PU Learning for Matrix Completion

Cho-Jui Hsieh CJHSIEH@CS.UTEXAS.EDU
Nagarajan Natarajan NAGA86@CS.UTEXAS.EDU
Inderjit S. Dhillon INDERJIT@CS.UTEXAS.EDU
Department of Computer Science, The University of Texas, Austin, TX 78721, USA

Abstract

In this paper, we consider the matrix completion problem when the observations are one-bit measurements of some underlying matrix M, and in particular the observed samples consist *only* of ones and no zeros. This problem is motivated by modern applications such as recommender systems and social networks where only "likes" or "friendships" are observed. The problem is an instance of PU (positive-unlabeled) learning, i.e. learning from only positive and unlabeled examples that has been studied in the context of binary classification. Under the assumption that M has bounded nuclear norm, we provide recovery guarantees for two different observation models: 1) M parameterizes a distribution that generates a binary matrix, 2) M is thresholded to obtain a binary matrix. For the first case, we propose a "shifted matrix completion" method that recovers M using only a subset of indices corresponding to ones; for the second case, we propose a "biased matrix completion" method that recovers the (thresholded) binary matrix. Both methods yield strong error bounds — if $M \in \mathbb{R}^{n\times n}$, the error is bounded as $O\big(\frac{1}{(1-\rho)n}\big)$, where $1-\rho$ denotes the fraction of ones observed. This implies a sample complexity of $O(n\log n)$ ones to achieve a small error, when M is dense and n is large. We extend our analysis to the inductive matrix completion problem, where rows and columns of M have associated features. We develop efficient and scalable optimization procedures for both the proposed methods and demonstrate their effectiveness for link prediction (on real-world networks consisting of over 2 million nodes and 90 million links) and semi-supervised clustering tasks.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

1. Introduction

The problem of recovering a matrix from a given subset of its entries arises in many practical problems of interest. The famous Netflix problem of predicting user-movie ratings is one example that motivates the traditional matrix completion problem, where we would want to recover the underlying (ratings) matrix given partial observations. Strong theoretical guarantees have been developed in the recent past for the low-rank matrix completion problem (Candès & Recht, 2009). An important variant of the matrix completion problem is to recover an underlying matrix from one-bit quantization of its entries. Modern applications of the matrix completion problem reveal a conspicuous gap between existing matrix completion theory and practice. For example, consider the problem of link prediction in social networks. Here, the goal is to recover the underlying friendship network from a given snapshot of the social graph consisting of observed friendships. We can pose the problem as recovering the adjacency matrix of the network A such that $A_{ij} = 1$ if users i and j are related and $A_{ij} = 0$ otherwise. In practice, we only observe positive relationships between users corresponding to 1's in A. Thus, there is not only one-bit quantization in the observations, but also a one-sided nature to the sampling process here — no "negative" entries are sampled. In the context of classification, methods for learning in the presence of positive and unlabeled examples only, called positive-unlabeled (PU in short) learning, have been studied in the past (Elkan & Noto, 2008; Liu et al., 2003). For matrix completion, can one guarantee recovery when only a subset of positive entries is observed? In this paper, we formulate the PU matrix completion problem and answer the question in the affirmative under different settings.

Minimizing squared loss on the observed entries corresponding to 1's, subject to the low-rank constraints, yields a degenerate solution — the rank-1 matrix with all its entries equal to 1 achieves zero loss. In practice, a popular heuristic used is to try and complete the matrix by treating some or all of the missing observations as true 0's, which seems to be a good strategy when the underlying matrix has a small number of positive examples, i.e., small number of 1's. This motivates viewing the problem of learning from only positive samples as a certain noisy matrix completion problem. Existing theory for noise-tolerant matrix completion (Candès & Plan, 2009; Davenport et al., 2012) does not sufficiently address recoverability under PU learning (see Section 2).

In our work, we assume that the true matrix $M \in \mathbb{R}^{m\times n}$ has a bounded nuclear norm $\|M\|_*$. The PU learning model for matrix completion is specified by a certain one-bit quantization process that generates a binary matrix Y from M and a one-sided sampling process that reveals a subset of positive entries of Y. In particular, we consider two recovery settings for PU matrix completion: The first setting is non-deterministic — M parameterizes a probability distribution which is used to generate the entries of Y. We show that it is possible to recover M using only a subset of positive entries of Y. The idea is to minimize an unbiased estimator of the squared loss between the estimated and the observed "noisy" entries, motivated by the approach in (Natarajan et al., 2013). We recast the objective as a "shifted matrix completion" problem that facilitates in obtaining a scalable optimization algorithm. The second setting is deterministic — Y is obtained by thresholding the entries of M, and then a subset of positive entries of Y is revealed. While recovery of M is not possible (see Section 2), we show that we can recover Y with low error. To this end, we propose a scalable biased matrix completion method where the observed and the unobserved entries of Y are penalized differently. Recently, an inductive approach to matrix completion was proposed (Jain & Dhillon, 2013) where the matrix entries are modeled as a bilinear function of real-valued features associated with the rows and the columns. We extend our methods under the two aforementioned settings to the inductive matrix completion problem and establish similar recovery guarantees. Our contributions are summarized below:

1. To the best of our knowledge, this is the first paper to formulate and study PU learning for matrix completion, necessitated by the applications of matrix completion. Furthermore, we extend our results to the recently proposed inductive matrix completion problem.
2. We provide strong guarantees for recovery; for example, in the non-deterministic setting, the error in recovering an $n \times n$ matrix is $O(\frac{1}{(1-\rho)n})$ for our method compared to $O(\frac{1}{(1-\rho)\sqrt{n}})$ implied by the method in (Davenport et al., 2012), where $(1-\rho)$ is the fraction of observed 1's.
3. Our results provide a theoretical insight for the heuristic approach used in practice, namely, biased matrix completion.
4. We give efficient, scalable optimization algorithms for our methods; experiments on real-world data (social networks consisting of over 2 million users and 90 million links) demonstrate the superiority of the proposed methods for the link prediction problem.

Outline of the paper. We begin by establishing some hardness results and describing our PU learning settings in Section 2. In Section 3, we propose methods and give recovery guarantees for the matrix completion problem under the different settings. We extend the results to PU learning for inductive matrix completion problem in Section 4. We describe efficient optimization procedures for the proposed methods in Section 5. Experimental results on synthetic and real-world data are presented in Section 6.

Related Work. In the last few years, there has been a tremendous amount of work on the theory of matrix completion since the remarkable result concerning recovery of low-rank matrices by (Candès & Recht, 2009). Strong results on recovery from noisy observations have also been established (Candès & Plan, 2009; Keshavan et al., 2010). Recently, (Davenport et al., 2012) studied the problem of recovering matrices from 1-bit observations, motivated by the nature of observations in domains such as recommender systems where matrix completion is heavily applied. Our work draws motivation from recommender systems as well, but differs from (Davenport et al., 2012) in that we seek to understand the case when only 1's in the matrix are observed. One of the algorithms we propose for PU matrix completion is based on using different costs in the objective for observed and unobserved entries. The approach has been used before, albeit heuristically, in the context of matrix completion in recommender system applications (Hu et al., 2008; Rendle et al., 2009; Sindhwani et al., 2010). Compressed sensing is a field that is closely related to matrix completion. Recently, compressed sensing theory has been extended to the case of single-bit quantization (Boufounos & Baraniuk, 2008). Here, the goal is to recover an s-sparse signal when the observations consist of only the signs of the measurements, and remarkable recovery guarantees have been proved for the single-bit quantization case (Boufounos & Baraniuk, 2008).

2. Problem Settings

We assume that the underlying matrix $M \in \mathbb{R}^{m\times n}$ has a bounded nuclear norm, i.e., $\|M\|_* \leq t$, where t is a constant independent of m and n. If $M_{ij} \in \{0,1\}$ for all (i,j), stating the PU matrix completion problem is straightforward: we only observe a subset Ω_1 randomly sampled from $\{(i,j) \mid M_{ij}=1\}$ and the goal is to recover M based on this "one-sided" sampling. We call this the "basic setting". However, in real world applications it is unlikely that the underlying matrix is binary. In the following, we consider two general settings, which include the basic setting as a special case.

2.1. Non-deterministic setting

In the non-deterministic setting, we assume M_{ij} has bounded values and without loss of generality we can assume $M_{ij} \in [0,1]$ for all (i,j) by normalizing it. We then consider each entry as a probability distribution which generates a **clean** 0-1 observation $Y \in \mathbb{R}^{m\times n}$:

$$P(Y_{ij}=1)=M_{ij},\ \ P(Y_{ij}=0)=1-M_{ij},$$

In the classical matrix completion setting, we will observe partial entries sampled randomly from Y; In our PU learning model, we assume only a subset of positive entries of Y is observed. More precisely, we observe a subset Ω_1 from Y where Ω_1 is sampled uniformly from $\{(i,j) \mid Y_{ij}=1\}$. We assume $|\Omega_1|=\bar{s}$ and denote the number of 1's in Y by s. With only Ω_1 given, the goal of PU matrix completion is to recover the underlying matrix M. Equivalently, letting $A\in\{0,1\}^{m\times n}$ to denote the observations, where $A_{\Omega_1}=1$ and $A_{ij}=0$ for all $(i,j)\notin\Omega_1$, the non-deterministic setting can be specified as observing A by the process:

$$P(A_{ij}\!=\!1)\!=\!M_{ij}(1\!-\!\rho),\ P(A_{ij}\!=\!0)\!=\!1\!-\!M_{ij}(1\!-\!\rho), \quad (1)$$

where $\rho\!=\!1\!-\!\bar{s}/s$ is the noise rate of flipping a 1 to 0 (or equivalently, $1\!-\!\rho$ is the sampling rate to obtain Ω_1 from Y).

