A Primal-Dual Analysis of Global Optimality in Nonconvex Low-Rank Matrix Recovery

Xiao Zhang [*1] **Lingxiao Wang** [*2] **Yaodong Yu** [1] **Quanquan Gu** [2]

Abstract

We propose a primal-dual based framework for analyzing the global optimality of nonconvex low-rank matrix recovery. Our analysis are based on the restricted strongly convex and smooth conditions, which can be verified for a broad family of loss functions. In addition, our analytic framework can directly handle the widely-used incoherence constraints through the lens of duality. We illustrate the applicability of the proposed framework to matrix completion and one-bit matrix completion, and prove that all these problems have no spurious local minima. Our results not only improve the sample complexity required for characterizing the global optimality of matrix completion, but also resolve an open problem in Ge et al. (2017) regarding one-bit matrix completion. Numerical experiments show that primal-dual based algorithm can successfully recover the global optimum for various low-rank problems.

1. Introduction

Low-rank matrix recovery has received increasing attention in recent years, due to its wide range of applications including signal processing, computer vision and collaborative filtering (Rennie & Srebro, 2005; Ahmed & Romberg, 2015). The objective is to estimate an unknown rank-r matrix $\mathbf{X}^* \in \mathbb{R}^{d_1 \times d_2}$ based on partially observed measurements. More formally, low-rank matrix recovery can be formulated as the following optimization problem

$$\min_{\mathbf{X} \in \mathcal{C}} \mathcal{F}_n(\mathbf{X}) \quad \text{subject to} \quad \text{rank}(\mathbf{X}) \leq r, \tag{1.1}$$

where $\mathcal{F}_n : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ denotes a general sample loss function with respect to n measurements, and $\mathcal{C}$ denotes a constraint set such that $\mathbf{X}^* \in \mathcal{C}$. For example, $\mathcal{C}$ is set to be $\mathbb{R}^{d_1 \times d_2}$ in matrix sensing (Recht et al., 2010; Negahban & Wainwright, 2011), and is chosen to be the set of incoherent matrices in matrix completion (Rohde et al., 2011; Koltchinskii et al., 2011; Negahban & Wainwright, 2012) and one-bit matrix completion (Cai & Zhou, 2013; Davenport et al., 2014).

Tremendous efforts have been made to efficiently solve (1.1), among which the most popular ones are nuclear norm relaxation based methods (Srebro et al., 2004; Candès & Tao, 2010; Rohde et al., 2011; Recht et al., 2010; Recht, 2011; Negahban & Wainwright, 2011; 2012; Gui & Gu, 2015). These methods can achieve near optimal sample complexity for recovery (Balcan et al., 2017), but a singular value decomposition (SVD) step, whose time complexity is $O(d^3)$[1], is required at each iteration, which is computationally expensive for large-scale datasets. To avoid using SVD, the most commonly-used technique is based on Burer-Monteiro factorization (Burer & Monteiro, 2003), which reparameterizes the low-rank matrix $\mathbf{X}$ as the product of two smaller matrices $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$ such that $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$. Instead of optimizing (1.1) directly, we turn to solve the following nonconvex optimization problem

$$\min_{\mathbf{U} \in \mathcal{C}_1, \mathbf{V} \in \mathcal{C}_2} \mathcal{F}_n(\mathbf{U}\mathbf{V}^\top), \tag{1.2}$$

where $\mathcal{C}_1 \subseteq \mathbb{R}^{d_1 \times r}, \mathcal{C}_2 \subseteq \mathbb{R}^{d_2 \times r}$ are some constraint sets induced by $\mathcal{C}$ (c.f. Section 2.1). Note that (1.2) automatically ensures the low-rankness of the estimated matrix.

A line of research (Bach et al., 2008; Keshavan et al., 2009; Lee et al., 2013; Jain et al., 2013; Bach, 2013; Hardt, 2014; Hardt & Wootters, 2014; Netrapalli et al., 2014; Jain & Netrapalli, 2014; Haeffele et al., 2014; Sun & Luo, 2015; Bhojanapalli et al., 2015; Chen & Wainwright, 2015; Zhao et al., 2015; Tu et al., 2015; Chen & Wainwright, 2015; Zheng & Lafferty, 2015; 2016; Park et al., 2016b; Jin et al., 2016; Gu et al., 2016; Wang et al., 2017a;b; Xu et al., 2017; Zhang et al., 2018) proposed to solve (1.2) based on gradient

[*]Equal contribution [1]Department of Computer Science, University of Virginia, Charlottesville, VA 22904, USA. [2]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA 90095, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[1]We assume $d_1 = d_2 = d$ when discussing the sample complexity for simplicity.

descent and/or alternating minimization, and established a locally linear convergence property provided that the initial solution falls into a basin of attraction, i.e., a small neighbourhood around the optimum. Recently, another line of research (Bhojanapalli et al., 2016; Ge et al., 2016; Park et al., 2016c; Li et al., 2016; Zhu et al., 2017a; Ge et al., 2017) directly characterized the optimization landscape of (1.2) and proved that various low-rank matrix recovery problems, including matrix sensing (Bhojanapalli et al., 2016; Park et al., 2016c; Zhu et al., 2017a), matrix completion (Ge et al., 2016), and robust PCA (Ge et al., 2017), have no spurious local minima, i.e., all local minima are global ones. Based on existing results on finding local minimum for certain nonconvex problems (Ge et al., 2015; Carmon et al., 2016; Agarwal et al., 2016; Jin et al., 2017), they further showed that (1.2) can be successfully solved by saddle-avoiding algorithms, such as perturbed gradient descent. However, none of the aforementioned work is generic enough to cover objective functions beyond square loss, e.g., the sample loss function for one-bit matrix completion (Davenport et al., 2014).

Following the second line of research, we propose a primal-dual analysis to characterize the landscape of general objective functions in nonconvex low-rank matrix recovery including both square loss and beyond. By using restricted strongly convex and smooth conditions (Negahban et al., 2009; Negahban & Wainwright, 2011), we are able to characterize a large family of low-rank problems. To incorporate the widely-used incoherence constraints for low-rank matrix estimation, we propose to analyze the corresponding Lagrangian function rather than the primal objective function and use the Karush-Kuhn-Tucker (KKT) condition (Nocedal & Wright, 2006) to characterize the local minima of (1.2). Our analysis shows that the optimization landscape of (1.2) is well-behaved, i.e., there are no spurious local minima. In addition, we demonstrate empirically that the primal-dual based algorithm proposed in (Nocedal & Wright, 2006) can recover the ground truth matrix successfully. Our major contributions are further highlighted as follows.

