrices where the optimal solution \( X^\star \) is low-rank. To reduce the computational burden, we factorize the variable \( X \) into a product of two smaller matrices and optimize over these two matrices instead of \( X \). Despite the resulting nonconvexity, recent studies in matrix completion and sensing have shown that the factored problem has no spurious local minima and obeys the so-called strict saddle property (the function has a directional negative curvature at all critical points but local minima). We analyze the global geometry for a general and yet well-conditioned objective function \( f(X) \) whose restricted strong convexity and restricted strong smoothness constants are comparable. In particular, we show that the reformulated objective function has no spurious local minima and obeys the strict saddle property. These geometric properties imply that a number of iterative optimization algorithms (such as gradient descent) can provably solve the factored problem with global convergence.

Keywords—Low-rank matrix optimization, matrix sensing, nonconvex optimization, optimization geometry, strict saddle

I. INTRODUCTION

Consider the minimization of a general objective function \( f(X) \) over all \( n \times m \) matrices:

\[
\minimize_{X \in \mathbb{R}^{n \times m}} f(X),
\]

which we suppose admits a low-rank solution \( X^\star \in \mathbb{R}^{n \times m} \) with \( \text{rank}(X^\star) = r^\star \). Low-rank matrix optimizations of the form (1) appear in a wide variety of applications, including quantum tomography [1], [2], collaborative filtering [3], [4], sensor localization [5], low-rank matrix recovery from compressive measurements [6], [7], and matrix completion [8], [9]. In order to find a low-rank solution, the nuclear norm is widely used in matrix inverse problems [10], [7] arising in machine learning [11], signal processing [12], and control [13]. Although nuclear norm minimization enjoys strong statistical guarantees [8], its computational complexity is very high (as most algorithms require performing an expensive singular value decomposition (SVD) in each iteration), prohibiting it from scaling to practical problems.

To relieve the computational bottleneck, recent studies propose to factorize the variable into the Burer-Monteiro type decomposition [14], [15] with \( X = UV^T \), and optimize over the \( n \times r \) and \( m \times r \) \( (r \geq r^\star) \) matrices \( U \) and \( V \). With this parameterization of \( X \), we can recast (1) into the following program:

\[
\minimize_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} h(U, V) := f(UV^T).
\]

The bilinear nature of the parameterization renders the objective function of (2) nonconvex even when \( f(X) \) is a convex function. Hence, the objective function in (2) can potentially have spurious local minima (i.e., local minimizers that are not global minimizers) or “bad” saddle points that prevent a number of iterative algorithms from converging to the global solution. By analyzing the landscape of nonconvex functions, several recent works have shown that with an exact factorization \( (r = r^\star) \), the factored objective function \( h(U, V) \) in certain matrix inverse problems has no spurious local minima [16], [17], [18].

We generalize this line of work by focusing on a general objective function \( f(X) \) in the optimization (1), not necessarily a quadratic loss function coming from a matrix inverse problem. By focusing on a general objective function, we attempt to provide a unifying framework for low-rank matrix optimizations with the factorization approach. We provide a geometric analysis for the factored program (2) and show that, under certain conditions of \( f(X) \), all critical points of the objective function \( h(U, V) \) are well-behaved. Our characterization of the geometry of the objective function ensures that a number of iterative optimization algorithms converge to a global minimum.

A. Summary of Results

The purpose of this paper is to analyze the geometry of the factored problem \( h(U, V) \) in (2). In particular, we attempt to understand the behavior of all of the critical points of the objective function in the reformulated problem (2).

Before presenting our main results, we lay out the necessary assumptions on the objective function \( f(X) \). As is known, without any assumptions on the problem, even minimizing traditional quadratic objective functions is challenging. For this purpose, we focus on the model where \( f(X) \) is \( (2r, 4r) \)-restricted strongly convex and smooth, i.e., for any \( n \times m \) matrices \( X, G \) with \( \text{rank}(X) \leq 2r \) and \( \text{rank}(G) \leq 4r \), the Hessian of \( f(X) \) satisfies

\[
\alpha \|G\|^2_F \leq \|\nabla^2 f(X)\|_{2,2} \leq \beta \|G\|^2_F
\]

for some positive \( \alpha \) and \( \beta \). A similar assumption is also utilized in [19] Conditions 5.3 and 5.4. With this assumption on \( f(X) \), we summarize our main results in the following informal theorem.

**Theorem 1.** (informal) Suppose the function \( f(X) \) satisfies the \((2r, 4r)\)-restricted strongly convexity and smoothness condition (3) and the program (1) has a low-rank minimizer program:

\[
\minimize_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} h(U, V) := f(UV^T).
\]

The bilinear nature of the parameterization renders the objective function of (2) nonconvex even when \( f(X) \) is a convex function. Hence, the objective function in (2) can potentially have spurious local minima (i.e., local minimizers that are not global minimizers) or “bad” saddle points that prevent a number of iterative algorithms from converging to the global solution. By analyzing the landscape of nonconvex functions, several recent works have shown that with an exact factorization \( (r = r^\star) \), the factored objective function \( h(U, V) \) in certain matrix inverse problems has no spurious local minima [16], [17], [18].

We generalize this line of work by focusing on a general objective function \( f(X) \) in the optimization (1), not necessarily a quadratic loss function coming from a matrix inverse problem. By focusing on a general objective function, we attempt to provide a unifying framework for low-rank matrix optimizations with the factorization approach. We provide a geometric analysis for the factored program (2) and show that, under certain conditions of \( f(X) \), all critical points of the objective function \( h(U, V) \) are well-behaved. Our characterization of the geometry of the objective function ensures that a number of iterative optimization algorithms converge to a global minimum.

A. Summary of Results

The purpose of this paper is to analyze the geometry of the factored problem \( h(U, V) \) in (2). In particular, we attempt to understand the behavior of all of the critical points of the objective function in the reformulated problem (2).

Before presenting our main results, we lay out the necessary assumptions on the objective function \( f(X) \). As is known, without any assumptions on the problem, even minimizing traditional quadratic objective functions is challenging. For this purpose, we focus on the model where \( f(X) \) is \( (2r, 4r) \)-restricted strongly convex and smooth, i.e., for any \( n \times m \) matrices \( X, G \) with \( \text{rank}(X) \leq 2r \) and \( \text{rank}(G) \leq 4r \), the Hessian of \( f(X) \) satisfies

\[
\alpha \|G\|^2_F \leq \|\nabla^2 f(X)\|_{2,2} \leq \beta \|G\|^2_F
\]

for some positive \( \alpha \) and \( \beta \). A similar assumption is also utilized in [19] Conditions 5.3 and 5.4. With this assumption on \( f(X) \), we summarize our main results in the following informal theorem.

**Theorem 1.** (informal) Suppose the function \( f(X) \) satisfies the \((2r, 4r)\)-restricted strongly convexity and smoothness condition (3) and the program (1) has a low-rank minimizer...
\(X^* \in \mathbb{R}^{n \times m} \) with \(\text{rank}(X^*) = r^* \leq r\). Then the factored objective function \(h(U, V)\) (with an additional regularizer) see Theorem 3 in Section II) has no spurious local minima and obeys the strict saddle property (see Definition 3 in Section II).

Remark 1. The above result implies that we can recover the rank-\(r^*\) global minimizer \(X^*\) of (1) by many iterative algorithms (such as the trust region method [20], stochastic gradient descent [21]) even from a random initialization. This is because 1) as guaranteed by Theorem 2, the strict saddle property ensures local search algorithms converge to a local minimum, and 2) there are no spurious local minima.

Remark 2. Since our main result only requires the \((2r, 4r)\)-restricted strong convexity and smoothness property (3), aside from low-rank matrix recovery [10], it can also be applied to many other low-rank matrix optimization problems [22] which do not necessarily involve quadratic loss functions. Typical examples include robust PCA [23], 1-bit matrix completion [25], [26] and Poisson principal component analysis (PCA) [27].

B. Related Works

Compared with the original program (1), the factored form (2) typically involves many fewer variables (or variables with much smaller size) and can be efficiently solved by simple but powerful methods (such as gradient descent [21], [28], the trust region method [29], and alternating methods [30]) for large-scale settings, though it is nonconvex. In recent years, tremendous effort has been devoted to analyzing nonconvex optimizations by exploiting the geometry of the corresponding objective functions. These works can be separated into two types based on whether the geometry is analyzed locally or globally. One type of works analyze the behavior of the objective function in a small neighborhood containing the global optimum and require a good initialization that is close enough to a global minimum. Problems such as phase retrieval [31], matrix sensing [32], and semi-definite optimization [33] have been studied.

Another type of works attempt to analyze the landscape of the objective function and show that it obeys the strict saddle property. If this particular property holds, then simple algorithms such as gradient descent and the trust region method are guaranteed to converge to a local minimum from a random initialization [28], [21], [34] rather than requiring a good guess. We approach low-rank matrix optimization with general objective functions (1) via a similar geometric characterization. Similar geometric results are known for a number of problems including complete dictionary learning [34], phase retrieval [20], orthogonal tensor decomposition [21], and matrix inverse problems [16], [17], [35]. Empirical evidence also supports using the factorization approach for estimating a low-rank PSD matrix from a set of rank-one measurements corrupted by arbitrary outliers [36] and for recovering a dynamically evolving low-rank matrix from incomplete observations [37].

