Submitted to the Annals of Statistics

SUPPLEMENT TO "LIKELIHOOD ESTIMATION OF SPARSE TOPIC DISTRIBUTIONS IN TOPIC MODELS AND ITS APPLICATIONS TO WASSERSTEIN DOCUMENT DISTANCE CALCULATIONS"

BY XIN BING[1], FLORENTINA BUNEA[2,*], SETH STRIMAS-MACKEY[2,†] AND MARTEN WEGKAMP[3]

[1]*Department of Statistical Sciences, University of Toronto, xin.bing@utoronto.ca*

[2]*Department of Statistics and Data Science, Cornell University, *fb238@cornell.edu; †scs324@cornell.edu*

[3]*Departments of Mathematics, and of Statistics and Data Science, Cornell University, mhw73@cornell.edu*

Appendix A contains results on the support recovery of T_* for both known A and unknown A. Appendix B contains the data analysis of the IMDB data set while its supplementary results are collected in Appendix C. Simulation results on estimation of T_* and Π_* are presented in Appendix D while semi-synthetic simulations to compare document-distance estimation rates are stated in Appendix E. All the proofs are collected in Appendices F – I. Appendix J contains the algorithm used for estimating the word-topic matrix A. Appendix K states guarantees on estimation of A based on some existing results. Finally, discussion on the ℓ_2 convergence rate of estimating T_* is stated in Appendix L.

APPENDIX A: RECOVERY OF THE SUPPORT OF T_*

A.1. Support recovery when A is known. We discuss the consistent support recovery of the estimator $\widehat{T}_{\mathrm{mle}}$, and introduce another simple consistent estimator of $S_* = \mathrm{supp}(T_*)$ in the presence of anchor words.

In light of Theorem 5, establishing consistent support recovery for $\widehat{T}_{\mathrm{mle}}$ also requires the other direction, $\mathrm{supp}(T_*) \subseteq \mathrm{supp}(\widehat{T}_{\mathrm{mle}})$, for which we provide a simple sufficient condition below in the presence of anchor words.

PROPOSITION A.1 (Consistent support recovery of $\widehat{T}_{\mathrm{mle}}$). *Suppose there exists at least one anchor word j_k for each topic $k \in S_*$ such that $\Pi_{*j_k} \geq 2\varepsilon_{j_k}$ with ε_{j_k} defined in (7). Then, with probability $1 - 2p^{-1}$,*

$$\mathrm{supp}(T_*) \subseteq \mathrm{supp}(\widehat{T}_{\mathrm{mle}}).$$

Furthermore, if additionally (24) holds, then, with probability $1 - 2p^{-1} - 6s^{-1} - 2K^{-1}$,

$$\mathrm{supp}(T_*) = \mathrm{supp}(\widehat{T}_{\mathrm{mle}}).$$

Proposition A.1 imposes a signal condition on the frequency of the anchor words corresponding to the non-zero topics. Recall ε_{j_k} from (7) that the signal condition simply requires

$$\Pi_{*j_k} \gtrsim \frac{\log(p)}{N}, \quad \text{for one anchor word } j_k \text{ of topic } k \in S_*.$$

In addition to the above signal condition, if Assumption 1 holds (or equivalently, there exists at least one anchor word for each of the zero topics, that is, the topic $k \in S_*^c$), then the following simple estimator

$$\widehat{S} := \{k \in [K] : \exists X_j > 0 \text{ corresponding to anchor word } j \text{ of topic } k\} \tag{A.1}$$

consistently estimates S_*, as stated in the following proposition.

PROPOSITION A.2. *Under Assumption 1, we have $\widehat{S} \subseteq S_*$ with probability one. Furthermore, if additionally there exists at least one anchor word j_k for each topic $k \in S_*$ such that $\Pi_{*j_k} \geq 2\varepsilon_{j_k}$ with ε_{j_k} defined in (7). Then,*

$$\mathbb{P}\{\widehat{S} = S_*\} \geq 1 - 2p^{-1}.$$

PROOF. To show $\widehat{S} \subseteq S_*$, if $k \in \widehat{S}$, then we must have $k \in S_*$. This is because if $k \notin S_*$, with probability one, we couldn't have observed any anchor word $X_j > 0$ of topic k as $\Pi_{*j} = A_{jk}T_{*k} = 0$. Conversely, to show $S_* \subseteq \widehat{S}$, if $k \in S_*$ and there exists a $\Pi_{*j_k} > 2\varepsilon_{j_k}$, then on the event $\mathcal{E}$, $X_{j_k} \geq \Pi_{*j_k} - |X_{j_k} - \Pi_{*j_k}| > \varepsilon_{j_k} > 0$, that is, $k \in \widehat{S}$. This completes the proof. □

The estimator $\widehat{S}$ simply collects the topics for which we have observed anchor words. Proposition A.1 ensures that we always have $\widehat{S} \subseteq S_*$ under Assumption 1. In practice, this property is helpful to check whether $\mathcal{E}_{\text{supp}}$ holds. Specifically, if Assumption 1 holds and we find $\text{supp}(\widehat{T}_{\text{mle}}) \subseteq \widehat{S}$, then we necessarily have $\text{supp}(\widehat{T}_{\text{mle}}) \subseteq \text{supp}(T_*)$.

