THE NONCONVEX GEOMETRY OF LOW-RANK MATRIX OPTIMIZATIONS WITH GENERAL OBJECTIVE FUNCTIONS

Qiuwei Li and Gongguo Tang

Department of Electrical Engineering, Colorado School of Mines, Golden, CO USA

ABSTRACT

This work considers the minimization of a general convex function $f(X)$ over the cone of positive semi-definite matrices whose optimal solution $X^\star$ is of low-rank. Standard first-order convex solvers require performing an eigenvalue decomposition in each iteration, severely limiting their scalability. A natural nonconvex reformulation of the problem factors the variable X into the product of a rectangular matrix with fewer columns and its transpose. For a special class of matrix sensing and completion problems with quadratic objective functions, local search algorithms applied to the factored problem have been shown to be much more efficient and, in spite of being nonconvex, to converge to the global optimum. The purpose of this work is to extend this line of study to general convex objective functions $f(X)$ and investigate the geometry of the resulting factored formulations. Specifically, we prove that when $f(X)$ satisfies the restricted well-conditioned assumption, each critical point of the factored problem either corresponds to the optimal solution $X^\star$ or a strict saddle where the Hessian matrix has a strictly negative eigenvalue. Such a geometric structure of the factored formulation ensures that many local search algorithms can converge to the global optimum with random initializations.

Index Terms— Burer-Monteiro factorization, low-rank matrix optimization, nonconvex optimization, strict saddle property

1. INTRODUCTION

Consider a general semi-definite program (SDP) where a convex objective function $f(X)$ is minimized over the cone of positive semi-definite (PSD) matrices:

$$\underset{X\in\mathbb{R}^{n\times n}}{\text{minimize}}\, f(X) \text{ subject to } X \succeq 0. \tag{1}$$

For this problem, even fast first-order methods, such as the projected gradient descent algorithm [2], require performing an expensive eigenvalue decomposition in each iteration. These expensive operations form the major computational bottleneck of the algorithms and prevent them from scaling to scenarios with millions of variables, a typical situation in a diverse of applications, including quantum state tomography [3], user preferences prediction [4], and pairwise distances estimation in sensor localization [5].

When the SDP (1) admits a low-rank solution $X^\star$, in their pioneer work [6], Burer and Monteiro proposed to factorize the variable $X = UU^T$, where $U \in \mathbb{R}^{n\times r}$ with $r \ll n$, and solved a factored nonconvex problem

$$\underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}}\, g(U), \text{ where } g(U) := f(UU^T). \tag{2}$$

There, they dealt with standard SDPs with a linear objective function and several linear constraints, and argued that when the factorization $X = UU^T$ is overparameterized, *i.e.*, $r > r^\star := \text{rank}(X^\star)$, any local minimum of (2) corresponds to the solution $X^\star$, provided some regularity conditions are satisfied. Unfortunately, these regularity conditions are generally hard to verify for specific SDPs arising in applications. Our work differs in that the convex objective function $f(X)$ is generally not linear and there are no additional linear constraints.

The past few years have seen renewed interest in the Burer-Monteiro factorization for solving low-rank matrix recovery inverse problems. With technical innovations in analyzing the nonconvex landscape of the factored objective function, several recent works have shown that with exact parameterization (*i.e.*, $r = r^\star$) the factored objective function $g(U)$ in has no spurious local minima or degenerate saddle points [7–12]. An important implication is that local search algorithms, such as gradient descent and its variants, are able to converge to the global optimum with even random initialization [13].

We generalize this line of work by assuming a general objective function $f(X)$ in the optimization (1). Viewing the factored problem (2) as a way to solve the convex optimization (1) to the global optimum, frees us from rederiving the statistical performances of the factored optimization (2). Instead, its performance inherits from that of the convex optimization (1), whose performance can be developed using a suite of powerful convex analysis techniques accumulated from several decades of research. As a specific example, the optimal sampling complexity [14] and minimax denoising rate [15] need not to be rederived once one knows the equivalence between the convex and the factored formulations.

Full version appears as [1]. This work was supported by NSF grant CCF-1464205. Email: {qiuli, gtang}@mines.edu.

