Optimal Statistical and Computational Rates for One Bit Matrix Completion

Renkun Ni
Department of Statistics
University of Virginia

Quanquan Gu
Department of Systems and Information Engineering
University of Virginia

Abstract

We consider one bit matrix completion under rank constraint. We present an estimator based on rank constrained maximum likelihood estimation, and an efficient greedy algorithm to solve it approximately based on an extension of conditional gradient descent. The output of the proposed algorithm converges at a linear rate to the underlying true low-rank matrix up to the optimal statistical estimation error rate, i.e., $\mathcal{O}(\sqrt{rn\log(n)/|\Omega|})$, where n is the dimension of the underlying matrix and $|\Omega|$ is the number of observed entries. Our work establishes the first computationally efficient approach with provable guarantee for optimal estimation in one bit matrix completion. Our theory is supported by thorough numerical results.

1 Introduction

Matrix completion [7, 20, 18] has received increasing attention in the past decade, and it has wide applications in data mining and computer vision. For example, in the recommendation system/collaborative filtering, we aim to predict the unknown preference of a set of users on a set of items, provided partial observed ratings. The most popular methods for matrix completion are based on empirical risk minimization with nuclear norm penalty [20, 18]. Other types of estimators with nonconvex penalties [12], max norm [5] and rank constraints [24, 26] have also been investigated. In some applications, the observation matrix is in terms of single bit, and thus standard matrix completion cannot be applied. To overcome this problem, Davenport et al. [9] proposed one bit matrix completion, which recovers a low rank matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ from a set of $|\Omega|$ noisy sign measurements. Particularly, they proposed a trace norm constrained maximum likelihood estimator. They showed that their estimator is minimax rate optimal under the uniform sampling model. Later on, Cai and Zhou [6] proposed a max norm [25] constrained maximum likelihood estimator for one bit matrix completion under a general (e.g., non-uniform) sampling model. They also proved that the estimation error bound is rate optimal because it matches the minimax lower bound established under the general sampling model. It is also worth noting that one bit matrix completion is relevant to one bit compressed sensing, which was originally proposed by Boufounos and Baraniuk [3] and has been widely studied in recent work [23, 28, 30].

While one bit matrix completion has been studied under different constraints [9, 6, 2] in theory, existing optimization algorithms for one bit matrix completion remain heuristic and lack of provable guarantee. For example, Cai and Zhou [6] proposed a simple first order method which is a special case of the projected gradient algorithm for solving the max norm constrained convex program. The non-monotonic spectral projected gradient (SPG) algorithm was implemented in [9] to solve the one bit matrix completion problem under nuclear norm constraint. However, neither of the algorithms has been proved to converge for one bit matrix completion.

In this paper, to overcome the limitations of existing methods, we present a new estimator for one bit matrix completion with rank constraint under uniform random sampling model. Note that the resulting optimization problem for our proposed estimator is non-convex and in general NP hard. In order to obtain an approximate solution to the estimator, we propose a greedy algorithm based on an extension of conditional gradient descent [14, 10, 8, 4]. We rigorously prove the statistical and computational error bounds for this rank constrained maximum likelihood estimator. In particular, we show that the statistical estimation error bound of proposed estimator is $\mathcal{O}(\sqrt{rn\log(n)/|\Omega|})$, which at-

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

tains the minimax lower bound [6] up to logarithmic factor. In addition, we show that the output of the proposed greedy algorithm converges to the true low-rank matrix at a linear rate up to the statistical error. Therefore, our proposed estimator and algorithm attain both optimal statistical and computational rates for one bit matrix completion. The numerical results on both simulation and real world datasets show that our proposed estimator consistently outperforms the exiting estimators with max norm and nuclear norm constraints.

The remainder of this paper is organized as follows. In Section 2, we present an estimator for one bit matrix completion with rank constraint. Also, an extension of conditional gradient descent algorithm is proposed to solve the optimization problem corresponding to one bit matrix completion. Section 3 is devoted to the main theory of the proposed estimator and algorithm. Both statistical and optimization error bounds are established. In Section 4, we outline the proof of some key theoretical results. Numerical simulations are provided in Section 5. We conclude the paper in Section 6.

Notation For a set Ω, we denote by $|\Omega|$ the cardinality of this set. For a vector $\mathbf{x} \in \mathbb{R}^m$, we denote by $\|\mathbf{x}\|_p = (\sum_{i=1}^m |x_i|^p)^{1/p}$ its ℓ_p norm, and denote by $\|\mathbf{x}\|_\infty = \max_{i=1,\ldots,d} |x_i|$ its ℓ_∞ norm. The inner product is denoted by $\langle \cdot, \cdot \rangle$. For a matrix $\mathbf{X} = (X_{ij}) \in \mathbb{R}^{m\times n}$, we use $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n X_{ij}^2}$ to denote its Frobenius norm and let $\|\mathbf{X}\|_\infty = \max_{i,j} |X_{ij}|$ be the elementwise ℓ_∞ norm. We use $\|\mathbf{X}\|_2 = \sigma_1(\mathbf{X})$ to denote the spectral norm, where $\sigma_1(\mathbf{X})$ is the largest singular value of matrix $\mathbf{X}$. We use $\|\mathbf{X}\|_* = \sum_{i=1}^{d_1} \sigma_i(\mathbf{X})$ to denote the nuclear norm, where $d_1 = \min(m, n)$.

2 The Proposed Estimator and Algorithm

In this section, we introduce the probabilistic model for one bit matrix completion, followed by the rank constrained maximum likelihood estimator. We also propose an algorithm for solving the corresponding optimization problem.

