# Towards Faster Rates and Oracle Property for Low-Rank Matrix Estimation

**Huan Gui** HUANGUI2@ILLINOIS.EDU
**Jiawei Han** HANJ@ILLINOIS.EDU
Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

**Quanquan Gu**[*] QG5W@VIRGINIA.EDU
Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22904, USA

## Abstract

We present a unified framework for low-rank matrix estimation with nonconvex penalty. A proximal gradient homotopy algorithm is developed to solve the proposed optimization problem. Theoretically, we first prove that the proposed estimator attains a faster statistical rate than the traditional low-rank matrix estimator with nuclear norm penalty. Moreover, we rigorously show that under a certain condition on the magnitude of the nonzero singular values, the proposed estimator enjoys oracle property (*i.e.*, exactly recovers the true rank of the matrix), besides attaining a faster rate. Extensive numerical experiments on both synthetic and real world datasets corroborate our theoretical findings.

## 1. Introduction

Statistical estimation of low-rank matrices (Srebro et al., 2004; Candès & Tao, 2010; Rohde et al., 2011; Koltchinskii et al., 2011a; Candès & Recht, 2012; Jain et al., 2013; Hardt, 2014; Jain & Netrapalli, 2014) has received increasing interest in the past decade. It has broad applications in many fields such as data mining and computer vision. For example, in the recommendation systems, one aims to predict the unknown preferences of a set of users over a set of items, provided a partially observed rating matrix. Another application of low-rank matrix estimation is image inpainting, to recover missing pixels based on a portion of pixels being observed.

Since it is not tractable to minimize the rank of a matrix directly, many surrogate loss functions of the matrix rank have been proposed (*e.g.*, nuclear norm (Srebro et al., 2004; Candès & Tao, 2010; Recht et al., 2010; Negahban & Wainwright, 2011; Koltchinskii et al., 2011a), Schatten-$p$ norm (Rohde et al., 2011; Nie et al., 2012), max norm (Srebro & Shraibman, 2005; Cai & Zhou, 2013), the von Neumann entropy (Koltchinskii et al., 2011b)). Among those surrogate losses for rank, nuclear norm is probably the most widely used penalty for low-rank matrix estimation (Negahban & Wainwright, 2011; Koltchinskii et al., 2011a), since it is the tightest convex relaxation of the matrix rank.

On the other hand, it is now well-known that $\ell_1$ penalty in Lasso (Fan & Li, 2001; Zhang, 2010; Zou, 2006) introduces a bias into the resulting estimator, which compromises the estimation accuracy. In contrast, nonconvex penalties such as smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010) are favored in terms of estimation accuracy and variable selection consistency (Wang et al., 2013b). Due to the close connection between $\ell_1$ norm and nuclear norm (nuclear norm can be seen as an $\ell_1$ norm defined on the singular values of a matrix), nonconvex penalties for low-rank matrix estimation have recently received increasing attention for low-rank matrix estimation. Typical examples of nonconvex approximation of the matrix rank include Schatten $\ell_p$-norm ($0 < p < 1$) (Nie et al., 2012), the truncated nuclear norm (Hu et al., 2013), and the MCP penalty defined on the singular values of a matrix (Wang et al., 2013a; Liu et al., 2013). Although good empirical results have been observed in these studies (Nie et al., 2012; Hu et al., 2013; Wang et al., 2013a; Liu et al., 2013; Lu et al., 2014; Yao et al., 2015), little is known about the theory of nonconvex penalty for low-rank matrix estimation. The theoretical justification for the nonconvex surrogates of matrix rank is still an open problem.

In this paper, to bridge the gap between practice and theory of low-rank matrix estimation, we propose a unified

[*] Corresponding Author

*Proceedings of the $33^{rd}$ International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

framework for low-rank matrix estimation with nonconvex penalty. A proximal gradient homotopy method is presented to solve the proposed estimator. We prove that our proposed estimator, by taking advantage of singular values with large magnitude, attains faster statistical convergence rates, compared with the conventional estimator with nuclear norm penalty. Furthermore, under a mild assumption on the magnitude of the singular values, we rigorously show that the proposed estimator enjoys oracle property, which exactly recovers the true rank of the underlying matrix, as well as attains a faster rate. Our theoretical results are verified through both simulations and thorough experiments on real world datasets for collaborative filtering and image inpainting.

**Notation.** We use lowercase letters $(a, b, \ldots)$ to denote scalars, bold lower case letters $(\mathbf{a}, \mathbf{b}, \ldots)$ for vectors, and bold upper case letters $(\mathbf{A}, \mathbf{B}, \ldots)$ for matrices. For a real number $a$, we denote by $\lfloor a \rfloor$ the largest integer that is no greater than $a$. For a vector $\mathbf{x}$, define vector norm as $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$. Considering matrix $\mathbf{A}$, we denote by $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ the largest and smallest eigenvalue of $\mathbf{A}$, respectively. For a pair of matrices $\mathbf{A}, \mathbf{B}$ with commensurate dimensions, $\langle \mathbf{A}, \mathbf{B} \rangle$ denotes the trace inner product on matrix space that $\langle \mathbf{A}, \mathbf{B} \rangle := \text{trace}(\mathbf{A}^\top \mathbf{B})$. Given a matrix $\mathbf{A} \in \mathbb{R}^{m_1 \times m_2}$, its (ordered) singular values are denoted by $\gamma_1(\mathbf{A}) \geq \gamma_2(\mathbf{A}) \geq \cdots \geq \gamma_m(\mathbf{A}) \geq 0$ where $m = \min\{m_1, m_2\}$. Moreover, $M = \max\{m_1, m_2\}$. We also define $\|\cdot\|$ for various norms defined on matrices, based on the singular values, including nuclear norm $\|\mathbf{A}\|_* = \sum_{i=1}^m \gamma_i(\mathbf{A})$, spectral norm $\|\mathbf{A}\|_2 = \gamma_1(\mathbf{A})$, and the Frobenius norm $\|\mathbf{A}\|_F = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle} = \sqrt{\sum_{i=1}^m \gamma_i^2(\mathbf{A})}$. In addition, we define $\|\mathbf{A}\|_\infty = \max_{1 \leq j \leq m_1, 1 \leq k \leq m_2} A_{jk}$, where $A_{jk}$ is the element of $\mathbf{A}$ at row $j$, column $k$.

## 2. Low-rank Matrix Estimation with Nonconvex Penalty

In this section, we present a unified framework for low-rank matrix estimation with nonconvex penalty, followed by the theoretical analysis of the proposed estimator.

### 2.1. The Observation Model

We consider a generic observation model as follows:

$$y_i = \langle \mathbf{X}_i, \mathbf{\Theta}^* \rangle + \epsilon_i \quad \text{for } i = 1, 2, \ldots, n, \qquad (2.1)$$

where $\{\mathbf{X}_i\}_{i=1}^n$ is a sequence of observation matrices, and $\{\epsilon_i\}_{i=1}^n$ are i.i.d. zero mean sub-Gaussian observation noise with variance $\sigma^2$. Moreover, the observation model can be rewritten in a more compact way as $\mathbf{y} = \mathfrak{X}(\mathbf{\Theta}^*) + \boldsymbol{\epsilon}$, where $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$, and $\mathfrak{X}(\cdot)$ is a linear operator that $\mathfrak{X}(\mathbf{\Theta}^*) := (\langle \mathbf{X}_1, \mathbf{\Theta}^* \rangle, \langle \mathbf{X}_2, \mathbf{\Theta}^* \rangle, \cdots, \langle \mathbf{X}_n, \mathbf{\Theta}^* \rangle)^\top$. In addition, we define the adjoint of the operator $\mathfrak{X}$ as $\mathfrak{X}^* : \mathbb{R}^n \rightarrow \mathbb{R}^{m_1 \times m_2}$, which is defined as $\mathfrak{X}^*(\boldsymbol{\epsilon}) = \sum_{i=1}^n \epsilon_i \mathbf{X}_i$. It is worth noting that the observation model presented in (2.1), by which many matrix estimation problems can be unified, has also been considered before by Koltchinskii et al. (2011a); Negahban & Wainwright (2011).

