Entry-Specific Bounds for Low-Rank Matrix Completion under Highly Non-Uniform Sampling

Xumei Xi*, Christina Lee Yu* and Yudong Chen†
*School of Operations Research and Information Engineering, Cornell University, Ithaca, NY, USA,
{xx269, cleeyu}@cornell.edu
†Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA, yudong.chen@wisc.edu

***Abstract*—Low-rank matrix completion concerns the problem of estimating unobserved entries in a matrix using a sparse set of observed entries. We consider the non-uniform setting where the observed entries are sampled with highly varying probabilities, potentially with different asymptotic scalings. We show that under structured sampling probabilities, it is often better and sometimes optimal to run estimation algorithms on a smaller submatrix rather than the entire matrix. In particular, we prove error upper bounds customized to each entry, which match the minimax lower bounds under certain conditions. Our bounds characterize the hardness of estimating each entry as a function of the localized sampling probabilities. We provide numerical experiments that confirm our theoretical findings.**

I. Introduction

Matrix completion concerns estimating a low-rank matrix given partial and potentially noisy observations of its entries [10], [12]. This problem has applications such as in collaborative filtering [23], system identification [21] and sensor localization [4]. Many algorithms with provable guarantees have been developed, including convex relaxation [7], [8], alternating minimization [16], [19] and spectral algorithms [17].

The early literature in matrix completion primarily focused on settings in which observations are uniformly distributed across the matrix, and the goal was to either derive conditions for exact recovery in the noiseless setting (e.g. [8], [17]) or characterize the mean squared error averaged across entries under observation noise (e.g. [18]). In recent years, there has been a growing interest towards relaxing the unrealistic uniform sampling requirements as well as obtaining more fine-grained, entrywise error bounds, especially as downstream decisions may be made by comparing estimates of individual entries. While these two goals have been pursued separately, few results address both simultaneously. In particular, when sampling is non-uniform, one expects that *some entries can be better estimated than the others.* Existing work falls short of capturing this phenomenon.

In this paper, we tackle the above two goals jointly to answer the following research questions. Can we obtain refined *entry-specific* error bounds under highly non-uniform sampling that correctly identifies the hardness of estimating each entry? Can we develop a computationally simple algorithm that is statistically efficient for estimating individual entries? When the sampling probabilities in different regions of the matrix have asymptotically different orders of magnitude, one would hope that we can retain high performance for entries in regions of the matrix with high sampling probabilities, while still providing optimal estimates for entries in regions with low sampling probabilities. Our results provide entry-specific error guarantees as a function of the localized sampling probabilities. We further show that our bounds are minimax optimal for structured sampling probabilities.

We design a meta algorithm that can be combined with any matrix estimation method; for concreteness, we use Singular Value Thresholding (SVT) [5]. For each target entry (i, j), our method chooses a submatrix to input into SVT (or any matrix estimation algorithm), with the goal of obtaining a better estimate of (i, j) than applying SVT to the entire matrix. This algorithm allows us to obtain a more refined estimation error bound that has varying rates across entries. We perform numerical experiments on synthetic datasets that confirm our theoretical findings.

A. Related Work

Several recent works consider matrix completion with the non-uniform observation pattern. Using graph limit theory, [9] shows that a sequence of matrices is asymptotically recoverable in mean squared error if the deterministic sampling patterns converge to a graphon[1] that is nonzero almost everywhere. This requirement implies the sampling cannot be too non-uniform or sparse. Meanwhile, [13] considers non-uniform deterministic sampling patterns and proposes a simple algorithm with a weighted mean-squared error guarantee dependent on a dissimilarity function between the weights and the sampling pattern. The work [6] studies max-norm relaxation method and shows that it achieves minimax Frobenius norm error under moderately non-uniform sampling.

A related line of work uses structured graphs to construct the sampling pattern or the weight matrix. For example, [3] uses a bipartite graph with a large spectral gap as the sampling pattern, whereas [15] uses expander graph and other graph sparsifiers. Complementarily, [20] considers the setting where the sampling pattern is fixed and aims to choose a weight matrix that yields weighted mean-squared error bounds.

Going beyond (weighted) mean-squared error, several recent works consider entrywise ℓ_∞ error, that is, the worst-case estimation error across all entries. The work [1] provides entrywise error bounds for SVT using ℓ_∞ eigenspace perturbation

[1]A graphon is a symmetric measurable function, which serves as the limit of a sequence of dense graphs. Interested readers can consult [22].

analysis. Under uniform sampling, [11] provides entrywise guarantees for convex relaxation and non-convex approach. The paper [2] considers deterministic sampling, and their algorithm searches for an almost fully observed submatrix containing the entry to be estimated. While their approach bears some similarities with ours, we note that our results consider random sampling and allow for significantly sparser observations.

II. Problem setup

Notation: We use c, C etc. to denote positive absolute constants, which might change from line to line. Let $[n] := \{1, 2, \ldots, n\}$. For non-negative sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ when $a_n \leq Cb_n, \forall n$, and write $a_n \asymp b_n$ or $a_n = \Theta(b_n)$ when both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. Let $M_{\mathcal{U},\mathcal{V}}$ denote the submatrix of $M \in \mathbb{R}^{n\times m}$ indexed by $\mathcal{U} \subseteq [n]$ and $\mathcal{V} \subseteq [m]$, and $\|M\|_\infty = \max_{i,j}|M_{ij}|$ the entrywise ℓ_∞ norm.

A. Latent Variable Model

Our goal is to estimate the entries of a low-rank signal matrix $M^* \in \mathbb{R}^{n\times m}$ given noisy partial observations. We consider a latent variable model, where M^* is generated via

$$M^*_{ij} = \langle a^*_i, b^*_j \rangle, \tag{1}$$

and the row latent variables $a^*_i \in \mathbb{R}^r, i \in [n]$ are sampled i.i.d. from some distribution; similarly for the column latent variables $b^*_j \in \mathbb{R}^r, j \in [m]$. If the distributions of $\{a_i\}$ and $\{b_j\}$ are sufficiently regular (e.g., sub-exponential with a non-degenerate covariance matrix), then with high probability the matrix M^* is rank-r and has a bounded incoherence parameters [25]. For concreteness, we consider Gaussian latent factors: $a^*_i \overset{\text{i.i.d.}}{\sim} N(0, I_r)$ and $b^*_j \overset{\text{i.i.d.}}{\sim} N(0, I_r)$. We are given a noisy and partially observed matrix

$$Y = \Omega \circ (M^* + E), \tag{2}$$

where $E \in \mathbb{R}^{n\times m}$ is additive noise with $E_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, and $\Omega \in \{0, 1\}^{n\times m}$ is the sampling/mask matrix generated as $\Omega_{ij} \sim \text{Bernoulli}(P_{ij})$, independently across entries. Given Y, the goal is to estimate the entries of M^*. We assume $m \asymp n$ and the rank $r \ll n$ is known.

B. Monotone Sampling Probabilities

In the above model, the observations are non-uniform as determined by the sampling probability matrix $P = (P_{ij}) \in [0, 1]^{n\times m}$. We assume that P is known, which is a reasonable assumption in settings where the learner has (partial) control over the sampling process or can estimate P from data. Without restriction on P, this setting includes arbitrary deterministic sampling pattern as a special case—just let P_{ij} be binary—under which matrix completion is NP-hard [14]. Therefore, we further assume P has a monotone structure: there exist permutations $\pi_n : [n] \to [n]$ and $\pi_m : [m] \to [m]$ such that $P_{\pi_n(i)\pi_m(j)} \geq P_{\pi_n(i')\pi_m(j')}$ whenever $\pi_n(i) \leq \pi_n(i')$ and $\pi_m(j) \leq \pi_m(j')$. Without loss of generality, we may assume both π_n and π_m are the identity:[2]

[2]If π_n and π_m exist, they can be found by sorting the rows of P and then the columns.