2. MAIN THEOREM

Before presenting our main result, we provide several necessary definitions. We call a vector $\boldsymbol{x}$ a *critical point* of some differentiable function $f(\cdot)$ if the gradient $\nabla f(\boldsymbol{x}) = \mathbf{0}$. When $f(\cdot)$ is twice continuously differentiable, a critical point $\boldsymbol{x}$ is called a *strict saddle* or *riddable saddle* [16] if the Hessian has a strictly negative eigenvalue, i.e., $\lambda_{\min}(\nabla^2 f(\boldsymbol{x})) < 0$. A twice continuously differentiable function satisfies the *strict saddle property* if every critical point is either a local minimum or is a strict saddle [7].

Heuristically, the strict saddle property describes a geometric structure of the landscape: if a critical point is not a local minimum, then it is a strict saddle, which implies the Hessian matrix at this point has a strictly negative eigenvalue. Hence, we can continue to decrease the function value at this point along the negative-curvature direction.

Theorem 1 (Local convergence [13, 17, 18]). *The strict saddle property allows many local search algorithms to escape all the saddle points and converge to a local minimum.*

Our governing assumption on the objective function $f(X)$ is the $(2r, 4r)$-restricted well-conditioned assumption:

$$m \leq [\nabla^2 f(X)](D, D)/\|D\|_F^2 \leq M \text{ with } \frac{M}{n} \leq 1.5 \qquad (3)$$

for any D of $\operatorname{rank}(D) \leq 4r$ and any PSD matrix X with $\operatorname{rank}(X) \leq 2r$. Here, $[\nabla^2 f(X)](D, D)$ is the directional curvature along D, defined as $\sum_{i,j,l,k} \frac{\partial^2 f(X)}{\partial X_{ij} \partial X_{lk}} D_{ij} D_{lk}$. This restricted well-conditioned assumption (3) is standard in matrix inverse problem [19, 20]. We show that if the original objective function $f(X)$ is $(2r, 4r)$-restricted well conditioned, then each critical point of the factored objective function $g(U)$ either corresponds to the low-rank global solution of the original convex program or is a strict saddle where the Hessian $\nabla^2 g(U)$ has a strictly negative eigenvalue. This implies the factored objective function $g(U)$ satisfies the strict saddle property.

Theorem 2 (Global landscape). *Suppose the function $f(X)$ in (1) is twice continuously differentiable and restricted well-conditioned (3). Assume $X^\star$ is an optimal solution of the minimization (1) with $\operatorname{rank}(X^\star) = r^\star$. Set $r \geq r^\star$ in (2). Let U be any critical point of $g(U)$ satisfying $\nabla g(U) = \mathbf{0}$. Then U either corresponds to a square-root factor of $X^\star$,* i.e.,

$$X^\star = UU^T; \qquad (4)$$

or is a strict saddle of the factored problem (2):

$$\lambda_{\min}(\nabla^2 g(U)) \leq \begin{cases} -0.24 m\tau & \text{when } r > r^\star \\ -0.19 m\rho(X^\star) & \text{when } r = r^\star \\ -0.24 m\rho(X^\star) & \text{when } U = \mathbf{0} \end{cases} \qquad (5)$$

with $\tau := \min\{\rho(U)^2, \rho(X^\star)\}$ and $\rho(W)$ denoting the smallest nonzero singular value.

Remarks. First, the matrix D is the direction from the saddle point U to its closest globally optimal factor $U^\star R$ of the same size as U. Second, our result covers both over-parameterization where $r > r^\star$ and exact parameterization where $r = r^\star$. Third, this strict saddle property ensures that many iterative algorithms, for example, stochastic gradient descent [17], trust-region method [18], and gradient descent with sufficiently small stepsize [13], all converge to a square-root factor of $X^\star$, even with random initialization.

3. APPLICATIONS

Our main result only relies on the restricted well-conditioned property. Therefore, in addition to the traditional low-rank matrix recovery problems with a quadratic loss function, it is also applicable to a lot of other low-rank matrix optimization problems with possibly non-quadratic loss functions. We compiled the following list of applications that are covered by our theory.

Weighted PCA Problem. Formally, in the weighted-PCA problem, given a pointwisely-weighted observation of a PSD matrix X, *i.e.*, $Y = W \odot X$ where $\odot$ is the Hadamard product or its perturbed version with W being the sensing matrix, one aims to recover the principle component U by minimizing the nonconvex objective function $g(U) = \|Y - W \odot (UU^T)\|_F^2$. The weighted-PCA problem has no known analytic solution and it is shown to be NP-hard [21]. Fortunately, by defining $f(X) = \|Y - W \odot X\|_F^2$, we can compute its directional curvature as $[\nabla^2 f(X)](D, D) = \|W \odot D\|_F^2$. Hence, as long as the weights have a smaller dynamic range: $\frac{\max W_{ij}^2}{\min W_{ij}^2} \leq 1.5$, it is guaranteed to recover U through local search algorithms.

