Rate-Optimal Subspace Estimation on Random Graphs

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the theory of random bipartite graph whose adjacency matrix is generated according to a connectivity matrix $\mathbf{M}$. We consider the bipartite graph to be sparse, i.e., the entries of $\mathbf{M}$ are upper bounded by certain sparsity parameter. We show that the performance of estimating the connectivity matrix $\mathbf{M}$ depends on the sparsity of the graph. We focus on two measurement of performance of estimation: the error of estimating $\mathbf{M}$ and the error of estimating the column space of $\mathbf{M}$. In the first case, we consider the operator norm and Frobenius norm of the difference between the estimation and the true connectivity matrix. In the second case, the performance will be measured by the difference between the estimated projection matrix and the true projection matrix in operator norm and Frobenius norm. We will show that the estimators we propose achieve the minimax optimal rate.

1 Introduction

There have been many fruitful results on subspace estimation principal component analysis. Researchers have been considering random matrix with Gaussian and sub-Gaussian noise. In the present work, we focus on random *bipartite graph*. A bipartite graph is a graph whose vertices can be divided into two disjoint sets ,$\mathcal{U}$ and $\mathcal{V}$, such that every edge connects a vertex in $\mathcal{U}$ to one in $\mathcal{V}$. In a directed random bipartite graph, an edge connecting node $i \in \mathcal{U}$ to node $j \in \mathcal{V}$ exists with probability $\mathbf{M}_{ij}$ independently. Suppose $|\mathcal{U}| = n_1$ and $|\mathcal{V}| = n_2$, a bipartite graph can be represented by an *adjacency matrix* $\mathbf{A}$, whose elements $\mathbf{A}_{ij}$, $(i, j) \in [n_1] \times [n_2]$ are independent Bernoulli random variables with latent parameters $\mathbf{M}_{ij}$. In other words, we can focus on the random adjacency matrix $\mathbf{A}$ with entries

$$\mathbf{A}_{ij} \sim \text{Bernoulli}(\mathbf{M}_{ij}) \quad \text{for } (i, j) \in [n_1] \times [n_2] \text{ independently.} \tag{1}$$

$\mathbf{M}$ elements $\mathbf{M}_{ij}$ is called the *connectivity matrix*. The goal of this paper is to show the optimality of a regularized singular value decomposition for the latent connectivity matrix $\mathbf{M}$ and its singular space, under the assumption that $\mathbf{M}$ follows certain low-rank structure. Formally speaking, for the task of estimating $\mathbf{M}$, we aim to the matrix estimation task, we consider the following parameter space:

$$\mathbf{\Theta}_1(n_1, n_2, p, r) = \{\mathbf{M} \in [0, p]^{n_1 \times n_2} : \text{rank}(\mathbf{M}) \leq r\} \quad \text{for positive integers } n_1, n_2, r. \tag{2}$$

In other words, $\mathbf{\Theta}_1(n_1, n_2, p, r)$ is consists of all $n_1 \times n_2$ matrices with entries between 0 and p and rank at most r. To avoid empty parameter space, we always assume $r \leq \min(n_1, n_2)$. The parameter p controls the sparisty of the bipartite graph.

Matrix estimation has been studied in many settings. In Chatterjee et al. [2015], the author assumes the random matrix has bounded entries. The universal singular value thresholding (USVT) algorithm is proved to be optimal in the parameter space with bounded nuclear norm. In Koltchinskii et al. [2011], the authors considers the parameter space with subexponential random matrices with bounded

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

means. However, their approach is not optimal under our random graph assumption. One-bit matrix completion is considered in Cai and Zhou [2013], Davenport et al. [2014], but parameters therein are not necessarily of low rank matrices, so the minimax error is not comparable with that of the family (2) in the present paper. In network analysis problems, the density of networks are expected to be sparse. The upper bound of connection probability p in our model (2) plays a key role in error rates of the estimation.

When we consider the singular space (SS) estimation problem, an extra assumption about the smallest non-zero singular value is required. We define the corresponding parameter space as follows:

$$\mathbf{\Theta}_2(n_1, n_2, p, r, \sigma_*) = \{\mathbf{M} \in [0, p]^{n_1 \times n_2} : \text{rank}(\mathbf{M}) = r, \sigma_r(\mathbf{M}) \geq \sigma_*\} \tag{3}$$

where $\sigma_r(\mathbf{M})$ is the rth largest singular value of $\mathbf{M}$. It is clear that $\mathbf{\Theta}_2(n_1, n_2, p, r, \sigma_*) \subset \mathbf{\Theta}_1(n_1, n_2, p, r)$. In other words, in the singular space estimation problem, we consider parameters in $\mathbf{\Theta}_1(n_1, n_2, p, r)$ whose smallest nonzero singular value is bounded below by σ_*.

Singular space estimation in the Gaussian case has been considered in Cai et al. [2018]. To show the optimality of their estimation, they use packing number of Grassmanian manifold provided in Szarek [1982]. In the bipartite graph model that we consider in this paper, the connectivity matrix $\mathbf{M}$ in (3) is a merely proper subset of the Grassmanian manifold because the entries of $\mathbf{M}$ are contained in the interval $[0, p]$. As a consequence, the previous approach fails to calculate the lower bound of the packing number of the parameter space $\mathbf{\Theta}_2(n_1, n_2, p, r, \sigma_*)$. In the literature of community detection, spectral methods have been applied in clustering tasks on random graphs Lei et al. [2015], Chin et al. [2015], Gao et al. [2018], Zhou and Amini [2019], but these works have not focused on the optimality of singular space estimation. With the help of recent development in non-asymptotic analysis of random matrices, we propose a stochastic approach to compute the minimax lower bound of subspace estimation over the parameter space (3).