2.1 Probabilistic Models

One bit matrix completion is a problem that recovers the underlying low-rank matrix based on a subset of one bit measurements. Instead of observing actual entries from the underlying matrix $\mathbf{X}^* \in \mathbb{R}^{m\times n}$, the sign of a random subset of the noisy entries of $\mathbf{X}^*$ is observed. In this paper, without loss of generality, we assume $m \le n$. In addition, we consider one bit matrix completion under uniform random sampling model. Given a low rank matrix $\mathbf{X}^* \in \mathbb{R}^{m\times n}$ with $\text{rank}(\mathbf{X}^*) = r$, a probability mass function $p(\cdot): \mathbb{R} \to [0,1]$ and a subset of indices $\Omega \subseteq (m \times n)$, we observe a subset of binary matrix $\mathbf{Y}$ depending on $\mathbf{X}^*$ by the following probabilistic model:

$$Y_{ij} = \begin{cases} +1, & \text{with probability } p(X_{ij}^*), \\ -1, & \text{with probability } 1 - p(X_{ij}^*), \end{cases} \forall (ij) \in \Omega. \tag{2.1}$$

If $p(\cdot)$ is the cumulative distribution function of $-Z_{ij}$, where $\mathbf{Z} = (Z_{ij}) \in \mathbb{R}^{m\times n}$ is a noise matrix with i.i.d. entries, we can rewrite the above the model as

$$Y_{ij} = \begin{cases} +1, & \text{if } X_{ij}^* + Z_{ij} > 0, \\ -1, & \text{if } X_{ij}^* + Z_{ij} < 0, \end{cases} \forall (ij) \in \Omega. \tag{2.2}$$

There are many possible choices for the probability mass function $p(\cdot)$. Here we give two widely used examples: logistic function and probit function.
1. Logistic model: The logistic function is defined as $p(X_{ij}) = e^{X_{ij}}/(1 + e^{X_{ij}})$, which is equivalent to the fact that the i.i.d. noise Z_{ij} in (2.2) follows the standard logistic distribution.
2. Probit model: The probability function in (2.1) can be chosen as $p(X_{ij}) = \Phi\big(X_{ij}/\sigma\big)$, where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution $N(0,1)$. It is equivalent to the fact that the i.i.d. noise Z_{ij} in (2.2) follows Gaussian distribution $N(0,\sigma)$.

2.2 Constrained Maximum Likelihood Estimator

Given the probabilistic model in (2.1) for one bit matrix completion, we can estimate the underlying matrix $\mathbf{X}^*$ by minimizing the negative log-likelihood function under certain constraints. In particular, the negative log-likelihood function is

$$f_\Omega(\mathbf{X}) = -\sum_{(i,j)\in\Omega} \bigg\{ \mathbb{1}_{(Y_{ij}=1)} \log\big(p(X_{ij})\big) + \mathbb{1}_{(Y_{ij}=-1)} \log\big(1 - p(X_{ij})\big) \bigg\}, \tag{2.3}$$

where the probability mass function $p(X)$ can be selected as the examples we introduced before. Given different probability mass functions, we will achieve different likelihood functions, thus leading to different optimization problems.

In this paper, instead of using trace norm or max norm based convex relaxation for the rank [9, 6], we impose an exact rank constraint for one bit matrix completion. In addition, it is observed in [11] that if $\mathbf{X}^*$ is equal to zero in nearly all rows or columns, then it is impossible to recovery $\mathbf{X}^*$ unless all of its entries are sampled. In other words, there will always be some low-rank matrices which are too spiky 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 [22]. It is argued that the spikiness condition is much less restricted than the incoherence condition imposed in exact low-rank matrix completion [22, 17]. Thus, we consider the class of low-rank matrices with exact rank constraint and infinity norm constraint as follows

$$\mathcal{C}_{\mathrm{rank}}(\alpha, r) = \big\{\mathbf{X} \in \mathbb{R}^{m\times n} : \|\mathbf{X}\|_\infty \leq \alpha, \mathrm{rank}(\mathbf{X}) \leq r\big\}.$$

Given the above negative log-likelihood function and the class of low-rank matrices, we present an estimator for one bit matrix completion as follows

$$\widehat{\mathbf{X}} = \operatorname*{argmin}_{\mathbf{X}} f_\Omega(\mathbf{X}) \quad \text{subject to} \quad \mathbf{X} \in \mathcal{C}_{\mathrm{rank}}(\alpha, r). \tag{2.4}$$

Note that the rank constraint has been used for matrix completion [26, 24] and applied to many applications such as collaborative filtering and sensor network localization [16].

From (2.3), we can find that if the probability mass function $p(\mathbf{X})$ is log-concave, the objective function $f_\Omega(\mathbf{X})$ will be a convex function. In fact, both of the choices in the examples are log-concave functions. However, due to non-convexity of the set $\mathcal{C}_{\mathrm{rank}}(\alpha, r)$, the problem in (2.4) is a non-convex optimization problem and hard to solve. It is worth noting that a similar estimator was also studied in [2], where a heuristic algorithm was proposed to solve it without any provable guarantee. In what follows, we will present an efficient greedy algorithm based on a variant of conditional gradient descent for solving this problem with provable guarantee.

Algorithm 1 One Bit Matrix Completion with Rank Constraint via Conditional Gradient Descent

1: **Inialize:** $\mathbf{X}^0 = \mathbf{0}$
2: **for** $t = 0$ to T **do**
3: $\mathbf{V}^t \leftarrow \mathrm{argmin}_{\mathbf{V}\in\mathcal{D}} \langle \mathbf{V}, \nabla f_\Omega(\mathbf{X}^t)\rangle$
4: $(l_t, q_t) \leftarrow \mathrm{argmin}_{l_t\geq 0, q_t \geq 0} f_\Omega(l_t\mathbf{X}^t + q_t\mathbf{V}^t)$, subject to $\|l_t\mathbf{X}^t + q_t\mathbf{V}^t\|_\infty \leq \alpha$
5: $\sigma_i^{(t)} \leftarrow l_t \cdot \sigma_i^{(t-1)}, \forall i < t, \sigma_t^{(t)} \leftarrow q_t, \mathbf{A}^t \leftarrow \mathbf{V}^t$
6: $\mathbf{X}^{t+1} \leftarrow \sum_{i=0}^t \sigma_i^{(t)}\mathbf{A}^i$
7: **end for**

Problems with rank constraints can be solved by various algorithms such as heuristic methods [19], projected gradient methods [15] and conditional gradient descent [27, 29]. For the heuristic methods, most of them do not have optimality guarantees, and for projected gradient method, the computation is slow when the scale of problem is very large. Moreover, conditional gradient methods in [27, 29] cannot be applied directly to our problem since we imposed an infinity norm constraint on $\mathbf{X}^t$. In our paper, we propose an extension of conditional gradient descent algorithm to address the infinity norm constraint. The basic idea of conditional gradient descent is first iteratively looking for a search direction and then updating the current iterate. The algorithm is outlined in Algorithm 1. By this algorithm, we can guarantee the convergence to the global optimum $\widehat{\mathbf{X}}$ at a linear rate.