Symmetric Robust PCA. In the symmetric variant of robust PCA, the observed matrix $Y = X + S$ with S being sparse and X being PSD. Traditionally, we recover X by minimizing $||Y - X||_1 = \sum_{ij} |Y_{ij} - S_{ij}|$ subject to a PSD constraint. However, this formulation doesn't fit into our framework naively due to the non-smoothness of the ℓ_1 norm. An interesting bypass would be solving X by minimizing $\sum_{ij} h_a(Y_{ij} - S_{ij})$ where $h_a(.)$ is chosen to be a convex smooth approximation to the absolute value function. A possible choice is $h_a(x) = a \log((\exp(x/a) + \exp(-x/a))/2)$, which is shown to be strictly convex and smooth in [22, Lemma A.1].

1-Bit Matrix Recovery. Given quantized measurements: $y_j = \operatorname{bit}(A_j \bullet X^\star)$ where $\bullet$ denotes the inner product and $\operatorname{bit}(x)$ outputs 0 or 1 in a probabilistic manner, we attempt to recover $X^\star \in \mathbb{R}^{n \times m}$ by minimizing $f(X) = -\sum_j (y_j \log(\sigma(A_j \bullet X)) + (1 - y_j) \log(1 - \sigma(A_j \bullet X)))$, where $\sigma(x) = \frac{e^x}{1+e^x}$ is the logistic regression function [23]. Moreover, the Hessian quadratic form of $f(X)$ is $[\nabla^2 f(X)](D, D)$

$= \sum_j \sigma'(A_j \bullet X)(A_j \bullet D)^2$. Then as long as the number of Gaussian measurements is comparable to the degrees of freedom in $X^\star$ and the elements of A_j concentrate in a certain range (which ensures $\sigma'(\cdot)$ has small dynamic range), 1-bit matrix recovery fits into our framework. In particular, when we get full measurements, *i.e.,* we have $y_{ij} = \mathrm{bit}(E_{ij} \bullet X^\star), \forall i \in [n], j \in [m]$ with E_{ij} being the canonical basic in $\mathbb{R}^{n\times m}$. For this special case, the Hessian quadratic form is given as $[\nabla^2 f(X)](D,D) = \sum_{ij} \sigma'(X_{ij}) D_{ij}^2$, so we have $\min_{ij} \sigma'(X_{ij})\|D\|_F^2 \le [\nabla^2 f(X)](D,D) \le \max_{ij} \sigma'(X_{ij})\|D\|_F^2$. Direct computations give that the derivative $\sigma'(x)$ has a smaller dynamic range for small x: $\frac{\max_{x\in[-1,1]}\sigma'(x)}{\min_{x\in[-1,1]}\sigma'(x)} \le 1.27$. Therefore, when we restrict the values $|X_{ij}|$ to be small: $\max_{ij}|X_{ij}| \le 1$, 1-bit matrix recovery fits into our framework. Such an constraint on $\max_{ij}|X_{ij}|$ is also required in [23] to obtain an accurate estimate of $X^\star$.

Low-rank Matrix Recovery with Non-Gaussian Noise. Consider a matrix sensing or PCA problem. When the noise is from the normal distribution, the according maximum likelihood estimation (MLE) is the minimizer of a squared loss function. However, in practice, the noise in data is often from other distributions. In this case the resulting MLE is obtained by minimizing the negative log-likelihood function, which is not the square loss. Such a noise-adaptive estimator is more effective than square-loss minimization. To have a strongly convex and smooth objective function so that our theory can apply, the noise distribution should be log-strongly-concave. The reference [24] contains many examples of such distributions.

4. PROBLEM FORMULATIONS AND PRELIMINARIES

This paper considers the problem (1) of minimizing a convex function $f(X)$ over the PSD cone. Let $X^\star$ be an optimal solution of (1) of rank $r^\star$. When the PSD variable X is reparameterized as

$$X = \phi(U) := UU^T,$$

where $U \in \mathbb{R}^{n\times r}$ with $r \ge r^\star$ is a rectangular, matrix square root of X, the convex program is transformed into the factored problem (2) whose objective function is $g(U) := f(\phi(U))$. Inspired by the lifting technique in constructing SDP relaxations, we refer to the variable X as the lifted variable, and the variable U as the factored variable. Similar naming conventions apply to the optimization problems, their domains, and objective functions.