Assumption 1 (Monotonicity): The probability matrix P satisfies $P_{ij} \geq P_{i'j'}$ if $i \leq i'$ and $j \leq j'$.

Assumption 1 is satisfied, e.g., in a movie rating setting, where the probability of user i rating movie j is determined by the activeness of the user and the popularity of the movie. In fact, this example corresponds to a special case of Assumption 1 where P has rank one, as stated below. We sometimes consider this stronger assumption.

Assumption 2 (Rank-one P): There exist vectors $\alpha = (\alpha_1, \ldots, \alpha_n)$ and $\beta = (\beta_1, \ldots, \beta_m)$ such that

$$P_{ij} = \alpha_i \beta_j, \quad \forall (i, j) \in [n] \times [m],$$

where $1 \geq \alpha_1 \geq \cdots \geq \alpha_n \geq 0$ and $1 \geq \beta_1 \geq \cdots \geq \beta_m \geq 0$.

III. Algorithm: submatrix completion

Let $\texttt{MC}(\cdot)$ be a black-box matrix completion subroutine, e.g., SVT. If $\texttt{MC}(\cdot)$ were applied on the entire observed matrix Y, it outputs an estimate $\hat{M} = \texttt{MC}(Y) \in \mathbb{R}^{n\times m}$ of the true signal matrix. Our algorithm, *submatrix completion*, instead applies $\texttt{MC}(\cdot)$ to carefully chosen submatrices of Y. In particular, for each target entry (i, j) to be estimated, we compute an index

$$k^* \equiv k^*(i, j) = \operatorname*{argmax}_{k \leq \min\{n,m\}} k \cdot \min\{P_{\max\{i,k\},k}, P_{k,\max\{j,k\}}\}. \tag{3}$$

We then apply $\texttt{MC}(\cdot)$ on the submatrix indexed by $[k^*] \cup \{i\}$ and $[k^*] \cup \{j\}$, and use the corresponding entry of the output $\texttt{MC}(Y_{[k^*]\cup\{i\},[k^*]\cup\{j\}})$ as an estimate of the target entry M^*_{ij}. In the optimization problem (3), the variable k corresponds to the size of the submatrix used to estimate (i, j), $P_{\max\{i,k\},k}$ is the smallest probability on the last row of the submatrix excluding entry (i, j), and $P_{k,\max\{j,k\}}$ is the smallest probability on the last column. As will become clear in Section IV-A, k^* is chosen to minimize an upper bound on the entrywise estimation error of the submatrix.

For illustration and ease of analysis, we adopt SVT as the matrix completion subroutine $\texttt{MC}(\cdot)$. Given the observation Y, SVT forms the rescaled observation matrix $\bar{Y} = (Y_{ij}/P_{ij})_{i\in[n],j\in[m]}$ (which is an unbiased estimator of M^*), and then computes the best rank-r approximation $\hat{M}$ of $\bar{Y}$. Explicitly, if $\bar{Y}$ has singular value decomposition (SVD) $\bar{Y} = \bar{U}\bar{\Sigma}\bar{V}^\top$, where $\bar{\Sigma}$ is a diagonal matrix containing the singular values of $\bar{Y}$ in descending order, then $\hat{M} = \bar{U}_{\cdot[r]}\bar{\Sigma}_{[r],[r]}\bar{V}^\top_{\cdot[r]}$.

A. Illustrating example

We illustrate how our algorithm works with a concrete example. To this end, let us first derive a useful characterization of the solution $k^*(i, j)$ to (3). Define

$$i^* := \operatorname*{argmax}_i iP_{ii}. \tag{4}$$

We call the submatrix indexed by $[i^*] \times [i^*]$ the *core submatrix*. Under the monotone Assumption 1, for all entries (i, j) in the core submatrix, the solution $k^*(i, j)$ coincides with i^*:

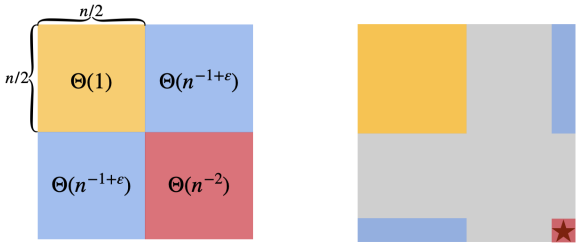
Lemma 1: Under Assumption 1, if $i \leq i^*$ and $j \leq i^*$, then $k^*(i, j) = i^*$.

Proof: Note that $i^* P_{i^* i^*} \geq k P_{kk}$ for all k by definition. Additionally, we have $i^* P_{i^* i^*} \geq k P_{kk} \geq k P_{ik}$ for all $k < i$. For all $k < j$, we also have $i^* P_{i^* i^*} \geq k P_{kk} \geq k P_{kj}$. ∎

For our example, suppose the probability matrix P can be divided into four equal-size blocks, where the probabilities inside each block are the same up to constants, having the following structure:

$$P_{ij} = \begin{cases} \Theta(1) & i \leq n/2, j \leq n/2 \\ \Theta(n^{-2}) & i > n/2, j > n/2 \\ \Theta(n^{-1+\varepsilon}) & \text{otherwise,} \end{cases}$$

for some $0 < \varepsilon < 1$. See Fig. 1(a) for an illustration of P.



(a) Sturcture of P. (b) Submatrix completion.

Fig. 1. An example of using submatrix completion with monotone P.

Since the smallest probability (red block) scales as $p_{\min} \asymp n^{-2}$, existing entrywise error bound [1] gives $\|\hat{M} - M^*\|_\infty = O(1/\sqrt{p_{\min} n})$, which is vacuous as $p_{\min}$ is so small.[3] In contrast, our submatrix completion algorithm achieves much better guarantees. In particular, it can be verified that the yellow block is the core submatrix. By Lemma 1, to estimate entries inside the yellow block, our algorithm will apply SVT on this block itself and achieve an entrywise error bound of $O(1/\sqrt{n})$. For each entry (i, j) in the red block, our algorithm will use the submatrix $Y_{[n/2]\cup\{i\},[n/2]\cup\{j\}}$ (see Fig. 1(b)) and achieve an $O(1/\sqrt{n^\varepsilon})$ error. Note that our error bounds are independent of $p_{\min}$.

This example provides intuition for why it can be beneficial to only use a subset of the observations for estimation. SVT applied to the entire matrix would try too hard to fit the (noisy) observations in the blue and red blocks with low sampling probabilities (and hence high variances). As a result, the estimation error for the high probability yellow block would be worse than using observations only from this block.

IV. Theoretical Guarantees

In this section, we present entry-specific error upper bounds for our submatrix completion algorithm coupled with SVT. We also derive entry-specific minimax lower bounds, which match the upper bound for structured P.

A. Upper bound for our algorithm

We derive an error bound for estimating a specific entry (i, j). Set $p^*(i, j) = \min\{P_{\max\{i,k^*\},k^*}, P_{k^*,\max\{j,k^*\}}\}$ with $k^* \equiv k^*(i, j)$ being the solution to (3). Let $\hat{M}_{ij}$ be the estimate of M^*_{ij} given by our submatrix completion algorithm.