A.2. Support recovery when A is unknown. Regarding the consistent support recovery of $\widehat{T}$, we remark that the results in Section A.1 continue to hold provided that the anchor words can be consistently estimated. Consistent estimation of the anchor words has been fully established in Bing, Bunea and Wegkamp (2020a). Also, see, Bittorf et al. (2012); Arora et al. (2013) for other procedures of estimating anchor words.

APPENDIX B: APPLICATION: IMDB MOVIE REVIEWS

In this section we demonstrate our proposed approach of estimating topic proportions for use in document distance estimation. Using a popular movie-review dataset (Maas et al., 2011), we perform the following steps:

1. Estimate the word-topic matrix A using the method in Bing, Bunea and Wegkamp (2020b). For the reader's convenience, we restate the procedure in Appendix J. Use anchor words defined via $\widehat{A}$ to give an initial interpretation of each topic.
2. Estimate the topic distributions $\widehat{T}^{(i)}$ from $\widehat{A}$ and $X^{(i)}$, for each document $i \in [n]$, by solving (29). Use these estimates, in the context of the corpus, to adjust and refine the initial topic interpretation.
3. Calculate document distances (51) – (53), along with other candidate distances, and compare their ability to capture similarity between the documents.

Data and preprocessing. We use a collection of 50K IMDB movie reviews designed for unsupervised learning from the Large Movie Review Dataset (Maas et al., 2011). We preprocess the data by removing stop words and words that have document frequency of less than 1%. Among the remaining 1685 words, we keep only the 500 most common (by term frequency), for ease of interpretation of the topics (we found qualitatively similar results and reached the same conclusions when including all 1685 words). We also only keep documents with greater than 50 words. After preprocessing, we end up with a $p \times n$ word-count matrix $\boldsymbol{X}$, where $p = 500$, $n = 20,605$.

REMARK B.1.

(1) We recall from Section 1.1 that one motivation of our theoretical analysis of the estimation of $T_*^{(i)}$ is to address the case when $\Pi_{*j}^{(i)} = 0$ for a document i and word j. After

TABLE 1

Excerpts from documents that are estimated to be exclusively generated from Topics 3 and 5 (formally, documents with $\widehat{T}_k^{(i)} = 1$ for each topic k). The third column gives the ID number in the original dataset (Maas et al., 2011). See Table 6 in Appendix C for further excerpts for all 6 topics.

Topic	Interpretation	Movie ID	Document excerpt
Topic 3	Video Games	23,753	*This game really is worth the ridiculous prices out there...*
		12,261	*I remember playing this game at a friend...*
Topic 5	TV Shows	32,315	*I used to watch this show when I was a little girl...*
		10,454	*I've watched the TV show Hex twice over and I still can not get enough of it. The show is excellent...*

preprocessing, the total number of distinct words in each review in this dataset is 63 on average, much less than the vocabulary size $p = 500$. Thus, for each document i there are typically many words j with $X_j^{(i)} = 0$. For at least some of these words, it is possible that $\Pi_{*j}^{(i)} = 0$. For example, we find reviews of films in genres such as horror and comedy that have no relation to 'war', one of the 500 words in the vocabulary: for these reviews, it is reasonable to expect the word 'war' to have cell probability $\Pi_{*j}^{(i)} = 0$. These observations provide a real-data example further motivating the need for a theoretical analysis allowing for this case.

(2) We also emphasize that our discrete mixture probability estimates allow us to construct non-zero estimates of non-zero $\Pi_{*j}^{(i)}$, even when $X_j^{(i)} = 0$. In fact, we find that the average number of non-zero entries in the estimator $\widetilde{\Pi}^{(i)}$, over all documents $i \in [n]$, is 490, much larger than the average number of non-zero entries of $X^{(i)}$ (which we recall was 68). In most cases, we found zero entries of $\widetilde{\Pi}^{(i)}$ correspond to anchor words for topics that are not present in document i. This demonstrates that $\widetilde{\Pi}^{(i)}$ is able to produce zero estimates for words that we expect to have no chance of occurring in document i, while still producing non-zero estimates corresponding to words that could occur in that document, but were not observed in that particular sample.