In detail, in step 3 of Algorithm 1, we start from linearizing the negative log-likelihood function $f_\Omega(\cdot)$ and finding a descent direction $\mathbf{V}^t$ by solving the following subproblem

$$\mathbf{V}^t = \operatorname*{argmin}_{\mathbf{V}\in\mathcal{D}} \langle \mathbf{V}, \nabla f_\Omega(\mathbf{X}^t)\rangle, \tag{2.5}$$

where $\mathcal{D}$ is the set of rank-1 matrices with unit Frobenius norm, i.e.,

$$\mathcal{D} = \{\mathbf{X} : \mathrm{rank}(\mathbf{X}) \leq 1, \|\mathbf{X}\|_F = 1\}.$$

Notice that each rank-one matrix $\mathbf{V}$ with unit Frobenius norm can be written as the product of two unit vectors, $\mathbf{V} = \mathbf{u}\mathbf{v}^\top$ for some $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathbb{R}^n$ with $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$. The subproblem in (2.5) can be equivalently solved as follows:

$$(\mathbf{u}^t, \mathbf{v}^t) = \operatorname*{argmin}_{\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1} \mathbf{u}^\top \nabla f_\Omega(\mathbf{X}^t)\mathbf{v}, \tag{2.6}$$

and set $\mathbf{V}^t = \mathbf{u}^t\mathbf{v}^{t\top}$. It is easy to show that the minimum value of (2.6) is $-\sigma_1^t$, where σ_1^t is the largest singular value of the matrix $\nabla f_\Omega(\mathbf{X}^t)$. In particular, suppose $(\mathbf{u}_1^t, \mathbf{v}_1^t)$ is a pair of left and right singular vectors corresponding to the largest singular value σ_1^t of the matrix $\nabla f_\Omega(\mathbf{X}^t)$. We can choose $\mathbf{u}^t = -\mathbf{u}_1^t$ and $\mathbf{v}^t = \mathbf{v}_1^t$ and therefore $\mathbf{V}^t = -\mathbf{u}_1^t\mathbf{v}_1^{t\top}$. This subproblem can be efficiently solved by power method in $\mathcal{O}(mn)$ time, where m, n is the dimension of the matrix. Note that similar subproblem has been formulated and solved in orthogonal rank-1 matrix pursuit [26] for standard matrix completion.

In step 4 of Algorithm 1, we find the optimal step size along the descent direction under the infinity norm constraint. Here, we have to add the infinity norm constraint in order to make sure every iterate satisfies the condition $\|\mathbf{X}\|_\infty \leq \alpha$. In particular, we aim at finding l_t and q_t such that

$$\mathbf{X}^{t+1} \leftarrow l_t\mathbf{X}^t + q_t\mathbf{V}^t,$$

minimizes the negative log-likelihood function and satisfies $\|l_t\mathbf{X}^t + q_t\mathbf{V}^t\|_\infty \leq \alpha$ simultaneously. This is a two dimensional optimization problem and can be solved very efficiently. It can be seen that, in Algorithm 1, the iterate $\mathbf{X}^{t+1}$ in each iteration is always belonging to the class $\mathcal{C}_{\text{rank}}(\alpha, r)$ when $t \leq r$. We will prove that Algorithm 1 is guaranteed to obtain the global optimum at a linear convergence rate in Section 3.

3 Main Theory

In this section, we are going to present the main theoretical results for the estimator proposed in Section 2.2. Before we lay out the main theoretical results, we first make several definitions and assumptions, which are essential to our theory.

Definition 3.1. Define a sampling operator $(\cdot)_\Omega$, which observes an index subset Ω of the entries from the underlying matrix $\mathbf{M}$,

$$\mathbf{M}_\Omega = \begin{cases} M_{ij}, & \text{if } (i,j) \in \Omega, \\ 0, & \text{if } (i,j) \notin \Omega. \end{cases}$$

We make the following assumption to obtain the lower and upper bounds on the eigenvalues of the Hessian for the objective function.

Assumption 3.2. Given the probability mass function $p(x)$, there exist $\gamma_\alpha, \mu_\alpha$ and L_α such that

$$\gamma_\alpha \leq \min\bigg(\inf_{|x|\leq\alpha}\bigg\{\frac{p'^2(x)}{p^2(x)} - \frac{p''(x)}{p(x)}\bigg\}, \inf_{|x|\leq\alpha}\bigg\{\frac{p'^2(x)}{(1-p(x))^2} + \frac{p''(x)}{1-p(x)}\bigg\}\bigg),$$
$$\mu_\alpha \geq \max\bigg(\sup_{|x|\leq\alpha}\bigg\{\frac{p'^2(x)}{p^2(x)} - \frac{p''(x)}{p(x)}\bigg\}, \sup_{|x|\leq\alpha}\bigg\{\frac{p'^2(x)}{(1-p(x))^2} + \frac{p''(x)}{1-p(x)}\bigg\}\bigg),$$
$$L_\alpha \geq \sup_{|x|\leq\alpha}\bigg\{\frac{|p'(x)|}{p(x)\big(1-p(x)\big)}\bigg\},$$

where α is the upper bound of the absolute value for every entry X_{ij}.

Here L_α reflects the steepness of function $f_\Omega(\cdot)$, μ_α and γ_α control the quadratic lower and upper bound on the second order Taylor expansion of $f_\Omega(\cdot)$. When α is a fixed constant, and $f_\Omega(\cdot)$ is specified, L_α, μ_α and γ_α are all fixed constants which do not depend on the dimension. For instance, we have $L_\alpha = 1$, $\gamma_\alpha = e^\alpha/(1+e^\alpha)^2$ and $\mu_\alpha = 1/4$ for the logistic model.

We are now ready to present an upper bound for our proposed estimator in the following theorem.

Theorem 3.3 (Statistical error)**.** Suppose Assumption 3.2 holds and $\mathbf{X}^* \in \mathcal{C}_{\text{rank}}(\alpha, r)$. A subset Ω of entries of underlying matrix $\mathbf{X}^*$ is sampled, and the binary matrix $\mathbf{Y}$ in (2.1) is generated based on the log-concave probability mass function $p(\mathbf{X})$. With probability at least $1 - (C_1+1)/n$, the optimal solution of (2.4) satisfies

$$\frac{\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F}{\sqrt{mn}} \leq C_2 \max(1, \alpha)\sqrt{\frac{rn\log(n)}{|\Omega|}},$$

where C_1, C_2 are universal constants.