Theorem 1: Under Assumption 1, with probability at least $1 - \delta$, for each (i, j) satisfying $p^*(i, j) \geq \frac{c \log(n/\delta)}{k^*(i,j)}$, we have

$$\left|\hat{M}_{ij} - M^*_{ij}\right| \leq Cr(r + \sigma)\sqrt{\frac{\log^5(n/\delta)}{k^*(i, j) p^*(i, j)}}. \tag{5}$$

In the above upper bound, the denominator inside the square root is a function of the index (i, j), where $k^*(i, j)$ is chosen precisely to maximize this denominator and hence optimize the error bound. As the denominator is increasing in the size of the submatrix but decreasing as the minimum probability decreases, our bound highlights the tension in the choice of the submatrix. A large submatrix may have a large size but a small minimum probability; a small submatrix has a small size but could have a larger minimum probability. This flexibility of choosing an appropriate submatrix enables us to obtain fine-grained error bounds that are specific to each entry. Compared with the uniform worst-case entrywise bound stated in Theorem 3, our bound is valid even when $p_{\min}$ does not meet the condition therein. Furthermore, our bound is able to capture the potential order-wise difference between the estimation quality of different entries, as demonstrated in the example from Section III.

B. Minimax lower bound

We present an entry-specific minimax lower bound on the estimation error. The following theorem is valid for any sampling probability matrix P.

Theorem 2: Fix $i \in [n]$ and $j \in [m]$. There exists an absolute constant $C > 0$ such that

$$\inf_{\hat{M}_{ij}} \sup_{M^*} \mathbb{E}\left[\left|\hat{M}_{ij} - M^*_{ij}\right|\right] \geq C\sigma\sqrt{\frac{r}{\min\{\sum_{i'} P_{i'j}, \sum_{j'} P_{ij'}\}}}, \tag{6}$$

with probability at least $\frac{1}{2}$. Here, the infimum is over all estimators of $M^*_{i,j}$, the supremum is over all rank-r M^*, the expectation is w.r.t. the additive noise E, and the probability is w.r.t. the sampling mask Ω.

We prove Theorem 2 by reduction from noisy linear regression. Consider estimating the entry $M^*_{ij} = \langle a^*_i, b^*_j \rangle$. If the row latent factors $\{a^*_i\}_{i\in[n]}$ were known, then estimating M^*_{ij} is the same as the linear regression problem of estimating the j-th column latent factor b^*_j given noisy observations $Y_{i'j} = \langle a^*_{i'}, b^*_j \rangle + E_{i'j}$ for $i' \in [m] : \Omega_{i'j} = 1$; note that we have $\sum_{i'} P_{i'j}$ such observations in expectation. As the original problem is at least as hard as the regression problem, we can use standard minimax lower bounds for linear regression to derive a lower bound for our problem. A similar lower bound can be derived by assuming the column factors were known. Taking the larger of these two bounds proves Theorem 2.

Perhaps surprisingly, for quite general settings of monotone P, the simple lower bound above matches the upper bound in Theorem 1 (up to logarithmic factors), in which case our algorithm is information-theoretically optimal for estimating each entry. We discuss such a setting below.

[3] We ignore logarithmic factors and dependence on the rank r.

C. Example: block-structured P

We discuss a generalization of the example from Section III-A. Suppose P can be partitioned into 4 blocks: [4] $P = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}$, where $Q_{11} \in \mathbb{R}^{n_1 \times n_1}, Q_{12} \in \mathbb{R}^{n_1 \times n_2}, Q_{21} \in \mathbb{R}^{n_2 \times n_1}$, $Q_{22} \in \mathbb{R}^{n_2 \times n_2}$, and $n_1 + n_2 = n$. Inside each block, the probabilities are of the same order but can be otherwise different. Let the minimum probabilities of the 4 blocks be $q_{11}, q_{12}, q_{21}, q_{22}$. Assume the probabilities satisfy $n_1 q_{11} \gtrsim n q_{12}$, $n_1 q_{11} \gtrsim n q_{21}$, $n_1 q_{12} \gtrsim n q_{22}$, and $n_1 q_{21} \gtrsim n q_{22}$. (This assumption is satisfied when, for example, P is monotone and $n_1 \gtrsim n_2$.) One may verify that for estimating entry (i, j), our algorithm will pick the submatrix $Y_{[n_1]\cup\{i\},[n_1]\cup\{j\}}$. Applying Theorems 1 and 2 to each block, we obtain the following matching upper and lower bounds (omitting log factors):[5]

1) When $i \le n_1, j \le n_1$, the upper bound is $1/\sqrt{n_1 q_{11}}$, and the lower bound is $1/\sqrt{n_1 q_{11} + n_2 q_{12}}$.
2) When $i \le n_1, j > n_1$, the upper bound is $1/\sqrt{n_1 q_{12}}$, and the lower bound is $1/\sqrt{n_1 q_{12} + n_2 q_{22}}$.
3) When $i > n_1, j \le n_1$, the upper bound is $1/\sqrt{n_1 q_{21}}$, and the lower bound is $1/\sqrt{n_1 q_{21} + n_2 q_{22}}$.
4) When $i > n_1, j > n_2$, the upper bound is $1/\sqrt{n_1 \min\{q_{12}, q_{21}\}}$, and the lower bound is $1/\sqrt{n_1 \min\{q_{12}, q_{21}\} + n_2 q_{22}}$.

V. Key ideas of proof

In this section, we present the tools for proving the upper bound in Theorem 1. The high-level idea is to apply the entrywise error bound for SVT from [1] to the submatrix chosen by our algorithm. To do so, their result needs to be adapted to the setting where the submatrix may have non-uniform sampling probabilities and one entry (the target entry to be estimated) may have an arbitrarily small probability.

We consider estimating a deterministic rank-r matrix $A^* \in \mathbb{R}^{n \times m}$ given noisy observations $Y = \Omega \circ (A^* + E)$, where the mask Ω and noise E are the same as before. Here, A^* can be either the whole matrix M^* introduced earlier or a submatrix of M^*. Let the rank-r SVD of A^* be $A^* = U^* \Sigma^* V^{*\top}$. Define $\kappa = \frac{\sigma_1^*}{\sigma_r^*}$ and $\eta = (\|U^*\|_{2\to\infty} \vee \|V^*\|_{2\to\infty})$. Recall that the SVT algorithm forms the rescaled matrix $A_{ij} = (Y_{ij}/P_{ij})$ and computes the rank-r truncated SVD $U \Sigma V^\top$ of A.

A. Guarantee for SVT with non-uniform observations

Leveraging the results from [1], we establish the following entrywise error bound for SVT under non-uniform sampling probabilities P. Let $p_{\min} := \min_{i,j} P_{ij}$.

Theorem 3: Suppose $p_{\min} \ge \frac{c \log(n/\delta)}{n+m}$ and $\kappa \frac{(\|A^*\|_\infty + \sigma)}{\sigma_r^*} \sqrt{\frac{(n+m)\log(n/\delta)}{p_{\min}}} \le 1$ for some $\delta > 0$. With probability at least $1 - \delta$, we have

$$\left\|U\Sigma V^\top - A^*\right\|_\infty \le C\eta^2\kappa^4(\|A^*\|_\infty + \sigma)\sqrt{\frac{(n+m)\log(n/\delta)}{p_{\min}}}. \tag{7}$$

[4] This example can be generalized to P with $O(1)$ blocks.