The nonlinear parametrization $X = \phi(U)$ makes $g(U)$ a nonconvex function and introduces additional critical points (*i.e.*, those U with $\nabla g(U) = \mathbf{0}$ that are not global optima of the factored optimization (2)). Our goal is to show that each critical points either corresponds to $X^\star$ or is a strict saddle where the Hessian has a strictly negative eigenvalue.

4.1. Metrics in the Lifted and Factored Spaces

Since for any U, $\phi(U) = \phi(UR)$ where $R \in \mathbb{O}_r$ with $\mathbb{O}_r$ being all $r \times r$ orthonormal matrices, the domain of the factored objective function $g(U)$ is stratified into equivalent classes and can be viewed as a quotient manifold. The matrices in each of these equivalent classes differ by an orthogonal transformation (not necessarily unique when the rank of U is less than r). One implication is that, when working in the factored space, we should consider all factorizations of $X^\star$:

$$\mathcal{A}^\star = \{U^\star \in \mathbb{R}^{n\times r} : X^\star = \phi(U^\star)\}.$$

A second implication is that when considering the distance between two points U_1 and U_2, one should use the distance between their corresponding equivalent classes:

$$\mathrm{d}(U_1, U_2) = \min_{R\in\mathbb{O}_r} \|U_1 - U_2 R\|_F. \tag{6}$$

For any two matrices $U_1, U_2 \in \mathbb{R}^{n\times r}$, the following lemma relates the distance $\|\phi(U_1) - \phi(U_2)\|_F$ in the lifted space to the distance $\mathrm{d}(U_1, U_2)$ in the factored space, with the proof deferred to [1]:

Lemma 1. *Assume that $U_1, U_2 \in \mathbb{R}^{n\times r}$. Then*

$$\|\phi(U_1) - \phi(U_2)\|_F \ge \min\{\rho(U_1), \rho(U_2)\}\mathrm{d}(U_1, U_2).$$

5. PROOF: CONNECTING THE OPTIMALITY CONDITIONS

The proof is inspired by connecting the optimality conditions for the two programs (1) and (2). First of all, as a constrained convex optimization, all critical points of (1) are global optima and are characterized by the necessary and sufficient KKT condition [2]:

$$\nabla f(X^\star) \succeq 0, \nabla f(X^\star)X^\star = \mathbf{0}, X^\star \succeq 0. \tag{7}$$

The factored optimization (2) is unconstrained, whose critical points are specified by the zero gradient condition:

$$\nabla g(U) = 2\nabla f(\phi(U))U = \mathbf{0}. \tag{8}$$

To classify the critical points, we compute the Hessian bilinear form $[\nabla^2 g(U)](D,D)$ as:

$$\begin{aligned}[\nabla^2 g(U)](D,D) &= 2\langle \nabla f(\phi(U)), DD^T\rangle \\ &+ [\nabla^2 f(\phi(U))](DU^T + UD^T, DU^T + UD^T).\end{aligned} \tag{9}$$

For any critical point U of $g(U)$, the corresponding lifted variable $\phi(U) = UU^T$ is PSD and satisfies $\nabla f(\phi(U))\phi(U) = \mathbf{0}$. On one hand, if $\phi(U)$ further satisfies $\nabla f(\phi(U)) \succeq 0$, then in view of the KKT conditions (7) and noting $\mathrm{rank}(\phi(U)) \le r$, we must have $\phi(U) = X^\star$, the global optimum of (1). On the other hand, if $\phi(U) \ne X^\star$, implying $\nabla f(\phi(U)) \not\succeq 0$ due

to the necessity of (7), then additional critical points can be introduced into the factored space. Fortunately, $\nabla f(\phi(U)) \not\succeq 0$ also implies that the first quadratic form in (9) might be negative for a properly chosen direction D. To sum up, the critical points of $g(U)$ can be classified into two categories: the global optima in the optimal factor set $\mathcal{A}^\star$ with $\nabla f(\phi(U)) \succeq 0$ and those with $\nabla f(\phi(U)) \not\succeq 0$. For the latter case, by choosing a proper direction D, we will argue that the Hessian quadratic form (9) has a strictly negative eigenvalue, and hence moving along D in a short distance will decrease the value of $g(U)$, implying that they are strict saddles and are not local minima.