### 2.2. Examples

Low-rank matrix estimation has broad applications. We briefly review two examples: matrix completion and matrix sensing. For more examples, please refer to Koltchinskii et al. (2011a); Negahban & Wainwright (2011).

**Example 2.1** (Matrix Completion)**.** In the setting of matrix completion with noise, one uniformly observes partial entries of the unknown matrix $\mathbf{\Theta}^*$ with noise. In detail, the observation matrix $\mathbf{X}_i \in \mathbb{R}^{m_1 \times m_2}$ is in the form of $\mathbf{X}_i = \mathbf{e}_{j_i}(m_1)\mathbf{e}_{k_i}(m_2)^\top$, where $\mathbf{e}_{j_i}(m_1)$ and $\mathbf{e}_{j_i}(m_2)$ are the canonical basis vectors in $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$, respectively.

**Example 2.2** (Matrix Sensing)**.** In the setting of matrix sensing, one observes a set of random projections of the unknown matrix $\mathbf{\Theta}^*$. More specifically, the observation matrix $\mathbf{X}_i \in \mathbb{R}^{m_1 \times m_2}$ has i.i.d. standard normal $N(0, 1)$ entries, so that one makes observations of the form $y_i = \langle \mathbf{X}_i, \mathbf{\Theta}^* \rangle + \epsilon_i$. It is obvious that matrix sensing is an instance of the model (2.1).

### 2.3. The Proposed Estimator

We now propose an estimator that is naturally designed for estimating low-rank matrices. Given a collection of $n$ samples $\mathcal{Z}_1^n = \{(y_i, \mathbf{X}_i)\}_{i=1}^n$, which is assumed to be generated from the observation model (2.1), the unknown low-rank matrix $\mathbf{\Theta}^* \in \mathbb{R}^{m_1 \times m_2}$ can be estimated by solving the following optimization problem

$$\widehat{\mathbf{\Theta}} = \underset{\mathbf{\Theta} \in \mathbb{R}^{m_1 \times m_2}}{\text{argmin}} \; \frac{1}{2n} \big\| \boldsymbol{y} - \mathfrak{X}(\mathbf{\Theta}) \big\|_2^2 + \mathcal{P}_\lambda(\mathbf{\Theta}), \qquad (2.2)$$

which includes two components: (i) the empirical loss function $\mathcal{L}_n(\mathbf{\Theta}) = (2n)^{-1} \|\boldsymbol{y} - \mathfrak{X}(\mathbf{\Theta})\|_2^2$; and (ii) the nonconvex penalty (Fan & Li, 2001; Zhang, 2010; Zhang et al., 2012) $\mathcal{P}_\lambda(\mathbf{\Theta})$ with regularization parameter $\lambda$, which helps to enforce the low-rank structure constraint on the regularized M-estimator $\widehat{\mathbf{\Theta}}$. Considering the low rank assumption on the matrices, we apply the nonconvex regularization on the singular values of $\mathbf{\Theta}$, which induces sparsity of singular values, and therefore low-rankness of the matrix. For singular values of $\mathbf{\Theta}$, $\boldsymbol{\gamma}(\mathbf{\Theta}) = \big(\gamma_1(\mathbf{\Theta}), \gamma_2(\mathbf{\Theta}), \ldots, \gamma_m(\mathbf{\Theta})\big)$, where $\gamma_1(\mathbf{\Theta}) \geq \ldots \geq \gamma_m(\mathbf{\Theta}) \geq 0$, we define $\mathcal{P}_\lambda(\mathbf{\Theta}) = \sum_{i=1}^n p_\lambda\big(\gamma_i(\mathbf{\Theta})\big)$, where $p_\lambda$ is a univariate nonconvex function. There is a line of research on nonconvex regularization and various nonconvex penalties have been proposed, such as SCAD (Fan & Li, 2001) and MCP (Zhang, 2010). We take SCAD and MCP penalties as illustrations.

Hence, for SCAD, the function $p_\lambda(\cdot)$ is defined as follows

$$p_\lambda(t) = \begin{cases} \lambda|t|, & \text{if } |t| \le \lambda, \\ -\frac{t^2 - 2b\lambda|t| + \lambda^2}{2(b-1)}, & \text{if } \lambda < |t| \le b\lambda, \\ (b+1)\lambda^2/2, & \text{if } |t| > b\lambda, \end{cases}$$

where $b > 2$ and $\lambda > 0$. The SCAD penalty corresponds to a quadratic spline function with knots at $t = \lambda$ and $t = b\lambda$. Regarding MCP, we have

$$\begin{aligned} p_\lambda(t) &= \lambda \int_0^{|t|} \Big(1 - \frac{z}{\lambda b}\Big)_+ dz \\ &= \Big(\lambda|t| - \frac{t^2}{2b}\Big)\mathbf{1}(|t| \le b\lambda) + \frac{b\lambda^2}{2}\mathbf{1}(|t| > b\lambda), \end{aligned}$$

where $b > 0$ is a fix parameter.

In addition, the nonconvex penalty $p_\lambda(t)$ can be further decomposed as $p_\lambda(t) = \lambda|t| + q_\lambda(t)$, where $|t|$ is the $\ell_1$ penalty and $q_\lambda(t)$ is a concave component. For the SCAD penalty, $q_\lambda(t)$ can be obtained as follows,

$$\begin{aligned} q_\lambda(t) = &-\big(|t| + \lambda\big)^2/\big(2(b-1)\big)\mathbf{1}(\lambda < |t| \le b\lambda) \\ &+ \big(1/2(b+1)\lambda^2 - \lambda|t|\big)\mathbf{1}(|t| > b\lambda). \end{aligned}$$

For MCP, the concave part is

$$q_\lambda(t) = -\frac{t^2}{2b}\mathbf{1}(|t| \le b\lambda) + \bigg(\frac{b\lambda^2}{2} - \lambda|t|\bigg)\mathbf{1}(|t| > b\lambda).$$

Since the regularization term $\mathcal{P}_\lambda(\mathbf{\Theta})$ is imposed on the vector of singular values, hence, the decomposability of $p_\lambda(t)$ is equivalent to the decomposability of $\mathcal{P}_\lambda(\mathbf{\Theta})$ as $\mathcal{P}_\lambda(\mathbf{\Theta}) = \lambda\|\mathbf{\Theta}\|_* + \mathcal{Q}_\lambda(\mathbf{\Theta})$, where $\mathcal{Q}_\lambda(\mathbf{\Theta})$ is the concave component, $\mathcal{Q}_\lambda(\mathbf{\Theta}) = \sum_{i=1}^m q_\lambda\big(\gamma_i(\mathbf{\Theta})\big)$, and $\|\mathbf{\Theta}\|_*$ is the nuclear norm.

### 2.4. Optimization Algorithm

In this section, we present a proximal gradient homotopy algorithm, which is adapted from Xiao & Zhang (2013), as shown in Algorithm 1, to solve the optimization problem with nonconvex penalty (2.2).

The main idea of proximal gradient homotopy method (PGH) is to solve the optimization problem with an initial regularization parameter $\lambda = \lambda_0$ that is sufficiently large and then gradually decrease $\lambda$ until the target regularization parameter $\lambda_{\text{tgt}}$ is attained, which will be given in Theorem 3.4 and Theorem 3.5, respecting different conditions.

In addition, we have $\lambda_t = \eta^t\lambda_0$, where $\eta$ is an absolute constant. The number of iterations for the homotopy algorithm is $K = \lfloor \ln(\lambda_0/\lambda_{tgt})/\ln(1/\eta) \rfloor$. For the final stage of the proximal gradient homotopy method, we need to solve up to high precision with $\epsilon_{\text{opt}} \ll \lambda_{\text{tgt}}/4$. The key component in Algorithm 1 is the function ProxGrad() (Line 6 and 8), a proximal gradient method tailored for the M-estimator with nonconvex penalty, as shown in Algorithm 2. The details of the proximal gradient algorithm are introduced as follows.