Our work is most closely related to certain recent works in low-rank matrix optimization. Bhojanapalli et al. [16] showed that the low-rank, positive semi-definite (PSD) matrix sensing problem has no spurious local minima and obeys the strict saddle property. Similar results were exploited for PSD matrix completion [17], PSD matrix factorization [38] and low-rank, PSD matrix optimization problems with generic objective functions [39]. Our work extends this line of research to low-rank matrix (not necessary PSD or even square) optimization problems. Another closely related work considers the low-rank, non-square matrix sensing problem with the factorization approach [18], [40]. We note that our general objective function framework includes the low-rank matrix sensing problem as a special case (see Section II-C). Furthermore, our result covers both over-parameterization where \(r > r^*\) and exact parameterization where \(r = r^*\). Wang et al. [19] also considered the factored low-rank matrix minimization problem with a general objective function which satisfies the restricted strong convexity and smoothness condition. Their algorithms require good initializations for global convergence since they characterized only the local landscapes around the global optima. By categorizing the behavior of all the critical points, our work differs from [19] in that we instead characterize the global landscape of the factored objective function.

This paper continues in Section IV with formal definitions for strict saddles and the strict saddle property. We present the main results and their implications in matrix sensing, weighted low-rank approximation and 1-bit matrix completion in Section V. The proof of our main results is given in Section VI. We conclude the paper in Section VI.

II. Preliminaries

A. Notation

To begin, we first briefly introduce some notation used throughout the paper. The symbols I and 0 respectively represent the identity matrix and zero matrix with appropriate sizes. The set of \(r \times r\) orthonormal matrices is denoted by \(O_r := \{R \in \mathbb{R}^{r \times r} : R^T R = I\}\). If a function \(h(U, V)\) has two arguments, \(U \in \mathbb{R}^{n \times r}\) and \(V \in \mathbb{R}^{m \times r}\), we occasionally use the notation \(h(W)\) when we put these two arguments into a new one as \(W = \begin{bmatrix} U \\ V \end{bmatrix}\). For a scalar function \(f(Z)\) with a matrix variable \(Z \in \mathbb{R}^{n \times m}\), its gradient is an \(n \times m\) matrix whose \((i,j)\)-th entry is \(\nabla f(Z)_{ij} = \frac{\partial f(Z)}{\partial z_{ij}}\) for all \(i \in [n], j \in [m]\). Here \([n] = \{1, 2, \ldots, n\}\) for any \(n \in \mathbb{N}\) and \([m] = \{1, 2, \ldots, m\}\). The Hessian of \(f(Z)\) can be viewed as an \(nm \times nm\) matrix \(\nabla^2 f(Z)\) for all \(i, j \in [nm]\), where \(z_i\) is the \(i\)-th entry of the vectorization of \(Z\). An alternative way to represent the Hessian is by a bilinear form defined via \(\nabla^2 f(Z)(A, B) = \sum_{i,j,k,l} \frac{\partial^2 f(Z)}{\partial z_i \partial z_j} A_{ij} B_{kl}\) for any \(A, B \in \mathbb{R}^{n \times m}\). The bilinear form for the Hessian is widely utilized through the paper.

B. Strict Saddle Property

Suppose \(h : \mathbb{R}^n \rightarrow \mathbb{R}\) is a twice continuously differentiable objective function. We begin with the notion of strict saddles and the strict saddle property.
Definition 1 (Critical points). We say \( x \) a critical point if the gradient at \( x \) vanishes, i.e., \( \nabla h(x) = 0 \).

Definition 2 (Strict saddles). A critical point \( x \) is a strict saddle if the Hessian matrix evaluated at this point has a strictly negative eigenvalue, i.e., \( \lambda_{\min}(\nabla^2 h(x)) < 0 \).

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

Intuitively, the strict saddle property requires a function to have a directional negative curvature at all critical points but local minima. This property allows a number of iterative algorithms such as noisy gradient descent \([21]\) and the trust region method \([21]\) to further decrease the function value at all the strict saddles and thus converge to a local minimum.

Theorem 2. \([29], [21], [28]\) (informal) For a twice continuously differentiable objective function satisfying the strict saddle property, a number of iterative optimization algorithms (such as gradient descent and the trust region method) can find a local minimum.

### III. Problem Formulation and Main Results

#### A. Problem Formulation

This paper considers the problem \((1)\) of minimizing a general function \( f(X) \) admitting a low-rank solution \( X^\star \) with \( \text{rank}(X^\star) = r^\star < r \). We factorize the variable \( X = UV^T \) with \( U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r} \) and transform \((1)\) into its factored counterpart \((2)\). Through the paper, \( X, W \) and \( \tilde{W} \) are matrices depending on \( U \) and \( V \):

\[
W = \begin{bmatrix} U \\ V \end{bmatrix}, \quad \tilde{W} = \begin{bmatrix} U \\ -V \end{bmatrix}, \quad X = UV^T.
\]

Although the new variable \( W \) has much smaller size than \( X \) when \( r \ll \min\{n, m\} \), the objective function in the factored problem \((2)\) may have a much more complicated landscape due to the bilinear form about \( U \) and \( V \). The reformulated objective function \( h(U, V) \) could introduce spurious local minima or degenerate saddle points even when \( f(X) \) is convex. Our goal is to guarantee that this does not happen.

Let \( X^\star = Q_U \Sigma^\star Q_V^T \) denote an SVD of \( X^\star \), where \( Q_U \in \mathbb{R}^{n \times r} \) and \( Q_V \in \mathbb{R}^{m \times r} \) are orthonormal matrices of appropriate sizes, and \( \Sigma^\star \in \mathbb{R}^{r \times r} \) is a diagonal matrix with non-negative diagonal (but with some zeros on the diagonal if \( r > r^\star = \text{rank}(X^\star) \)). We denote

\[
U^\star = Q_U \Sigma^{1/2}, \quad V^\star = Q_V \Sigma^{1/2},
\]

where \( X^\star = U^\star V^\star^T \) forms a balanced factorization of \( X^\star \) since \( U^\star \) and \( V^\star \) have the same singular values. Throughout the paper, we utilize the following two ways to stack \( U^\star \) and \( V^\star \) together:

\[
W^\star = \begin{bmatrix} U^\star \\ V^\star \end{bmatrix}, \quad \tilde{W}^\star = \begin{bmatrix} U^\star \\ -V^\star \end{bmatrix}.
\]

Before moving on, we note that for any solution \((U, V)\) to \((2)\), \((U \Phi, V \Phi)\) is also a solution to \((3)\) for any \( \Phi \in \mathbb{R}^{r \times r} \) such that \( U \Phi \Phi^T V^T = U V^T \). In order to address this ambiguity (i.e., to reduce the search space of \( W \) for \((2)\)), we utilize the trick in \([22], [18], [19]\) by introducing a regularizer

\[
g(U, V) = \frac{\mu}{4} \left\| U^T U - V^T V \right\|_F^2
\]

and solving the following problem

\[
\text{minimize}_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} \rho(U, V) := f(UV^T) + g(U, V), \quad (5)
\]

where \( \mu > 0 \) controls the weight for the term \( \left\| U^T U - V^T V \right\|_F^2 \), which will be discussed soon.

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

\[
\mathcal{E} := \left\{ W = \begin{bmatrix} U \\ V \end{bmatrix} : U^T U - V^T V = 0 \right\}.
\]

Informally, we can view \((5)\) as finding a point from \( \mathcal{E} \) that also minimizes \( f(UV^T) \). This is formally established in Theorem 3.

#### B. Main Results

Our main argument is that, under certain conditions on \( f(X) \), the objective function \( \rho(W) \) has no spurious local minima and satisfies the strict saddle property. This is equivalent to categorizing all the critical points into two types: 1) the global minima which correspond to the global solution of the original convex problem \((1)\) and 2) strict saddles such that the Hessian matrix \( \nabla^2 \rho(W) \) evaluated at these points has a strictly negative eigenvalue. We formally establish this in the following theorem, whose proof is given in the next section.

Theorem 3. For any \( \mu > 0 \), each critical point \( W = \begin{bmatrix} U \\ V \end{bmatrix} \) of \( \rho(W) \) defined in \((5)\) satisfies

\[
U^T U - V^T V = 0.
\]

Furthermore, suppose the function \( f(X) \) satisfies the \((2r, 4r)\)-restricted strong convexity and smoothness condition \((3)\) with positive constants \( \alpha \) and \( \beta \) satisfying \( \beta \leq 1.5 \). Set \( \mu \leq \frac{\alpha}{16\pi} \) for the factored problem \((5)\). Then \( \rho(W) \) has no spurious local minima, i.e., any local minimum of \( \rho(W) \) is a global minimum corresponding to the global solution of the original convex problem \((1)\): \( U V^T = X^\star \). In addition, \( \rho(W) \) obeys

---

\footnote{We note that this minimizer need not be unique, and in particular that there could exist other minimizers that are high-rank (however, under our scenarios of interest, Proposition \( [4] \) forbids the existence of other low-rank solutions.)}
the strict saddle property that any critical point not being a local minimum is a strict saddle with
\[ \lambda_{\text{min}}(\nabla^2(\rho(W))) \leq \begin{cases} 
-0.08\alpha \sigma_r((X^*)) & , \ r = r^* \\
-0.5\alpha \min\{\sigma^2_r(W), 2\sigma_r, ((X^*))\} & , \ r > r^* \\
-0.1\alpha \sigma_r((X^*)) & , \ r = 0,
\end{cases} \] (8)
where \( r^* \leq r \) is the rank of \( W \), \( \lambda_{\text{min}}(\cdot) \) represents the smallest eigenvalue, and \( \sigma_r(\cdot) \) denotes the \( \ell^2 \)-th largest singular value.