It is important to note that although we have a rank constraint for the underlying matrix $\text{rank}(\mathbf{X}^*) \leq r$, the theorem still holds as r increases.

Remark 3.4. Under uniform sampling model [9] and general weighted sampling model [6], the estimation error bound of nuclear norm and max norm constrained estimators for one bit matrix completion is $\mathcal{O}(\sqrt{rn\log(n)/|\Omega|})$. Davenport et al. [9], Cai and Zhou [6] also proved the minimax lower bound for one bit matrix completion, which is $\mathcal{O}(\sqrt{rn/|\Omega|})$. From Theorem 3.3, our estimation error bound is $\mathcal{O}(\sqrt{rn\log(n)/|\Omega|})$, which matches the minimax lower bound up to a logarithmic factor $\log(n)$. Thus our proposed estimator is statistically optimal. Note that in [2], the authors also considered a rank constrained estimator. However, their theory only applies to a specific sampling model corresponding to a d-regular bipartite graph. The bipartite d-regular graph based sampling operator is too restrictive, since, in practice, it is difficult to guarantee that every row of $\mathbf{X}^*$ has d entries being observed. Moreover, their estimation error bound is $\mathcal{O}(m^3\sqrt{r^3 n}/|\Omega|^2)$, which does not match the minimax lower bound and thus it is not optimal.

In Theorem 3.3, we have shown that our statistical error bound attains the minimax lower bound, which is the same as nuclear and max norm constrained estimators in [9] and [6]. Next, we will further analyze the optimization error of Algorithm 1, which is used to solve our proposed estimator in (2.4). We will show that it is guaranteed to converge to the global optimum at a linear convergence rate in the following theorem.

Theorem 3.5 (Optimization error)**.** Let $\widehat{\mathbf{X}} = \sum_{i=0}^t \sigma_i^* \mathbf{A}_i^*$ be an optimal solution to the problem (2.4) of minimizing the negative log-likelihood function, and $\{\mathbf{X}^t\}$ is a sequence of iterates generated by Algorithm 1. Denote $s^* = \sum_{\mathbf{X}\in\mathcal{S}^*} \sigma_i^*$. Then, for any $t \geq 1$, we have

$$f_\Omega(\mathbf{X}^{t+1}) - f_\Omega(\widehat{\mathbf{X}}) \leq \bigg(1 - \min\bigg(\frac{1}{2}, \frac{2\gamma_\alpha(s^*)^2}{\mu_\alpha D^2 k}\bigg)\bigg)^{t+1} \cdot \big(f_\Omega(\mathbf{X}^0) - f_\Omega(\widehat{\mathbf{X}})\big),$$
$$\|(\mathbf{X}^{t+1} - \widehat{\mathbf{X}})_\Omega\|_F^2 \leq \frac{2}{\gamma_\alpha}\bigg(1 - \min\bigg(\frac{1}{2}, \frac{2\gamma_\alpha(s^*)^2}{\mu_\alpha D^2 k}\bigg)\bigg)^{t+1} \cdot \big(f_\Omega(\mathbf{X}^0) - f_\Omega(\widehat{\mathbf{X}})\big),$$

where μ_α and γ_α are defined in Assumption 3.2, $D = \max_{\mathbf{V},\mathbf{X}\in\mathcal{D}} \|\mathbf{X}-\mathbf{V}\|_F^2$, and k is the number of non-zero entries in $\{\sigma_i^*\}$.

Theorem 3.5 characterizes the optimization error bounds for both the objective function value and the iterate restricted on the observed entries. Note that in both bounds, the last term $f_\Omega(\mathbf{X}^0) - f_\Omega(\widehat{\mathbf{X}})$ is a known constant and the term $(1-\min\{1/2, 2\gamma_\alpha(s^*)^2/\mu_\alpha D^2 k\})$ is strictly smaller than one. Thus, when t increases, the bound decays to zero exponentially.

As a direct consequence of Theorem 3.3 and Theorem 3.5, the following theorem summarizes the total estimation error for the output of Algorithm 1.

Theorem 3.6. Suppose Assumption 3.2 holds and $\mathbf{X}^* \in \mathcal{C}_{\text{rank}}(\alpha, r)$. Given the binary matrix (2.1) for one bit matrix completion and index subset Ω sampled from uniform model with probability $|\Omega|/mn$. $\{\mathbf{X}^t\}$ is a sequence of iterates generated by Algorithm 1. For any $t \geq 1$, with probability at least $1-(C_1+1)/n$, we have

$$\frac{\|\mathbf{X}^{t+1}-\mathbf{X}^*\|_F}{\sqrt{mn}} \leq \underbrace{C_2 \max(1,\alpha)\sqrt{\frac{rn\log n}{|\Omega|}}}_{\text{statistical error}} + \underbrace{C_3\sqrt{\frac{1}{|\Omega|\gamma_\alpha}\kappa^{t+1}\big(f_\Omega(\mathbf{X}^0) - f_\Omega(\widehat{\mathbf{X}})\big)}}_{\text{optimization error}},$$

where $\kappa = \big(1-\min\{1/2, (2\gamma_\alpha(s^*)^2)/(\mu_\alpha D^2 k)\}\big)$, and C_1, C_2, C_3 are universal constants.

4 Proof of The Main Theory

In this section, we present the sketched proof of Theorem 3.3 in previous section. The proofs of the other theorems are deferred in the supplemental material.

We begin with a lemma, which shows that the negative likelihood function $f_\Omega(\cdot)$, satisfies the restricted strong convexity (RSC) condition and restricted strong smoothness (RSS) condition over the set $\mathcal{C}_{\text{rank}}(\alpha, r)$.

