A Unified Computational and Statistical Framework for Nonconvex Low-Rank Matrix Estimation

Lingxiao Wang* University of Virginia

Xiao Zhang* University of Virginia

Quanquan Gu University of Virginia

Abstract

We propose a unified framework for estimating low-rank matrices through nonconvex optimization based on gradient descent algorithm. Our framework is quite general and can be applied to both noisy and noiseless observations. In the general case with noisy observations, we show that our algorithm is guaranteed to linearly converge to the unknown low-rank matrix up to a minimax optimal statistical error, provided an appropriate initial estimator. While in the generic noiseless setting, our algorithm converges to the unknown low-rank matrix at a linear rate and enables exact recovery with optimal sample complexity. In addition, we develop a new initialization algorithm to provide the desired initial estimator, which outperforms existing initialization algorithms for nonconvex low-rank matrix estimation. We illustrate the superiority of our framework through three examples: matrix regression, matrix completion, and one-bit matrix completion. We also corroborate our theory through extensive experiments on synthetic data.

1 INTRODUCTION

Low-rank matrix estimation has broad applications in many fields such as collaborative filtering (Srebro et al., 2004). Numerous efforts have been made in order to efficiently estimate the unknown low-rank matrix, among which nuclear norm relaxation based methods (Srebro et al., 2004; Candès and Tao, 2010; Rohde et al., 2011; Recht et al., 2010; Recht, 2011; Negahban and Wainwright, 2011, 2012; Gui and Gu, 2015) are most popular. Although nuclear norm based methods enjoy nice theoretical guarantees for recovering low-rank matrices, the computational complexities of these methods are very high. For example, to estimate a rank-r matrix, most of these algorithms require to compute a rank-r singular value decomposition per-iteration, which is computationally prohibitive for huge matrices. In order to get over such a computational barrier, many recent studies proposed to estimate the unknown low-rank matrix via matrix factorization, or more generally speaking, nonconvex optimization. Specifically, for a rank-r matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, it can be factorized as $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$, $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$, and such a reparametrization automatically enforces the low-rankness of the unknown matrix. While matrix factorization makes the optimization problem nonconvex, it can significantly improve the computational efficiency. A series of work (Jain et al., 2013; Zhao et al., 2015; Chen and Wainwright, 2015; Zheng and Lafferty, 2015; Tu et al., 2015; Bhojanapalli et al., 2015; Park et al., 2016a,b) has been carried out to analyze different nonconvex optimization algorithms for various low-rank matrix estimation problems.

In this paper, we propose a unified framework for nonconvex low-rank matrix estimation, which integrates both optimization-theoretic and statistical analyses. Instead of considering specific low-rank matrix estimation problems, we consider general ones, which correspond to optimizing a family of loss functions which satisfies restricted strong convexity and smoothness conditions (Negahban et al., 2009). We highlight our major contributions as follows:

1. We propose a general algorithm, which is applicable to both low-rank matrix estimation with noisy observations and that with noiseless observations. We establish the linear convergence rate to the unknown low-rank matrix for our algorithm. In particular, for noisy observations, our algorithm achieves statistical

*Equal Contribution

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

error that matches the minimax lower bound (Negahban and Wainwright, 2012; Koltchinskii et al., 2011). While in the noiseless case, our algorithm enables exact recovery of the global optimum (i.e., unknown low-rank matrix) and achieves optimal sample complexity (Recht et al., 2010; Tu et al., 2015).

2. We develop a new and generic initialization algorithm to provide suitable initial estimator. We prove that our initialization procedure relaxes the stringent requirement on condition number of the objective function, assumed in recent studies (Bhojanapalli et al., 2015; Park et al., 2016a,b), thereby resolving an open question in Bhojanapalli et al. (2015).

3. We apply our unified framework to specific problems, such as matrix regression, matrix completion and one-bit matrix completion. We establish the linear convergence rates and optimal statistical error bounds of our method for each examples. We also demonstrate the superiority of our approach over the state-of-the art methods via thorough experiments.

Notation. We use $[d]$ to denote the index set $\{1, 2, \ldots, d\}$. For any index set $\Omega \subseteq [d_1] \times [d_2]$, denote $\Omega_{i,*} = \big\{(i,j) \in \Omega \,\big|\, j \in [d_2]\big\}$, and $\Omega_{*,j} = \big\{(i,j) \in \Omega \,\big|\, i \in [d_1]\big\}$. For any matrix $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$, we denote the i-th row and j-th column of $\mathbf{A}$ by $\mathbf{A}_{i,*}$ and $\mathbf{A}_{*,j}$, respectively. The (i,j)-th entry of $\mathbf{A}$ is denoted by A_{ij}. Denote the ℓ-th largest singular value of $\mathbf{A}$ by $\sigma_\ell(\mathbf{A})$. Let $\mathbf{x} = [x_1, x_2, \cdots, x_d]^\top \in \mathbb{R}^d$ be a d-dimensional vector. For $0 < q < \infty$, denote the ℓ_q vector norm by $\|\mathbf{x}\|_q = (\Sigma_{i=1}^d |x_i|^q)^{1/q}$. As usual, let $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|_2$ be the Frobenius norm and the spectral norm of matrix $\mathbf{A}$, respectively. The element-wise infinity norm of $\mathbf{A}$ is defined as $\|\mathbf{A}\|_\infty = \max_{i,j} |A_{ij}|$. Besides, we define the largest ℓ_2 norm of its rows as $\|\mathbf{A}\|_{2,\infty} = \max_i \|\mathbf{A}_{i,*}\|_2$.

2 RELATED WORK

In recent years, a surge of nonconvex optimization algorithms for estimating low-rank matrices have been established. For example, Jain et al. (2013) analyzed the convergence of alternating minimization approach for matrix regression and matrix completion. Zhao et al. (2015) provided a more unified analysis by proving that, with a reasonable initial solution, a broad class of nonconvex optimization algorithms, including alternating minimization and gradient-type methods, can successfully recover the unknown low-rank matrix. However, they also required a stringent form of the restricted isometry property that is similar to Jain et al. (2013). Recently, Zheng and Lafferty (2015, 2016) analyzed the gradient descent based approach for matrix regression and matrix completion. They showed that their algorithm is guaranteed to converge linearly to the global optimum with an appropriate initial solution, and improves the alternating minimization algorithm in terms of both computational complexity and sample complexity. Tu et al. (2015) provided an improved analysis of matrix regression via gradient descent, compared to Zheng and Lafferty (2015), through a more sophisticated initialization procedure and a refined restricted isometry assumption on the measurements.

