Journal of Machine Learning Research 21 (2020) 1-45 Submitted 1/20; Published 07/20

Optimal Estimation of Sparse Topic Models

Xin Bing XB43@CORNELL.EDU

Florentina Bunea FB238@CORNELL.EDU
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14850, USA

Marten Wegkamp MHW73@CORNELL.EDU
Department of Mathematics and Department of Statistics and Data Science
Cornell University
Ithaca, NY 14850, USA

Editor: Arnak Dalalyan

Abstract

Topic models have become popular tools for dimension reduction and exploratory analysis of text data which consists in observed frequencies of a vocabulary of p words in n documents, stored in a $p \times n$ matrix. The main premise is that the mean of this data matrix can be factorized into a product of two non-negative matrices: a $p \times K$ word-topic matrix A and a $K \times n$ topic-document matrix W.

This paper studies the estimation of A that is possibly element-wise sparse, and the number of topics K is unknown. In this under-explored context, we derive a new minimax lower bound for the estimation of such A and propose a new computationally efficient algorithm for its recovery. We derive a finite sample upper bound for our estimator, and show that it matches the minimax lower bound in many scenarios. Our estimate adapts to the unknown sparsity of A and our analysis is valid for any finite n, p, K and document lengths.

Empirical results on both synthetic data and semi-synthetic data show that our proposed estimator is a strong competitor of the existing state-of-the-art algorithms for both non-sparse A and sparse A, and has superior performance is many scenarios of interest.

Keywords: topic models, minimax estimation, sparse estimation, adaptive estimation, high dimensional estimation, non-negative matrix factorization, separability, anchor words

1. Introduction

Topic modeling has been a popular and powerful statistical model during the last two decades in machine learning and natural language processing for discovering thematic structures from a corpus of documents. Topic models have wide applications beyond the context in which was originally introduced, to genetics, neuroscience and social science (Blei, 2012), to name just a few areas in which they have been successfully employed.

In the computer science and machine learning literature, topic models were first introduced as *latent semantic indexing models* by Deerwester et al. (1990); Papadimitriou et al. (1998); Hofmann (1999); Papadimitriou et al. (2000). For uniformity and clarity, we explain our methodology in the language typically associated with this set-up. A corpus of n

documents is assumed to follow generative models based on the bag-of-word representation. Specifically, each document $X_i \in \mathbb{R}^p$ is a vector containing empirical (observed) frequencies of p words from a pre-specified dictionary, generated as

$$X_i \sim \frac{1}{N_i}\text{Multinomial}_p\left(N_i, \Pi_i\right), \quad \text{for each } i \in [n] := \{1, 2, \ldots, n\}. \tag{1}$$

Here N_i denotes the length (or the number of sampled words) in the ith document. The expected frequency vector $\Pi_i \in \mathbb{R}^p$ is called the word-document vector, and is a convex combination of K word-topic vectors with weights corresponding to the allocation of K topics. Mathematically, one postulates that

$$\Pi_i = \sum_{k=1}^{K} A_{\cdot k} W_{ki} \tag{2}$$

where $A_{\cdot k} = (A_{1k}, \ldots, A_{pk})$ is the word-topic vector for the kth topic and $W_{\cdot i} = (W_{1i}, \ldots, W_{Ki})$ is the allocation of K topics in this ith document. From a probabilistic point of view, equation (2) has the conditional probability interpretation

$$\underbrace{\mathbb{P}(\text{word } j \mid \text{document } i)}_{\Pi_{ji}} = \sum_{k=1}^{K} \underbrace{\mathbb{P}(\text{word } j \mid \text{topic } k)}_{A_{jk}} \cdot \underbrace{\mathbb{P}(\text{topic } k \mid \text{document } i)}_{W_{ki}} \tag{3}$$

for each $j \in [p]$, justified by Bayes' theorem. As a result, the (expected) word-document frequency matrix $\Pi = (\Pi_1, \ldots, \Pi_n) \in \mathbb{R}^{p \times n}$ has the following decomposition

$$\Pi = AW = A(W_1, \ldots, W_n). \tag{4}$$

The entries of the columns of Π, A and W are probabilities, so they are non-negative and sum to one:

$$\sum_{j=1}^{p} \Pi_{ji} = 1, \quad \sum_{j=1}^{p} A_{jk} = 1, \quad \sum_{k=1}^{K} W_{ki} = 1, \quad \text{for any } k \in [K] \text{ and } i \in [n]. \tag{5}$$

Since the number of topics, K, is typically much smaller than p and n, the matrix Π exhibits a low-rank structure. In the topic modeling literature, the main interest is to recover the matrix A when only the $p \times n$ frequency matrix $X = (X_1, \ldots, X_n)$ and the document lengths $N_1, \ldots, N_n$ are observed.

One direction of a large body of work is of Bayesian nature, and the most commonly used prior distribution on W is the Dirichlet distribution (Blei et al., 2003). Posterior inference on A is then typically conducted via variational inference (Blei et al., 2003), or sampling techniques involving MCMC-type solvers (Griffiths and Steyvers, 2004). We refer to Blei (2012) for a in-depth review.

The computational intensive nature of Bayesian approaches, in high dimensions, motivated a separate line of recent work that develops efficient algorithms, with theoretical guarantees, from a frequentist perspective. Anandkumar et al. (2012) proposes an estimation method, with provable guarantees, that employs the third moments of Π via a

tensor-decomposition. However, the success of this approach requires the topics not be correlated, and in many situations there is strong evidence suggesting the contrary (Blei and Lafferty, 2007; Li and McCallum, 2006).

This motivated another line of work, similar in spirit with the work presented in this paper, which relies on the following *separability* condition on A, and allows for correlated topics.

Assumption 1 (separability) *For each topic $k \in [K]$, there exists at least one word j such that $A_{jk} > 0$ and $A_{j\ell} = 0$ for any $\ell \neq k$.*

The separability condition was first introduced by Donoho and Stodden (2004) to ensure uniqueness in the Non-negative Matrix Factorization (NMF) framework. Arora et al. (2012) introduce the separability condition to the topic model literature with the interpretation that, for each topic, there exist some words which *only* occur in this topic. These special words are called *anchor words* (Arora et al., 2012) and guarantee recovery of A, coupled with the following condition on W (Arora et al., 2012).

Assumption 2 *Assume the matrix $n^{-1}WW^\top$ is strictly positive definite.*

Finding anchor words is the first step towards the recovery of the desired target A. Many algorithms are developed for this purpose, see, for instance, Arora et al. (2012); Recht et al. (2012); Arora et al. (2013); Ding et al. (2013); Ke and Wang (2017). All these works require the number of topics K be *known*, yet in practice K is rarely known. This motivated us Bing et al. (2020) to develop a method that estimates K consistently from the data under the *incoherence* Condition 3 on the topic-document matrix W given in Section 5. We defer to this for further discussion of other existing methods for finding anchor words.

Despite the wide-spread interest and usage of topic models, most of the existing works are mainly devoted to the computational aspects of estimation, and relatively few works provide statistical guarantees for estimators of A. An exception is Arora et al. (2012, 2013) that provide upper bounds for the ℓ_1-loss $\|\widehat{A} - A\|_1 = \sum_{j=1}^{p}\sum_{k=1}^{K}|\widehat{A}_{jk} - A_{jk}|$ of their estimator. Their analysis allows K, p and N_i to grow with n. Unfortunately, the convergence rate of their estimator is not optimal (Ke and Wang, 2017; Bing et al., 2020). The recent work of Ke and Wang (2017) is the first to establish the minimax lower bound for the estimator of A in topic models for known, fixed K. Their estimator provably achieves the minimax optimal rate under appropriate conditions. When K is allowed to grow with n, the minimax optimal rate of $\|\widehat{A} - A\|_1$ is established in Bing et al. (2020) and an optimal estimation procedure is proposed.

Despite these recent advances, all the aforementioned results are established for a *fully dense* matrix A. In the modern big data era, the dictionary size p, the number of documents n and the number of topics K are large, as evidenced by real data in Section 6. Sparsity is likely to happen for large dictionaries (p) and when the number of topics K is large, one should expect that there are many words *not* occurring in all topics, that is, $A_{jk} = \mathbb{P}(\text{word } j \mid \text{topic } k) = 0$ for some k.

To the best of our knowledge, the minimax lower bound of $\|\widehat{A} - A\|_1$ in the topic model is unknown when the word-topic matrix A is element-wise sparse and no estimation procedure exists tailored to this scenario of sparse A and unknown K.

1.1. Our Contributions

We summarize our contributions in this paper.

New minimax lower bound for $\|\widehat{A} - A\|_1$, when A is sparse. To understand the difference of estimating a dense A and a entry-wise sparse A in topic models, we first establish the minimax lower bound of estimators of A in Theorem 1 of Section 2. It shows that

$$\inf_{\widehat{A}} \sup_{A} \mathbb{P}_A \left\{ \|\widehat{A} - A\|_1 \geq c_0 \|A\|_1 \sqrt{\frac{\|A\|_0}{nN}} \right\} \geq c_1.$$

for some constants $c_0 > 0$ and $c_1 \in (0, 1]$, by assuming $N = N_1 = N_2 = \cdots = N_n$ for ease of presentation. The infimum is taken over all estimators $\widehat{A}$ while the supremum is over a prescribed parameter space $\mathcal{A}$ defined in (7) below. We have $\|A\|_1 = K$ by (5) for all A. The term $\|A\|_0$ characterizes the overall sparsity of A, and the minimax rate of A becomes faster as A gets more sparse. When the rows $A_{j\cdot}$ of non-anchor words j are dense in the sense $\|A_{j\cdot}\|_0 = K$, our result reduces to that in Bing et al. (2020). Our minimax lower bound is valid for all p, K, N and n and, to the best of our knowledge, the lower bound with dependency on the sparsity of A is new in the topic model literature.

A new estimation procedure for sparse A. To the best of our knowledge, the only minimax-optimal estimation procedure, for *dense* A and K large and unknown, is offered in Bing et al. (2020). While the procedure is computationally very fast, it is impractical to adjust it in simple ways in order to obtain a sparse estimator of A, that would hopefully be minimax-optimal.

For instance, simply thresholding an estimator $\widehat{A}$ to encourage sparsity will require threshold levels that vary from row to row, resulting in too many tuning parameters. We propose a new estimation procedure in Section 3 that adapts to this unknown sparsity. To motivate our procedure, we start with the recovery of A in the noise-free case in Section 3.1, under Assumptions 1 and 2. Since several existing algorithms, including Bing et al. (2020), provably select the anchor words, we mainly focus on the estimation of the portion of A corresponding to non-anchor words.

In the presence of noise, we propose our estimator in Section 3.2 and summarize the procedure in Algorithm 1. The new algorithm requires the solution of a quadratic program for each non-anchor row. Except for a ridge-type tuning parameter (which can often be set to zero), the procedure is devoid of any (further) tuning parameters. We give detailed comparisons with other methods in the topic model literature in Section 3.3.

Adaptation to sparsity. We provide finite sample upper bounds on the ℓ_1 loss of our new estimator in Section 4, valid for all p, K, n and N. As shown in Theorem 2, our estimator adapts to the unknown sparsity of A. To the best of our knowledge, our estimator is the first computationally fast estimator shown to adapt to the unknown sparsity of A. We further show in Corollary 3 that it is minimax optimal under reasonable scenarios.

Simulation study. In Section 6, we provide experimental results based on both synthetic data and semi-synthetic data. We compare our new estimator with existing state-of-the-art algorithms. The effect of sparsity on the estimation of A is verified in Section 6.1 for

synthetic data, while we analyze two semi-synthetic data sets based on a corpus of NIPs articles and a corpus of New York Times (NYT) articles in Section 6.2.

1.2. Notation

We introduce notation that we use throughout the paper. The integer set $\{1, \ldots, n\}$ is denoted by $[n]$. We use $\mathbf{1}_d$ to denote the d-dimensional vector with entries equal to 1 and use $\{e_1, \ldots, e_K\}$ to denote the canonical basis vectors in $\mathbb{R}^K$. For a generic set S, we denote $|S|$ as its cardinality. For a generic vector $v \in \mathbb{R}^d$, we let $\|v\|_q$ denote the vector ℓ_q norm, for $q = 0, 1, 2, \ldots, \infty$, and let $\mathrm{supp}(v)$ denote its support. We write $\|v\|_2 = \|v\|$ for brevity. We denote by $\mathrm{diag}(v)$ a $d \times d$ diagonal matrix with diagonal elements equal to v. For a generic matrix $Q \in \mathbb{R}^{d\times m}$, we write $\|Q\|_1 = \sum_{1\le i\le d, 1\le j\le m} |Q_{ij}|$ and $\|Q\|_{\infty,1} = \max_{1\le i\le d} \sum_{1\le j\le m} |Q_{ij}|$. For the submatrix of Q, we let $Q_{i\cdot}$ and $Q_{\cdot j}$ be the ith row and jth column of Q. For a set S, we let Q_S and $Q_{\cdot S}$ denote its $|S| \times m$ and $d \times |S|$ submatrices. For a symmetric matrix Q, we denote its smallest eigenvalue by $\lambda_{\min}(Q)$. We use $a_n \lesssim b_n$ to denote there exists an absolute constant $c > 0$ such that $a_n \le cb_n$, and write $a_n \asymp b_n$ if there exists two absolute constants $c, c' > 0$ such that $cb_n \le a_n \le c'b_n$. In the probabilities of our results, we might write $c''a_n$ as $O(a_n)$ for some absolute constant $c'' > 0$. Finally, we write $a_n = o_p(b_n)$ if $a_n/b_n \to 0$ with probability tending to 1.

For a given word-topic matrix A, we let $I := I(A)$ be the set of anchor words, and $\mathcal{I}$ be its partition relative to the K topics. That is,

$$I_k := \{j \in [p] : A_{jk} > 0,\ A_{j\ell} = 0 \ \text{ for all } \ell \neq k\}, \quad I := \bigcup_{k=1}^{K} I_k, \quad \mathcal{I} := \{I_1, \ldots, I_K\}. \tag{6}$$

We further write $J := [p] \setminus I$ to denote the set of non-anchor words. For the convenience of our analysis, we assume all documents have the same number of sampled words, that is, $N := N_1 = \cdots = N_n$, while our results can be extended to the general case.

2. Minimax lower bounds of $\|\widehat{A} - A\|_1$

In this short section, we establish the minimax lower bound of $\|\widehat{A} - A\|_1$ based on model (4) for any estimator $\widehat{A}$ of A over the parameter space

$$\mathcal{A} := \Big\{A \in \mathbb{R}_+^{p\times K} : A^\top \mathbf{1}_p = \mathbf{1}_K,\ A \text{ satisfies Assumption 1 with } \|A\|_0 \le nN\Big\}. \tag{7}$$

To prove the lower bound, it suffices to choose one particular W. We let

$$W^0 = \{\underbrace{e_1, \ldots, e_1}_{n_1}, \underbrace{e_2, \ldots, e_2}_{n_2}, \ldots, \underbrace{e_K, \ldots, e_K}_{n_K}\} \tag{8}$$

with $\sum_{k=1}^K n_k = n$ and $|n_k - n_{k'}| \le 1$ for $k, k' \in [K]$. Note that W^0 satisfies Assumption 2. Denote by $\mathbb{P}_A$ the joint distribution of $(X_1, \ldots, X_n)$ under model (4), for the chosen W^0.

Theorem 1 *Under topic model (4), assume (1). Then, there exist constants $c_0 > 0$ and $c_1 \in (0, 1]$ such that*

$$\inf_{\widehat{A}} \sup_{A\in\mathcal{A}} \mathbb{P}_A \left\{\|\widehat{A} - A\|_1 \ge c_0 \|A\|_1 \sqrt{\frac{\|A\|_0}{nN}}\right\} \ge c_1. \tag{9}$$

The infimum is taken over all estimators $\widehat{A}$ of A.

Remark 1 The estimate constructed in the next section achieves this lower bound in many scenarios. The lower bound rate of $\|\widehat{A} - A\|_1$ in (9) becomes faster as $\|A\|_0$ decreases, that is, if A becomes more sparse. Since each of the K columns of A sum to one, we always have $\|A\|_1 = K$. If the submatrix A_J, corresponding to the non-anchor words, is dense in the sense that $\|A_J\|_0 = K|J|$, Theorem 1 reduces to the result in (Bing et al., 2020, Theorem 6) for $K = K(n)$, and the result in (Ke and Wang, 2017, Theorem 2.2) for fixed K.

3. Estimation of A

In this section, we present our procedure for estimating A when a subset of anchor words $L = \bigcup_{k=1}^{K} L_k$ and its partition $\mathcal{L} = \{L_1, \ldots, L_K\}$ are given. Moreover, we assume that, for each $k \in [K]$, $L_k \subseteq I_{\pi(k)}$ for some group permutation $\pi : [K] \to [K]$. For simplicity of presentation, we assume π is identity such that

$$L_k \subseteq I_k, \qquad \text{for each } k \in [K]. \tag{10}$$

We discuss methods for selecting L and $\mathcal{L}$ in Section 5. We start with the noise-free case, that is, we observe the expected word-document frequency matrix Π, in Section 3.1. Our strategy is as follows. Instead of inferring A from Π directly, we consider the scaled version B that has all its rows sum to 1, but critically retains the same sparsity pattern of A. The submatrix B_L of B with rows corresponding to the representative set L is then immediate to find, as the i-th row of B equals the unit vector $\mathbf{e}_k$, for each $i \in L_k$. Next, we show that the remaining submatrix B_{L^c} of the sparse matrix B can be found, row-by-row, using a quadratic optimization over the probability simplex, making essential use of the fact that each row of B sums to one. Finally, we recover the original matrix A by column-wise renormalization of the obtained matrix B. Motivated by the developed algorithm in the noise-free case that recovers A, we propose the estimation procedure of A in Section 3.2 when we have access to X only, requiring slight modifications from the procedure for the noiseless case, including a hard-thresholding step to handle words with extremely low frequencies. The optimization problem is an efficient $K \times K$ quadratic program, which gives it an edge over previous works such as Arora et al. (2013).