1.1. Contributions

- Our general framework can be applied to any loss function that satisfies the restricted strongly convex and smooth conditions (c.f. Section 3), which covers a broad family of loss functions for low-rank problems. All the existing theoretical analyses (Bhojanapalli et al., 2016; Ge et al., 2016; Park et al., 2016c; Li et al., 2016; Zhu et al., 2017a; Ge et al., 2017) are limited to square loss, thus we resolve an open problem raised in Ge et al. (2017) regarding the characterization of global geometry for one-bit matrix completion.

- Our primal-dual analysis is applicable to various noisy low-rank problems. In particular, our analysis suggests there are no spurious local minima in noisy matrix completion, provided that the number of observations is $O(r^2 d \log d)$. Compared with existing studies (Ge et al., 2016; 2017) whose sample complexity scales to the fourth power with the rank, the sample requirement of our method matches the best-known sample complexity of matrix completion using nonconvex optimization algorithm (Zheng & Lafferty, 2016) under the incoherence condition.

- Compared with the seminal work (Ge et al., 2016; 2017) along this line that makes use of ad hoc regularizer to deal with incoherence constraints, our primal-dual analytic framework directly characterizes the global geometry of constrained nonconvex optimization problem for low-rank matrix recovery using duality theory. We believe the Lagrangian based proof technique is of independent interest, which can be extended to handle more general inequality constraints in other nonconvex problems.

1.2. Related Work

Characterizing the landscape of various objective functions has attracted more and more attention in recent years. For instance, Sun et al. (2015) studied the nonconvex geometry of complete dictionary recovery problem, and proved that all local minima are global ones. Sun et al. (2016) showed that a nonconvex fourth-order polynomial objective for phase retrieval has no spurious local minimum and all global minima are equivalent. Lee et al. (2016) showed that gradient descent converges to local minimum almost surely, using the stable manifold theorem from dynamical system. Ge et al. (2016) proved that the commonly used nonconvex objective function for positive semidefinite matrix completion has no spurious local minimum. In an independent work, Bhojanapalli et al. (2016) proved that positive semidefinite (PSD) matrix sensing, a very related problem to matrix completion, has no spurious local minima under the restricted isometry property (RIP). Later on, Park et al. (2016c) extended the geometric analysis of matrix sensing from PSD matrices to rectangular matrices. Zhu et al. (2017a) provided a unified geometric analysis for objective functions satisfying the restricted strong convexity/smoothness property, but their work cannot deal with the constrained optimization, e.g., matrix completion and one-bit matrix completion.

Most recently, several studies attempted to unify the global geometry analyses for nonconvex low-rank matrix recovery problems. For instance, Li et al. (2016) proposed a general theory to characterize the global geometry of positive semidefinite low-rank matrix factorization problem. Zhu et al. (2017b) further extended the geometric analysis in Li et al. (2016) to rectangular matrix factorization problem.

Nevertheless, both of their analyses require the objective function to be quadratic (i.e., square loss function), which is not applicable to constrained low-rank matrix recovery problems. The most relevant work to ours is Ge et al. (2017), which proposed a general framework to characterize the landscape of nonconvex low-rank matrix recovery problem. More specifically, they incorporated the constraints for matrix completion and robust PCA by a specifically designed regularizer, to make the solution lie in a desired region (e.g., the set of incoherent matrices). However, their framework still requires the loss function to be quadratic, thus unable to analyze low-rank problems with general objective function, such as one-bit matrix completion.

1.3. Organization and Notation

The remainder of this paper is organized as follows. We formally state the general low-rank matrix recovery problem and introduce two specific applications in Section 2. In Section 3, we lay out conditions for the proposed primal-dual based framework and present our main theoretical results. In Section 4, we apply the general results to two specific low-rank problems. The primal-dual based method and the numerical experiments are illustrated in Sections 5 and 6, respectively. We conclude in Section 7 and defer the detailed proofs to the supplementary materials.

We use capital symbols such as $\mathbf{A}$ to denote matrices and $[d]$ to denote index set $\{1, 2, \ldots, d\}$. Let the i-th row, j-th column and (i, j)-th entry of $\mathbf{A}$ be $\mathbf{A}_{i,*}$, $\mathbf{A}_{*,j}$ and A_{ij} respectively. Denote the i-th standard basis by $\mathbf{e}_i$ and the ℓ-th largest singular value of $\mathbf{A}$ by $\sigma_\ell(\mathbf{A})$. We use $\mathrm{vec}(\mathbf{A})$ to denote the vectorization of matrix $\mathbf{A}$. For any vector $\mathbf{x}$, we use $\|\mathbf{x}\|_2$ to denote its ℓ_2 norm. Let $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|_2$ be the Frobenius norm and the spectral norm of matrix $\mathbf{A}$, respectively. We define the largest ℓ_2 norm of its rows as $\|\mathbf{A}\|_{2,\infty} = \max_i \|\mathbf{A}_{i,*}\|_2$. For any two sequences $\{a_n\}$ and $\{b_n\}$, if there exists a constant $0 < C < \infty$ such that $a_n \leq C b_n$, then we denote $a_n = O(b_n)$.

2. Constrained Nonconvex Optimization for Low-Rank Matrix Recovery

In this section, we introduce our general framework for low-rank matrix recovery, along with several applications.

2.1. Generic Framework

Let the singular value decomposition of the unknown low-rank matrix be $\mathbf{X}^* = \overline{\mathbf{U}}^* \mathbf{\Sigma} \overline{\mathbf{V}}^{*\top}$ and $\mathbf{U}^*$, $\mathbf{V}^*$ be the underlying factorized matrices such that $\mathbf{U}^* = \overline{\mathbf{U}}^* \mathbf{\Sigma}^{1/2}$, $\mathbf{V}^* = \overline{\mathbf{V}}^* \mathbf{\Sigma}^{1/2}$. Denote the sorted singular values of $\mathbf{X}^*$ by $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$. Recall that we aim to characterize the global optimality of the general nonconvex optimization problem (1.2). Formally, we define the general constraint sets $\mathcal{C}_1, \mathcal{C}_2$ as follows

$$\begin{aligned} \mathcal{C}_1 &= \big\{ \mathbf{U} \in \mathbb{R}^{d_1 \times r} \,\big|\, \|\mathbf{U}\|_{2,\infty} \leq \alpha_1 \big\}, \\ \mathcal{C}_2 &= \big\{ \mathbf{V} \in \mathbb{R}^{d_1 \times r} \,\big|\, \|\mathbf{V}\|_{2,\infty} \leq \alpha_2 \big\}, \end{aligned} \tag{2.1}$$

where α_1, α_2 will be specified for different examples. To guarantee optimization problem (1.2) is well-defined, we assume $\mathbf{U}^* \in \mathcal{C}_1$ and $\mathbf{V}^* \in \mathcal{C}_2$. It is worth noting that (2.1) can cover a wide range of low-rank matrix recovery problems. For instance, in matrix completion, constraint sets in the form of (2.1) are proposed to ensure the estimator $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ is incoherent.