**Remark 3.** Equation (8) shows that any critical point \( W \) belongs to \( \mathcal{E} \) for the objective function in the factored problem (5) with any positive \( \mu \). This demonstrates the reason for adding the regularizer \( g(U, V) \). Thus, any iterative optimization algorithm converging to some critical point of \( \rho(W) \) results in a solution within \( \mathcal{E} \). Furthermore, the strict saddle property along with the lack of spurious local minima ensures that a number of optimization algorithms find the global minimum.

**Remark 4.** For any critical point \( W \in \mathbb{R}^{(n+m)\times r} \) that is not a local minimum, the right hand side of (8) is strictly negative, implying \( W \) is a strict saddle. We also note that Theorem 3 not only covers exact parameterization where \( r = r^* \), but also includes over-parameterization where \( r > r^* \).

**Remark 5.** The constants appearing in Theorem 3 are not optimized. We use \( \mu \leq \frac{1}{3\alpha} \) simply to include \( \mu = \frac{1}{4} \) which is utilized for the matrix sensing problem in [52]. If the ratio between the restricted strong convexity and smoothness constants \( \frac{\mu}{\alpha^3} \leq 1.4 \), then we can show that \( \rho(W) \) has no spurious local minima and obeys the strict saddle property for any \( \mu \leq \frac{1}{4\alpha} \) (where \( \mu = \frac{1}{4} \) is utilized for the matrix sensing problem in [18]). In all cases, a smaller \( \mu \) yields a more negative constant in (8); see Section IV for more discussion on this. This implies that when the restricted strong convexity constant \( \alpha \) is not provided a priori, one can always choose a small \( \mu \) to ensure the strict saddle property holds, and hence guarantee the global convergence of many iterative optimization algorithms.

**Remark 6.** We finally remark that although Theorem 3 requires the additional regularizer [4], empirical evidence (see experiments in Section IV) shows we can get rid of this regularizer for many iterative algorithms with random initialization.

We prove Theorem 3 in Section IV. Before proceeding, we present two stylized applications of Theorem 3 in matrix sensing and weighted low-rank approximation.

**C. Stylized Applications**

1) **Matrix Sensing:** We first consider the implication of Theorem 3 in the matrix sensing problem where
\[ f(X) = \frac{1}{2} \| A(X - X^*) \|_F^2. \]
Here \( A : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^p \) is a known measurement operator satisfying the following restricted isometry property.

**Definition 4.** *(Restricted Isometry Property (RIP) [7])* The map \( \mathcal{A} : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^p \) satisfies the \( r \)-RIP with constant \( \delta_r \) if
\[ (1 - \delta_r) \| X \|_F^2 \leq \| A(X) \|_2^2 \leq (1 + \delta_r) \| X \|_F^2 \] (9)
holds for any \( n \times m \) matrix \( X \) with \( \text{rank}(X) \leq r \).

Note that in this case, the Hessian quadrature form \( \nabla^2 f(X)[Y, Y] \) for any \( n \times m \) matrices \( X \) and \( Y \) is given by
\[ \nabla^2 f(X)[Y, Y] = \| A(Y) \|_2^2. \]
If \( \mathcal{A} \) satisfies the \( 4r \)-restricted isometry property with constant \( \delta_{4r} \), then \( f(X) \) satisfies the \( (2r, 4r) \)-restricted strong convexity and smoothness condition [3] with constants \( \alpha = 1 - \delta_{4r} \) and \( \beta = 1 - \delta_{4r} \) since
\[ (1 - \delta_{4r}) \| Y \|_F^2 \leq \| A(Y) \|_2^2 \leq (1 + \delta_{4r}) \| Y \|_F^2 \]
for any rank-\( 4r \) matrix \( Y \). Now, applying Theorem 3 we can characterize the geometry for the following matrix sensing problem with the factorization approach:
\[ \min_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} \frac{1}{2} \| A(UV^T - X^*) \|_2^2 + g(U, V), \] (10)
where \( g(U, V) \) is the added regularizer defined in [4].

**Corollary 1.** Suppose \( \mathcal{A} \) satisfies the \( 4r \)-RIP with constant \( \delta_{4r} \leq \frac{1}{2} \), and set \( \mu = \frac{1}{1 - 2\delta_{4r}} \). Then the objective function in (10) has no spurious local minima and satisfies the strict saddle property.

This result follows directly from Theorem 3 by noting that \( \frac{\mu}{\alpha^3} = \frac{1 + 2\delta_{4r}}{1 - 2\delta_{4r}} \leq 1.5 \) if \( \delta_{4r} \leq \frac{1}{2} \). We remark that Park et al. [18, Theorem 4.3] provided a similar geometric result for (10). Compared to their result which requires \( \delta_{4r} \leq \frac{1}{200} \), our result has a much weaker requirement on the RIP of the measurement operator.

2) **Weighted Low-Rank Matrix Factorization:** We now consider the implication of Theorem 3 in the weighted matrix factorization problem [42], where
\[ f(X) := \frac{1}{2} \| \Omega \circ (X - X^*) \|_F^2. \]
Here \( \Omega \) is an \( n \times m \) weight matrix consisting of positive elements and \( \circ \) denotes the point-wise product between two matrices. In this case, the Hessian quadrature form \( \nabla^2 f(X)[Y, Y] \) for any \( n \times m \) matrices \( X \) and \( Y \) is given by
\[ \nabla^2 f(X)[Y, Y] = \| \Omega \circ Y \|_F^2. \]
Thus \( f(X) \) satisfies the \( (2r, 4r) \)-restricted strong convexity and smoothness condition [3] with constants \( \alpha = \| \Omega \|_{\text{min}} \) and \( \beta = \| \Omega \|_{\text{max}} \) since
\[ \| \Omega \|_{\text{min}} \| Y \|_F^2 \leq \| \Omega \circ Y \|_F^2 \leq \| \Omega \|_{\text{max}} \| Y \|_F^2, \]
where \( \| \Omega \|_{\text{min}} \) and \( \| \Omega \|_{\text{max}} \) represent the smallest and largest entries in \( \Omega \), respectively. Now we consider the following weighted matrix factorization problem:
\[ \min_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} \frac{1}{2} \| \Omega \circ (UV^T - X^*) \|_F^2 + g(U, V), \] (11)
where \( g(U, V) \) is the added regularizer defined in [4]. For an arbitrary weight matrix \( \Omega \), it is proved that the weighted low-rank factorization can be NP-hard [43] and has spurious local minima. When the elements in the weight matrix \( \Omega \) are
Corollary 2. Suppose \( \Omega \) satisfies \( \frac{\|\Omega\|_2^2}{\|\Omega\|_{\min}} \leq 1.5 \). Set \( \mu \leq \frac{\|\Omega\|_2^2}{\|\Omega\|_{\min}} \). Then the objective function in (11) has no spurious local minima and satisfies the strict saddle property.

3) 1-bit Matrix Completion: Finally, we consider the problem of completing a low-rank matrix from a subset of 1-bit measurements [25]. Given \( X^\circ \in \mathbb{R}^{n \times m} \), a subset of indices \( \Omega \subset [m] \times [n] \), and a differentiable function \( q : \mathbb{R} \to [0, 1] \), we observe

\[
Y_{i,j} = \begin{cases} 
+1 & \text{with probability } q(x^\circ_{i,j}), \\
-1 & \text{with probability } 1 - q(x^\circ_{i,j}),
\end{cases}
\tag{12}
\]

for all \((i, j) \in \Omega\). Typical choices for \( q \) include the logistic regression model where \( q(x) = \frac{1}{1 + e^{-x}} \) and the probit regression model where \( q(x) = \Phi(-x/\sigma) = \Phi(x/\sigma) \). Here \( \Phi \) is the cumulative distribution function (CDF) of a mean-zero Gaussian distribution with variance \( \sigma^2 \). In [25], the authors attempt to recover \( X^\circ \) from the incomplete nonlinear measurements \( \{Y_{ij}\}_{(i,j) \in \Omega} \) by minimizing the negative log-likelihood function

\[
F_{\Omega, Y}(X) := - \sum_{(i,j) \in \Omega} \left( I(Y_{i,j} = 1) \log(q(X^\circ_{i,j})) + I(Y_{i,j} = -1) \log(1 - q(X^\circ_{i,j})) \right)
\]

which results in a maximum likelihood (ML) estimate.

We note that \( F_{\Omega, Y} \) is a convex function for both the logistic model and the probit model. The following result also establishes that \( F_{\Omega, Y} \) satisfies the restricted strong convexity and smoothness condition if we observe full 1-bit measurements, i.e., \( \Omega = [n] \times [m] \).