3.1. Recovery of A in the Noise-free Case

Suppose that Π is given and write $D_\Pi := n^{-1}\mathrm{diag}(\Pi \mathbf{1}_n)$ and $D_W := n^{-1}\mathrm{diag}(W\mathbf{1}_n)$. We recover A via its row-wisely normalized version

$$B = D_\Pi^{-1} A D_W \tag{11}$$

as B enjoys the following three properties:

$$\mathrm{supp}(B) = \mathrm{supp}(A), \quad B_{jk} \in [0,1], \quad \|B_{j\cdot}\|_1 = 1, \qquad \text{for all } j \in [p], k \in [K]. \tag{12}$$

The row-wise sum-to-one property is critical in the later estimation step to adapt to the unknown sparsity of B (or equivalently, the sparsity of A). From $\mathcal{L} = \{L_1, \ldots, L_K\}$ and (12), we can directly recover B_L by setting

$$B_{i\cdot} = e_k, \qquad \text{for any } i \in L_k, k \in [K].$$

To recover B_{L^c} with $L^c := [p] \setminus L$, let

$$R := D_\Pi^{-1}\Theta D_\Pi^{-1} = B\left(D_W^{-1}\frac{1}{n}WW^\top D_W^{-1}\right)B^\top := BMB^\top$$

be a normalized version of

$$\Theta := n^{-1}\Pi\Pi^\top.$$

Since R has the decomposition

$$R_{LL} = B_L M B_L^\top, \qquad R_{L^cL} = B_{L^c} M B_L^\top.$$

and Assumption 2 implies M is invertible, we arrive at the expressions

$$M = (B_L^\top B_L)^{-1}B_L^\top R_{LL}B_L(B_L^\top B_L)^{-1}, \tag{13}$$
$$B_{L^c} = R_{L^cL}B_L(B_L^\top B_L)^{-1}M^{-1}. \tag{14}$$

Display (14) implies that

$$MB_{L^c}^\top = (B_L^\top B_L)^{-1}B_L^\top R_{LL^c} := H,$$

whence $M\beta = h$ for each column β of $B_{L^c}^\top$ (which is a *row* of B_{L^c}) and corresponding column h of H. Given M and H, the solution β of the equation $M\beta = h$ is the minimizer of $\beta^\top M\beta - 2\beta^\top h$ over $\beta \geq 0$ and $\|\beta\|_1 = 1$. This formulation will be used in the next subsection.

After recovering $B^\top = (B_L^\top, B_{L^c}^\top)$, display (11) implies that A can be recovered by normalizing columns of $D_\Pi B$ to unit sums.

3.2. Estimation of A in the Noisy Case

The estimation procedure of A follows the same idea of the noise-free case. We first estimate B defined in (11) by using the estimate

$$\widehat{R} = D_X^{-1}\widehat{\Theta}D_X^{-1} \tag{15}$$

of R, based on $D_X = n^{-1}\mathrm{diag}(X\mathbf{1}_n)$ and the unbiased estimator

$$\widehat{\Theta} = \frac{1}{n}\sum_{i=1}^n\left[\frac{N_i}{N_i-1}X_iX_i^\top - \frac{1}{N_i-1}\mathrm{diag}(X_i)\right] \tag{16}$$

of the matrix Θ. We estimate B_L by

$$\widehat{B}_{i\cdot} = e_k, \quad \text{for any } i \in L_k, k \in [K]. \tag{17}$$

Based on

$$\widehat{M} = (\widehat{B}_L^\top\widehat{B}_L)^{-1}\widehat{B}_L^\top\widehat{R}_{LL}\widehat{B}_L(\widehat{B}_L^\top\widehat{B}_L)^{-1}, \qquad \widehat{H} = (\widehat{B}_L^\top\widehat{B}_L)^{-1}\widehat{B}_L^\top\widehat{R}_{LL^c}, \tag{18}$$

we estimate row-by-row the remainder of the matrix B. We compute, for each $j \in L^c$,

$$\widehat{B}_{j\cdot} = 0, \qquad \text{if } (D_X)_{jj} \le 7\log(n \vee p)/(nN), \tag{19}$$

$$\widehat{B}_{j\cdot} = \arg \min_{\beta \ge 0,\ \|\beta\|_1 = 1} \beta^\top (\widehat{M} + \lambda \boldsymbol{I}_K)\beta - 2\beta^\top \widehat{h}^{(j)}, \quad \text{otherwise}, \tag{20}$$

where $\widehat{h}^{(j)}$ is the corresponding column of $\widehat{H}$. We set $\lambda = 0$ whenever $\widehat{M}$ is invertible and otherwise choose λ large enough such that $\widehat{M} + \lambda \boldsymbol{I}_K$ is invertible. We detail the exact rate of λ when $\widehat{M}$ is not invertible in Section 4. Finally, we estimate A via normalizing $D_X\widehat{B}$ to unit column sums.

Remark 2 In our procedure, the hard-thresholding step in (19) is critical to obtain the optimal rate of the final estimator that does not rely on a lower bound condition on the word-frequencies. In contrast, the analysis of Arora et al. (2013) requires a lower bound for all word-frequencies. The thresholding level in (19) is carefully chosen from the element-wise control of the difference $D_X - D_\Pi$.

For the reader's convenience, the estimation procedure is summarized in Algorithm 1.

Algorithm 1 Sparse Topic Model solver (STM)

Require: frequency data matrix $X \in \mathbb{R}^{p\times n}$ with document lengths $N_1, \dots, N_n$; the partition of anchor words $\mathcal{L}$

procedure
- compute $D_X = n^{-1}\mathrm{diag}(X\mathbf{1}_n)$, $\widehat{\Theta}$ from (16) and $\widehat{R}$ from (15)
- compute $\widehat{B}_L$ from (17)
- compute $\widehat{M}$ and $\widehat{H}$ from (18)
- solve $\widehat{B}_{L^c}$ from (19) – (20) by using λ in (29)
- compute $\widehat{A}$ by normalizing $D_X\widehat{B}$ to unit column sums
- **return** $\widehat{A}$

3.3. Comparison with Existing Methods

In this section, we provide comparisons between our estimation procedure and two existing methods, which are seemingly close to our procedure.

3.3.1. Comparison with Arora et al. (2013)

This algorithm also estimates the same target B defined in (11) first. For a given set L of anchor words, there are two main differences for estimating B.

1. The algorithm in Arora et al. (2013) uses *only one* anchor word per topic to estimate B whereas our estimation procedure utilizes all anchor words. The benefit of using multiple anchor words per topic is substantial and verified in our simulation in Section 6.

2. The algorithm in Arora et al. (2013) is based on different quadratic programs with more parameters (pK versus K^2). This makes it more computationally intensive

and less accurate than the algorithm proposed here. This is verified in our simulations in Section 6.2. Specifically, write $\widetilde{\Theta} := D_\Theta^{-1}\Theta = D_\Theta^{-1}A(n^{-1}WW^\top)A^\top$, $Q := (n^{-1}WW^\top)A^\top$ and $\widetilde{Q} := D_Q^{-1}Q$ with $D_\Theta = \mathrm{diag}(\Theta\mathbf{1}_p)$ and $D_Q = \mathrm{diag}(Q\mathbf{1}_p)$. Arora et al. (2013) utilizes the following observation

$$\widetilde{\Theta} = D_\Theta^{-1}AQ = D_\Theta^{-1}AD_Q\widetilde{Q} = B\widetilde{Q}$$

by noting that $D_\Theta = D_\Pi$ and $D_Q = D_W$ from (5). Based on the observation that $\widetilde{\Theta}_{j\cdot} \in \mathbb{R}^p$ is a convex combination of $\widetilde{\Theta}_{\widetilde{L}} = \widetilde{Q} \in \mathbb{R}^{K\times p}$ for any $j \in [p] \setminus \widetilde{L}$, Arora et al. (2013) proposes to estimate $B_{j\cdot}$ by solving

$$\widehat{B}_{j\cdot} = \arg\min_{\beta\geq 0, \|\beta\|_1=1} \left\| \widehat{\widetilde{\Theta}}_{j\cdot} - \beta^\top \widehat{\widetilde{Q}} \right\|^2 \tag{21}$$

where $\widehat{\widetilde{\Theta}}_{j\cdot}$ and $\widehat{\widetilde{Q}}$ are the corresponding estimates of $\widetilde{\Theta}_{j\cdot}$ and $\widetilde{Q}$. The matrix $\widetilde{Q}$ contains $p \times K$ entries, while our estimation procedure in (20) only requires to estimate $M \in \mathbb{R}^{K\times K}$ which has fewer parameters. The analysis of Arora et al. (2013) only holds for invertible estimates $\widetilde{Q}\widetilde{Q}^\top$ and the rate of the estimator from (21) depends on $\lambda_{\min}(\widetilde{Q}\widetilde{Q}^\top)$. Our result holds as long as $\lambda_{\min}(M) > 0$ due to the ridge-type estimator in (20) and the rate of our estimator in (20) depends on $\lambda_{\min}(M)$. Lemma 20 in the Appendix shows that

$$\lambda_{\min}(M)\lambda_{\min}(n^{-1}WW^\top) \min_{k\in[K], i\in I_k} A_{ik}^2 \leq \lambda_{\min}(\widetilde{Q}\widetilde{Q}^\top) \leq \lambda_{\min}(M).$$

Since $0 < \lambda_{\min}(n^{-1}WW^\top) \leq 1/K$ as shown in Lemma 21 and $0 < \min_{i\in I_k, k\in[K]} A_{ik}^2 < 1$, it is easy to see that $\lambda_{\min}(\widetilde{Q}\widetilde{Q}^\top)$ could be much smaller comparing to $\lambda_{\min}(M)$. This suggests that our procedure in (20) should be more accurate than (21), which is confirmed in our simulations in Section 6.2.

3.3.2. Comparison with Bing et al. (2020)

Although both methods are based on the normalized second moment R, they differ significantly in estimating A.

1. The algorithm in Bing et al. (2020) uses R *only* to estimate the anchor words and relies on Θ for the estimation of B. Specifically, by observing

$$\Theta_{\cdot\widetilde{L}} := ACA_{\widetilde{L}} = AA_{\widetilde{L}}^{-1}A_{\widetilde{L}}CA_{\widetilde{L}} = AA_{\widetilde{L}}^{-1}\Theta_{\widetilde{L}\widetilde{L}} := \widetilde{A}\Theta_{\widetilde{L}\widetilde{L}}$$

with $\widetilde{L}$ being a set that contains one anchor word per topic and $A_{\widetilde{L}} \in \mathbb{R}^{K\times K}$ being a diagonal matrix, Bing et al. (2020) proposes to first estimate $\widetilde{A}$ by $\widehat{\Theta}_{\cdot\widetilde{L}}\widehat{\Omega}$. Here $\widehat{\Omega}$ is an estimator of $\Theta_{\widetilde{L}\widetilde{L}}^{-1}$ obtained via solving a linear program. Instead of $\widetilde{A}$, we propose here to first estimate B defined in (11). This is a different scaled version of A with more desirable structures (12).

2. Furthermore, our estimation of B is done row-by-row via quadratic programming instead of simple matrix multiplication. While this is more computationally expensive than estimating $\Theta_{\widetilde{L}\widetilde{L}}^{-1}$, it gives more accurate row-wise control of $\widehat{B}-B$. This control is the key to obtain a faster rate of $\|\widehat{A}-A\|_1$ that adapts to the unknown sparsity.

3. Finally, we emphasize that it is impractical to modify the estimator of Bing et al. (2020) to adapt to the sparsity of A. For instance, further thresholding the estimator of $\widetilde{A}$ to encourage sparsity, will require the thresholding levels to vary row-by-row. This would involve too many tuning parameters.

4. Upper Bounds of $\|\widehat{A}-A\|_1$

To simplify notation and properly adjust the scales, for each $j\in[p]$ and $k\in[K]$, we define

$$\mu_j := \frac{p}{n}\sum_{i=1}^{n}\Pi_{ji}, \qquad \gamma_k := \frac{K}{n}\sum_{i=1}^{n}W_{ki}, \qquad \alpha_j := p\max_{1\le k\le K}A_{jk}, \tag{22}$$

such that $\sum_{j=1}^{p}\mu_j=p$, $\sum_{k=1}^{K}\gamma_k=K$ and $p\le\sum_{j=1}^{p}\alpha_j\le pK$ from (5). For given set L satisfying (10), we further set

$$\underline{\mu}_L=\min_{i\in L}\mu_i, \quad \overline{\gamma}=\max_{1\le k\le K}\gamma_k, \quad \underline{\gamma}=\min_{1\le k\le K}\gamma_k, \quad \underline{\alpha}_L=\min_{i\in L}\alpha_i, \quad \rho_j=\alpha_j/\underline{\alpha}_L. \tag{23}$$

For future reference, we note that

$$\overline{\gamma}\ge 1\ge\underline{\gamma}.$$

As our procedure depends whether the inverse of $\widehat{M}$ defined in (18) exists, we first give a critical bound on the control for the operator norm of $\widehat{M}-M$ and provide insight on the choice of λ in (20).

Lemma 3 *Consider the topic model (4) under assumption 1 and*

$$\min_{i\in L}\frac{1}{n}\sum_{i=1}^{n}\Pi_{ji}\ge\frac{c_0\log(n\vee p)}{N}, \qquad \min_{i\in L}\max_{1\le i\le n}\Pi_{ji}\ge\frac{c_1\log^2(n\vee p)}{N} \tag{24}$$

for some sufficiently large constants $c_0,c_1>0$. Then, with probability $1-O((n\vee p)^{-1})$, we have

$$\|\widehat{M}-M\|_{\mathrm{op}}\lesssim\frac{K}{\underline{\gamma}}\sqrt{\frac{pK\log(n\vee p)}{\underline{\mu}_L nN}}. \tag{25}$$

Remark 4 Arora et al. (2013) observe that the smallest frequency of anchor words plays an important role in the estimation of A. Condition (24) prevents the frequency of anchor words from being too small and also appeared in Bing et al. (2020).

In case the matrix $\widehat{M}$ cannot be inverted, we select $\lambda\ge\|\widehat{M}-M\|_{\mathrm{op}}$ in (20). Lemma 3 thus suggests to choose λ as

$$\lambda=c\cdot\frac{K}{\underline{\gamma}}\sqrt{\frac{pK\log(n\vee p)}{\underline{\mu}_L nN}}, \tag{26}$$

for some absolute constant $c > 0$. Let $\widehat{A}$ be obtained via choosing λ as (26). The following theorem states the upper bound of $\|\widehat{A} - A\|_1$. Our procedure, its theoretical performance and its proof differ from those in Bing et al. (2020). While the proof borrows some preliminary lemmas from Bing et al. (2020), it requires a more refined analysis (see Lemmas 15 – 19 in the Appendix).

We define $s_j = \|A_{j\cdot}\|_0$ for $j \in [p]$, $s_J = \sum_{j\in J} s_j$ and $\widetilde{s}_J := \sum_{j\in L^c}(\alpha_j/\underline{\alpha}_L)s_j = \sum_{j\in L^c}\rho_j s_j$.

Theorem 2 *Under model (4), assume Assumptions 1, 2 with* $\lambda_{\min} := \lambda_K(n^{-1}WW^\top) > 0$ *and (24). Then, with probability* $1 - O((n\vee p)^{-1})$*, we have*

$$\|\widehat{A} - A\|_1 \lesssim \mathrm{I} + \mathrm{II} + \mathrm{III},$$

where

$$\begin{aligned}
\mathrm{I} &= \frac{K}{\underline{\gamma}}\sqrt{\frac{p\log(n\vee p)}{nN}} + \frac{pK\log(n\vee p)}{\underline{\gamma}nN} \\
\mathrm{II} &= \frac{\overline{\gamma}^2}{\underline{\gamma}K\lambda_{\min}}\left\{\max\{s_J + |I| - |L|, \widetilde{s}_J\}\left(\frac{K\log(n\vee p)}{\underline{\gamma}nN} + \sqrt{\frac{p\log^4(n\vee p)}{\underline{\mu}_L nN^3}}\right)\right. \\
&\qquad \left. + K\sqrt{\max\{s_J + |I| - |L|, \widetilde{s}_J\}\frac{\log(n\vee p)}{\underline{\gamma}nN}}\right\} \\
\mathrm{III} &= K\sqrt{K\widetilde{s}_J\cdot\frac{\overline{\gamma}}{\underline{\gamma}}\cdot\frac{\log(n\vee p)}{\underline{\gamma}nN}}
\end{aligned}$$

Furthermore, if

$$\lambda_{\min} \geq c_2\frac{\overline{\gamma}^2}{\underline{\gamma}}\sqrt{\frac{p\log(n\vee p)}{\underline{\mu}_L KnN}} \tag{27}$$

for some sufficiently large constant $c_2 > 0$*, then, with probability* $1-O((n\vee p)^{-1})$*,* $\widehat{A}$ *obtained via* $\lambda = 0$ *enjoys the same rate with* $\mathrm{III} = 0$*.*

Remark 5 The estimation error of A consists of three parts: I, II and III. Each part reflects errors made at different stages of our estimation procedure. Recall that $\widehat{A}$ first uses a hard-thresholding step in (19) and then relies on the estimates $\widehat{B}$ and D_X of B and D_Π, respectively. The first term in I quantifies the error of $D_X - D_\Pi$, while the second term is due to the hard-thresholding step. The second term II is due to the error of $\widehat{B}_j - B_j$ for those $j \in [p] \setminus L$ that pass the test (19). Finally, III stems from the error incurred by the regularization choice of λ.

Remark 6 Condition (27) is a lower bound for the smallest eigenvalue of the matrix $n^{-1}WW^\top$. If it holds, we can set $\lambda = 0$ with high probability and the rate of $\|\widehat{A} - A\|_1$ is improved by (at most) a factor of $\sqrt{K(\overline{\gamma}/\underline{\gamma})}$. Under (24), inequality (27) follows from

$$\lambda_{\min} \geq \frac{c_2}{\sqrt{c_0}}\frac{\overline{\gamma}^2}{\underline{\gamma}}\sqrt{\frac{1}{Kn}}.$$

The following corollary provides sufficient conditions that guarantee that our estimator $\widehat{A}$ constructed in Section 3.2 achieves the optimal minimax rate.

Corollary 3 (Attaining the minimax rate) *Consider the topic model (4) with Assumptions 1 and 2. Suppose further that*

(i) $\|A\|_0 \log(n \vee p) \lesssim nN$;

(ii) $\overline{\gamma} \asymp \underline{\gamma}, \quad \lambda_{\min} \asymp 1/K$;

(iii) $\sum_{j \in L^c} \rho_j s_j \lesssim s_J + |I|$

hold. Further, assume (24) holds with the condition on $\min_{i \in L} n^{-1} \sum_{i=1}^n \Pi_{ji}$ *replaced by*

$$\min_{i \in L} \frac{1}{n} \sum_{i=1}^n \Pi_{ji} \geq c_0 \max \left\{ 1, \frac{(s_J + |I| - |L|) \log^2(n \vee p)}{K^2 N} \right\} \frac{\log(n \vee p)}{N}. \tag{28}$$

Then, with probability $1 - O((n \vee p)^{-1})$, *we have*

$$\|\widehat{A} - A\|_1 \lesssim \|A\|_1 \sqrt{\frac{\|A\|_0 \log(n \vee p)}{nN}}.$$

Remark 7 (Conditions in Corollary 3)

1. Condition (i) is natural (up to the multiplicative logarithmic factor) as $\|A\|_0$ is the effective number of parameters to estimate while nN is the total sample size.

2. The first part of condition (ii), $\underline{\gamma} \asymp \overline{\gamma}$, requires that all topics have similar frequency. The ratio $\overline{\gamma}/\underline{\gamma}$ is called the *topic imbalance* (Arora et al., 2012) and is expected to affect the estimation rate of A.

3. The second part of condition (ii), $\lambda_{\min} \asymp 1/K$, requires that topics are not too correlated. This is expected even for known W, playing the same role of the design matrix in the classical regression setting.

