# Non-Convex Low-Rank Matrix Recovery from Corrupted Random Linear Measurements

Yuanxin Li, Yuejie Chi
Department of ECE
The Ohio State University, Columbus, OH 43210
{li.3822, chi.97}@osu.edu

Huishuai Zhang, Yingbin Liang
Department of EECS
Syracuse University, Syracuse, NY 13244
{hzhan23, yliang06}@syr.edu

***Abstract*—Recent work has demonstrated the effectiveness of gradient descent for recovering low-rank matrices from random linear measurements in a globally convergent manner. However, their performance is highly sensitive in the presence of outliers that may take arbitrary values, which is common in practice. In this paper, we propose a truncated gradient descent algorithm to improve the robustness against outliers, where the truncation is performed to rule out the contributions from samples that deviate significantly from the *sample median*. A restricted isometry property regarding the sample median is introduced to provide a theoretical footing of the proposed algorithm for the Gaussian orthogonal ensemble. Extensive numerical experiments are provided to validate the superior performance of the proposed algorithm.**

***Index Terms*—median, gradient descent, outliers, low-rank matrix recovery**

## I. INTRODUCTION

A considerable amount of work has been done on low-rank matrix recovery in recent years, and it is shown that low-rank matrices can be recovered accurately and efficiently from a much smaller number of observations than their ambient dimensions [1]–[5]. An extensive overview can be found in [6]. It has been well recognized that convex relaxation is a popular strategy which replaces the low-rank constraint by a convex surrogate, such as nuclear norm minimization [7]–[9]. However, despite statistical (near-)optimality, their computational costs are prohibitive for high-dimensional problems.

In practice, a widely used alternative, pioneered by Burer and Monteiro [10], is to directly estimate the factors of a low-rank matrix, which has a much lower-dimensional representation and therefore admits more computationally and memory efficient algorithms. This typically leads to a non-convex loss function. Recently, a series of work has demonstrated that, starting from a careful initialization, simple algorithms such as gradient descent [11]–[15] enjoy global convergence guarantees under a near-optimal sample complexity. On the other hand, the global geometry of non-convex low-rank matrix estimation has been investigated in [16]–[19], and it is proven that no spurious local optima, except strict saddle points, exist under suitable conditions, which implies global convergence from random initialization, provided the algorithm of choice can escape saddle points [20].

In this paper, we focus on low-rank matrix recovery from random linear measurements in the presence of outliers, which is formulated in a way similar to [11]. Specifically, the low-rank matrix of interest is a positive semidefinite (PSD) matrix and the sensing matrices are drawn i.i.d. from the Gaussian orthogonal ensemble. Moreover, we assume the measurements are corrupted by sparse outliers, possibly in an adversarial fashion with arbitrary amplitudes. Although convex optimization can be still effective [21], [22], our goal is to develop fast and robust non-convex alternatives that are globally convergent.

Unfortunately, the vanilla gradient descent algorithm in [11] is not robust in the presence of outliers, as the outliers can perturb the search directions arbitrarily. In [23], a median-truncated gradient descent algorithm is proposed for non-convex robust phase retrieval, where the sample median is exploited to control both the initialization and the gradient step, where only a subset of samples are selected to contribute to the search direction in each iteration. Inspired by [23], we design a median-truncated gradient descent algorithm for low-rank matrix recovery, where we carefully set the truncation strategy to mitigate the impact of outliers. The sample median is a highly robust object against adversarial outliers, which can be computed in linear time [24]. A highlight of the proposed algorithm is that it does not assume a priori information regarding the outliers. A restricted isometry property (RIP) of the sample median is established to provide theoretical grounds of the proposed algorithm. Numerical experiments demonstrate the excellent empirical performance of the proposed algorithm for low-rank matrix recovery from outlier-corrupted measurements, which significantly outperforms existing algorithms that are not resilient to outliers [11].