Here are some notations we will use in this paper. For real numbers a and b, $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. $\sigma_r(\mathbf{M})$ is the rth largest singular value of matrix $\mathbf{M}$. $\|\mathbf{M}\|_{\text{op}} = \sigma_1(\mathbf{M})$, i.e., the largest singular value, is the spectral norm of $\mathbf{M}$. $\|\mathbf{M}\|_{\text{F}} = \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbf{M}_{ij}^2\right)^{1/2}$ is the Frobenius norm of $\mathbf{M}$. We note that $\|\mathbf{M}\|_{\text{F}}^2 = \sum_{i=1}^{n_1 \wedge n_2} \sigma_i(\mathbf{M})^2$. $\text{Col}(\mathbf{M})$ is the column space of $\mathbf{M}$. Suppose $\mathbf{a} = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$, then $\text{diag}(\mathbf{a})$ is an $n \times n$ diagonal matrix with entries $a_1, a_2, \ldots, a_n$ on the diagonal. For integer n, $[n] := \{1, 2, \ldots, n\}$. For positive sequence a_n and b_n, if there exists a constant C such that $a_n \leq C b_n$ for all n, then we write $a_n = O(b_n)$ or $a_n \lesssim b_n$. If $a_n = O(b_n)$ and $b_n = O(a_n)$, then we write $a_n = \Theta(b_n)$. If $a_n / b_n \to 0$ as $n \to \infty$, then we write $a_n = o(b_n)$ or $a_n \ll b_n$. When we discuss asymptotic properties, we consider $\min(n_1, n_2) \to \infty$, i.e., the size the bipartite graph tends to infinity.

The rest of this paper will be organized as follows. In Section 2, we introduce the algorithm for matrix estimation problem, and provide theoretical analysis for the error rate and minimax lower bound of this problem. Applying the result of matrix estimation, we propose the algorithm for column and row space estimation in Section 3. Then we show that the algorithm outputs estimator with minimax optimal error rate. Numerical experiments on the algorithms appear in Section 4. The proofs of the theorems can be found in appendix in the supplementary material.

2 Low Rank Matrix Estimation of Random Graph

Given a random bipartite graph, we aim to estimate true connecting probability of the adjacency matrix assuming that the connectivity matrix has low rank. To be more specific, we consider the task of low rank matrix estimation parameters belong to the space

$$\mathbf{\Theta}_1(n_1, n_2, p, r) = \{\mathbf{M} \in [0, p]^{n_1 \times n_2} : \text{rank}(\mathbf{M}) \leq r\}.$$

Without loss of generality, we assume $n_1 \geq n_2$ in this section. We recall that a random bipartite graph with adjacency matrix $\mathbf{A}$ is observed, and $\mathbf{A}_{ij} \sim \text{Bernoulli}(\mathbf{M}_{ij})$ independently for $(i, j) \in [n_1] \times [n_2]$. We will first proposed the algorithm to estimate $\mathbf{M}$, then we will show the optimality of this estimator in the minimax sense.

2.1 Hard Singular Value Thresholding

We first introduce the algorithm to compute the estimator $\hat{\mathbf{M}}$. The hard thresholding procedure is similar as universal singular value thresholding proposed in Chatterjee et al. [2015]. However, our

algorithm is specifically designed for random graph, which leads to tighter error rate. The algorithm first applies degree truncation on bipartite adjacency matrix $\mathbf{A}$, then compute the best low rank approximation on the resulting matrix by singular value decomposition (SVD). For $n_1 \times n_2$ matrix $\mathbf{X}$, and recall that $n_1 \geq n_2$, the SVD of $\mathbf{X}$ is defined by

$$\mathbf{X} = \sum_{i=1}^{n_2} \sigma_i \mathbf{U}_i \mathbf{V}_i^\top = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$$

where

$$\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_{n_2}), \quad \mathbf{\Lambda} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{n_2}), \quad \text{and } \mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_{n_2})$$

with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{n_2}$ of $\mathbf{X}$. Now we are ready to present the algorithm for estimating $\mathbf{M}$ by hard singular value thresholding.

Algorithm 1 Hard Singular Value Thresholding

1: **Input:** adjacency matrix $\mathbf{A}$, sparsity parameter p, rank r.
2: **Output:** the estimate of $\mathbf{M}$, $\hat{\mathbf{M}}$.
3: Set $r' = r \wedge \lfloor n_2 p \rfloor$.
4: If the degree of any row or column of $\mathbf{A}$ is greater than $2n_1 p$, replace a subset of 1's in such rows and columns by 0 so that the column-degrees and the row degrees of the resulting matrix are bounded by $2n_1 p$. We call this regularized matrix $\mathbf{A}_{\text{re}}$.
5: Let $\hat{\mathbf{M}}$ be the best rank-r' approximation of $\mathbf{A}_{\text{re}}$ composed of the first r' SVD components with the largest singular values. (More details of this step will be discussed later.)