4. Condition (iii) puts a mild constraint on the word-topic matrix A between the selected anchor words and the other words (anchor and non-anchor). It is implied by

$$\sum_{j \in L^c} \frac{s_j}{\sum_{j \in L^c} s_j} \|A_{j\cdot}\|_\infty \lesssim \min_{i \in L} \|A_{i\cdot}\|_1,$$

which in turn is implied by

$$\max_{1 \leq k \leq K} \mathbb{P}\{\text{word } j \mid \text{topic } k\} \lesssim \sum_{k=1}^K \mathbb{P}\{\text{word } i \mid \text{topic k}\}$$

for any $i \in L$ and $j \notin L$. The latter condition prevents the selected anchor words from being much less frequent than the other words.

5. Finally, condition (28) strengthens (24) by requiring a slightly larger lower bound for the frequency of selected anchor words. It is implied by

$$N \geq \frac{\|A\|_0 \log^2(n \vee p)}{K^2} \geq \frac{(s_J + |I| - |L|) \log^2(n \vee p)}{K^2}$$

under (24). As discussed in Arora et al. (2012, 2013); Bing et al. (2020), usage of infrequent anchor words often leads to inaccurate estimation of A.

5. Practical Aspects of the Algorithm

We discuss two practical concerns of our proposed algorithm in Section 3.2:

1. Selection of the number of topics K and subset of anchor words L
2. Data-driven choice of the tuning parameter λ in (26).

5.1. Selection of K and L

Several existing algorithms with theoretical guarantees for finding anchor words in the topic model exist. Most methods rely on finding the vertices of a simplex structure, *provided that the number of topics K is known beforehand*. For known K, Recht et al. (2012) make clever use of the appropriately defined simplex structure on $\Theta = n^{-1}\Pi\Pi^\top$ implied by Assumption 1. However, their method needs to solve a linear program in dimension $p \times p$, which becomes rapidly computationally intractable. Arora et al. (2013) proposes a faster combinatorial algorithm which returns one anchor word per topic. The returned anchor words are shown to be *close to* anchor words within a specified tolerance level. Recently, Ke and Wang (2017) proposes another algorithm for finding anchor words by utilizing the simplex structure of the singular vectors of the word-document frequency matrix. However, their algorithm runs much slower than that of Arora et al. (2013).

In practice, K is rarely known in advance. This situation is addressed in Bing et al. (2020). This work proposes a method that provably estimates K from the data, provided that the topic-document matrix W satisfies the following incoherence condition.

Assumption 3 *The inequality*

$$\cos\left(\angle(W_{i\cdot}, W_{j\cdot})\right) < \frac{\zeta_i}{\zeta_j} \wedge \frac{\zeta_j}{\zeta_i} \qquad \textit{for all } 1 \leq i \neq j \leq K,$$

holds, with $\zeta_i := \|W_{i\cdot}\|_2 / \|W_{i\cdot}\|_1$.

This additional assumption is not needed in the aforementioned work when K is known. When columns of W are i.i.d. samples of Dirichlet distribution, Assumption 3 holds with high probability under mild conditions on the hyper-parameter of Dirichlet distribution (Bing et al., 2020, Lemma 25 in the Supplement). In addition to the estimation of K, the algorithm in Bing et al. (2020) estimates both the set and the partition of *all* anchor words for each topic. This sets it further apart from Arora et al. (2013), as the latter only recovers *one* approximate anchor word for each topic. The algorithm of finding anchor words in Bing et al. (2020) is optimization-free and runs as fast as that in Arora et al. (2013).

Hence, for selecting L, we can use Algorithm 4 in Arora et al. (2013) when K is known and Algorithm 2 in Bing et al. (2020) if K is known or needs to be estimated.

5.2. Data-driven Choice of λ

The precise rate for λ in (26) contains unknown quantities $\underline{\gamma}$ and $\underline{\mu}_L$. We proceed via cross-validation over a specified grid. We prove in Lemma 22 in the Appendix that $|\min_{i\in L}(D_X)_{ii} - \underline{\mu}_L/p| = o_p(\sqrt{\log(n\vee p)/(nN)})$ with $D_X = n^{-1}\mathrm{diag}(X\mathbf{1}_n)$. We recommend the following procedure for selecting λ. For some constant c_0 (our empirical study suggests the choice $c_0 = 0.01$), we take

$$t^* = \arg\min\left\{t \in \{0,1,2,\ldots\} : \widehat{M} + \lambda(t)\boldsymbol{I}_K \text{ is invertible}\right\},$$

with

$$\lambda(t) = t \cdot c_0 \cdot K\left(\frac{K\log(n\vee p)}{[\min_{i\in L}(D_X)_{ii}]n} \cdot \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i}\right)^{1/2}. \tag{29}$$

6. Experimental Results

In this section, we report on the empirical performance of the new algorithm proposed and compare it with existing competitors on both synthetic and semi-synthetic data.

Notation. Recall that n denotes the number of documents, N denotes the number of words drawn from each document, p denotes the dictionary size, K denotes the number of topics, and $|I_k|$ denotes the cardinality of anchor words for topic k. We write $\xi := \min_{k\in[K], i\in I_k} A_{ik}$ for the minimal frequency of anchor words. Larger values of ξ are more favorable for estimation.

Methodology. For competing algorithms, we consider Latent Dirichlet Allocation (LDA) (Blei et al., 2003)[1], the algorithm (AWR) proposed in Arora et al. (2013) and the TOP algorithm proposed in Bing et al. (2020). We use the default values of hyper-parameters for all algorithms. Both LDA and AWR need to specify the number of topics K. In our proposed Algorithm 1 (STM), we choose λ according to (29) and we select the anchor words either via AWR with specified K or via TOP (Bing et al., 2020), and proceed with the estimation of A as described in Section 3.2. We name the resulting estimates STM-AWR and STM-TOP, respectively.

6.1. Synthetic Data

In this section, we use synthetic data to demonstrate the effect of the sparsity of A on the estimation error $\|\widehat{A} - A\|_1/K$ for AWR, TOP, STM-AWR and STM-TOP. Both AWR and STM-AWR are given the correct K, while TOP and STM-TOP estimate K.

To simulate synthetic data, we generate A satisfying Assumption 1 by the following strategy.

- Generate anchor words by $A_{ik} := \xi$ for any $i \in I_k$ and $k \in [K]$.
- Each entry of non-anchor words is sampled from Uniform$(0, 1)$.

1. We use the code of LDA from Riddell et al. (2016) implemented via the fast collapsed Gibbs sampling with the default of 1,000 iterations

- Normalize each sub-column $A_{Jk} \subset A_{\cdot k}$ to have sum $1 - \sum_{i \in I} A_{ik}$.
- Draw columns of W from the symmetric Dirichlet distribution with parameter 0.3.
- Simulate N words from Multinomial$_p(N; AW)$.

To change the sparsity of A, we randomly set $s = \lfloor \eta K \rfloor$ entries of each row in A_J to zero, for a given sparsity proportion $\eta \in (0, 1)$. Normalizing the thresholded matrix gives $A(\eta)$ and the sparsity of $A(\eta)$ is calculated as $s(\eta) = \|A(\eta)\|_0/(pK)$. We set

$$N = 1500,\ p = n = 1000,\ K = 20,\ |I_k| = p/200 \text{ and } \xi = K/p.$$

For each $\eta \in \{0, 0.1, 0.2, \ldots, 0.9\}$, we generate 50 data sets based on $A(\eta)$ and report in Figure 1 the average estimation errors $\|\widehat{A} - A(\eta)\|_1/K$ of the four different algorithms. The x-axis represents the corresponding sparsity level $s(\eta)$. Since the selected anchor words are up to a group permutation, we align the columns of $\widehat{A}$ before calculating the estimation error.

Conclusion. STM-TOP has the best performance overall. Both STM-AWR and STM-TOP perform increasingly better as A becomes sparser. The performance of AWR improves only if the sparsity level is sufficiently large, say $s(\eta) < 0.5$. As expected, TOP does not adapt to the sparsity.

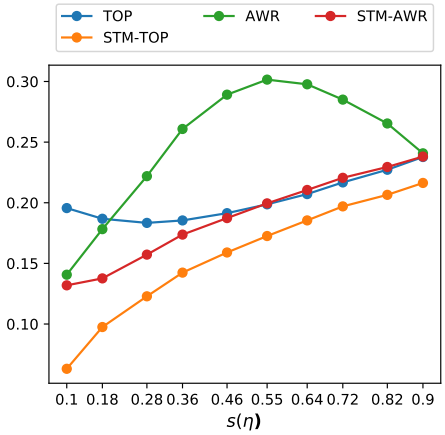

Figure 1: Plots of the estimation error $\|\widehat{A} - A(\eta)\|_1/K$ for $\eta \in \{0, 0.1, 0.2, \ldots, 0.9\}$.

6.2. Semi-synthetic Data

We evaluate two real-world data sets, a corpus of NIPs articles and a corpus of New York Times (NYT) articles (Dheeru and Karra Taniskidou, 2017). Following (Arora et al., 2013),

1. We removed common stopping words and rare words occurring in less than 150 documents.
2. For each preprocessed data set, we apply LDA with $K = 100$ and obtain an estimated word-topic matrix $A^{(0)}$.
3. For each document $i \in [n]$, we generate the topics W_i from a specified distribution.
4. We sample N words from Multinomial$_p(N; A^{(0)}W)$.

6.2.1. NIPs corpus

After this preprocessing stop, the NIPs data set consists of $n = 1,500$ documents with dictionary size $p = 1,253$ and mean document length 847.

1. We set $N = 850$ and vary $n \in \{2000, 4000, 6000, 8000, 10000\}$ for generating semi-synthetic data.

2. While the estimated $A^{(0)}$ from LDA does not have exact zero entries, we calculate *the approximate sparsity level* of A by

$$\textbf{sparsity} = \frac{1}{pK}\sum_{j=1}^{p}\sum_{k=1}^{K} 1\{A_{jk} \geq 10^{-3}p^{-1}\} \approx 0.696. \tag{30}$$

   The calculated **sparsity** indicates that the posterior $A^{(0)}$ from LDA has many entries close to 0.

3. As in Arora et al. (2013), we manually add $|I_k| = m$ anchor words for each topic with $m \in \{1, 5\}$. After adding m anchor words, we re-normalize the columns to obtain $A^{(m)}$.

4. The columns of W are generated from the symmetric Dirichlet distribution with parameter 0.03. We sample N words from $\text{Multinomial}_p(N; A^{(m)}W)$.

For each combination of n and m, we generate 20 data sets and the average estimation errors $\|\widehat{A} - A\|_1/K$ of different algorithms are shown in Figure 2. The bars represent the standard deviations across 20 repetitions. Again, LDA, AWR and STM-AWR are given the correct K, while TOP and STM-TOP estimate K.

Conclusion. STM-TOP has best overall performance and STM-AWR has the second best result. LDA is dominated by all other algorithms, although increasing the number of iterations might boost the performance of LDA. Both STM-TOP and TOP have better performance when one has more anchor words.

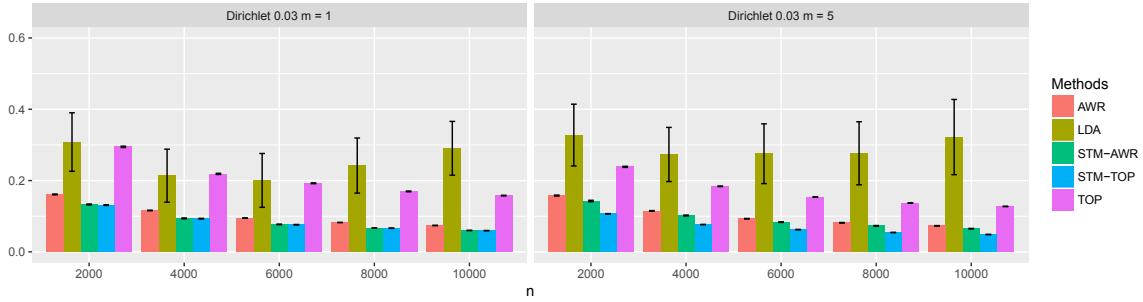

Figure 2: Plots of the estimation errors $\|\widehat{A} - A\|_1/K$

We also investigate the effect of the correlation among topics on the estimation of A. Following Arora et al. (2013), we simulate W from a log-normal distribution with block diagonal covariance matrix and different within-block correlation. To construct the block

diagonal covariance structure, we divide 100 topics into 10 groups. For each group, the off-diagonal elements of the covariance matrix of topics is set to ρ, while the diagonal entries are set to 1. The parameter $\rho \in \{0.03, 0.3\}$ reflects the magnitude of correlation among topics. We take the case $m = 1$ and the estimation errors of the algorithms are shown in Figure 3.

Conclusion. STM-TOP has the best performance in all settings. As long as the number of documents n is large, STM-AWR is more robust to the correlation among topics than AWR. LDA and AWR are comparable.

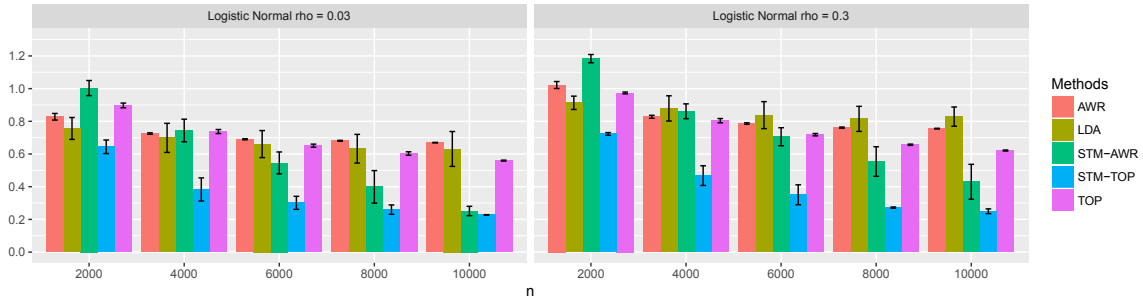

Figure 3: Plots of the estimation errors $\|\widehat{A} - A\|_1/K$ for $\rho = 0.03$ and $\rho = 0.3$.

Finally, we report the running times of the various algorithms in Table 1. As one can see, LDA is the slowest and does not scale well with n. On the other hand, TOP is the fastest and the other three algorithms (AWR, STM-AWR and STM-TOP) have comparable running times.

Table 1: Running time (seconds) of different algorithms.

	TOP	STM-TOP	AWR	STM-AWR	LDA
n = 2000	35.2	614.3	393.8	500.7	1918.7
n = 4000	32.8	611.2	447.0	466.2	3724.5
n = 6000	41.8	610.9	455.0	416.7	5616.6
n = 8000	44.7	605.1	458.4	463.5	7358.8
n = 10000	52.0	609.0	482.8	517.9	9130.6

6.2.2. New York Times (NYT) data set

After the same preprocessing step, the NYT data set contains $n = 299,419$ documents with dictionary size $p = 3,079$ and mean document length 210. We choose $N = 300$ and vary $n \in \{30000, 40000, \ldots, 70000\}$. The estimated $A^{(0)}$ from LDA has **sparsity** ≈ 0.679 calculated from (30). As in the NIPs corpus earlier, we manually add $|I_k| = m \in \{1, 5\}$ anchor words per topic. For each m and n, we generate 20 data sets where columns of W are generated from the symmetric Dirichlet distribution with parameter 0.03. The average estimation errors $\|\widehat{A} - A\|_1/K$ are shown in Figure 4. We also study the effect of correlation among topics on the estimation errors for the case $m = 1$ and with the columns of W generated from the log-normal distribution with block diagonal correlation and $\rho = \{0.1, 0.3\}$.

The result is shown in Figure 5.

Conclusion. From Figure 4, in the presence of anchor words, we see that STM-TOP has the best overall performance and STM-AWR outperforms AWR. The errors of STM-TOP and TOP decrease if more anchor words are introduced. In Figure 5, STM-TOP outperforms the other algorithms in all cases. TOP has the second best performance while the other three algorithms are comparable.

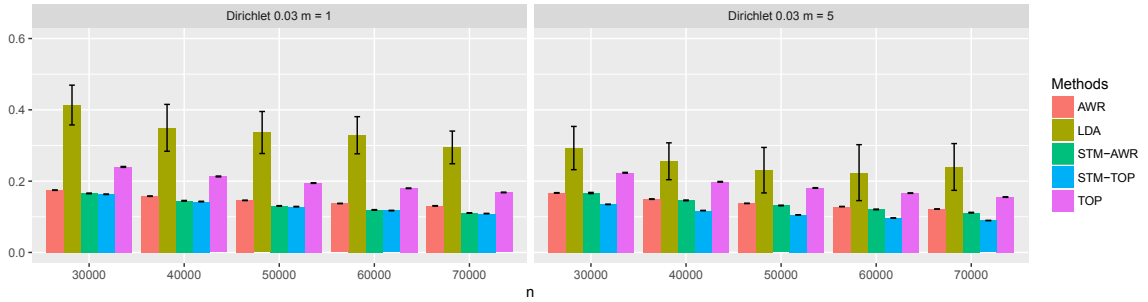

Figure 4: Plots of the estimation errors $\|\widehat{A} - A\|_1/K$

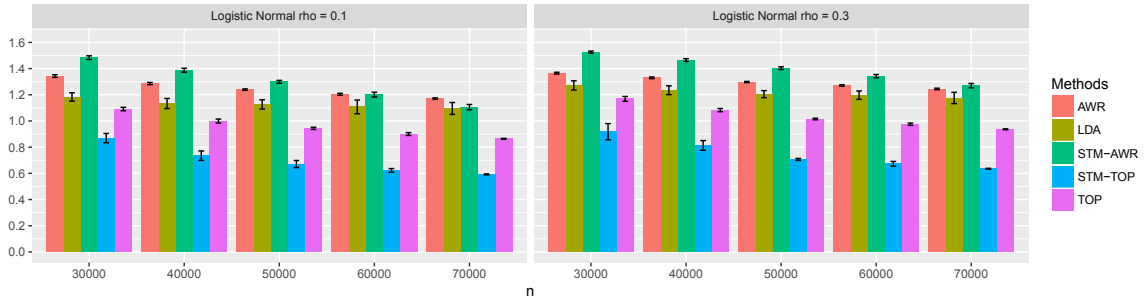

Figure 5: Plots of the estimation errors $\|\widehat{A} - A\|_1/K$ for $\rho = 0.1$ and $\rho = 0.3$.

7. Conclusion

We have studied estimation of the word-topic matrix A when it is possibly entry-wise sparse and the number of topics K is unknown, under the *separability* condition. A new minimax lower bound of $\|\widehat{A} - A\|_1$ is derived and a computationally efficient procedure (STM) for estimating A is proposed. The estimator provably achieves the minimax lower bound (modulo a logarithmic factor) and adapts to the unknown sparsity. Extensive simulations corroborate the superior performance of our new estimation procedure in tandem with the existing algorithm in Bing et al. (2020) for selecting anchor words.

Acknowledgments

Bunea and Wegkamp are supported in part by NSF grants DMS-1712709 and DMS-2015195.

Appendix A. Proofs

The proofs rely on some lemmas in Bing et al. (2020). For the reader's convenience, we restate them in Section A.2 and use similar notations for simplicity.