B.1. Estimating topic distributions for a refined understanding of the topics covered by a document corpus. We run the method in Bing, Bunea and Wegkamp (2020b) on $\boldsymbol{X}$ to estimate A for this dataset, with tuning parameter $C_1 = 4$, and denote the output $\widehat{A}$. The number of topics is estimated to be $\widehat{K} = 6$. In Table 4 in Appendix C, we show the anchor words for each of the 6 topics, from which we can give an initial interpretation to the topics (shown in the third column of the table). In particular, the only anchor words for Topics 3 and 5 are 'game' and 'episode' respectively, despite this dataset nominally being composed of reviews of full-length movies.

To further interpret the topics (in particular Topics 3 and 5), we compute the estimated topic proportions $\widehat{T}^{(i)}$ from $\widehat{A}$ and $X^{(i)}$ for each document $i \in [n]$ using (29). Table 6 in Appendix C shows, for each $k \in [\widehat{K}]$, examples of documents such that $\widehat{T}_k = 1$; namely, documents that are generated entirely from topic k. This table demonstrates the usefulness of estimating the topic proportions $\widehat{T}^{(i)}$: inspecting these topic-specific documents provides detailed information on what each topic captures. For space limitations, we only give an excerpt of Table 6 here in Table 1, featuring Topic 3 and 5. We find that the documents displayed for Topics 3 and 5 are in fact not movie reviews, but reviews of video games and TV shows, respectively.

Besides these non-movie reviews, we confirm that the examples from Topics 1, 4, and 6 are indeed book adaptations, horror films, and films related to war and history, respectively. We see that Topic 2 is indeed related to sentiment in these examples, with both reviews being very negative. All details can be found in Table 6 in Appendix C.

In summary, we have demonstrated that the estimated topic proportions $\widehat{\boldsymbol{T}}$ are useful tools for topic interpretation, and a needed companion to the estimation of A, on the basis of which one gives the initial definition of the topics.

B.2. Estimating the 1-Wasserstein distance between documents. We recall that, by abuse of terminology, but for clarity of exposition, we refer to distances between probabilistic representations of documents as distances between documents.

We now compare a set of candidate document distance measures, including our proposed methods. We select several representative documents among the documents kept from the IMDB dataset after preprocessing, and compute the distance between them. We recall that in order to compute the 1-Wasserstein distance between two documents represented via their respective topic-distributions, we need to first calculate the distance between elements on their supports, the topics, which in turn are probability distributions on words, estimated by the columns of $\widehat{A}$. Therefore, with $\widehat{K} = 6$, we first compute (52) and (53), which we repeat here for clarity:

$$\text{(B.1)} \qquad \widehat{D}_W^{\text{topic}}(k,l) = W_1(\widehat{A}_{\cdot k}, \widehat{A}_{\cdot l}; D^{\text{word}}), \qquad \widehat{D}_{TV}^{\text{topic}}(k,l) = \frac{1}{2}\|\widehat{A}_{\cdot k} - \widehat{A}_{\cdot l}\|_1,$$

for all $k, l \in \{1, \ldots, 6\}$. To compute D^{word}, we use open-source word embeddings from Google[1] that come pre-trained using the word2vec model (Mikolov et al., 2013) on a Google News corpus of around 100 billion words. These word embeddings contain a word vector x_i for each the 500 words in our dictionary, except one item in the vocabulary (the number '10', common in movie ratings out of 10), for which we remove the corresponding row from $\widehat{A}$ (then re-normalize to have unit column sums) when computing $\widehat{D}_W^{\text{topic}}$. We follow standard practice of normalizing all word-embeddings to unit length. The distance $D^{\text{word}}(i,j)$ between words i and j is then computed as $D^{\text{word}}(i,j) = \|x_i - x_j\|_2 / \max_{i,j} \|x_i - x_j\|_2$. We divide by the normalizing factor $\max_{i,j} \|x_i - x_j\|_2$ so that the elements of D^{word} are in the range $[0,1]$. This results in $\widehat{D}_W^{\text{topic}}$ also being in the range $[0,1]$, and so on the same scale as $\widehat{D}_{TV}^{\text{topic}}$.

See Table 2 for details on each document, including the estimated topic proportions, and Table 3 for the computed distances. We make several remarks based on the results in Table 3.