Hardness of recovering M: The 1-bit matrix completion approach of (Davenport et al., 2012) can be applied to this setting — Given a matrix M, a subset Ω is sampled uniformly at random from M, and the observed values are "quantized" by a known probability distribution. We can transform our problem to the 1-bit matrix completion problem by assuming all the unobserved entries are zeros. When $M\in\mathbb{R}^{n\times n}$, the following maximum likelihood estimator can be used to recover the groundtruth:

$$\hat{M} = \operatorname*{argmax}_{X:\|X\|_*\le t}\Bigg(\sum_{(i,j)\in\Omega \text{ s.t. } Y_{ij}=1}\log(f(X_{ij})) + \sum_{(i,j)\in\Omega \text{ s.t. } Y_{ij}=0}\log(1-f(X_{ij}))\Bigg). \quad (2)$$

and the following error bound for the estimator can be obtained from the result in (Davenport et al., 2012):

$$\frac{1}{n^2}\|\hat{M}-M\|_F^2 = O\bigg(\frac{\sqrt{r}}{(1-\rho)\sqrt{n}}\bigg), \quad (3)$$

where $\mathrm{rank}(M)\le r$ (See Appendix B for details). The main drawback of using this approach for PU matrix completion is computation — time complexity of solving (32) is $O(n^2)$ which makes the approach prohibitive for large matrices. Moreover, the average error on each element is $O(1/\sqrt{n})$ (in contrast, our algorithm has $O(1/n)$ average error). To see how this affects sample complexity for recovery, assume $\sum_{i,j}M_{i,j}=O(n^2)$ (number of 1's are of the same order as the number of 0's in the original matrix) and $O(\log n)$ 1's are observed. Then $(1-\rho)=O(\frac{\log n}{n})$ and the average error according to (34) is $\frac{1}{n^2}\|\hat{M}-M\|_F^2 = O(\frac{\sqrt{rn}}{\log n})$, which diverges as $n\to\infty$. In contrast, the average error of our estimator will diminish to 0 as $n\to\infty$.

2.2. Deterministic setting

In the deterministic setting, a **clean** 0-1 matrix Y is observed from M by the thresholding process: $Y_{ij}=I(M_{ij}>q)$, where $I(\cdot)$ is the indicator function and $q\in\mathbb{R}$ is the threshold. Again, in our PU learning model, we assume only a subset of positive entries of Y are observed, i.e. we observe Ω_1 from Y where Ω_1 is sampled uniformly from $\{(i,j)\mid Y_{ij}=1\}$. Equivalently, we will use A to denote the observations, where $A_{ij}=1$ if $(i,j)\in\Omega_1$, and $A_{ij}=0$ otherwise.

It is impossible to recover M even if we observe all the entries of Y. A trivial example is that all the matrices $\eta\boldsymbol{e}\boldsymbol{e}^T$ will give $Y=\boldsymbol{e}\boldsymbol{e}^T$ if $\eta>q$, and we cannot recover η from Y. Therefore, in the deterministic setting we can only hope to recover the underlying 0-1 matrix Y from the given observations. To the best of our knowledge, there is no existing work that gives a reasonable guarantee of recovering Y. For example, if we apply the noisy matrix completion algorithm proposed in (Candès & Plan, 2009), the estimator $\hat{Y}$ has an error bound $\|\hat{Y}-Y\|\le\|A-Y\|$, which indicates the error in $\hat{Y}$ is not guaranteed to be better than the trivial estimator A (see Appendix C for details).

3. Proposed Algorithms for PU Matrix Completion

In this section, we introduce two algorithms: shifted matrix completion for non-deterministic PU matrix completion, and biased matrix completion for deterministic PU matrix completion. All proofs are deferred to Appendix.

3.1. Shifted Matrix Completion for Non-deterministic Setting (ShiftMC)

We want to find a matrix X such that the loss $\|M-X\|_F^2$ is bounded, using the noisy observation matrix A generated from M by (1). Observe that conditioned on Y, the noise in A_{ij} is *asymmetric*, i.e. $P(A_{ij}=0|Y_{ij}=1)=\rho$ and $P(A_{ij}=1|Y_{ij}=0)=0$. Asymmetric label noise has been studied in the context of binary classification, and recently (Natarajan et al., 2013) proposed a method of unbiased estimator to bound the true loss using only noisy observations. In our case, we aim to find a matrix minimizing the unbiased estimator defined on each element, which leads to the following optimization problem:

$$\min_X \sum_{i,j}\tilde{\ell}(X_{ij},A_{ij}) \quad (4)$$
$$\text{such that } \|X\|_*\le t,\ \ 0\le X_{ij}\le 1\ \forall(i,j).$$

$$\text{where } \tilde{\ell}(X_{ij},A_{ij}) = \begin{cases}\frac{(X_{ij}-1)^2-\rho X_{ij}^2}{1-\rho} & \text{if } A_{ij}=1,\\ X_{ij}^2 & \text{if } A_{ij}=0.\end{cases} \quad (5)$$

The bound constraint on X in the above estimator ensures the loss has bounded Lipschitz constant. This optimization problem is equivalent to the traditional trace-norm regularization problem

$$\min_X \sum_{i,j} \tilde{\ell}(X_{ij}, A_{ij}) + \lambda \|X\|_*, \tag{6}$$
$$\text{such that } 0 \le X_{ij} \le 1 \; \forall (i,j),$$

where λ has a one-to-one mapping to t. We use $\tilde{\ell}$ instead of the original loss ℓ because it is the unbiased estimator of the underlying squared loss $\ell(X_{ij}, M_{ij}) = (X_{ij} - M_{ij})^2$, as formalized below. Thus, we use $\tilde{\ell}$ on the observed A_{ij}, we minimize the loss w.r.t. Y_{ij} in expectation.

Lemma 1. *For any $X \in \mathbb{R}^{m \times n}$, $\frac{1}{mn} E\big[\sum_{i,j}(X_{ij} - Y_{ij})^2\big] = \frac{1}{mn} E\big[\sum_{i,j} \tilde{\ell}(X_{ij}, A_{ij})\big]$.*

Interestingly, we can rewrite $\tilde{\ell}$ as $\tilde{\ell}(X_{ij}, 1) = \big(X_{ij} - \frac{1}{1-\rho}\big)^2 - \frac{\rho}{(1-\rho)^2}$. Therefore, (6) can be rewritten as the following "shifted matrix completion" problem:

$$\hat{X} = \operatorname*{argmin}_X \sum_{i,j:A_{ij}=1} \left(X_{ij} - \frac{1}{1-\rho}\right)^2 + \sum_{i,j:A_{ij}=0} X_{ij}^2 + \lambda \|X\|_*$$
$$\text{s.t. } 0 \le X_{ij} \le 1 \; \forall (i,j). \tag{7}$$

We want to show that the average error of the ShiftMC estimator $\hat{X}$ decays as $O(1/n)$. In order to do so, we first need to bound the difference between the expected error and the empirical error. We define the hypothesis space to be $\mathcal{X} := \{X \mid X \in \mathbb{R}^{m \times n} \text{ and } \|W\|_* \le t\}$. The expected error can be written as $E_A[R_{\tilde{\ell}}(W)] = E_A[\frac{1}{mn}\sum_{i,j} \tilde{\ell}(W_{ij}, A_{ij})]$, and the empirical error is $\hat{R}_{\tilde{\ell}}(W) = \frac{1}{mn}\sum_{i,j} \tilde{\ell}(W_{ij}, A_{ij})$. We first show that the difference between expected error and empirical error can be upper bounded:

Theorem 1. *Let $\mathcal{X} := \{X \in \mathbb{R}^{m \times n} \mid \|X\|_* \le t, 0 \le X \le 1\}$, then*

$$\max_{X \in \mathcal{X}} \left| E_A[R_{\tilde{\ell}}(X)] - \hat{R}_{\tilde{\ell}}(X) \right|$$
$$\le tC \frac{\sqrt{n} + \sqrt{m} + \sqrt[4]{s}}{(1-\rho)mn} + 3\frac{\sqrt{\log(2/\delta)}}{\sqrt{mn}(1-\rho)}$$

with probability at least $1 - \delta$, where C is a constant, $E[R_{\tilde{\ell}}(X)] := E[\frac{1}{mn}\sum_{i,j} \tilde{\ell}(X_{ij}, A_{ij})]$ is the expected error, and $\hat{R}_{\tilde{\ell}}(X) = \frac{1}{mn}\sum_{i,j} \tilde{\ell}(X_{ij}, A_{ij})$ is the empirical error.

Combining Lemma 1 and Theorem 1, we have our first main result:

Theorem 2 (Main Result 1)**.** *With probability at least $1-\delta$,*

$$\frac{1}{mn}\sum_{i,j}(M_{ij} - \hat{X}_{ij})^2 \le 6\frac{\sqrt{\log(2/\delta)}}{\sqrt{mn}(1-\rho)} + 2Ct\frac{\sqrt{n} + \sqrt{m} + \sqrt[4]{s}}{(1-\rho)mn}.$$

The average error is of the order of $O(\frac{1}{n(1-\rho)})$ when $M \in \mathbb{R}^{n \times n}$, where $1 - \rho$ denotes the ratio of observed 1's. This shows that even when we only observe a very small ratio of 1's in the matrix, we can still estimate M accurately when n is large enough.