Lemma 1. Suppose \( \Omega = [n] \times [m] \). Let

\[
\alpha_{q, \gamma} = \min_{|x| \leq \gamma} \left( \frac{(q'(x))^2 - q(x)q''(x)}{q'^2(x)}, \frac{(q'(x))^2 + (1 - q(x))q''(x)}{(1 - q(x))^2} \right)
\]

and

\[
\beta_{q, \gamma} = \max_{|x| \leq \gamma} \left( \frac{(q'(x))^2 - q(x)q''(x)}{q'^2(x)}, \frac{(q'(x))^2 + (1 - q(x))q''(x)}{(1 - q(x))^2} \right).
\]

Then \( F_{\Omega, Y} \) satisfies the restricted strong convexity and smoothness condition:

\[
\alpha_{q, \gamma} \|G\|_F^2 \leq \|\nabla^2 F_{\Omega, Y}(X)\| (G, G) \leq \beta_{q, \gamma} \|G\|_F^2
\]

for any \( G \in \mathbb{R}^{n \times m} \) and \( \|X\|_\infty \leq \gamma \).

The proof of Lemma 1 is given in Appendix A. Now we consider the logistic regression model where \( q(x) = \frac{e^x}{1 + e^x} \).

Corollary 3. Suppose \( \Omega = [n] \times [m] \) and \( \gamma \leq 1.3 \). Consider the logistic regression model where \( q(x) = \frac{e^x}{1 + e^x} \). Then \( F_{\Omega, Y} \) satisfies the restricted strong convexity and smoothness condition with

\[
\frac{\beta_{q, \gamma}}{\alpha_{q, \gamma}} \leq 1.5.
\]

Proof of Corollary 3 Applying Lemma 1 with direct calculation gives

\[
\alpha_{q, \gamma} = q'(\gamma) = \frac{e^\gamma}{(1 + e^\gamma)^2},
\beta_{q, \gamma} = q'(0) = \frac{e^\gamma}{(1 + e^\gamma)^2} = \frac{1}{4},
\]

where \( q'(x) = \frac{e^x}{(1 + e^x)^2} \). Now if we restrict \( \|X\|_\infty \leq 1.3 \), we have

\[
\frac{\beta_{q, \gamma}}{\alpha_{q, \gamma}} = 4 \frac{e^{1.3}}{(1 + e^{1.3})^2} \leq 1.5.
\]

Under the assumption that \( X^\circ \) is low-rank, a nuclear norm constraint is utilized in [25] to force a low-rank solution. Corollary 3 implies that we can apply matrix factorization for 1-bit matrix recovery given that the elements of \( X \) are bounded. Empirical evidence (see Section V-C) supports that matrix factorization also works for 1-bit matrix completion where \( \Omega \) is only a subset of \([n] \times [m]\).

IV. PROOF OF THEOREM 3

In this section, we provide a formal proof of Theorem 3. The main argument involves showing that each critical point of \( \rho(W) \) either corresponds to the global solution of (1) or is a strict saddle whose Hessian \( \nabla^2 \rho(W) \) has a strictly negative eigenvalue. Specifically, we show that \( W \) is a strict saddle by arguing that the Hessian \( \nabla^2 \rho(W) \) has a strictly negative curvature along \( \Delta := W - W^* R \), i.e., \( \nabla^2 \rho(W)(\Delta, \Delta) \leq -\tau \|\Delta\|_F^2 \) for some \( \tau > 0 \). Here \( R \) is an \( r \times r \) orthonormal matrix such that the distance between \( W \) and \( W^* \) rotated through \( R \) is as small as possible.

A. Supporting Results

We first present some useful results. Because of the restricted strong convexity and smoothness condition [3], the following result establishes that if (1) has an optimal solution \( X^* \) with \( \text{rank}(X^*) \leq r \), then this is the unique global minimum of rank at most \( r \).

Proposition 1. Suppose \( f(X) \) satisfies the \((2r, 4r)\)-restricted strong convexity and smoothness condition [3] with positive \( \alpha \) and \( \beta \). Assume \( X^* \) is a global minimum of \( f(X) \) with \( \text{rank}(X^*) = r^* \leq r \). Then there is no other optimum of (1) that has rank less than or equal to \( r \).
Proof of Proposition 1. First note that if $X^*$ is the globally optimal solution of the original unconstrained program (1), then
\[ \nabla f(X^*) = 0. \]
Now suppose there exists another global minimum $X' \neq X^*$ with rank$(X') \leq r$ such that $f(X') = f(X^*)$. On the other hand, the second order Taylor expansion gives
\[ f(X') = f(X^*) + \langle \nabla f(X^*), X' - X^* \rangle + \frac{1}{2} \| \nabla^2 f(X)(X' - X^*, X' - X^*) \|, \]
where $\tilde{X} = tX^* + (1-t)X'$ for some $t \in [0,1]$. This Taylor expansion together with $f(X^*) = f(X')$ and $\nabla f(X^*) = 0$ gives
\[ \| \nabla^2 f(X)(X' - X^*, X' - X^*) \| = 0, \]
which contradicts (3) since $X' \neq X^*$ and both $\tilde{X}$ and $X' - X^*$ have rank at most 2r.

The $(2r, 4r)$-restricted strong convexity and smoothness assumption (3) also implies the following isometry property, whose proof is given in Appendix D.

Proposition 2. Suppose the function $f(X)$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (3) with positive $\alpha$ and $\beta$. Then for any $n \times m$ matrices $Z, G, H$ of rank at most 2r, we have
\[ \frac{2}{\alpha + \beta} \| \nabla^2 f(Z) \|_F (G, H) - \langle G, H \rangle \leq \frac{\beta - \alpha}{\beta + \alpha} \| \nabla f \|_F \| H \|_F. \]

The following result provides an upper bound on the energy of the difference $WW^T - W^*W^{*T}$ when projected onto the column space of $W$. Its proof is given in Appendix C.

Lemma 2. Suppose $f(X)$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (3). For any critical point $W$ of (5), let $P_W \in \mathbb{R}^{(m+n) \times (m+n)}$ be the orthogonal projector onto the column space of $W$. Then
\[ \| (WW^T - W^*W^{*T})P_W \|_F \leq 2 \frac{\beta - \alpha}{\beta + \alpha} \| X - X^* \|_F. \]

We remark that Lemma 2 is a variant of [18, Lemma 3.2]. While the result here requires the 4r-RIP condition of the objective function, our result depends on the $(2r, 4r)$-restricted strong convexity and smoothness condition. Our result is also slightly tighter than [18, Lemma 3.2].

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

Lemma 3. For any matrices $C, D \in \mathbb{R}^{n \times r}$ with ranks $r_1$ and $r_2$, respectively, let $R = \arg \min_{R \in \mathbb{C}^{r \times r}} \| C - DR \|_F$. Then
\[ \| CC^T - DD^T \|_F^2 \| C - DR \|_F^2 \geq \max \left\{ 2(\sqrt{2} - 1)\sigma_2^2(D), \min \{ \sigma_{r_1}^2(C), \sigma_{r_2}^2(D) \} \right\}. \]

If $C = 0$, then we have
\[ \| CC^T - DD^T \|_F^2 \| C - DR \|_F^2 \geq \frac{1}{2} \| C - DR \|_F^2. \]

We present one more useful result in the following Lemma.

Lemma 4. [39, Lemma 3] For any matrix $C, D \in \mathbb{R}^{n \times r}$, let $P_C$ be the orthogonal projector onto the range of $C$. Let $R = \arg \min_{R \in \mathbb{C}^{r \times r}} \| C - DR \|_F$. Then
\[ \| C(C - DR) \|_F^2 \leq \frac{1}{8} \| CC^T - DD^T \|_F^2 \]
\[ + (3 + \frac{1}{2(\sqrt{2} - 1)}) \| CC^T - DD^T \|_F^2. \]

Finally, we provide the gradient and Hessian expressions for $\rho(W)$. The gradient of $\rho(W)$ is given by
\[ \nabla \rho(U, V) = \nabla f(X) V + \mu U (U^T U - V^T V) \]
\[ - \nabla \rho(U, V) = \nabla f(X)^T U - \mu V (U^T U - V^T V). \]

Standard computations give the the Hessian quadrature form of $\nabla^2 \rho(W)(\Delta, \Delta)$ for any $\Delta = \begin{bmatrix} \Delta_U \\ \Delta_V \end{bmatrix}$ where $\Delta_U \in \mathbb{R}^{n \times r}, \Delta_V \in \mathbb{R}^{m \times r}$.

\[ \| \nabla^2 \rho(W)(\Delta, \Delta) \|_F = \| \nabla^2 f(X)(\Delta_U V^T + U \Delta_V^T, \Delta_U V^T + U \Delta_V^T) \|_F \]
\[ + 2 \| \nabla f(X), \Delta_U, \Delta_V \|_F^2 + \| \nabla^2 g(W)(\Delta, \Delta) \|_F. \]

where
\[ \| \nabla^2 g(W)(\Delta, \Delta) \|_F = \mu \langle \nabla \rho(W), \Delta_U \Delta_V^T \rangle \]
\[ + \mu \langle \nabla \rho(W), \Delta_U \Delta_V^T \rangle + \mu \langle \nabla \rho(W), \Delta_U \Delta_V^T \rangle. \]

B. The Formal Proof

Proof of Theorem 3. Any critical point $W$ of $\rho(W)$ satisfies $\nabla \rho(W) = 0$, i.e.,
\[ \nabla f(X) V + \mu U (U^T U - V^T V) = 0, \tag{13} \]
\[ \nabla f(X)^T U - \mu V (U^T U - V^T V) = 0. \tag{14} \]

By (14), we obtain
\[ U^T \nabla f(X) = \mu (U^T U - V^T V) V^T. \]

Multiplying (14) by $U^T$ and plugging in the expression for $U^T \nabla f(X)$ from the above equation $V^T$ gives
\[ (U^T U - V^T V) V^T V + U^T(U^T U - V^T V) = 0, \]
which further implies
\[ U^T U U^T U = V^T V V^T V. \]

Note that $U^T U$ and $V^T V$ are the principal square roots (i.e., PSD square roots) of $U^T U U^T U$ and $V^T V V^T V$.\[ \]

6
respectively. Utilizing the result that a PSD matrix has a unique principal square root \( [43] \), we obtain
\[
U^T U = V^T V. \tag{15}
\]
Thus, we can simplify (13) and (14) by
\[
\nabla \rho(U, V) = \nabla f(X) V = 0, \tag{16}
\]
\[
\nabla \rho(U, V) = \nabla f(X)^T U = 0. \tag{17}
\]
Now we turn to prove the strict saddle property and that there are no spurious local minima.