A.1. Notations and two useful expressions

From the topic model specifications, the matrices Π, A and W are all scaled as

$$\sum_{j=1}^{p} \Pi_{ji} = 1, \quad \sum_{j=1}^{p} A_{jk} = 1, \quad \sum_{k=1}^{K} W_{ki} = 1 \tag{31}$$

for any $1 \le j \le p$, $1 \le i \le n$ and $1 \le k \le K$. In order to adjust their scales properly, we denote

$$m_j = p \max_{1\le i\le n} \Pi_{ji}, \quad \mu_j = \frac{p}{n}\sum_{i=1}^{n} \Pi_{ji}, \quad \alpha_j = p \max_{1\le k\le K} A_{jk}, \quad \gamma_k = \frac{K}{n}\sum_{i=1}^{n} W_{ki}, \tag{32}$$

so that

$$\sum_{k=1}^{K} \gamma_k = K, \qquad \sum_{j=1}^{p} \mu_j = p. \tag{33}$$

Recall that $\rho_j = \alpha_j/\underline{\alpha}_L$ and $\widetilde{s}_J := \sum_{j\in L^c} \rho_j s_j$. We define

$$\frac{\widehat{\mu}_j}{p} = \frac{1}{n}\sum_{t=1}^{n} X_{jt}, \qquad \text{for all } 1 \le j \le p. \tag{34}$$

We write $d := n \vee p$ throughout the proof. Finally, note that Assumption 3 implies $K < n$.

From model specifications (31) and (32), we derive three useful facts that are later repeatedly invoked.

(a) For any $j \in [p]$, by using (32),

$$\mu_j = \frac{p}{n}\sum_{i=1}^{n} \Pi_{ji} = \frac{p}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} A_{jk}W_{ki} = \frac{p}{K}\sum_{k=1}^{K} A_{jk}\gamma_k \quad \Rightarrow \quad \frac{p}{K}\sum_{k=1}^{K} A_{jk}\cdot\underline{\gamma} \le \mu_j \le \alpha_j. \tag{35}$$

In particular, for any $j \in I_k$ with any $k \in [K]$,

$$\mu_j = \frac{p}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} A_{jk}W_{ki} = \frac{p}{K}A_{jk}\gamma_k \overset{(32)}{=} \frac{\alpha_j\gamma_k}{K}. \tag{36}$$

(b) For any $j \in [p]$,

$$m_j \overset{(32)}{=} p \max_{1\le i\le n} \Pi_{ji} = p \max_{1\le i\le n}\sum_{k=1}^{K} A_{jk}W_{ki} \le p \max_{1\le k\le K} A_{jk} \overset{(32)}{=} \alpha_j \quad \Rightarrow \quad \mu_j \le m_j \le \alpha_j, \tag{37}$$

by using $0 \le W_{ki} \le 1$ and $\sum_k W_{ki} = 1$ for any $k \in [K]$ and $i \in [n]$.

(c) For any $j \in [p]$ and $k \in [K]$, define

$$\psi_{jk} = \sum_{a=1}^{K} A_{ja} C_{ak} \text{ with } C = n^{-1} W W^\top. \tag{38}$$

We have

$$\sum_{j=1}^{p} \psi_{jk} = \sum_{j=1}^{p} \sum_{a=1}^{K} A_{ja} C_{ak} = \sum_{a=1}^{K} C_{ak} = \frac{1}{n} \sum_{t=1}^{n} \sum_{a=1}^{K} W_{kt} W_{at} \overset{(31)}{=} \frac{1}{n} \sum_{t=1}^{n} W_{kt} \overset{(32)}{=} \frac{\gamma_k}{K}. \tag{39}$$

A.2. Useful results from Bing et al. (2020)

Let $\varepsilon_{ji} := X_{ji} - \Pi_{ji}$, for $1 \le i \le n$ and $1 \le j \le p$ and assume $N_1 = \ldots = N_n = N$ for ease of presentation since similar results for different N can be derived by using the same arguments.

Lemma 8 *With probability $1 - 2d^{-1}$, we have*

$$\frac{1}{n} \left| \sum_{i=1}^{n} \varepsilon_{ji} \right| > 2 \sqrt{\frac{\mu_j \log(d)}{npN}} + \frac{4 \log(d)}{nN}, \quad \textit{uniformly in } 1 \le j \le p. \tag{40}$$

If $\min_{1 \le j \le p} \mu_j / p \ge \log(d)/(nN)$ holds, with probability $1 - 2d^{-1}$,

$$\frac{1}{n} \left| \sum_{i=1}^{n} \varepsilon_{ji} \right| \le 6 \sqrt{\frac{\mu_j \log(d)}{npN}}, \quad \textit{uniformly in } 1 \le j \le p.$$

Lemma 9 *Recall $\Theta = n^{-1} \Pi \Pi^\top$. With probability $1 - 2d^{-1}$,*

$$\frac{1}{n} \left| \sum_{i=1}^{n} \Pi_{\ell i} \varepsilon_{ji} \right| \le \sqrt{\frac{6 m_\ell \Theta_{j\ell} \log(d)}{npN}} + \frac{2 m_\ell \log(d)}{npN}, \quad \textit{uniformly in } 1 \le j, \ell \le p.$$

Lemma 10 *If $\min_{1 \le j \le p} \mu_j / p \ge 2 \log(d)/(3N)$, then with probability $1 - 4d^{-1}$,*

$$\frac{1}{n} \left| \sum_{i=1}^{n} (\varepsilon_{ji} \varepsilon_{\ell i} - \mathbb{E}[\varepsilon_{ji} \varepsilon_{\ell i}]) \right| \le 12\sqrt{6} \sqrt{\Theta_{j\ell} + \frac{(\mu_j + \mu_\ell) \log(d)}{pN}} \sqrt{\frac{\log^3(d)}{nN^2}} + 4d^{-3},$$

holds, uniformly in $1 \le j, \ell \le p$.

Lemma 11 *Assume model (4) and*

$$\min_{1 \le j \le p} \frac{1}{n} \sum_{i=1}^{n} \Pi_{ji} \ge \frac{c \log(d)}{nN} \tag{41}$$

for some sufficiently large constant $c > 0$. With probability greater than $1 - O(d^{-1})$,

$$|\widehat{\Theta}_{j\ell} - \Theta_{j\ell}| \le c_0 \eta_{j\ell}, \qquad |\widehat{R}_{j\ell} - R_{j\ell}| \le c_1 \delta_{j\ell}, \quad \textit{for all } 1 \le j, \ell \le p$$

for some constant $c_0, c_1 > 0$, where

$$\eta_{j\ell} = \sqrt{\frac{\Theta_{j\ell}\log(d)}{nN}}\sqrt{\frac{m_j+m_\ell}{p} \vee \frac{\log^2(d)}{N}} + \frac{(m_j+m_\ell)}{p}\frac{\log(d)}{nN} + \sqrt{\frac{\log^4(d)}{nN^3}}\sqrt{\frac{\mu_j+\mu_\ell}{p} \vee \frac{\log(d)}{N}} \tag{42}$$

and

$$\delta_{j\ell} := \frac{p^2\eta_{j\ell}}{\mu_j\mu_\ell} + \frac{p^2\Theta_{j\ell}}{\mu_j\mu_\ell}\left(\sqrt{\frac{p}{\mu_j}} + \sqrt{\frac{p}{\mu_\ell}}\right)\sqrt{\frac{\log(d)}{nN}}. \tag{43}$$

A.3. Proof of Theorem 1 in Section 2

We first choose $\{I_1, \ldots, I_K\}$ such that $||I_k| - |I_{k'}|| \le 1$ for any $k, k' \in [K]$. This also implies $|I_k| \le 2|I|/K$. Further choose the integer set $\{g_1, \ldots, g_K\}$ such that $\sum_{k=1}^K g_k = s_J$ and $|g_k - g_{k'}| \le 1$ for any $k, k' \in [K]$, further implying $g_k \le 2s_J/K$. We first choose $A^{(0)}$. Let

$$\widetilde{A}^{(0)} = \begin{bmatrix} \mathbf{1}_{|I_1|} & & & \\ & \mathbf{1}_{|I_2|} & & \\ & & \ddots & \\ & & & \mathbf{1}_{|I_K|} \\ \widetilde{\mathbf{1}}_{g_1} & \widetilde{\mathbf{1}}_{g_2} & \cdots & \widetilde{\mathbf{1}}_{g_K} \end{bmatrix} \tag{44}$$

where, for any $k \in [K]$, $\widetilde{\mathbf{1}}_{g_k} = \mathbf{1}_{g_k}$ if $g_k = |J|$ and $\widetilde{\mathbf{1}}_{g_k} = (\mathbf{1}_{g_k}^\top, 0^\top)^\top$ otherwise. We then set

$$A^{(0)} = \widetilde{A}^{(0)} \begin{bmatrix} \frac{1}{|I_1|+g_1} & & & \\ & \frac{1}{|I_2|+g_2} & & \\ & & \ddots & \\ & & & \frac{1}{|I_K|+g_K} \end{bmatrix}.$$

We start by constructing a set of "hypotheses" of A. Assume $|I_k|+g_k$ is even for $1 \le k \le K$. Let

$$\mathcal{M} := \{0,1\}^{(|I|+s_J)/2}.$$

Following the Varshamov-Gilbert bound in Lemma 2.9 in Tsybakov (2009), there exists $w^{(j)} \in \mathcal{M}$ for $j = 0, 1, \ldots, T$, such that

$$\left\|w^{(i)} - w^{(j)}\right\|_1 \ge \frac{|I|+s_J}{16}, \quad \text{for any } 0 \le i \ne j \le T, \tag{45}$$

with $w^{(0)} = 0$ and

$$\log(T) \ge \frac{\log(2)}{16}(|I|+s_J). \tag{46}$$

For each $w^{(j)} \in \mathcal{M}$, we divide it into K chunks as $w^{(j)} = \left(w_1^{(j)}, w_2^{(j)}, \ldots, w_K^{(j)}\right)$ with $w_k^{(j)} \in \mathbb{R}^{(|I_k|+g_k)/2}$. For each $w_k^{(j)}$, we write $\widetilde{w}_k^{(j)} \in \mathbb{R}^p$ as its augumented counterpart such that

$[\widetilde{w}_k^{(j)}]_{S_k} = [w_k^{(j)}, -w_k^{(j)}]$ and $[\widetilde{w}_k^{(j)}]_\ell = 0$ for any $\ell \notin S_k$, where $S_k := \mathrm{supp}(A_k^{(0)})$. For $1 \le j \le T$, we choose $A^{(j)}$ as

$$A^{(j)} = A^{(0)} + \gamma \left[\widetilde{w}_1^{(j)} \quad \cdots \quad \widetilde{w}_K^{(j)} \right] \tag{47}$$

with

$$\gamma = \sqrt{\frac{\log(2)}{4^5(1+c_0)}} \sqrt{\frac{K^2}{nN(|I|+s_J)}} \tag{48}$$

for some constant $c_0 > 0$. Under $|I| + s_J \le nN$, it is easy to verify that $A^{(j)} \in \mathcal{A}(|I|, s_J)$ for all $0 \le j \le T$.

In order to apply Theorem 2.5 in Tsybakov (2009), we need to check the following conditions:

(a) $\mathrm{KL}(\mathbb{P}_{A^{(j)}}, \mathbb{P}_{A^{(0)}}) \le \log(T)/16$, for each $i = 1, \ldots, T$.

(b) $\|A^{(i)} - A^{(j)}\|_1 \ge c_1 K \sqrt{(|I|+s_J)/(nN)}$, for $0 \le i < j \le T$ and some constant $c_1 > 0$.

We first show part (a). Fix $1 \le j \le T$ and choose $D^{(j)} = A^{(j)} W^0$ where W^0 is defined in (8). Let m_k be the set such that $|m_k| = n_k$ and $W_i^0 = e_k$, for all $i \in m_k$ and $k \in [K]$. Since $|I_k| + g_k \le 2(|I| + s_J)/K$, it follows that

$$D_{\ell i}^{(0)} = \sum_{k=1}^K A_{\ell k}^{(0)} W_{ki}^0 = \begin{cases} 1/(|I_k| + g_k) \ge 2^{-1} K/(|I| + s_J), & \text{if } \ell \in S_k, i \in m_k, k \in [K] \\ 0, & \text{otherwise} \end{cases}. \tag{49}$$

for any $i \in [n]$ and $\ell \in [p]$. Similarly, we have

$$\left| D_{\ell i}^{(j)} - D_{\ell i}^{(0)} \right| = \gamma \left| \sum_{k=1}^K [\widetilde{w}_k^{(j)}]_\ell W_{ki}^0 \right| \le \begin{cases} \gamma, & \text{if } \ell \in S_k, i \in m_k, k \in [K] \\ 0, & \text{otherwise} \end{cases}. \tag{50}$$

Thus, by $|I| + s_J \le nN$, we have

$$\max_{(\ell,i) \in \mathcal{T}^c} \frac{|D_{\ell i}^{(j)} - D_{\ell i}^{(0)}|}{D_{\ell i}^{(0)}} \le 2\gamma \frac{|I| + s_J}{K} < 1, \qquad \text{for any } 1 \le j \le T$$

where $\mathcal{T} := \{(\ell, i) \in [p] \times [n] : D_{\ell i}^{(0)} = 0\}$ and $\mathcal{T}^c := [p] \times [n] \setminus \mathcal{T}$. Observe that $D_{\ell i}^{(j)} = 0$ for any $(\ell, i) \in \mathcal{T}$ and $1 \leq j \leq T$, and invoke Lemma 12 to get

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{A^{(j)}}, \mathbb{P}_{A^{(0)}}) &\leq (1+c_0)\, N \sum_{(\ell,i)\in\mathcal{T}} \frac{|D_{\ell i}^{(j)} - D_{\ell i}^{(0)}|^2}{D_{\ell i}^{(0)}} \\
&\leq (1+c_0)\, N \sum_{k=1}^{K} \sum_{i\in m_k} \sum_{\ell \in S_k} \gamma^2 (|I_k| + g_k) \\
&= (1+c_0)\, N \sum_{k=1}^{K} \sum_{i\in m_k} \gamma^2 (|I_k| + g_k)^2 \qquad (\text{by } |S_k| = |I_k| + g_k) \\
&\leq 4\,(1+c_0)\, N n \gamma^2 \frac{(|I| + s_J)^2}{K^2} \\
&\overset{(46)}{\leq} \frac{1}{16} \log T.
\end{aligned}$$

The second inequality uses (49) and (50) and the fourth line uses $|I_k| + g_k \leq 2(|I| + s_J)/K$. This verifies (a).

To show (b), (47) yields

$$\begin{aligned}
\|A^{(j)} - A^{(\ell)}\|_1 &= \sum_{k=1}^{K} \left\| A_{\cdot k}^{(j)} - A_{\cdot k}^{(\ell)} \right\|_1 \\
&= 2\gamma \sum_{k=1}^{K} \left\| w_k^{(j)} - w_k^{(\ell)} \right\|_1 \\
&= 2\gamma \left\| w^{(j)} - w^{(\ell)} \right\|_1 \\
&\overset{(45)}{\geq} \frac{\gamma}{8} (|I| + s_J).
\end{aligned}$$

After we plug this into the expression of γ, we obtain (b). Invoking (Tsybakov, 2009, Theorem 2.5) concludes the proof when $|I_k| + g_k$ is even for all $k \in [K]$. The complementary case is easy to derive with slight modifications. Specifically, denote by $\mathcal{S}_{odd} := \{1 \leq k \leq K : |I_k| + g_k \text{ is odd}\}$. Then we change $\mathcal{M} := \{0,1\}^{Card}$ with

$$Card = \sum_{k\in\mathcal{S}_{odd}} \frac{|I_k| + g_k - 1}{2} + \sum_{k\in\mathcal{S}_{odd}^c} \frac{|I_k| + g_k}{2}.$$

For each $w^{(j)}$, we write it as $w^{(j)} = (w_1^{(j)}, \ldots, w_K^{(j)})$ and each $w_k^{(j)}$ has length $(|I_k| + g_k - 1)/2$ if $k \in \mathcal{S}_{odd}$ and $(|I_k| + g_k)/2$ otherwise. We then construct $A_k^{(j)} = A_k^{(0)} + \gamma \widetilde{w}_k^{(j)}$ where $\widetilde{w}_k^{(j)} \in \mathbb{R}^p$ is the same augumented ccounterpart of $w_k^{(j)}$. The result follows from the same arguments and the proof is complete. ■

The upper bound of Kullback-Leibler divergence between two multinomial distributions is studied in (Ke and Wang, 2017, Lemma 6.7). We use the following modification of their bound.

Lemma 12 *Let D and D' be two $p \times n$ matrices such that each column of them is a weight vector. Under model (4), let $\mathbb{P}$ and $\mathbb{P}'$ be the probability measures associated with D and D', respectively. Let $\mathcal{T}$ be the set such that*

$$\mathcal{T} := \big\{(j,i) \in [p] \times [n] : D_{ji} = D'_{ji} = 0\big\}$$

Let $\mathcal{T}^c := ([p] \times [n]) \setminus \mathcal{T}$ and

$$\eta = \max_{(j,i)\in\mathcal{T}^c} \frac{|D'_{ji} - D_{ji}|}{D_{ji}}$$

and assume $\eta < 1$. There exists a universal constant $c_0 > 0$ such that

$$KL(\mathbb{P}', \mathbb{P}) \le (1 + c_0\eta)N \sum_{(j,i)\in\mathcal{T}^c} \frac{|D'_{ji} - D_{ji}|^2}{D_{ji}}.$$

Proof With the convention that $0/0 = 1$, we have

$$KL(\mathbb{P}', \mathbb{P}) = N \sum_{i=1}^{n}\sum_{j=1}^{p} D'_{ji} \log\left(\frac{D'_{ji}}{D_{ji}}\right) = N \sum_{(j,i)\in\mathcal{T}^c} D'_{ji} \log\left(1 + \eta_{ji}\right).$$

Then the proof follows by the same arugments in Ke and Wang (2017). ■

A.4. Proofs of Section 4

We first give the proof of Lemma 3 and then prove our main Theorem 2.