In addition, following Tu et al. (2015); Zheng & Lafferty (2016); Park et al. (2016a); Wang et al. (2017a), we add an additional regularization term $\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2$ to (1.2) such that the solutions $\mathbf{U}$, $\mathbf{V}$ are in similar scale. Specifically, we propose to analyze the following constrained optimization problem with respect to the stacked matrix $\mathbf{Z} = [\mathbf{U}; \mathbf{V}]$ in the lifted space $\mathbb{R}^{(d_1+d_2)\times r}$

$$\min_{\mathbf{Z} \in \mathcal{D}} \mathcal{G}(\mathbf{Z}) = \mathcal{F}_n(\mathbf{U}\mathbf{V}^\top) + \frac{\gamma}{4}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2, \tag{2.2}$$

where the constraint set $\mathcal{D}$ is defined as $\mathcal{D} = \{\mathbf{Z} \in \mathbb{R}^{(d_1+d_2)\times r} \mid \|\mathbf{Z}\|_{2,\infty} \leq \alpha\}$, where $\alpha = \max\{\alpha_1, \alpha_2\}$, and γ denotes the regularization parameter.

2.2. Specific Examples

We briefly introduce two specific examples: noisy matrix completion and one-bit matrix completion.

Noisy matrix completion. The goal of matrix completion (Rohde et al., 2011; Koltchinskii et al., 2011; Negahban & Wainwright, 2012) is to estimate the unknown low-rank matrix $\mathbf{X}^*$ from its partially observed (noisy) entries. More specifically, we consider the uniform observation model

$$Y_{jk} = \begin{cases} X^*_{jk} + E_{jk}, & \text{for any } (j,k) \in \Omega; \\ *, & \text{otherwise.} \end{cases} \tag{2.3}$$

where $\Omega \subseteq [d_1] \times [d_2]$ denotes the observed index set such that for any $(j, k) \in \Omega$, $j \sim \text{uniform}([d_1])$ and $k \sim \text{uniform}([d_2])$. Here, $\mathbf{Y} \in \mathbb{R}^{d_1 \times d_2}$ denotes the observed data matrix, $\mathbf{E} \in \mathbb{R}^{d_1 \times d_2}$ denotes a random noise matrix such that each entry of $\mathbf{E}$ follows i.i.d. Gaussian distribution with variance $\nu^2/(d_1 d_2)$.

As discussed in previous work (Gross, 2011; Negahban & Wainwright, 2012), it is impossible to recover the unknown low-rank matrix $\mathbf{X}^*$ if it is too sparse. To avoid such issue, we impose the following incoherence condition (Candès & Recht, 2009) on the singular spaces of $\mathbf{X}^*$

$$\|\overline{\mathbf{U}}^*\|_{2,\infty} \leq \sqrt{\frac{\beta r}{d_1}} \quad \text{and} \quad \|\overline{\mathbf{V}}^*\|_{2,\infty} \leq \sqrt{\frac{\beta r}{d_2}}, \tag{2.4}$$

where r denotes the rank of $\mathbf{X}^*$, and β denotes the incoherence parameter. Note that based on the incoherence condition (2.4), we can derive $\|\mathbf{X}^*\|_{\infty,\infty} \leq \|\overline{\mathbf{U}}^*\|_{2,\infty} \cdot \|\mathbf{\Sigma}\|_2 \cdot \|\overline{\mathbf{V}}^*\|_{2,\infty} \leq \beta r \sigma_1 / \sqrt{d_1 d_2}$, which leads to a dimension-free signal-to-noise ratio in observation model (2.3).

More specifically, we consider the following constrained optimization problem for matrix completion

$$\min_{\mathbf{Z}\in\mathcal{D}} \frac{1}{2p} \sum_{(j,k)\in\Omega} (\mathbf{U}_{j,*}\mathbf{V}_{k,*}^\top - Y_{jk})^2 + \frac{\gamma}{4}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2, \tag{2.5}$$

where $\mathcal{D}$ is defined in Section 2.1 and $p = |\Omega|/(d_1 d_2)$ denotes the sampling rate. In order to guarantee $\mathbf{Z}^* = [\mathbf{U}^*; \mathbf{V}^*] \in \mathcal{D}$, we set $\alpha = \sqrt{\beta r \sigma_1^*/d}$, where d_1 and d_2 are in the same order $O(d)$ for simplicity.

One-bit matrix completion. The objective of one-bit matrix completion (Davenport et al., 2014; Cai & Zhou, 2013; Ni & Gu, 2016) is to recover the unknown low-rank matrix $\mathbf{X}^*$ from a set of binary observations. In detail, the dependence of the observed binary matrix $\mathbf{Y} \in \mathbb{R}^{d_1\times d_2}$ on $\mathbf{X}^*$ is presented as follows

$$Y_{jk} = \begin{cases} +1, & \text{if } X_{jk}^* + E_{jk} > 0, \\ -1, & \text{if } X_{jk}^* + E_{jk} < 0, \end{cases} \tag{2.6}$$

where $\mathbf{E} \in \mathbb{R}^{d_1\times d_2}$ denotes a random noise matrix. Let f be the cumulative distribution function of $-E_{jk}$, then the observation model (2.6) can be recast as the following probabilistic model

$$Y_{jk} = \begin{cases} +1, & \text{with probability } f(X_{jk}^*), \\ -1, & \text{with probability } 1 - f(X_{jk}^*). \end{cases} \tag{2.7}$$

In addition, we consider the same uniform sampling model for the observed index set Ω as in matrix completion, and impose the incoherence condition (2.4) on $\mathbf{X}^*$ to avoid the overly sparse matrices. Specifically, we aim to solve the following optimization problem for one-bit matrix completion

$$\min_{\mathbf{Z}\in\mathcal{D}} -\frac{1}{|\Omega|} \sum_{(j,k)\in\Omega} \Big\{ \mathbb{1}_{(Y_{jk}=1)} \log\big(f(\mathbf{U}_{j*}\mathbf{V}_{k*}^\top)\big) + \mathbb{1}_{(Y_{jk}=-1)} \log\big(1 - f(\mathbf{U}_{j,*}\mathbf{V}_{k,*}^\top)\big) \Big\} + \frac{\gamma}{4}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2, \tag{2.8}$$

where $\Omega \subseteq [d_1] \times [d_2]$ denotes the observed index set with cardinality $|\Omega|$, and we also set the parameter $\alpha = \sqrt{\beta r \sigma_1^*/d}$ in the constraint set $\mathcal{D}$ to ensure optimization probelm (2.8) is well-defined.