1. Consider the distance between documents D_1 and D_2, which have $\widehat{T}^{(1)} = \widehat{T}^{(2)}$ and are both entirely generated from the Horror topic. Since $\widehat{T}^{(1)} = \widehat{T}^{(2)}$, all distances between the topic proportions (panels (a), (b), and (e)) are equal to zero. Since $\widehat{T}^{(1)} = \widehat{T}^{(2)}$ implies $\widetilde{\Pi}^{(1)} = \widetilde{\Pi}^{(2)}$, the distance based on the latter estimators is also zero (Table 3, panel (d)). The only distance that does not capture this underlying topical similarity is the Word Mover's Distance (WMD), which has a value of 0.56 between D_1 and D_2.
2. Compare the distances between D_4 (a video game review) and each other document (all movie reviews) to the distances between pairs of movie reviews. For the two Wasserstein distances between the topic proportions, as well as the distance based on $\widetilde{\Pi}$ (Table 3, panels (a), (b), and (d), respectively), the distance between the video game review D_4 and any other review is much greater than the distance between any two movie reviews. (The one exception is that the distance between D_4 and D_6 is not large, since D_6 has substantial weight on the Video game topic). Thus, these methods are able to detect the difference between video game and movie reviews. We similarly find that these methods detect documents from the TV show topic as outliers from full-length movie reviews, but don't include this in Table 3 for simplicity of presentation.

[1] https://code.google.com/archive/p/word2vec/

In contrast, the WMD in panel (c) computes the distances between all pairs of distinct documents to be all relatively close together. In fact, based on the WMD, the Horror film review D_1 is the same distance to the Video game review D_4 as the War & History film review D_3; we note that this is perhaps unsurprising, given that the WMD is not designed to capture similarity based on topics. On the other hand, the TV distance in panel (e) computes the distance between any two documents with disjoint topics to be the maximum value of 1, not distinguishing between topics that are more or less similar.

3. The Wasserstein distance based on $\widehat{D}^{\text{topic}}_{TV}$ (panel (a) of Table 3) gives qualitatively similar results to the other two model-based Wasserstein distances (panels (b) and (d)), while obviating the need for the pre-trained word embeddings used to compute D^{word} and $\widehat{D}^{\text{topic}}_{W}$, and the calculation of any p-dimensional Wasserstein distances, which are computationally expensive.

In summary, the three Wasserstein-based distances defined with the estimated parameters of the topic model (panels (a), (b), (d) in Table 3) are the most successful in capturing topic-based document similarity, and the distance based on $\widehat{D}^{\text{topic}}_{TV}$ (panel (a)) has the further benefit of not requiring the use of pre-trained word embeddings or p-dimensional optimization.

TABLE 2

For each document in Table 3, we give the document ID from the original IMDB dataset, the topic proportions (estimated using (29)), and the interpretations of each topic in the document.

Document	ID	Topic proportions	Topic interpretations
D_1	29,114	$\widehat{T} = (0, 0, 0, 1, 0, 0)$	Horror
D_2	3,448	$\widehat{T} = (0, 0, 0, 1, 0, 0)$	Horror
D_3	26,918	$\widehat{T} = (0, 0, 0, 0, 0, 1)$	War & History
D_4	23,753	$\widehat{T} = (0, 0, 1, 0, 0, 0)$	Video games
D_5	4,058	$\widehat{T} = (0, 0, 0, 0.5, 0, 0.5)$	Horror + War & History
D_6	5,977	$\widehat{T} = (0, 0, 0.5, 0.5, 0, 0)$	Horror + Video games

APPENDIX C: SUPPLEMENTARY RESULTS ON IMDB DATA

In this section we present further details of our analysis in Section B of the IMDB movie review dataset. We first give Table 4, which gives an initial interpretation to each estimated topic based on its anchor words.

TABLE 4

Anchor words for each topic in the IMDB dataset, along with an initial interpretation of each topic.

Topic	Anchor words	Initial interpretation
1	book, read, version	Book adaptations
2	crap, talent	Sentiment
3	game	Game-related
4	blood, dark, dead, evil, fans, flick, genre, gore, horror, house, killer, sequel, strange	Horror films
5	episode	TV Shows
6	history, war	History & war films

Table 5 gives the computed values of the two topic-distance matrices (B.1) and shows that these distances qualitatively capture the same similarity relationships between the topics. This is despite the fact that $\widehat{D}^{\text{topic}}_{W}$ incorporates word similarity from pre-trained word embeddings, whereas $\widehat{D}^{\text{topic}}_{TV}$ depends only parameters estimated directly from the IMDB corpus.

TABLE 3
Distances between documents using various metrics.