We argue that a good choice of D is the direction from current U to its closest point in the optimal factor set $\mathcal{A}^\star$. Formally, $D = U - U^\star R$ where $R = \operatorname{argmin}_{\tilde{R}\in\mathbb{O}_r} \|U - U^\star\tilde{R}\|_F$ is the optimal rotation for the orthogonal Procrustes problem. Plugging D into the first term of (9), we simplify it as

$$\begin{aligned}
&\langle \nabla f(\phi(U)), DD^T\rangle \\
&= \langle \nabla f(\phi(U)), U^\star U^{\star T} - U^\star R U^T - U(U^\star R)^T + UU^T\rangle \\
&= \langle \nabla f(\phi(U)), U^\star U^{\star T}\rangle \\
&= \langle \nabla f(\phi(U)), \phi(U^\star) - \phi(U)\rangle, \qquad (10)
\end{aligned}$$

where in the second equality the last three terms involving U were canceled and in the last equality the term $-UU^T$ was reintroduced both due to the critical point property $\nabla f(\phi(U))U = \mathbf{0}$. To build intuition on why (10) is negative while the second term in (9) remains small, we consider a simple example: the matrix Principal Component Analysis (PCA) problem.

Example 1. *Matrix PCA Problem.* Consider the PCA problem for symmetric PSD matrices:

$$\underset{X\in\mathbb{R}^{n\times n}}{\text{minimize}}\, f(X) := \frac{1}{2}\|X - X^\star\|_F^2 \text{ subject to } X \succeq 0,$$

where $X^\star$ is a symmetric PSD matrix of rank $r^\star$. Apparently, the optimal solution is $X = X^\star$. Now consider the factored problem:

$$\underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}}\, g(U) := f(\phi(U)) = \frac{1}{2}\|\phi(U) - \phi(U^\star)\|_F^2,$$

where $U^\star \in \mathbb{R}^{n\times r}$ satisfies $\phi(U^\star) = X^\star$. Our goal is to show that any critical point U such that $\phi(U) \neq X^\star$ is a strict saddle. Since $\nabla f(\phi(U)) = \phi(U) - \phi(U^\star)$, by (10), the first term of $[\nabla^2 g(U)](D, D)$ in (9) becomes

$$\begin{aligned}
2\langle \nabla f(\phi(U)), DD^T\rangle &= 2\langle \nabla f(\phi(U)), \phi(U^\star) - \phi(U)\rangle \\
&= 2\langle \phi(U) - \phi(U^\star), \phi(U^\star) - \phi(U)\rangle \\
&= -2\|\phi(U) - \phi(U^\star)\|_F^2, \qquad (11)
\end{aligned}$$

which is strictly negative.

The second term $[\nabla^2 f(\phi(U))](DU^T + UD^T, DU^T + UD^T)$, essentially vanishes since we will see in the next that $DU^T = \mathbf{0}$ (hence $UD^T = \mathbf{0}$). For this purpose, let $X^\star = \sum_{i=1}^{r^\star} \lambda_i \boldsymbol{q}_i \boldsymbol{q}_i^T$ be the reduced eigenvalue decomposition of $X^\star$, where $\boldsymbol{q}_i$'s are orthonormal and $\lambda_i > 0$. Similarly, let $\phi(U) = \sum_{i=1}^{r'} \mu_i \boldsymbol{v}_i \boldsymbol{v}_i^T$ be the reduced eigenvalue decomposition of $\phi(U)$, where $r' = \operatorname{rank}(U)$. The critical point U satisfies $-\nabla g(U) = 2(X^\star - \phi(U))U = \mathbf{0}$, *i.e.,*

$$\mathbf{0} = (X^\star - \sum_{i=1}^{r'} \mu_i \boldsymbol{v}_i \boldsymbol{v}_i^T)\boldsymbol{v}_j = X^\star \boldsymbol{v}_j - \mu_j \boldsymbol{v}_j, j = 1, \ldots, r'$$

implying $\{\mu_j, \boldsymbol{v}_j\}_{j=1}^{r'}$ form eigenvalue-eigenvector pairs of $X^\star$ (so $r' \le r^\star$). Then, by reordering the indices if necessary,

$$\mu_j = \lambda_j \text{ and } \boldsymbol{v}_j = \boldsymbol{q}_j, j = 1, \ldots, r'.$$