[5] We impose the mild assumption $\min\{q_{11}, q_{12}, q_{21}\} \gtrsim \frac{1}{n_1}$, so that the problem is non-trivial with at least one observation in each row/column.

The bound (7) depends on the smallest sampling probability $p_{\min}$. This bound becomes vacuous when just a single entry (i, j) has a very small sampling probability P_{ij}. We next present an improved bound, which is unaffected by a few entries with small probabilities. This improvement plays a crucial role in proving our main Theorem 1.

B. Improved bound under a few small probabilities

Let s be a non-negative integer. Let $p_{(1)} \ge p_{(2)} \ge \cdots \ge p_{(nm)}$ denote the probabilities $\{P_{ij}\}$ sorted in descending order. Note that $p_{(nm-s)}$ is the $(s+1)$-th smallest value in $\{P_{ij}\}$ and in particular $p_{(nm)} = p_{\min}$. The following theorem gives an error bound that only depends on $p_{(nm-s)}$. In order for the argument to hold, we need to slightly modify the way we rescale the observation matrix Y. Suppose $(i_{nm-s'}, j_{nm-s'})$ indicates the position of the probability $p_{(nm-s')}$ for $0 \le s' \le s-1$. Let $A_{i_{nm-s'}, j_{nm-s'}} = 2Y_{i_{nm-s'}, j_{nm-s'}}$, where we replace the probability $p_{nm-s'}$ in the denominator with $\frac{1}{2}$, enabling us to ignore probabilities smaller than $p_{(nm-s)}$.

Theorem 4: Suppose $p_{(nm-s)} \ge \frac{c \log(n/\delta)}{n+m}$ and $\kappa \frac{(\|A^*\|_\infty + \sigma)}{\sigma_r^*} \sqrt{\frac{(n+m)\log(n/\delta)}{p_{(nm-s)}}} \le 1$ for some $\delta > 0$. With probability at least $1 - \delta$, we have

$$\left\|U\Sigma V^\top - A^*\right\|_\infty \le C\eta^2\kappa^4(\|A^*\|_\infty + \sigma)\sqrt{\frac{(n+m)(s+\log(n/\delta))}{p_{(nm-s)}}}. \tag{8}$$

Compared with the bound (7), the denominator on the right hand side of (8) improves from $p_{(nm)}$ to $p_{(nm-s)}$, at the cost of the numerator increasing by s. This cost is negligible whenever $s = O(\log(n/\delta))$. To see the benefit, consider applying Theorem 4 with $s = 1$ to the submatrix in Fig. 1(b), for which we obtain an error bound that depends on $p_{(nm-1)} = \Theta(n^{-1+\epsilon})$ instead of $p_{\min} = \Theta(n^{-2})$.

VI. Numerical experiments

In this section, we numerically evaluate our algorithm. We compare two algorithms: (i) our algorithm SVT-sub, which applies SVT to submatrices, and (ii) SVT-whole, which applies SVT to the entire matrix. We consider two monotone probability matrices P. For each case, we randomly generate a 100-by-100 signal matrix M^* of rank $r = 2$ according to the latent variable model in Section II-A with noise standard deviation $\sigma = 0.1$. This is repeated for 100 trials. We record the error e_{ij}^{sub} and e_{ij}^{whole} for estimating each entry (i, j) using SVT-sub and SVT-whole, respectively, averaged over 100 trials.

A. Block-constant P matrix

We first consider a block-constant $P \in [0, 1]^{100\times 100}$ with

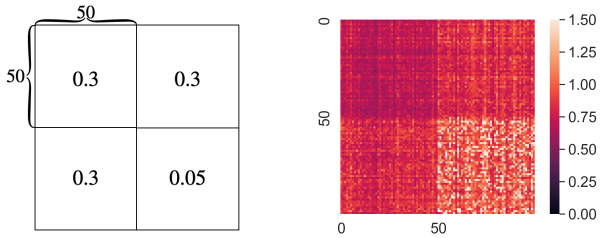
$$P_{ij} = \begin{cases} 0.3 & i \le 50 \text{ or } j \le 50, \\ 0.05 & i > 50 \text{ and } j > 50, \end{cases}$$

which is visualized in Fig. 2(a). In this case, our algorithm SVT-sub uses the submatrix $M_{[50]\cup\{i\},[50]\cup\{j\}}$ to estimate each entry (i, j). We plot the heatmaps of the errors e_{ij}^{sub} and e_{ij}^{whole} in Fig. 2(b) and (c), respectively. We observe that SVT-sub

achieves a smaller error, especially in the three 0.3 blocks. We further compute the *relative improvement* for estimating each entry (i, j), defined as $(e_{ij}^{\text{whole}} - e_{ij}^{\text{sub}})/e_{ij}^{\text{whole}}$, which represents the percentage of improvement of SVT-sub over SVT-whole. We plot the relative improvements in Fig. 2(d), which shows that the most substantial improvement happens in the two off-diagonal blocks. In particular, SVT-sub improves upon SVT-whole by 12.7% on average over the top-left block, by 21.3% over the two off-diagonal blocks, and by 14.5% over the bottom-right block. We also plot the histogram of the relative improvements in Fig. 4(a), showing a strong trend of positive improvement.

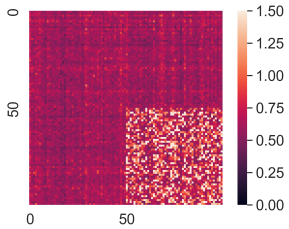
B. Rank-one P matrix

We consider a rank-one $P = ab^\top$, where a and b are sampled randomly from the same distribution. In particular, for $i \le 80$, we have $a_i \sim 0.5 \cdot \text{Beta}(5, 2) + 0.5$; for $i > 80$, we have $a_i \sim 0.5 \cdot \text{Beta}(5, 2)$. We sort the entries of a and b in descending order for better visualization. One realization of P is shown in Fig 3(a), in which we observe a clear drop in the probabilities near the 80th row/column. In this realization, the largest and smallest values of P_{ij} are 0.989 and 0.025, respectively, hence the sampling probabilities are highly non-uniform. The core matrix (see Section III-A) obtained by solving (4) is a 74-by-74 submatrix. We plot the heatmaps of the entrywise error e_{ij}^{sub} and e_{ij}^{whole} in Fig. 3(b) and (c), respectively, as well as the relative improvements in Fig. 3(d) and Fig. 4(b). We observe that the majority of the entries benefit from using our SVT-sub algorithm. On average, the relative improvement is 17.7%.

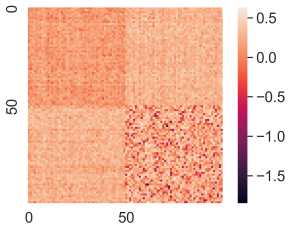


(a) 2-by-2 block-constant P.

(b) Error by SVT-whole.



(c) Error by SVT-sub.



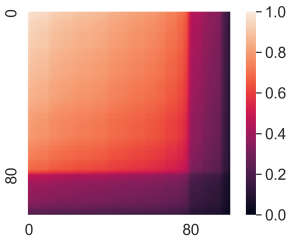
(d) Relative improvement.

Fig. 2. Heatmaps for Subsection VI-A.