Lemma 4.1. Under Assumption 3.2, the negative likelihood function $f_\Omega(\mathbf{X})$ in the one bit matrix completion satisfies restricted strong convexity and smoothness condition over the set $\mathcal{C}_{\text{rank}}(\alpha, r)$:

$$f_\Omega(\mathbf{X}) \geq f_\Omega(\mathbf{M}) + \langle \nabla f_\Omega(\mathbf{M}), \mathbf{X}-\mathbf{M}\rangle + \frac{\gamma_\alpha}{2}\|(\mathbf{X}-\mathbf{M})_\Omega\|_F^2,$$
$$f_\Omega(\mathbf{X}) \leq f_\Omega(\mathbf{M}) + \langle \nabla f_\Omega(\mathbf{M}), \mathbf{X}-\mathbf{M}\rangle + \frac{\mu_\alpha}{2}\|(\mathbf{X}-\mathbf{M})_\Omega\|_F^2,$$

where γ_α, μ_α are defined in Assumption 3.2.

In view of Lemma 4.1, we obtain a lower bound for the second order term in Taylor's expansion of the negative log-likelihood function. Since we want to get the bound of statistical error $\|\widehat{\mathbf{X}}-\mathbf{X}^*\|_F$, we still need to obtain an upper bound for the first order term in Taylor's expansion. In what follows, we will show that the bound holds with high probability.

Lemma 4.2. Suppose Assumption 3.2 holds. For any $\mathbf{M} \in \mathcal{C}_{\text{rank}}(\alpha, r)$, with probability at least $1-1/n$, we have

$$\|\nabla f_\Omega(\mathbf{M})\|_2 \leq 2L_\alpha\sqrt{\frac{3|\Omega|\log(n)}{m}},$$

where L_α is defined in Assumption 3.2.

In what follows, we will establish the relationship between the statistical error on the whole matrix and the statistical error restricted on the observed entries $\|(\widehat{\mathbf{X}}-\mathbf{X}^*)_\Omega\|_F$. In order to show the relationship, we need to introduce spikiness ratio and low-rank ratio, which have been previously defined in in [22].

Definition 4.3. Let $\mathbf{X} \in \mathbb{R}^{m\times n}$, we define the spikiness ratio $\alpha_{\text{sp}}(\mathbf{X}) = (\sqrt{mn}\|\mathbf{X}\|_\infty)/\|\mathbf{X}\|_F$, and define the low-rank ratio $\beta_{\text{ra}}(\mathbf{X}) = \|\mathbf{X}\|_*/\|\mathbf{X}\|_F$.

From this definition, we can observe that $\alpha_{\text{sp}}(\mathbf{X})$ reflects the spikiness of a certain matrix and the rank ratio β_{ra} is a tractable measurement of how close the underlying matrix $\mathbf{X}$ is to a low-rank matrix. Based on these two ratios and the notation $d = (m+n)/2$, we will present another constraint set

$$\mathcal{C}'(|\Omega|, C_0) = \Big\{\Delta \in \mathbb{R}^{m\times n}, \Delta \neq 0 \,\Big|\, \alpha_{\text{sp}}(\Delta)\beta_{\text{ra}}(\Delta) \leq \frac{1}{C_0}\sqrt{|\Omega|/d\log(d)}\Big\},$$

where $|\Omega|$ is the sample size and C_0 is a constant. Note that as the simple size $|\Omega|$ increases, this set allows for matrices with larger values of spikiness. The following lemma characterizes the relationship between the Frobenius norm of a matrix and its Frobenius norm restricted on a subset of entries, for any matrices belong to the constraint set $\mathcal{C}'(|\Omega|, C_0)$.

Lemma 4.4. [22] There are universal constants C, C_0, C_1 such that as long as $|\Omega| > Cd\log(d)$, we have, for all $\Delta \in \mathcal{C}'(|\Omega|, C_0)$,

$$\frac{\|\Delta_\Omega\|_F}{\sqrt{|\Omega|}} \geq \frac{1}{8}\frac{\|\Delta\|_F}{\sqrt{mn}}\Big\{1 - \frac{128\alpha_{\text{sp}}(\Delta)}{\sqrt{|\Omega|}}\Big\}, \qquad (4.1)$$

with probability at least $1-C_1/n$.

Note that, this bound guarantees that the sampling operator obtains a substantial component of any matrix $\Delta \in \mathcal{C}'(|\Omega|, C_0)$, which is not overly spiky. In particular, if $128\alpha_{\text{sp}}(\Delta)/\sqrt{|\Omega|} \leq 1/2$, the bound in (4.1) implies that $\|\Delta_\Omega\|_F^2/|\Omega| \geq \|\Delta\|_F^2/(256mn)$.

Now we are ready to prove Theorem 3.3.

Proof of Theorem 3.3. According to the RSC condition in Lemma 4.1, we have

$$
\begin{aligned}
& f_\Omega(\widehat{\mathbf{X}}) \\
& \geq f_\Omega(\mathbf{X}^*) + \langle \nabla f_\Omega(\mathbf{X}^*), \widehat{\mathbf{X}} - \mathbf{X}^* \rangle + \frac{\gamma_\alpha}{2} \|(\widehat{\mathbf{X}} - \mathbf{X}^*)_\Omega\|_F^2 \\
& \geq f_\Omega(\mathbf{X}^*) - \|\nabla f_\Omega(\mathbf{X}^*)\|_2 \cdot \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_* + \frac{\gamma_\alpha}{2} \|(\widehat{\mathbf{X}} - \mathbf{X}^*)_\Omega\|_F^2 \\
& \geq f_\Omega(\mathbf{X}^*) - \sqrt{2r} \|\nabla f_\Omega(\mathbf{X}^*)\|_2 \cdot \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F \\
& \quad + \frac{\gamma_\alpha}{2} \|(\widehat{\mathbf{X}} - \mathbf{X}^*)_\Omega\|_F^2,
\end{aligned}
$$

where the second inequality follows by Hölder's inequality. By the definition of $\widehat{\mathbf{X}} = \mathrm{argmin}_{\mathbf{X}} f_\Omega(\mathbf{X})$, we obtain

$$
\begin{aligned}
& 0 \geq f_\Omega(\widehat{\mathbf{X}}) - f_\Omega(\mathbf{X}^*) \\
& \quad \geq -\sqrt{2r} \|\nabla f_\Omega(\mathbf{X}^*)\|_2 \cdot \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F + \frac{\gamma_\alpha}{2} \|(\widehat{\mathbf{X}} - \mathbf{X}^*)_\Omega\|_F^2.
\end{aligned} \tag{4.2}
$$

From Lemma 4.4, we have the relation between the Frobenius norm of the observed entries and the Frobenius norm based on the whole matrix over the constraint set $\mathcal{C}'(|\Omega|, C_0)$. Now apply Lemma 4.4 to (4.2), and consider the following cases for this equation.