3. Results for Generic Framework

Before presenting our main theoretical results, we first lay out the formal definition of local minimizer and the basic necessary optimality conditions with respect to constrained optimization problem (2.2) .

Definition 3.1. $\mathbf{Z}$ is a local minimizer of constrained optimization (2.2), if $\mathbf{Z}$ satisfies the following conditions: (i) $\mathbf{Z} \in \mathcal{D}$. (ii) There exists a neighbourhood $\mathcal{N}$ of $\mathbf{Z}$ such that for any $\widetilde{\mathbf{Z}} \in \mathcal{N} \cap \mathcal{D}$, $\mathcal{G}(\widetilde{\mathbf{Z}}) \geq \mathcal{G}(\mathbf{Z})$ holds.

For general constrained optimization (2.2), we provide the fundamental first-order necessary condition as follows.

Lemma 3.2. Suppose $\mathbf{Z}$ is a local minimizer of constrained optimization problem (3.1). Then for all feasible directions[2] $\mathbf{\Delta} \in \mathbb{R}^{(d_1+d_2)\times r}$, the following inequality holds:

$$\langle \nabla\mathcal{G}(\mathbf{Z}), \mathbf{\Delta}\rangle \geq 0.$$

Recall that the constraint set $\mathcal{D}$ is defined as $\mathcal{D} = \{\mathbf{Z} \in \mathbb{R}^{(d_1+d_2)\times r} \mid \|\mathbf{Z}\|_{2,\infty} \leq \alpha\}$. Thus, according to the definition of $\|\cdot\|_{2,\infty}$, we can reformulate (2.2) as the following equivalent standard form

$$\begin{aligned} \min_{\mathbf{Z}\in\mathbb{R}^{(d_1+d_2)\times r}} \ & \mathcal{G}(\mathbf{Z}) = \mathcal{F}_n(\mathbf{U}\mathbf{V}^\top) + \frac{\gamma}{4}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2 \\ \text{subject to } \ & h_i(\mathbf{Z}) \leq 0, \ \text{ for all } i \in [d_1+d_2], \end{aligned} \tag{3.1}$$

where $h_i(\mathbf{Z}) = \|\mathbf{e}_i^\top\mathbf{Z}\|_2^2 - \alpha^2$, and $\mathbf{e}_i$ represents the i-th natural basis. Accordingly, the Lagrangian with respect to (3.1) is given as follows

$$\mathcal{L}(\mathbf{Z}, \boldsymbol{\lambda}) = \mathcal{G}(\mathbf{Z}) + \sum_{i=1}^{d_1+d_2} \lambda_i h_i(\mathbf{Z}),$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots \lambda_{(d_1+d_2)}]^\top$ denotes the Lagrange multiplier vector. Based on the Lagrangian, we give the following necessary optimality conditions (Nocedal & Wright, 2006) for constrained optimization problem (3.1).

Lemma 3.3. Let $\mathbf{Z}$ be a local minimizer of constrained optimization problem (3.1). Define the set of active constraint gradients at $\mathbf{Z}$ as $\mathcal{A}(\mathbf{Z}) = \{i \in [d_1+d_2] \mid h_i(\mathbf{Z}) = 0\}$. Suppose the active set $\{\nabla h_i(\mathbf{Z}) \mid i \in \mathcal{A}(\mathbf{Z})\}$ is linearly independent. Then there exists a Lagrange multiplier vector $\boldsymbol{\lambda}$ such that $(\mathbf{Z}, \boldsymbol{\lambda})$ satisfies the following Karush-Kuhn-Tucker (KKT) conditions

1. $h_i(\mathbf{Z}) \leq 0, \quad$ for all $i \in [d_1+d_2]$,
2. $\boldsymbol{\lambda} \geq \mathbf{0}$,
3. $\lambda_i h_i(\mathbf{Z}) = 0, \quad$ for all $i \in [d_1+d_2]$,
4. $\nabla_{\mathbf{Z}}\mathcal{L}(\mathbf{Z}, \boldsymbol{\lambda}) = \mathbf{0}$.

[2] We say $\mathbf{\Delta}$ is a feasible direction for constraint set $\mathcal{D}$ at $\mathbf{Z}$, if there exists $t > 0$ such that $\mathbf{Z} + s\mathbf{\Delta} \in \mathcal{D}$ for all $s \in [0, t]$.

Lemma 3.4. Under the same conditions as in Lemma 3.3, let $\boldsymbol{\lambda}$ be a Lagrange multiplier vector such that $(\mathbf{Z}, \boldsymbol{\lambda})$ satisfies the KKT condition. For any $\boldsymbol{\Delta} \in \mathbb{R}^{(d_1+d_2)\times r}$ satisfying the condition that $\langle \nabla h_i(\mathbf{Z}), \boldsymbol{\Delta} \rangle \leq 0$ holds for all $i \in \mathcal{A}(\mathbf{Z})$, $(\mathbf{Z}, \boldsymbol{\lambda})$ also satisfies the following inequality

$$\mathrm{vec}(\boldsymbol{\Delta})^\top \nabla^2_{\mathbf{Z}} \mathcal{L}(\mathbf{Z}, \boldsymbol{\lambda}) \mathrm{vec}(\boldsymbol{\Delta}) \geq 0.$$

The aforementioned necessary optimality conditions characterize the properties of local minima with respect to constrained optimization problem. As will be seen in later analysis, these optimality conditions are essential to show that there are no spurious local minima in (3.1).

Next, we lay out several conditions for the general sample loss function $\mathcal{F}_n$. To begin with, we present the restricted strong convexity and smoothness conditions (Negahban et al., 2009; Loh & Wainwright, 2013).

Condition 3.5 (Restricted Strong Convexity)**.** The sample loss function $\mathcal{F}_n$ is μ-restricted strongly convex, i.e., for all matrices $\mathbf{Y}, \mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ with rank at most $6r$

$$\mathcal{F}_n(\mathbf{Y}) \geq \mathcal{F}_n(\mathbf{W}) + \langle \nabla \mathcal{F}_n(\mathbf{W}), \mathbf{Y} - \mathbf{W} \rangle + \frac{\mu}{2} \|\mathbf{Y} - \mathbf{W}\|_F^2.$$

Condition 3.6 (Restricted Strong Smoothness)**.** The sample loss function $\mathcal{F}_n$ is L-restricted strongly smooth, i.e., for all matrices $\mathbf{Y}, \mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ with rank at most $6r$

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

In addition, we assume there is an upper bound for the gradient of the sample loss function $\nabla \mathcal{F}_n$ with respect to the unknown low-rank matrix $\mathbf{X}^*$.