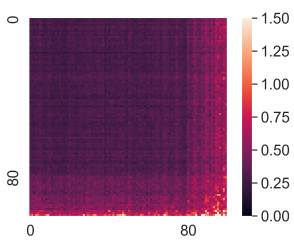
VII. Discussion

We propose a submatrix completion algorithm, which hand-picks a submatrix for estimating a specific entry based on the sampling probabilities and then applies the matrix estimation subroutine to the selected submatrix. Using SVT as the subroutine, we establish entry-specific upper bound and minimax lower bound on the estimation error. Under certain sampling probability patterns, the upper and lower bounds match up to log factors. We also present numerical experiments that demonstrate the benefit of our algorithm. Future directions include combining our algorithm with other matrix estimation algorithms, as well as extending the results to more general probability patterns, such as those that are not monotone globally but may satisfy local monotonicity.

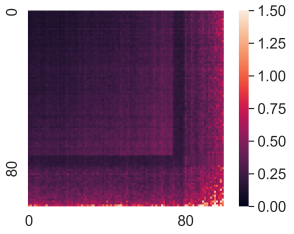
Acknowledgement: C. Yu is partially supported by NSF grants CCF-1948256 and CNS-1955997, AFOSR grant FA9550-23-1-0301, and by an Intel Rising Stars award. Y. Chen is partially supported by NSF grants CCF-1704828 and CCF-2047910.



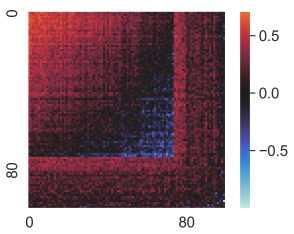
(a) Rank-1 P.



(b) Error by SVT-whole.

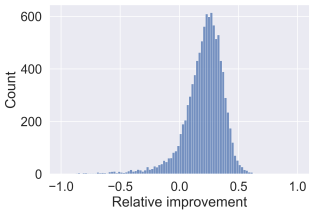


(c) Error by SVT-sub.

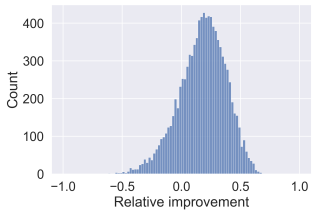


(d) Relative improvement.

Fig. 3. Heatmaps for Subsection VI-B.



(a) For Subsection VI-A.



(b) For Subsection VI-B.

Fig. 4. Histograms of relative improvement.

REFERENCES

[1] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, 48 3:1452–1474, 2020.

[2] Anish Agarwal, Munther Dahleh, Devavrat Shah, and Dennis Shen. Causal matrix completion. *arXiv preprint arXiv:2109.15154*, 2021.

[3] Srinadh Bhojanapalli and Prateek Jain. Universal matrix completion. In *International Conference on Machine Learning*, pages 1881–1889. PMLR, 2014.

[4] Pratik Biswas, Tzu-Chen Lian, Ta-Chung Wang, and Yinyu Ye. Semidefinite programming based algorithms for sensor network localization. *ACM Trans. Sen. Netw.*, 2(2):188–220, may 2006.

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

[6] T Tony Cai and Wen-Xin Zhou. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10:1493–1525, 2016.

[7] Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

[8] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[9] Sourav Chatterjee. A deterministic theory of low rank matrix completion. *IEEE Transactions on Information Theory*, 66(12):8046–8055, 2020.

[10] Yudong Chen and Yuejie Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31, 2018.

[11] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121, 2020.

[12] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[13] Simon Foucart, Deanna Needell, Reese Pathak, Yaniv Plan, and Mary Wootters. Weighted matrix completion from non-random, non-uniform sampling patterns. *IEEE Transactions on Information Theory*, 67(2):1264–1290, Feb 2021.

[14] Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. In *Conference on Learning Theory*, pages 703–725. PMLR, 2014.

[15] Eyal Heiman, Gideon Schechtman, and Adi Shraibman. Deterministic algorithms for matrix completion. *Random Struct. Algorithms*, 45(2):306–317, sep 2014.

[16] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, page 665–674, New York, NY, USA, 2013. Association for Computing Machinery.

[17] Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Advances in neural information processing systems*, 22, 2009.

[18] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

[19] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[20] Troy Lee and Adi Shraibman. Matrix completion from any given set of observations. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[21] Zhang Liu and Lieven Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2010.

[22] László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

[23] Jasson D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 713–719, New York, NY, USA, 2005. Association for Computing Machinery.

[24] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[25] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[26] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

APPENDIX A
TECHNICAL LEMMAS

Before we present the technical lemmas, we define some notations. For a vector $x \in \mathbb{R}^n$, we define $\|x\|_2 = \sqrt{\sum_{i\in[n]} x_i^2}$ and $\|x\|_\infty = \max_{i\in[n]}|x_i|$. For a matrix $M \in \mathbb{R}^{n\times m}$, let $M_{i\cdot}$ denote its i-th row and $M_{\cdot j}$ its j-th column. Let the operator norm of matrix $M \in \mathbb{R}^{n\times m}$ be $\|M\|_{\mathrm{op}} = \max_{\|x\|_2=1,\|y\|_2=1} x^\top My$ and the $2\to\infty$ norm be $\|M\|_{2\to\infty} = \max_{\|x\|_2=1}\|Mx\|_\infty = \max_i \|M_{i\cdot}\|_2$.

Recall that in the latent variable model introduced in Subsection II-A, the signal matrix $M^* \in \mathbb{R}^{n\times m}$ is generated via $M^*_{ij} = \langle a_i^*, b_j^*\rangle$. The latent variables are standard Gaussians: $a_i^* \overset{\text{i.i.d.}}{\sim} N(0, I_r)$ and $b_j^* \overset{\text{i.i.d.}}{\sim} N(0, I_r)$, independently. We introduce the following lemmas to show the signal matrix is low-rank and bounded with high probability, the proof of which is deferred to Appendix A-A and Appendix A-B.

Lemma 2 (Low-rank signal): Assume that $\delta > 0$ satisfy $r + \log(4/\delta) \le \frac{1}{C}\min\{n, m\}$, for a sufficiently large absolute constant $C > 0$. Then, with probability at least $1-\delta$, matrix M^* is rank-r.

Lemma 3 (Bounded signal): With probability at least $1-\delta$, we have

$$\|M^*\|_\infty \le 2r\log\frac{nmr^2}{\delta}. \tag{9}$$

The next lemma shows that the signal matrix M^* is incoherent and well-conditioned, with its proof in Appendix A-C.