3.2. Biased Matrix Completion for Deterministic Setting (BiasMC)

In the deterministic setting, we propose to solve the matrix completion problem with label-dependent loss (Scott, 2012). Let $\ell(x, a) = (x - a)^2$ denote the squared loss, for $a \in \{0, 1\}$. The α-weighted loss is defined by

$$\ell_\alpha(x, a) = \alpha 1_{a=1}\ell(x, 1) + (1-\alpha)1_{a=0}\ell(x, 0), \tag{8}$$

where $1_{a=1}, 1_{a=0}$ are indicator functions. We then recover the groundtruth by solving the following biased matrix completion (biasMC) problem:

$$\hat{X} = \operatorname*{argmin}_{X:\|X\|_* \le t} \sum_{i,j} \ell_\alpha(X_{ij}, A_{ij}) \tag{9}$$
$$= \operatorname*{argmin}_{X:\|X\|_* \le t} \alpha \sum_{i,j:A_{ij}=1} (X_{ij} - 1)^2 + (1-\alpha) \sum_{i,j:A_{ij}=0} X_{ij}^2$$

The underlying binary matrix Y is then recovered by the thresholding operator $\bar{X}_{ij} = I(\hat{X}_{ij} > q)$.

A similar formulation has been used in (Sindhwani et al., 2010) to recommend items to users in the "who-bought-what" network. Here, we show that this biased matrix factorization technique can be used to provably recover Y. For convenience, we define the thresholding operator $\text{thr}(x) = 1$ if $x > q$, and $\text{thr}(x) = 0$ if $x \le q$. We first define the recovery error as $R(X) = \frac{1}{mn}\sum_{i,j} 1_{\text{thr}(X_{ij}) \ne Y_{ij}}$, where Y is the underlying 0-1 matrix. Define the label-dependent error:

$$U_\alpha(x, a) = (1-\alpha)1_{\text{thr}(x)=1}1_{a=0} + \alpha 1_{\text{thr}(x)=-1}1_{a=1}. \tag{10}$$

and α-weighted expected error:

$$R_{\alpha,\rho}(X) = E\big[\sum_{i,j} U_\alpha(X_{ij}, A_{ij})\big],$$

The following lemma is a special case of Theorem 9 in (Natarajan et al., 2013), showing that $R(X)$ and $R_{\alpha,\rho}(X)$ can be related by a linear transformation:

Lemma 2. *For the choice $\alpha^* = \frac{1+\rho}{2}$ and $\beta = \frac{1+\rho}{2}$, there exists a constant b that is independent of X such that, for any matrix X, $R_{\alpha^*,\rho}(X) = \beta R(X) + b$.*

Therefore, minimizing the α-weighed expected error in the partially observed situation is equivalent to minimizing the true recovery error R. By further relating $R_{\alpha^*,\rho}(X)$ and $R_{\ell_{\alpha^*},\rho}(X) := E\big[\sum_{i,j} l_{\alpha^*}(X_{ij}, A_{ij})\big]$, we can show:

Theorem 3 (Main Result 2). *Let $\hat{X}$ be the minimizer of (9), and $\bar{X}$ be the thresholded 0-1 matrix of $\hat{X}$, then with probability at least $1-\delta$, we have*

$$R(\bar{X}) \leq \frac{2\eta}{1+\rho}\left(Ct\frac{\sqrt{n}+\sqrt{m}+\sqrt[4]{s}}{mn}+3\frac{\sqrt{\log(2/\delta)}}{\sqrt{mn}(1-\rho)}\right),$$

where $\eta = \max(1/q^2, 1/(1-q)^2)$ and C is a constant.

The average error is of the order of $O(\frac{1}{n(1-\rho)})$ when $M \in \mathbb{R}^{n\times n}$, where $1-\rho$ denotes the ratio of observed 1's, similar to the ShiftMC estimator.

4. PU Inductive Matrix Completion

In this section, we extend our approaches to inductive matrix completion problem, where in addition to the samples, row and column features $F_u \in \mathbb{R}^{m\times d}, F_v \in \mathbb{R}^{n\times d}$ are also given. In the standard inductive matrix completion problem (Jain & Dhillon, 2013), the observations A_Ω are sampled from the groundtruth $M \in \mathbb{R}^{m\times n}$, and we want to recover M by solving the following optimization problem:

$$\min_{D\in\mathbb{R}^{d\times d}} \sum_{i,j\in\Omega}(A_{ij}-(F_uDF_v^T)_{ij})^2+\lambda\|D\|_*. \tag{11}$$

Matrix completion is a special case of inductive matrix completion when $F_u = I, F_v = I$. In the multi-label learning problem, M represents the label matrix and F_u corresponds to examples (typically $F_v = I$) (Yu et al., 2014; Xu et al., 2013). This technique has also been applied to gene-disease prediction (Natarajan & Dhillon, 2014), semi-supervised clustering (Yi et al., 2013), and theoretically studied in (Jain & Dhillon, 2013).

The problem is fairly recent and we wish to extend PU learning analysis to this problem, which is also well motivated in many real world applications. For example, in multi-label learning with partially observed labels, negative labels are usually not available. In the experiments, we will consider another interesting application — semi-supervised clustering problem with only positive and unlabeled relationships.

4.1. Shifted Inductive Matrix Completion for Non-deterministic Setting (ShiftIMC)

In the non-deterministic setting, we consider the inductive version of ShiftMC:

$$\min_{D\in\mathbb{R}^{d\times d}} \sum_{i,j}\tilde{\ell}((F_uDF_v^T)_{ij}, A_{ij}) \tag{12}$$
$$\text{s. t. } \|D\|_* \leq t,\ 1 \geq F_uDF_v^T \geq 0,$$

where the unbiased estimator of loss $\tilde{\ell}(\cdot)$ is defined in (5). Note that we can assume that F_u, F_v are orthogonal (otherwise we can conduct a preprocessing step to normalize it). Let $\boldsymbol{u}_i$ be the i-th row of F_u (the feature for row i) and $\boldsymbol{v}_j$ be the j-th row of F_v. We define constants $\mathcal{X}_u = \max_i \|\boldsymbol{u}_i\|, \mathcal{X}_v = \max_j \|\boldsymbol{v}_j\|$. Since the output of inductive matrix completion is F_uDF_v, it can only recover the original matrix when the underlying matrix M can be written in such form. Following (Xu et al., 2013; Yi et al., 2013), we assume the features are good enough such that $M = F_u(F_u)^TMF_v(F_v)^T$. Recall $\|M\|_* \leq t$. We now extend Theorem 2 to PU inductive matrix completion.

Theorem 4. *Assume $\hat{D}$ is the optimal solution of (12) and the groundtruth M is in the subspace formed by F_u and F_v: $M = F_u(F_u)^TMF_v(F_v)^T$, and let $\hat{M} = F_u^T\hat{D}F_v$, then, with probability at least $1-\delta$:*

$$\frac{1}{mn}\|M-\hat{M}\|_F^2 \leq 6\frac{\sqrt{\log(2/\delta)}}{\sqrt{mn}(1-\rho)}+\frac{4t\sqrt{\log 2d}}{\sqrt{mn}\sqrt{1-\rho}}\mathcal{X}_u\mathcal{X}_v. \tag{13}$$

Therefore if t and d are bounded, the mean square error of ShiftIMC is $O(1/n)$ in the non-deterministic setting.

4.2. Biased Inductive Matrix Completion for Deterministic Setting (BiasIMC)

In the deterministic setting, we propose to solve the inductive version of BiasMC:

$$\hat{D} = \arg\min_{D:\|D\|_*\leq t} \alpha \sum_{i,j:A_{ij}=1}((F_uDF_v^T)_{ij}-1)^2 + (1-\alpha)\sum_{i,j:A_{ij}=0}(F_uDF_v^T)_{ij}^2. \tag{14}$$

The clean 0-1 matrix Y can then be recovered by $\hat{Y}_{ij} = I((F_u\hat{D}F_v^T)_{ij} > q)$.

$$\hat{Y}_{ij} = \begin{cases} 1 & \text{if } (F_u\hat{D}F_v^T)_{ij} \geq q \\ 0 & \text{if } (F_u\hat{D}F_v^T)_{ij} < q. \end{cases} \tag{15}$$

Similar to the case of matrix completion, Lemma 2 shows that the expected 0-1 error $R(X)$ and the α-weighted expected error in noisy observation $R_{\alpha,\rho}(X)$ can be related by a linear transformation when $\alpha^* = \frac{1+\rho}{2}$. With this choice of α^*, Lemma 2 continues to hold in this case, which allows us to extend Theorem 3 to PU inductive matrix completion and bound $R(\hat{Y}) = \frac{1}{mn}\|Y-\hat{Y}\|_F^2$:

Theorem 5. *Let $\hat{D}$ be the minimizer of (14) with $\alpha^* = (1+\rho)/2$, and let $\hat{Y}$ be generated from $\hat{D}$ by (15), then with probability at least $1-\delta$, we have*

$$R(\hat{Y}) \leq \frac{2\eta}{1+\rho}\left(\frac{4t\sqrt{\log 2d}}{\sqrt{mn}\sqrt{1-\rho}}\mathcal{X}_u\mathcal{X}_v+6\frac{\sqrt{\log(2/\delta)}}{\sqrt{mn}(1-\rho)}\right),$$

where $\eta = \max(1/q^2, 1/(1-q)^2)$.