The remainder of this paper is organized as follows. In Section II, we mathematically formulate the low-rank PSD matrix recovery problem. The details of the proposed algorithm are given in Section III. Numerical experiments are provided in Section IV to validate the performance of the proposed algorithm. Finally, we conclude and discuss the future work in Section V. Throughout this paper, we denote vectors by bold lowercases and matrices by bold capitals. The transpose of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^T$, and $\|\boldsymbol{A}\|_F$ represents the Frobenius norm. $\mathrm{med}(\boldsymbol{y})$ denotes the median of the entries in vector $\boldsymbol{y}$, and $|\boldsymbol{y}|$ denotes entry-wise absolute value. Besides, the inner product between two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is defined as $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{Tr}\left(\boldsymbol{B}^T \boldsymbol{A}\right)$, where $\mathrm{Tr}(\cdot)$ denotes the trace.

978-1-5386-1565-2/17/$31.00 

## II. PROBLEM FORMULATION

Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a rank-$r$ PSD matrix that can be written as $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{X}^T$, where $\boldsymbol{X} \in \mathbb{R}^{n\times r}$. Denote the set of sensing matrices by $\{\boldsymbol{A}_i\}_{i=1}^m$, where $\boldsymbol{A}_i \in \mathbb{R}^{n\times n}$ is the $i$th symmetric sensing matrix, generated i.i.d. from the Gaussian orthogonal ensemble with $(\boldsymbol{A}_i)_{k,k} \sim \mathcal{N}(0,2)$, $(\boldsymbol{A}_i)_{k,t} \sim \mathcal{N}(0,1)$ for $k < t$, and $(\boldsymbol{A}_i)_{k,t} = (\boldsymbol{A}_i)_{t,k}$.

Denote the index set of corrupted measurements by $\mathcal{S}$, and correspondingly, the index set of clean measurements is given as the complementary set $\mathcal{S}^c$. Mathematically, the measurements $\boldsymbol{y} = \{y_i\}_{i=1}^m$ can be represented as

$$y_i = \begin{cases} \langle \boldsymbol{A}_i, \boldsymbol{M}\rangle, & \text{if} \quad i \in \mathcal{S}^c; \\ \eta_i, & \text{if} \quad i \in \mathcal{S}, \end{cases} \tag{1}$$

where $\boldsymbol{\eta} = \{\eta_i\}_{i\in\mathcal{S}}$ is the set of outliers that can take arbitrary values. Further assume the cardinality of $\mathcal{S}$ as $|\mathcal{S}| = s \cdot m$, where $0 \le s < 1$ is the fraction of outliers. Our goal is to recover $\boldsymbol{M}$ from the corrupted measurements, without a priori knowledge of the outliers, in a computationally efficient and provably accurate manner.

## III. MEDIAN-TRUNCATED GRADIENT DESCENT

Instead of recovering $\boldsymbol{M}$, we aim to directly recover its low-rank factor $\boldsymbol{X}$. It is straightforward that for any orthonormal matrix $\boldsymbol{P} \in \mathbb{R}^{r\times r}$, we have $(\boldsymbol{X}\boldsymbol{P})(\boldsymbol{X}\boldsymbol{P})^T = \boldsymbol{X}\boldsymbol{X}^T$, and consequently, $\boldsymbol{X}$ can be recovered only up to orthonormal transformations. Furthermore, we introduce the shorthand notations for the linear maps $\mathcal{A}_i(\boldsymbol{U}) = \{\mathbb{R}^{n\times r} \mapsto \mathbb{R} : \langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^T\rangle\}$, and $\mathcal{A}(\boldsymbol{U}) = \{\mathbb{R}^{n\times r} \mapsto \mathbb{R}^m : \{\mathcal{A}_i(\boldsymbol{U})\}_{i=1}^m\}$.