Case 1: Suppose $\widehat{\mathbf{X}} - \mathbf{X}^* \notin \mathcal{C}'(|\Omega|, C_0)$, then according to the definition of $\mathcal{C}'(|\Omega|, c_0)$ we have,

$$
\alpha_{\mathrm{sp}}(\widehat{\mathbf{X}} - \mathbf{X}^*) \beta_{\mathrm{ra}}(\widehat{\mathbf{X}} - \mathbf{X}^*) \geq \frac{1}{C_0} \sqrt{\frac{|\Omega|}{d \log(d)}}.
$$

Based on the definition of $\alpha_{\mathrm{sp}}(\widehat{\mathbf{X}} - \mathbf{X}^*)$ and $\beta_{\mathrm{ra}}(\widehat{\mathbf{X}} - \mathbf{X}^*)$, we can rewrite the above inequality,

$$
\sqrt{mn} \frac{\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_* \cdot \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_\infty}{\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2} \geq \frac{1}{C_0} \sqrt{\frac{|\Omega|}{d \log(d)}}.
$$

By reorganizing the inequality, we can obtain

$$
\begin{aligned}
& \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2 \\
& \leq \sqrt{mn} \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_* \cdot \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_\infty C_0 \sqrt{\frac{d \log(d)}{|\Omega|}} \\
& \leq 2\sqrt{2mnr} \alpha C_0 \sqrt{\frac{d \log(d)}{|\Omega|}} \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F,
\end{aligned}
$$

where the last inequality holds as the fact that $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_* \leq \sqrt{2r} \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F$ and the fact $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_\infty \leq 2\alpha$. Thus, we can achieve the statistical error bound in Case 1 as follows,

$$
\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F \leq 2\sqrt{2} \alpha C_0 n \sqrt{\frac{rm \log(n)}{|\Omega|}}, \tag{4.3}
$$

where we used the fact $d = (m+n)/2$ and $m \leq n$.

For the case $\widehat{\mathbf{X}} - \mathbf{X}^* \in \mathcal{C}'(|\Omega|, C_0)$, we will have another two more sub cases.

Case 2.1: Suppose $\widehat{\mathbf{X}} - \mathbf{X}^* \in \mathcal{C}'(|\Omega|, C_0)$, and $128\alpha_{\mathrm{sp}}(\widehat{\mathbf{X}} - \mathbf{X}^*)/\sqrt{|\Omega|} > 1/2$, then we obtain

$$
\frac{128\alpha_{\mathrm{sp}}(\widehat{\mathbf{X}} - \mathbf{X}^*)}{\sqrt{|\Omega|}} = \frac{128\sqrt{mn} \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_\infty}{\sqrt{|\Omega|} \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F} > \frac{1}{2}.
$$

By rearranging the above inequality, we obtain the bound

$$
\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F < \frac{512\sqrt{mn}\alpha}{\sqrt{|\Omega|}}. \tag{4.4}
$$

Case 2.2: Suppose $\widehat{\mathbf{X}} - \mathbf{X}^* \in \mathcal{C}'(|\Omega|, C_0)$, and $128\alpha_{\mathrm{sp}}(\widehat{\mathbf{X}} - \mathbf{X}^*)/\sqrt{|\Omega|} \leq 1/2$, taking this condition into Lemma 4.4 yields

$$
\frac{\|(\widehat{\mathbf{X}} - \mathbf{X}^*)_\Omega\|_F}{\sqrt{|\Omega|}} \geq \frac{1}{16} \frac{\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F}{\sqrt{mn}},
$$

with probability at least $1 - (C_1+1)/n$. By reorganizing the inequality, we have

$$
\|(\widehat{\mathbf{X}} - \mathbf{X}^*)_\Omega\|_F \geq \frac{\sqrt{|\Omega|}}{16\sqrt{mn}} \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F. \tag{4.5}
$$

Combining (4.2), (4.5) and Lemma 4.2, we achieve

$$
\begin{aligned}
& 0 \geq f_\Omega(\widehat{\mathbf{X}}) - f_\Omega(\mathbf{X}^*) \\
& \quad \geq -(2\sqrt{2r} L_\alpha \sqrt{3|\Omega| \log(n)/m}) \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F \\
& \quad + \frac{|\Omega|\gamma_\alpha}{32mn} \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2 \\
& \quad = \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F \\
& \qquad \cdot \left(-2\sqrt{2r} L_\alpha \sqrt{\frac{3|\Omega| \log(n)}{m}} + \frac{|\Omega|\gamma_\alpha}{32mn} \|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F \right).
\end{aligned}
$$

We have to let $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F \leq (2\sqrt{2r} L_\alpha \sqrt{3|\Omega| \log(n)/m})/(|\Omega|\gamma_\alpha/(32mn))$ such that $f_\Omega(\widehat{\mathbf{X}}) - f_\Omega(\mathbf{X}^*) \leq 0$. As a conclusion for all the cases, we have

$$
\begin{aligned}
\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F \leq \max \Bigg(& 2\sqrt{2} \alpha C_0 n \sqrt{\frac{rm \log(n)}{|\Omega|}}, \frac{512\sqrt{mn}\alpha}{\sqrt{|\Omega|}}, \\
& 64\sqrt{6} L_\alpha n \sqrt{\frac{rm \log(n)}{|\Omega|}} \Bigg),
\end{aligned} \tag{4.6}
$$

with probability at least $1 - C_1/n$. Compare the three terms in (4.6), we can see that both the first and the third terms have the same order, which is slower than the second term $512\sqrt{mn}\alpha/\sqrt{|\Omega|}$. As a result, we have with probability at least $1 - (C_1+1)/n$ that

$$
\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F \leq C_2 \max(1, \alpha) n \sqrt{\frac{rm \log(n)}{|\Omega|}},
$$