Remark 1. *We note that the regularization in Step 4 is necessary to achieve the optimal error rate. It is worth noting that the regularization method presented in the algorithm is not the unique one. Besides replacing the entries by zeroes, we can also rescale the columns and rows by a small constant so that the degree is bounded by $2n_1 p$. In this way, the regularized adjacency matrix loses less information, so the estimation is more accurate. The threshold of regularization can be replaced by $cn_1 p$ rather than $2n_1 p$. In most cases, the best c can be much smaller than 2, though the theoretical estimation error rate cannot be improved by choosing the optimal c. We will discuss the choice of c by simulation in Section 4.*

In step 5, suppose $\mathbf{A}_{\text{re}}$ has SVD

$$\mathbf{A}_{\text{re}} = \sum_{i=1}^{n_2} \sigma_i \mathbf{U}_i \mathbf{V}_i^\top,$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{n_2}$, then we define $\hat{\mathbf{M}}$ to be the best rank-r' approximation $\mathbf{A}_{\text{re}}$:

$$\hat{\mathbf{M}} = \sum_{i=1}^{r'} \sigma_i \mathbf{U}_i \mathbf{V}_i^\top.$$

The algorithm has difficulty to provide good estimation if $\mathbf{M}$ is not sufficiently low-rank. To judge if the algorithm benefits from the low rank-structure, we check if the rank is smaller than $n_2 p$ in step 3. Suppose $r' = \lfloor n_2 p \rfloor$, then the output of the algorithm is the same as random guessing. Suppose $\mathbf{M}$ has low rank, then graph regularization in step 4 followed by hard singular value thresholding in step 5 can return better estimation. Now we are ready to present the high probability bounds for estimation of $\mathbf{M}$. The error will be measured by both operator norm $\|\hat{\mathbf{M}} - \mathbf{M}\|_{\text{op}}$ and $\|\hat{\mathbf{M}} - \mathbf{M}\|_{\text{F}}$. The upper bounds of the error rate are shown in the following theorem.

Theorem 1. *Given the adjacency matrix $\mathbf{A}$ generated by parameter $\mathbf{M}$ from space $\mathbf{\Theta}_1(n_1, n_2, p, r)$ and assuming $n_1 \geq n_2$, then the estimator $\hat{\mathbf{M}}$ obtained from Algorithm 1 satisfies and*

$$\sup_{\mathbf{M} \in \mathbf{\Theta}_1(n_1, n_2, p, r)} \mathbb{P}\big(\|\hat{\mathbf{M}} - \mathbf{M}\|_{op}^2 \lesssim (n_1 p) \wedge (n_1 n_2 p^2)\big) \geq 1 - n_1^{-1}, \tag{4}$$

and

$$\sup_{\mathbf{M} \in \mathbf{\Theta}_1(n_1, n_2, p, r)} \mathbb{P}\big(\|\hat{\mathbf{M}} - \mathbf{M}\|_{F}^2 \lesssim (n_1 p r) \wedge (n_1 n_2 p^2)\big) \geq 1 - n_1^{-1}. \tag{5}$$

This theorem shows that if we consider the average loss of estimating entries of $\mathbf{M}$,

$$\frac{1}{n_1 n_2}\|\hat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}}^2 \lesssim \frac{pr}{n_2} \wedge p^2$$

with high probability given the dimension of the matrix tends to infinity. As long as $pr = o(n_2)$, Algorithm 1 provides consistent estimation of the connectivity matrix $\mathbf{M}$.

Remark 2. *Without the Step 4 in Algorithm 1,* (4) *can still achieve the error rate* $[(n_1 p) \vee (\log n_1)] \wedge (n_1 n_2 p^2)$ *with high probability. One can check the technical reason from the concentration of random graph without regularization in Feige and Ofek [2005]. However, this error rate cannot achieve the minimax lower bound in Theorem 3.*

2.2 Soft Singular Value Thresholding

Soft singular value thresholding introduced in Cai et al. [2010] can also be applied in estimating the connectivity matrix $\mathbf{M}$. In hard thresholding, the algorithm truncates all the smallest $n - r'$ singular value to 0, where r' is defined in Step 3 of Algorithm 1. In the soft thresholding approach, the algorithm shrink all singular values towards zero.

Algorithm 2 Soft Singular Value Thresholding

1: **Input:** adjacency matrix $\mathbf{A}$, sparsity parameter p, rank r.
2: **Output:** the estimate of $\mathbf{M}$, $\hat{\mathbf{M}}$.
3: Set $r' = r \wedge \lfloor n_2 p \rfloor$.
4: Apply regularization (step 4 of Algorithm 1) on $\mathbf{A}$ to obtain $\mathbf{A}_{\text{re}}$.
5: Let $\mathbf{A}_{\text{re}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ be the SVD of $\mathbf{A}_{\text{re}}$ where $\mathbf{\Lambda} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{n_2})$.
6: Let $\hat{\mathbf{\Lambda}} = \text{diag}((\sigma_1 - \sigma_{r'+1}) \vee 0, (\sigma_2 - \sigma_{r'+1}) \vee 0, \ldots, (\sigma_{n_1 \wedge n_2} - \sigma_{r'+1}) \vee 0)$.
7: Let $\hat{\mathbf{M}} = \mathbf{U}\hat{\mathbf{\Lambda}}\mathbf{V}^\top$.

Theorem 2. *Under the same assumptions in Theorem 1, the estimator* $\hat{\mathbf{M}}$ *obtained from Algorithm 2 satisfies* (4) *and* (5)*.*

Theorem 1 and Theorem 2 show that both hard and soft singular values thresholding methods provides consistent estimation of $\mathbf{M}$ up to a constant factor. Essentially, the estimator $\hat{\mathbf{M}}$ in each algorithm is a low rank matrix with $\|\mathbf{M} - \hat{\mathbf{M}}\|_{\text{op}}^2 \lesssim (n_1 p) \wedge (n_1 n_2 p^2)$ with high probability.

This upper bound has also been studied in Gao et al. [2018], Zhou and Amini [2019]. The authors applied their result for community detection when the number of communities are sufficiently smaller than the average degree of the graph. However, the term $n_1 n_2 p^2$ is not included in previous works.