Condition 3.7. Given a fixed sample size n and tolerance parameter $\delta \in (0, 1)$. Let $\epsilon(n, \delta)$ be the smallest scalar such that with probability at least $1 - \delta$, we have

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

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

As will be clear in the next section and in the proofs, Conditions 3.5, 3.6 and 3.7 can be verified for a wide range of sample loss functions, such as the objective functions in matrix completion and one-bit matrix completion.

Finally, we present the main results regarding the global optimality of optimization problem (3.1). In particular, we are going to show that under proper conditions, (3.1) has no spurious local minima.

Theorem 3.8. Assume the sample loss function $\mathcal{F}_n$ satisfies Conditions 3.5, 3.6 and 3.7. Under condition that $L/\mu \in (1, 18/17)$, for all local minima $\mathbf{Z} = [\mathbf{U}; \mathbf{V}]$ of optimization problem (3.1) with regularization parameter γ satisfying $\mu - L/2 \leq \gamma < \min\{(22\mu - 19L)/4, (3L - 2\mu)/2\}$, the reconstruction error satisfies

$$\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^*\|_F^2 \leq \Gamma r \epsilon^2(n, \delta), \tag{3.2}$$

with probability at least $1 - \delta$, where Γ is a constant depending on μ, L and γ.

Remark 3.9. Theorem 3.8 suggests that for all local minima $\mathbf{Z} = [\mathbf{U}; \mathbf{V}]$ of (3.1), the reconstructed matrix $\mathbf{U}\mathbf{V}^\top$ lies in a small neighbourhood of $\mathbf{X}^*$, and the radius of such neighbourhood decreases as the sample size n increases. While in the noiseless case ($\epsilon(n, \delta) = 0$), the right hand side of (3.2) is 0, which suggests that all local minima are global ones, i.e., $\mathbf{U}\mathbf{V}^\top = \mathbf{X}^*$. Note that we require the condition number L/μ to be close to 1 in Theorem 3.8. As will be clear in the next section and in the proofs, this assumption can be verified for the specific examples discussed in Section 2.1. Similar assumption has been imposed in existing work on matrix sensing (Bhojanapalli et al., 2016; Ge et al., 2017), in that the restricted isometry property parameter is required to be in a small range around 0.

4. Implications for Specific Examples

In this section, we illustrate how to apply our general framework to two specific low-rank problems: noisy matrix completion and one-bit matrix completion. Note that given the general results in Section 3, we only need to verify Conditions 3.5, 3.6, 3.7 and the assumption regarding the restricted condition number L/μ for each specific example.

4.1. Results for Noisy Matrix Completion

Recall that for noisy matrix completion, we aim to optimize (2.5) under the uniform sampling model (2.3) and incoherence condition (2.4). Specifically, we verify Conditions 3.5, 3.6 and 3.7 for $\mathcal{F}_n$ in the following corollary to characterize the global optimality of noisy matrix completion.

Corollary 4.1. Consider noisy matrix completion problem (2.5) under the uniform sampling model. Suppose the unknown rank-r matrix $\mathbf{X}^*$ satisfies incoherence condition (2.4) and each entry of the noise matrix $\mathbf{E}$ follows i.i.d. Gaussian distribution with variance $\nu^2/(d_1 d_2)$. Provided the number of observed samples $|\Omega| \geq c_1 r^2 d \log d$, with regularization parameter $\gamma = 1/2$, all local minima $\mathbf{Z} = [\mathbf{U}; \mathbf{V}]$ of optimization problem (2.5) satisfy

$$\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^*\|_F \leq c_2 \max\left\{\nu, \sqrt{r}\beta\sigma_1\right\} \sqrt{\frac{rd \log d}{n}},$$

with probability at least $1 - c_2/d$, where c_1, c_2 are both universal constants.

Remark 4.2. Due to the existence of noise, it is impossible to exactly recover the unknown low-rank matrix $\mathbf{X}^*$ for noisy matrix completion. However, we show that the reconstruction $\mathbf{U}\mathbf{V}^\top$ from any local minimizer of (2.5) is actually a good estimator of $\mathbf{X}^*$. The estimation error is in the order of $O(\sqrt{r^2 d \log d/n})$, which suggests that the more observations we have, the smaller estimation error we can achieve. It is worth noting that we only need $O(r^2 d \log d)$ observed entries of $\mathbf{X}^*$ to ensure (3.1) has no spurious local minima, in sharp contrast to existing work (Ge et al., 2016; 2017) whose sample complexity is at least $O(r^4 d \log d)$.

4.2. Results for One-Bit Matrix Completion

Recall that the objective of one-bit matrix completion is to solve optimization problem (2.8). We assume the standard regularity condition (Cai & Zhou, 2013) on the cumulative distribution function f in (2.7) as follows

$$\sup_{|x|\leq\beta'} \big\{|f'(x)|/\big(f(x)(1-f(x))\big)\big\} \leq \gamma_{\beta'}, \tag{4.1}$$

where $\gamma_{\beta'}$ reflects the steepness of the sample loss function, and when f and β' are given, $\gamma_{\beta'}$ is a fixed constant. We note that this condition holds for a large family of distributions, such as Logistic distribution, Gaussian distribution, and Laplacian distribution. To apply the results in the unified framework, it suffices to prove Conditions 3.5, 3.6 and 3.7 for one-bit matrix completion, respectively.

Corollary 4.3. Assume that the observed matrix $\mathbf{Y}$ follows the binary observation model (2.7) generated based on a cumulative distribution function f satisfying (4.1). Suppose the unknown low-rank matrix $\mathbf{X}^*$ satisfies incoherence condition (2.4) and the observed index set Ω follows the uniform sampling model. If the sample complexity $|\Omega|$ exceeds $c_1 r^2 d \log d$, and the regularization parameter is set as $\gamma = 1/2$, then with probability at least $1 - c_2/d$, all local minima $\mathbf{Z} = [\mathbf{U}; \mathbf{V}]$ of optimization problem (2.8) satisfy

$$\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^*\|_F \leq c_3 \max\{\gamma_{\beta'}, \sqrt{r}\beta\sigma_1\}\sqrt{\frac{rd\log d}{n}},$$

where c_1, c_2, c_3 are all universal constants.

Remark 4.4. Corollary 4.3 shows all local minima $\mathbf{Z} = [\mathbf{U}; \mathbf{V}]$ of one-bit matrix completion satisfy the condition $\mathbf{U}\mathbf{V}^\top$ lies in a close neighborhood around $\mathbf{X}^*$ with radius $O(\sqrt{r^2 d \log d/n})$. This suggests that as long as the number of observations is sufficient, we can obtain a good estimator for $\mathbf{X}^*$ by solving (2.8). To the best of our knowledge, all the existing studies on the characterization of global geometry for low-rank problems require the objective function to be square loss, thus our work is the first that can characterize the global optimality for one-bit matrix completion, which resolves an open problem in Ge et al. (2017).