Again, we have that if t and d are bounded, the mean square error of BiasIMC is $O(1/n)$.

5. Optimization Techniques for PU Matrix Completion

In this section, we show that BiasMC can be solved very efficiently for large-scale (millions of rows and columns) datasets, and that ShiftMC can be solved efficiently after a relaxation.

First, consider the optimization problem for BiasMC:

$$\begin{aligned}\operatorname*{argmin}_X \ & \alpha \sum_{i,j:A_{ij}=1} (X_{ij}-1)^2 + (1-\alpha) \sum_{i,j:A_{ij}=0} X_{ij}^2 + \lambda \|X\|_* \\ & := f_b(X) + \lambda \|X\|_*, \end{aligned} \tag{16}$$

which is equivalent to the constrained problem (9) with suitable λ. The typical proximal gradient descent update is $X \leftarrow \mathcal{S}(X - \eta \nabla f_b(X), \lambda)$, where η is the learning rate and $\mathcal{S}$ is the soft thresholding operator on singular values (Ji & Ye, 2009). The (approximate) SVD of $G := (X - \eta \nabla f_b(X))$ can be computed efficiently using power method or Lanczos algorithm if we have a fast procedure to compute GP for a tall-and-thin matrix $P \in \mathbb{R}^{n \times k}$. In order to do so, we first rewrite $f_b(X)$ as

$$f_b(X) = (1-\alpha)\|X-A\|_F^2 + (2\alpha-1) \sum_{i,j:A_{ij}=1} (X_{ij}-A_{ij})^2. \tag{17}$$

Assume the current solution is stored in a low-rank form $X = WH^T$ and $R = (X - A)_{\Omega_1}$ is the residual on Ω_1, then

$$\begin{aligned} GP &= XP - 2\eta \left[(1-\alpha)(X-A) + (2\alpha-1)R\right] P \\ &= (1-2\eta(1-\alpha))WH^TP + 2\eta[(1-\alpha)A - (2\alpha-1)R]P, \end{aligned}$$

where the first term can be computed in $O(mk^2 + nk^2)$ flops, and the remaining terms can be computed in $O(|\Omega_1|k)$ flops. With this approach, we can efficiently compute the proximal operator. This can also be applied to other faster nuclear norm solvers (for example, (Hsieh & Olsen, 2014)).

Next we show that the non-convex form of BiasMC can also be efficiently solved, and thus can scale to millions of nodes and billions of observations. It is well known that the nuclear norm regularized problem $\min_X f_b(X) + \lambda \|X\|_*$ is equivalent to

$$\min_{W \in \mathbb{R}^{m\times k}, H \in \mathbb{R}^{n \times k}} f_b(WH^T) + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2) \tag{18}$$

when k is sufficiently large. We can use a trick similar to (17) to compute the gradient and Hessian efficiently:

$$\begin{aligned} \frac{1}{2}\nabla_W f_b(WH^T) &= [(1-\alpha)(WH^T - A) + (2\alpha-1)R_\Omega]H, \\ \frac{1}{2}\nabla^2_{W_{i,:}} f_b(WH^T) &= (1-\alpha)H^TH + (2\alpha-1)H_{\Omega_i}^T H_{\Omega_i}, \end{aligned}$$

where H_{Ω_i} is the sub-matrix with columns $\{\boldsymbol{h}_j : j \in \Omega_i\}$, and Ω_i is the column indices of observations in the i-th row. Thus, we can efficiently apply Alternating Least Squares (ALS) or Coordinate Descent (CD) for solving (18). For example, when applying CCD++ in (Yu et al., 2013), each coordinate descent update only needs $O(|\Omega_i| + k)$ flops. We apply this technique to solve large-scale link prediction problems (see Section 6).

The optimization problem for ShiftMC is harder to solve because of the bounded constraint. We can apply the bounded matrix factorization technique (Kannan et al., 2014) to solve the non-convex form of (6), where the time complexity is $O(mn)$ because of the constraint $0 \le (WH^T)_{ij} \le 1$ for all (i,j). To scale it to large datasets, we relax the bounded constraint and solve:

$$\begin{aligned} \min_{W \in \mathbb{R}^{m\times k}, H \in \mathbb{R}^{n\times k}} & \|A - WH^T\|_F^2 + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2) \\ & \text{s. t. } 0 \le W, H \le \sqrt{1/k} \end{aligned} \tag{19}$$

This approach (ShiftMC-relax) is easy to solve by ALS or CD with $O(|\Omega|k)$ complexity per sweep (similar to the BiasMC). In our experiments, we show ShiftMC-relax performs even better than shiftMC in practice.

6. Experiments

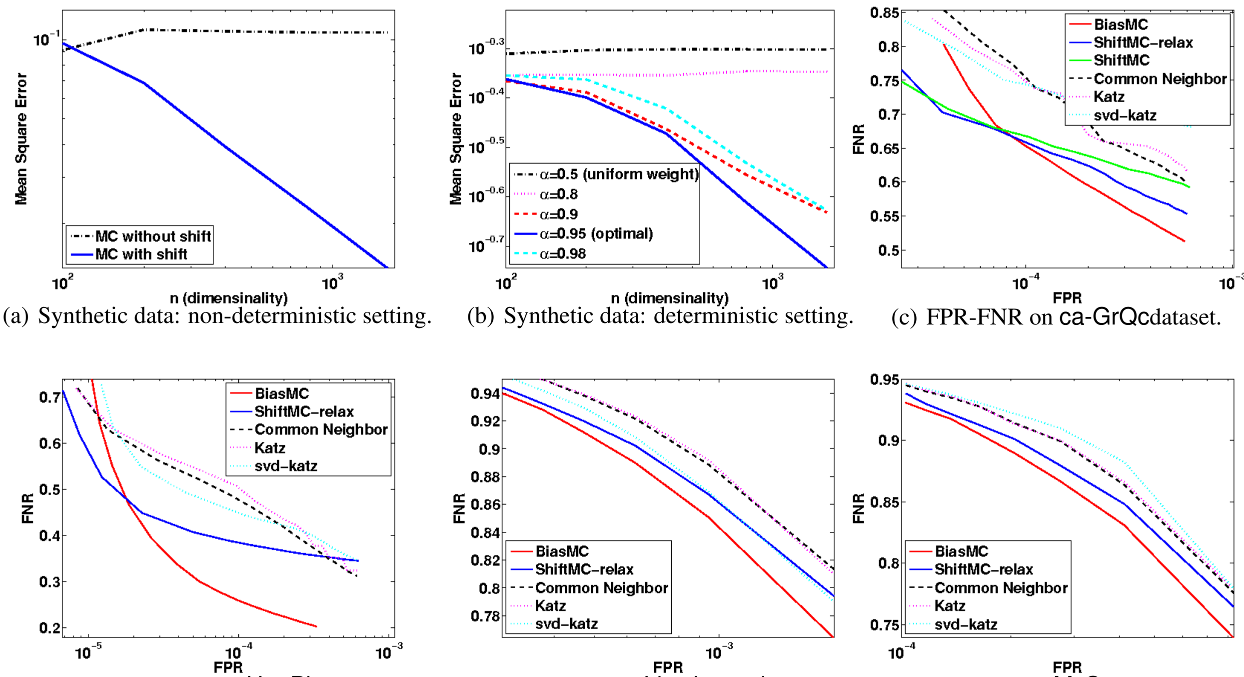
We first use synthetic data to show that our bounds are meaningful and then demonstrate the effectiveness of our algorithms in real world applications.

6.1. Synthetic Data

We assume the underlying matrix $M \in \mathbb{R}^{n\times n}$ is generated by UU^T, where $U \in \mathbb{R}^{n \times k}$ is the orthogonal basis of a random Gaussian n by k matrix with mean 0 and variance 1. For the non-deterministic setting, we linearly scale M to have values in $[0, 1]$, and then generate training samples as described Section 2. For deterministic setting, we choose q so that Y has equal number of zeros and ones. We fix $\rho = 0.9$ (so that only 10% 1's are observed). From Lemma 2, $\alpha = 0.95$ is optimal. We fix $k = 10$, and test our algorithms with different sizes n. The results are shown in Figure 1(a)-(b). Interestingly, the results reflect our theory: error of our estimators decreases with n; in particular, error linearly decays with n in log-log scaled plots, which suggests a rate of $O(1/n)$, as shown in Theorems 2 and 3. Directly minimizing $\|A - X\|_F^2$ gives very poor results. For BiasMF, we also plot the performance of estimators with various α values in Figure 1(b). As our theory suggests, $\alpha = \frac{1+\rho}{2}$ performs the best. We also observe that the error is well-behaved in a certain range of α. A principled way of selecting α is an interesting problem for further research.

6.2. Parameter Selection

Before showing the experimental results on real-world problems, we discuss the selection of the parameter ρ in our PU matrix completion model (see eq (1)). Note that



(a) Synthetic data: non-deterministic setting. (b) Synthetic data: deterministic setting. (c) FPR-FNR on ca-GrQcdataset.

(d) FPR-FNR on ca-HepPhdataset. (e) FPR-FNR on LiveJournaldataset. (f) FPR-FNR on MySpacedataset.

Figure 1. (a)-(b): Recovery error of ShiftMC and BiasMC on synthetic data. We observe that without shifting or biasing, error does not decrease with n (the black lines). The error of our estimators decreases approximately as $\frac{1}{n}$, as proved in Theorems 2 and 3. (c)-(f): Comparison of link prediction methods. ShiftMC and BiasMC consistently perform better than the rest.