2.3 Minimax Lower Bound

Theorem 1 and Theorem 2 provide a high-probability upper bound for the estimation in $\mathbf{M}$. We now establish lower bounds for the estimation error measured by operator norm and Frobenius distance. We will show that the high probability bounds in (4) and (5) are minimax optimal in the next theorem.

Theorem 3. *Suppose* $n_1 \geq n_2$ *and* n_1 *is sufficiently large, then*

$$\inf_{\hat{\mathbf{M}}} \sup_{\mathbf{M} \in \mathbf{\Theta}_1(n_1, n_2, p, r)} \mathbb{P}\Big(\|\hat{\mathbf{M}} - \mathbf{M}\|_{op}^2 > \frac{(n_1 p) \wedge (n_1 n_2 p^2)}{512}\Big) \geq \frac{1}{4} \tag{6}$$

and

$$\inf_{\hat{\mathbf{M}}} \sup_{\mathbf{M} \in \mathbf{\Theta}_1(n_1, n_2, p, r)} \mathbb{P}\Big(\|\hat{\mathbf{M}} - \mathbf{M}\|_{F}^2 > \frac{(n_1 pr) \wedge (n_1 n_2 p^2)}{512}\Big) \geq \frac{1}{4}. \tag{7}$$

Here, the infimum is over all estimators depending on $\mathbf{A}$, p and r. (6) shows that there is no estimator can achieve high probability bound better than $\Theta((n_1 p) \wedge (n_1 n_2 p^2))$ when we measure the estimation error by operator norm. Similarly, (7) shows the optimality for Frobenius norm.

These lower bounds are comparable with [Koltchinskii et al., 2011, Theorem 5]. Although in their paper, the authors consider Gaussian noise rather than Bernoulli random variables, the construction

of packing number in the proofs are similar. Nevertheless, the upper bound and lower bound in their paper does not match if the entrywise standard deviation and $\max_{i,j}|M_{ij}|$ are not asymptotically equivalent. In the situation of random graph, we assume that $\max_{i,j}|M_{ij}| = p$, then $p \ll \sqrt{p(1-p)}$ if the random graph is sparse, i.e., $p \ll 1$. In this case, the minimax lower bound in Koltchinskii et al. [2011] is suboptimal. Theorem 3 resolves this issue and show that Algorithm 1 and Algorithm 2 return optimal estimator in the random bipartite graph problem.

3 Singular Space Estimation

In this section, we will study the estimation of column and row subspaces of parameter $\mathbf{M}$. This task is closely related to the estimation of $\mathbf{M}$ which has been discussed in the previous section. Compared with the parameter space $\mathbf{\Theta}_1(n_1, n_2, p, r)$, it additionally requires $\sigma_r(\mathbf{M})$ to be bounded away from zero, where $\sigma_r(\mathbf{M})$ is the rth largest singular value of $\mathbf{M}$. This is a necessary condition to distinguish the subspace associated with top singular values from its normal space. We recall the parameter space from (3):

$$\mathbf{\Theta}_2(n_1, n_2, p, r, \sigma_*) = \{\mathbf{M} \in [0, p]^{n_1 \times n_2} : \text{rank}(\mathbf{M}) = r, \sigma_r(\mathbf{M}) \geq \sigma_*\},$$

For $\mathbf{M} \in \mathbf{\Theta}_2(n_1, n_2, p, r, \sigma_*)$, $\mathbf{M}^\top\mathbf{M} \in [0, n_2p^2]^{n_1 \times n_1}$. Therefore,

$$n_1 n_2 p^2 \geq \text{tr}(\mathbf{M}^\top\mathbf{M}) = \sum_{i=1}^{n_1 \wedge n_2} \sigma_i^2(\mathbf{M}) \geq \sigma_*^2 r. \tag{8}$$

Hence, $\sigma_* \leq \sqrt{\frac{n_1 n_2}{r}}p$ is a necessary condition for the parameter space $\mathbf{\Theta}_2(n_1, n_2, p, r, \sigma_*)$ being nonempty. Recall that we assume the observed bipartite adjacency matrix follows the distribution

$$\mathbf{A}_{ij} \sim \text{Bernoulli}(\mathbf{M}_{ij}) \quad \text{for } (i, j) \in [n_1] \times [n_2] \text{ independently.} \tag{9}$$

We aim to recover the top r left and right singular vectors. Due to the symmetry of the column and row space, we will focus on left singular vectors in the following discussion. Let

$$\mathbf{M} = \sum_{i=1}^{r} \sigma_i(\mathbf{M})\mathbf{U}_i\mathbf{V}_i^\top$$

be the singular value decomposition of $\mathbf{M}$. Our goal is to recover the subspace spanned by $\{\mathbf{U}_1, \ldots, \mathbf{U}_r\}$. Let $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_r)$. For an estimator $\hat{\mathbf{U}}$ of $\mathbf{U}$, we measure the loss of the estimation by the operator norm and Frobenius norm between the projection matrices of $\hat{\mathbf{U}}$ and $\mathbf{U}$:

$$L_{\text{op}}(\hat{\mathbf{U}}, \mathbf{U}) = \|\hat{\mathbf{U}}(\hat{\mathbf{U}}^\top\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}^\top - \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\|_{\text{op}}^2.$$

and

$$L_{\text{F}}(\hat{\mathbf{U}}, \mathbf{U}) = \|\hat{\mathbf{U}}(\hat{\mathbf{U}}^\top\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}^\top - \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\|_{\text{F}}^2.$$

We note that suppose $\mathbf{U}$ and $\hat{\mathbf{U}}$ have orthonormal columns, i.e., $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_r$ and $\hat{\mathbf{U}}^\top\hat{\mathbf{U}} = \mathbf{I}_r$, then the loss function can be reduced to

$$L_{\text{op}}(\hat{\mathbf{U}}, \mathbf{U}) = \|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_{\text{op}}^2 \tag{10}$$

and

$$L_{\text{F}}(\hat{\mathbf{U}}, \mathbf{U}) = \|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_{\text{F}}^2 \tag{11}$$

We can similarly define the loss of row space estimator by $L_{\text{op}}(\hat{\mathbf{V}}, \mathbf{V})$ and $L_{\text{F}}(\hat{\mathbf{V}}, \mathbf{V})$.