(a) $W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}, \widehat{D}^{\text{topic}}_{TV})$

	D_1	D_2	D_3	D_4	D_5	D_6
D_1	0	0	0.14	0.21	0.07	0.10
D_2	·	0	0.14	0.21	0.07	0.10
D_3	·	·	0	0.23	0.07	0.18
D_4	·	·	·	0	0.22	0.10
D_5	·	·	·	·	0	0.11
D_6	·	·	·	·	·	0

(b) $W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}, \widehat{D}^{\text{topic}}_{W})$

	D_1	D_2	D_3	D_4	D_5	D_6
D_1	0	0	0.10	0.16	0.05	0.08
D_2	·	0	0.10	0.16	0.05	0.08
D_3	·	·	0	0.17	0.05	0.13
D_4	·	·	·	0	0.16	0.08
D_5	·	·	·	·	0	0.09
D_6	·	·	·	·	·	0

(c) $W_1(\widehat{\Pi}^{(i)}, \widehat{\Pi}^{(j)}, D^{\text{word}})$ (WMD)

	D_1	D_2	D_3	D_4	D_5	D_6
D_1	0	0.56	0.62	0.62	0.56	0.60
D_2	·	0	0.66	0.68	0.63	0.64
D_3	·	·	0	0.71	0.61	0.67
D_4	·	·	·	0	0.65	0.63
D_5	·	·	·	·	0	0.59
D_6	·	·	·	·	·	0

(d) $W_1(\widetilde{\Pi}^{(i)}, \widetilde{\Pi}^{(j)}, D^{\text{word}})$

	D_1	D_2	D_3	D_4	D_5	D_6
D_1	0	0	0.10	0.16	0.05	0.08
D_2	·	0	0.10	0.16	0.05	0.08
D_3	·	·	0	0.17	0.05	0.12
D_4	·	·	·	0	0.16	0.08
D_5	·	·	·	·	0	0.09
D_6	·	·	·	·	·	0

(e) $TV(\widehat{T}^{(i)}, \widehat{T}^{(j)})$

	D_1	D_2	D_3	D_4	D_5	D_6
D_1	0	0	1	1	0.50	0.50
D_2	·	0	1	1	0.50	0.50
D_3	·	·	0	1	0.50	1
D_4	·	·	·	0	1	0.50
D_5	·	·	·	·	0	0.50
D_6	·	·	·	·	·	0

Finally, we give Table 6, which gives excerpts of two documents for each topic that are estimated to be exclusively generated from that topic. These excerpts allow for further interpretation of the topics.

APPENDIX D: SIMULATIONS ON THE ESTIMATION OF T_* AND Π_*

In this section we present a simulation study of the estimation of topic proportions T_* and word-distribution Π_* to accompany our theoretical analysis in Section 2.

We perform simulations to study the performance of the MLE in (4) (known A) and the estimator in (29) (unknown A) for estimating T_*, and compare to the Restricted Least Squares (RLS) estimator,

$$\widehat{T}_{\text{rls}} := \min_{T \in \Delta_K} \|X - \widehat{A}T\|_2^2, \tag{D.1}$$

as well as the iterative weighted restricted least squares (IWRLS) estimator, both mentioned in Remark 5. To compute the IWRLS estimator for *known* A, we use the following steps (for unknown A, we just replace A by $\widehat{A}$ estimated using the Sparse-TOP method of Bing, Bunea and Wegkamp (2020b)). We use the parameter $\varepsilon = 10^{-8}$ to avoid division by zero, $\delta = 10^{-4}$ as a stopping criterion, and $m_{it} = 1000$ as the maximum number of iterations.

1. Compute $\widehat{T}_{\text{rls}}$ from (D.1) and set $\widehat{T} = \widehat{T}_{\text{rls}}$.

TABLE 5
The two $\widehat{K} \times \widehat{K}$ matrices of distances between topics.

(a) $\widehat{\boldsymbol{D}}_{TV}^{\text{topic}}$

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**
Topic 1	0	0.14	0.22	0.13	0.20	0.14
Topic 2	0.14	0	0.22	0.14	0.22	0.17
Topic 3	0.21	0.22	0	0.21	0.22	0.23
Topic 4	0.13	0.14	0.21	0	0.21	0.14
Topic 5	0.20	0.22	0.22	0.21	0	0.22
Topic 6	0.14	0.17	0.23	0.14	0.22	0

(b) $\widehat{\boldsymbol{D}}_{W}^{\text{topic}}$

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**
Topic 1	0	0.10	0.16	0.10	0.15	0.10
Topic 2	0.10	0	0.17	0.10	0.16	0.12
Topic 3	0.16	0.17	0	0.16	0.17	0.17
Topic 4	0.10	0.10	0.16	0	0.16	0.10
Topic 5	0.15	0.16	0.17	0.16	0	0.17
Topic 6	0.10	0.12	0.17	0.10	0.17	0

2. Let
$$\widehat{D} = \text{diag}(d_1, \ldots, d_p), \qquad \text{with} \quad d_j = \frac{1}{\sqrt{(A\widehat{T})_j \vee \varepsilon}}.$$
3. Update $\widehat{T}$ as
$$\widehat{T} \leftarrow \underset{T \in \Delta_K}{\arg\min} \|\widehat{D}(X - AT)\|_2^2.$$
4. Repeat Steps 2 and 3 until either the ℓ_1 distance between $\widehat{T}$ from the current step and the previous step is less than δ, or a maximum of m_{it} iterations have been completed, then take $\widehat{T}_{\text{iwrls}} = \widehat{T}$ as the final estimator.