5. The Primal-Dual Algorithm

So far, we have shown that all local minima of inequality constrained optimization problem (3.1) belong to a close neighbourhood of the ground truth matrix, with applications to two specific low-rank problems. It remains to find an efficient and effective algorithm that can find a local minimizer of (3.1) successfully.

Recall that our characterization of global optimality with respect to (3.1) is based on the Lagrange function and the duality theory. This motivates us to search for local minima of (3.1) from the primal-dual perspective. It has been proved in Di Pillo et al. (2011) that a particular primal-dual based algorithm can converge to a solution that satisfies the necessary optimality Conditions 3.3 and 3.4 for general nonlinear inequality constrained optimization problems. It immediately suggests that their algorithm can be directly applied to solve the optimization problem (3.1), and is guaranteed to find a local minimizer. Nevertheless, the algorithm in Di Pillo et al. (2011) requires to access the Hessian information, which is computationally very expensive for large scale problems. Thus, a more practical primal-dual algorithm is preferred.

Witnessing the empirical success of Augmented Lagrangian Method (Nocedal & Wright, 2006), a first-order primal-dual method for general constrained optimization problem, we propose to use the augmented Lagrangian method to solve the inequality constrained optimization problem (3.1), as displayed in Algorithm 1. More specifically, we introduce a slack variable $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_{d_1+d_2}]^\top$ to transform the inequality constraints in (3.1) into equality ones. Define the augmented Lagrange function $\widetilde{\mathcal{L}}$ as follows

$$\widetilde{\mathcal{L}}(\mathbf{Z}, \boldsymbol{\xi}, \boldsymbol{\lambda}, \mu) = \mathcal{G}(\mathbf{Z}) + \sum_{i=1}^{d_1+d_2} \lambda_i c_i(\mathbf{Z}) + \frac{\mu}{2}\sum_{i=1}^{d_1+d_2} c_i^2(\mathbf{Z}),$$

where $c_i(\mathbf{Z}) = h_i(\mathbf{Z}) + \xi_i^2$. At each iteration of Algorithm 1, we solve the minimization sub-problem in line 2 based on gradient descent with respect to $\mathbf{Z}$ and $\boldsymbol{\xi}$, and update the dual variable $\boldsymbol{\lambda}$ and the penalty parameter μ as suggested in Nocedal & Wright (2006). Here, we let $\mathbf{h}(\mathbf{Z}) = [h_1(\mathbf{Z}), \ldots, h_{d_1+d_2}(\mathbf{Z})]^\top$ in line 3.

Algorithm 1 Augmented Lagrangian Method

Input: Augmented Lagrangian function $\widetilde{\mathcal{L}}$; parameters $T, \boldsymbol{\xi}_0, \boldsymbol{\lambda}_0, \mu_0 > 0$ and $\rho \geq 1$; initial estimator $\mathbf{Z}_0$

1: **for** $t = 0, \ldots, T$ **do**
2: Solve $\big(\mathbf{Z}_{t+1}, \boldsymbol{\xi}_{t+1}\big) = \operatorname{argmin}_{\mathbf{Z},\boldsymbol{\xi}} \widetilde{\mathcal{L}}(\mathbf{Z}, \boldsymbol{\xi}, \boldsymbol{\lambda}_t, \mu_t)$
3: $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \mu_t\big(\mathbf{h}(\mathbf{Z}_{t+1}) + \boldsymbol{\xi}_{t+1}^2\big)$
4: $\mu_{t+1} = \rho\mu_t$
5: **end for**

Output: $\mathbf{Z}_{T+1}$

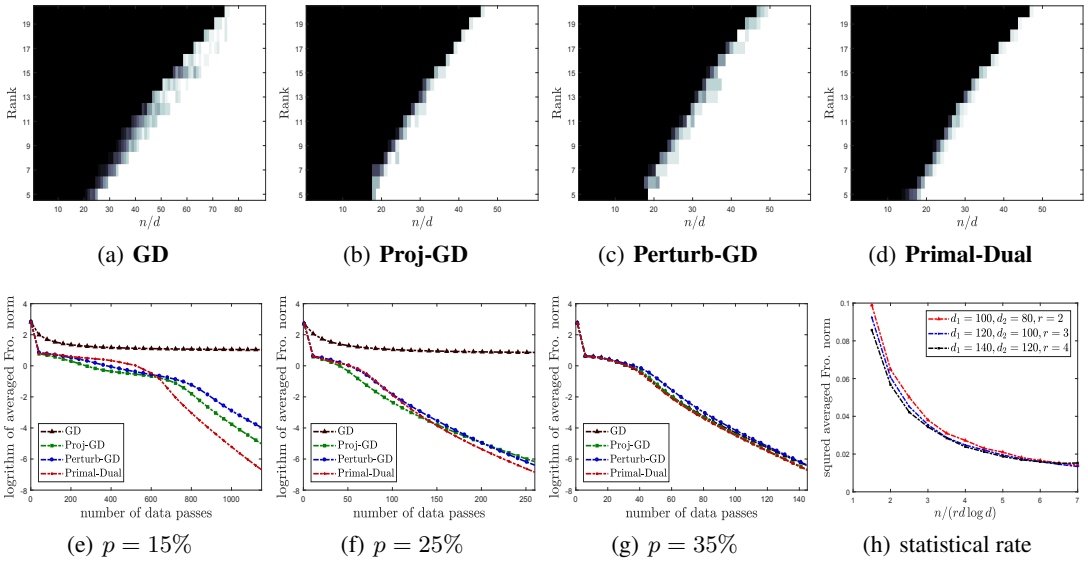


Figure 1. Simulation results for matrix completion: (i) Top panel: plots of the recovery probability based on the compared methods under the setting $d_1 = 120$, $d_2 = 100$ with different sample size n and rank r. White block indicates success and black block indicates failure. We set $d = \max\{d_1, d_2\}$. (ii) Bottom panel (e),(f),(g): comparison results on convergence rate in the noiseless case under the setting $d_1 = 2000$, $d_2 = 1500$, $r = 20$ with sampling rate p varied in $\{15\%, 25\%, 35\%\}$. (iii) Bottom panel (h): Plot of the squared averaged Frobenius norm error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/(d_1 d_2)$ vs. the rescaled sample size $n/(rd\log d)$ based on our approach under different settings.