First, note that as guaranteed by Proposition [1], \( X^* \) is the unique \( n \times m \) matrix with rank at most \( r \). Also the gradient of \( f(X) \) vanishes at \( X^* \) since \( [1] \) is an unconstrained optimization problem. Denote the set of critical points of \( \rho(W) \) by
\[
C := \left\{ W \in \mathbb{R}^{(n+m)\times r} : \nabla \rho(W) = 0 \right\}.
\]
We separate \( C \) into two subsets:
\[
C_1 := C \cap \left\{ W \in \mathbb{R}^{(n+m)\times r} : UV^T = X^* \right\},
\]
\[
C_2 := C \cap \left\{ W \in \mathbb{R}^{(n+m)\times r} : UV^T \neq X^* \right\},
\]
satisfying \( C = C_1 \cup C_2 \). Since any critical point \( W \) satisfies (15), \( q(W) \) achieves its global minimum at \( W \). Also \( f(X) \) achieves its global minimum at \( X^* \). We conclude that \( W \) is the globally optimal solution of \( \rho \) for any \( W \in C_1 \). If we show that any \( W \in C_2 \) is a strict saddle, then we prove that there are no spurious local minima as well as the strict saddle property. Thus, the remaining part is to show that \( C_2 \) is the set of strict saddles.

To show that \( C_2 \) is the set of strict saddles, it is sufficient to find a direction \( \Delta \) along which the Hessian has a strictly negative curvature for each of these points. We construct \( \Delta = W - W^* R \), the difference from \( W \) to its nearest global factor \( W^* \), where
\[
R = \arg \min_{R \in O_r} \| W - W^* R \|_F.
\]
Such \( \Delta \) satisfies \( \Delta \neq 0 \) since \( X \neq X^* \) implying \( WW^T \neq W^* W^T \cdot \) Then we evaluate the Hessian bilinear form along the direction \( \Delta \):
\[
[\nabla^2 \rho(W)](\Delta, \Delta) = 2 \langle \nabla f(X), \Delta_L \Delta_L^T \rangle_{\Pi_1} + 2 \langle \nabla f(X), (\Delta_L V^T + U \Delta_L^T)^T \rangle_{\Pi_2} + \mu \langle W \Delta_L^T, \Delta L \Delta_L^T \rangle_{\Pi_3} + \mu \langle WW^T, \Delta \Delta^T \rangle_{\Pi_4}, \tag{18}
\]

The following result (which is proved in Appendix I) states that \( \Pi_1 \) is strictly negative, while the remaining terms are relatively small, though they may be nonnegative:
\[
\Pi_1 \leq -\alpha \| X - X^* \|_F^2, \quad \Pi_2 \leq \beta \| W \Delta^T \|_F^2, \quad \Pi_3 \leq \| W \Delta^T \|_F^2, \quad \Pi_4 \leq 2 \| X - X^* \|_F^2. \tag{19}
\]
Now, substituting (19) into (18) gives
\[
[\nabla^2 \rho(W)](\Delta, \Delta) = 2\Pi_1 + \Pi_2 + \mu \Pi_3 + \mu \Pi_4 \leq -2\alpha \| X - X^* \|_F^2 + \beta \mu \| W \Delta^T \|_F^2 + 2\mu \| X - X^* \|_F^2 \leq ( -2\alpha + 2\mu ) \| X - X^* \|_F^2 + ( \beta + \mu ) \| X - X^* \|_F^2 \leq -0.2\alpha \| X - X^* \|_F^2, \tag{20}
\]
where (i) utilizes Lemmas 2 and 3 (ii) utilizes the following inequality (which is proved in Appendix I)
\[
\| WW^T - W^* W^* \|_F^2 \leq 4 \| X - X^* \|_F^2, \tag{21}
\]
and (ii) holds because \( \beta \leq 1.5 \) and \( \mu \leq \frac{1}{10} \alpha \). Thus, if \( X \neq X^* \), \( [\nabla^2 \rho(X)](\Delta, \Delta) \) is always negative. This implies that \( W \) is a strict saddle.

To complete the proof, we utilize Lemma 3 to further bound the last term in (20):
\[
[\nabla^2 \rho(W)](\Delta, \Delta) \leq -0.05\alpha \| WW^T - W^* W^* \|_F^2 \leq -0.05\alpha \| \Delta \|_F^2 \times \begin{cases} 2(\sqrt{2} - 1)\sigma^2_r(W^*), & r = r^*, \\ \min\{\sigma^2_r(W), \sigma^2_r(W^*)\}, & r > r^*, \\ \sigma^2_r(W^*), & r_c = 0, \end{cases} \tag{22}
\]
where \( r^c \) is the rank of \( W \), the fist inequality utilizes (21), and the second inequality follows from Lemma 3. We complete the proof of Theorem 3 by noting that \( \sigma^2_r(W^*) = 2\sigma_r(X^*) \) for all \( \ell \in \{1, \ldots, r^*\} \) since
\[
W^* = \left[ Q_U, \Sigma^{1/2} \right] \cdot \left[ Q_V, \Sigma^{1/2} \right]^T = \left[ Q_U^{1/2}, \sqrt{2} \Sigma^{1/2} \right] \cdot \left[ Q_U^{1/2}, \sqrt{2} \Sigma^{1/2} \right]^T I
\]
is an SVD of \( W^* \), where we recall that \( X^* = Q_U \cdot \Sigma^* Q_V^* \) is an SVD of \( X^* \).

**Remark 7.** From (22), we observe that a smaller \( \mu \) yields a more negative bound on \( [\nabla^2 \rho(X)](\Delta, \Delta) \). This can be explained intuitively as follows. First note that any critical point \( W \) satisfies (15) provided \( \mu > 0 \), no matter how large or small \( \mu \) is. The Hessian information about \( g(W) \) is represented by the terms \( \Pi_3 \) and \( \Pi_4 \). We have
\[
\Pi_3 + \Pi_4 = \langle \tilde{W} \tilde{W}^T, \Delta \Delta \rangle + \langle \tilde{W}^T \Delta, \Delta \tilde{W} \rangle = \langle \tilde{W} \tilde{W}^T, \Delta \Delta \rangle + \langle \tilde{W}^T \Delta, \Delta \tilde{W} \rangle = \langle \tilde{W}^T \Delta, \tilde{W}^T \Delta + \Delta \tilde{W} \rangle \geq 0,
\]
where the last line holds since for any $r \times r$ matrix $A$,
\[
\begin{align*}
\langle A, A + A^T \rangle \\
= \frac{1}{2} \langle A + A^T, A + A^T \rangle + \frac{1}{2} \langle A - A^T, A + A^T \rangle \\
= \frac{1}{2} \left\| A + A^T \right\|_F^2 \geq 0.
\end{align*}
\]

Thus the Hessian of $\rho$ evaluated at any critical point $W$ is a PSD matrix instead of having a negative eigenvalue. In low-rank, PSD matrix optimization problems, the corresponding objective function (without any regularizer as such $g(W)$) is proved to have the strict saddle property \cite{lee2018first}. Therefore, $h(W)$ is also expected to have the strict saddle property, and so is $\rho(W)$ when $\mu$ is small, i.e., the Hessian of $g(W)$ has little influence on the Hessian of $\rho(W)$ when $\mu$ is small. Our results also indicate that when the restricted strict convexity constant $\alpha$ is not provided a priori, we can always choose a small $\mu$ to ensure the strict saddle property of $\rho(W)$ is met, and hence we are guaranteed the global convergence of a number of local search algorithms applied to $\rho(W)$.

\section*{V. Experiments}

In this section, we present a set of experiments on matrix sensing, matrix completion and 1-bit matrix completion to demonstrate the performance of gradient descent for low-rank matrix optimization. Unless noted otherwise, we denote the matrix factorization approach by NVX and use the minFunc package\footnote{Software available at https://www.cs.ubc.ca/~schmidtm/Software/minFunc.html} to perform the local search algorithms for the factorized problem.