**Algorithm 1** $\{\mathbf{\Theta}^t\}_{t=1}^{K+1} \leftarrow \text{PGH}(\lambda_0, \lambda_{\text{tgt}}, \epsilon_{\text{opt}}, L_{\min})$

```
input $\lambda_0 > 0, \lambda_{\text{tgt}} > 0, \epsilon_{\text{opt}} > 0, L_{\min} > 0$
1: parameters $\eta \in (0,1), \delta \in (0,1)$
2: initialize $\mathbf{\Theta}^0 \leftarrow \mathbf{0}, L_0 \leftarrow L_{\min}, K \leftarrow \Big\lfloor \frac{\ln(\lambda_0/\lambda_{tgt})}{\ln(1/\eta)} \Big\rfloor$
3: for $t = 0, 1, 2, \ldots, K-1$ do
4:     $\lambda_{t+1} \leftarrow \eta\lambda_t$
5:     $\epsilon_{t+1} \leftarrow \lambda_t/4$
6:     $\{\mathbf{\Theta}^{t+1}, L_{t+1}\} \leftarrow \text{ProxGrad}(\lambda_{t+1}, \epsilon_{t+1}, \mathbf{\Theta}^t, L_t)$
7: end for
8: $\{\mathbf{\Theta}^{K+1}, L_{K+1}\} \leftarrow \text{ProxGrad}(\lambda_{\text{tgt}}, \epsilon_{\text{opt}}, \mathbf{\Theta}^K, L_K)$
9: return $\{\mathbf{\Theta}^t\}_{t=1}^{K+1}$
```

Recall that $\mathcal{P}_\lambda(\mathbf{\Theta}) = \lambda\|\mathbf{\Theta}\|_* + \mathcal{Q}_\lambda(\mathbf{\Theta})$. We define

$$\phi_\lambda(\mathbf{\Theta}) = \mathcal{L}_n(\mathbf{\Theta}) + \mathcal{P}_\lambda(\mathbf{\Theta}) = \widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}) + \lambda\|\mathbf{\Theta}\|_*, \quad (2.3)$$

where $\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}) = \mathcal{L}_n(\mathbf{\Theta}) + \mathcal{Q}_\lambda(\mathbf{\Theta})$. For any fixed matrix $\mathbf{M}$ and a given regularization parameter $\lambda$, we define a local model of $\phi_\lambda(\mathbf{\Theta})$ around $\mathbf{M}$ using a simple quadratic approximation of $\widetilde{\mathcal{L}}_{n,\lambda}(\cdot)$ as follows:

$$\begin{aligned} \psi_{L,\lambda}(\mathbf{\Theta}; \mathbf{M}) = \widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{M}) + \nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{M})^\top(\mathbf{\Theta} - \mathbf{M}) \\ + \frac{L}{2}\|\mathbf{\Theta} - \mathbf{M}\|_F^2 + \lambda\|\mathbf{\Theta}\|_*. \quad (2.4) \end{aligned}$$

Suppose $\mathcal{T}_{L,\lambda}(\mathbf{M})$ is the unique minimize of $\psi_{L,\lambda}(\mathbf{\Theta}; \mathbf{M})$,

$$\mathcal{T}_{L,\lambda}(\mathbf{M}) = \operatorname*{argmin}_{\mathbf{\Theta}} \psi_{L,\lambda}(\mathbf{\Theta}; \mathbf{M}). \quad (2.5)$$

Via exploiting the structure of the nuclear norm regularization in (2.4), the optimization problem in (2.5) can be easily solved by singular value thresholding method (Ji & Ye, 2009; Cai et al., 2010).

Suppose $\widehat{\mathbf{\Theta}}$ is the global solution to the optimization problem (2.2). According to the optimality condition, there exists $\mathbf{\Upsilon} \in \partial\|\widehat{\mathbf{\Theta}}\|_*$ such that, for all $\mathbf{\Theta} \in \mathbb{R}^{m_1 \times m_2}$,

$$(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})^\top\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\widehat{\mathbf{\Theta}}) + \lambda\mathbf{\Upsilon}\big) \le 0. \quad (2.6)$$

Hence, based on the optimality condition in (2.6), we measure the suboptimality of a $\mathbf{\Theta} \in \mathbb{R}^{m_1 \times m_2}$ using

$$\begin{aligned} \omega_\lambda(\mathbf{\Theta}) &= \min_{\mathbf{\Upsilon}' \in \partial\|\widehat{\mathbf{\Theta}}\|_*} \max_{\mathbf{\Theta}'} \bigg\{ \frac{(\mathbf{\Theta} - \mathbf{\Theta}')^\top\big(\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}) + \lambda\mathbf{\Upsilon}'\big)}{\|\mathbf{\Theta} - \mathbf{\Theta}'\|_*} \bigg\} \\ &= \min_{\mathbf{\Upsilon}' \in \partial\|\widehat{\mathbf{\Theta}}\|_*} \Big\{ \big\|\nabla\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}) + \lambda\mathbf{\Upsilon}'\big\|_2 \Big\}, \end{aligned}$$

where the second equality follows from the duality between $\|\cdot\|_*$ and $\|\cdot\|_2$. The main idea of the suboptimality is that, if $\boldsymbol{\Theta}$ is an exact optimum, by the optimality condition (2.6), we have $\omega_\lambda(\boldsymbol{\Theta}) < 0$; otherwise, if $\boldsymbol{\Theta}$ is close to the optimum, $\omega_\lambda(\boldsymbol{\Theta})$ is likely to be a small positive value.

To use Algorithm 2, we need to choose an initial optimistic estimate $L_{\min}$ for the Lipschitz constant $L_{\widetilde{\mathcal{L}}_{n,\lambda}}$, such that $0 < L_{\min} \leq L_{\widetilde{\mathcal{L}}_{n,\lambda}}$. The detailed discussion on Lipschitz constant $L_{\widetilde{\mathcal{L}}_{n,\lambda}}$ will be presented in Section 3.

**Algorithm 2** $\{\widetilde{\boldsymbol{\Theta}}, \widehat{L}\} \leftarrow \text{ProxGrad}(\lambda, \widehat{\epsilon}, \boldsymbol{\Theta}^0, L_0)$

```
input $\lambda > 0, \widehat{\epsilon} > 0, \boldsymbol{\Theta}^0 \in \mathbb{R}^{m_1 \times m_2}, L_0 > 0, k = 0$
1: repeat
2:     $k \leftarrow k + 1$
3:     $\{\boldsymbol{\Theta}^k, N_k\} \leftarrow$ LineSearch$(\lambda, \boldsymbol{\Theta}^{k-1}, L_{k-1})$
4:     $L_k \leftarrow \max\{L_{\min}, N_k/2\}$
5: until $\omega_\lambda(\boldsymbol{\Theta}^k) \leq \widehat{\epsilon}$
6: $\widetilde{\boldsymbol{\Theta}} \leftarrow \boldsymbol{\Theta}^k, \widehat{L} \leftarrow L_k$
7: return $\{\widetilde{\boldsymbol{\Theta}}, \widehat{L}\}$
```

Line 3 in Algorithm 2 is the line search algorithm (Algorithm 3), adaptively searching for the best quadratic coefficient $L_k$ for the local quadratic approximation in (2.4).

**Algorithm 3** $\{\boldsymbol{\Theta}, N\} \leftarrow \text{LineSearch}(\lambda, \mathbf{M}, L)$

```
input $\lambda > 0, \boldsymbol{\Theta} \in \mathbb{R}^{m_1 \times m_2}, L > 0$
repeat
    $\boldsymbol{\Theta} \leftarrow \mathcal{T}_{L,\lambda}(\mathbf{M})$
    if $\phi_\lambda(\boldsymbol{\Theta}) > \psi_{L,\lambda}(\boldsymbol{\Theta}; \mathbf{M})$ then
        $L \leftarrow 2L$
    end if
until $\phi_\lambda(\boldsymbol{\Theta}) \leq \psi_{L,\lambda}(\boldsymbol{\Theta}; \mathbf{M})$
$N \leftarrow L$
return $\{\boldsymbol{\Theta}, N\}$
```

Particularly, following the analysis in Xiao & Zhang (2013); Wang et al. (2013b), the iterative solution sequence $\{\boldsymbol{\Theta}^t\}_{t=1}^{K+1}$, which is obtained by Algorithm 1, convergences at geometric rate towards $\widehat{\boldsymbol{\Theta}}$, as defined in (2.2).