ρ indicates the noise rate of flipping a 1 to 0. If there are equal number of positive and negative elements in the underlying matrix Y, we will have $\rho = 1 - 2s$ where $s =$ (# positive entries)/(# total entries). In practice (e.g., link prediction problems) number of 1's are usually less than number of 0 in the underlying matrix, but we do not know the ratio. Therefore, in all the experiments we chose ρ from the set $\{1-2s, 10(1-2s), 100(1-2s), 1000(1-2s)\}$ based on a random validation set, and use the corresponding α in the optimization problems.

6.3. Matrix completion for link prediction

One of the important applications that motivated our analysis in this paper is the link prediction problem. Note that matrix completion has been used for link prediction on signed network (Chiang et al., 2014), but the application to unsigned network has not been discussed before. Here, we are given n nodes (users) and a set of edges Ω_{train} (relationships) and the goal is to predict *missing* edges, i.e. Ω_{test}. We use 4 real-world datasets: 2 co-author networks ca-GrQc(4,158 nodes and 26,850 edges) and ca-HepPh(11,204 nodes and 235,368 edges), where we randomly split edges into training and test such that $|\Omega_{train}| = |\Omega_{test}|$; 2 social networks LiveJournal(1,770,961 nodes, $|\Omega_{train}|$ = 83,663,478 and $|\Omega_{test}|$ = 2,055,288) and MySpace(2,137,264 nodes, $|\Omega_{train}|$ = 90,333,122 and $|\Omega_{test}|$ = 1,315,594), where train/test split is done using timestamps. For our proposed methods BiasMC, ShiftMC and ShiftMC-relax, we solve the non-convex form with $k = 50$ for ca-GrQc, ca-HepPhand $k = 100$ for LiveJournaland MySpace. The α and λ values are chosen by a validation set.

We compare with competing link prediction methods (Kiben-Nowell & Kleinberg, 2003) Common Neighbors, Katz, and SVD-Katz (compute Katz using the rank-k approximation, $A \approx U_k \Sigma_k V_k$). Note that the classical matrix factorization approach in this case is equivalent to SVD on the given 0-1 training matrix, and SVD-Katz slightly improves over SVD by further computing the Katz values based on the low rank approximation (see (Shin et al., 2012)), so we omit the SVD results in the figures. Note that the algorithm ldNMF proposed in (Sindhwani et al., 2010) is time consuming because there are $O(n^2)$ hidden variables to estimate. Therefore, we compare with it only on a $n = 500$ subset of ca-GrQc. On this subset, BiasMC got 1.14% and ldNMF got 1.08% top 10 prediction accuracy.

Based on the training matrix, each link prediction method will output a list of k candidate entries. We evaluate the quality of the top-k entries by computing the False Positive Rate (FPR) and False Negative Rate (FNR) on the

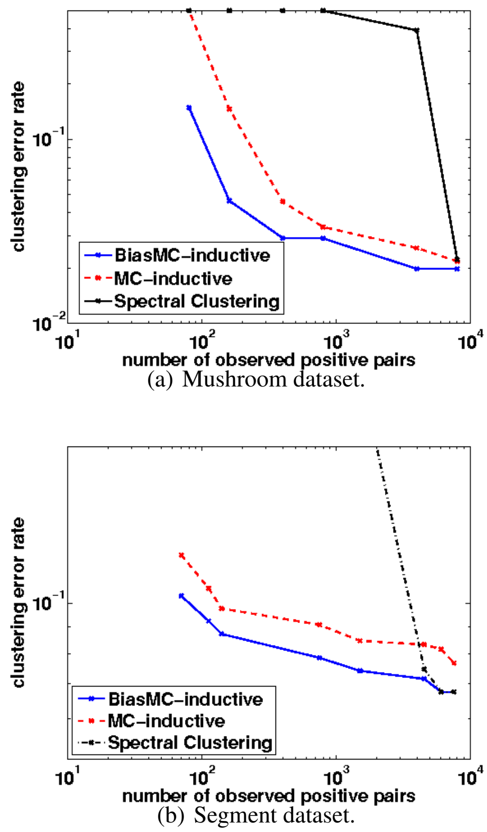


Figure 2. Semi-supervised clustering performance of BiasMC-inductive on two real datasets. BiasMC-inductive performs better than MC-inductive (treats unlabeled relationships as zeros) and spectral clustering (does not use features). BiasMC-inductive achieves under 10% error using just 100 samples.

test snapshot. The results are shown in Figure 1 (c)-(f). ca-GrQcis a small dataset, so we can solve the original ShiftMC problem accurately, although ShiftMC-relax achieves a similar performance here. For larger datasets, we show only the performance of ShiftMC-relax. In general BiasMC performs the best, and ShiftMC tends to perform better in the beginning. Overall, our methods achieve lower FPR and FNR comparing to other methods, which indicate that we obtain a better link prediction model by solving the PU matrix completion problem. Also, BiasMC is highly efficient — it takes 516 seconds for 10 coordinate descent sweeps on the largest dataset (MySpace), whereas computing top 100 eigenvectors using `eigs` in MATLAB requires 2408 seconds.

6.4. Inductive matrix completion

We use the semi-supervised clustering problem to evaluate our PU inductive matrix completion methods. PU inductive matrix completion can be applied to many real-world problems, including recommender systems with features and 0-1 observations, and the semi-supervised clustering problem when we can only observed positive relationships. Here we use the latter as an example to demonstrate the usefulness of our algorithm.

In semi-supervised clustering problems, we are given n samples with features $\{\boldsymbol{x}_i\}_{i=1}^n$ and pairwise relationships $A \in \mathbb{R}^{n\times n}$, where $A_{ij} = 1$ if two samples are in the same cluster, $A_{ij} = -1$ if they are in different clusters, and $A_{ij} = 0$ if the relationship is unobserved. Note that the groundtruth matrix $M \in \{+1, -1\}^{n\times n}$ exhibits a simple structure and is a low rank as well as low trace norm matrix; it is shown in (Yi et al., 2013) that we can recover M using IMC when there are both positive and negative observations. We consider the setting where only positive relationships are observed, so A is a 0-1 matrix. We show that biased IMC can recover M using very few positive relationships. We test the algorithms on two datasets: the Mushroom dataset with 8142 samples, 112 features, and 2 classes; the Segment dataset with 2310 samples, 19 features, and 7 classes. The results are presented in Figure 2. We compare BiasMC-inductive with (a) MC-inductive, which considers all the unlabeled pairs as zeros and minimizes $\|F_u D F_v^T - A\|_F^2$, and (b) spectral clustering, which does not use feature information. Since the data is from classification datasets, the ground truth M is known and can be used to evaluate the results. In Figure 2, the vertical axis is the clustering error rate defined as the fraction of entries in M predicted with correct sign. Figure 2 shows that BiasMC-inductive is much better than other approaches for this task.

7. Conclusions

Motivated by modern applications of matrix completion, our work attempts to bridge a gap between the theory of matrix completion and practice. We have shown that even when there is noise in the form of one-bit quantization as well as a one-sided sampling process that reveals the measurements, the underlying matrix can be accurately recovered. We consider two recovery settings, both of which are natural for PU learning, and provide similar recovery guarantees for both. Our error bounds are strong and useful in practice. Our work serves to provide the first theoretical insight into the biased matrix completion approach that has been employed as a heuristic for similar problems in the past. Experimental results on synthetic data conform to our theory and the effectiveness of the methods are evident for the link prediction task in real-world networks. A principled way of selecting or estimating the bias α in BiasMC seems worthy of exploration given our encouraging results.

Acknowledgements

This research was supported by NSF grants CCF-1320746 and CCF-1117055. C.-J.H also acknowledges support from an IBM PhD fellowship.

References

Boufounos, P. T and Baraniuk, R. G. 1-bit compressive sensing. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pp. 16–21. IEEE, 2008.

Candès, E. J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2009.

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

Chiang, K. Y., Hsieh, C. J., Natarajan, N., Tewari, A., and Dhillon, I. S. Prediction and clustering in signed networks: A local to global perspective. *Journal of Machine Learning Research (JMLR)*, 15:1177 – 1213, March 2014.

Davenport, M. A., Plan, Y., Berg, E., and Wootters, M. 1-bit matrix completion. *arXiv preprint arXiv:1209.3672*, 2012.

Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–220. ACM, 2008.

Hsieh, C.-J. and Olsen, P. A. Nuclear norm minimization via active subspace selection. In *ICML*, 2014.

Hu, Y., Koren, Y., and Volinsky, C. Collaborative filtering for implicit feedback datasets. In *ICDM*, 2008.

Jain, P. and Dhillon, I. S. Provable inductive matrix completion. *CoRR*, abs/1306.0626, 2013.

Ji, S. and Ye, J. An accelerated gradient method for trace norm minimization. In *ICML*, 2009.

Kakade, Sham M, Sridharan, Karthik, and Tewari, Ambuj. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, volume 21, pp. 793–800, 2008.

Kannan, R., Ishteva, M., and Park, H. Bounded matrix factorization for recommender system. *Knowledge and Information Systems*, 2014.

Keshavan, Raghunandan H, Montanari, Andrea, and Oh, Sewoong. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(2057-2078):1, 2010.

Kiben-Nowell, D. and Kleinberg, J. The link prediction problem for social networks. In *CIKM*, 2003.

Latala, R. Some estimates of norms of random matrices. *Proceedings of the AMS*, 2005.