We then compare model-based estimators of Π_* based on estimates of T_* to the empirical estimate of Π_*. In terms of notation, recall that N is the number of words in a document, n is the number of documents in the corpus, K is the number of topics, and p is the dictionary size.

Data generating mechanism. For fixed anchor word sets $I_1, \ldots, I_K \subset [p]$, we generate the $p \times K$ matrix A as follows. We set $A_{ik} = K/p$ for all $i \in I_k$ and $k \in [K]$. Draw all entries of non-anchor words from Uniform$(0, 1)$, then normalize each sub-column $A_{I^c k}$ to have sum $1 - \sum_{i \in I_k} A_{ik}$ where $I^c = [p] \setminus (\cup_k I_k)$. We choose balanced anchor word sets such that $|I_k| = m_{anc}$ for all $k \in [K]$. We choose $m_{anc} = 5$, $K = 20$, and $p = 1500$ for all experiments in this section.

For fixed support size s, we generate $T_*^{(1)}, \ldots, T_*^{(n)}$ identically and independently as follows. For each $T_*^{(i)}$, select a subset $S \subset [K]$ with $|S| = s$ by drawing elements from $[K]$ uniformly at random without replacement. Set $[T_*^{(i)}]_{S^c} = 0$ and generate each entry of $[T_*^{(i)}]_S$ independently from Uniform$(0, 1)$. Finally, normalize $T_*^{(i)}$ so its entries sum to 1. The result is that $T_*^{(i)}$ has support s for all $i \in [n]$.

TABLE 6

Excerpts from documents that are estimated to be exclusively generated from Topics 3 and 5 (formally, documents with $\widehat{T}_k^{(i)} = 1$ for each topic k). Excerpts from two separate documents with this property are given for each topic. The third column gives the ID number in the original dataset (Maas et al., 2011) for reference. We find that the "movie reviews" corresponding to Topics 3 and 5 are in fact reviews of video games and TV shows, respectively.

Topic	**Interpretation**	**Movie ID**	**Document excerpt**
Topic 1	Book Adaptations	37,123	*This was an OK movie, at best, outside the context of the book. But having read and enjoyed the book quite a bit it was a real disappointment in comparison...*
		15,709	*This has always been one of my favourite books. I was thrilled when I saw that the book had been made into a movie, for the first time since it was written, over 50 years before...*
Topic 2	Negative reviews	12,445	*This was the most disappointing films I have ever seen recently. And I really hardly believe that people say goods things about this very bottom film!...*
		32,442	*Acting was awful. Photography was awful. Dialogue was awful. Plot was awful. (I'm not being mean here...It really was this bad.)...*
Topic 3	Video Games	23,753	*This game really is worth the ridiculous prices out there. The graphics really are great for the SNES, though the magic spells don't look particularly great...*
		12,261	*I remember playing this game at a friend. Watched him play a bit solo until we decided to try play 2 and 2, which we found out how to do...*
Topic 4	Horror	29,114	*After watching such teen horror movies as Cherry Falls and I know what you did last summer, I expected this to be similar...*
		3,448	*Being a HUGE fan of the horror genre, I have come to expect and appreciate cheesey acted, plot-holes galore, bad scripts...*
Topic 5	TV Shows	32,315	*I used to watch this show when I was a little girl...*
		10,454	*I'v watched the TV show Hex twice over and I still can not get enough of it. The show is excellent...*
Topic 6	War & History	6,709	*Carlo Levi, an Italian who fought against the arrival of Fascism in his native Torino, was arrested for his activities...*
		26,918	*As directed masterfully by Clint Eastwood, "Flags of Our Fathers" plays both as a war film and a sensitive human drama...*

For each choice of A and $T_*^{(1)}, \ldots, T_*^{(n)}$, we then set $\Pi_*^{(i)} = AT_*^{(i)}$ for $1 \le i \le n$ and generate $NX^{(i)} \sim \text{Multinomial}_p(N, \Pi_*^{(i)})$. We report the ℓ_1 error of the estimation of T_* by each method averaged over all 100 repetitions of the simulation.