Hence $U = [\sqrt{\lambda_1}\boldsymbol{q}_1 \ \cdots \ \sqrt{\lambda_{r'}}\boldsymbol{q}_{r'} \ \mathbf{0}_{(r-r')\times n}]V^T$ for some orthonormal matrix $V \in \mathbb{R}^{r\times r}$. Without loss of generality, we can choose $U^\star = [\sqrt{\lambda_1}\boldsymbol{q}_1 \ \cdots \ \sqrt{\lambda_{r^\star}}\boldsymbol{q}_{r^\star} \ \mathbf{0}_{(r-r^\star)\times n}]$. By the Procrustes Lemma in [25], we get $R = V^T$. Finally,

$$DU^T = UU^T - U^\star R U^T = \sum_{j=1}^{r'} \lambda_j \boldsymbol{q}_j \boldsymbol{q}_j^T - \sum_{j=1}^{r'} \lambda_j \boldsymbol{q}_j \boldsymbol{q}_j^T = \mathbf{0}.$$

Hence $[\nabla^2 g(U)](D, D)$ is simply determined by its first term (11):

$$\begin{aligned}
[\nabla^2 g(U)](D, D) &= -2\|\phi(U) - \phi(U^\star)\|_F^2 \\
&\le -2\min\{\rho(U)^2, \rho(U^\star)^2\}\|D\|_F^2 \\
&= -2\rho(X^\star)\|D\|_F^2,
\end{aligned}$$

where the inequality follows from Lemma 1 and the last equality holds since all eigenvalues of $\phi(U)$ come from those of $X^\star$. This further implies $\lambda_{\min}(\nabla^2 g(U)) \le -2\rho(X^\star)$.

This simple example is ideal in several ways, particularly the gradient $\nabla f(\phi(U)) = \phi(U) - \phi(U^\star)$, which directly establishes the negativity of the first term in (9); and by choosing $D = U - U^\star R$ and using $DU^T = \mathbf{0}$, the second term vanishes. Both are not true any more for general objective functions $f(X)$. However, the example does suggest that the direction $D = U - U^\star R$ is a good choice to show $[\nabla^2 g(U)](D, D) \le -\tau\|D\|_F^2$ for some $\tau > 0$. For a formal proof, we will also use the direction $D - U - U^\star R$ to show those critical points U not corresponding to $X^\star$ have a negative directional curvature for general factored objective function $g(U)$.

6. CONCLUSIONS

This work investigates the minimization of a convex function $f(X)$ over the cone of PSD matrices. To improve computational efficiency, we focus on the factored problem, and show it has a benign landscape: each critical point is either a factor of the globally optimal solution $X^\star$, or a strict saddle.

7. REFERENCES

[1] Q. Li and G. Tang, "The nonconvex geometry of low-rank matrix optimizations with general objective functions," *arXiv:1611.03060*, 2016.

[2] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[3] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Physical review letters*, vol. 105, no. 15, p. 150401, 2010.

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

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

[8] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," *arXiv preprint arXiv:1704.00708*, 2017.

[9] 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.

[10] Q. Li, Z. Zhu, and G. Tang, "Geometry of factored nuclear norm regularization," *arXiv preprint arXiv:1704.01265*, 2017.

[11] 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.

[12] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "Global optimality in low-rank matrix optimization," *arXiv preprint arXiv:1702.07945*, 2017.

[13] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent converges to minimizers," *University of California, Berkeley*, vol. 1050, p. 16, 2016.

[14] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[15] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[16] J. Sun, Q. Qu, and J. Wright, "When are nonconvex problems not scary?," *arXiv preprint arXiv:1510.06096*, 2015.

[17] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle pointsłonline stochastic gradient for tensor decomposition," in *Proceedings of The 28th Conference on Learning Theory*, pp. 797–842, 2015.

[18] J. Sun, *When Are Nonconvex Optimization Problems Not Scary?* PhD thesis, COLUMBIA UNIVERSITY, 2016.

[19] A. Agarwal, S. Negahban, and M. J. Wainwright, "Fast global convergence rates of gradient methods for high-dimensional statistical recovery," in *Advances in Neural Information Processing Systems*, pp. 37–45, 2010.

[20] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *Journal of Machine Learning Research*, vol. 13, no. May, pp. 1665–1697, 2012.

[21] 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.

[22] 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.

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

[24] J. Wellner, "Log-concave distributions: definitions, properties, and consequences," 2012.

[25] N. Higham and P. Papadimitriou, "Matrix procrustes problems," *Rapport technique, University of Manchester*, 1995.