As will be seen in the next section, we demonstrate through numerical experiments that the primal-dual based Algorithm 1 can efficiently solve the constrained nonconvex optimization problem (3.1) given enough observations.

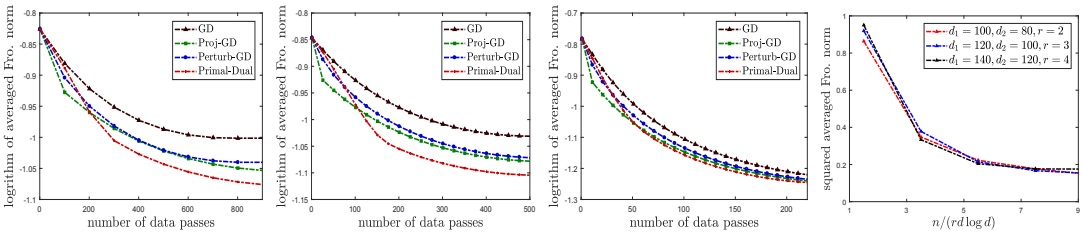
6. Experiments

In this section, we provide simulation results of the primal-dual based method, as discussed in Section 5, for matrix completion and one-bit matrix completion. Under random initialization, we compare our primal-dual based Algorithm 1 (**Primal-Dual**) with existing gradient-based methods, including vanilla gradient descent (**GD**), projected gradient descent (**Proj-GD**) and perturbed gradient descent (Jin et al., 2017) (**Perturb-GD**). We remark that **GD** and **Proj-GD** have been proposed in Ma et al. (2017) and Zheng & Lafferty (2016) for matrix completion respectively, but they all require a specially designed initialization procedure. Here, we are interested in evaluating the performance of all these algorithms with random initialization. All of the aforementioned algorithms are implemented in Matlab, and all the following experimental results are based on the optimal parameters, which are selected by cross validation and averaged over 20 trials.

6.1. Matrix Completion

We generate the observed data matrix according to the uniform observation model (2.3). In particular, the unknown low-rank matrix $\mathbf{X}^* \in \mathbb{R}^{d_1 \times d_2}$ is generated via $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^{*\top}$, where each entry of $\mathbf{U}^* \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V}^* \in \mathbb{R}^{d_2 \times r}$ is generated independently from standard Gaussian distribution, and we scale them to ensure $\max\{\|\mathbf{U}^*\|_{2,\infty}, \|\mathbf{V}^*\|_{2,\infty}\} \le \alpha$, where $\alpha = 2$. The noise matrix $\mathbf{E}$ is set as $\mathbf{0}$ in the noiseless case, while under the noisy setting, we generate each element of the noise matrix $\mathbf{E}$ from i.i.d. centered Gaussian distribution with variance $\sigma^2 = 0.25$. Note that due to random initialization, the initial estimators may not satisfy the incoherence constraint. In the sequel, we are going to evaluate the recovery performance of different algorithms under the noiseless setting, and investigate the statistical rate of our method.

To begin with, we compare the sample complexities required by the aforementioned algorithms under the setting $d_1 = 120$ and $d_2 = 100$ with sample size n and rank r varied. We say the final estimator $\widehat{\mathbf{X}}$ successfully recover the ground truth matrix $\mathbf{X}^*$, if the relative error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F/\|\mathbf{X}^*\|_F \le 10^{-3}$. Figures 1(a)-1(d) illustrate the recovery probability of different methods. Here, the white block indicates successful recovery and the black block denotes failure. As for **GD**, all the cases have phase transition around $n = 4rd$. While our method has phase transition around $n = 2.5rd$ under all the settings, and similar results are observed in Figure 1(b) for **Proj-GD** and Figure 1(c) regarding **Perturb-GD**. These results suggest that the sample complexity for all these methods could be linear in both d and r.



(a) $p = 10\%$ (b) $p = 20\%$ (c) $p = 50\%$ (d) statistical rate

Figure 2. Numerical results for one-bit matrix completion: (a),(b),(c) Convergence rate of different methods in the noiseless case under the setting that $d_1 = 2000$, $d_2 = 1500$ and $r = 20$ with sampling rate $p \in \{10\%, 20\%, 50\%\}$. (d) Plot of the squared averaged Frobenius norm error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/(d_1 d_2)$ vs. the rescaled sample size $n/(rd\log d)$ based on our algorithm under different settings.

Next, we compare the convergence rate of different methods under the setting $d_1 = 2000, d_2 = 1500, r = 20$ with sampling rate $p = |\Omega|/(d_1 d_2)$ chosen from $\{15\%, 25\%, 35\%\}$. The experimental results in terms of the averaged Frobenius norm error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F/\sqrt{d_1 d_2}$ versus the number of data passes are demonstrated in Figures 1(e)-1(g). It can be seen that our primal-dual based algorithm achieves lower estimation error than **GD** after the same number of data passes, especially when the observed sample size is small. On the other hand, compared with **Proj-GD** and **Perturb-GD**, our method achieves comparable performance, which verifies the effectiveness of Algorithm 1 for solving nonconvex optimization problem (2.5) to find a local minimum.

Finally, we study the statistical rate of our method under the following settings: (i) $d_1 = 100, d_2 = 80, r = 2$; (ii) $d_1 = 120, d_2 = 100, r = 3$; (iii) $d_1 = 140, d_2 = 120, r = 4$. The results are displayed in Figure 1(h). In detail, the vertical axis represents the estimation error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/(d_1 d_2)$, and the horizontal axis is the rescaled sample size $n/(rd\log d)$. The results show that the estimation error and the rescaled sample size align well under different settings, which suggests that the statistical rate of our method is $O(rd\log d/n)$.

6.2. One-Bit Matrix Completion

We generate the data matrix $\mathbf{Y}$ based on probability model (2.7) with $f(X_{ij}) = \Phi(X_{ij}/\sigma)$, where Φ is the cumulative distribution function of the standard Gaussian distribution, and we choose $\sigma = 0.5$ as the noise level. For the unknown low-rank matrix $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^{*\top}$ with rank r, we follow the same generative procedure as in Davenport et al. (2014); Bhaskar & Javanmard (2015); Ni & Gu (2016). More specifically, $\mathbf{U}^* \in \mathbb{R}^{d_1 \times r}$, $\mathbf{V}^* \in \mathbb{R}^{d_2 \times r}$ are randomly generated from a uniform distribution on $[-1/2, 1/2]$, and are scaled properly such that the incoherence constraint $\max\{\|\mathbf{U}^*\|_{2,\infty}, \|\mathbf{V}^*\|_{2,\infty}\} \le \alpha$ is satisfied, where we set $\alpha = 1$. In addition, we sample the observed index set Ω according to the uniform sampling model.