Liu, Bing, Dai, Yang, Li, Xiaoli, Lee, Wee Sun, and Yu, Philip S. Building text classifiers using positive and unlabeled examples. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pp. 179–186. IEEE, 2003.

Natarajan, N. and Dhillon, I. S. Inductive matrix completion for predicting gene-disease associations. In *Bioinformatics*, 2014.

Natarajan, N., Tewari, A., Dhillon, I. S., and Ravikumar, P. Learning with noisy labels. In *NIPS*, 2013.

Rendle, S., C. Freudenthaler, Z. Gantner, and Schmidt-Thieme, L. BPR: Bayesian personalized ranking from implicit feedback. In *UAI*, 2009.

Scott, C. Calibrated asymmetric surrogate losses. *Electronic J. of Stat.*, 6(958–992), 2012.

Shamir, O. and Shalev-Shwartz, S. Collaborative filtering with the trace norm: Learning, bounding, and transducing. In *COLT*, 2011.

Shawe-Taylor, J. and Cristianini, N. Kernel methods for pattern analysis. *Cambridge University Press*, 2004.

Shin, D., Si, S., and Dhillon, I. S. Multi-scale link prediction. In *Proceedings of the 21st ACM Conference on Information and Knowledge Management(CIKM)*, October 2012.

Sindhwani, V., Bucak, S. S., Hu, J., and Mojsilovic, A. One-class matrix completion with low-density factorization. In *ICDM*, 2010.

Xu, M., Jin, R., and Zhou, Z.-H. Speedup matrix completion with side information: Application to multi-label learning. In *NIPS*, 2013.

Yi, J., Zhang, L., Jin, R., Qian, Q., and Jain, A. K. Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion. In *ICML*, 2013.

Yu, H.-F., Hsieh, C.-J., Si, S., and Dhillon, I. S. Parallel matrix factorization for recommender systems. *Knowledge and Information Systems*, 2013.

Yu, H.-F., Jain, P., Kar, P., and Dhillon, I. S. Large-scale multi-label learning with missing labels. In *ICML*, pp. 593–601, 2014.

A. Proofs

A.1. Proof of Lemma 1

Proof.

$$\frac{1}{mn}E\Big[\sum_{i,j}\tilde{l}(X_{ij},A_{ij})\Big]$$
$$=\frac{1}{mn}\sum_{i,j}E\Big[\tilde{l}(X_{ij},A_{ij})\Big]$$
$$=\frac{1}{mn}\sum_{i,j}\Big(P(Y_{ij}=0)X_{ij}^2+P(Y_{ij}=1)\Big(\rho X_{ij}^2$$
$$+(1-\rho)(\frac{(X_{ij}-1)^2-\rho X_{ij}^2}{1-\rho})\Big)\Big)$$
$$=\frac{1}{mn}\sum_{i,j}\big(P(Y_{ij}=0)X_{ij}^2+P(Y_{ij}=1)(X_{ij}-1)^2\big)$$
$$=\frac{1}{mn}E\Big[\sum_{i,j}(X_{ij}-Y_{ij})^2\Big]=\frac{1}{mn}E\Big[l(X_{ij},Y_{ij})\Big].$$

□

A.2. Proof of Theorem 1

Proof. We want to bound $\sup_{X\in\mathcal{X}}\Big|\hat{R}_{\tilde{l}}(X)-E_A[R_{\tilde{l}}(X)]\Big|$. First,

$$E_A[R_{\tilde{l}}(X)]\le\hat{R}_{\tilde{l}}(X)+\sup_{X\in\mathcal{X}}\Big(E_A[\frac{1}{mn}\sum_{i,j}\tilde{l}(X_{ij},A_{ij})]-\frac{1}{mn}\sum_{i,j}\tilde{l}(X_{ij},A_{ij})\Big).$$

Apply McDiarmid's Theorem in (Shawe-Taylor & Cristianini, 2004); since each $\tilde{l}(X_{ij},A_{ij})$ can be either X_{ij}^2 or $\frac{(X_{ij}-1)^2-\rho X_{ij}^2}{1-\rho}$, when changing one random variable A_{ij}, in the worst case the quantity $\sup_{X\in\mathcal{X}}\Big(E_A[R_{\tilde{l}}(X)]-\frac{1}{mn}\sum_{i,j}\tilde{l}(X_{ij},A_{ij})\Big)$ can be changed by

$$\Big|X_{ij}^2-\frac{(X_{ij}-1)^2-\rho X_{ij}^2}{1-\rho}\Big|\le\Big|\frac{2X_{ij}+1}{1-\rho}\Big|\le\frac{3}{1-\rho}.$$

So by McDiarmid's Theorem, with probability $1-\delta/2$,

$$\sup_{X\in\mathcal{X}}\Big(E_A(R_{\tilde{l}}(X))-\frac{1}{mn}\sum_{i,j}\tilde{l}(X_{ij},A_{ij})\Big)$$
$$\le E_A\left[\sup_{X\in\mathcal{X}}\big(E_A(R_{\tilde{l}}(X))-\frac{1}{mn}\sum_{i,j}\tilde{l}(X_{ij},A_{ij})\big)\right]$$
$$+3\frac{\sqrt{\log(2/\delta)}}{\sqrt{mn}(1-\rho)}.$$

Also,

$$E_A\left[\sup_{X\in\mathcal{X}}\Big(E_A[R_{\tilde{l}}(X)]-\frac{1}{mn}\sum_{i,j}\tilde{l}(X_{ij},A_{ij})\Big)\right] \tag{20}$$
$$\le E_{A,\tilde{A}}\left[\sup_{X\in\mathcal{X}}\Big(\frac{1}{mn}\sum_{i,j}\tilde{l}(X_{ij},\tilde{A}_{ij})-\frac{1}{mn}\sum_{i,j}\tilde{l}(X_{ij},A_{ij})\Big)\right] \tag{21}$$
$$=E_{A,\tilde{A}}\left[\sup_{X\in\mathcal{X}}\frac{1}{mn}\sum_{i,j}(\tilde{l}(X_{ij},\tilde{A}_{ij})-\tilde{l}(X_{ij},A_{ij}))\right] \tag{22}$$
$$=\frac{1}{mn}E_{A,A',\sigma}\left[\sup_{X\in\mathcal{X}}\sum_{i,j:M_{ij}=1}\sigma_{ij}(\tilde{l}(X_{ij},\tilde{A}_{ij})-\tilde{l}(X_{ij},A_{ij}))\right] \tag{23}$$
$$\le\frac{1}{mn}E_{A,\sigma}\left[\sup_{X\in\mathcal{X}}\sum_{i,j:M_{ij}=1}\sigma_{ij}\tilde{l}(X_{ij},A_{ij})\right] \tag{24}$$

where σ_{ij} are random variables with half chance to be +1 and half chance to be -1. Where from (22) to (23) we use the fact that $A_{ij}=0$ with probability 1 if $M_{ij}=0$. Next we want to bound the Rademacher complexity $E_{A,\sigma}\Big[\sup_{X\in\mathcal{X}}\sum_{i,j:M_{ij}=1}\sigma_{ij}\tilde{l}(X_{ij},A_{ij})\Big]$. When $M_{ij}=1$,

$$\tilde{l}(X_{ij},A_{ij})=\begin{cases}X_{ij}^2 & \text{with probability } \rho\\ \frac{(X_{ij}-1)^2-\rho X_{ij}^2}{1-\rho} & \text{with probability } 1-\rho\end{cases}$$

Since $0\le X_{ij}\le 1$, the Lipschitz constant for $\tilde{l}(X_{ij},A_{ij})$ is at most $1/(1-\rho)$, so

$$(24)\le\frac{1}{mn}E_\sigma(\sup_{X\in\mathcal{X}}\sum_{i,j:M_{ij}=1}\sigma_{ij}X_{ij})$$
$$\le\frac{1}{(1-\rho)mn}E_\sigma\left[\sup_{X\in\mathcal{X}}\|\mathcal{P}_{M_{ij}=1}(\sigma)\|_2\|X\|_*\right].$$

As pointed out in (Shamir & Shalev-Shwartz, 2011), we can then apply the main Theorem in (Latala, 2005), when Z is an independent zero mean random matrix,

$$E[\|Z\|_2]\le C\Bigg(\max_i\sqrt{\sum_j E[Z_{ij}^2]}+\max_j\sqrt{\sum_i E[Z_{ij}^2]}+\sqrt[4]{\sum_{ij}E[Z_{ij}^4]}\Bigg)$$

with a universal constant C.