Lemma 4 (Incoherence & condition number guarantee): Let the SVD of M^* be $M^* = U^*\Sigma^*V^{*\top}$. There exists a sufficiently large absolute constant $C > 0$ such that with probability at least $1-\delta$, we have

$$\kappa(M^*) \le \frac{1 + C\sqrt{\frac{r+\log(2(n+m+2)/\delta)}{\min\{n,m\}}}}{1 - C\sqrt{\frac{r+\log(4(n+m+2)/\delta)}{\min\{n,m\}}}}. \tag{10}$$

Furthermore, M^* is incoherent:

$$\begin{aligned}\|U^*\|_{2\to\infty} &\le C\frac{\sqrt{r} + \sqrt{\log(2(n+m+2)/\delta)}}{\sqrt{n}}\\ \|V^*\|_{2\to\infty} &\le C\frac{\sqrt{r} + \sqrt{\log(2(n+m+2)/\delta)}}{\sqrt{m}}.\end{aligned} \tag{11}$$

A. Proof of Lemma 2

Write $A = \begin{pmatrix} a_1^\top \\ a_2^\top \\ \vdots \\ a_n^\top \end{pmatrix} \in \mathbb{R}^{n\times r}$ and $B = \begin{pmatrix} b_1^\top \\ b_2^\top \\ \vdots \\ b_m^\top \end{pmatrix} \in \mathbb{R}^{m\times r}$. We rescale the Gaussian vectors and define $U = \frac{1}{\sqrt{n}}A$ and $V = \frac{1}{\sqrt{m}}B$. Then, we have $U^\top U = \frac{1}{n}\sum_{i\in[n]} a_i a_i^\top$ and $V^\top V = \frac{1}{m}\sum_{i\in[m]} b_i b_i^\top$. By a standard argument of covariance estimation [26], there exists a sufficiently large absolute constant $C > 0$ such that

$$\left\|U^\top U - I\right\|_{\mathrm{op}} \le C\sqrt{\frac{r + \log(4/\delta)}{n}}, \tag{12}$$

with probability at least $1 - \delta/2$. By the same argument, we have

$$\left\|V^\top V - I\right\|_{\mathrm{op}} \le C\sqrt{\frac{r + \log(4/\delta)}{m}}, \tag{13}$$

with probability at least $1 - \delta/2$. Applying the union bound on the aforementioned two events yields that with probability at least $1-\delta$, we have $\mathrm{rank}(U) = r$ and $\mathrm{rank}(V) = r$, since all singular values of U and V are concentrated around 1, by Weyl's inequality. Because $M^* = \sqrt{nm}UV^\top$, we immediately have $\mathrm{rank}(M) \le \min\{\mathrm{rank}(U), \mathrm{rank}(V)\} = r$. Finally, invoking Sylvester's rank inequality, we get $\mathrm{rank}(M^*) \ge \mathrm{rank}(U) + \mathrm{rank}(V) - r = r$. As a result, $\mathrm{rank}(M^*) = r$ with probability at least $1-\delta$.

B. Proof of Lemma 3

Note that for each $a_{ik}^* \sim N(0,1)$ and $b_{j\ell}^* \sim N(0,1)$, we have the Gaussian tail bound

$$\mathbb{P}(|a_{ik}^*| \ge t) \le \exp\left(-\frac{t^2}{2}\right). \tag{14}$$

We then derive the tail bound on the maximum of $|a_{ik}^*|$ and $\left|b_{j\ell}^*\right|$ as

$$\begin{aligned}&\mathbb{P}(\max\{\max_{i\in[n],k\in[r]}|a_{ik}^*|, \max_{j\in[m],\ell\in[r]}\left|b_{j\ell}^*\right|\} \ge t)\\ \le&\mathbb{P}(\{\cup_{i,k}\{|a_{ik}^*| \ge t\}\} \cup \{\cup_{j,\ell}\{\left|b_{j\ell}^*\right| \ge t\}\})\\ \le&\sum_{i,k}\mathbb{P}(|a_{ik}^*| \ge t) + \sum_{j,\ell}\mathbb{P}(\left|b_{j\ell}^*\right| \ge t)\\ \le&nmr^2\exp\left(-\frac{t^2}{2}\right).\end{aligned}$$

Let $t = \sqrt{2\log\frac{nmr^2}{\delta}}$ and we get

$$\mathbb{P}\left(\max\{\max_{i\in[n],k\in[r]}|a_{ik}^*|, \max_{j\in[m],\ell\in[r]}\left|b_{j\ell}^*\right|\} \ge \sqrt{2\log\frac{nmr^2}{\delta}}\right) \le \delta.$$

Hence, we conclude that $\max_{ij}\left|M_{ij}^*\right| \le 2r\log\frac{nmr^2}{\delta}$ with probability at least $1-\delta$.

C. Proof of Lemma 4

Continuing from the previous proof, we have $M^* = \sqrt{nm}UV^\top$ where U and V satisfy (12) and (13). Let the (reduced) QR decomposition of U and V be

$$\begin{aligned}U &= Q^U R^U\\ V &= Q^V R^V.\end{aligned}$$

Plugging the above decomposition into $M^* = \sqrt{nm}UV^\top$, we have

$$\begin{aligned}M^* &= \sqrt{nm}Q^U(R^U R^{V\top})Q^{V\top}\\ &= \sqrt{nm}Q^U(P^R\Sigma^R Q^{R\top})Q^{V\top},\end{aligned}$$

where in the last line, we further decompose matrix $R^U R^{V\top}$ by its singular vectors and singular values. Rewriting the last line, we get

$$M^* = \sqrt{nm}(Q^U P^R)\Sigma^R(Q^V Q^R)^\top = U^*\Sigma^*V^{*\top},$$

where Note that both $Q^U P^R$ and $Q^V Q^R$ are orthogonal. Hence, we derive the two-to-infinity norm bound of U^* and V^* by

$$\|U^*\|_{2\to\infty} \le \|Q^U\|_{2\to\infty}\|P^R\|_{\rm op} = \|Q^U\|_{2\to\infty}$$
$$\|V^*\|_{2\to\infty} \le \|Q^V\|_{2\to\infty}\|Q^R\|_{\rm op} = \|Q^V\|_{2\to\infty},$$

using the fact that orthogonal matrices are norm-preserving. Note that Q^U and Q^V are comprised of (right) singular vectors of U and V, respectively. We can bound the two-to-infinity norm of Q^U by the two-to-infinity norm of U and the inverse of the largest singular value of U:

$$\begin{aligned}\|Q^U\|_{2\to\infty} &= \|UR^{U,-1}\|_{2\to\infty} \\ &\le \|U\|_{2\to\infty}\|R^{U,-1}\|_{\rm op} \\ &= \|U\|_{2\to\infty}\|U\|_{\rm op}^{-1}.\end{aligned}$$

Similarly, we get

$$\|Q^V\|_{2\to\infty} \le \|V\|_{2\to\infty}\|V\|_{\rm op}^{-1}.$$

Next, we condition on the union of several high-probability events to upper bound the norms of U and V. Recall that for standard normal vectors of dimension r, its norm will concentrate around $\sqrt{r}$. There exists a sufficiently large absolute constant $C > 0$ such that with probability at least $1-\delta$, the following is true.

$$\|U^\top U - I\|_{\rm op} \le C\sqrt{\frac{r+\log(2(n+m+2)/\delta)}{n}}$$
$$\|V^\top V - I\|_{\rm op} \le C\sqrt{\frac{r+\log(2(n+m+2)/\delta)}{m}}$$
$$\|U\|_{2\to\infty} \le \frac{\sqrt{r}+\sqrt{\log(2(n+m+2)/\delta)}}{\sqrt{n}}$$
$$\|V\|_{2\to\infty} \le \frac{\sqrt{r}+\sqrt{\log(2(n+m+2)/\delta)}}{\sqrt{m}}.$$

Under the above event, we conclude that

$$\|U^*\|_{2\to\infty} \lesssim \|U\|_{2\to\infty} \lesssim \sqrt{\frac{r}{n}}$$
$$\|V^*\|_{2\to\infty} \lesssim \|V\|_{2\to\infty} \lesssim \sqrt{\frac{r}{m}}.$$

Finally, the condition number of M can be upper bounded because $\sigma_{\max}(UV^\top) \le \sigma_{\max}(U)\sigma_{\max}(V) = 1 + O(\sqrt{\frac{r}{n}}) + O(\sqrt{\frac{r}{m}})$ and $\sigma_{\min}(UV^\top) \ge \sigma_{\min}(U)\sigma_{\min}(V) = 1 + O(\sqrt{\frac{r}{n}}) + O(\sqrt{\frac{r}{m}})$.