where C_2 is a universal constant. This completes the proof. $\square$

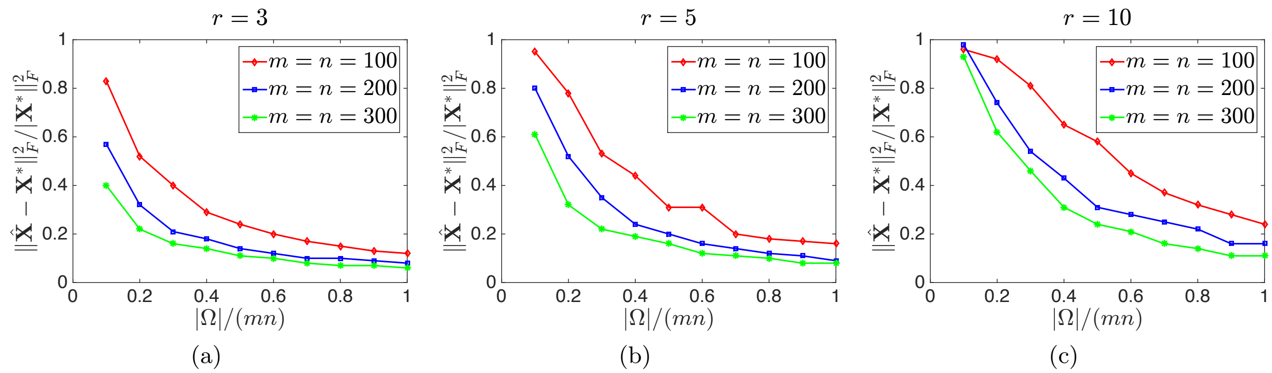

Figure 1: Simulation results for one bit matrix completion. The dimension of matrix is $m \times n$, and the rank is $r \in \{3, 5, 10\}$. Figure 1(a)- 1(c) correspond to the performance of our proposed estimator for different dimensions under the same rank.

5 Experiments

In this section, we will evaluate the performance of our proposed estimator for one bit matrix completion on both synthetic and real datasets, and compare it with nuclear norm constrained estimator [9] and max norm constrained estimator [6].

5.1 Simulation

We begin by investigating the performance of our proposed estimator over a range of different dimensions of the underlying matrix $\mathbf{X}^*$. First, we construct the underlying matrix by $\mathbf{X}^* = \mathbf{U}\mathbf{V}^\top$, where both $\mathbf{U} \in \mathbb{R}^{m\times r}, \mathbf{V} \in \mathbb{R}^{n\times r}$ are generated randomly by uniform distribution $[-1/2, 1/2]$, such that $\text{rank}(\mathbf{X}^*) = r$. Then we scaled the underlying matrix by $\|\mathbf{X}^*\|_\infty$ such that the infinity-norm of $\mathbf{X}^*$ equals one. Here, we choose to work with the Probit model under uniform sampling, namely $p(X_{ij}) = \Phi(\mathbf{X}_{ij}/\sigma)$. In this experiment, we set rank $r \in \{3, 5, 10\}$, dimension $m = n \in \{100, 200, 300\}$ and the noise at a moderate level $\sigma = 0.18$. For every settings, we repeat 20 trials and measure the performance of the estimator using the squared Frobenius norm of the estimation error normalized by the squared Frobenius norm of underlying matrix, i.e., $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/\|\mathbf{X}^*\|_F^2$ over all trials. The numerical results are shown in Figure 1(a)-1(c). We can naturally observe that the normalized error decreases as the percentage of observed entries ($|\Omega|/mn$) increases in every situation. Figure 1 also shows that with the same percentage of observed entries, the normalized error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/\|\mathbf{X}^*\|_F^2$ decreases as the dimension m, n increases, which is consistent with the convergence rate $\mathcal{O}(\sqrt{(rn)/(|\Omega|)})$ we obtained in Section 3.

We also conducted experiments comparing the performance of rank, max norm and nuclear norm constrained estimators for one bit matrix completion by using the same criterion as in [6] and [9]. We plot the normalized error versus the percentage of observed entries for three different ranks $r \in \{5, 10, 15\}$, three different matrix dimensions $m = n \in \{100, 200, 300\}$ and versus a range of different values of noise level σ on a logarithmic scale. Each setting is repeated for 20 times. According to Figure 2(a)-2(e), our proposed estimator performs consistently better than both trace norm and max norm based estimators, especially under the case $r = 10, m = n = 100$. From Figure 2(f), when the noise level σ is really small, all of the performances of the estimators are poor. However, rank constrained estimator performs consistently better than the other two methods except when the noise level equals $10^{-1.5}$. This again verifies the advantage of our proposed estimator and backs up our theory.

5.2 Real Data

In this experiment, we applied our proposed estimator to collaborative filtering on real-world dataset, and compare the performance with some existing methods, including nuclear norm and max norm constrained one bit matrix completion, as well as standard matrix completion method [1]. We use the MovieLens (100k) dataset, which is available for download at `http://www.grouplens.org/node/73`. This dataset contains 100,000 movie ratings from 943 users on 1682 movies, and each rating occurs on a scale from 1 to 5. We use the logistic model to generate one bit data from the rating matrix. Our goal is to recover the underlying rating matrix. We use 50%, 70% and 90% entries