So in our case $E[\|\sigma\|_2]\le C\,(\sqrt{n}+\sqrt{m}+\sqrt[4]{s})$, so (24) $\le tC\frac{\sqrt{n}+\sqrt{m}+\sqrt[4]{s}}{(1-\rho)mn}$. □

A.3. Proof of Theorem 2

Let $\hat{X}$ be the minimizer of (4), and

$$P := tC\frac{\sqrt{n}+\sqrt{m}+\sqrt[4]{s}}{(1-\rho)mn} + 3\frac{\sqrt{\log(2/\delta)}}{\sqrt{mn}(1-\rho)},$$

we have

$$\begin{aligned}
&E\big[\frac{1}{mn}\sum_{i,j}(\hat{X}_{ij}-Y_{ij})^2\big] \\
&= E\big[\frac{1}{mn}\sum_{i,j}\tilde{l}(\hat{X}_{ij},A_{ij})\big] \quad \text{(Lemma 1)} \\
&\leq \hat{R}_{\tilde{l}}(\hat{X}) + P \quad \text{(Theorem 1)} \\
&\leq \hat{R}_{\tilde{l}}(M) + P \quad \text{(by the definition of } \hat{X}) \\
&\leq E\big[\frac{1}{mn}\tilde{l}(M_{ij},A_{ij})\big] + 2P \quad \text{(Theorem 1)} \\
&= E\big[\frac{1}{mn}\sum_{i,j}(M_{ij}-Y_{ij})^2\big] \quad \text{(Lemma 1)}
\end{aligned}$$

Therefore

$$\frac{1}{mn}\sum_{i,j}E\big[(X_{ij}-Y_{ij})^2-(M_{ij}-Y_{ij})^2\big] \leq 2P.$$

Since $P(Y_{ij}=1)=M_{ij}$, we have

$$\begin{aligned}
&E\big[(X_{ij}-Y_{ij})^2-(M_{ij}-Y_{ij})^2\big] \\
&= M_{i,j}\big((X_{ij}-1)^2-(M_{ij}-1)^2\big) + (1-M_{ij})(X_{ij}^2-M_{ij}^2) \\
&= (X_{ij}-M_{ij})^2,
\end{aligned}$$

therefore

$$\frac{1}{mn}\sum_{i,j}(X_{ij}-M_{ij})^2 \leq 2P.$$

A.4. Proof of Theorem 3

Proof. We want to show

$$R_{\alpha,\rho}(X)-\min_X R_{\alpha,\rho}(X) \leq \eta(R_{l_\alpha,\rho}(X)-\min_X R_{l_\alpha,\rho}(X)), \tag{25}$$

where $\eta=\max(1/q^2,1/(1-q)^2)$. Consider the following two cases. If $Y_{ij}=0$, then

$$\begin{aligned}
&R_{\alpha,\rho}(X_{ij}) = \alpha 1_{X_{ij}>q}, \ \ \min_{X_{ij}} R_{\alpha,\rho}(X_{ij}) = 0 \\
&R_{l_\alpha,\rho}(X_{ij}) = \alpha X_{ij}^2, \ \ \min_{X_{ij}} R_{l_\alpha,\rho}(X_{ij}) = 0,
\end{aligned}$$

so the left hand side of (25) is $\alpha 1_{X_{ij}>q}$ and the right hand side is αX_{ij}^2. Therefore we can simply verify that (25) holds with $\eta=1/q^2$. For the second case if $Y_{ij}=1$,

$$\begin{aligned}
R_{\alpha,\rho}(X_{ij}) &= \rho(1-\alpha^*)1_{X_{ij}>q} + (1-\rho)\alpha^* 1_{X_{ij}<q} \\
&= \frac{(1-\rho)(1+\rho)}{2}1_{X_{ij}<q} + \frac{\rho(1-\rho)}{2}1_{X_{ij}>q}, \\
R_{l_\alpha,\rho}(X_{ij}) &= \frac{(1-\rho)(1+\rho)}{2}(X_{ij}-1)^2 + \frac{\rho(1-\rho)}{2}X_{ij}^2.
\end{aligned}$$

We can see $(1-\rho)(1+\rho)1_{X_{ij}<q} \leq \frac{1}{(1-q)^2}(1-\rho)(1+\rho)(X_{ij}-1)^2$ and $\rho(1-\rho)1_{X_{ij}<q} < \frac{1}{q^2}\rho(1-\rho)X_{ij}^2$, so both will satisfied by our chosen η.

Next we compute $\min_{X_{ij}} R_{\alpha,\rho}(X_{ij})$ and $\min_{X_{ij}} R_{l_\alpha,\rho}(X_{ij})$. By definition

$$R_{\alpha^*,\rho}(X_{ij}) = \begin{cases} \rho(1-\alpha^*) = \frac{\rho(1-\rho)}{2} & \text{if } X_{ij}>q \\ (1-\rho)\alpha^* = \frac{(1+\rho)(1-\rho)}{2} & \text{if } X_{ij}<q. \end{cases}$$

Therefore

$$\min_{X_{ij}} R_{\alpha,\rho}(X_{ij}) = \frac{\rho(1-\rho)}{2}. \tag{26}$$

On the other hand,

$$\begin{aligned}
R_{l_\alpha,\rho}(x) &= \rho(1-\alpha^*)x^2 + (1-\rho)\alpha^*(x-1)^2 \\
&= \frac{\rho(1-\rho)}{2}x^2 + \frac{(1-\rho)(1+\rho)}{2}(x-1)^2.
\end{aligned}$$

Taking gradient equals to zero we get $x^*=\frac{\rho+1}{2\rho+1}$, and therefore

$$\min_x R_{l_\alpha,\rho}(x) = \frac{(1+\rho)\rho(1-\rho)}{2} \leq \rho(1-\rho). \tag{27}$$

Combining (26) and (27), we have

$$\min_x R_{\alpha,\rho} \geq \min_x R_{l_\alpha,\rho}/2,$$

therefore we need $\eta \geq 2$. But this is satisfied by $\eta=\max(1/q^2,1/(1-q)^2)$. Combining the above arguments, we proved that (25) holds.

Next we show an upper bound of (25). Using the proof similar to Theorem 1 we have

$$\begin{aligned}
&R_{l_\alpha,\rho}(X) - \min_X R_{l_\alpha,\rho}(X) \\
\leq& Ct\frac{\sqrt{n}+\sqrt{m}+\sqrt[4]{s}}{mn} + 3\frac{\sqrt{\log(2/\delta)}}{\sqrt{mn}(1-\rho)}
\end{aligned} \tag{28}$$

Now for the left hand side $R_{\alpha,\rho}(X)-\min_X R_{\alpha,\rho}(X)$. By Theorem 2, we know that

$$R_{\alpha^*,\rho}(X) - \min_X R_{\alpha^*,\rho}(X) = \left(\frac{1+\rho}{2}\right)R(X). \tag{29}$$

Here we use the fact that $\min_X R(X)=0$ because $R(Y)=0$, and the term B vanished because it is a constant for both sides. Combining (29), (28) and (25), we have

$$\left(\frac{1+\rho}{2}\right)R(X) \leq \eta\left(Ct\frac{\sqrt{n}+\sqrt{m}+\sqrt[4]{s}}{mn} + 3\frac{\sqrt{\log(2/\delta)}}{\sqrt{mn}(1-\rho)}\right),$$

therefore

$$R(X) \leq \frac{2\eta}{1+\rho}\left(Ct\frac{\sqrt{n}+\sqrt{m}+\sqrt[4]{3s}}{mn} + 3\frac{\sqrt{\log(2/\delta)}}{\sqrt{mn}(1-\rho)}\right).$$

$\square$

A.5. Proof of Theorem 4

Proof. For convenience, we let $X = F_u D F_v^T$. We first apply the same argument in to the proof in Appendix A.2 to get (24). Now we want to bound the Rademacher compelxity $E_{A,\sigma}[\sup_{w\in\mathcal{W}} \sum_{i,j} \sigma_{ij}\tilde{l}(X_{ij}, A_{ij})]$ (upper bound of (24)). Since $\tilde{l}(X_{ij}, A_{ij})$ is Lipchitz continuous with constant $\frac{1}{1-\rho}$ (use the fact that X_{ij} is bounded between 0 and 1), we have $\tilde{l}(X_{ij}, A_{ij}) \le \frac{1}{1-\rho}(X_{ij} - A_{ij})$. Therefore,

$$
\begin{aligned}
&E_{A,\sigma}[\sup_{w\in\mathcal{W}} \sum_{i,j:A_{ij}=1} \sigma_{ij}\tilde{l}((F_u D F_v^T)_{ij}, A_{ij})] \\
&\le E_{A,\sigma}[\sup_{w\in\mathcal{W}} \sum_{i,j:A_{ij}=1} \frac{\sigma_{ij}}{1-\rho}(F_u D F_v^T)_{ij}] + E_{A,\sigma}[\frac{\sigma_{ij}}{1-\rho}] \\
&= \frac{1}{1-\rho} E_{A,\sigma}[\sup_{w\in\mathcal{W}} \sum_{i,j:A_{ij}=1} \sigma_{ij}\mathrm{tr}(\boldsymbol{u}_i \boldsymbol{v}_j^T D)]
\end{aligned}
$$

We then use the following Lemma, which is a special case of Theorem 1 in (Kakade et al., 2008) when taking $\|\cdot\|$ to be the matrix 2 norm and $\|\cdot\|_*$ (the dual norm) is the trace norm:

Lemma 3. *Let $\mathcal{D} := \{D \mid D \in \mathbb{R}^{d\times d}$ and $\|D\|_* \le \mathcal{D}_1\}$ (where $\|W\|_*$ is the trace norm of W), and $\mathcal{W} = \max_i \|W_i\|_2$, then*

$$
E_\sigma[\sup_{D\in\mathcal{D}} \frac{1}{p}\sum_{i=1}^{m} \sigma_i tr(DW_i)] \le 2\mathcal{W}\mathcal{D}_1\sqrt{\frac{\log 2d}{p}}.
$$