A.4.1. Proof of Lemma 3

From (18), we have

$$\widehat{M}_{ab} = \frac{1}{|L_a||L_b|} \sum_{i\in L_a, j\in L_b} \widehat{R}_{ij}.$$

Further notice that

$$\frac{1}{|L_a||L_b|} \sum_{i\in L_a, j\in L_b} R_{ij} = M_{ab}.$$

Using the fact that $\|Q\|_{\mathrm{op}} \le \|Q\|_{\infty,1}$ for any symmetric matrix Q, yields

$$\begin{aligned}
\|\widehat{M} - M\|_{\mathrm{op}} &\le \|\widehat{M} - M\|_{\infty,1} \\
&= \max_{1\le k\le K} \sum_{a=1}^{K} \left| \frac{1}{|L_a||L_b|} \sum_{i\in L_a, j\in L_b} (\widehat{R}_{ij} - R_{ij}) \right| \\
&\le \max_{1\le k\le K} \max_{i\in L_k} \sum_{a=1}^{K} \max_{j\in L_a} |\widehat{R}_{ij} - R_{ij}|.
\end{aligned}$$

Invoking Lemma 11 for all $i, j \in L$ under condition (24), with probability $1 - O(d^{-1})$, we have

$$\|\widehat{M} - M\|_{\text{op}} \leq \max_{1\leq k\leq K} \max_{i\in L_k} \sum_{a=1}^{K} \max_{j\in L_a} \delta_{ij}.$$

The result follows by invoking Lemma 14. ■

A.4.2. Proof of Theorem 2

As our estimation procedure uses a thresholding step in (19), we first define

$$T := \left\{ j \in L^c : \frac{1}{n}\sum_{i=1}^{n} \Pi_{ji} < \frac{\log(d)}{nN} \right\}, \quad \widehat{T} := \left\{ j \in L^c : \frac{1}{n}\sum_{i=1}^{n} X_{ji} < \frac{7\log(d)}{nN} \right\} \tag{51}$$

and write $T^c := [p] \setminus T$ and $\widehat{T}^c := [p] \setminus \widehat{T}$.
Recall that our final estimator $\widehat{A}$ is obtained by normalizing $\widehat{\bar{B}} = D_X \widehat{B}$ to unit column sums with $D_X = \text{diag}\,(\widehat{u}_1/p, \ldots, \widehat{u}_p/p)$ where $\widehat{u}_j/p$ is defined in (34) for $1 \leq j \leq p$. For any $j \in [p]$ and $k \in [K]$, we have

$$\widehat{A}_{jk} - A_{jk} = \frac{\widehat{\bar{B}}_{jk}}{\|\widehat{\bar{B}}_k\|_1} - \frac{\bar{B}_{jk}}{\|\bar{B}_k\|_1}$$

where $\bar{B} = D_\Pi B$. Summing over $1 \leq j \leq p$ yields

$$\begin{aligned}
\|\widehat{A}_k - A_k\|_1 &= \sum_{j=1}^{p} \left| \frac{\widehat{\bar{B}}_{jk}}{\|\widehat{\bar{B}}_k\|_1} - \frac{\widehat{\bar{B}}_{jk}}{\|\bar{B}_k\|_1} + \frac{\widehat{\bar{B}}_{jk} - \bar{B}_{jk}}{\|\bar{B}_k\|_1} \right| \\
&\leq \frac{|\|\bar{B}_k\|_1 - \|\widehat{\bar{B}}_k\|_1|}{\|\bar{B}_k\|_1} + \frac{\|\widehat{\bar{B}}_k - \bar{B}_k\|_1}{\|\bar{B}_k\|_1} \\
&\leq \frac{2\|\widehat{\bar{B}}_k - \bar{B}_k\|_1}{\|\bar{B}_k\|_1} \\
&= \frac{2K}{\gamma_k} \|\widehat{\bar{B}}_k - \bar{B}_k\|_1.
\end{aligned}$$

In the last equality, we use

$$\|\bar{B}_k\|_1 = \sum_{j=1}^{p} A_{jk} \frac{1}{n}\sum_{t=1}^{n} W_{kt} = \frac{\gamma_k}{K},$$

by observing that $\bar{B} = D_\Pi B = AD_W$. Further recall that $\widehat{\bar{B}}_{jk} = \widehat{\mu}_j \widehat{B}_{jk}/p$ for $j \in [p]$ and $\widehat{B}_{j\cdot} = 0$ for any $j \in \widehat{T}$. We have

$$\begin{aligned}
\|\widehat{A}_k - A_k\|_1 &= \frac{2K}{\gamma_k} \sum_{j=1}^{p} \left| \frac{\widehat{\mu}_j}{p} \widehat{B}_{jk} - \frac{\mu_j}{p} B_{jk} \right| \\
&\leq \frac{2K}{\gamma_k} \sum_{j=1}^{p} \left\{ \widehat{B}_{jk} \frac{|\widehat{\mu}_j - \mu_j|}{p} + \frac{\mu_j}{p} |\widehat{B}_{jk} - B_{jk}| \right\} \\
&= \frac{2K}{\gamma_k} \left\{ \sum_{j \in \widehat{T}^c} \left(\widehat{B}_{jk} \frac{|\widehat{\mu}_j - \mu_j|}{p} + \frac{\mu_j}{p} |\widehat{B}_{jk} - B_{jk}| \right) + \sum_{j \in \widehat{T}} \frac{\mu_j B_{jk}}{p} \right\} \\
&= \frac{2K}{\gamma_k} \sum_{j \in \widehat{T}^c} \left(\widehat{B}_{jk} \frac{|\widehat{\mu}_j - \mu_j|}{p} + \frac{\mu_j}{p} |\widehat{B}_{jk} - B_{jk}| \right) + 2 \sum_{j \in \widehat{T}} A_{jk}.
\end{aligned}$$

We use $A_{jk} = (\mu_j/p) B_{jk} (K/\gamma_k)$ in the last line. Summing over $1 \leq k \leq K$ gives

$$\begin{aligned}
\|\widehat{A} - A\|_1 &\leq \frac{2K}{\underline{\gamma}} \sum_{j \in \widehat{T}^c} \left(\frac{|\widehat{\mu}_j - \mu_j|}{p} + \frac{\mu_j}{p} \|\widehat{B}_{j\cdot} - B_{j\cdot}\|_1 \right) + 2 \sum_{j \in \widehat{T}} \|A_{j\cdot}\|_1 \\
&= \frac{2K}{\underline{\gamma}} \left\{ \sum_{j \in \widehat{T}^c} \frac{|\widehat{\mu}_j - \mu_j|}{p} + \sum_{j \in \widehat{T}^c \setminus L} \frac{\mu_j}{p} \|\widehat{B}_{j\cdot} - B_{j\cdot}\|_1 \right\} + 2 \sum_{j \in \widehat{T}} \|A_{j\cdot}\|_1.
\end{aligned}$$

We use $\|\widehat{B}_{jk}\|_1 = 1$ in the first line and the fact $\widehat{B}_{j\cdot} = B_{j\cdot}$ for all $j \in L$ in the second line.

Next, we study the three terms on the right hand side. To bound the first term, we observe that

$$\mathbb{P}\{T \subseteq \widehat{T}\} = \mathbb{P}\{\widehat{T}^c \subseteq T^c\} = 1 - 2d^{-1},$$

by Lemma 13. This fact, the second part of Lemma 8 and the inequality

$$\min_{j \in T^c} \frac{\mu_j}{p} \geq \frac{\log(d)}{nN}. \tag{52}$$

yield

$$\mathbb{P}\left\{ \sum_{j \in \widehat{T}^c} \frac{|\widehat{\mu}_j - \mu_j|}{p} \leq \sum_{j \in T^c} \frac{|\widehat{\mu}_j - \mu_j|}{p} \leq \sum_{j \in T^c} 6 \sqrt{\frac{\mu_j \log(d)}{npN}} \right\} \geq 1 - 4d^{-1}.$$

Further, the Cauchy-Schwarz inequality and $\sum_{j \in T^c} \mu_j/p \leq \sum_{j=1}^{p} \mu_j/p = 1$ yield

$$\mathbb{P}\left\{ \sum_{j \in \widehat{T}^c} \frac{|\widehat{\mu}_j - \mu_j|}{p} \leq 6 \sqrt{\frac{|T^c| \log(d)}{nN}} \leq 6 \sqrt{\frac{p \log(d)}{nN}} \right\} \geq 1 - 4d^{-1}. \tag{53}$$

To bound the third term, Lemma 13 yields

$$\mathbb{P}\left\{\sum_{j\in\widehat{T}}\|A_{j\cdot}\|_1 \le \frac{20K|\widehat{T}|\log(d)}{\underline{\gamma}nN} \le \frac{20Kp\log(d)}{\underline{\gamma}nN}\right\} \ge 1-2d^{-1}. \tag{54}$$

The proof of the upper bound for the second term is more involved. We work on the intersection of the event $\{\widehat{T}^c \subseteq T^c\}$ with

$$\mathcal{E}_M := \left\{\lambda_{\min}(\widehat{M}+\lambda \boldsymbol{I}_K) \ge \lambda_{\min}(M)+\lambda-\|\widehat{M}-M\|_{\mathrm{op}} \ge \lambda_{\min}(M)\right\}$$

to establish an upper bound for

$$\sum_{j\in T^c\setminus L}\frac{\mu_j}{p}\|\widehat{B}_{j\cdot}-B_{j\cdot}\|_1.$$

Lemma 3 and the choice of λ guarantee $\mathbb{P}(\mathcal{E}_M)=1-O(d^{-1})$. Pick any $j\in T^c\setminus L$ and recall that $\widehat{B}_{j\cdot}$ is estimated via (20). Starting with

$$\widehat{B}_{j\cdot}^\top(\widehat{M}+\lambda\boldsymbol{I}_K)\widehat{B}_{j\cdot}-2\widehat{B}_{j\cdot}^\top\widehat{h}^{(j)} \le B_{j\cdot}^\top(\widehat{M}+\lambda\boldsymbol{I}_K)^{-1}B_{j\cdot}-2B_{j\cdot}^\top\widehat{h}^{(j)},$$

standard arguments yield

$$\begin{aligned}(\Delta^{(j)})^\top(\widehat{M}+\lambda\boldsymbol{I}_K)\Delta^{(j)} &\le 2\left|(\Delta^{(j)})^\top(\widehat{h}^{(j)}-\widehat{M}B_{j\cdot}-\lambda B_{j\cdot})\right| \\ &\le 2\left\{|(\Delta^{(j)})^\top(\widehat{h}^{(j)}-h^{(j)})|+|(\Delta^{(j)})^\top(h^{(j)}-\widehat{M}B_{j\cdot})|+\lambda\|\Delta^{(j)}\|\|B_{j\cdot}\|\right\}\end{aligned}$$

by writing $\Delta^{(j)}:=\widehat{B}_{j\cdot}-B_{j\cdot}$. Hence, on the event $\mathcal{E}_M$, we have

$$\|\Delta^{(j)}\| \le \frac{2}{\lambda_{\min}(M)}\left\{\frac{|(\Delta^{(j)})^\top(\widehat{h}^{(j)}-h^{(j)})|}{\|\Delta^{(j)}\|}+\frac{|(\Delta^{(j)})^\top(h^{(j)}-\widehat{M}B_{j\cdot})|}{\|\Delta^{(j)}\|}+\lambda\|B_{j\cdot}\|\right\}. \tag{55}$$

Let $s_j=\|B_{j\cdot}\|_0$ and $S_j=\mathrm{supp}(B_{j\cdot})$. Since

$$0=\|B_{j\cdot}\|_1-\|\widehat{B}_{j\cdot}\|_1=\|B_{jS_j}\|_1-\|\widehat{B}_{jS_j}\|_1-\|\widehat{B}_{jS_j^c}\|_1 \le \|\Delta^{(j)}_{S_j}\|_1-\|\Delta^{(j)}_{S_j^c}\|_1,$$

we have

$$\|\Delta^{(j)}\|_1 \le 2\|\Delta^{(j)}_{S_j}\|_1 \le 2\sqrt{s_j}\|\Delta^{(j)}_{S_j}\| \le 2\sqrt{s_j}\|\Delta^{(j)}\|. \tag{56}$$

Combination of (56) with (55) gives

$$\begin{aligned}&\sum_{j\in T^c\setminus L}\frac{\mu_j}{p}\|\widehat{B}_{j\cdot}-B_{j\cdot}\|_1 \\ &\le 2\sum_{j\in T^c\setminus L}\frac{\mu_j}{p}\sqrt{s_j}\|\Delta^{(j)}\| \\ &\le \frac{4}{\lambda_{\min}(M)}\sum_{j\in T^c\setminus L}\sqrt{s_j}\cdot\frac{\mu_j}{p}\left\{\frac{|(\Delta^{(j)})^\top(\widehat{h}^{(j)}-h^{(j)})|}{\|\Delta^{(j)}\|}+\frac{|(\Delta^{(j)})^\top(h^{(j)}-\widehat{M}B_{j\cdot})|}{\|\Delta^{(j)}\|}+\lambda\|B_{j\cdot}\|\right\}\end{aligned} \tag{57}$$

The results of Lemmas 15 and 16 and the inequality $\lambda_{\min}(M) \geq \lambda_{\min} K^2/\overline{\gamma}^2$ give

$$\begin{aligned}
&\sum_{j \in T^c \setminus L} \frac{\mu_j}{p} \|\widehat{B}_{j\cdot} - B_{j\cdot}\|_1 \\
&\lesssim \frac{\overline{\gamma}^2}{K^2 \lambda_{\min}} \left\{ \max\left\{ s_J + |I| - |L|, \widetilde{s}_J \right\} \left(\frac{K \log(d)}{\underline{\gamma} nN} + \sqrt{\frac{p \log^4(d)}{\underline{\mu}_L nN^3}} \right) \right. \\
&\left. + K \sqrt{\max\left\{ s_J + |I| - |L|, \widetilde{s}_J \right\} \frac{\log(d)}{\underline{\gamma} nN}} + \lambda \sum_{j \in T^c \setminus L} \sqrt{s_j} \frac{\mu_j}{p} \|B_{j\cdot}\| \right\}. \qquad (58)
\end{aligned}$$

Finally, (53), (54) and (58) together imply that

$$\begin{aligned}
\|\widehat{A} - A\|_1 &\lesssim \frac{K}{\underline{\gamma}} \sqrt{\frac{p \log(d)}{nN}} + \frac{pK \log(d)}{\underline{\gamma} nN} \\
&+ \frac{\overline{\gamma}^2}{\underline{\gamma} K \lambda_{\min}} \left\{ \max\left\{ s_J + |I| - |L|, \widetilde{s}_J \right\} \left(\frac{K \log(d)}{\underline{\gamma} nN} + \sqrt{\frac{p \log^4(d)}{\underline{\mu}_L nN^3}} \right) \right. \\
&\left. + K \sqrt{\max\left\{ s_J + |I| - |L|, \widetilde{s}_J \right\} \frac{\log(d)}{\underline{\gamma} nN}} + \lambda \sum_{j \in T^c \setminus L} \sqrt{s_j} \frac{\mu_j}{p} \|B_{j\cdot}\| \right\}. \qquad (59)
\end{aligned}$$

holds with probability $1 - O(d^{-1})$. After we invoke the result of Lemma 17, the proof of the first result follows. The second result follows by setting $\lambda = 0$ in (59) as

$$\mathbb{P}\left\{ \lambda_{\min}(\widehat{M}) \geq \lambda_{\min}(M) - \|\widehat{M} - M\|_{\text{op}} \geq c \lambda_{\min}(M) \right\} \geq 1 - O(d^{-1}).$$

A.5. Lemmas used in the proof of Theorem 2

Lemma 13 *Let T and $\widehat{T}$ be defined in (51). With probability $1 - 2d^{-1}$, we have $T \subseteq \widehat{T}$ and, for any $1 \leq j \leq p$, if*

$$\frac{1}{n} \sum_{i=1}^{n} X_{ji} < \frac{7 \log(d)}{nN},$$

we further have

$$\|A_{j\cdot}\|_1 \leq \frac{19K \log(d)}{\underline{\gamma} nN}.$$

Proof Recall that $X_{ji} = \Pi_{ji} + \varepsilon_{ji}$ such that $\widehat{\mu}_j/p = \mu_j/p + n^{-1} \sum_{i=1}^{n} \varepsilon_{ji}$. We work on the event

$$\mathcal{E}_1 := \bigcap_{j=1}^{p} \left\{ \frac{1}{n} \left| \sum_{i=1}^{n} \varepsilon_{ji} \right| < 2 \sqrt{\frac{\mu_j \log(d)}{npN}} + \frac{4 \log(d)}{nN} \right\}$$

which holds with probability $1 - 2d^{-1}$ from Lemma 8. Since, for any $j \in T$,

$$\frac{\widehat{u}_j}{p} \leq \frac{\mu_j}{p} + \frac{|\widehat{\mu}_j - \mu_j|}{p} \overset{\mathcal{E}_1}{<} \frac{\log(d)}{nN} + 2 \sqrt{\frac{\mu_j \log(d)}{npN}} + \frac{4 \log(d)}{nN} < \frac{7 \log(d)}{nN},$$

we have $j \in \widehat{T}$, hence $T \subseteq \widehat{T}$.

To prove the second statement, for any j such that $\widehat{u}_j/p \le 7\log(d)/(nN)$, we have

$$\frac{\mu_j}{p} \le \frac{\widehat{\mu}_j}{p} + \frac{1}{n}\left|\sum_{i=1}^n \varepsilon_{ji}\right| < \frac{7\log(d)}{nN} + \frac{1}{n}\left|\sum_{i=1}^n \varepsilon_{ji}\right|.$$

For this j, since

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^n \varepsilon_{ji}\right| < 2\sqrt{\frac{\mu_j \log(d)}{npN}} + \frac{4\log(d)}{nN}\right\} \ge \mathbb{P}(\mathcal{E}_1) = 1 - 2d^{-1},$$

we have, with probability $1 - 2d^{-1}$,

$$\frac{\mu_j}{p} < 2\sqrt{\frac{\mu_j \log(d)}{npN}} + \frac{11\log(d)}{nN},$$

which implies $\mu_j/p \le 19\log(d)/(nN)$. The result then follows by using (35). ■

Lemma 14 *Let δ_{ij} be defined in (43) for any $i, j \in [p]$. Let ψ_{jk} be defined in (38) for any $j \in [p]$, $k \in [K]$. Under condition (24), we have*

$$\max_{1\le k\le K} \max_{i\in L_k} \sum_{a=1}^K \max_{j\in L_a} \delta_{ij} \lesssim \frac{K}{\underline{\gamma}}\sqrt{\frac{pK\log(d)}{\underline{\mu}_L nN}}$$

and, for any $k \in [K]$ and $j \in [p]$,

$$\sqrt{\frac{p\log(d)}{\underline{\mu}_L nN}} \max_{i\in L_k} \sum_{a=1}^K \frac{A_{ja}\gamma_a}{K} \max_{\ell\in L_a} \delta_{i\ell} \lesssim K\sqrt{\frac{\rho_j \psi_{jk}\log(d)}{\underline{\gamma}\gamma_k nN}} + \sqrt{\|A_{j\cdot}\|_1}\sqrt{\frac{\rho_j \log(d)}{\gamma_k nN}}.$$

For any $j \in [p]$, if

$$\frac{1}{n}\sum_{t=1}^n \Pi_{jt} \ge \frac{c\log(d)}{nN},$$

for some constant $c > 0$, we further have

$$\frac{\mu_j}{p}\max_{i\in L_k}\delta_{ij} \lesssim (1+\rho_j)\frac{K\log(d)}{\gamma_k nN} + \sqrt{\frac{p\log^4(d)}{\underline{\mu}_L nN^3}} + \frac{K}{\gamma_k}\sqrt{(1+\rho_j)\frac{\psi_{jk}\log(d)}{nN}} + \sqrt{\frac{\mu_j}{p}\frac{K\log(d)}{\gamma_k nN}}.$$

Proof For any $i \in L_k$ and $j \in L_a$ with $a, k \in [K]$, we start with the expressions in (42) and (43). Note that (35), (37) and (64) imply

$$\frac{m_i + m_j}{p} \ge \frac{2c_1\log^2(d)}{p}, \quad \frac{\mu_i + \mu_j}{p} = \frac{1}{n}\sum_{t=1}^n(\Pi_{it} + \Pi_{jt}) \ge \frac{2c_0\log(d)}{N}. \tag{60}$$