\subsection*{A. Matrix Sensing}

We first present some experiments to illustrate the performance of gradient descent for matrix sensing problem with the factorization approach \cite{recht2010guaranteed}. In these experiments, we set $n = 50$, $m = 50$ and vary the rank $r$ from 1 to 19. We generate a rank-$r$ $n \times m$ random matrix $X^*$ by setting $X^* = UV^T$ where $U$ and $V$ are respectively $n \times r$ and $m \times r$ matrixes of normally distributed random numbers. We then obtain $p$ random measurements $y = A(X^*)$ with
\[y_i = \langle X^*, Y_i \rangle,\]
where the entries of each $n \times m$ matrix $Y_i$ are independent and identically distributed (i.i.d.) normal random variables with zero mean and variance $\frac{1}{2}$ for $i \in \{1, 2, \ldots, p\}$. For each pair of $r$ and the number of measurements, 10 Monte Carlo trials are carried out and for each trial, we claim matrix recovery to be successful if the relative reconstruction error satisfies
\[\frac{\|X^* - \hat{X}\|_F}{\|X^*\|_F} \leq 10^{-4},\]

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

\subsection*{B. Matrix Completion}

We compare the performance of the matrix factorization approach with SVP \cite{wright2013svd}, the convex approach, and singular value thresholding (SVT) \cite{ recht2010guaranteed} for matrix completion where we want to recover a low-rank matrix $X^*$ from incomplete measurements $\{X^*_{ij}\}_{(i,j) \in \Omega}$, where $\Omega \subset [n] \times [m]$. Let $P_\Omega$ denote the projection onto the index set $\Omega$. The convex

\[\min_X \|X\|_*, \quad \text{subject to } y = A(X)\]
approach (denoted by CVX) attempts to use the nuclear norm as a convex relaxation of the rankness and solves
\begin{equation}
\begin{aligned}
\min_{X} & \|X\|_*, \\
\text{subject to} & \quad P_\Omega(X) = P_\Omega(X^*).
\end{aligned}
\tag{23}
\end{equation}

To make the recovery of $X^*$ well-posed, we require $X^*$ to be incoherent such that the information in $X$ is not concentrated in a small number of entries \[^6\]. A matrix $X \in \mathbb{R}^{n \times m}$ with singular value decomposition $X = LUQ^T$ is $\mu$-incoherent if \[^{45}\] Definition 2.1
\[\max_{i,j} |L_{ij}| \leq \sqrt{\frac{\mu}{n}}, \quad \max_{i,j} |Q_{ij}| \leq \sqrt{\frac{\mu}{m}}.\]

Though $P_\Omega$ does not satisfy the $r$-RIP \[^{9}\] for all low-rank matrices $X$, it satisfies the RIP when restricted to low-rank incoherent matrices.

**Theorem 4.** \[^{45}\] Theorem 4.2] Without loss of generality, assume $n \geq m$. There exists a constant $C \geq 0$ such that for $\Omega \in [n] \times [m]$ chosen according to the Bernoulli model with density greater than $C n^2 r^2 \log n / \delta^2 m$, with probability at least $1 - e^{-n \log n}$, the RIP holds for all $\mu$-incoherent matrices $X$ of rank at most $r$.

Thus, if the iterates of local search algorithms (such as gradient descent) remain incoherent, then Theorem \[^3\] guarantees the global convergence of the matrix factorization approach with these algorithms. We note that this hypothesis is also required for SVP \[^{45}\]. Though we can add a regularizer for incoherence as in \[^{17}\], empirical evidence supports this hypothesis that the iterates in gradient descent are incoherent.

In the first set of experiments, we set $n = m = 100$ and vary the rank $r$ from 1 to 30. Similar to the setup for matrix sensing in Section \[^{5}\], we generate a rank-$r$ random matrix and randomly obtain $p$ entries, i.e., $|\Omega| = p$. Figure 2 displays the phase transition for gradient descent with a random initialization, the SVP \[^{45}\], the singular value thresholding (SVT) \[^{46}\], and the convex approach. As can been seen, the matrix factorization approach has similar phase transition to SVP, and is slightly better than SVT and the convex approach in terms of the number of measurements needed for successful recovery.

In the second set of experiments, we set $r = 5$ and $p = 3r(2n - r)$ (3 times the number of degrees of freedom within a rank-$r$ $n \times n$ matrix), and vary $n$ from 40 to 5120. We compare the time needed for the four approaches in Figure 3, our matrix factorization approach is much faster than the other methods. The time savings for the matrix factorization approach comes from avoiding performing the SVD, which is needed both for SVT and SVP in each iteration. We also observe that convex approach has highest computational complexity and is not scalable (which is the reason that we only present its time for $n$ up to 640).

**C. 1-bit Matrix Completion**

In the last set of experiments, we compare the performance of the matrix factorization approach with the convex approach\[^\star\].

\[^\star\]Software available at |\url{http://mdav.ece.gatech.edu/software/}|

\[^{25}\] for 1-bit matrix completion. We first note that to make the recovery problem well-posed, a constraint on $\|X\|_\infty$ (the entry-wise maximum of the matrix $X$) is applied in \[^{25}\] to require that the matrix is not too “spiky”. Instead of using the constraint on $\|X\|_\infty$, we add a smooth regularizer $\|X\|_F^2$ and turn to minimize the following objective function
\[f_{\Omega, X}(X) = F_{\Omega, X}(X) + \frac{n}{2} \|X\|_F^2,\]

which is also a convex function over $X$ and satisfies a similar restricted strong convexity and smoothness condition to $F_{\Omega, Y}$ in Lemma \[^1\]. Thus, we factorize $X$ into $UV^T$ and solve the following optimization problem over the $n \times r$ and $m \times r$ matrices $U$ and $V$:
\begin{equation}
\begin{aligned}
\min_{U, V} & \quad \rho_{\Omega, Y}(U, V) = f_{\Omega, Y}(UV^T).
\end{aligned}
\tag{24}
\end{equation}
To evaluate the performance of this factorization approach on 1-bit matrix completion, we generate \( n \times r \) matrices \( U^\circ \) and \( V^\circ \) with entries drawn i.i.d. from a uniform distribution on \([-\frac{1}{2}, \frac{1}{2}]\) and construct a random \( n \times n \) matrix \( X^\circ \) with rank \( r \). Similar to the setup in \([23]\), the matrix is then scaled so that \( \|X^\circ\| = 1 \). We obtain 1-bit observations \( \{Y_{i,j}\}_{(i,j) \in \Omega} \) by adding Gaussian noise of variance \( \sigma^2 \) and recording the sign of the resulting value \([12]\), where the subset of indices \( \Omega \) is chosen at random with \( E[\Omega] = p \). We compare the performance of this factorization approach on \( \Omega \) over a range of different values of \( n, p, r \) or \( \sigma \) using matrices of the error \( \|X^\circ - X^\hat{\circ}\|_F \) (where \( X^\hat{\circ} \) denotes the reconstructed matrix) and average the results over 10 draws of Monte Carlo trials. We observe that matrix factorization approach has a similar performance to the convex approach for 1-bit matrix completion \([23]\).

![Figure 4](image-url)

Figure 4. The performance in terms of the relative Frobenius norm of the error for the matrix factorization approach (denoted by NVX) with (a) varying \( n \) and \( \sigma = 0.3 \), \( r = 4 \), \( p = 0.25n^2 \); (b) varying \( p \) and \( \sigma = 0.3 \), \( n = 200 \), \( r = 4 \); (c) varying \( r \) and \( \sigma = 0.3 \), \( n = 200 \), \( p = 0.25n^2 \); (d) varying \( n \) and \( n = 200 \), \( p = 0.25n^2 \).

**VI. Conclusion**

This paper considers low-rank matrix optimization with general objective functions. By focusing on the general objective functions, we provide a unifying framework for low-rank matrix optimizations with the factorization approach. Although the resulting optimization problem is not convex, we show that the reformulated optimization function has a simple landscape: there are no spurious local minima and any critical point not being a local minimum is a strict saddle such that the Hessian evaluated at this point has a strictly negative eigenvalue. These properties guarantee that a number of iterative optimization algorithms (such as gradient descent and the trust region method) will converge to the global optimum from a random initialization.

**APPENDIX A**

**Proof of Lemma 1**

**Proof of Lemma 1** We compute the partial derivative of \( F_{i,j,Y} \) in terms of \( X_{i,j} \) as

\[
\frac{\partial F_{i,j,Y}}{\partial X_{i,j}} = q'(X_{i,j}) - q(X_{i,j})
\]

which implies

\[
\frac{\partial^2 F_{i,j,Y}}{\partial X_{i,j} \partial Y_{i,j}} = 0
\]

for all \((k, \ell) \neq (i, j)\). Thus, the bilinear form for the Hessian of \( \nabla^2 F_{i,j,Y}(X) \) can be computed as

\[
[\nabla^2 F_{i,j,Y}(X)](G, G) = \sum_i \sum_j \frac{\partial^2 F_{i,j,Y}}{\partial X_{i,j} \partial X_{i,j}} G_{i,j}^2
\]

for any \( G \in \mathbb{R}^{n \times m} \). Now since by assumption \( \|X\|_\infty \leq \gamma \), we have

\[
\alpha_{\gamma, \gamma} \|G\|_F^2 \leq [\nabla^2 F_{i,j,Y}(X)](G, G) \leq \beta_{\gamma, \gamma} \|G\|_F^2.
\]

**APPENDIX B**

**Proof of Proposition 2**

**Proof of Proposition 2** This proof follows similar steps to the proof of \([47]\) Lemma 2.1. First note that the bilinear form \( \nabla^2 f(Z)(G, H) = \sum_{i,j,k,l} \partial^2 f(Z)_{i,j,k,l} G_{i,j} H_{k,l} \) implies \( \nabla^2 f(Z)(G, H) \) is invariant under all scalings for both \( G \) and \( H \), i.e.,

\[
[\nabla^2 f(Z)](aG, bH) = ab[\nabla^2 f(Z)](G, H)
\]

for any \( a, b \in \mathbb{R} \). If either \( G \) or \( H \) is zero, \([3]\) holds since both sides are 0.