Remark 3. *Due to the identifiability issue, we cannot directly compare the true basis of the column space of* $\mathbf{M}$ *with the estimator* $\hat{\mathbf{U}}$*. In this paper, we use the difference between the projection matrices corresponding to different subspaces to measure the estimation error. There are other norms using in literature, for instance, the* $\sin\Theta$ *distance*

$$\|\sin\Theta(\hat{\mathbf{U}}, \mathbf{U})\| = \|\mathbf{U}^\top\hat{\mathbf{U}}_\perp\|$$

where $\hat{\mathbf{U}}_\perp$ *has orthonormal columns which spans the normal space of Col*$(\hat{\mathbf{U}})$*. Another possible measurement of distance between* $\mathbf{U}$ *and* $\hat{\mathbf{U}}$ *is*

$$\inf_{\mathbf{O} \in \mathbb{O}_r} \|\mathbf{U} - \hat{\mathbf{U}}\mathbf{O}\|.$$

Here, the norm $\|\cdot\|$ *can be either the operator norm or the spectral norm, and* $\mathbb{O}_r$ *represents the set of* $r \times r$ *orthogonal matrix. It can be shown that these two norms are equivalent to the projection loss defined in* (10) *and* (11) *[Cai et al., 2018].*

3.1 Singular Subspace Estimation and Upper Bound

We consider the rank truncation algorithm. In the Algorithm 3, we first obtain the estimation $\hat{\mathbf{M}}$ of the connectivity matrix $\mathbf{M}$ by Algorithm 1 or Algorithm 2, then we use the top singular vectors as of $\hat{\mathbf{M}}$ as estimation of the ones of $\mathbf{M}$.

Algorithm 3 Singular space estimation

1: **Input:** adjacency matrix $\mathbf{A}$, sparsity parameter p, rank r.
2: **Output:** the estimate of the column space and row space of $\mathbf{M}$.
3: Let $\hat{\mathbf{M}}$ be the output of Algorithm 1 or Algorithm 2.
4: Let $\hat{\mathbf{U}} = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$ where $\mathbf{u}_i$'s are the left singular vectors of $\hat{\mathbf{M}}$ corresponding to top-r singular values.
5: Let $\hat{\mathbf{V}} = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$ where $\mathbf{v}_i$'s are the right singular vectors of $\hat{\mathbf{M}}$ corresponding to top-r singular values.
6: Output $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$.

We combine the upper bound in Theorem 1 with Davis-Kahan theorem, then obtain the upper bound of loss in subspace estimation given by Algorithm 3.

Theorem 4. *Given the adjacency matrix $\mathbf{A}$ generated by connectivity matrix $\mathbf{M}$ from parameter space $\mathbf{\Theta}_2(n_1, n_2, p, r, \sigma_*)$, the output of $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ from Algorithm 3 satisfies*

$$\sup_{\mathbf{M}\in\mathbf{\Theta}_2(n_1,n_2,p,r,\sigma_*)} \mathbb{P}\Big(\max(L_{op}(\hat{\mathbf{U}}, \mathbf{U}), L_{op}(\hat{\mathbf{V}}, \mathbf{V})) \lesssim \frac{(n_1+n_2)p}{\sigma_*^2} \wedge 1\Big) \geq 1 - (n_1 \vee n_2)^{-1},$$

and

$$\sup_{\mathbf{M}\in\mathbf{\Theta}_2(n_1,n_2,p,r,\sigma_*)} \mathbb{P}\Big(\max(L_F(\hat{\mathbf{U}}, \mathbf{U}), L_F(\hat{\mathbf{V}}, \mathbf{V})) \lesssim \frac{(n_1+n_2)pr}{\sigma_*^2} \wedge r\Big) \geq 1 - (n_1 \vee n_2)^{-1}.$$

3.2 Minimax Lower Bound

In this section, we introduce the lower bound for the estimation in Theorem 4. Let us denote the projection matrix of the left singular space of $\mathbf{M}$ by

$$\mathbf{P}_{\mathbf{M}} = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top.$$

In the next theorem, we will show that under regularity conditions on the parameter space, the estimation error rate of $\mathbf{P}_{\mathbf{M}}$ in the previous theorem is optimal.