The most related work to ours is Chen and Wainwright (2015); Bhojanapalli et al. (2015); Park et al. (2016b). In detail, Chen and Wainwright (2015) proposed a projected gradient descent framework to recover the positive semidefinite low-rank matrices. Although their work can be applied to a wide range of problems, the iteration complexity derived from their optimization framework is very high for many specific examples. Bhojanapalli et al. (2015) proposed a factorized gradient descent algorithm for nonconvex optimization over positive semidefinite matrices. They proved that, when the general empirical loss function is both strongly convex and smooth, their algorithm can recover the unknown low-rank matrix at a linear convergence rate. Built upon Bhojanapalli et al. (2015), Park et al. (2016b) derived the theoretical guarantees of the factorized gradient descent algorithm for rectangular matrix factorization problem under similar conditions. Nevertheless, their analyses (Bhojanapalli et al., 2015; Park et al., 2016b) are limited to the optimization perspective, and do not support the case with noisy observations. Our proposed framework, on one hand, simplifies the conditions of nonconvex low-rank matrix estimation to restricted strong convexity and smoothness, and on the other hand, integrates both optimization-theoretic and statistical analyses. In fact, it achieves the best of both worlds, and provides a simple but powerful toolkit to analyze various low-rank matrix estimation problems. Furthermore, our proposed initialization algorithm relaxes the strict constraint on condition number of the objective function, which is imposed by Bhojanapalli et al. (2015); Park et al. (2016a,b), thereby resolving an open question in Bhojanapalli et al. (2015).

We also note that in order to get rid of the disadvantages of initialization procedure, Bhojanapalli et al. (2016); Park et al. (2016c) proved that for matrix regression, all local minima of the nonconvex optimization based on matrix reparametrization are close to a global optimum under the restricted isometry property assumption. And for positive semidefinite matrix completion, Ge et al. (2016) proved a similar result. However, for general low-rank matrix completion such as one-bit matrix completion, it is still unclear whether the global optimality holds for all local minima.

3 LOW-RANK MATRIX ESTIMATION

In this section, we provide a general problem setup for low-rank matrix estimation, together with several illustrative examples to show the applicability of our general framework.

3.1 General Problem Setup

Let $\mathbf{X}^* \in \mathbb{R}^{d_1 \times d_2}$ be an unknown low-rank matrix with rank r. Our goal is to estimate $\mathbf{X}^*$ through a collection of n observations. Let $\mathcal{L}_n : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ be the sample loss function, which measures the fitness of any matrix $\mathbf{X}$ with respect to the given observations. Thus, the low-rank matrix estimation can be formulated as the following optimization problem

$$\min_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}} \mathcal{L}_n(\mathbf{X}), \text{ subject to } \mathbf{X} \in \mathcal{C}, \ \text{rank}(\mathbf{X}) \leq r,$$

where $\mathcal{C} \subseteq \mathbb{R}^{d_1 \times d_2}$ is a feasible set, such that $\mathbf{X}^* \in \mathcal{C}$.

In order to solve the low-rank matrix estimation problem more efficiently, following Jain et al. (2013); Tu et al. (2015); Zheng and Lafferty (2016); Park et al. (2016a), we reparameterize $\mathbf{X}$ as $\mathbf{U}\mathbf{V}^\top$, and solve the following nonconvex optimization problem

$$\min_{\substack{\mathbf{U} \in \mathbb{R}^{d_1 \times r} \\ \mathbf{V} \in \mathbb{R}^{d_2 \times r}}} \mathcal{L}_n(\mathbf{U}\mathbf{V}^\top), \text{ subject to } \mathbf{U} \in \mathcal{C}_1, \mathbf{V} \in \mathcal{C}_2, \quad (3.1)$$

where $\mathcal{C}_1 \subseteq \mathbb{R}^{d_1 \times r}, \mathcal{C}_2 \subseteq \mathbb{R}^{d_2 \times r}$ are the corresponding rotation-invariant[1] feasible sets implied by $\mathcal{C}$. Suppose $\mathbf{X}^*$ can be decomposed as $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^{*\top}$, we need to ensure that $\mathbf{U}^* \in \mathcal{C}_1$ and $\mathbf{V}^* \in \mathcal{C}_2$.

3.2 Illustrative Examples

Here we briefly introduce matrix regression, matrix completion and one-bit matrix completion as three examples, to demonstrate the applicability of our generic framework.

Matrix Regression. In matrix regression (Recht et al., 2010; Negahban and Wainwright, 2011), our goal is to estimate the unknown rank-r matrix $\mathbf{X}^* \in \mathbb{R}^{d_1 \times d_2}$ based on a set of noisy measurements $\boldsymbol{y} = \mathcal{A}(\mathbf{X}^*) + \boldsymbol{\epsilon}$, where $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^n$ is a linear operator such that $\mathcal{A}(\mathbf{X}^*) = (\langle \mathbf{A}_1, \mathbf{X}^* \rangle, \langle \mathbf{A}_2, \mathbf{X}^* \rangle, \ldots, \langle \mathbf{A}_n, \mathbf{X}^* \rangle)^\top$, and $\boldsymbol{\epsilon}$ is a noise vector with i.i.d. sub-Gaussian entries with parameter ν. Specifically, each random matrix $\mathbf{A}_i \in \mathbb{R}^{d_1 \times d_2}$ has i.i.d. standard normal entries. As discussed before, in order to estimate the low-rank matrix more efficiently, we consider the following nonconvex optimization problem

$$\min_{\substack{\mathbf{U} \in \mathbb{R}^{d_1 \times r} \\ \mathbf{V} \in \mathbb{R}^{d_2 \times r}}} \mathcal{L}_n(\mathbf{U}\mathbf{V}^\top) := \frac{1}{2n} \|\mathbf{y} - \mathcal{A}(\mathbf{U}\mathbf{V}^\top)\|_2^2.$$

Note that here the convex feasible sets $\mathcal{C}_1$ and $\mathcal{C}_2$ in (3.1) are both $\mathbb{R}^{d_1 \times r}$, which give rise to an unconstrained optimization.