APPENDIX B
PROOFS FOR SECTION V

A. Proof of Theorem 3

To apply Theorem 2.1 in [1], we follow their proof for the symmetric matrix completion and extend it to the general case via the "symmetric dilation" trick. Let $A^* \in \mathbb{R}^{n\times n}$ be symmetric, $A = (A_{ij})$ with $A_{ij} = (A^*_{ij} + E_{ij})\Omega_{ij}/P_{ij}$ and $\bar{A} = (\bar{A}_{ij})$ with $\bar{A}_{ij} = A^*_{ij}\Omega_{ij}/P_{ij}$. We check spectral norm concentration in Lemma 5 and row norm concentration in Lemma 6. Assume $p_{\min} \ge \frac{c\log(n/\delta)}{n}$.

Lemma 5: There exists a constant $C > 0$ such that with probability at least $1-\delta$,

$$\|\bar{A} - A^*\|_{\rm op} \le C\|A^*\|_\infty\sqrt{\frac{n\log(n/\delta)}{p_{\min}}}$$
$$\|A - \bar{A}\|_{\rm op} \le C\sigma\sqrt{\frac{n\log(n/\delta)}{p_{\min}}}.$$

Proof: WLOG, assume $\|A^*\|_\infty = 1$. To prove the first inequality, we invoke the matrix Bernstein inequality [24] and obtain

$$\mathbb{P}\Big(\|\bar{A} - A^*\|_{\rm op} \ge t\Big) \le 2n\exp\Big(\frac{-p_{\min}t^2/2}{n + t/3}\Big).$$

Setting $t = \sqrt{\frac{2n\log(2n/\delta)}{p_{\min}}}$ yields the desired result.

To prove the second inequality, we define $Z_{ij} = E_{ij}\mathbb{1}_{\{|E_{ij}|\le c_1\sqrt{\log(n/\delta)}\}}$ and $\tilde{A}_{ij} = (A^*_{ij} + Z_{ij})\Omega_{ij}/P_{ij}$. By applying matrix Bernstein's inequality, we get

$$\mathbb{P}\Big(\Big\|\tilde{A} - \bar{A}\Big\|_{\rm op} \ge t\Big) \le 2n\exp\Big(\frac{-p_{\min}t^2/2}{n\sigma^2 + t\log(n/\delta)/3}\Big).$$

By setting $t = \sigma\sqrt{\frac{n\log(n/\delta)}{p_{\min}}}$, we obtain

$$\Big\|\tilde{A} - \bar{A}\Big\|_{\rm op} \lesssim \sigma\sqrt{\frac{n\log(n/\delta)}{p_{\min}}},$$

with probability at least $1-\delta$. Using the standard Gaussian tail bound, we can also show that

$$\begin{aligned}\mathbb{P}(\tilde{A} - A \ne 0) &\le \cup_{i,j}\mathbb{P}(|\varepsilon_{ij}| \ge c_1\sqrt{\log n/\delta}) \\ &\le n^2\exp\Big(-\frac{c_1\log(n/\delta)}{2}\Big) \le \delta,\end{aligned}$$

provided c_1 is sufficiently large, which completes the proof. ∎

In the next lemma, we adopt similar notations as [1] and define $\bar{\varphi}(x)$ and $\tilde{\varphi}(x)$ exactly the same. Our proof is not too different from theirs. So, we first check some inequalities in the non-uniform setting and then refer readers to their proof.

Lemma 6: For any fixed $W \in \mathbb{R}^{n\times r}$ and $k \in [n]$, we have

$$\|(\bar{A} - A^*)_{k\cdot}W\|_2 \le C\|W\|_{2\to\infty}\bar{\varphi}\Big(\frac{\|W\|_F}{\sqrt{n}\|W\|_{2\to\infty}}\Big)$$
$$\|(A - \bar{A})_{k\cdot}W\|_2 \le C\|W\|_{2\to\infty}\tilde{\varphi}\Big(\frac{\|W\|_F}{\sqrt{n}\|W\|_{2\to\infty}}\Big),$$

with probability at least $1-\delta$.

Proof: Let $X_i = (\bar{A} - A^*)_{ki}W_{i\cdot}$ for all $i \in [n]$. We derive that

$$\|X_i\|_2 \le \|A^*\|_\infty\|W\|_{2\to\infty}/P_{ki} \le \|A^*\|_\infty\|W\|_{2\to\infty}/p_{\min}$$

and

$$\mathbb{E}\|X_i\|_2^2 = \|W_{i\cdot}\|_2^2\mathbb{E}\bar{A}_{ki}^2$$

$$\leq \|A^*\|_\infty^2 \|W_{i\cdot}\|_2^2 / P_{ki}$$
$$\leq \|A^*\|_\infty^2 \|W_{i\cdot}\|_2^2 / p_{\min}.$$

Applying matrix Bernstein's inequality as in [1], we obtain the desired bound.

To prove the second inequality, we define $S_k = \{i \in [n] \colon \Omega_{ki} = 0\}$ as the observed entries in the k-th row. Then, we have

$$(A - \bar{A})_{k\cdot} W = \sigma \sum_{i \in S_k} \frac{E_{ki}}{\sigma} \cdot \frac{W_{i\cdot}}{P_{ki}}$$
$$= \frac{\sigma}{p_{\min}} \sum_{i \in S_k} \frac{E_{ki}}{\sigma} \cdot \frac{p_{\min} W_{i\cdot}}{P_{ki}}.$$

Given S_k, we note that $\{E_{ki}/\sigma\}_{i \in S_k}$ are i.i.d. $N(0,1)$. Additionally, we have

$$\sum_{i \in S_k} \left\| \frac{p_{\min} W_{i\cdot}}{P_{ki}} \right\|_2^2 \leq \sum_{i \in S_k} \|W_{i\cdot}\|_2^2.$$

Furthermore, using Chernoff's bound, we can also derive

$$\mathbb{P}(|S_k| \geq 2 \sum_{k \in [n]} P_{ki}) \leq \delta,$$

since $p_{\min} \geq \frac{c \log(n/\delta)}{n}$. The rest of the proof follows [1], where they first show the concentration of a matrix Gaussian sequence given S_k and then invoke Chernoff's inequality to control the cardinality of S_k. ∎

B. Proof of Theorem 4

Let $\mathcal{Z}$ be the set of indices of the s-th smallest probabilities. Consider a probability matrix $P' \in \mathbb{R}^{n \times m}$ that satisfies

$$P'_{ij} = \begin{cases} \frac{1}{2} & (i,j) \in \mathcal{Z} \\ P_{ij} & (i,j) \notin \mathcal{Z}. \end{cases}$$

The matrix P' agrees with P on all entries in $\mathcal{Z}^{\complement}$ and equals $\frac{1}{2}$ otherwise. WLOG, assume $p_{(nm-s)} \leq \frac{1}{2}$. (Otherwise, change $\frac{1}{2}$ to $p_{(nm-s)}$ in the above definition.) By this construction, we know $\min_{ij} P'_{ij} = p_{(nm-s)}$. Let E be the event that the following bound holds

$$\left\| \hat{A} - A^* \right\|_\infty \lesssim \eta^2 \kappa^4 (\|A^*\|_\infty + \sigma) \sqrt{\frac{(n+m)\log(1/\delta)}{p_{(nm-s)}}}.$$