### A. Algorithm description

To begin, consider the following *oracle* loss function,

$$f_{\text{oracle}}(\boldsymbol{U}) = \frac{1}{4m}\sum_{i\in\mathcal{S}^c}\left(y_i - \mathcal{A}_i(\boldsymbol{U})\right)^2, \tag{2}$$

which aims to minimize the quadratic loss over *clean* measurements only. However, since the oracle information regarding the support of outliers is absent, we cannot directly minimize $f_{\text{oracle}}(\boldsymbol{U})$. Moreover, $f_{\text{oracle}}(\boldsymbol{U})$ is non-convex. To proceed, define the sample-wise loss function as

$$f_i(\boldsymbol{U}) = \frac{1}{4m}\left(y_i - \mathcal{A}_i(\boldsymbol{U})\right)^2, \tag{3}$$

whose gradient with respect to $\boldsymbol{U}$ can be written as

$$\nabla f_i(\boldsymbol{U}) = \frac{1}{m}(\mathcal{A}_i(\boldsymbol{U}) - y_i)\boldsymbol{A}_i\boldsymbol{U}. \tag{4}$$

In sharp contrast to the gradient descent approach in [11], we propose to control the initialization and the search directions more carefully in order to adaptively eliminate outliers. For initialization, we adopt the spectral method, which uses the top eigenvectors of the sample-weighted matrix in (6), where only the samples whose values do not significantly digress from the sample median are included. In the gradient loop, we update the estimate via (7), which can be viewed as a truncated gradient descent update:

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \tilde{\mu}_t \sum_{i\in\left\{i|\mathcal{E}_i^t \text{ is true}\right\}} \nabla f_i(\boldsymbol{U}_t), \tag{5}$$

where only samples whose measurement residuals at the current iteration do not digress from the sample median significantly are included. Note that the set $\{i|\mathcal{E}_i^t \text{ is true}\}$ varies per iteration, and therefore can adaptively prune the outliers. Details of the proposed algorithm are provided in Algorithm 1, where the stopping criterion is simply set as reaching a preset maximum number of iterations. In practice, it is also possible to set the stopping criteria by examining the progress between iterations.

The computational advantage of gradient descent has been justified in the earlier work, e.g. [11]. It is worthwhile to note that the key difference between the proposed Algorithm 1 and the one in [11] is the truncation strategy used in (6) and (7), both of which improve the robustness of the algorithm guided by the sample median.

**Algorithm 1** Median-Truncated Gradient Descent Algorithm

**Parameters:** Thresholds $\alpha_y$ and $\alpha_h$, the step size $\mu_t$, and the rank $r$.

**Input:** The measurements $\boldsymbol{y} = \{y_i\}_{i=1}^m$, and the sensing matrices $\{\boldsymbol{A}_i\}_{i=1}^m$.

**Initialization:** $\boldsymbol{U}_0 = \boldsymbol{Z}\boldsymbol{\Sigma}$, where the columns of $\boldsymbol{Z}$ contain the normalized eigenvectors corresponding to the $r$ largest eigenvalues in terms of absolute values, i.e. $|\sigma_1| \ge |\sigma_2| \ge \cdots \ge |\sigma_r|$, of the matrix

$$\boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^m y_i\boldsymbol{A}_i\mathbf{1}_{\left\{|y_i|\le\alpha_y^2\text{med}(|\boldsymbol{y}|)/0.9539\right\}}, \tag{6}$$

and $\boldsymbol{\Sigma}$ is a $r\times r$ diagonal matrix, with the $i$th diagonal entry given as $\sqrt{|\sigma_i|/2}$.

**Gradient Loop:** For $t = 0:1:T-1$ do

$$\begin{aligned}\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \mu_t \cdot &\frac{1}{\sum_{k=1}^r |\sigma_k|/2} \\ &\cdot \frac{1}{m}\sum_{i=1}^m \left(\mathcal{A}_i(\boldsymbol{U}_t) - y_i\right)\boldsymbol{A}_i\boldsymbol{U}_t\mathbf{1}_{\mathcal{E}_i^t},\end{aligned} \tag{7}$$

where

$$\mathcal{E}_i^t = \left\{|y_i - \mathcal{A}_i(\boldsymbol{U}_t)| \le \alpha_h \text{med}\left(|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}_t)|\right)\right\}. \tag{8}$$

**Output:** $\hat{\boldsymbol{X}} = \boldsymbol{U}_T$.

### B. RIP-like property for sample median

RIP plays a critical role in the analysis of both convex [7] and non-convex [12] procedures for low-rank matrix recovery. In particular, it is shown in [23] that a similar property for the sample median holds for the phase retrieval problem. Below, for the problem of low-rank matrix recovery, we establish that the sample median also possesses RIP-like property, even when a constant fraction of the measurements are arbitrarily

corrupted, as long as the sample complexity is on the order of $nr \log n$.