## 3. Main Theory

In this section, we are going to present the main theoretical results for the proposed estimator in (2.2). We first lay out the assumptions made on the empirical loss function and the nonconvex penalty.

Suppose the SVD of $\boldsymbol{\Theta}^*$ is $\boldsymbol{\Theta}^* = \mathbf{U}^*\boldsymbol{\Gamma}^*\mathbf{V}^{*\top}$, where $\mathbf{U}^* \in \mathbb{R}^{m_1 \times r}$, $\mathbf{V}^* \in \mathbb{R}^{m_2 \times r}$ and $\boldsymbol{\Gamma}^* = \text{diag}(\boldsymbol{\gamma}_i^*) \in \mathbb{R}^{r \times r}$. We can construct the subspaces $\mathcal{F}$ and $\mathcal{F}^\perp$ as follows

$$\mathcal{F}(\mathbf{U}^*, \mathbf{V}^*) := \{\boldsymbol{\Delta} | \text{row}(\boldsymbol{\Delta}) \subseteq \mathbf{V}^* \text{ and } \text{col}(\boldsymbol{\Delta}) \subseteq \mathbf{U}^*\},$$
$$\mathcal{F}^\perp(\mathbf{U}^*, \mathbf{V}^*) := \{\boldsymbol{\Delta} | \text{row}(\boldsymbol{\Delta}) \perp \mathbf{V}^* \text{ and } \text{col}(\boldsymbol{\Delta}) \perp \mathbf{U}^*\}.$$

Shorthand notations $\mathcal{F}$ and $\mathcal{F}^\perp$ are used whenever $\mathbf{U}^*, \mathbf{V}^*$ are clear from context. It is worth noting that $\mathcal{F}$ is the span of the row and column space of $\boldsymbol{\Theta}^*$, and $\boldsymbol{\Theta}^* \in \mathcal{F}$ consequently. In addition, $\Pi_{\mathcal{F}}(\cdot)$ is the projection operator that projects matrices into the subspace $\mathcal{F}$.

To begin with, we impose two conditions on the empirical loss function $\mathcal{L}_n(\cdot)$ over a restricted set, known as restricted strong convexity (RSC) and restricted strong smoothness (RSS), respectively. Those two assumptions assume that there exist a quadratic lower bound and a quadratic upper bound, respectively, on the remainder of the first order Taylor expansion of $\mathcal{L}_n(\cdot)$. The RSC condition has been discussed extensively in previous work (Negahban et al., 2012; Loh & Wainwright, 2013), which guarantees the strong convexity of the loss function in the restricted set and helps to control the estimation error $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F$. In particular, we define the following subset, which is a cone of a restricted set of directions,

$$\mathcal{C} = \big\{\boldsymbol{\Delta} \in \mathbb{R}^{m_1 \times m_2} \big| \|\Pi_{\mathcal{F}^\perp}(\boldsymbol{\Delta})\|_* \leq 5\|\Pi_{\mathcal{F}}(\boldsymbol{\Delta})\|_*\big\}.$$

**Assumption 3.1** (Restricted Strong Convexity)**.** For operator $\mathfrak{X}$, there exists some $\kappa(\mathfrak{X}) > 0$ such that, for all $\boldsymbol{\Delta} \in \mathcal{C}$,

$$\mathcal{L}_n(\boldsymbol{\Theta} + \boldsymbol{\Delta}) \geq \mathcal{L}_n(\boldsymbol{\Theta}) + \langle \nabla \mathcal{L}_n(\boldsymbol{\Theta}), \boldsymbol{\Delta} \rangle + \kappa(\mathfrak{X})/2 \|\boldsymbol{\Delta}\|_F^2.$$

**Assumption 3.2** (Restricted Strong Smoothness)**.** For operator $\mathfrak{X}$, there exists some $\infty > \rho(\mathfrak{X}) \geq \kappa(\mathfrak{X})$ such that, for all $\boldsymbol{\Delta} \in \mathcal{C}$,

$$\mathcal{L}_n(\boldsymbol{\Theta}) + \langle \nabla \mathcal{L}_n(\boldsymbol{\Theta}), \boldsymbol{\Delta} \rangle + \rho(\mathfrak{X})/2 \|\boldsymbol{\Delta}\|_F^2 \geq \mathcal{L}_n(\boldsymbol{\Theta} + \boldsymbol{\Delta}).$$

Recall that $\mathcal{L}_n(\boldsymbol{\Theta}) = (2n)^{-1}\|\boldsymbol{y} - \mathfrak{X}(\boldsymbol{\Theta})\|_2$. It can be verified that with high probability $\mathcal{L}_n(\boldsymbol{\Theta})$ satisfies both RSC and RSS conditions for different applications, including matrix completion and matrix sensing. We will establish the results for RSC and RSS conditions in Section 3.2.

Further, we impose several regularity conditions on the nonconvex penalty $\mathcal{P}_\lambda(\cdot)$, in terms of the univariate functions $p_\lambda(\cdot)$ and $q_\lambda(\cdot)$.

**Assumption 3.3.**

(i) On the nonnegative real line, there exits a constant $\nu$ that function $p_\lambda(t)$ satisfies $p'_\lambda(t) = 0, \forall\, t \geq \nu > 0$.
(ii) On the nonnegative real line, $q'_\lambda(t)$ is monotone and Lipschitz continuous, *i.e.*, for $t' \geq t$, there exists a constant $\zeta_- \geq 0$ such that $q'_\lambda(t') - q'_\lambda(t) \geq -\zeta_-(t' - t)$.
(iii) Both function $q_\lambda(t)$ and its derivative $q'_\lambda(t)$ pass through the origin, *i.e.*, $q_\lambda(0) = q'_\lambda(0) = 0$.
(iv) On the nonnegative real line, $|q'_\lambda(t)|$ is upper bounded by $\lambda$, *i.e.*, $|q'_\lambda(t)| \leq \lambda$.

Note that condition (ii) is a type of curvature property which determines concavity level of $q_\lambda(\cdot)$, and the nonconvexity level of $p_\lambda(\cdot)$ consequently. These conditions are satisfied by many widely used nonconvex penalties, such as SCAD and MCP. For instance, it is easy to verify that SCAD penalty satisfies the conditions in Assumption 3.3 with $\nu = b\lambda$ and $\zeta_- = 1/(b-1)$; while for MCP, we have those conditions satisfied with $\nu = b\lambda$ and $\zeta_- = 1/b$. Based on Assumption 3.2, if $b$ is chosen such that $\kappa(\mathfrak{X}) > \zeta_-$, it can be shown that the Lipschitz constant is $L_{\widetilde{\mathcal{L}}_{n,\lambda}} = \rho(\mathfrak{X}) - \zeta_-$, and the parameter $L_{\min}$ for Algorithm 1 can be chosen such that $L_{\min} \le \rho(\mathfrak{X}) - \zeta_-$.

### 3.1. Results for the Generic Observation Model

We first present a deterministic error bound of the estimator for the generic observation model, as stated in Theorem 3.4. In particular, our results implies that matrix completion via nonconvex penalty achieves a faster statistical convergence rate than the convex penalty, by taking advantage of large singular values.