Now the set $\mathcal{D}$ is $\{D : \|D\|_* \le t\}$ and number of terms that $A_{ij} = 1$ is $p = n^2(1-\rho)$, so using the above lemma we have

$$
\begin{aligned}
&E_{A,\sigma}[\sup_{D\in\mathcal{D}} \sum_{i,j:A_{ij}=1} \sigma_{ij}\tilde{l}((F_u D F_v^T)_{ij}, A_{ij})] \\
&\le \frac{2}{1-\rho} t \left(\max_{i,j} \|\boldsymbol{u}_i \boldsymbol{v}_j^T\|_2\right)\sqrt{\log 2d}\sqrt{p} \\
&\le \frac{2}{1-\rho} t \mathcal{X}_u \mathcal{X}_v \sqrt{\log 2d}\sqrt{mn}\sqrt{1-\rho} \\
&= \frac{2\sqrt{mn}t\sqrt{\log 2d}}{\sqrt{1-\rho}} \mathcal{X}_u \mathcal{X}_v.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\frac{1}{mn} E_{A,\sigma}[\sup_{D\in\mathcal{D}} \sum_{i,j:A_{ij}=1} \sigma_{ij}\tilde{l}((F_u D F_v^T)_{ij}, A_{ij})] \\
\le& \frac{2t\sqrt{\log 2d}}{\sqrt{mn}\sqrt{1-\rho}} \mathcal{X}_u \mathcal{X}_v.
\end{aligned}
$$

Combined with other part of the proof of Theorem 1 we have

$$
\begin{aligned}
&\frac{1}{mn}\sum_{i,j}(M_{ij} - F_u \hat{D} F_v^T)^2 \\
&\le 6\frac{\sqrt{\log(2/\delta)}}{\sqrt{mn}(1-\rho)} + \frac{4t\sqrt{\log 2d}}{\sqrt{mn}\sqrt{1-\rho}}\mathcal{X}_u \mathcal{X}_v.
\end{aligned}
$$

□

A.6. Proof of Theorem 5

Proof. We follow the proof for Theorem 3. Again let $X = F_u D F_v^T$. First, we show that (25) is still true for the inductive case. The only difference here is to show that $\min_{X:\|X\|_*\le t} R_{\alpha,\rho} \ge \eta \min_{X:\|X\|_*\le t} R_{l_\alpha,\rho}(X)$ because now all the (i,j) elements are dependent. However, as discussed in the previous proof, if we treat each (i,j) independently, the optimal value for each (i,j) elements will be

$$
Z_{ij} = \begin{cases} > q \text{ if } A_{ij} = 1, \\ \le q \text{ if } A_{ij} = 0. \end{cases}
$$

By assumption we know there exists an D with $\|F_u D F_v^T\|_* = \|D\|_* \le t$ and $X = F_u D F_v^T$ satisfies the above condition. Therefore the value of $\min_{X:\|X\|_*\le t} R_{\alpha,\rho}$ still takes the same value with Theorem 3. On the other hand, since now we enforce a more strict constraint that $X = F_u D F_v^T$, Theorem 3 gives an upper bound of $\min_{D:\|D\|_*\le t} R_{l_\alpha,\rho}(F_u D F_v^T)$. Therefore, equation (25) still holds.

We then also have

$$
\begin{aligned}
&R_{l_\alpha,\rho}(F_u D F_v^T) - \min_D R_{l_\alpha,\rho}(F_u D F_v^T) \\
&\le 6\frac{\sqrt{\log(2/\delta)}}{\sqrt{mn}(1-\rho)} + \frac{4t\sqrt{\log 2d}}{\sqrt{mn}\sqrt{1-\rho}}\mathcal{X}_u \mathcal{X}_v \qquad (30)
\end{aligned}
$$

using the similar proof to Theorem 4.

Combining (30), (25) and Theorem 2, the proof is complete. □

B. Hardness of recovering M in the non-deterministic setting

The 1-bit matrix completion approach of (Davenport et al., 2012) can be applied to this setting — Given a matrix M, a subset Ω is sampled uniformly at random from M, and the observed values are "quantized" by a known probability distribution. We can transform our problem to the 1-bit matrix completion problem by assuming all the unobserved entries are zeros. For convenience we assume $M \in \mathbb{R}^{n\times n}$. We will show that the one-bit matrix completion approach in (Davenport et al., 2012) is not satisfactory for PU matrix completion in the non-deterministic setting. In (Davenport et al., 2012), the underlying matrix M is assumed to satisfy $\|M\|_* \le t$ and $\|M\|_\infty \le \alpha$; we are given a subset (chosen uniformly random) Ω with $|\Omega| = m$ and we observe the following quantity on Ω:

$$Y_{i,j} = \begin{cases} 1 & \text{with probability } f(M_{ij}), \\ -1 & \text{with probability } 1 - f(M_{ij}). \end{cases} \tag{31}$$

By setting $f(M_{ij}) = (1-\rho)M_{ij}$ and $1 \geq M_{ij} \geq 0$, and assuming Ω contains all the n^2 entries, it is equivalent to our problem.

The estimator is obtained by solving the following optimization problem:

$$\hat{M} = \operatorname*{argmax}_{X:\|X\|_* \leq t} \sum_{i,j\in\Omega} \Bigg(\sum_{i,j:Y_{ij}=1} \log(f(X_{ij})) + \sum_{i,j:Y_{ij}=0} \log(1 - f(X_{ij})) \Bigg). \tag{32}$$

The following result shows that $\hat{M}$ is close to M:

Theorem 6 ((Davenport et al., 2012)). *Assume $\|M\|_* \leq t$, and Y is generated by (31), then*

$$\frac{1}{n^2}\|\hat{M} - M\|_F^2 \leq \sqrt{2} C_\alpha \sqrt{\frac{2nr}{m}}, \tag{33}$$

where $m = |\Omega|$, $C_\alpha := C_2 \alpha L_\alpha \beta_\alpha$ where C_2 is a constant and

$$L_\alpha = \sup_{|x|\leq\alpha} \frac{|f'(x)|}{f(x)(1-f(x))} \text{ and } \beta_\alpha = \sup_{|x|\leq\alpha} \frac{f(x)(1-f(x))}{(f'(x))^2}.$$

By substituting $f(x) = (1-\rho)x$ into the above formulas we can find $L_\alpha \geq \frac{1}{\alpha(1-(1-\rho)\alpha)}$ and $\beta_\alpha \geq \frac{\alpha(1-(1-\rho)\alpha)}{1-\rho}$, so $L_\alpha \beta_\alpha \geq \frac{1}{1-\rho}$. Therefore the above theorem suggests that

$$\frac{1}{n^2}\|\hat{M} - M\|_F^2 \leq O\bigg(\frac{\sqrt{nr}}{(1-\rho)\sqrt{m}}\bigg).$$

In our setting, $m = n^2$, so we have

$$\frac{1}{n^2}\|\hat{M} - M\|_F^2 \leq O\bigg(\frac{\sqrt{r}}{(1-\rho)\sqrt{n}}\bigg). \tag{34}$$

Thus the recovery error is $O\big(\frac{1}{(1-\rho)\sqrt{n}}\big)$, which implies that the sample complexity for recovery using this approach is quite high: For example, observing $O(n\log n)$ 1's, when M is dense, is not sufficient.

The main drawback of using this approach for PU matrix completion is computation — time complexity of solving (32) is $O(n^2)$ which makes the approach prohibitive for large matrices. Moreover, the average error on each element is $O(1/\sqrt{n})$ (in contrast, our algorithm has $O(1/n)$ average error). To see how this affects sample complexity for recovery, assume $\sum_{i,j} M_{i,j} = O(n^2)$ (number of 1's are of the same order as the number of 0's in the original matrix) and $O(n\log n)$ 1's are observed. Then $(1-\rho) = O(\frac{\log n}{n})$ and the average error according to (34) is $\frac{1}{n^2}\|\hat{M} - M\|_F^2 = O(\frac{\sqrt{rn}}{\log n})$, which diverges as $n \to \infty$. In contrast, we will show that the average error of our estimator vanishes as $n \to \infty$.

C. Hardness of applying noisy matrix completion to our deterministic setting

An easy way to model PU matrix completion problem in the deterministic setting is to think of it as a traditional matrix completion problem with n^2 "noisy" observed entries. In (Candès & Plan, 2009), it is assumed that $A = M + Z$ where Z is noise and $\delta = \|Z\|_F$. The idea is to solve:

$$\min \|X\|_* \text{ such that } \|\mathcal{P}_\Omega(X - A)\|_F \leq \delta, \tag{35}$$

where δ is total amount of noise. (Candès & Plan, 2009) established the following recovery guarantee:

Theorem 7 ((Candès & Plan, 2009)). *Let $M \in \mathbb{R}^{n\times n}$ be a fixed matrix of rank r, and assume M is μ-incoherent, i.e.,*

$$\|\boldsymbol{u}_i\|_\infty \leq \sqrt{\mu/n} \text{ and } \|\boldsymbol{v}_i\|_\infty \leq \sqrt{\mu/n}, \tag{36}$$

where $\boldsymbol{u}_i, \boldsymbol{v}_i$ are eigenvectors of M. Suppose we observe m entries of M with locations sampled uniformly at random, and

$$m \geq C\mu^2 nr \log^6 n, \tag{37}$$

where C is a numerical constant, then

$$\|\hat{M} - M\| \leq 4\sqrt{\frac{(2+p)n}{p}}\delta + 2\delta, \tag{38}$$

where $p = m/n^2$.

If we apply Theorem 7 to our case, $\delta = \|Z\|_F = \frac{\rho}{1-\rho}\bar{s}$, the error of the recovered matrix $\hat{M}$ using (35) can be bounded as:

$$\|\hat{M} - M\|_F \leq \|A - M\|_F, \tag{39}$$

and clearly this bound is not very useful.