Estimation of T_* with known A. We first compare the MLE in (4), RLS, and IWRLS when A is known. In this case we can take $n = 1$ and drop the superscript on T_*, X, and Π_*. To see the impact of sparsity in these methods, we also include a baseline estimator for each method that corresponds to the support of T_* being exactly known. To be precise, let S_* be the support of T_*, and $A_{\cdot S_*}$ the $p \times |S_*|$ submatrix of A with columns restricted to the support of T_*. The baseline estimators are computed by estimating T_{S_*} using the MLE, RLS, or IWRLS with $A_{\cdot S_*}$ and X as input, and estimating $T_{S_*^c}$ by zeroes. We plot the ℓ_1 error of estimation of T_* as a function of N and $s = |S_*|$ in Figure 1.

Results. In the left panel Figure 1 we see how the ℓ_1 error of all estimators decays as the document length increases. On the other hand, we see in the right panel that the error increases as the support size of T_* increases. We observe in both panels the remarkable feature that the MLE with unknown support has nearly identical risk to the MLE with exactly known support, empirically illustrating Theorem 5. In contrast, the RLS with unknown support performs substantially worse than with known support, illustrating that the RLS does not enjoy the same support recovery properties as the MLE. Comparing the MLE and RLS, both with

unknown support, we also observe that the risk of the MLE is uniformly lower than that of the RLS. This gives support to the MLE being a clearly superior estimator of T_*.

For the simulation parameters in Figure 1, IWRLS has approximately equal risk to that of the MLE. This in line with the fact the IWRLS is asymptotically (as $N \to \infty$) equivalent to the MLE; see, for instance, Bishop, Fienberg and Holland (2007) and Agresti (2012). However, in Figure 2, we plot the error of the MLE and IWRLS for smaller values of N, where the asymptotic equivalence of these two methods breaks down. We see that the MLE has lower error for small N, a regime of practical interest corresponding to short documents. Furthermore, from Table 7, we observe that in the worst case, the IWRLS has a computation time of around two orders of magnitude more than that of the MLE (while the RLS has the lowest computation time of all). Lastly, from Table 8 we see that for small N, the IWLS did not converge within 1000 iterations for a large proportion of runs (we found similar results, with even longer run times, when increasing the maximum allowed iterations for the IWRLS). The increased computation time of the IWRLS relative to the MLE, along with our observation that the error is either equal (for large N) or greater (for small N), give strong support for the MLE being preferred as an estimator of T_*.

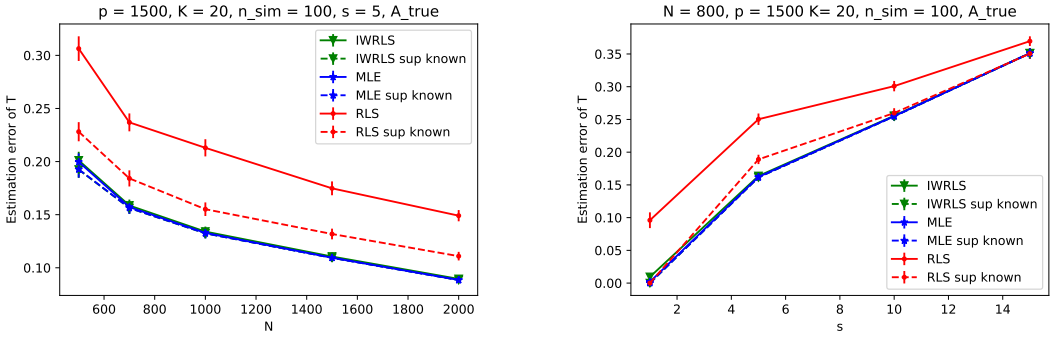

Fig 1: ℓ_1 error of the estimation of T_* for the MLE, RLS, and IWRLS, as a function of document length N (left) and support size s (right), when A is known. Dashed lines correspond to the predictors when the true support of T_* is known. The error for the MLE and IWRLS are approximately equal for these simulation settings, for both known and unknonwn support.

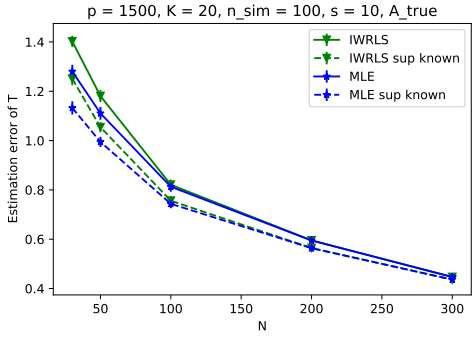

Fig 2: ℓ_1 error of the estimation of T_* for the MLE and IWRLS, for small values of N, when A is known. Dashed lines correspond to the predictors when the true support of T_* is known.

Method	$N = 30$	$N = 50$	$N = 100$	$N = 200$	$N = 300$
MLE	0.68	0.64	0.60	0.57	0.54
RLS	0.08	0.08	0.08	0.08	0.08
IWRLS	61.87	43.31	31.19	9.63	3.44

TABLE 7

Average computation time for each method (in seconds) from the simulation in Figure 2.