Matrix Completion. In the noisy matrix completion (Rohde et al., 2011; Koltchinskii et al., 2011; Negahban and Wainwright, 2012), our goal is to recover the unknown rank-r matrix $\mathbf{X}^* \in \mathbb{R}^{d_1 \times d_2}$ based on a set of randomly observed noisy entries from $\mathbf{X}^*$. For instance, one uniformly observes each entry independently with probability $p \in (0, 1)$. Specifically, we represent these observations by a random matrix $\mathbf{Y} \in \mathbb{R}^{d_1 \times d_2}$ such that

$$Y_{jk} := \begin{cases} X^*_{jk} + Z_{jk}, & \text{with probability } p, \\ *, & \text{otherwise,} \end{cases}$$

where $\mathbf{Z} = (Z_{jk}) \in \mathbb{R}^{d_1 \times d_2}$ is a noise matrix with i.i.d. entries, such that each entry Z_{jk} follows sub-Gaussian distribution with parameter ν. Let $\Omega \subseteq [d_1] \times [d_2]$ be the index set of the observed entries, then we can estimate the low-rank matrix $\mathbf{X}^*$ by solving the following nonconvex optimization problem

$$\min_{\substack{\mathbf{U} \in \mathbb{R}^{d_1 \times r} \\ \mathbf{V} \in \mathbb{R}^{d_2 \times r}}} \mathcal{L}_\Omega(\mathbf{U}\mathbf{V}^\top) := \frac{1}{2p} \sum_{(j,k) \in \Omega} (\mathbf{U}_{j*}\mathbf{V}_{k*}^\top - Y_{jk})^2,$$

where $p = |\Omega|/(d_1 d_2)$. Here the feasible sets $\mathcal{C}_1$ and $\mathcal{C}_2$ in (3.1) are defined as follow $\mathcal{C}_i = \big\{ \mathbf{A} \in \mathbb{R}^{d_i \times r} \,\big|\, \|\mathbf{A}\|_{2,\infty} \leq \gamma \big\}$, where $i \in \{1, 2\}$, and $\gamma > 0$ is a constant, which will be defined in later analysis.

One-Bit Matrix Completion. In one-bit matrix completion (Davenport et al., 2014; Cai and Zhou, 2013), we observe the sign of a random subset of noisy entries from the unknown rank-r matrix $\mathbf{X}^* \in \mathbb{R}^{d_1 \times d_2}$, instead of observing the actual entries. In particular, we consider one-bit matrix completion problem under the uniform random sampling model (Davenport et al., 2014; Cai and Zhou, 2013; Ni and Gu, 2016). Given a differentiable function $f : \mathbb{R} \to [0, 1]$ and an index set $\Omega \subseteq [d_1] \times [d_2]$, we observe the corresponding set of entries from a binary matrix $\mathbf{Y}$ according to the following probabilistic model:

$$Y_{jk} = \begin{cases} +1, & \text{with probability } f(X^*_{jk}), \\ -1, & \text{with probability } 1 - f(X^*_{jk}). \end{cases} \quad (3.2)$$

If f is the cumulative distribution function of $-Z_{jk}$, where $\mathbf{Z} = (Z_{jk}) \in \mathbb{R}^{d_1 \times d_2}$ is a noise matrix with i.i.d.

[1] We say $\mathcal{C}_1$ is rotation-invariant, if for any $\mathbf{A} \in \mathcal{C}_1$, $\mathbf{A}\mathbf{R} \in \mathcal{C}_1$,where $\mathbf{R}$ is an arbitrary r-by-r orthogonal matrix.

entries, then we can rewrite the above model as

$$Y_{jk} = \begin{cases} +1, & \text{if } X^*_{jk} + Z_{jk} > 0, \\ -1, & \text{if } X^*_{jk} + Z_{jk} < 0. \end{cases} \tag{3.3}$$

One widely-used function is the logistic function $f(X_{jk}) = e^{X_{jk}}/(1+e^{X_{jk}})$, which is equivalent to the fact that Z_{jk} in (3.3) follows the standard logistic distribution. Given the function f, the objective loss function for one-bit matrix completion is given by

$$\mathcal{L}_\Omega(\mathbf{X}) := -\frac{1}{p}\sum_{(j,k)\in\Omega}\Big\{\mathbb{1}\big\{Y_{jk}=1\big\}\log\big(f(X_{jk})\big) + \mathbb{1}\big\{Y_{jk}=-1\big\}\log\big(1-f(X_{jk})\big)\Big\},$$

where $p = |\Omega|/(d_1 d_2)$. Similar to the previous case, we can efficiently estimate $\mathbf{X}^*$ by solving a nonconvex optimization problem through matrix factorization.

4 THE PROPOSED ALGORITHM

In this section, we propose an optimization algorithm to solve (3.1) based on gradient descent. It is important to note that the optimal solution to (3.1) is not unique. To be specific, for any solution $(\mathbf{U}, \mathbf{V})$ to the optimization problem (3.1), $\big(\mathbf{UP}, \mathbf{V}(\mathbf{P}^{-1})^\top\big)$ is also a valid solution, where $\mathbf{P} \in \mathbb{R}^{r\times r}$ can be any invertible matrix. In order to address this issue, following Tu et al. (2015); Zheng and Lafferty (2016); Park et al. (2016b), we consider the following optimization problem, which has an additional regularizer to force the two factors to be balanced:

$$\min_{\mathbf{U}\in\mathcal{C}_1,\mathbf{V}\in\mathcal{C}_2} \mathcal{L}_n(\mathbf{UV}^\top) + \frac{1}{8}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2. \tag{4.1}$$

We propose a gradient descent algorithm to solve the proposed estimator in (4.1), which is displayed in Algorithm 1.

Algorithm 1 Gradient Descent (GD)

Input: Loss function $\mathcal{L}_n$, step size η, number of iterations T, initial solutions $\mathbf{U}^0, \mathbf{V}^0$.
  for: $t = 0, 1, 2, \ldots, T-1$ **do**
    $\mathbf{U}^{t+1} = \mathbf{U}^t - \eta\big(\nabla_U\mathcal{L}_n(\mathbf{U}^t\mathbf{V}^{t\top}) - \frac{1}{2}\mathbf{U}^t(\mathbf{U}^{t\top}\mathbf{U}^t - \mathbf{V}^{t\top}\mathbf{V}^t)\big)$
    $\mathbf{V}^{t+1} = \mathbf{V}^t - \eta(\nabla_V\mathcal{L}_n(\mathbf{U}^t\mathbf{V}^{t\top}) - \frac{1}{2}\mathbf{V}^t(\mathbf{V}^{t\top}\mathbf{V}^t - \mathbf{U}^{t\top}\mathbf{U}^t)\big)$
    $\mathbf{U}^{t+1} = \mathcal{P}_{\mathcal{C}_1}(\mathbf{U}^{t+1})$
    $\mathbf{V}^{t+1} = \mathcal{P}_{\mathcal{C}_2}(\mathbf{V}^{t+1})$
  end for
Output: $\mathbf{X}^T = \mathbf{U}^T\mathbf{V}^{T\top}$

Here $\mathcal{P}_{\mathcal{C}_i}$ denotes the projection operator onto the feasible set $\mathcal{C}_i$, where $i \in \{1, 2\}$. Algorithm 1 is more general than Tu et al. (2015); Zheng and Lafferty (2015, 2016), because it applies to a larger family of loss functions. Therefore, various low-rank matrix estimation problems including those examples discussed in Section 3.2 can be solved by Algorithm 1. Compared with the algorithm proposed by Park et al. (2016b), we include a projection step to ensure the estimators lie in a feasible set, which is essential for many low-rank matrix recovery problems such as matrix completion and one-bit matrix completion.