Theorem 5. *Let $1 \leq r < n_2 \leq n_1$. Suppose that for some positive constant C_0, $r \leq n_1/C_0$, $r\sigma_*^2 \leq n_1 n_2 p^2/C_0$ for $r > 1$ and $n_1 n_2 (4p/10)^2 \leq r\sigma_*^2 \leq n_1 n_2 (6p/10)^2$ for $r = 1$. Then, for sufficiently large C_0 there exist a prior π^* such that*

$$\mathbb{P}_{\pi^*}\left\{\mathbf{M} \in [0.3p, 0.7p]^{n_1\times n_2}, \sigma_r(\mathbf{M}) \geq \sigma_*, \textit{rank}(\mathbf{M}) = r\right\} = 1 \tag{12}$$

and the risk of the Bayes estimator $\hat{\mathbf{P}}^* = \mathbb{E}_{\pi^*}\left[\mathbf{P}_{\mathbf{M}}\middle|\mathbf{A}\right]$ under $\mathbb{P}_{\pi^*}$ is bounded from below by

$$\mathbb{E}_{\pi^*}\left[\|\hat{\mathbf{P}}^* - \mathbf{P}_{\mathbf{M}}\|_{op}^2\right] \geq \frac{1}{50}\left(\frac{n_1 p}{\sigma_*^2} \wedge 1\right) \tag{13}$$

and

$$\mathbb{E}_{\pi^*}\left[\|\hat{\mathbf{P}}^* - \mathbf{P}_{\mathbf{M}}\|_F^2\right] \geq \frac{r}{50}\left(\frac{n_1 p}{\sigma_*^2} \wedge 1\right). \tag{14}$$

Remark 4. *The assumptions on the parameters are mild and necessary (up to the choice of constants). r/n_1 has to be sufficiently small so that the parameter $\mathbf{M}$ is low-rank. In* (8)*, we have shown that $n_1 n_2 p^2/(r\sigma_*^2)$ is sufficiently large so that the parameter space is nonempty. It requires the parameter space to be large enough to construct the prior π_*. We need slightly different regularity conditions for $r = 1$ and $r > 1$ because of the different construction of π_* in each case.*

Theorem 5 presents the lower bound of the Bayes error rate. By the fact that a minimax estimator should be a Bayes estimator with respect to a least favorable prior of the parameter space, it is straightforward to obtain

$$\inf_{\hat{\mathbf{P}}} \sup_{\mathbf{M}\in\mathbf{\Theta}_2(n_1,n_2,p,r,\sigma_*)} \mathbb{E}\Big[\|\hat{\mathbf{P}}-\mathbf{P}_{\mathbf{M}}\|_{\text{op}}^2\Big] \geq \frac{1}{50}\left(\frac{n_1 p}{\sigma_*^2}\wedge 1\right)$$

and

$$\inf_{\hat{\mathbf{P}}} \sup_{\mathbf{M}\in\mathbf{\Theta}_2(n_1,n_2,p,r,\sigma_*)} \mathbb{E}\Big[\|\hat{\mathbf{P}}-\mathbf{P}_{\mathbf{M}}\|_{\text{F}}^2\Big] \geq \frac{r}{50}\left(\frac{n_1 p}{\sigma_*^2}\wedge 1\right).$$

from (13) and (14). We have been focus on the lower bound of the left subspace estimation in the discussion so far. The result of Theorem 5 can be extend to the right subspace estimation. We skip the presentation of this similar result.

4 Numerical Experiments

In previous sections, we have shown that the algorithms returns rate optimal estimation in different tasks. The choice of hyperparameter does not affect the error rate in the theorems, but it can change the performance of the algorithms in practice. In this section, each experiment will repeat 100 times. In each iteration, the randomization procedure follows these steps:

1. Randomly generate matrices $\mathbf{M}_1 \in \mathbb{R}^{n_1\times(r-1)}$ $\mathbf{M}_2 \in \mathbb{R}^{(r-1)\times n_2}$ so that each entries follows normal distribution independently.
2. $\mathbf{M}_3 = \mathbf{M}_1\mathbf{M}_2$ is an $n_1 \times n_2$ matrix with rank $r-1$.
3. Rescale $\mathbf{M}_3$ so that all entries belongs to the interval $[0, p]$. We note that this rescaling step increases the rank of the matrix by 1. The resulting matrix is the connectivity matrix $\mathbf{M}$ in the current iteration.
4. Generate the adjacency matrix of the random bipartite graph with parameter $\mathbf{M}$.

4.1 Simulation on Matrix Estimation

Algorithm 1 and Algorithm 2 have the same theoretical error rate. We consider the following parameters in $\mathbf{\Theta}_1(n_1, n_2, r, p)$. $n_1 = n_2 = 1000, r = 3, p = 0.01, 0.03, 0.05$. In the following experiments, we varies the regularization constant from 0.2 to 1, where the default constant equals to 2 in the algorithms presented in the previous section. When we apply regularization on the adjacency matrix, we rescale the columns and the rows corresponding to nodes with large degrees in the bipartite graph (as described in Remark 1).

From the simulation results in Figure 1, we can conclude following observations:

1. Compared with soft singular value thresholding in Algorithm 2, the performance of hard singular value thresholding in Algorithm 1 more heavily relies on the choice of regularization parameter. If we are able to choose a good regularization parameter, Algorithm 1 outperforms Algorithm 2 when the error is measured by operator norm. On the other hand, if we consider the Frobenius loss, Algorithm 2 performs relatively better. Suppose we are unable to choose a good regularization constant, Algorithm 2 performs better in most cases.
2. Since the density parameter $p = 0.01, 0.03$ and 0.05 on different columns from left to right, we compare the plots in the first row and observe that regularization has smaller effect as the density of the random bipartite graph increases. This phenomenon indeed coincides with the theoretical background described in Remark 2. Even if we do not apply regularization on the adjacency matrix, the loss rate can be bounded above by $(n_1 p) \vee (\log n_1)$. As p increases, $\log n_1$ can take on the role of the upper bound.
3. Regularization has little affect on soft singular value thresholding (Algorithm 2). When $p = 0.01$, regularization constant = 0.8 works slightly better than 0.7 and 0.9. For $p = 0.03$ and 0.05, the performances are almost identical when the regularization constant is greater than 0.6. The performance of Algorithm 2 in this experiment is slightly different from our expectation, but it does not violate any theorems in the previous sections. It could be an interesting future work to explain why regularization does not substantially improve the soft singular value thresholding method.