Also, using $m_i \le \alpha_i$ from (37) and $\mu_i = \alpha_i\gamma_k/K$ from (36), together with

$$\Theta_{ij} = \frac{1}{n}\sum_{t=1}^{n} A_{ik}W_{kt}W_{at}A_{ja} = A_{ik}A_{ja}C_{ka} \overset{(22)}{=} \frac{\alpha_i\alpha_j C_{ka}}{p^2}, \tag{61}$$

we obtain

$$\begin{aligned}\delta_{ij} \lesssim \frac{K^2}{\gamma_k\gamma_a}\Bigg\{&\sqrt{C_{ka}\left(\frac{1}{\alpha_i}+\frac{1}{\alpha_j}\right)\frac{p\log(d)}{nN}} + \left(\frac{1}{\alpha_i}+\frac{1}{\alpha_j}\right)\frac{p\log(d)}{nN} \\ &+ \sqrt{\frac{\alpha_i\gamma_k+\alpha_j\gamma_a}{\alpha_i^2\alpha_j^2}}\sqrt{\frac{p^3\log^4(d)}{KnN^3}} + C_{ka}\left(\sqrt{\frac{1}{\alpha_i\gamma_k}}+\sqrt{\frac{1}{\alpha_j\gamma_k}}\right)\sqrt{\frac{pK\log(d)}{nN}}\Bigg\}.\end{aligned} \tag{62}$$

Using the Cauchy-Schwarz inequality and the fact that

$$\sum_{a=1}^{K} C_{ka} = \frac{1}{n}\sum_{t=1}^{n}\sum_{a=1}^{K} W_{kt}W_{at} = \frac{1}{n}\sum_{t=1}^{n} W_{kt} \overset{(22)}{=} \frac{\gamma_k}{K}, \tag{63}$$

we further have, after a bit of algebra,

$$\max_{i\in L_k}\sum_{a=1}^{K}\max_{j\in L_a}\delta_{ij} \lesssim \frac{K^2}{\underline{\gamma}}\left\{\sqrt{\frac{p\log(d)}{\underline{\alpha}_L\underline{\gamma}nN}} + \frac{pK\log(d)}{\underline{\alpha}_L\underline{\gamma}nN} + \sqrt{\frac{p\log(d)}{\underline{\alpha}_L\underline{\gamma}KnN}} + \sqrt{\frac{p^3K\log^4(d)}{\underline{\alpha}_L^3\underline{\gamma}^2nN^3}}\right\}$$

where we also use $\alpha_i \ge \underline{\alpha}_L$, $\gamma_a \ge \underline{\gamma}$. Note that the first term on the right-hand side dominates the other three as

$$\frac{pK^2\log(d)}{\underline{\alpha}_L\underline{\gamma}nN} \le \frac{1}{c_0}, \quad \frac{p^2K\log^3(d)}{\underline{\alpha}_L^2\underline{\gamma}N^2} \le \frac{p\log^2(d)}{c_0\underline{\alpha}_L N} \le \frac{1}{c_0c_1}$$

by using $K < n$ and the following observation from (24),

$$\frac{\underline{\alpha}_L}{p} \overset{(37)}{\ge} \min_{i\in L}\frac{m_i}{p} \ge \frac{c_1\log^2(d)}{N}, \qquad \frac{\underline{\alpha}_L}{pK} \ge \frac{\underline{\alpha}_L\underline{\gamma}}{pK} \overset{(36)}{=} \frac{\underline{\mu}_L}{p} \ge \frac{c_0\log(d)}{N}. \tag{64}$$

The first result then follows by using $\underline{\mu}_L = \underline{\alpha}_L\underline{\gamma}/K$ from (36).

To prove the second result, we argue

$$\begin{aligned}&\sum_{a=1}^{K}\frac{A_{ja}\gamma_a}{K}\max_{i\in L_k,\ell\in L_a}\delta_{i\ell} \\ &\lesssim \frac{K}{\gamma_k}\sum_{a=1}^{K}A_{ja}\left\{\sqrt{\frac{C_{ka}p\log(d)}{\underline{\alpha}_L nN}} + \sqrt{\frac{p^3\overline{\gamma}\log^4(d)}{K\underline{\alpha}_L^3nN^3}} + C_{ka}\sqrt{\frac{pK\log(d)}{\underline{\alpha}_L\underline{\gamma}nN}} + \frac{p\log(d)}{\underline{\alpha}_L nN}\right\} \\ &\lesssim \frac{K}{\gamma_k}\sum_{a=1}^{K}A_{ja}\left\{\sqrt{\frac{C_{ka}p\log(d)}{\underline{\alpha}_L nN}} + C_{ka}\sqrt{\frac{pK\log(d)}{\underline{\alpha}_L\underline{\gamma}nN}} + \frac{p\log(d)}{\underline{\alpha}_L nN}\right\} \\ &\le \frac{K}{\gamma_k}\sqrt{\frac{\rho_j\psi_{jk}K\log(d)}{nN}} + K\sqrt{\frac{\rho_j\psi_{jk}K\log(d)}{\underline{\gamma}\gamma_k nN}} + \|A_{j\cdot}\|_1\frac{pK\log(d)}{\underline{\alpha}_L\gamma_k nN} \\ &\le 2K\sqrt{\frac{\rho_j\psi_{jk}K\log(d)}{\underline{\gamma}\gamma_k nN}} + \|A_{j\cdot}\|_1\frac{pK\log(d)}{\underline{\alpha}_L\gamma_k nN}\end{aligned}$$

The second line follows from (62), the third line uses

$$\frac{p^2\log^2(d)}{\underline{\alpha}_L^2 n^2 N^2}\Big/\frac{p^3\overline{\gamma}\log^4(d)}{K\underline{\alpha}_L^3 nN^3} = \frac{\underline{\alpha}_L KN}{\overline{\gamma}p\log^2(d)} \overset{(64)}{\geq} c_1\frac{K}{\overline{\gamma}} \geq c_1,$$

and the fourth line uses the Cauchy-Schwarz inequality together with $\rho_j = \alpha_j/\underline{\alpha}_L$, $C_{ka} \leq \gamma_k/K$ and (38). Since

$$\sqrt{\frac{p\log(d)}{\underline{\mu}_L nN}} \overset{(64)}{\geq} \sqrt{\frac{1}{c_0 n}} \geq \sqrt{\frac{1}{c_0 K}},$$

we have

$$\sqrt{\frac{p\log(d)}{\underline{\mu}_L nN}}K\sqrt{\frac{\rho_j\psi_{jk}K\log(d)}{\underline{\gamma}\gamma_k nN}} \leq K\sqrt{\frac{\rho_j\psi_{jk}\log(d)}{c_0\underline{\gamma}\gamma_k nN}}.$$

The result now follows after observing that

$$\begin{aligned}
\|A_{j\cdot}\|_1\frac{pK\log(d)}{\underline{\alpha}_L\gamma_k nN}\sqrt{\frac{p\log(d)}{\underline{\mu}_L nN}} &\leq \|A_{j\cdot}\|_1\frac{pK\log(d)}{\underline{\alpha}_L\underline{\gamma} nN}\sqrt{\frac{pK\log(d)}{\underline{\alpha}_L\gamma_k nN}} \\
&\leq \sqrt{\|A_{j\cdot}\|_1\frac{K\alpha_j}{p}}\frac{p\log(d)}{\underline{\mu}_L nN}\sqrt{\frac{pK\log(d)}{\underline{\alpha}_L\gamma_k nN}} \\
&\overset{(64)}{\leq} \sqrt{\|A_{j\cdot}\|_1\frac{K\alpha_j}{p}}\frac{1}{c_0 n}\sqrt{\frac{pK\log(d)}{\underline{\alpha}_L\gamma_k nN}} \\
&\leq \sqrt{\|A_{j\cdot}\|_1}\sqrt{\frac{\rho_j\log(d)}{\gamma_k nN}} \quad (\text{by } K < n).
\end{aligned}$$

We proceed to prove the third result. Fix any $j \in [p]$ and $i \in L_k$ with $k \in [K]$ and note that (60) still holds by replacing the constants 2 by 1. Since

$$\Theta_{ij} = A_{ik}\frac{1}{n}\sum_{t=1}^{n}W_{kt}\sum_{a=1}^{K}A_{ja}W_{at} \overset{(38)}{=} A_{ik}\psi_{jk}, \tag{65}$$

and $m_j \leq \alpha_j$ (37), $\mu_i = \alpha_i\gamma_k/K$ from (36) and $\rho_j = \alpha_j/\underline{\alpha}_L$, the expressions of (42) and (43) yield

$$\begin{aligned}
\frac{\mu_j}{p}\delta_{ij} \lesssim \frac{K}{\gamma_k}\sqrt{(1+\rho_j)\frac{\psi_{jk}\log(d)}{nN}} + (1+\rho_j)\frac{K\log(d)}{\gamma_k nN} + \sqrt{\frac{p\log^4(d)}{\underline{\mu}_L nN^3}} \\
+ \frac{K}{\gamma_k}\sqrt{\frac{\mu_j}{p}\frac{p^2\log^4(d)}{\underline{\alpha}_L^2 nN^3}} + \frac{K\psi_{jk}}{\gamma_k}\sqrt{\frac{pK\log(d)}{\underline{\alpha}_L\gamma_k nN}} + \frac{K\psi_{jk}}{\gamma_k}\sqrt{\frac{p\log(d)}{\mu_j nN}}.
\end{aligned}$$

We now simplify the three terms in the second line. Since

$$\psi_{jk} = \frac{1}{n}\sum_{t=1}^{n}\sum_{a=1}^{K}A_{ja}W_{at}W_{kt} \leq \frac{1}{n}\sum_{t=1}^{n}\sum_{a=1}^{K}A_{ja}W_{at} = \frac{1}{n}\sum_{t=1}^{n}\Pi_{jt} = \frac{\mu_j}{p},$$

we have

$$\frac{K\psi_{jk}}{\gamma_k}\sqrt{\frac{p\log(d)}{\mu_j nN}} \le \frac{K}{\gamma_k}\sqrt{\frac{\psi_{jk}\log(d)}{nN}}.$$

Also note that (72) yields

$$\frac{K\psi_{jk}}{\gamma_k}\sqrt{\frac{pK\log(d)}{\underline{\alpha}_L\gamma_k nN}} \le \frac{K}{\gamma_k}\sqrt{\frac{\alpha_j\psi_{jk}\log(d)}{\underline{\alpha}_L nN}} = \frac{K}{\gamma_k}\sqrt{\frac{\rho_j\psi_{jk}\log(d)}{nN}}.$$

Finally, by using

$$\frac{p^2\log^4(d)}{\underline{\alpha}_L^2 nN^3} \le \frac{p\log^2(d)}{c_1\underline{\alpha}_L nN^2} \le \frac{\gamma_k\log(d)}{c_0c_1KnN}$$

from (64) and $\gamma_k \ge \underline{\gamma}$, we can upper bound $\max_{i\in L_k}(\mu_j/p)\delta_{ij}$ by

$$(1+\rho_j)\,\frac{K\log(d)}{\gamma_k nN} + \sqrt{\frac{p\log^4(d)}{\underline{\mu}_L nN^3}} + \frac{K}{\gamma_k}\sqrt{(1+\rho_j)\,\frac{\psi_{jk}\log(d)}{nN}} + \sqrt{\frac{\mu_j}{p}\frac{K\log(d)}{\gamma_k nN}}, \tag{66}$$

which completes the proof. ■

The following three lemmas provide upper bounds for the three terms on the right-hand-side of (57). Recall that $\rho_j = \alpha_j/\underline{\alpha}_L$, $\widetilde{s}_J = \sum_{j\in L^c}\rho_j s_j$ and $\psi_{jk} = \sum_{a=1}^K A_{ja}C_{ak}$ for any $j \in [p]$ and $k \in [K]$.

Lemma 15 *Under conditions of Theorem 2, with probability* $1 - O(d^{-1})$,

$$\begin{aligned}
&\sum_{j\in T^c\setminus L}\sqrt{s_j}\cdot\frac{\mu_j}{p}\frac{|(\Delta^{(j)})^\top(\widehat{h}^{(j)}-h^{(j)})|}{\|\Delta^{(j)}\|}\\
&\quad\lesssim \max\{s_J+|I|-|L|,\widetilde{s}_J\}\left\{\frac{K\log(d)}{\underline{\gamma}nN}+\sqrt{\frac{p\log^4(d)}{\underline{\mu}_L nN^3}}\right\}\\
&\quad + K\sqrt{\max\{s_J+|I|-|L|,\widetilde{s}_J\}\,\frac{\log(d)}{\underline{\gamma}nN}}.
\end{aligned}$$

Proof Pick any $j \in T^c \setminus L$. From the definition of $\widehat{h}^{(j)}$ in (18), we have

$$\begin{aligned}
\frac{\mu_j}{p}|(\Delta^{(j)})^\top(\widehat{h}^{(j)}-h^{(j)})| &\le \sum_{k=1}^K|\Delta_k^{(j)}|\cdot\frac{\mu_j}{p}\left|\widehat{h}_k^{(j)}-h_k^{(j)}\right|\\
&\le \sum_{k=1}^K|\Delta_k^{(j)}|\cdot\frac{\mu_j}{p}\left|\frac{1}{|L_k|}\sum_{i\in L_k}\left(\widehat{R}_{ij}-R_{ij}\right)\right|\\
&\le c_1\sum_{k=1}^K|\Delta_k^{(j)}|\cdot\frac{\mu_j}{p}\max_{i\in L_k}\delta_{j\ell},
\end{aligned}$$

with probability $1 - O(d^{-1})$, invoking Lemma 11 and inequality (52). Application of the third part of Lemma 14 further gives

$$\begin{aligned}\frac{\mu_j}{p}|(\Delta^{(j)})^\top(\widehat{h}^{(j)} - h^{(j)})| &\le c_1\|\Delta^{(j)}\|\left[\sum_{k=1}^K \left(T_2^{(jk)}\right)^2\right]^{1/2} + c_1\|\Delta^{(j)}\|_1 \max_{1\le k\le K} T_1^{(jk)} \\ &\overset{(56)}{\le} c_1\|\Delta^{(j)}\|\left[\sum_{k=1}^K \left(T_2^{(jk)}\right)^2\right]^{1/2} + 2c_1\sqrt{s_j}\|\Delta^{(j)}\| \max_{1\le k\le K} T_1^{(jk)},\end{aligned}$$

where

$$T_1^{(jk)} = (1+\rho_j)\,\frac{K\log(d)}{\gamma_k nN} + \sqrt{\frac{p\log^4(d)}{\underline{\mu}_L nN^3}} \tag{67}$$

$$T_2^{(jk)} = \frac{K}{\gamma_k}\sqrt{(1+\rho_j)\,\frac{\psi_{jk}\log(d)}{nN}} + \sqrt{\frac{\mu_j}{p}\frac{K\log(d)}{\gamma_k nN}}. \tag{68}$$

Hence, by the Cauchy-Schwarz inequality,

$$\begin{aligned}&\sum_{j\in T^c\setminus L}\sqrt{s_j}\cdot\frac{\mu_j}{p}\frac{|(\Delta^{(j)})^\top(\widehat{h}^{(j)} - h^{(j)})|}{\|\Delta^{(j)}\|} \\ &\lesssim \sqrt{\sum_{j\in T^c\setminus L}(1+\rho_j)s_j}\left\{\sum_{j\in T^c\setminus L}\sum_{k=1}^K\left(\frac{K^2\psi_{jk}\log(d)}{\gamma_k^2 nN} + \frac{\mu_j K\log(d)}{\gamma_k npN}\right)\right\}^{\frac{1}{2}} \\ &\quad + \sum_{j\in T^c\setminus L} s_j \max_{k\in[K]} T_1^{(jk)}.\end{aligned}$$

We conclude our proof by observing that

$$\begin{aligned}\sum_{j\in T^c\setminus L} s_j &\le s_J + |I| - |L| \\ \sum_{j\in T^c\setminus L}\sum_{k=1}^K \frac{K^2\psi_{jk}}{\gamma_k^2} &\le \sum_{k=1}^K\frac{K^2}{\gamma_k^2}\sum_{j=1}^p\psi_{jk} \overset{(39)}{\le} \frac{K^2}{\underline{\gamma}}, \\ \sum_{j\in T^c\setminus L}\sum_{k=1}^K \frac{\mu_j}{p}\frac{K\log(d)}{\gamma_k nN} &\le \frac{K^2\log(d)}{\underline{\gamma}nN}\end{aligned}$$

by $\sum_{j=1}^p \mu_j = p$. ∎

Lemma 16 *Under conditions of Theorem 2, with probability* $1 - O(d^{-1})$,

$$\begin{aligned}&\sum_{j\in T^c\setminus L}\sqrt{s_j}\cdot\frac{\mu_j}{p}\frac{|(\Delta^{(j)})^\top(h^{(j)} - \widehat{M}B_{j\cdot})|}{\|\Delta^{(j)}\|} \\ &\lesssim \widetilde{s}_J\sqrt{\frac{p\log^4(d)}{\underline{\mu}_L nN^3}} + \frac{K\widetilde{s}_J\log(d)}{\underline{\gamma}nN} + K\sqrt{\frac{\widetilde{s}_J\log(d)}{\underline{\gamma}nN}}.\end{aligned}$$

Proof We work on the event

$$\mathcal{E} := \bigcap_{i\in L}\left\{\frac{1}{n}\left|\sum_{t=1}^{n}\varepsilon_{it}\right| \le 6\sqrt{\frac{\mu_i \log(d)}{npN}}\right\}\bigcap\left\{\bigcap_{i,\ell\in L}\left\{|\widehat{R}_{i\ell}-R_{i\ell}| \le c_1\delta_{i\ell}\right\}\right\}. \tag{69}$$