As will be seen in our theoretical analysis, it is guaranteed to converge to the unknown parameters $\mathbf{U}^*$ and $\mathbf{V}^*$, only if the initial solutions $\mathbf{U}^0$ and $\mathbf{V}^0$ are sufficiently close to $\mathbf{U}^*$ and $\mathbf{V}^*$. Thus, inspired by Jain et al. (2010), we propose an initialization algorithm, which is displayed in Algorithm 2, to satisfy this requirement. For any matrix $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$, we denote its rank-$r$ singular value decomposition by $\mathrm{SVD}_r(\mathbf{X})$. Moreover, if $\mathrm{SVD}_r(\mathbf{X}) = [\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}]$, then we denote the best rank-r approximation of $\mathbf{X}$ by $\mathcal{P}_r(\mathbf{X}) = \mathbf{U\Sigma V}^\top$, where $\mathcal{P}_r$ is a projection operator onto the rank-r matrix subspace.

Algorithm 2 Initialization

Input: Loss function $\mathcal{L}_n$, parameter τ, number of iterations S.
  $\mathbf{X}_0 = \mathbf{0}$
  for: $s = 1, 2, 3, \ldots, S$ **do**
    $\mathbf{X}_s = \mathcal{P}_r\big(\mathbf{X}_{s-1} - \tau\nabla\mathcal{L}_n(\mathbf{X}_{s-1})\big)$
  end for
  $[\overline{\mathbf{U}}^0, \mathbf{\Sigma}^0, \overline{\mathbf{V}}^0] = \mathrm{SVD}_r(\mathbf{X}_S)$
  $\mathbf{U}^0 = \overline{\mathbf{U}}^0(\mathbf{\Sigma}^0)^{1/2}$, $\mathbf{V}^0 = \overline{\mathbf{V}}^0(\mathbf{\Sigma}^0)^{1/2}$
Output: $\mathbf{U}^0, \mathbf{V}^0$

5 MAIN THEORY

In this section, we are going to present our main theoretical results for the proposed algorithms. To begin with, we introduce some notations and facts to simplify our proof.

Let the singular value decomposition (SVD) of $\mathbf{X}^*$ be $\mathbf{X}^* = \overline{\mathbf{U}}^*\mathbf{\Sigma}^*\overline{\mathbf{V}}^{*\top}$, where $\overline{\mathbf{U}}^* \in \mathbb{R}^{d_1\times r}$, $\overline{\mathbf{V}}^* \in \mathbb{R}^{d_2\times r}$ are orthonormal matrices, and $\mathbf{\Sigma}^* \in \mathbb{R}^{r\times r}$ is a diagonal matrix. Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$ be the sorted nonzero singular values of $\mathbf{X}^*$, and denote the condition number of $\mathbf{X}^*$ by κ, i.e., $\kappa = \sigma_1/\sigma_r$. Besides, let $\mathbf{U}^* = \overline{\mathbf{U}}^*(\mathbf{\Sigma}^*)^{1/2}$ and $\mathbf{V}^* = \overline{\mathbf{V}}^*(\mathbf{\Sigma}^*)^{1/2}$, then following Tu et al. (2015); Zheng and Lafferty (2016), we can lift the low-rank matrix $\mathbf{X}^* \in \mathbb{R}^{d_1\times d_2}$ to a positive semidefinite matrix $\mathbf{Y}^* \in \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}$ in higher dimension

$$\mathbf{Y}^* = \begin{bmatrix} \mathbf{U}^*\mathbf{U}^{*\top} & \mathbf{U}^*\mathbf{V}^{*\top} \\ \mathbf{V}^*\mathbf{U}^{*\top} & \mathbf{V}^*\mathbf{V}^{*\top} \end{bmatrix} = \mathbf{Z}^*\mathbf{Z}^{*\top},$$

where $\mathbf{Z}^*$ is defined as $\mathbf{Z}^* = [\mathbf{U}^*; \mathbf{V}^*] \in \mathbb{R}^{(d_1+d_2)\times r}$. Observant readers may have already noticed that the symmetric factorization of $\mathbf{Y}^*$ is not unique. In order to address this issue, it is convenient to define a solution set, which can be seen as an equivalent class of the optimal solutions. Thus, we define the solution sets with respect to the true parameter $\mathbf{Z}^*$ as

$$\mathcal{Z} = \Big\{ \mathbf{Z} \in \mathbb{R}^{(d_1+d_2)\times r} \;\Big|\; \mathbf{Z} = \mathbf{Z}^*\mathbf{R} \text{ for some } \mathbf{R} \in \mathbb{Q}_r \Big\},$$

where $\mathbb{Q}_r$ denotes the set of r-by-r orthonormal matrices. Note that for any $\mathbf{Z} \in \mathcal{Z}$, we have $\mathbf{X}^* = \mathbf{Z}_U \mathbf{Z}_V^\top$, where $\mathbf{Z}_U$ and $\mathbf{Z}_V$ denote the top d_1 and bottom d_2 rows of matrix $\mathbf{Z} \in \mathbb{R}^{(d_1+d_2)\times r}$, respectively.

Definition 5.1. Define the estimation error $D(\mathbf{Z}, \mathbf{Z}^*)$ as the minimal Frobenius norm between $\mathbf{Z}$ and $\mathbf{Z}^*$ with respect to the optimal rotation, namely

$$D(\mathbf{Z}, \mathbf{Z}^*) = \min_{\widetilde{\mathbf{Z}} \in \mathcal{Z}} \|\mathbf{Z} - \widetilde{\mathbf{Z}}\|_F = \min_{\mathbf{R} \in \mathbb{Q}_r} \|\mathbf{Z} - \mathbf{Z}^*\mathbf{R}\|_F.$$

Definition 5.2. We denote the local region around optimum $\mathbf{Z}^*$ with radius R as

$$\mathbb{B}(R) = \Big\{ \mathbf{Z} \in \mathbb{R}^{(d_1+d_2)\times r} \;\Big|\; d(\mathbf{Z}, \mathbf{Z}^*) \le R \Big\}.$$

Before we present our main results, we first lay out several necessary conditions regarding the sample loss function $\mathcal{L}_n$. First, we impose two conditions on the sample loss function $\mathcal{L}_n$. These two conditions are known as restricted strong convexity (RSC) and restricted strong smoothness (RSS) conditions (Negahban et al., 2009; Loh and Wainwright, 2013), assuming that there are both quadratic lower bound and upper bound, respectively, on the remaining term of the first order Taylor expansion of $\mathcal{L}_n$.