**Proposition 1** (RIP of sample median)**.** *Suppose $s \leq s_0$, where $s_0$ is a small enough constant. Fix a small $\epsilon \in (0,1)$. If $m \geq c_0 \left(\epsilon^{-2} \log \epsilon^{-1}\right) nr \log n$, we have*

$$(\gamma_1 - \epsilon) \left\|\boldsymbol{X}\boldsymbol{X}^T\right\|_F \leq \operatorname{med}\left(|\boldsymbol{y}|\right) \leq (\gamma_2 + \epsilon) \left\|\boldsymbol{X}\boldsymbol{X}^T\right\|_F,$$

*and*

$$\operatorname{med}\left(|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U})|\right) \geq (\gamma_3 - \epsilon) \left\|\boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{U}\boldsymbol{U}^T\right\|_F;$$
$$\operatorname{med}\left(|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U})|\right) \leq (\gamma_4 + \epsilon) \left\|\boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{U}\boldsymbol{U}^T\right\|_F,$$

*with probability at least $1 - c_1 \exp\left(-c_2 m \epsilon^2\right)$ for all matrices $\boldsymbol{U}, \boldsymbol{X} \in \mathbb{R}^{n \times r}$, where both $\gamma_i$, $i = 1,2,3,4$, and $c_i$, $i = 0,1,2$, are some universal constants only depending on $s_0$. Specifically, when $s_0 = 0$, we can set $\gamma_1 = \gamma_2 = 0.9539$ and $\gamma_3 = \gamma_4 = 0.9539$.*

Our ongoing investigation suggests that this critical concentration property provides theoretical footings on the success of Algorithm 1, whose complete theoretical guarantee will be presented in a later draft.

## IV. NUMERICAL EXPERIMENTS

In this section, we numerically evaluate the performance of the proposed Algorithm 1. In all the experiments, we pick $\alpha_y = 3$, $\alpha_h = 5$ and consider a constant step size $\mu_t = 0.4$. We set the maximum number of iterations as $T = 10^4$.

Let $n = 40$, $r = 4$ and $m = 480$. We randomly generate a rank-$r$ PSD matrix as $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{X}^T$, where $\boldsymbol{X}$ is composed of i.i.d. standard Gaussian variables. The $i$-th sensing matrix $\boldsymbol{A}_i$ is generated as $\boldsymbol{A}_i = \left(\boldsymbol{B}_i + \boldsymbol{B}_i^T\right)/\sqrt{2}$, where $\boldsymbol{B}_i \in \mathbb{R}^{n \times n}$ is composed of i.i.d. standard Gaussian variables, $i = 1, 2, \ldots, m$. The support of the outliers are uniformly selected at random and their values are i.i.d. generated following the distribution $\mathcal{N}(0, 10^4 \|\boldsymbol{X}\boldsymbol{X}^T\|_F^2)$. The normalized reconstruction error is defined as $\|\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^T - \boldsymbol{X}\boldsymbol{X}^T\|_F / \|\boldsymbol{X}\boldsymbol{X}^T\|_F$, where $\hat{\boldsymbol{X}}$ is the estimate of the matrix factor.