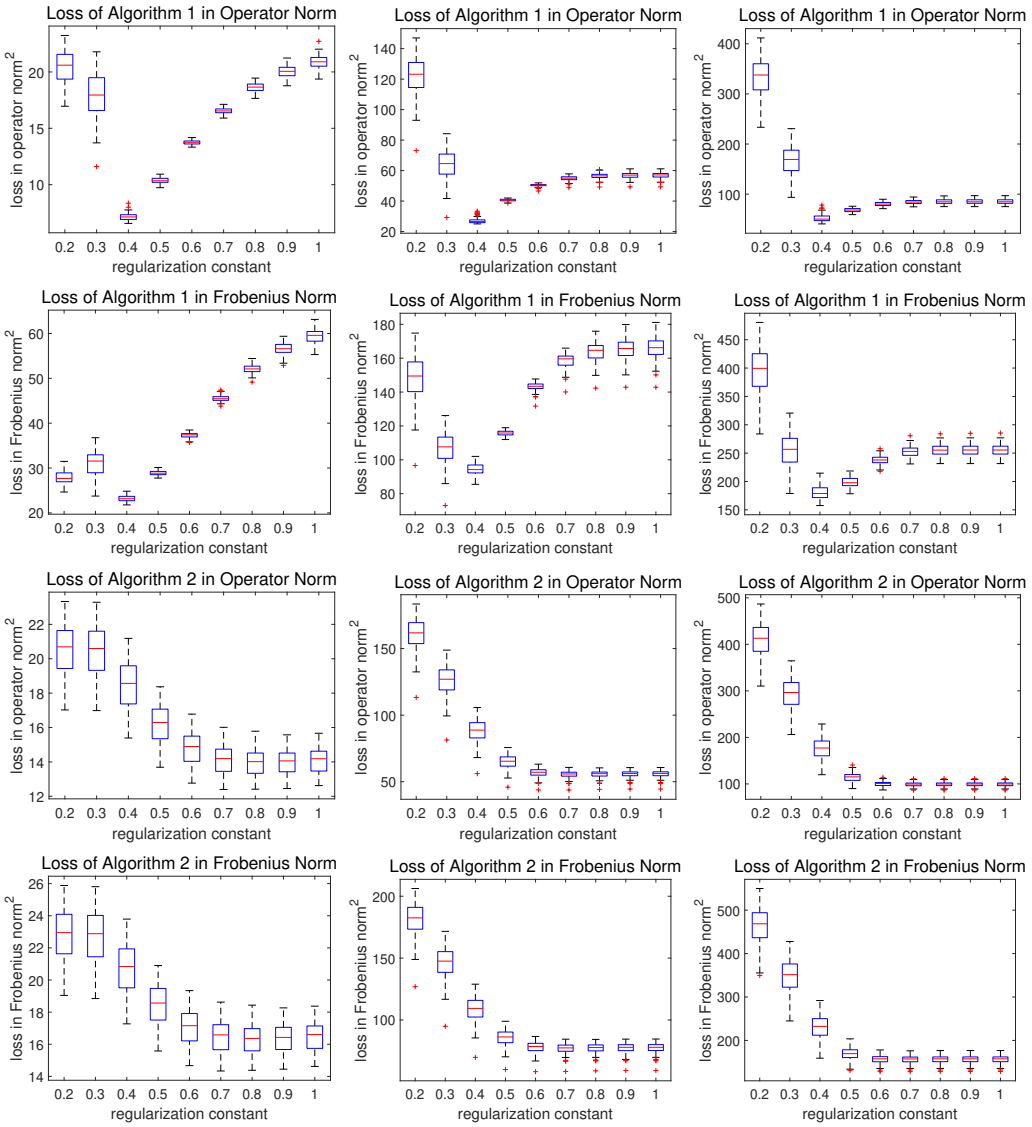

Figure 1: Numerical experiments of Algorithm 1 and Algorithm 2. The probability matrix $\mathbf{M}$'s are generated from the parameter space $\mathbf{\Theta}_1(n_1, n_2, r, p)$ with $n_1 = n_2 = 1000, r = 3$. $p = 0.01, 0.03, 0.05$ on different columns (from left to right). The horizontal axis represents the choice of regularization parameter.

4.2 Simulation on Subspace Estimation

The simulation results of Algorithm 3 are presented in Figure 2. Due to the simulation procedure, it is difficult to control the smallest nonzero singular value of the connectivity matrix $\mathbf{M}$. Therefore, we plot the loss measured by $\sigma_*^2 L_{\text{op}}(\hat{\mathbf{U}}, \mathbf{U})$ and $\sigma_*^2 L_{\text{F}}(\hat{\mathbf{U}}, \mathbf{U})$. We observe that the regularization constant does not have a significant effect on the performance, but we can still conclude that the algorithm performs slightly better if we set the constant to be around 0.5 when we measure the loss by Frobenius norm.

5 Discussion

In this section, we will discuss some possible future works.