Condition 5.3 (Restricted Strong Convexity)**.** For a given sample size n, $\mathcal{L}_n$ is restricted strongly convex with parameter μ, such that for all matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d_1\times d_2}$ with rank at most $3r$

$$\mathcal{L}_n(\mathbf{Y}) \ge \mathcal{L}_n(\mathbf{X}) + \langle \nabla \mathcal{L}_n(\mathbf{X}), \mathbf{Y} - \mathbf{X} \rangle + \frac{\mu}{2}\|\mathbf{Y} - \mathbf{X}\|_F^2.$$

Condition 5.4 (Restricted Strong Smoothness)**.** Given a fixed sample size n, $\mathcal{L}_n$ is restricted strongly smooth with parameter L, such that for all matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d_1\times d_2}$ with rank at most $3r$

$$\mathcal{L}_n(\mathbf{Y}) \le \mathcal{L}_n(\mathbf{X}) + \langle \nabla \mathcal{L}_n(\mathbf{X}), \mathbf{Y} - \mathbf{X} \rangle + \frac{L}{2}\|\mathbf{Y} - \mathbf{X}\|_F^2.$$

Both Conditions 5.3 and 5.4 can be verified for the illustrative examples discussed in Section 3.2.

Next, we assume the gradient of the sample loss function at $\mathbf{X}^*$ is upper bounded, in terms of spectral norm.

Condition 5.5. For a given sample size n and tolerance parameter $\delta \in (0, 1)$, we let $\epsilon(n, \delta)$ be the smallest scalar such that with probability at least $1-\delta$, we have

$$\|\nabla \mathcal{L}_n(\mathbf{X}^*)\|_2 \le \epsilon(n, \delta),$$

where $\epsilon(n, \delta)$ depends on sample size n and δ.

Condition 5.5 is essential to derive the statistical error of the estimator returned by our algorithm.

5.1 Results for the Generic Model

In this subsection, we first provide the theoretical guarantees of our proposed algorithm for the generic model, where the sample loss function $\mathcal{L}_n$ satisfies Conditions 5.3, 5.4 and 5.5.

Theorem 5.6 (Gradient Descent)**.** Recall that $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^{*\top}$ is the unknown rank-r matrix. For any $\mathbf{Z}^0 \in \mathbb{B}(c_2\sqrt{\sigma_r})$, where $c_2 \le \min\{1/4, \sqrt{2\mu'/5(4L+1)}\}$, if the sample size n is large enough such that $\epsilon^2(n, \delta) \le c_2^2\mu'\sigma_r^2/(10c_3 r)$, where $\mu' = \min\{\mu, 1\}$ and $c_3 = 2/L + 4/\mu$, then with step size $\eta = c_1/\sigma_1$, where $c_1 \le \min\{1/(64L), 1/32\}$, the estimator at iteration t of Algorithm 1 satisfies

$$D^2(\mathbf{Z}^{t+1}, \mathbf{Z}^*) \le \left(1 - \frac{c_1\mu'}{10\kappa}\right) D^2(\mathbf{Z}^t, \mathbf{Z}^*) + \eta c_3 r \epsilon^2(n, \delta),$$

with probability at least $1 - \delta$. If we let $\rho = 1 - c_1\mu'/(10\kappa)$, then the iterates $\{\mathbf{Z}^t\}_{t=0}^\infty$ satisfy

$$d^2(\mathbf{Z}^t, \mathbf{Z}^*) \le \rho^t d^2(\mathbf{Z}^0, \mathbf{Z}^*) + \frac{10c_3 r}{\mu'\sigma_r}\epsilon^2(n, \delta),$$

with probability at least $1 - \delta$.

Thus, it is sufficient to perform $T = O\big(\kappa \log(1/\epsilon)\big)$ iterations for $\mathbf{Z}^T$ to converge to a close neighborhood of $\mathbf{Z}^*$, where ϵ depends on the statistical error term $r\epsilon^2(n, \delta)$. Note that in Theorem 5.6, the step size η is chosen according to $1/\sigma_1$. In practice, we can set the step size as $\eta = c'/\|\mathbf{Z}^0\|_2^2$, where c' is a small constant, since $\sqrt{\sigma_1} \le \|\mathbf{Z}^0\|_2 \le 2\sqrt{\sigma_1}$ holds as long as $\mathbf{Z}^0 \in \mathbb{B}(\sqrt{\sigma_r}/4)$. Moreover, the reconstruction error $\|\mathbf{X}^T - \mathbf{X}^*\|_F^2$ can be upper bounded by $C\sigma_1 D^2(\mathbf{Z}^T, \mathbf{Z}^*)$, where C is a universal constant. Therefore, $\mathbf{X}^T$ is indeed a good estimator for $\mathbf{X}^*$.

Theorem 5.7 (Initialization)**.** Recall that $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^{*\top}$ is the unknown rank-r matrix. Consider $\mathbf{U}^0$, $\mathbf{V}^0$ produced in the initialization Algorithm 2, and let $\mathbf{X}^0 = \mathbf{U}^0\mathbf{V}^{0\top}$. If $L/\mu \in (1, 4/3)$, then with step size $\tau = 1/L$, we have

$$\|\mathbf{X}^0 - \mathbf{X}^*\|_F \le \rho^S \|\mathbf{X}^*\|_F + \frac{2\sqrt{3r}\epsilon(n, \delta)}{L(1-\rho)},$$

with probability at least $1 - \delta$, where $\rho = 2\sqrt{1 - \mu/L}$ is the contraction parameter.

Remark 5.8. According to Lemma 5.14 in Tu et al. (2015), if we have $\|\mathbf{X}^0 - \mathbf{X}^*\|_F \leq c\sigma_r$, where $c \leq \min\{1/2, 2c_2\}$, then the following inequality holds

$$d^2(\mathbf{Z}, \mathbf{Z}^*) \leq \frac{\sqrt{2}-1}{2} \frac{\|\mathbf{X}^0 - \mathbf{X}^*\|_F^2}{\sigma_r} \leq c_2^2 \sigma_r.$$

Therefore, in order to satisfy the initial assumption $\mathbf{Z}^0 \in \mathbb{B}(c_2\sqrt{\sigma_r})$ in Theorem 5.6, it is sufficient to make sure $\mathbf{X}^0$ is close enough to the unknown rank-r matrix $\mathbf{X}^*$, i.e., $\|\mathbf{X}^0 - \mathbf{X}^*\|_F \leq c\sigma_r$. In addition, we can assume the sample size n is large enough such that $\epsilon(n, \delta) \leq cL(1-\rho)\sigma_r/(2\sqrt{3r})$, which has the same order as the error bound in Theorem 5.6. Thus according to Theorem 5.7, it is sufficient to perform $S = \log(c'\sigma_r/\|\mathbf{X}^*\|_F)/\log(\rho)$ number of iterations in Algorithm 2 to make sure $\|\mathbf{X}^0 - \mathbf{X}^*\|_F \leq c\sigma_r$. Furthermore, our initialization algorithm only requires the condition $L/\mu \in (1, 4/3)$, which significantly relaxes the condition required in Park et al. (2016b), i.e.,

$$\frac{L}{\mu} \leq 1 + \frac{\sigma_r^2}{4608\|\mathbf{X}^*\|_F^2}.$$

5.2 Results for Specific Examples

The deterministic results in Theorem 5.6 are fairly abstract in nature. Here, we consider the specific examples of low-rank matrix estimation in Section 3.2, and demonstrate how to apply our general results in Section 5.1 to these examples. In the following discussions, we denote $d = \max\{d_1, d_2\}$.