**Theorem 3.4** (Deterministic Bound for General Singular Values)**.** Under Assumption 3.1, suppose that $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \in \mathcal{C}$ and the nonconvex penalty $\mathcal{P}_\lambda(\mathbf{\Theta}) = \sum_{i=1}^m p_\lambda\big(\gamma_i(\mathbf{\Theta})\big)$ satisfies Assumption 3.3. Under the condition that $\kappa(\mathfrak{X}) > \zeta_-$, for any optimal solution $\widehat{\mathbf{\Theta}}$ of (2.2) with regularity parameter $\lambda \ge 2\|\mathfrak{X}^*(\boldsymbol{\epsilon})\|_2/n$, it holds that, for $r_1 = |S_1|, r_2 = |S_2|$,

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F \le \underbrace{\frac{\tau\sqrt{r_1}}{\kappa(\mathfrak{X}) - \zeta_-}}_{S_1:\gamma_i^* \ge \nu} + \underbrace{\frac{3\lambda\sqrt{r_2}}{\kappa(\mathfrak{X}) - \zeta_-}}_{S_2:\nu > \gamma_i^* > 0}, \qquad (3.1)$$

where $\tau = \big\|\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2$, where $\mathcal{F}_{S_1}$ is a subspace of $\mathcal{F}$ associated with $S_1$.

It is important to note that the upper bound on the Frobenius norm-based estimation error includes two parts corresponding to different magnitudes of the singular values of the true matrix, *i.e.*, $\gamma_i^*$: (i) $S_1$ corresponds to the set of singular values with larger magnitudes; and (ii) $S_2$ corresponds to the set of singular values with smaller magnitudes. By setting $\zeta_- = \kappa(\mathfrak{X})/2$, we have

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F \le 2\tau\sqrt{r_1}/\kappa(\mathfrak{X}) + 6\lambda\sqrt{r_2}/\kappa(\mathfrak{X}).$$

We can see that provided that $r_1 > 0$, the rate of the proposed estimator is faster than the nuclear norm based one, i.e, $\mathcal{O}\big(\lambda\sqrt{r}/\kappa(\mathfrak{X})\big)$(Negahban & Wainwright, 2011), in light of the fact that $\tau = \big\|\mathbf{\Pi}_{\mathcal{F}_{S_1}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2$ is order of magnitude smaller than $\big\|\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big\|_2 = \lambda$. This would be demonstrated in more detail for specific examples, *i.e.*, matrix completion and matrix sensing, in Section 3.2. In particular, if $\gamma_r^* \ge \nu$, meaning that all the nonzero singular values are larger than $\nu$, the proposed estimator attains the best-case convergence rate of $2\tau\sqrt{r}/\kappa(\mathfrak{X})$.

In Theorem 3.4, we have shown that the convergence rate of nonconvex penalty based estimator is faster than the nuclear norm based one. In the following, we show that under certain assumptions on the magnitudes of the singular values, the estimator in (2.2) enjoys the oracle properties, namely, the obtained M-estimator performs as well as if the underlying model were known beforehand. Before presenting the results on the oracle property, we first formally introduce the oracle estimator,

$$\widehat{\mathbf{\Theta}}_O = \mathop{\mathrm{argmin}}_{\mathbf{\Theta}\in\mathcal{F}(\mathbf{U}^*,\mathbf{V}^*)} \mathcal{L}_n(\mathbf{\Theta}). \qquad (3.2)$$

Remark that the objective function in (3.2) only includes the empirical loss term because the optimization program is constrained in the rank-$r$ subspace $\mathcal{F}(\mathbf{U}^*,\mathbf{V}^*)$. Since it is impossible to get $\mathbf{U}^*, \mathbf{V}^*$ and the rank $r$ in practice, *i.e.*, $\mathcal{F}(\mathbf{U}^*,\mathbf{V}^*)$ is unknown, the oracle estimator defined above is not a practical estimator. We analyze the estimator in (2.2) when $\kappa(\mathfrak{X}) > \zeta_-$, under which condition $\widetilde{\mathcal{L}}_{n,\lambda}(\mathbf{\Theta}) = \mathcal{L}_n(\mathbf{\Theta}) + \mathcal{P}_\lambda(\mathbf{\Theta})$ is strongly convex over the restricted set $\mathcal{C}$ and $\widehat{\mathbf{\Theta}}$ is the unique global optimal solution for the optimization problem. Moreover, the following theorem shows that under suitable conditions, the estimator in (2.2) is identical to the oracle estimator.

**Theorem 3.5** (Oracle Property)**.** Under Assumption 3.1 and 3.2, suppose that $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \in \mathcal{C}$ and $\mathcal{P}_\lambda(\mathbf{\Theta}) = \sum_{i=1}^r p_\lambda(\gamma_i(\mathbf{\Theta}))$ satisfies regularity condition (i), (ii), (iii) in Assumption 3.3. If $\kappa(\mathfrak{X}) > \zeta_-$ and $\boldsymbol{\gamma}^*$ satisfies the condition that

$$\min_{i\in S}\big|(\boldsymbol{\gamma}^*)_i\big| \ge \nu + \frac{2\sqrt{r}\|\mathfrak{X}^*(\boldsymbol{\epsilon})\|_2}{n\kappa(\mathfrak{X})}, \qquad (3.3)$$

where $S = \mathrm{supp}(\boldsymbol{\gamma}^*)$. For the estimator in (2.2) with choice of regularization parameter $\lambda \ge 2n^{-1}\big\|\mathfrak{X}^*(\boldsymbol{\epsilon})\big\|_2 + 2n^{-1}\sqrt{r}\rho(\mathfrak{X})\big\|\mathfrak{X}^*(\boldsymbol{\epsilon})\big\|_2/\kappa(\mathfrak{X})$, we have that $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}}_O$, indicating $\mathrm{rank}(\widehat{\mathbf{\Theta}}) = \mathrm{rank}(\widehat{\mathbf{\Theta}}_O) = \mathrm{rank}(\mathbf{\Theta}^*) = r$. Moreover, we have,

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F \le 2\sqrt{r}\tau/\kappa(\mathfrak{X}), \qquad (3.4)$$

where $\tau = \big\|\Pi_{\mathcal{F}}\big(\nabla\mathcal{L}_n(\mathbf{\Theta}^*)\big)\big\|_2$.

Theorem 3.5 implies that, with a suitable choice of regularization parameter $\lambda$, if the magnitude of the smallest nonzero singular value is sufficiently large, *i.e.*, satisfying (3.3), the proposed estimator in (2.2) is identical to the oracle estimator. This is a very strong result because we do not even know the subspace $\mathcal{F}$. The direct consequence is that the M-estimator exactly recovers the rank of the true matrix, $\mathbf{\Theta}^*$. Moreover, as Theorem 3.5 is a specific case of Theorem 3.4 with $r_1 = r$, we immediately have that the convergence rate in Theorem 3.5 corresponds to the best-case convergence rate in (3.1), which is identical to the statistical rate of the oracle estimator.

## 3.2. Results for Specific Examples

The deterministic results in Theorem 3.4 and Theorem 3.5 are fairly abstract in nature. In what follows, we consider the two specific examples of low-rank matrix estimation as in Section 2.2, and show how the results obtained so far yield concrete and interpretable results. More importantly, we rigorously demonstrate the improvement of the proposed estimator on statistical convergence rate over the traditional one with nuclear norm penalty. More results on oracle property can be found in Appendix, Section E.

### 3.2.1. Matrix Completion

We first analyze the example of matrix completion, as discussed earlier in Example 2.1. It is worth noting that under a suitable condition on spikiness ratio[1], we can establish the restricted strongly convexity, as stated in Assumption 3.1.