Now suppose both \( G \) and \( H \) are nonzero. By the scaling invariance property of both sides in \([3]\), we assume \( \|G\|_F = \|H\|_F = 1 \) without loss of generality. Note that the \((2r, 4r)\)-restricted strong convexity and smoothness condition \([3]\) implies

\[
\alpha \|G \pm H\|_F^2 \leq [\nabla^2 f(X)](G \pm H, G \pm H) \leq \beta \|G \pm H\|_F^2.
\]
Thus we have
\[ -\frac{\beta - \alpha}{2} \left( \|G\|_F^2 + \|H\|_F^2 \right) \]
\[ \leq 2 \left\{ \nabla^2 f(Z) \right\} (G, H) - (\alpha + \beta) \langle G, H \rangle \]
\[ \leq \frac{\beta - \alpha}{2} \left( \|G\|_F^2 + \|H\|_F^2 \right), \]
which further implies
\[ 2 \left\{ \nabla^2 f(Z) \right\} (G, H) - (\alpha + \beta) \langle G, H \rangle \leq \beta - \alpha = (\beta - \alpha) \|G\|_F \|H\|_F. \]

\[ \text{Appendix C} \]

**Proof of Lemma 2**

First recall the notation \( X = UV^T \), \( X^* = U^*V^* \), and
\[ W = \begin{bmatrix} U & V \end{bmatrix}, \quad \hat{W} = \begin{bmatrix} U & -V \end{bmatrix}, \quad W^* = \begin{bmatrix} U^* & V^* \end{bmatrix}, \quad \hat{W}^* = \begin{bmatrix} U^* & -V^* \end{bmatrix}. \]

It follows from (16) and (17) that any critical point \( W \) satisfies
\[ \begin{bmatrix} 0 & \nabla f(X)^T \nabla f(X) \end{bmatrix} W = 0, \]

which gives
\[ 0 = \begin{bmatrix} 0 & \nabla f(X)^T \nabla f(X) \end{bmatrix}, ZW^T \]
\[ = \begin{bmatrix} 0 & \nabla f(X)^T \nabla f(X) - \nabla f(X^*) \end{bmatrix}, ZW^T \]
\[ = \left( \nabla f(X) - \nabla f(X^*) - \frac{\alpha + \beta}{2} (X - X^*), ZU V^T + UZ V^T \right) \]
\[ + \frac{\alpha + \beta}{2} \left( X - X^*, ZU V^T + UZ V^T \right) \gamma_2 \]

(25)

for any \( Z = \begin{bmatrix} Z_U & Z_V \end{bmatrix} \in \mathbb{R}^{(n+m)\times r} \). Here the second line utilizes the fact \( \nabla f(X^*) = 0 \). We bound \( \gamma_1 \) by first using integral form of the mean value theorem for \( \nabla f(X) \):
\[ \gamma_1 = \int_0^1 \left[ \nabla^2 f(tX + (1-t)X^*) \right] (X - X^*, ZU V^T + UZ V^T)dt \]
\[ - \frac{\alpha + \beta}{2} \left( X - X^*, ZU V^T + UZ V^T \right). \]

Noting that all the three matrices \( tX + (1-t)X^* \), \( X - X^* \)

and \( ZU V^T + UZ V^T \) have rank at most \( 2r \), it follows from Proposition 2 that
\[ |\gamma_1| \leq \frac{\beta - \alpha}{2} \|X - X^*\|_F \left\| ZU V^T + UZ V^T \right\|_F, \]
which when plugged into (25) gives
\[ \frac{\alpha + \beta}{2} \gamma_2 = -\gamma_1 \]
\[ \leq \frac{\beta - \alpha}{2} \|X - X^*\|_F \left\| ZU V^T + UZ V^T \right\|_F. \]

(26)

Now let \( Z = (WW^T - W^*W^*)W^T \), which gives \( ZW^T = (WW^T - W^*W^*)P_W \). Here \( \dagger \) denotes the pseudoinverse of a matrix and \( P_W \) is the orthogonal projector onto the range of \( W \). Utilizing the fact \( \hat{W}^TW = 0 \) from (7), we further connect the left hand side of (26) with
\[ \left\| (WW^T - W^*W^*)P_W \right\|_F^2 \]
\[ = \frac{\alpha + \beta}{2} \gamma_2 = \frac{\alpha + \beta}{2} \left( \hat{W}^TW, ZW^T \right) \]
\[ = \frac{\alpha + \beta}{4} \left( WW^T - W^*W^*P_W \right) + \frac{\alpha + \beta}{4} \left( \hat{W}^*P_W, (WW^T - W^*W^*)P_W \right) \]
\[ \geq \frac{\alpha + \beta}{4} \left( WW^T - W^*W^*P_W, (WW^T - W^*W^*)P_W \right) \]
\[ = \frac{\alpha + \beta}{4} \left| (WW^T - W^*W^*)P_W \right|_F^2, \]

(27)

where the inequality follows because
\[ \left( \hat{W}^*P_W, (WW^T - W^*W^*)P_W \right) = 0 \]

and \( \left( \hat{W}^*P_W, (WW^T - W^*W^*)P_W \right) \geq 0 \) since it is the inner product between two PSD matrices.

On the other hand, we give an upper bound on the right hand side of (26):
\[ \|X - X^*\|_F \left\| ZU V^T + UZ V^T \right\|_F \]
\[ \leq \|X - X^*\|_F \sqrt{2 \left\| ZU V^T \right\|_F^2 + 2 \left\| UZ V^T \right\|_F^2} \]
\[ \leq \|X - X^*\|_F \left\| (WW^T - W^*W^*)P_W \right\|_F, \]

where the last line follows because \( \left\| ZU V^T \right\|_F^2 + \left\| ZV U^T \right\|_F^2 = \|U^T U = V^TV\|_F^2 \), implying \( 2 \left\| ZU V^T \right\|_F^2 + 2 \left\| UZ V^T \right\|_F^2 = \|ZW^T \|_F^2 \).

This together with (26) and (27) completes the proof.

\[ \text{Appendix D} \]

**Proof of Lemma 3**

When \( C \neq 0 \), the proof follows directly from the following results.

**Lemma 5.** [59, Lemma 2] For any matrix \( C, D \in \mathbb{R}^{n\times r} \) with rank \( r_1 \) and \( r_2 \), respectively, let \( R = \arg \min_{R \in \mathcal{O}} \|C - DR\|_F \). Then
\[ \|CC^T - DD^T\|_F \geq \min \{ \sigma_{r_1}(C), \sigma_{r_2}(D) \} \cdot \|C - DR\|_F. \]
Lemma 6. [54] Lemma 5.4] For any matrix $C, D \in \mathbb{R}^{n \times r}$ with $\text{rank}(D) = r$, let $R = \arg \min_{\mathcal{R}_{C_0}} \|C - DR\|_F$. Then
$$\left\|CC^T - DD^T\right\|_F^2 \geq 2(\sqrt{2} - 1)\sigma_2^2(D) \|C - DR\|_F^2.$$
If $C = 0$, then we have
$$\left\|CC^T - DD^T\right\|_F^2 = \left\|DD^T\right\|_F^2 = \sum_{i=1}^{r_2} \sigma_i^2(D) \geq \sigma_2^2(D) \sum_{i=1}^{r_2} \sigma_i^2(D) = \sigma_2^2(D) \|C - DR\|_F^2.$$ 

APPENDIX E

Proof of (19)

Proof of (19): We prove the upper bounds for the four terms as follows. Bounding term $\Pi_1$: Utilizing the fact that $\Delta_U = U - U^* R$ and $\Delta_V = V - V^* R$, we have
$$\Pi_1 = \left\langle \nabla f(X), \Delta U \Delta_V^T \right\rangle = \left\langle \nabla f(X), (U - U^* R)(V - V^* R)^T \right\rangle = \left\langle \nabla f(X), X + X^* - U^* R V^T - U R^T V^* T \right\rangle$$
$$= -\left\langle \nabla f(X), X - X^* \right\rangle$$
$$\leq -\alpha \|X - X^*\|_F^2,$$
where (i) follows from (16) and (17), (ii) utilizes $\nabla f(X^*) = 0$, and (iii) follows by using the $(2r, 4r)$-restricted strict convexity property (3):
$$\langle \nabla f(X) - \nabla f(X^*), X - X^* \rangle$$
$$= \int_0^1 \left[ \nabla^2 f(tX + (1-t)X^*) \right] (X - X^* , X - X^*) dt$$
$$\geq \int_0^1 \alpha (X - X^*, X - X^*) dt$$
$$= \alpha \|X - X^*\|_F^2,$$
where the first line follows from the integral form of the mean value theorem for vector-valued functions, and the second line uses the fact that both $tX + (1-t)X^*$ and $X - X^*$ have rank at most $2r$, and the $(2r, 4r)$-restricted strong convexity of the Hessian $\nabla^2 f(\cdot)$.