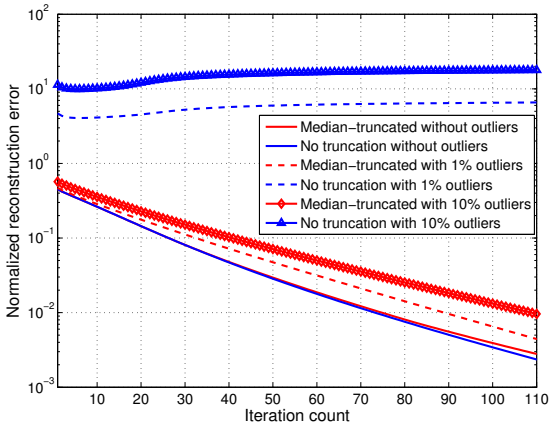



Fig. 1. Normalized reconstruction errors with respect to iteration count using Algorithm 1 and the algorithm in [11] under different corruption scenarios, when $n = 40$, $r = 4$ and $m = 480$.

Fig. 1 shows the normalized reconstruction errors with respect to the iteration count using Algorithm 1 and the algorithm in [11], which does not employ the median-truncation strategy, under different fractions of outliers. It can be seen that both algorithms yield comparable convergence rates in the absence of outliers. However, even with very few outliers (e.g. 1%), the algorithm in [11] suffers a dramatic performance degradation, while Algorithm 1 is much more robust and can still converge at a linear rate.

We next examine the phase transitions of Algorithm 1. Fix $n = 40$. Each trial is deemed a success if the normalized reconstruction error is below $10^{-6}$, and the success rate is calculated by averaging over 10 Monte Carlo trials. Fig. 2 (a) shows the success rates of Algorithm 1 with respect to the number of measurements and the rank, when the percent of outliers is fixed as $s = 5\%$, and (b) shows the success rates with respect to the percent of outliers and the rank, when the number of measurements is fixed as $m = 360$. It implies that the sample complexity of Algorithm 1 is near-optimal even under a constant fraction of outliers, and it is capable of tolerating a larger fraction of outliers when the rank is small.

Finally, we examine Algorithm 1 when the measurements are contaminated by both sparse outliers and dense noise. Fix $n = 40$, $r = 4$ and $s = 5\%$. The dense noise is generated with i.i.d. Gaussian entries following $\mathcal{N}\left(0, 0.01\|\boldsymbol{X}\boldsymbol{X}^T\|_F\right)$. Fig. 3 depicts the average normalized reconstruction errors with respect to the number of measurements using both Algorithm 1 and the algorithm in [11]. The performance of Algorithm 1 is comparable to that of the algorithm in [11] without outliers, therefore, it can handle outliers in an oblivious fashion. Moreover, the performance keeps stable as long as an upper bound of the true rank is provided.

## V. CONCLUSION

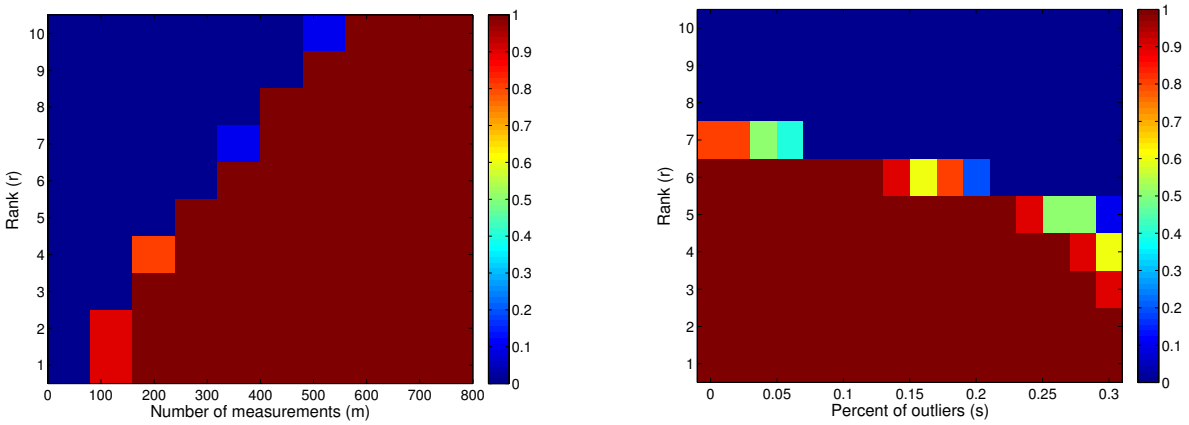
We propose an efficient median-truncated gradient descent algorithm to improve the efficacy and robustness of low-rank PSD matrix recovery from random linear measurements in the presence of sparse outliers, with possibly arbitrary magnitudes. The effectiveness of the proposed algorithm is validated through extensive numerical experiments. We also provide some initial evidence towards the theoretical guarantee of the proposed algorithm. The complete theoretical analysis will be presented in a future publication, together with the extension to the general rectangular low-rank matrix recovery problem.