**Corollary 3.6.** Suppose that $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \in \mathcal{C}$, the nonconvex penalty $\mathcal{P}_\lambda(\mathbf{\Theta})$ satisfies Assumption 3.3, and $\mathbf{\Theta}^*$ satisfies spikiness assumption, *i.e.*, $\|\mathbf{\Theta}^*\|_\infty \le \alpha^*$, then for any optimal solution $\widehat{\mathbf{\Theta}}$ to the slight modification of (2.2), *i.e.*,

$$\widehat{\mathbf{\Theta}} = \operatorname*{argmin}_{\mathbf{\Theta}\in\mathbb{R}^{m_1\times m_2}} \frac{1}{2n}\big\|\boldsymbol{y} - \mathfrak{X}(\mathbf{\Theta})\big\|_2^2 + \mathcal{P}_\lambda(\mathbf{\Theta}),$$
$$\text{subject to} \quad \|\mathbf{\Theta}\|_\infty \le \alpha^*,$$

there are universal constants $C_1, \ldots, C_5,$ with regularity parameter $\lambda \ge C_3\sigma\sqrt{\log M/(nm)}$ and $\kappa = C_4/(m_1m_2) > \zeta_-$, it holds with probability at least $1 - C_5/M$ that

$$\frac{1}{\sqrt{m_1m_2}}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$$
$$\le \max\{\alpha^*, \sigma\}\bigg[C_1r_1\sqrt{\frac{\log M}{n}} + C_2\sqrt{\frac{r_2M\log M}{n}}\bigg].$$

**Remark 3.7.** Corollary 3.6 is a direct result of Theorem 3.4. Recall the convergence rate[2] of matrix completion with nuclear norm penalty due to Koltchinskii et al. (2011a); Gunasekar et al. (2014), which is as follows

$$\frac{\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F}{\sqrt{m_1m_2}} = \mathcal{O}\bigg(\max\{\alpha^*, \sigma\}\sqrt{\frac{rM\log M}{n}}\bigg). \quad (3.5)$$

It is evident that if $r_1 > 0$, *i.e.*, we have $r_1$ singular values that are larger than $\nu$, the convergence rate obtained by a nonconvex penalty is faster than the one obtained with the convex penalty. In the worst case, when all the singular values are smaller than $\nu$, our result reduced to (3.5) with $r_2 = r$. Meanwhile, if the magnitude of singular values satisfies the condition that $\min_{i\in S}\gamma_i^* \ge v$, *i.e.*, $r_1 = r$ $(S_1 = S)$, the convergence rate of our results is $\mathcal{O}\big(\sqrt{r^2\log M/n}\big)$. In Koltchinskii et al. (2011a); Negahban & Wainwright (2012), the authors proved a minimax lower bound for matrix completion, which is $O(\sqrt{rM/n})$. Our result is not contradictory to the minimax lower bound, because the lower bound is proved for the general class of low rank matrices, while our result takes advantage of the large singular values. In other words, we consider a specific (potentially smaller) class of low rank matrices with both large and small singular values.

### 3.2.2. Matrix Sensing With Dependent Sampling

In the example of matrix sensing, a more general model with dependence among the entries of $\mathbf{X}_i$ is considered. Denote $\mathrm{vec}(\mathbf{X}_i) \in \mathbb{R}^{m_1m_2}$ as the vectorization of $\mathbf{X}_i$. For a symmetric positive definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{m_1m_2\times m_1m_2}$, it is called $\mathbf{\Sigma}$-Ensemble (Negahban & Wainwright, 2011) if the elements of observation matrices $\mathbf{X}_i$'s are sampled from $\mathrm{vec}(\mathbf{X}_i) \sim N(\mathbf{0}, \mathbf{\Sigma})$. Define $\pi^2(\mathbf{\Sigma}) = \sup_{\|\mathbf{u}\|_2=1,\|\mathbf{v}\|_2=1}\mathrm{Var}(\mathbf{u}^\top\mathbf{X}\mathbf{v})$, where $\mathbf{X} \in \mathbb{R}^{m_1\times m_2}$ is a random matrix sampled from the $\mathbf{\Sigma}$-Ensemble. Specifically, when $\mathbf{\Sigma} = \mathbf{I}$, it can be verified that $\pi(\mathbf{I}) = 1$, corresponding to the classical matrix sensing model where the entries of $\mathbf{X}_i$ are independent from each other.

**Corollary 3.8.** Suppose that $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \in \mathcal{C}$ and the nonconvex penalty $\mathcal{P}_\lambda(\mathbf{\Theta})$ satisfies Assumption 3.3, if the random design matrix $\mathbf{X}_i \in \mathbb{R}^{m_1\times m_2}$ is sampled from the $\mathbf{\Sigma}$-ensemble and $\lambda_{\min}(\mathbf{\Sigma})$ is the minimal eigenvalue of $\mathbf{\Sigma}$, there are universal constants $C_1, \ldots, C_6$, such that, if $\kappa(\mathfrak{X}) = C_3\lambda_{\min}(\mathbf{\Sigma}) > \zeta_-$ for any optimal solution $\widehat{\mathbf{\Theta}}$ of (2.2) with $\lambda \ge C_4\sigma\pi(\mathbf{\Sigma})\big(\sqrt{m_1/n} + \sqrt{m_2/n}\big)$, it holds with probability at least $1 - C_5\exp\big(-C_6(m_1 + m_2)\big)$ that

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F \le \frac{\sigma\pi(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})\sqrt{n}}\big[C_1r_1 + C_2\sqrt{r_2M}\big].$$

**Remark 3.9.** Similarly, Corollary 3.8 is a direct consequence of Theorem 3.4. The problem has been studied by (Negahban & Wainwright, 2011) via convex relaxation, with the following estimator error bound

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F = \mathcal{O}\bigg(\frac{\sigma\pi(\mathbf{\Sigma})\sqrt{r_2M}}{\lambda_{\min}(\mathbf{\Sigma})\sqrt{n}}\bigg). \quad (3.6)$$

When there are $r_1 > 0$ singular values that are larger than $\nu$, the result obtained in Corollary 3.8 implies that the convergence rate of the proposed estimator is faster than (3.6). When $r_1 = r$, we obtain the best-case convergence rate of $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F = \mathcal{O}\big(\sigma\pi(\mathbf{\Sigma})r/(\sqrt{n}\lambda_{\min}(\mathbf{\Sigma}))\big)$. In the worst case, when $r_1 = 0$ and $r_2 = r$, the results in Corollary 3.8 reduce to (3.6).

[1] It is insufficient to recover the low-rank matrices due to its infeasibility of recovering overly "spiky" matrices which has very few large entries. Additional assumption on spikiness ratio is needed. Details on spikiness are given in Appendix, Section E.1.

[2] Similar statistical convergence rate was obtained in Negahban & Wainwright (2012) for nonuniform sampling schema.

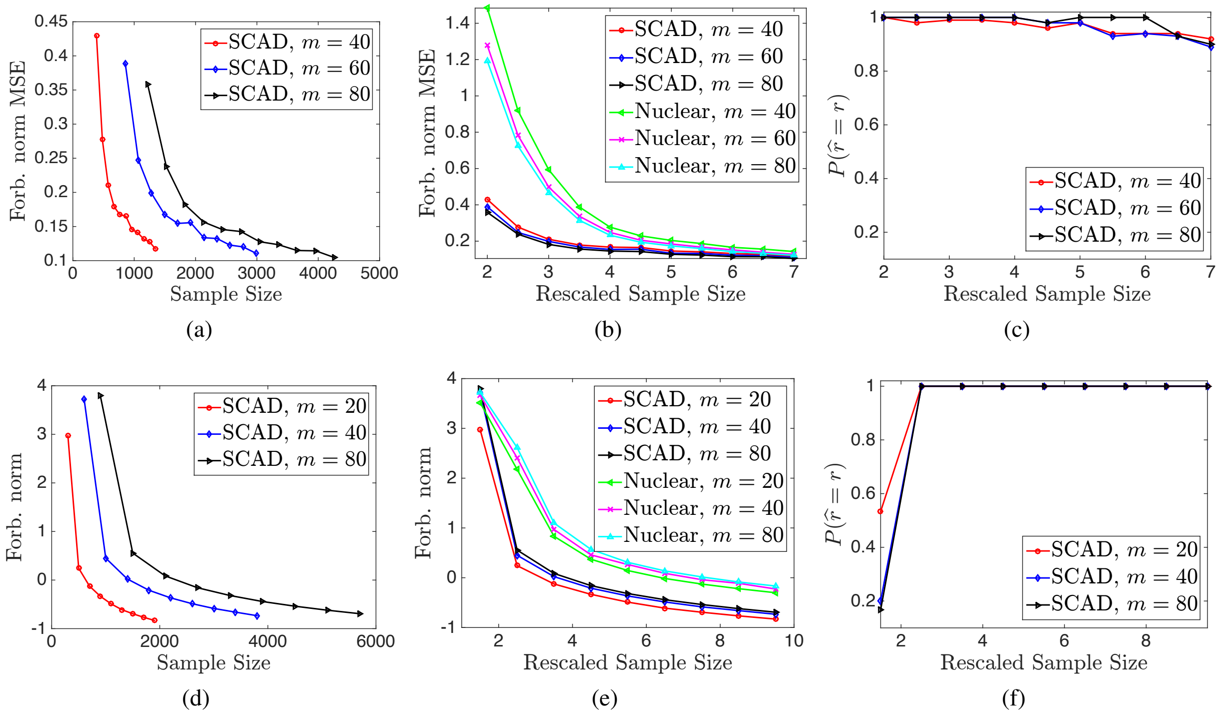

*Figure 1.* Simulation Results for Matrix Completion and Matrix Sensing with SCAD penalty. The size of matrix is $m \times m$. Figure 1(a)-1(c) correspond to matrix completion, with the rank $r = \lfloor \log^2 m \rfloor$, where the rescaled sample size is $N = n/(rm \log m)$. Figure 1(d)-1(f) correspond to matrix sensing, with the rank $r = 10$, where the rescaled sample size is $N = n/(rm)$.