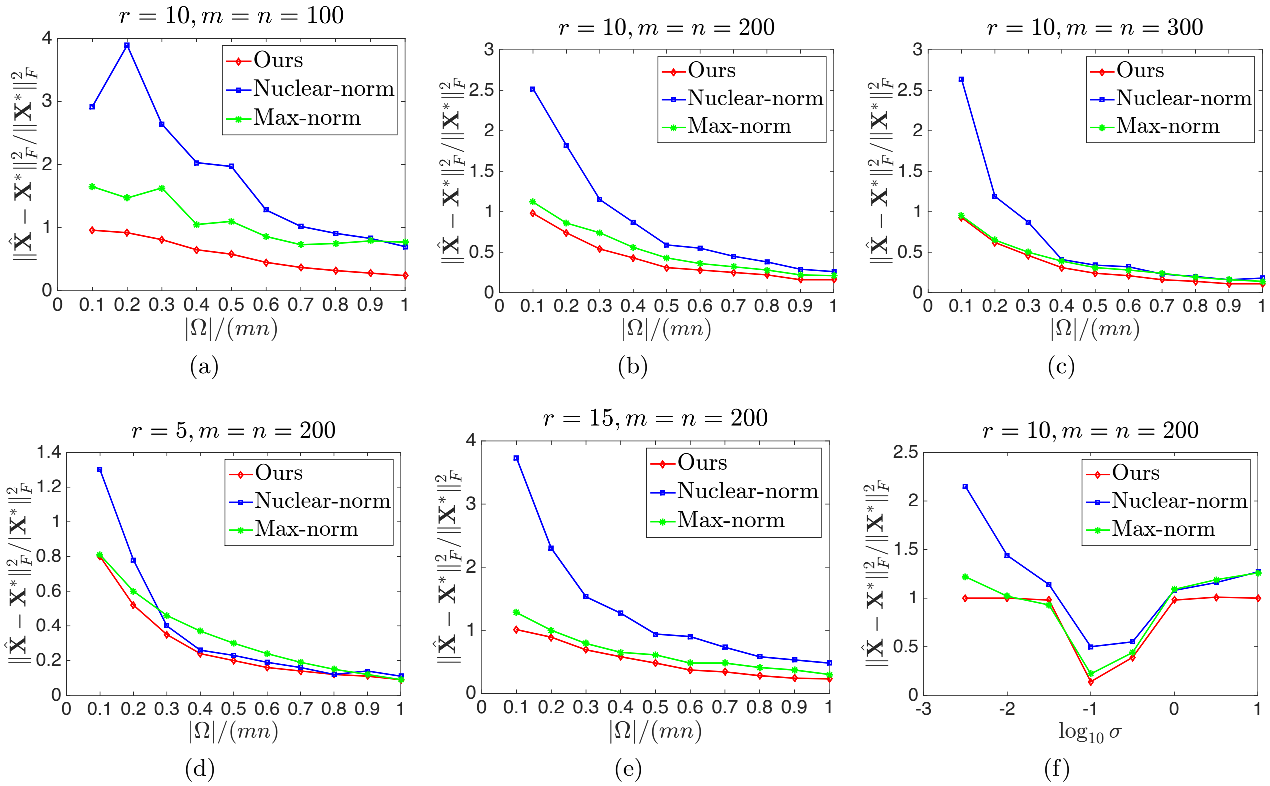

Figure 2: Plot for the performance of rank constraint, max norm and nuclear norm based methods under different cases $r \in \{5, 10, 15\}$ and $m = n \in \{100, 200, 300\}$. Figure 2(a)-2(e) plot the normalized error with respect to the percentage of observed entries for different $r \in \{5, 10, 15\}$ and $m = n \in \{100, 200, 300\}$. Figure 2(f) plots the normalized error versus noise level on logarithm scale, where $r = 10$ and $m = n = 200$.

of the one bit data as training data, and use the rest entries of the one bit data as test data. Each trail is repeated 10 times. Here, we use mean absolute error (MAE) to evaluate the performance of each estimator. MAE is defined as MAE $= \sum_{(i,j)\in\mathcal{T}} |X_{ij} - \widehat{X}_{ij}|/|\mathcal{T}|$, where X_{ij} denotes the rating user i gives to the item j, $\widehat{X}_{ij}$ denotes the predicted rating user i gives to the item j and $\mathcal{T}$ is the index set for the testing dataset. The results are shown in Table 1. From Table 1, we can find that standard matrix completion method performs worst in recovering the underlying rating matrix, since it is not specifically designed for matrix recovery based on one-bit information. Among the one bit matrix completion methods, our estimator consistently performs better than nuclear norm and max norm constrained estimators, which is consistent with the findings on the simulation datasets and backup our theory.

6 Conclusions

In this paper, we proposed a unified framework for one bit matrix completion with rank constraints under uniform random sampling model. We establish a unified estimation error analysis for one bit matrix completion, which integrates the statistical error of the estimator and the optimization error of the algorithm. Experiments on both synthetic datasets and real datasets indicate that the proposed estimator performs better than existing estimators, which is consistent to the theoretical analysis.

Table 1: Quantitative comparison of different estimators on the MovieLens dataset in terms of MAE.

Method	50% train	70% train	90% train
Standard MC	4.41±0.65	3.76±0.57	2.83±0.48
Nuclear-norm	2.12±0.04	1.86±0.05	1.69±0.01
Max-norm	2.62±0.08	1.88±0.03	1.78±0.01
Ours	1.96±0.05	1.73±0.02	1.43±0.03

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. Research was sponsored by Gu's startup funding at Department of Systems and Information Engineering, University of Virginia.

References

[1] Stephen R Becker, Emmanuel J Candès, and Michael C Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.

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

[3] Petros T Boufounos and Richard G Baraniuk. 1-bit compressive sensing. In *42nd Annual Conference on Information Sciences and Systems (CISS)*, pages 16–21. IEEE, 2008.

[4] Kristian Bredies, Dirk A Lorenz, and Peter Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Applications*, 42(2):173–193, 2009.

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

[6] Tony Cai and Wen-Xin Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research*, 14(1):3619–3647, 2013.

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

[8] Kenneth L Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):63, 2010.

[9] Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

[10] Joseph C Dunn and S Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.

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

[12] Huan Gui and Quanquan Gu. Towards faster rates and oracle property for low-rank matrix estimation. *arXiv preprint arXiv:1505.04780*, 2015.

[13] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

[14] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML*, pages 427–435, 2013.

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

[16] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[17] Olga Klopp et al. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

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

[19] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD*, pages 426–434. ACM, 2008.

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

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

[22] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

[23] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

[24] Shai Shalev-Shwartz, Alon Gonen, and Ohad Shamir. Large-scale convex minimization with a low-rank constraint. *arXiv preprint arXiv:1106.1622*, 2011.

[25] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *NIPS*, pages 1329–1336, 2004.

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

[27] Adams Wei Yu, Wanli Ma, Yaoliang Yu, Jaime Carbonell, and Suvrit Sra. Efficient structured matrix rank minimization. In *NIPS*, pages 1350–1358, 2014.

[28] Lijun Zhang, Jinfeng Yi, and Rong Jin. Efficient algorithms for robust one-bit compressive sensing. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 820–828, 2014.

[29] Xinhua Zhang, Dale Schuurmans, and Yao-liang Yu. Accelerated training for matrix-norm regularization: A boosting approach. In *NIPS*, pages 2906–2914, 2012.

[30] Rongda Zhu and Quanquan Gu. Towards a lower sample complexity for robust one-bit compressed sensing. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 739–747, 2015.