5.2.1 Matrix Regression

For matrix regression, we obtain the restricted strong convexity and smoothness parameters $\mu = 4/9$ and $L = 5/9$. Moreover, we derive the upper bound of the gradient $\nabla\mathcal{L}_n$ at $\mathbf{X}^*$ in terms of spectral norm.

Corollary 5.9. Suppose the previously stated conditions are satisfied. There exist constants $\{c_i\}_{i=1}^5$ such that if we choose step size $\eta \leq c_1/\sigma_1$, for the output of Algorithm 1, we have, with probability at least $1 - c_2\exp\big(-c_3 d\big)$, that

$$D^2(\mathbf{Z}^{t+1}, \mathbf{Z}^*) \leq \left(1 - \frac{2\sigma_r\eta}{45}\right) D^2(\mathbf{Z}^t, \mathbf{Z}^*) + \eta c_4 \nu^2 \frac{rd}{n},$$

for any initial solution $\mathbf{Z}^0 \in \mathbb{B}(c_5\sqrt{\sigma_r})$.

Remark 5.10. In the noisy case, Corollary 5.9 suggests that, after $O\big(\kappa\log(n/(rd))\big)$ number of iterations, the output of our algorithm achieves $O(\sqrt{rd/n})$ statistical error, which matches the minimax lower bound for matrix regression (Negahban and Wainwright, 2011). While in the noiseless case, in order to satisfy restricted strong convexity and smoothness conditions, we require the sample size $n = O(rd)$, which achieves the optimal sample complexity for matrix regression (Recht et al., 2010; Tu et al., 2015).

5.2.2 Matrix Completion

For matrix completion, we consider a partially observed setting, such that we only observe entries of $\mathbf{X}^*$ over a subset $\mathcal{X} \subseteq [d_1] \times [d_2]$. We assume a uniform sampling model such that $\forall(j,k) \in \mathcal{X}$, $j \sim \text{uniform}([d_1])$, $k \sim \text{uniform}([d_2])$. It is observed in Gross (2011) that if $\mathbf{X}^*$ is equal to zero in nearly all elements, it is impossible to recover $\mathbf{X}^*$ unless all of its entries are sampled. In other words, there will always be some low-rank matrices, which are too spiky (Negahban and Wainwright, 2012; Gunasekar et al., 2014) to be recovered without sampling the whole matrix. In order to avoid the overly spiky matrices in matrix completion, we add an infinity norm constraint $\|\mathbf{X}^*\|_\infty \leq \alpha$ into our estimator, which is known as spikiness condition (Negahban and Wainwright, 2012). It is argued that the spikiness condition is much less restricted than incoherence conditions (Candès and Recht, 2009) imposed in exact low-rank matrix completion (Negahban and Wainwright, 2012; Klopp et al., 2014).

Therefore, we consider the class of low-rank matrices with infinity norm constraint as follows $\mathcal{C}(\alpha) = \big\{\mathbf{X} \in \mathbb{R}^{d_1\times d_2} \,\big|\, \|\mathbf{X}\|_\infty \leq \alpha\big\}$. Based on $\mathcal{C}(\alpha)$, we further define feasible sets $\mathcal{C}_i = \{\mathbf{A} \in \mathbb{R}^{d_i \times r} \,\big|\, \|\mathbf{A}\|_{2,\infty} \leq \sqrt{\alpha}\}$, where $i \in \{1, 2\}$. In this way, for any $\mathbf{U} \in \mathcal{C}_1$ and $\mathbf{V} \in \mathcal{C}_2$, we have $\mathbf{U}\mathbf{V}^\top \in \mathcal{C}(\alpha)$. By imposing spikiness condition, we can establish the restricted strong convexity and smoothness conditions with parameters $\mu = 8/9$ and $L = 10/9$. Moreover, we obtain the upper bound of $\|\nabla\mathcal{L}_n(\mathbf{X}^*)\|_2$.

Corollary 5.11. Suppose the previously stated conditions are satisfied and $\mathbf{X}^* \in \mathcal{C}(\alpha)$. There exist constants $\{c_i\}_{i=1}^4$ such that if we choose step size $\eta \leq c_1/\sigma_1$, for the output of Algorithm 1, we have, with probability at least $1 - c_2/d$, that

$$D^2(\mathbf{Z}^{t+1}, \mathbf{Z}^*) \leq \left(1 - \frac{4\sigma_r\eta}{45}\right) D^2(\mathbf{Z}^t, \mathbf{Z}^*) + \eta c_3 \max\{\nu^2, \alpha^2\} \frac{rd\log d}{p},$$

for any initial solution $\mathbf{Z}^0 \in \mathbb{B}(c_4\sqrt{\sigma_r})$.

Remark 5.12. For matrix completion with noisy observations, Corollary 5.11 suggests that after $O\big(\kappa\log(n/(rd\log d))\big)$ number of iterations, for the standardized error term $\|\mathbf{X}^T - \mathbf{X}^*\|_F/\sqrt{d_1 d_2}$, our algorithm attains $O(\sqrt{rd\log d/n})$ statistical error, which matches the minimax lower bound for matrix completion established in Negahban and Wainwright (2012);

Koltchinskii et al. (2011). While in the noiseless case, in order to guarantee restricted strong convexity and smoothness conditions, we require the sample size $n = O(rd\log d)$, which obtains optimal sample complexity for matrix completion (Candès and Recht, 2009; Recht, 2011; Chen et al., 2013).