## 4. Numerical Experiments

In this section, we study the performance of the proposed estimator by various simulations and numerical experiments on real-word datasets. It it worth noting that we study the proposed estimator with $\zeta_- < \kappa(\mathfrak{X})$, which can be attained by setting $b = 1 + 2/\kappa(\mathfrak{X})$ for the SCAD penalty. Similarly, the parameter for MCP penalty can be set that $b = 2/\kappa(\mathfrak{X})$.

### 4.1. Simulations

The simulation results demonstrate the close agreement between theoretical upper bound and the numerical behavior of the M-estimator. Simulations are performed for both matrix completion and matrix sensing. In both cases, we solved instances of optimization problem (2.2) for a square matrix $\mathbf{\Theta}^* \in \mathbb{R}^{m \times m}$. For $\mathbf{\Theta}^*$ with rank $r$, we generate $\mathbf{\Theta}^* = \mathbf{ABC}^\top$, where $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{m \times m}$ are the left and right singular vectors of a random matrix, and set $\mathbf{B}$ to be a diagonal matrix with $r$ nonzero entries, and the magnitude of each nonzero entries is above $\nu = \lambda b$, *i.e.*, $r_1 = r$. The regularization parameter $\lambda$ is chosen based on theoretical results with $\sigma^2$ assumed to be known.

In the following, we report detailed results on the estimation errors of the obtained estimators and the probability of exactly recovering the true rank (oracle property). Due to space limitation, we include the simulation results using MCP in the appendix.

**Matrix Completion.** We study the performance of estimators with both convex and nonconvex penalties for $m \in \{40, 60, 80\}$, and the rank $r = \lfloor \log^2 m \rfloor$. $\mathbf{X}_i$'s are uniformed sampled over $\mathcal{X}$, with the variance of observation noise $\sigma^2 = 0.25$. For every configuration, we repeat 100 trials and compute the averaged mean squared Frobenius norm error $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F^2/m^2$ over all trials.

Figure 1(a)-1(c) summarize the results for matrix completion. Particularly, Figure 1(a) plots the mean-squared Frobenius norm error versus the raw sample size, which shows the consistency that estimation error decreases when sample size increases, while Figure 1(b) plots the MSE against the *rescaled sample size* $N = n/(rm \log m)$. It is clearly shown in Figure 1(b) that, in terms of estimation error, the proposed estimator with SCAD penalty outperforms the one with nuclear norm, which aligns with our theoretical analysis. Finally, the probability of exactly recovering the rank of underlying matrix is plotted in Figure 1(c), which indicates that with high probability the rank of underlying matrix can be exactly recovered.

**Matrix Sensing.** For matrix sensing, we set the rank $r = 10$ for all $m \in \{20, 40, 80\}$. $\mathbf{\Theta}^*$ is generated similarly as in matrix completion. We set the observation noise variance $\sigma^2 = 1$ and $\mathbf{\Sigma} = \mathbf{I}$, *i.e.*, the entries of $\mathbf{X}_i$ are independent. Each setting is repeated for 100 times.

*Table 1.* Results on image recovery in terms of RMSE ($\times 10^{-2}$, mean $\pm$ std).

| IMAGE | SVP | SOFTIMPUTE | ALTMIN | TNC | R1MP | NUCLEAR | SCAD | MCP |
|---|---|---|---|---|---|---|---|---|
| LENNA | $3.84 \pm 0.02$ | $4.58 \pm 0.02$ | $4.43 \pm 0.11$ | $5.49 \pm 0.62$ | $3.91 \pm 0.03$ | $5.05 \pm 0.17$ | $2.79 \pm 0.02$ | $2.81 \pm 0.04$ |
| BARBARA | $4.49 \pm 0.04$ | $5.23 \pm 0.03$ | $5.05 \pm 0.05$ | $6.57 \pm 0.92$ | $4.71 \pm 0.06$ | $6.48 \pm 0.53$ | $4.74 \pm 0.02$ | $4.73 \pm 0.03$ |
| CLOWN | $3.75 \pm 0.03$ | $4.43 \pm 0.05$ | $5.44 \pm 0.41$ | $6.92 \pm 1.89$ | $3.89 \pm 0.05$ | $3.70 \pm 0.24$ | $2.77 \pm 0.01$ | $2.81 \pm 0.01$ |
| CROWD | $4.49 \pm 0.04$ | $5.35 \pm 0.07$ | $4.78 \pm 0.09$ | $7.44 \pm 1.23$ | $4.88 \pm 0.06$ | $4.44 \pm 0.18$ | $3.64 \pm 0.07$ | $3.68 \pm 0.09$ |
| GIRL | $3.35 \pm 0.03$ | $4.12 \pm 0.03$ | $5.01 \pm 0.66$ | $4.51 \pm 0.52$ | $3.06 \pm 0.02$ | $4.77 \pm 0.34$ | $2.06 \pm 0.01$ | $2.05 \pm 0.02$ |
| MAN | $4.42 \pm 0.04$ | $5.17 \pm 0.03$ | $5.17 \pm 0.17$ | $6.01 \pm 0.62$ | $4.61 \pm 0.03$ | $5.44 \pm 0.45$ | $3.42 \pm 0.04$ | $3.40 \pm 0.02$ |

*Table 2.* Recommendation results measured in term of the averaged RMSE.

| DATASET | SVP | SOFTIMPUTE | ALTMIN | TNC | R1MP | NUCLEAR | SCAD | MCP |
|---|---|---|---|---|---|---|---|---|
| JESTER1 | 4.7318 | 5.1211 | 4.8562 | 4.4803 | 4.3401 | 4.6910 | 4.1721 | 4.1719 |
| JESTER2 | 4.7712 | 5.1523 | 4.8712 | 4.4511 | 4.3721 | 4.5597 | 4.2002 | 4.1987 |
| JESTER3 | 8.7439 | 5.4532 | 9.5230 | 4.6712 | 4.9803 | 5.1231 | 4.6729 | 4.6740 |

Figure 1(d)-1(f) correspond to results of matrix sensing. The Frobenius norm $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ is reported in log scale. Figure 1(d) demonstrate how the estimation errors scale with $m$ and $n$, which aligns well with our theoretical findings. Also, as observed in Figure 1(e), the estimator with SCAD penalty has lower error bounds compared with the one of nuclear norm penalty. At last, it shows in Figure 1(f) that, empirically, the underlying rank is perfectly recovered by the nonconvex estimator when $n$ is sufficiently large ($n \geq 3rm$).

### 4.2. Experiments on Real World Datasets

In this section, we apply our proposed matrix completion estimator to two real-world applications, image inpainting and collaborative filtering, and compare it with some existing methods, including singular value projection (SVP) (Jain et al., 2010), Trace Norm Constraint (TNC) (Jaggi & Sulovský, 2010), alternating minimization (AltMin) (Jain et al., 2013), spectral regularization algorithm (SoftImpute) (Mazumder et al., 2010), rank-one matrix pursuit (R1MP) (Wang et al., 2014), and nuclear norm penalty (Negahban & Wainwright, 2011).