Invoking Theorem 3, we know that

$$\mathbb{P}_{P'}(E^{\complement}) \leq \delta.$$

On the other hand, for each subset $\mathcal{S} \subseteq \mathcal{Z}$, let $F_{\mathcal{S}}$ denote the event that the entries in $\mathcal{S}$ are observed and those in $\mathcal{Z} \setminus \mathcal{S}$ are unobserved; note that $\mathbb{P}_{P'}(F_{\mathcal{S}}) = 2^{-s}$. By the law of total probability, we have

$$\mathbb{P}_{P'}(E^{\complement}) = \sum_{\mathcal{S} \subseteq \mathcal{Z}} \mathbb{P}_{P'}(E^{\complement} \mid F_{\mathcal{S}}) \cdot \mathbb{P}_{P'}(F_{\mathcal{S}})$$
$$= \sum_{\mathcal{S} \subseteq \mathcal{Z}} \mathbb{P}_{P}(E^{\complement} \mid F_{\mathcal{S}}) \cdot 2^{-s},$$

where the last step follows from the fact that $\mathbb{P}_{P'}(\cdot | F_{\mathcal{S}}) = \mathbb{P}_P(\cdot | F_{\mathcal{S}}), \forall \mathcal{S}$ since P and P' are identical on entries outside $\mathcal{Z}$ and the observations are independent across entries. Combining the last two display equations, and lower bounding the sum by one summand, we have

$$\mathbb{P}_P(E^{\complement} | F_{\mathcal{S}}) \leq 2^s \delta, \quad \forall \mathcal{S} \subseteq \mathcal{Z}$$

and hence $\mathbb{P}_P(E^{\complement}) \leq 2^s \delta$. Letting $\delta' = 2^s \delta$, we conclude that under P, with probability at least $1 - \delta'$, it holds that

$$\left\| \hat{A} - A^* \right\|_\infty \lesssim \eta^2 \kappa^4 (\|A^*\|_\infty + \sigma) \sqrt{\frac{(n+m)\log(2^s/\delta')}{p_{(nm-s)}}}.$$

Appendix C
Proofs for Section IV

A. Proof of Theorem 1

We first invoke Lemma 2 and Lemma 3 to obtain M^* is both rank-r and bounded, with probability at least $1 - \delta/3$. For entry (i,j), we have a corresponding submatrix $W = M_{[k^*(i,j)] \cup \{i\}, [k^*(i,j)] \cup \{j\}}$ of size $k^* \times k^*$, $(k^*+1) \times k^*$, or $k^* \times (k^*+1)$. We apply Lemma 4 to W and get the incoherence and condition number guarantee with probability at least $1 - \delta/3$:

$$\kappa(W) \leq \frac{1 + C\sqrt{\frac{r + \log(6(n+m+2)/\delta)}{k^*}}}{1 - C\sqrt{\frac{r + \log(6(n+m+2)/\delta)}{k^*}}}.$$

and

$$\left\| U^W \right\|_{2 \to \infty} \leq C \left(\sqrt{\frac{r}{k^*}} + \sqrt{\frac{\log(6(n+m+2)/\delta)}{k^*}} \right)$$
$$\left\| V^W \right\|_{2 \to \infty} \leq C \left(\sqrt{\frac{r}{k^*}} + \sqrt{\frac{\log(6(n+m+2)/\delta)}{k^*}} \right).$$

We consider three separate cases. If $k^* \geq i$ and $k^* \geq j$, it means the smallest probability in $P_{[k^*] \cup \{i\}, [k^*] \cup \{j\}}$ is P_{k^*,k^*}, which is precisely $p^*(i,j)$. In this case, we invoke Theorem 3 and the union bound to get

$$\left| \hat{M}_{ij} - M^*_{ij} \right| \lesssim r(r+\sigma) \sqrt{\frac{\log^5(n/\delta)}{k^*(i,j) p^*(i,j)}},$$

with probability at least $1 - \delta$. If $k^* < i$ and $k^* < j$, then the smallest probability in $P_{[k^*] \cup \{i\}, [k^*] \cup \{j\}}$ becomes $P_{k^* k^*}$ and the second smallest is $\min\{P_{ik^*}, P_{k^* j}\} = p^*(i,j)$. We apply Theorem 4 and the union bound to get

$$\left| \hat{M}_{ij} - M^*_{ij} \right| \lesssim r(r+\sigma) \sqrt{\frac{\log^5(n/\delta)}{k^*(i,j) p^*(i,j)}},$$

with probability at least $1 - \delta$. In the complementary case, we assume $i > k^*$ and $j \leq k^*$ without loss of generality. The smallest probability in $P_{[k^*] \cup \{i\}, [k^*] \cup \{j\}}$ is $P_{ik^*} = p^*(i,j)$. Apply Theorem 3 and the same conclusion follows.

B. Proof of Theorem 2

Recall the standard linear regression problems: $y = X\theta^* + \varepsilon$, with known covariates $X \in \mathbb{R}^{d\times r}$, unknown parameter $\theta^* \in \mathbb{R}^r$ and i.i.d. Gaussian noise $\varepsilon_i \sim N(0, \sigma^2)$. The minimax risk of estimating θ^* is

$$\inf_{\hat{\theta}} \sup_{\theta^*} \mathbb{E}\left[\left\|\hat{\theta} - \theta^*\right\|_2^2\right] \geq \frac{Cr\sigma^2}{d}\left\|\frac{1}{\sqrt{n}}X\right\|_{\text{op}}. \tag{15}$$

The proof can be found in [26], which is an application of Fano's method.

Fix $i \in [n], j \in [m]$. Assume $\{a_k^*\}_{k\in[n]}$ is known and estimate b_j^* using $\{Y_{kj}, k \in [n]\}$. Applying (15) yields:

$$\inf_{\hat{b}_j} \sup_{b_j^*\in\mathbb{R}^r} \mathbb{E}[(\hat{b}_j - b_j^*)^2] \geq \frac{Cr\sigma^2}{\sum_{i'} P_{i'j}}, \tag{16}$$

with probability at least $\frac{1}{2}$. The case where the column latent factors are known is symmetric. We prove (16) and the desired result (6) follows. Let $X_k = \mathbb{1}_{\{(k,j) \text{ is observed}\}}$ for all $k \in [n]$. We have $\mathbb{E}[X_k] = P_{kj}$. Applying Markov's inequality, we get

$$\mathbb{P}\left(\sum_k X_k \geq 2\sum_k P_{kj}\right) \leq \frac{1}{2}.$$

Let $\mathcal{K} = \{k \in [n] : X_k = 1\}$ be the set of observed entries. Applying (15) with $y = Y_{\mathcal{K}j}$, $X = (a_k^*)_{k\in\mathcal{K}}$, $\theta^* = b_j^*$ and $d = \sum_k X_k = |\mathcal{K}|$, we obtain

$$\inf_{\hat{b}_j} \sup_{b_j^*} \mathbb{E}[(\hat{b}_j - b_j^*)^2] \geq \frac{Cr\sigma^2}{\sum_k X_k} \geq \frac{Cr\sigma^2}{2\sum_k P_{kj}},$$

with probability at least $\frac{1}{2}$ over the randomness of $\mathcal{K}$ and $\{a_k^*\}$.