5.2.3 One-Bit Matrix Completion

For one bit matrix completion, we establish the restricted strong convexity and smoothness condition with parameters $\mu = C_1\mu_\alpha$ and $L = C_2L_\alpha$, where C_1, C_2 are constants and μ_α, L_α satisfy

$$\mu_\alpha \leq \min\Bigg(\inf_{|x|\leq\alpha}\bigg\{\frac{f'^2(x)}{f^2(x)} - \frac{f''(x)}{f(x)}\bigg\}, \inf_{|x|\leq\alpha}\bigg\{\frac{f'^2(x)}{(1-f(x))^2} + \frac{f''(x)}{1-f(x)}\bigg\}\Bigg), \tag{5.1}$$

$$L_\alpha \geq \max\Bigg(\sup_{|x|\leq\alpha}\bigg\{\frac{f'^2(x)}{f^2(x)} - \frac{f''(x)}{f(x)}\bigg\}, \sup_{|x|\leq\alpha}\bigg\{\frac{f'^2(x)}{(1-f(x))^2} + \frac{f''(x)}{1-f(x)}\bigg\}\Bigg), \tag{5.2}$$

where α is the upper bound of the absolute value for every entry X_{jk}, and $f(x)$ is the differential function in (3.2). Given α and $f(x)$, μ_α and L_α are fixed constants which do not depend on dimension. For instance, we have $\mu_\alpha = e^\alpha/(1+e^\alpha)^2$ and $L_\alpha = 1/4$ for the logistic function. Another important quantity is γ_α, which reflects the steepness of the objective loss function $\gamma_\alpha \geq \sup_{|x|\leq\alpha}\big\{|f'(x)|/\big(f(x)(1-f(x))\big)\big\}$. Moreover, similar to the previous models, we obtain the upper bound of $\|\nabla\mathcal{L}_n(\mathbf{X}^*)\|_2$.

Corollary 5.13. Suppose the previously stated conditions are satisfied and $\mathbf{X}^* \in \mathcal{C}(\alpha)$. A subset of entries of the unknown matrix $\mathbf{X}^*$ is uniformly sampled with index set Ω, and the binary matrix $\mathbf{Y}$ in (3.2) is generated based on the log-concave function f. There exist constants $\{c_i\}_{i=1}^4$ such that if we choose step size $\eta \leq c_1/\sigma_1$, for the output of Algorithm 1, with probability at least $1 - c_2/d$, we have

$$D^2(\mathbf{Z}^{t+1}, \mathbf{Z}^*) \leq \bigg(1 - \frac{\mu\sigma_r\eta}{10}\bigg)D^2(\mathbf{Z}^t, \mathbf{Z}^*) + \eta c_3\max\{\gamma_\alpha^2, \alpha^2\}\frac{rd\log d}{p},$$

for any initial solution $\mathbf{Z}^0 \in \mathbb{B}(c_4\sqrt{\sigma_r})$.

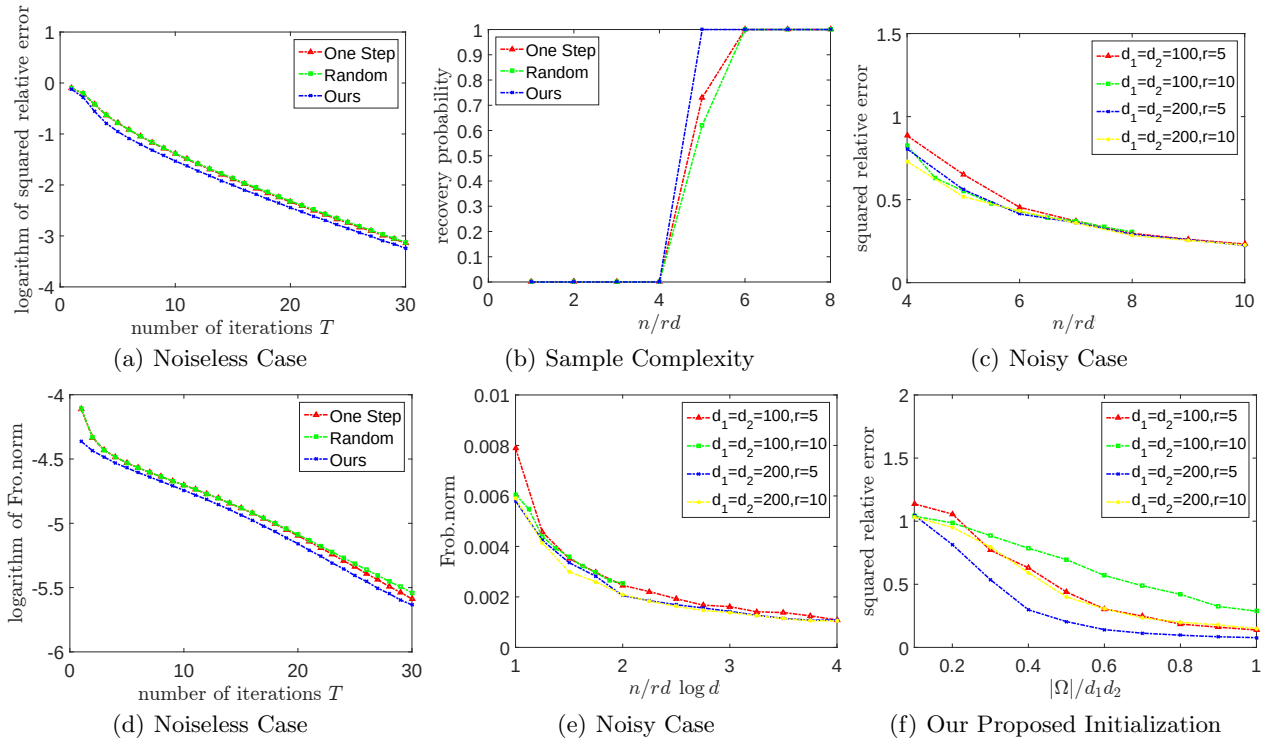
Remark 5.14. For one-bit matrix completion, Corollary 5.13 suggests that after $O\big(\kappa\log(n/(rd\log d))\big)$ number of iterations, for the standardized error term $\|\mathbf{X}^T - \mathbf{X}^*\|_F/\sqrt{d_1d_2}$, our algorithm obtains $O(\sqrt{rd\log d/n})$ statistical error, which matches the minimax lower bound of one-bit matrix completion problem provided by Davenport et al. (2014); Cai and Zhou (2013).

6 NUMERICAL EXPERIMENTS

In this section, we perform experiments on synthetic data to further illustrate the theoretical results of our method. We consider three approaches for initialization: (a) One step SVD of $\nabla\mathcal{L}_n(0)$ (One Step), which has been used in Bhojanapalli et al. (2015); Park et al. (2016a,b); (b) Random initialization (Random), which is suggested by Bhojanapalli et al. (2016); Park et al. (2016c); Ge et al. (2016); (c) Our proposed initialization Algorithm 2. We investigate the convergence rates of gradient descent under different initialization approaches, and evaluate the sample complexity that is required to recover the unknown low-rank matrices. All the results are based on 30 trials.

Matrix Regression and Matrix Completion. For matrix regression and matrix completion, we consider the unknowing matrix $\mathbf{X}^*$ in the following settings: (i) $d_1 = 100, d_2 = 100, r = 5$; (ii) $d_1 = 200, d_2 = 200, r = 10$; and (iii) $d_1 = 300, d_2 = 300, r = 20$. In all these settings, we first randomly generate $\mathbf{U}^* \in \mathbb{R}^{d_1\times r}, \mathbf{V}^* \in \mathbb{R}^{d_2\times r}$ to get $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^{*\top}$. Next, we generate measurements based on their observation models, respectively. For matrix regression, each entry of the observation matrix $\mathbf{A}_i$ follows i.i.d. standard Gaussian distribution. For both problems, we consider both (1) noisy case: the noise follows i.i.d. zero mean Gaussian distribution with standard deviation $\sigma = 0.1\cdot\|\mathbf{X}^*\|_\infty$; and (2) noiseless case.