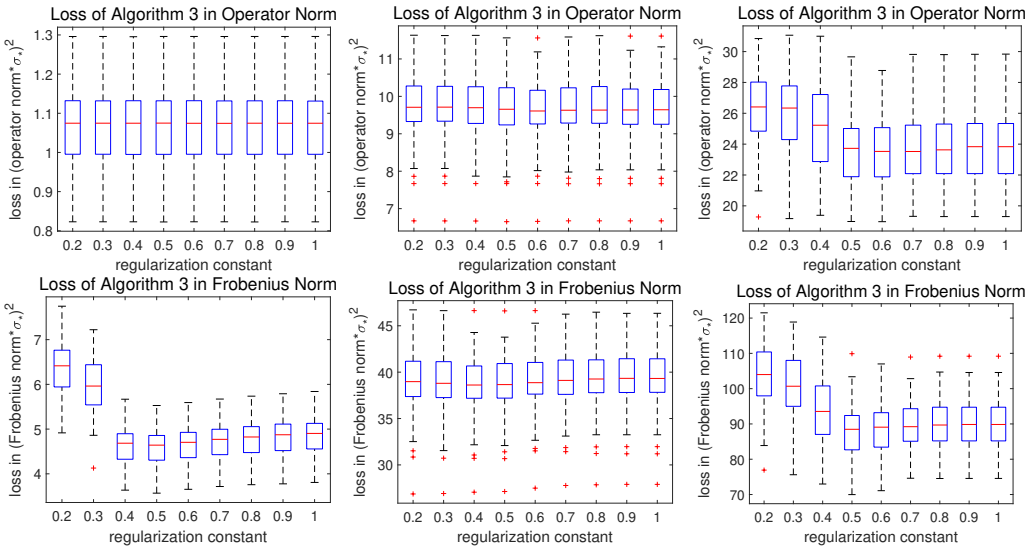

Figure 2: Numerical experiments of Algorithm 3. The probability matrix $\mathbf{M}$'s are generated from the parameter space $\mathbf{\Theta}_2(n_1, n_2, r, p, \sigma)$ with $n_1 = n_2 = 2000, r = 3$. $p = 0.01, 0.03, 0.05$ on different columns (from left to right). The horizontal axis represents the choice of regularization parameter.

5.1 Bipartite Graphs with Unbalanced Parts

Suppose the bipartite graphs has unbalanced parts, i.e., $n_1 \ll n_2$ for the dimension of the adjacency matrix, then the upper bound in Theorem 4 may not be optimal for $L_{\text{op}}(\mathbf{U}, \hat{\mathbf{U}})$. Intuitively, suppose there are more columns than rows in the adjacency matrix, then we are able to collect more information about the column space of the connectivity matrix $\mathbf{M}$. It is not clear how the assumption $n_1 \ll n_2$ affects the error rate of $L_{\text{op}}(\mathbf{U}, \hat{\mathbf{U}})$ in our theorem. On the other hand, increasing number of rows also changes the parameter of σ_*, the smallest nonzero singular value. Some recent works [Ndaoud et al., 2019, Cai et al., 2021] have focused on unbalanced random matrix, but the authors have not shown their results are optimal in the random bipartite graph setting. The generalized theorem in Cai et al. [2018] enable us to find a tight bound for random matrix with Gaussian entries, but the application on sparse random graphs is not straightforward. It is still an open problem to find the minimax error rate of $L_{\text{op}}(\mathbf{U}, \hat{\mathbf{U}})$ and $L_{\text{F}}(\mathbf{U}, \hat{\mathbf{U}})$.

5.2 Unspecified Parameters

To establish the minimax lower bound, we specify the sparsity parameter p and the rank r. However, these parameters are unknown in general in practice. A possible approach is to tune these parameters by cross validation. We randomly partition the bipartite graph into a training subgraph and a validation subgraph, then we apply the algorithms in the present paper with various parameters, and compute the likelihood on the validation subgraph. We can generalize this procedure to cross validation and find the parameters so that the average likelihood on validation subgraphs are maximized. The implementation of cross validation should be straightforward, while the risk of estimation without specified parameters can be an interesting future work.

6 Conclusion

In this paper, we study bipartite graphs with underlying low rank structures. We have introduced the algorithms for finding the connectivity matrix and its column and row space, using the techniques of regularization, hard and soft singular value thresholding. We provide theoretical analysis in Theorem 1, Theorem 2 and Theorem 4 for the algorithms we proposed. In addition, minimax lower bounds in Theorem 3 and Theorem 5 show that the algorithms return rate optimal estimators for corresponding tasks. In addition, simulation studies in Section 4 investigate the effect of regularization parameters in the algorithms. Finally, we discuss some possible future works in Section 5.

References

[1] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[2] Changxiao Cai, Gen Li, Yuejie Chi, H Vincent Poor, and Yuxin Chen. Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *The Annals of Statistics*, 49(2):944–967, 2021.

[3] 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.

[4] T Tony Cai, Anru Zhang, et al. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018.

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

[6] Sourav Chatterjee et al. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[7] Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory*, pages 391–423. PMLR, 2015.

[8] Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[9] Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

[10] Chao Gao, Zongming Ma, Anderson Y Zhang, Harrison H Zhou, et al. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185, 2018.

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

[12] Can M Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 2017.

[13] Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *Annals of Statistics*, 43(1):215–237, 2015.

[14] Mohamed Ndaoud, Suzanne Sigalla, and Alexandre B Tsybakov. Improved clustering algorithms for the bipartite stochastic block model. *arXiv preprint arXiv:1911.07987*, 2019.

[15] Stanislaw J Szarek. Nets of grassmann manifold and orthogonal group. In *Proceedings of research workshop on Banach space theory (Iowa City, Iowa, 1981)*, volume 169, page 185, 1982.

[16] JH Wilkinson. Elementary proof of the wielandt-hoffman theorem and of its generalization. 1970.

[17] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[18] Zhixin Zhou and Arash A Amini. Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *Journal of Machine Learning Research*, 20(47):1–47, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] The proofs can be found in the supplemental material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[TODO]** We will upload the code once this work is published.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]