To demonstrate the effectiveness of Algorithm 1, we compare our method with existing gradient-based algorithms including **GD**, **Proj-GD** and **Perturb-GD** with random initialization. In particular, we compute the logarithm of the averaged estimation error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F/\sqrt{d_1 d_2}$ and plot it with the number of data passes for different methods, which are illustrated in Figures 2(a)-2(c) under the setting $d_1 = 2000, d_2 = 1500, r = 20$ with varied sampling rate. These results again confirm that with random initialization, the primal-dual based algorithm can recover the unknown low-rank matrix $\mathbf{X}^*$ successfully. In addition, **Proj-GD** demonstrates a sharp decrease in estimation error after the first several data passes, which is due to the simple but effective projection mechanism.

In addition, we investigate the statistical rate for one-bit matrix completion based on our algorithm. Figure 2(d) plots the averaged estimation error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/(d_1 d_2)$ versus the rescaled sample size $n/rd\log d$ under different settings, which suggests that our primal-dual based approach achieves statistical rate with order $O(rd\log d/n)$.

7. Conclusions and Future Work

In this paper, we proposed a primal-dual based framework to characterize the global optimality for nonconvex low-rank matrix recovery with incoherence constraints. Based on duality, we proved that the optimization landscape of such problem is well-behaved. We further applied a primal-dual based algorithm to solve the nonconvex optimization problem and demonstrated its effectiveness via simulations.

There are still some open problems along this line of research. For example, how to prove the optimization guarantees for the primal-dual based algorithm? Another question is how to generalize our framework to other constrained nonconvex optimization beyond incoherence constraints. We hope this work can act as the first step towards understanding the global geometry of general constrained nonconvex optimization problems.

Acknowledgement

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

Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., and Ma, T. Finding local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*, 2016.

Ahmed, A. and Romberg, J. Compressive multiplexing of correlated signals. *IEEE Transactions on Information Theory*, 61(1): 479–498, 2015.

Bach, F. Convex relaxations of structured matrix factorizations. *arXiv preprint arXiv:1309.3117*, 2013.

Bach, F., Mairal, J., and Ponce, J. Convex sparse matrix factorizations. *arXiv preprint arXiv:0812.1869*, 2008.

Balcan, M.-F., Liang, Y., Woodruff, D. P., and Zhang, H. Optimal sample complexity for matrix completion and related problems via ell_2-regularization. *arXiv preprint arXiv:1704.08683*, 2017.

Bhaskar, S. A. and Javanmard, A. 1-bit matrix completion under exact low-rank constraint. In *Information Sciences and Systems (CISS), 2015 49th Annual Conference on*, pp. 1–6. IEEE, 2015.

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

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

Burer, S. and Monteiro, R. D. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

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

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6): 717–772, 2009.

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

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Accelerated methods for non-convex optimization. *arXiv preprint arXiv:1611.00756*, 2016.

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

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

Di Pillo, G., Liuzzi, G., and Lucidi, S. A primal-dual algorithm for nonlinear programming exploiting negative curvature directions. *Numerical Algebra, Control and Optimization*, 1(3):509–528, 2011.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *COLT*, pp. 797–842, 2015.

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

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

Gross, D. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3): 1548–1566, 2011.

Gu, Q., Wang, Z., and Liu, H. Low-rank and sparse structure pursuit via alternating minimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 600–609, 2016.

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

Haeffele, B., Young, E., and Vidal, R. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *International Conference on Machine Learning*, pp. 2007–2015, 2014.

Hardt, M. Understanding alternating minimization for matrix completion. In *FOCS*, pp. 651–660. IEEE, 2014.

Hardt, M. and Wootters, M. Fast matrix completion without the condition number. In *COLT*, pp. 638–678, 2014.

Jain, P. and Netrapalli, P. Fast exact matrix completion with finite samples. *arXiv preprint*, 2014.

Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674. ACM, 2013.

Jin, C., Kakade, S. M., and Netrapalli, P. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 4520–4528, 2016.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

Keshavan, R. H., Oh, S., and Montanari, A. Matrix completion from a few entries. In *2009 IEEE International Symposium on Information Theory*, pp. 324–328. IEEE, 2009.

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

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pp. 1246–1257, 2016.

Lee, K., Wu, Y., and Bresler, Y. Near optimal compressed sensing of sparse rank-one matrices via sparse power factorization. *arXiv preprint arXiv:1312.0525*, 2013.

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

Loh, P.-L. and Wainwright, M. J. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pp. 476–484, 2013.

Ma, C., Wang, K., Chi, Y., and Chen, Y. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

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

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

Negahban, S., Yu, B., Wainwright, M. J., and Ravikumar, P. K. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pp. 1348–1356, 2009.

Netrapalli, P., Niranjan, U., Sanghavi, S., Anandkumar, A., and Jain, P. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pp. 1107–1115, 2014.

Ni, R. and Gu, Q. Optimal statistical and computational rates for one bit matrix completion. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 426–434, 2016.

Nocedal, J. and Wright, S. J. *Sequential quadratic programming*. Springer, 2006.

Park, D., Kyrillidis, A., Bhojanapalli, S., Caramanis, C., and Sanghavi, S. Provable non-convex projected gradient descent for a class of constrained matrix optimization problems. *arXiv preprint arXiv:1606.01316*, 2016a.

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

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

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

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

Rennie, J. D. and Srebro, N. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pp. 713–719. ACM, 2005.

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

Srebro, N., Rennie, J., and Jaakkola, T. S. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pp. 1329–1336, 2004.

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery over the sphere. *arXiv preprint arXiv:1504.06785*, 2015.

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pp. 2379–2383. IEEE, 2016.

Sun, R. and Luo, Z.-Q. Guaranteed matrix completion via nonconvex factorization. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pp. 270–289. IEEE, 2015.

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

Wang, L., Zhang, X., and Gu, Q. A unified computational and statistical framework for nonconvex low-rank matrix estimation. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 981–990, 2017a.

Wang, L., Zhang, X., and Gu, Q. A unified variance reduction-based framework for nonconvex low-rank matrix recovery. In *International Conference on Machine Learning*, pp. 3712–3721, 2017b.

Xu, P., Ma, J., and Gu, Q. Speeding up latent variable gaussian graphical model estimation via nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 1930–1941, 2017.

Zhang, X., Wang, L., and Gu, Q. A unified framework for nonconvex low-rank plus sparse matrix recovery. In *International Conference on Artificial Intelligence and Statistics*, pp. 1097–1107, 2018.

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

Zheng, Q. and Lafferty, J. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pp. 109–117, 2015.

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

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

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. The global optimization geometry of nonsymmetric matrix factorization and sensing. *arXiv preprint arXiv:1703.01256*, 2017b.