Bounding term $\Pi_2$: By the smoothness condition (3), we have
$$\Pi_2 = \left\langle \nabla^2 f(X) \left( \Delta U V^T + U \Delta V^T + U \Delta U^T \right) \right\rangle$$
$$\leq \beta \left\| \Delta U V^T + U \Delta V^T + U \Delta U^T \right\|_F^2$$
$$\leq 2\beta \left( \left\| \Delta U V^T \right\|_F^2 + \left\| U \Delta V^T \right\|_F^2 \right)$$
$$= \beta \left\| W \Delta^T \right\|_F^2,$$
where the last line holds because $\left\| DU^T \right\|_F^2 = \left\| DV^T \right\|_F^2$ for any $D \in \mathbb{R}^{P \times r}$ with arbitrary $p \geq 1$ since any critical point $W$ satisfies $U^T U = V^T V$.

Bounding term $\Pi_3$: 
$$\Pi_3 = \left\langle \Delta U^T \Delta V \Delta_V^T \right\rangle + \left\langle V \Delta_V^T \Delta_V V^T \right\rangle - 2 \left\langle U \Delta_U^T \Delta_U V^T \right\rangle$$
$$\leq \left\| U \Delta_U^T \right\|_F^2 + \left\| V \Delta_V^T \right\|_F^2 + \left\| U \Delta_U^T \right\|_F^2 + \left\| V \Delta_U^T \right\|_F^2$$
$$= \left\| W \Delta^T \right\|_F^2.$$

Bounding term $\Pi_4$: 
$$\Pi_4 = \left\langle \nabla W^T, (W - W^* R) (W - W^* R)^T \right\rangle$$
$$= \left\langle \nabla W^T, W W^T - W^* W^* \right\rangle$$
$$\leq \left\langle \nabla W^T, W W^T - W^* W^* \right\rangle + \left\langle V \Delta V^T, W W^T - W^* W^* \right\rangle$$
$$\leq 2 \|X - X^*\|_F^2,$$
where (i) holds because $\nabla W^T W = 0$, and (ii) follows because $W^* W^* = 0$ and $\langle \nabla W^T, W W^T = 0 \rangle \geq 0$ since it is the inner product between two PSD matrices.

APPENDIX F

Proof of (21)

Proof of (21): To show (21), expanding the left hand side of (21), it is equivalent to show
$$\left\| UU^T - U^* U^* \right\|_F^2 + \left\| V V^T - V^* V^* \right\|_F^2 \leq 2 \|X - X^*\|_F^2.$$ 

Expanding both sides of the above equation and utilizing the fact $U^T U = V^T V$ and $U^* U^* = V^* V^*$, the remaining step is to show
$$\text{trace} \left( UU^T U^* U^* + V V^T V^* V^* \right)$$
$$\geq 2 \text{trace} \left( U V^T V^* U^* \right).$$ 

Thus, we obtain (21) by noting that the above equation is equivalent to
$$\text{trace} \left( (U^* U - V^* V)^2 \right) \geq 0.$$ 


REFERENCES

[1] S. Aaronson, “The learnability of quantum states,” in Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 463, pp. 3089–3114, The Royal Society, 2007.

[2] S. T. 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, vol. 14, no. 9, p. 095022, 2012.

[3] N. Srebro, J. Rennie, and T. S. Jaakkola, “Maximum-margin matrix factorization,” in Advances in Neural Information Processing Systems, pp. 1329–1336, 2004.

[4] D. DeCoste, “Collaborative prediction using ensembles of maximum margin matrix factorizations,” in Proceedings of the 23rd International Conference on Machine Learning, pp. 249–256, ACM, 2006.

[5] P. Biswas and Y. Ye, “Semidefinite programming for ad hoc wireless sensor network localization,” in Proceedings of the 3rd international symposium on Information processing in sensor networks, pp. 46–54, ACM, 2004.

[6] G. Tang and A. Nehorai, “Lower bounds on the mean-squared error of low-rank matrix reconstruction,” IEEE Transactions on Signal Processing, vol. 59, no. 10, pp. 4559–4571, 2011.

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

[8] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” Foundations of Computational Mathematics, vol. 9, no. 6, pp. 717–772, 2009.

[9] G. Liu and P. Li, “Low-rank matrix completion in the presence of high coherence,” IEEE Transactions on Signal Processing, vol. 64, no. 21, pp. 5623–5633, 2016.

[10] E. J. Candès and Y. Plan, “Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements,” IEEE Transactions on Information Theory, vol. 57, no. 4, pp. 2342–2359, 2011.

[11] Z. Harchaoui, M. Douze, M. Paulin, M. Dudík, and J. Malick, “Large-scale image classification with trace-norm regularization,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3386–3393, IEEE, 2012.

[12] M. A. Davenport and J. Romberg, “An overview of low-rank matrix recovery from incomplete observations,” IEEE Journal of Selected Topics in Signal Processing, vol. 10, no. 4, pp. 608–622, 2016.

[13] K. Mohan and M. Fazel, “Reweighted nuclear norm minimization with application to system identification,” in Proceedings of the 2010 American Control Conference, pp. 2939–2939, IEEE, 2010.

[14] S. Burer and R. D. Monteiro, “A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization,” Mathematical Programming, vol. 95, no. 2, pp. 329–357, 2003.

[15] S. Burer and R. D. Monteiro, “Local minima and convergence in low-rank semidefinite programming,” Mathematical Programming, vol. 103, no. 3, pp. 427–444, 2005.

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

[17] R. Ge, J. D. Lee, and T. Ma, “Matrix completion has no spurious local minimum,” arXiv preprint arXiv:1605.07272, 2016.

[18] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, “Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach,” arXiv preprint arXiv:1609.03240, 2016.

[19] L. Wang, X. Zhang, and Q. Gu, “A unified computational and statistical framework for nonconvex low-rank matrix estimation,” arXiv preprint arXiv:1610.05275, 2016.

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

[21] 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, pp. 797–842, 2015.

[22] M. Udell, C. Horn, R. Zadeh, and S. Boyd, “Generalized low rank models,” arXiv preprint arXiv:1410.0342, 2014.

[23] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?,” Journal of the ACM (JACM), vol. 58, no. 3, p. 11, 2011.

[24] B. Recht, M. Fazel, and P. A. Parrilo, “Guaranteed minimum-rank matrix completion using alternating minimization,” in Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing, pp. 665–674, ACM, 2013.

[25] E. J. Candès, X. Li, and M. Soltanolkotabi, “Phase retrieval via Wirtinger flow: Theory and algorithms,” IEEE Transactions on Information Theory, vol. 61, no. 4, pp. 1985–2007, 2015.

[26] S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht, “Low-rank solutions of linear matrix equations via Procrustes flow,” arXiv preprint arXiv:1507.03566, 2015.

[27] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi, “Dropping convexity for faster semi-definite optimization,” arXiv preprint, 2015.

[28] J. Sun, Q. Qu, and J. Wright, “Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method,” arXiv preprint arXiv:1511.04777, 2015.

[29] Q. Li, Z. Zhu, and G. Tang, “Geometry of factored nuclear norm regularization,” arXiv preprint arXiv:1704.01265, 2017.

[30] Y. Li, Y. Sun, and Y. Chi, “Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements,” IEEE Transactions on Signal Processing, vol. 65, no. 2, pp. 397–408, 2017.

[31] L. Xu and M. Davenport, “Dynamic matrix recovery from incomplete observations under an exact low-rank constraint,” in Advances in Neural Information Processing Systems, pp. 3585–3593, 2016.

[32] X. Li, Z. Wang, J. Lu, R. Arora, J. Haupt, H. Liu, and T. Zhao, “Symmetry, saddle points, and global geometry of nonconvex matrix factorization,” arXiv preprint arXiv:1612.09296, 2016.

[33] Q. Li and G. Tang, “The nonconvex geometry of low-rank matrix optimizations with general objective functions,” arXiv preprint arXiv:1611.03060, 2016.

[34] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, “The global optimization geometry of nonsymmetric matrix factorization and sensing,” arXiv preprint arXiv:1703.01256, 2017.

[35] A. R. Conn, N. I. Gould, and P. L. Toint, Trust region methods. SIAM, 2000.

[36] N. Srebro, T. Jaakkola, et al., “Weighted low-rank approximations,” in International Conference on Machine Learning, vol. 3, pp. 720–727, 2003.

[37] N. Gillis and F. Glineur, “Low-rank matrix approximation with weights or missing data is NP-hard,” SIAM Journal on Matrix Analysis and Applications, vol. 32, no. 4, pp. 1149–1165, 2011.
[44] C. R. Johnson, K. Okubo, and R. Reams, “Uniqueness of matrix square roots and an application,” *Linear Algebra and Its Applications*, vol. 323, no. 1, pp. 51–60, 2001.

[45] P. Jain, R. Meka, and I. S. Dhillon, “Guaranteed rank minimization via singular value projection,” in *Advances in Neural Information Processing Systems*, pp. 937–945, 2010.

[46] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[47] E. J. Candès, “The restricted isometry property and its implications for compressed sensing,” *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.