Lemmas 8, 11 and (24) guarantee that $\mathbb{P}(\mathcal{E}) \ge 1 - O(d^{-1})$. The event $\mathcal{E}$ and (24) further imply

$$c\frac{\mu_i}{p} \le \frac{\widehat{u}_i}{p} \le c'\frac{\mu_i}{p}, \qquad \text{for all} \quad i \in L, \tag{70}$$

for some constants $c, c' > 0$ and (64). Pick any $j \in T^c \setminus L$ and $k \in [K]$. Observe that $h^{(j)} = MB_{j\cdot}$ and

$$B_{ja} = \frac{p}{\mu_j}A_{ja}\frac{\gamma_a}{K}. \tag{71}$$

From (15) and (18), we have

$$\begin{aligned}
\frac{\mu_j}{p}\left|(\widehat{M}_{k\cdot}-M_{k\cdot})^\top B_{j\cdot}\right| &= \frac{1}{K}\left|\sum_{a=1}^{K}\frac{A_{ja}\gamma_a}{|L_k||L_a|}\sum_{i\in L_k,\ell\in L_a}\left(\widehat{R}_{i\ell}-R_{i\ell}\right)\right| \\
&= \frac{1}{K}\left|\sum_{a=1}^{K}\frac{A_{ja}\gamma_a}{|L_k||L_a|}\sum_{i\in L_k,\ell\in L_a}\left(\frac{p^2\widehat{\Theta}_{i\ell}}{\widehat{u}_i\widehat{u}_\ell}-\frac{p^2\Theta_{i\ell}}{\mu_i\mu_\ell}\right)\right| \\
&\le \frac{1}{K}\left|\sum_{a=1}^{K}\frac{A_{ja}\gamma_a}{|L_k||L_a|}\sum_{i\in L_k,\ell\in L_a}\frac{p^2(\widehat{\Theta}_{i\ell}-\Theta_{i\ell})}{\mu_i\mu_\ell}\right| \\
&\quad + \frac{1}{K}\left|\sum_{a=1}^{K}\frac{A_{ja}\gamma_a}{|L_k||L_a|}\sum_{i\in L_k,\ell\in L_a}\frac{(\mu_i\mu_\ell-\widehat{\mu}_i\widehat{\mu}_\ell)}{\mu_i\mu_\ell}\widehat{R}_{i\ell}\right| \\
&:= \mathrm{Rem}_1{}^{(jk)} + \mathrm{Rem}_2{}^{(jk)}.
\end{aligned}$$

For $\mathrm{Rem}_2{}^{(jk)}$, we find

$$\begin{aligned}
\mathrm{Rem}_2{}^{(jk)} &\le \left|\sum_{a=1}^{K}\frac{A_{ja}\gamma_a}{K}\frac{1}{|L_k||L_a|}\sum_{i\in L_k,\ell\in L_a}\frac{[\mu_i(\mu_\ell-\widehat{\mu}_\ell)+(\mu_i-\widehat{\mu}_i)\widehat{\mu}_\ell]}{\mu_i\mu_\ell}\widehat{R}_{i\ell}\right| \\
&\lesssim \sum_{a=1}^{K}\frac{A_{ja}\gamma_a}{K}\frac{1}{|L_k||L_a|}\sum_{i\in L_k,\ell\in L_a}\widehat{R}_{i\ell}\max_{i\in L_k,\ell\in L_a}\left(\frac{|\mu_\ell-\widehat{\mu}_\ell|}{\mu_\ell}+\frac{|\widehat{\mu}_i-\mu_i|}{\mu_i}\right) \\
&\lesssim \sum_{a=1}^{K}\frac{A_{ja}\gamma_a}{K}\left(\sqrt{\frac{pK\log(d)}{\underline{\alpha}_L\gamma_k nN}}+\sqrt{\frac{pK\log(d)}{\underline{\alpha}_L\gamma_a nN}}\right)\frac{1}{|L_k||L_a|}\sum_{i\in L_k,\ell\in L_a}(R_{i\ell}+c_1\delta_{i\ell}) \\
&\le 2\sum_{a=1}^{K}\frac{A_{ja}C_{ka}K}{\gamma_k}\sqrt{\frac{pK\log(d)}{\underline{\alpha}_L\underline{\gamma}nN}}+\sqrt{\frac{p\log(d)}{\underline{\mu}_L nN}}\sum_{a=1}^{K}\frac{A_{ja}\gamma_a}{K}\max_{i\in L_k,\ell\in L_a}\delta_{i\ell}.
\end{aligned}$$

We use (70) in the second line, the definition of the event $\mathcal{E}$ together with (36) in the third line and

$$R_{i\ell} = \frac{p^2\Theta_{i\ell}}{\mu_i\mu_\ell} = \frac{K^2C_{ka}}{\gamma_k\gamma_a}$$

(follows from (36) and (61)) in the fourth line. We bound the first term on the right as

$$\sum_{a=1}^{K}\frac{A_{ja}C_{ka}K}{\gamma_k}\sqrt{\frac{pK\log(d)}{\underline{\alpha}_L\underline{\gamma}nN}} = \frac{\psi_{jk}K}{\gamma_k}\sqrt{\frac{pK\log(d)}{\underline{\alpha}_L\underline{\gamma}nN}} \le K\sqrt{\frac{\rho_j\psi_{jk}\log(d)}{\underline{\gamma}\gamma_k nN}}$$

by using

$$\psi_{jk} = \frac{1}{n}\sum_{t=1}^{n}\sum_{a=1}^{K}A_{ja}W_{at}W_{kt} \le \|A_{j\cdot}\|_\infty\frac{1}{n}\sum_{t=1}^{n}\sum_{a=1}^{K}W_{at}W_{kt} = \frac{\alpha_j\gamma_k}{pK}. \tag{72}$$

Invoking the second result of Lemma 14 gives

$$\mathrm{Rem}_2{}^{(jk)} \lesssim K\sqrt{\frac{\rho_j\psi_{jk}\log(d)}{\underline{\gamma}\gamma_k nN}} + \sqrt{\frac{\rho_j\|A_{j\cdot}\|_1\log(d)}{\gamma_k nN}}. \tag{73}$$

We proceed to bound $\mathrm{Rem}_1{}^{(jk)}$. Recalling (71) and $\mu_\ell/p = A_{\ell a}\gamma_a/K$ from (36), we find

$$\begin{aligned}\mathrm{Rem}_1{}^{(jk)} &= \left|\sum_{a=1}^{K}\frac{A_{ja}}{|L_k||L_a|}\sum_{i\in L_k,\ell\in L_a}\frac{p(\widehat{\Theta}_{i\ell}-\Theta_{i\ell})}{\mu_i A_{\ell a}}\right| \\ &\le \max_{i\in L_k}\frac{p}{\mu_i}\left|\sum_{a=1}^{K}\frac{1}{|L_a|}\sum_{\ell\in L_a}\frac{A_{ja}}{A_{\ell a}}(\widehat{\Theta}_{i\ell}-\Theta_{i\ell})\right|.\end{aligned}$$

Since, for any $i\in L_k$, $j\in L_a$,

$$\begin{aligned}\widehat{\Theta}_{i\ell}-\Theta_{i\ell} &= \frac{N}{N-1}\left(\frac{1}{n}A_{ik}W_k^\top\varepsilon_\ell + \frac{1}{n}A_{\ell a}W_a^\top\varepsilon_i\right) + \frac{N}{N-1}\left(\frac{1}{n}\varepsilon_i^\top\varepsilon_\ell - \frac{1}{n}\mathbb{E}\left[\varepsilon_i^\top\varepsilon_\ell\right]\right) \\ &\quad - \frac{1}{N-1}\mathrm{diag}\left(\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{it}\right)1_{\{i=\ell\}},\end{aligned}$$

cf. (Bing et al., 2020, page 11 in the Supplement), we obtain

$$\begin{aligned}\mathrm{Rem}_1{}^{(jk)} &\lesssim \max_{i\in L_k}\frac{p}{\mu_i}\left\{A_{ik}\left|\sum_{a=1}^{K}\frac{1}{|L_a|}\sum_{\ell\in L_a}\frac{A_{ja}}{A_{\ell a}}\frac{1}{n}\sum_{t=1}^{n}W_{kt}\varepsilon_{\ell t}\right| + \left|\sum_{a=1}^{K}A_{ja}\frac{1}{n}\sum_{t=1}^{n}W_{at}\varepsilon_{it}\right|\right. \\ &\quad \left. + \left|\sum_{a=1}^{K}\frac{1}{|L_a|}\sum_{\ell\in L_a}\frac{A_{ja}}{A_{\ell a}}\left(\frac{1}{n}\varepsilon_i^\top\varepsilon_\ell - \frac{1}{n}\mathbb{E}\left[\varepsilon_i^\top\varepsilon_\ell\right]\right)\right| + \frac{A_{jk}}{NA_{ik}}\left|\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{it}\right|\right\} \\ &:= \mathrm{Rem}_{11}{}^{(jk)} + \mathrm{Rem}_{12}{}^{(jk)} + \mathrm{Rem}_{13}{}^{(jk)} + \mathrm{Rem}_{14}{}^{(jk)}.\end{aligned} \tag{74}$$

In the sequel, we provides separate bounds the each of the four terms. We start with the last term and obtain on the event $\mathcal{E}$

$$\mathrm{Rem}_{14}{}^{(jk)} \leq \max_{i\in L_k} \frac{p}{\mu_i}\frac{A_{jk}}{NA_{ik}}\left|\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{it}\right| \leq 6\rho_j\sqrt{\frac{p\log(d)}{\underline{\mu}_L nN^3}}. \tag{75}$$

by recalling that $\rho_j = \alpha_j/\underline{\alpha}_L$. Observing that $\sum_a A_{ja}W_{at} = \Pi_{jt}$, with probability $1 - O(d^{-1})$, the second term can be upper bounded by using Lemma 9 as

$$\begin{aligned}
\mathrm{Rem}_{12}{}^{(jk)} &= \max_{i\in L_k}\frac{p}{\mu_i}\left|\frac{1}{n}\sum_{t=1}^{n}\Pi_{jt}\varepsilon_{it}\right| \\
&\leq \max_{i\in L_k}\frac{p}{\mu_i}\left(\sqrt{\frac{6m_j\Theta_{ji}\log(d)}{npN}} + \frac{2m_j\log(d)}{npN}\right) \\
&\leq \max_{i\in L_k}\left(\frac{K}{\gamma_k}\sqrt{\frac{6\alpha_j\psi_{jk}\log(d)}{\alpha_i nN}} + \frac{2\alpha_j\log(d)}{\mu_i nN}\right) \\
&\leq \frac{K}{\gamma_k}\sqrt{\frac{6\rho_j\psi_{jk}\log(d)}{nN}} + \frac{2\rho_j K\log(d)}{\gamma_k nN}
\end{aligned} \tag{76}$$

where we also use(37) and (65) to derive the third line and use (36) to arrive at the last line. The upper bounds of $\mathrm{Rem}_{11}{}^{(jk)}$ and $\mathrm{Rem}_{13}{}^{(jk)}$ are proved in Lemmas 18 and 19. Combination of (73), (75), (76), (77) and (83) yields

$$\begin{aligned}
\frac{\mu_j}{p}\left|(\widehat{M}_{k\cdot} - M_{k\cdot})^\top B_{j\cdot}\right| \lesssim{}& \rho_j\sqrt{\frac{p\log^4(d)}{\underline{\mu}_L nN^3}} + \frac{2\rho_j K\log(d)}{\gamma_k nN} \\
&+ K\sqrt{\frac{\rho_j\psi_{jk}\log(d)}{\underline{\gamma}\gamma_k nN}} + \sqrt{\frac{\rho_j\|A_{j\cdot}\|_1\log(d)}{\gamma_k nN}} + \sqrt{\frac{\mu_j}{p}\frac{\rho_j K\log(d)}{\gamma_k nN}}
\end{aligned}$$

with probability $1 - O(d^{-1})$. Next we use similar arguments as in the proof of Lemma 15. Analogous to (67) – (68), we define

$$\begin{aligned}
\rho_j R_1^{(k)} &= \sqrt{\frac{p\log^4(d)}{\underline{\mu}_L nN^3}} + \frac{2K\log(d)}{\gamma_k nN} \\
\sqrt{\rho_j}R_2^{(jk)} &= K\sqrt{\frac{\psi_{jk}\log(d)}{\underline{\gamma}\gamma_k nN}} + \sqrt{\frac{\|A_{j\cdot}\|_1\log(d)}{\gamma_k nN}} + \sqrt{\frac{\mu_j}{p}\frac{K\log(d)}{\gamma_k nN}}.
\end{aligned}$$

We can obtain

$$\frac{\mu_j}{p}\left|(\Delta^{(j)})^\top(\widehat{M} - M)B_{j\cdot}\right| \lesssim \|\Delta^{(j)}\|\sqrt{\rho_j}\left[\sum_{k=1}^{K}\left(R_2^{(jk)}\right)^2\right]^{1/2} + \rho_j\sqrt{s_j}\|\Delta^{(j)}\|\max_{1\leq k\leq K}R_1^{(k)}$$

which, by the Cauchy-Schwarz inequality, further gives

$$\sum_{j\in T^c\setminus L}\sqrt{s_j}\cdot\frac{\mu_j}{p}\frac{|(\Delta^{(j)})^\top(h^{(j)} - \widehat{M}B_{j\cdot}|}{\|\Delta^{(j)}\|} \lesssim \sqrt{\widetilde{s}_J}\left[\sum_{j\in T^c\setminus L}\sum_{k=1}^{K}\left(R_2^{(jk)}\right)^2\right]^{\frac{1}{2}} + \widetilde{s}_J\max_{k\in[K]}R_1^{(k)}.$$

Finally, we calculate $\sum_{j\in T^c\setminus L}\sum_{k=1}^K (R_2^{(jk)})^2$ as

$$\begin{aligned}
&\sum_{j\in T^c\setminus L}\sum_{k=1}^K (R_2^{(jk)})^2 \\
&\le \sum_{j\in T^c\setminus L}\sum_{k=1}^K \Big\{ \frac{K^2\psi_{jk}\log(d)}{\underline{\gamma}\gamma_k nN} + \frac{\|A_{j\cdot}\|_1\log(d)}{\gamma_k nN} + \frac{\mu_j}{p}\frac{\rho_j K\log(d)}{\gamma_k nN}\Big\} \\
&\le \frac{3K^2\log(d)}{\underline{\gamma} nN}.
\end{aligned}$$

We use (33), (39) and $\sum_{j=1}^p \|A_{j\cdot}\|_1 = K$ to arrive at the last line. ■

Lemma 17 *Let λ be chosen as in (26). With probability $1-O(d^{-1})$,*

$$\lambda\sum_{j\in T^c\setminus L}\frac{\mu_j}{p}\sqrt{s_j}\|B_{j\cdot}\| \le cK\sqrt{\frac{\overline{\gamma}}{\underline{\gamma}}\cdot\frac{K\widetilde{s}_J\log(d)}{\underline{\gamma}nN}}.$$

Proof Recall that $B_{jk}\mu_j/p = A_{jk}\gamma_k/K$. We have

$$\begin{aligned}
\sum_{j\in T^c\setminus L}\frac{\mu_j}{p}\sqrt{s_j}\|B_{j\cdot}\| &= \frac{1}{K}\sum_{j\in T^c\setminus L}\sqrt{s_j}\left[\sum_{k=1}^K A_{jk}^2\gamma_k^2\right]^{1/2} \\
&\le \frac{1}{K}\sum_{j\in T^c\setminus L}\sqrt{s_j}\left[\sum_{k=1}^K A_{jk}\gamma_k\right]^{1/2}\sqrt{\frac{\alpha_j\overline{\gamma}}{p}}.
\end{aligned}$$

From $\underline{\mu}_L = \underline{\alpha}_L\underline{\gamma}/K$ and the choice of λ, it follows that, with probability $1-O(d^{-1})$,

$$\begin{aligned}
\lambda\sum_{j\in T^c\setminus L}\frac{\mu_j}{p}\sqrt{s_j}\|B_{j\cdot}\| &\le \frac{c}{\underline{\gamma}}\sqrt{\frac{pK^2\log(d)}{\underline{\alpha}_L\underline{\gamma}nN}}\sum_{j\in T^c\setminus L}\sqrt{s_j}\left[\sum_{k=1}^K A_{jk}\gamma_k\right]^{1/2}\sqrt{\frac{\alpha_j\overline{\gamma}}{p}} \\
&= c\sqrt{\frac{\overline{\gamma}K^2\log(d)}{\underline{\gamma}^3 nN}}\sum_{j\in T^c\setminus L}\sqrt{\rho_j s_j}\left[\sum_{k=1}^K A_{jk}\gamma_k\right]^{1/2} \\
&\le cK\sqrt{\frac{\overline{\gamma}K\log(d)}{\underline{\gamma}^3 nN}}\sqrt{\widetilde{s}_J}\left[\sum_{j\in T^c\setminus L}\sum_{k=1}^K\frac{A_{jk}\gamma_k}{K}\right]^{1/2} \\
&= cK\sqrt{\frac{\overline{\gamma}\widetilde{s}_J K\log(d)}{\underline{\gamma}^3 nN}}\left[\sum_{j\in T^c\setminus L}\frac{\mu_j}{p}\right]^{1/2} \\
&\le cK\sqrt{\frac{\overline{\gamma}\widetilde{s}_J K\log(d)}{\underline{\gamma}^3 nN}}.
\end{aligned}$$

Here we use the Cauchy-Schwarz inequality in the third line and the identity $\sum_{j=1}^{p} \mu_j = p$ in the last line. This completes the proof. ■

A.6. Lemmas used in the proof of Lemma 16

Let $\mathrm{Rem}_{11}{}^{(jk)}$ and $\mathrm{Rem}_{13}{}^{(jk)}$, $j \in [p]$, $k \in [K]$, be defined as (74).

Lemma 18 *Under conditions of Theorem 2, with probability* $1 - 2d^{-1}$,

$$\mathrm{Rem}_{11}{}^{(jk)} \leq \frac{K}{\gamma_k}\sqrt{\frac{6\rho_j \psi_{jk}\log(d)}{nN}} + \frac{4\rho_j K \log(d)}{\gamma_k nN} \tag{77}$$

uniformly for any $j \in [p]$ *and* $k \in [K]$.