To illustrate the convergence rate, we report the squared relative error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/\|\mathbf{X}^*\|_F^2$ and mean squared error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/(d_1d_2)$ for matrix regression and matrix completion respectively. For different settings, we generate $n = 0.2\cdot d_1d_2$ observations. To illustrate the sample complexity, we consider the empirical probability of exact recovery under different sample size. We get the output $\widehat{\mathbf{X}}$ of our algorithm given n random observations, and a trial is considered to be successful if the relative error is less than 10^{-3}. The convergence results under different initializations in the setting (i) are illustrated in Figures 1(a) and 1(d), which confirm the linear convergence rate of our algorithm. The results of empirical probability of exact recovery with different initializations for matrix regression in the setting (i) are displayed in Figure 1(b). We conclude that there exists a phase transition around $n = 4rd$, which implies that the sample complexity n is linear with rd. In the noisy case, the statistical error rates under our proposed initialization are shown in 1(c) and 1(e). Since the optimization error is eventually dominated by the statistical error, the total error converges to the statistical error rather than zero. Other experimental results can be found in the longer version of this paper.



(a) Noiseless Case (b) Sample Complexity (c) Noisy Case

(d) Noiseless Case (e) Noisy Case (f) Our Proposed Initialization

Figure 1: Simulation Results: (a)(d) Convergence rates for matrix regression and matrix completion in the noiseless case, which implies the linear convergence rate of our algorithm; (b) Empirical probability of exact recovery versus the rescaled sample size n/rd for matrix regression, which demonstrates the optimal sample complexity; (c)(e) Statistical error for matrix regression and matrix completion in the noisy case respectively, which confirms the statistical error bound; (f) Statistical error for one bit matrix completion under our proposed initialization, which is consistent with our theory.

One Bit Matrix Completion. For one-bit matrix completion, we consider the similar settings as in (Bhaskar and Javanmard, 2015; Davenport et al., 2014). We first generate the unknown low-rank matrix as $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^{*\top}$, where $\mathbf{U} \in \mathbb{R}^{d_1 \times r}, \mathbf{V} \in \mathbb{R}^{d_2 \times r}$ are randomly generated from a uniform distribution on $[-1/2, 1/2]$. Then, we scale $\mathbf{X}^*$ to make $\|\mathbf{X}^*\|_\infty = \alpha = 1$. Here we consider the Probit model under uniform sampling, namely $f(X_{ij}) = \Phi(\mathbf{X}_{ij}/\sigma)$ in (3.2), where Φ is the CDF of the standard Gaussian distribution. We set dimension $d_1 = d_2 \in \{100, 200\}$, rank $r \in \{5, 10\}$, and noise $\sigma = 0.18$.

In order to measure the performance of our estimator, we use the squared relative error which is defined as $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2 / \|\mathbf{X}^*\|_F^2$. The results are illustrated in Figure 1(f). It can be observed that the squared relative error decreases as the percentage of observed entries increases in all the settings. It also implies that under the same percentage of observed entries, the squared relative error decreases as the dimensionality increases, which further confirms the statistical rate.

7 Conclusions

In this paper, we developed a unified framework for estimating low-rank matrices, which integrates both optimization-theoretic and statistical analyses. Our algorithm and theory can be applied to low-rank matrix estimation based on both noisy observations and noiseless observations. In addition, we proposed a new initialization algorithm to provide a desired initial estimator, which outperforms existing initialization algorithms for nonconvex low-rank matrix estimation. Thorough experiments on synthetic data verified the advantages of our algorithm and theory.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This research was sponsored in part by the National Science Foundation IIS-1618948 and IIS-1652539. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Bhaskar, S. A. and Javanmard, A. (2015). 1-bit matrix completion under exact low-rank constraint. In *Information Sciences and Systems (CISS), 2015 49th Annual Conference on*. IEEE.

Bhojanapalli, S., Kyrillidis, A. and Sanghavi, S. (2015). Dropping convexity for faster semi-definite optimization. *arXiv preprint* .

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

Cai, T. and Zhou, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *Journal of Machine Learning Research* **14** 3619–3647.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics* **9** 717–772.

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

Chen, Y., Bhojanapalli, S., Sanghavi, S. and Ward, R. (2013). Completing any low-rank matrix, provably. *arXiv preprint arXiv:1306.2979* .

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

Davenport, M. A., Plan, Y., van den Berg, E. and Wootters, M. (2014). 1-bit matrix completion. *Information and Inference* **3** 189–223.

Ge, R., Lee, J. D. and Ma, T. (2016). Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272* .

Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory* **57** 1548–1566.

Gui, H. and Gu, Q. (2015). Towards faster rates and oracle property for low-rank matrix estimation. *arXiv preprint arXiv:1505.04780* .

Gunasekar, S., Ravikumar, P. and Ghosh, J. (2014). Exponential family matrix completion under structural constraints. In *ICML*.

Jain, P., Meka, R. and Dhillon, I. S. (2010). Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*.

Jain, P., Netrapalli, P. and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. In *STOC*.

Klopp, O. et al. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20** 282–303.

Koltchinskii, V., Lounici, K., Tsybakov, A. B. et al. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* **39** 2302–2329.

Loh, P.-L. and Wainwright, M. J. (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* 1069–1097.

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

Negahban, S., Yu, B., Wainwright, M. J. and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*.

Ni, R. and Gu, Q. (2016). Optimal statistical and computational rates for one bit matrix completion. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*.

Park, D., Kyrillidis, A., Bhojanapalli, S., Caramanis, C. and Sanghavi, S. (2016a). Provable burer-monteiro factorization for a class of norm-constrained matrix problems. *stat* **1050** 1.

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

Park, D., Kyrillidis, A., Caramanis, C. and Sanghavi, S. (2016c). Non-square matrix sensing without spurious local minima via the burer-monteiro approach. *arXiv preprint arXiv:1609.03240* .

Recht, B. (2011). A simpler approach to matrix completion. *Journal of Machine Learning Research* **12** 3413–3430.

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

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

Srebro, N., Rennie, J. and Jaakkola, T. S. (2004). Maximum-margin matrix factorization. In *Advances in neural information processing systems*.

Tu, S., Boczar, R., Soltanolkotabi, M. and Recht, B. (2015). Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566* .

Zhao, T., Wang, Z. and Liu, H. (2015). A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*.

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

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