## ACKNOWLEDGEMENT

This work of Y. Li and Y. Chi is supported in part by NSF under the grant ECCS-1650449 and by AFOSR under the grant FA9550-15-1-0205. The work of H. Zhang and Y. Liang is supported in part by NSF under grant ECCS 16-09916 and by AFOSR under grant FA9550-16-1-0077.

## REFERENCES

[1] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1548–1566, March 2011.




(a) (b)

Fig. 2. Phase transitions of low-rank PSD matrix recovery when $n = 40$, (a) with respect to the number of measurements $m$ and the rank $r$, when the percent of outliers is $s = 5\%$; (b) with respect to the percent of outliers $s$ and the rank $r$, when the number of measurements is $m = 360$.

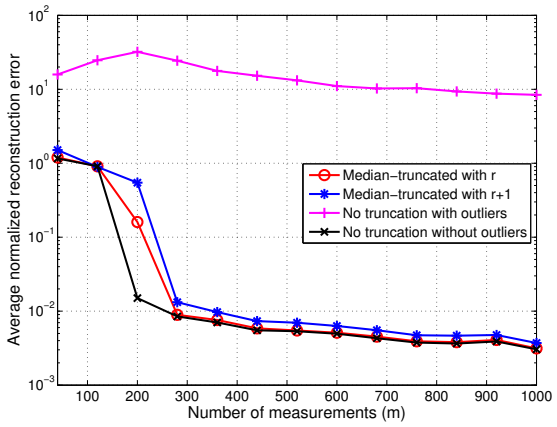



Fig. 3. Comparisons of average normalized reconstruction errors with respect to the number of measurements using Algorithm 1 and the algorithm in [11], when $n = 40$, $r = 4$, $s = 5\%$ and with additional Gaussian noise.


[2] S. Negahban and M. J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, vol. 39, no. 2, pp. 1069–1097, 2011.

[3] E. Candes and B. Recht, "Exact matrix completion via convex optimization," *Communications of the ACM*, vol. 55, no. 6, pp. 111–119, 2012.

[4] Y. Chen and Y. Chi, "Robust spectral compressed sensing via structured matrix completion," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6576–6601, 2014.

[5] Y. Chen, Y. Chi, and A. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4034–4059, July 2015.

[6] 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.

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

[9] E. J. Candes 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.

[10] 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.

[11] Q. Zheng and J. Lafferty, "A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[12] 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.

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

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

[15] D. Park, A. Kyrillidis, S. Bhojanapalli, C. Caramanis, and S. Sanghavi, "Provable non-convex projected gradient descent for a class of constrained matrix optimization problems," *stat*, vol. 1050, p. 4, 2016.

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

[19] 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.

[20] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—online stochastic gradient for tensor decomposition," *arXiv preprint arXiv:1503.02101*, 2015.

[21] 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, Jan 2017.

[22] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive principal component pursuit," *Information and Inference*, vol. 2, no. 1, pp. 32–68, 2013.

[23] H. Zhang, Y. Chi, and Y. Liang, "Provable non-convex phase retrieval with outliers: Median truncated wirtinger flow," *arXiv preprint arXiv:1603.03805*, 2016.

[24] P. J. Huber, *Robust statistics*. Springer, 2011.