Proof We upper bound $\mathrm{Rem}_{11}{}^{(jk)}$ by studying

$$\max_{i\in L_k} \frac{p}{\mu_i} A_{ik} \left| \sum_{a=1}^{K} \frac{1}{|L_a|} \sum_{\ell \in L_a} \frac{A_{ja}}{A_{\ell a}} \frac{1}{n} \sum_{t=1}^{n} W_{kt}\varepsilon_{\ell t} \right| \overset{(36)}{=} \frac{K}{\gamma_k} \left| \frac{1}{n} \sum_{t=1}^{n} W_{kt} \sum_{a=1}^{K} \sum_{\ell \in L_a} \frac{1}{|L_a|} \frac{A_{ja}}{A_{\ell a}} \varepsilon_{\ell t} \right|.$$

Recall that

$$\varepsilon_{\ell t} = \frac{1}{N} \sum_{r=1}^{N} Z_{rt}^{(\ell)} \tag{78}$$

where $Z_{rt}^{(\ell)}$ denotes the ℓth element of Z_{rt} and Z_{rt} has a centered Multinomial$_p(1;\Pi_t)$ (subtracted its mean M_t). Next we will use Bernstein's inequality to bound

$$\left| \frac{1}{n} \sum_{t=1}^{n} \sum_{r=1}^{N} W_{kt} \left(\sum_{a=1}^{K} \sum_{\ell \in L_a} \frac{1}{|L_a|} \frac{A_{ja}}{A_{\ell a}} Z_{rt}^{(\ell)} \right) \right| := \left| \frac{1}{n} \sum_{t=1}^{n} \sum_{r=1}^{N} W_{kt}\zeta_{rt} \right| \tag{79}$$

from above. Note that $\mathbb{E}[W_{kt}\zeta_{rt}] = 0$ and

$$|W_{kt}\zeta_{rt}| \leq \rho_j \sum_{a=1}^{K} \max_{\ell \in L_a} \left| Z_{rt}^{(\ell)} \right| \leq 2\rho_j. \tag{80}$$

To calculate the variance of $\sum_{t=1}^{n}\sum_{r=1}^{N} W_{kt}\zeta_{rt}$, observe that

$$\zeta_{rt} = \eta^\top Z_{rt}^{L}$$

with Z_{rt}^{L} denoting the sub-vector of Z_{rt} corresponding to L and

$$\eta = D_L \begin{bmatrix} \mathbf{1}_{|L_1|} & & \\ & \ddots & \\ & & \mathbf{1}_{|L_K|} \end{bmatrix} \begin{bmatrix} A_{j1}/|L_1| \\ \vdots \\ A_{jK}/|L_K| \end{bmatrix} \in \mathbb{R}^{|L|} \tag{81}$$

where $(D_L)_{\ell\ell} = 1/A_{\ell a}$ for any $\ell \in L_a$ and $a \in [K]$. We thus have

$$\begin{aligned}
\mathrm{Var}\left(\sum_{t=1}^{n}\sum_{r=1}^{N} W_{kt}\zeta_{rt}\right) &\le N\sum_{t=1}^{n} W_{kt}^2 \eta^\top \mathrm{diag}(\Pi_{Lt})\eta \\
&= nN\frac{1}{n}\sum_{t=1}^{n} W_{kt}^2 \sum_{a=1}^{K}\sum_{\ell\in L_a}\Pi_{\ell t}\left(\frac{A_{ja}}{A_{\ell a}}\frac{1}{|L_a|}\right)^2 \\
&\le nN\frac{1}{n}\sum_{t=1}^{n} W_{kt}\sum_{a=1}^{K}\frac{1}{|L_a|}\sum_{\ell\in L_a} W_{at}\frac{A_{ja}^2}{A_{\ell a}} \\
&\le \rho_j nN\psi_{jk}, \qquad (82)
\end{aligned}$$

using $W_{kt} \le 1$ and $|L_a| \ge 1$ in the third line and (38) in the last line. Invoke Lemma 23 with $B = 2\rho_j$ and $v = \rho_j nN\psi_{jk}$ to obtain, for any $t > 0$,

$$\mathbb{P}\left\{\frac{1}{nN}\left|\sum_{t=1}^{n}\sum_{r=1}^{N} W_{kt}\zeta_{rt}\right| > t\right\} \le 2\exp\left(-\frac{n^2N^2t^2/2}{\rho_j nN\psi_{jk} + 2nN\rho_j t/3}\right),$$

which further implies

$$\mathbb{P}\left\{\frac{1}{nN}\left|\sum_{t=1}^{n}\sum_{r=1}^{N} W_{kt}\zeta_{rt}\right| > \sqrt{\frac{\rho_j\psi_{jk}t}{nN}} + \frac{2\rho_j t}{3nN}\right\} \le 2e^{-t/2}, \quad \text{for any } t > 0.$$

Choosing $t = 6\log(d)$ yields

$$\mathrm{Rem}_{11}{}^{(jk)} \le \frac{K}{\gamma_k}\sqrt{\frac{6\rho_j\psi_{jk}\log(d)}{nN}} + \frac{4\rho_j K\log(d)}{\gamma_k nN}$$

with probability $1 - 2d^{-3}$. Taking the union bound for probabilities completes the proof. ■

Lemma 19 *Under conditions of Theorem 2, with probability $1 - 6d^{-1}$, we have*

$$\mathrm{Rem}_{13}{}^{(jk)} \lesssim \frac{K}{\gamma_k}\sqrt{\frac{\rho_j\psi_{jk}\log(d)}{nN}} + \sqrt{\frac{\mu_j}{p}\frac{\rho_j K\log(d)}{\gamma_k nN}} + \rho_j\sqrt{\frac{p\log^4(d)}{\underline{\mu}_L nN^3}} \qquad (83)$$

uniformly for $j \in [p]$ and $k \in [K]$.

Proof Recall that

$$\mathrm{Rem}_{13}{}^{(jk)} = \max_{i\in L_k}\frac{p}{\mu_i}\left|\frac{1}{n}\sum_{t=1}^{n}\left(\varepsilon_{it}\xi_t - \mathbb{E}\left[\varepsilon_{it}\xi_t\right]\right)\right|.$$

Using (78) and (79), we have

$$\xi_t := \sum_{a=1}^{K}\sum_{\ell\in L_a}\frac{A_{ja}}{A_{\ell a}|L_a|}\varepsilon_{\ell t} = \frac{1}{N}\sum_{r=1}^{N}\zeta_{rt}.$$

We will use similar truncation arguments in tandem with Hoeffding's inequality as in (Bing et al., 2020, proof of Lemma 15). This implies that, for any $i \in L$,

$$\mathbb{P}\left\{|\varepsilon_{it}| \le \sqrt{\frac{6\Pi_{it}\log(d)}{N}} + \frac{2\log(d)}{N} := T_t\right\} \ge 1 - 2d^{-3}.$$

To truncate ζ_{rt}, recall that $\mathbb{E}[\zeta_{rt}] = 0$, $|\zeta_{rt}| \le 2\rho_j$ from (80) and

$$\begin{aligned}
\mathrm{Var}\left(\sum_{r=1}^{N}\zeta_{rt}\right) &= N\eta^\top \mathrm{diag}(\Pi_{Lt})\eta \\
&= N\sum_{a=1}^{K}\sum_{\ell\in L_a}\Pi_{\ell t}\left(\frac{A_{ja}}{A_{\ell a}}\frac{1}{|L_a|}\right)^2 \\
&\le N\sum_{a=1}^{K}\frac{1}{|L_a|}\sum_{\ell\in L_a}W_{at}\frac{A_{ja}^2}{A_{\ell a}} \\
&\le N\rho_j\Pi_{jt}
\end{aligned}$$

where η is defined in (81). Invoking Lemma 23 with $B = 2\rho_j$ and $v = N\rho_j\Pi_{jt}$ yields

$$\mathbb{P}\left\{\frac{1}{N}\left|\sum_{r=1}^{N}\zeta_{rt}\right| \le \sqrt{\frac{6\rho_j\Pi_{jt}\log(d)}{N}} + \frac{4\rho_j\log(d)}{N} := T_t'\right\} \ge 1 - 2d^{-3}.$$

We define $Y_{it} = \varepsilon_{it}\mathbf{1}_{\mathcal{S}_t}$ with $\mathcal{S}_t := \{|\varepsilon_{it}| \le T_t\}$ and $Y_t' = \xi_t\mathbf{1}_{\mathcal{S}_t'}$ with $\mathcal{S}_t' := \{|\xi_t| \le T_t'\}$, for each $i \in [p]$ and $t \in [n]$, and set $\mathcal{S} := \cap_{i=1}^{p}\cap_{t=1}^{n}\mathcal{S}_t \cap \mathcal{S}_t'$. It follows that $\mathbb{P}(\mathcal{S}) \ge 1 - 4d^{-1}$. On the event $\mathcal{S}$, we have

$$\frac{1}{n}\left|\sum_{t=1}^{n}(\varepsilon_{it}\xi_t - \mathbb{E}[\varepsilon_{it}\xi_t])\right| \le \underbrace{\frac{1}{n}\left|\sum_{t=1}^{n}\left(Y_{it}Y_t' - \mathbb{E}[Y_{it}Y_t']\right)\right|}_{R_1} + \underbrace{\frac{1}{n}\left|\sum_{t=1}^{n}\left(\mathbb{E}[\varepsilon_{it}\xi_t] - \mathbb{E}[Y_{it}Y_t']\right)\right|}_{R_2}$$

Since

$$\mathbb{E}[\varepsilon_{it}\xi_t] = \mathbb{E}[Y_{it}Y_t'] + \mathbb{E}\left[Y_{it}\xi_t\mathbf{1}_{(\mathcal{S}_t')^c}\right] + \mathbb{E}\left[\varepsilon_{it}\mathbf{1}_{\mathcal{S}_t^c}\xi_t\right],$$

we have

$$\begin{aligned}
R_2 = \frac{1}{n}\left|\sum_{t=1}^{n}\left(\mathbb{E}[\varepsilon_{it}\xi_t] - \mathbb{E}[Y_{it}Y_t']\right)\right| &\le \frac{1}{n}\left|\sum_{t=1}^{n}\left(\mathbb{E}\left[Y_{it}\xi_t\mathbf{1}_{(\mathcal{S}_t')^c}\right] + \mathbb{E}\left[\varepsilon_{it}\mathbf{1}_{\mathcal{S}_t^c}\xi_t\right]\right)\right| \\
&\le \frac{1}{n}\sum_{t=1}^{n}2\rho_j\left(\mathbb{P}(\mathcal{S}_t^c) + \mathbb{P}((\mathcal{S}_t')^c)\right) \qquad (84) \\
&\le 8\rho_j d^{-3}
\end{aligned}$$

by using $|Y_{it}| \le |\varepsilon_{it}| \le 1$ and $|\xi_t| \le |\zeta_{rt}| \le 2\rho_j$ in the second inequality.

It remains to bound R_1. Since $|Y_{it}| \leq T_t$, we know $-2T_tT_t' \leq Y_{it}Y_t' - \mathbb{E}[Y_{it}Y_t'] \leq 2T_tT_t'$ for all $1 \leq t \leq n$. Applying Hoeffding's inequality (Lemma 24) with $a_t = -2T_tT_t'$ and $b_t = 2T_tT_t'$ gives

$$\mathbb{P}\left\{\left|\sum_{t=1}^{n}\left(Y_{it}Y_t' - \mathbb{E}[Y_{it}Y_t']\right)\right| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{8\sum_{t=1}^{n}T_t^2(T_t')^2}\right).$$

Taking $t = \sqrt{24\sum_{t=1}^{n}T_t^2(T_t')^2\log(d)}$ yields

$$R_1 = \frac{1}{n}\left|\sum_{t=1}^{n}\left(Y_{it}Y_t' - \mathbb{E}[Y_{it}Y_t']\right)\right| \leq 2\sqrt{6}\left(\frac{1}{n}\sum_{t=1}^{n}T_t^2(T_t')^2\cdot\frac{\log(d)}{n}\right)^{1/2} \tag{85}$$

with probability greater than $1 - 2d^{-3}$. Finally, note that

$$\begin{aligned}\frac{1}{4n}\sum_{i=1}^{n}T_t^2(T_t')^2 &\leq \frac{1}{n}\sum_{t=1}^{n}\left\{\frac{\Pi_{it}\rho_j\Pi_{jt}\log^2(d)}{N^2} + \frac{\rho_j^2\log^4(d)}{N^4} + \frac{\rho_j\Pi_{jt}\log^3(d)}{N^3} + \frac{\Pi_{it}\rho_j^2\log^3(d)}{N^3}\right\}\\ &= \frac{\rho_j\Theta_{ij}\log^2(d)}{N^2} + \frac{\rho_j^2\log^4(d)}{N^4} + \frac{\rho_j\mu_j\log^3(d)}{pN^3} + \frac{\rho_j^2\mu_i\log^3(d)}{pN^3}\\ &\lesssim \frac{\rho_j\alpha_i\psi_{jk}\log^2(d)}{pN^2} + \frac{\rho_j\mu_j\log^3(d)}{pN^3} + \frac{\rho_j^2\mu_i\log^3(d)}{pN^3}\end{aligned} \tag{86}$$

by using (32) in the second equality and (64) and (65) to obtain the last line. Finally, combining (84) - (86) gives

$$\begin{aligned}\mathrm{Rem}_{13}{}^{(jk)} &\lesssim \frac{K}{\gamma_k}\sqrt{\frac{\rho_j\psi_{jk}p\log^3(d)}{\underline{\alpha}_L nN^2}} + \frac{K}{\gamma_k}\sqrt{\rho_j\frac{\mu_j}{p}\frac{p^2\log^4(d)}{\underline{\alpha}_L^2 nN^3}} + \rho_j\sqrt{\frac{p\log^4(d)}{\underline{\mu}_L nN^3}} + \frac{p\rho_j}{\underline{\mu}_L d^3}\\ &\lesssim \frac{K}{\gamma_k}\sqrt{\frac{\rho_j\psi_{jk}\log(d)}{nN}} + \sqrt{\rho_j\frac{\mu_j}{p}\frac{K\log(d)}{\gamma_k nN}} + \rho_j\sqrt{\frac{p\log^4(d)}{\underline{\mu}_L nN^3}},\end{aligned}$$

for all j, k, with probability $1 - 6d^{-1}$. We also use (64) and $\underline{\alpha}_L\gamma_k \geq \underline{\alpha}_L\underline{\gamma} \geq c_0pK\log(d)/N$. This completes the proof. ■

A.7. Auxilliary lemmas

In this section, we state three lemmas which are used in the main paper. The following lemma gives the range of $\lambda_{\min}(\widetilde{Q}\widetilde{Q}^\top)$ where $\widetilde{Q} = D_Q^{-1}Q$ and $Q = CA^\top$.

Lemma 20 *Let $C = n^{-1}WW^\top$ and $M = D_W^{-1}CD_W^{-1}$. We have*

$$\left(\min_{k\in[K],i\in I_k}A_{ik}^2\right)\lambda_{\min}(C)\lambda_{\min}(M) \leq \lambda_{\min}(\widetilde{Q}\widetilde{Q}^\top) \leq \lambda_{\min}(M).$$

Proof Observe that

$$D_Q = Q\mathbf{1}_p = CA^\top \mathbf{1}_p = C\mathbf{1}_K \overset{(63)}{=} D_W,$$

whence

$$\lambda_{\min}(\widetilde{Q}\widetilde{Q}^\top) = \inf_{v\in\mathcal{S}^{K-1}} v^\top D_W^{-1} CA^\top ACD_W^{-1} v. \tag{87}$$

On the one hand,

$$\begin{aligned}\inf_{v\in\mathcal{S}^{K-1}} v^\top D_W^{-1} CA^\top ACD_W^{-1} v &\le \|C^{1/2}A^\top AC^{1/2}\|_{\mathrm{op}} \inf_{v\in\mathcal{S}^{K-1}} v^\top D_W^{-1} CD_W^{-1} v \\ &= \|ACA^\top\|_{\mathrm{op}} \cdot \lambda_{\min}(M).\end{aligned}$$

The upper bound now follows using (5) and $\|ACA^\top\|_{\mathrm{op}} \le 1$. The latter follows from the string of inequalities

$$\|ACA^\top\|_{\mathrm{op}} \le \|ACA^\top\|_{\infty,1} = \|ACA^\top \mathbf{1}_p\|_\infty = \|AC\mathbf{1}_K\|_\infty = \frac{1}{n}\|\Pi\mathbf{1}_n\|_\infty \le 1.$$

The lower bound follows immediately from

$$\begin{aligned}\lambda_{\min}(\widetilde{Q}\widetilde{Q}^\top) &\ge \lambda_{\min}(A^\top A)\lambda_{\min}(C)\lambda_{\min}(M) \\ &\ge \lambda_{\min}(A_I^\top A_I)\lambda_{\min}(C)\lambda_{\min}(M) \\ &\ge \left(\min_{k\in[K], i\in I_k} A_{ik}^2\right)\lambda_{\min}(C)\lambda_{\min}(M).\end{aligned}$$

■

Lemma 21 *Let $C = n^{-1}WW^\top$. We have*

$$\lambda_{\min}(C) \le \frac{1}{K}.$$

Proof From the definition of the smallest eigenvalue,

$$\lambda_{\min}(C) = \inf_{v\in\mathcal{S}^{K-1}} v^\top Cv \le \min_{1\le k\le K} C_{kk} = \min_{1\le k\le K} \frac{1}{n}\|W_{k\cdot}\|^2.$$

The result follows from

$$\frac{1}{n}\|W_{k\cdot}\|^2 \le \frac{1}{n}\|W_{k\cdot}\|_1 \overset{(22)}{=} \gamma_k$$

and

$$\min_k \gamma_k \le \sum_{k=1}^K \gamma_k/K = 1/K.$$

■

Lemma 22 *Under condition (24),*

$$\mathbb{P}\left\{\left|\min_{i\in L}(D_X)_{ii} - \frac{\mu_L}{p}\right| \le 6\sqrt{\frac{\log(d)}{nN}}\right\} \ge 1 - 2d^{-1}.$$

Proof For any $i \in L$, note that $(D_X)_{ii} - \mu_i/p = n^{-1}\sum_{t=1}^{n}\varepsilon_{it}$. The result follows from Lemma 8 and the inequalities $\mu_j/p \le 1$ for all $j \in [p]$ by (22). ■

For completeness, we state the well-known Bernstein and Hoeffding inequalities for bounded random variables.

Lemma 23 (Bernstein's inequality for bounded random variables) *For independent random variables $Y_1, \ldots, Y_n$ with bounded ranges $[-B, B]$ and zero means,*

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n} Y_i\right| > x\right\} \le 2\exp\left(-\frac{n^2x^2/2}{v + nBx/3}\right), \qquad \text{for any } x \ge 0,$$

where $v \ge var(Y_1 + \ldots + Y_n)$.

Lemma 24 (Hoeffding's inequality) *Let $Y_1, \ldots, Y_n$ be independent random variables with $\mathbb{E}[Y_i] = 0$ and $\mathbb{P}\{a_i \le Y_i \le b_i\} = 1$. For any $t \ge 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} Y_i\right| > t\right\} \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

References

Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-kai Liu. A spectral algorithm for latent dirichlet allocation. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 917–925. Curran Associates, Inc., 2012.

Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012, IEEE 53rd Annual Symposium*, pages 1–10. IEEE, 2012.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David M Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *ICML (2)*, pages 280–288, 2013.

Xin Bing, Florentina Bunea, and Marten Wegkamp. A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli*, 26(3):1765–1796, 08 2020. doi: 10.3150/19-BEJ1166. URL `https://doi.org/10.3150/19-BEJ1166`.

David M. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 55:77–84, 2012.

David M. Blei and John D. Lafferty. A correlated topic model of science. *Ann. Appl. Stat.*, 1(1):17–35, 06 2007. doi: 10.1214/07-AOAS114.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

Weicong Ding, Mohammad Hossein Rohban, Prakash Ishwar, and Venkatesh Saligrama. Topic discovery through data dependent and random projections. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1202–1210, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `http://proceedings.mlr.press/v28/ding13.html`.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. K. Saul, and P. B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 1141–1148. MIT Press, 2004.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0307752101.

Thomas Hofmann. Probabilistic latent semantic indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference*, 1999.

Tracy Zheng Ke and Minzhe Wang. A new svd approach to optimal topic estimation. *arXiv:1704.07016*, 2017.

Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML 2006, pages 577–584, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143917.

Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, PODS '98, pages 159–168, New York, NY, USA, 1998. ACM. ISBN 0-89791-996-3. doi: 10.1145/275487.275505.

Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.2000.1711.

Ben Recht, Christopher Re, Joel Tropp, and Victor Bittorf. Factoring nonnegative matrices with linear programs. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1214–1222. Curran Associates, Inc., 2012. URL `http://papers.nips.cc/paper/4518-factoring-nonnegative-matrices-with-linear-programs.pdf`.

Allen Riddell, Timothy Hopper, and Andreas Grivas. lda: 1.0.4, July 2016. URL `https://doi.org/10.5281/zenodo.57927`.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009. doi: 10.1007/b13794.