**Image Inpainting** We select 6 images [3] to test the performance of different algorithms. The matrices corresponding to selected images are of the size $512 \times 512$. We project the underlying matrices into the corresponding subspaces associated with the top $r = 200$ singular values of each matrix, by which we can guarantee that the problem being solved is a low-rank one. In addition, we randomly select 50% of the entries as observations. Each trial is repeated 10 times. The performance is measured by *root mean square error* (RMSE) (Jaggi & Sulovský, 2010; Shalev-Shwartz et al., 2011), summarized in Table 1. As shown in Table 1, the estimators obtained with nonconvex penalties, including SCAD penalty and MCP, achieve the best performance, and significantly outperform the other algorithms on all pictures, except for Barbara. It is worth noting that due to the similar properties of MCP and SCAD, the results of SCAD and MCP are comparable. Moreover, the estimators with nonconvex penalties have smaller RMSE for all pictures, compared with the nuclear norm based estimator, which backs up our theoretical analysis, and the improvement is significant compared with some specific algorithms.

**Collaborative Filtering** Considering the matrix completion algorithms for recommendations, we demonstrate using three datasets: Jester1[4], Jester2 and Jester3, which contain rating data of users on jokes, with real-valued rating scores ranging from $-10.0$ to 10.0. The sizes of these matrices are $\{24983, 23500, 24983\} \times 100$, containing $10^6$, $10^6$, $6 \times 10^5$ ratings, respectively. We randomly select 50% of the ratings as observations, and make predictions over the remaining 50%. Each run is repeated for 10 times. According to the numerical results summarized in Table 2, we observe that the proposed estimators (SCAD, MCP) have the best performance among all existing algorithms. In particular, the estimator with nonconvex penalties (*i.e.*, MCP, SCAD) is better than the estimator with nuclear norm penalty, which agrees well with the results obtained. Comparable results of MCP and SCAD are observed.

## 5. Conclusions

In this paper, we proposed a unified framework for low-rank matrix estimation with nonconvex penalty for a generic observation model. Our work serves as the bridge to connect practical applications of nonconvex penalty and theoretical analysis. Our theoretical results indicate that the convergence rate of estimators with nonconvex penalties is faster than the one with the convex penalty by taking advantage of the large singular values. In addition, we showed that the proposed estimator enjoys the oracle property when a mild condition on the magnitude of singular values is imposed. Extensive experiments demonstrate the close agreement between theoretical analysis and numerical behavior of the proposed estimator.

[3] The images can be downloaded from `http://www.utdallas.edu/~cxc123730/mh_bcs_spl.html`.

[4] The Jester dataset can be downloaded from `http://eigentaste.berkeley.edu/dataset/`.

**Acknowledgements**

We thank Zhaoran Wang and Cheng Wang for helpful comments on an earlier version of this manuscript. Research was sponsored in part by Quanquan Gu's startup funding at Department of Systems and Information Engineering, University of Virginia. Research was also sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1017362, IIS-1320617, and IIS-1354329, HDTRA1-10-1-0120, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the U.S. Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

## References

Cai, Jian-Feng, Candès, Emmanuel J, and Shen, Zuowei. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Cai, T. Tony and Zhou, Wenxin. Matrix completion via max-norm constrained optimization. *arXiv preprint arXiv:1303.0341*, 2013.

Candès, Emmanuel J. and Recht, Benjamin. Exact matrix completion via convex optimization. *Commun. ACM*, 55 (6):111–119, 2012.

Candès, Emmanuel J. and Tao, Terence. The power of convex relaxation: near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Fan, Jianqing and Li, Runze. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360, 2001.

Gunasekar, Suriya, Ravikumar, Pradeep, and Ghosh, Joydeep. Exponential family matrix completion under structural constraints. In *ICML*, pp. 1917–1925, 2014.

Hardt, Marcus. Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 651–660. IEEE, 2014.

Hu, Yao, Zhang, Debing, Ye, Jieping, Li, Xuelong, and He, Xiaofei. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(9):2117–2130, 2013.

Jaggi, Martin and Sulovský, Marek. A simple algorithm for nuclear norm regularized problems. In *ICML*, pp. 471–478, 2010.

Jain, Prateek and Netrapalli, Praneeth. Fast exact matrix completion with finite samples. *arXiv preprint arXiv:1411.1087*, 2014.

Jain, Prateek, Meka, Raghu, and Dhillon, Inderjit S. Guaranteed rank minimization via singular value projection. In *NIPS*, pp. 937–945, 2010.

Jain, Prateek, Netrapalli, Praneeth, and Sanghavi, Sujay. Low-rank matrix completion using alternating minimization. In *Symposium on Theory of Computing Conference*, pp. 665–674, 2013.

Ji, Shuiwang and Ye, Jieping. An accelerated gradient method for trace norm minimization. In *ICML*, pp. 457–464, 2009.

Koltchinskii, Vladimir, Lounici, Karim, Tsybakov, Alexandre B, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011a.

Koltchinskii, Vladimir et al. Von neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics*, 39(6):2936–2973, 2011b.

Liu, Dehua, Zhou, Tengfei, Qian, Hui, Xu, Congfu, and Zhang, Zhihua. A nearly unbiased matrix completion approach. In *ECML*, pp. 210–225, 2013.

Loh, Po-Ling and Wainwright, Martin J. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *NIPS*, pp. 476–484, 2013.

Lu, Canyi, Tang, Jinhui, Yan, Shuicheng, and Lin, Zhouchen. Generalized nonconvex nonsmooth low-rank minimization. In *2014 IEEE Conference on CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 4130–4137, 2014.

Mazumder, Rahul, Hastie, Trevor, and Tibshirani, Robert. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

Negahban, Sahand and Wainwright, Martin J. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2): 1069–1097, 04 2011.

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

Negahban, Sahand N., Ravikumar, Pradeep, Wainwright, Martin J., and Yu, Bin. A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 11 2012.

Nie, Feiping, Wang, Hua, Cai, Xiao, Huang, Heng, and Ding, Chris H. Q. Robust matrix completion via joint schatten p-norm and lp-norm minimization. In *ICDM*, pp. 566–574, 2012.

Recht, Benjamin, Fazel, Maryam, and Parrilo, Pablo A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Rohde, Angelika, Tsybakov, Alexandre B, et al. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

Shalev-Shwartz, Shai, Gonen, Alon, and Shamir, Ohad. Large-scale convex minimization with a low-rank constraint. In *ICML*, pp. 329–336, 2011.

Srebro, Nathan and Shraibman, Adi. Rank, trace-norm and max-norm. In *Proceedings of the 18th Annual Conference on Learning Theory*, pp. 545–560. Springer-Verlag, 2005.

Srebro, Nathan, Rennie, Jason D. M., and Jaakkola, Tommi. Maximum-margin matrix factorization. In *NIPS*, pp. 1329–1336, 2004.

Vershynin, Roman. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wang, Shusen, Liu, Dehua, and Zhang, Zhihua. Nonconvex relaxation approaches to robust matrix recovery. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 2013a.

Wang, Zhaoran, Liu, Han, and Zhang, Tong. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *arXiv preprint arXiv:1306.4960*, 2013b.

Wang, Zheng, Lai, Ming-Jun, Lu, Zhaosong, Fan, Wei, Davulcu, Hasan, and Ye, Jieping. Rank-one matrix pursuit for matrix completion. In *ICML*, pp. 91–99, 2014.

Weyl, Hermann. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen. *Mathematische Annalen*, 71(4):441–479, 1912.

Xiao, Lin and Zhang, Tong. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

Yao, Quanming, Kwok, James T, and Zhong, Wenliang. Fast low-rank matrix learning with nonconvex regularization. *arXiv preprint arXiv:1512.00984*, 2015.

Zhang, Cun-Hui. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pp. 894–942, 2010.

Zhang, Cun-Hui, Zhang, Tong, et al. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

Zou, Hui. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101: 1418–1429, December 2006.