$N = 30$	$N = 50$	$N = 100$	$N = 200$	$N = 300$
40%	31%	23%	5%	0%

TABLE 8

Percentage of IWRLS runs from the simulation in Figure 2 that reached the maximum number of iterations (1000) and did not converge.

Estimation of T_* with unknown A. We next compare the estimator in (29), RLS, and IWRLS when A is unknown and estimated by $\widehat{A}$ from the Sparse-TOP method Bing, Bunea and Wegkamp (2020b). For all values of n, we choose the first document $T_*^{(1)}$ to estimate (this choice is arbitrary, as $T_*^{(i)}$ is drawn from an identical distribution for all $i \in [n]$). In Figure 3, we plot the average ℓ_1 error of estimating $T_*^{(1)}$ as a function of N and s. We also include the MLE in (4) with known A for comparison. We refer to the estimator (29) as MLE-A-hat in the plot.

Results. Similarly to the case with known A, we see in Figure 3 that the estimator (29) uniformly outperforms the RLS, and for the (large) values of N in the plot, the IWLS has approximately equal risk to estimator (29). Comparing the MLE with known and unknown A, we observe in the left panel that the impact of not knowing A decreases as the document length increases, which is expected as longer documents improve the estimate $\widehat{A}$. We also observe in the right panel that not knowing A has less impact for T_* that are more sparse.

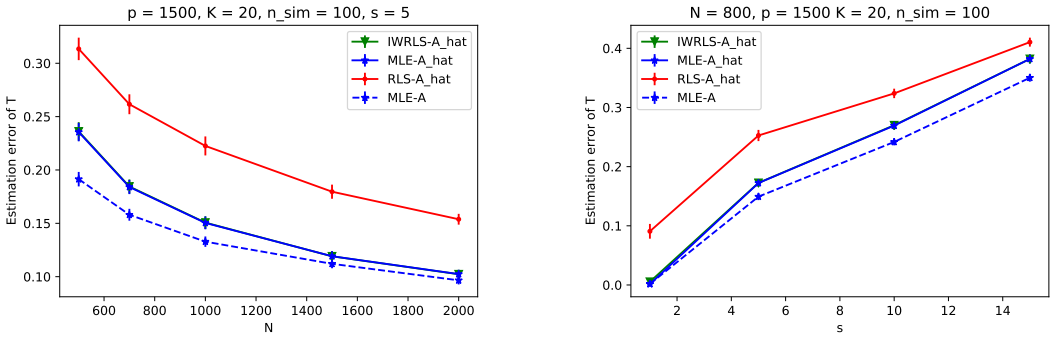

Fig 3: ℓ_1 error of the estimation of T_* for the MLE, RLS, and IWRLS, as a function of document length N (left) and support size s (right). Solid lines correspond to estimators using $\widehat{A}$, and the dashed line corresponds to the MLE with known A. The error for the MLE and IWRLS (both for unknown A) are approximately equal in these plots.

Estimation of Π_*. We compare the three estimators of Π_* presented in Section 2.3: the empirical estimate $\widehat{\Pi}$, and the model-based estimators $\widetilde{\Pi}_A = A\widehat{T}_{\text{mle}}$ and $\widehat{\Pi}_{\widehat{A}} = \widehat{A}\widehat{T}$. Setting $n = 1000$, we repeat the simulation 100 times and plot the average ℓ_1 error in estimating the first document $\Pi_*^{(1)}$ as a function of document length N in the top left panel of Figure

4. Note that $\widehat{\Pi}^{(1)}$ is a function only of the first document vector $X^{(1)}$, and ignores the other $n-1$ documents in the corpus. In contrast, the model-based estimator with unknown A, $\widetilde{\Pi}^{(1)}_{\widehat{A}}$, uses all n documents in the corpus via the estimation of $\widehat{A}$. We drop the superscript 1 in the remainder of this discussion for ease of notation.

Recall the definitions $\bar{J} := \{j : \Pi_{*j} > 0\}$ and $J := \{j : X_j > 0\}$ from Section 2.3. As discussed in that section, a particular advantage of the model-based estimators $\widetilde{\Pi}_A$ and $\widetilde{\Pi}_{\widehat{A}}$ over the empirical estimate $\widehat{\Pi}$ is their ability to non-trivially estimate non-zero cell probabilities with zero counts (Π_{*j} with $j \in \bar{J} \setminus J$), while still estimating the zero cell probabilities $j \in \bar{J}^c$ nearly as well as $\widehat{\Pi}$. We here conduct a simulation to empirically study the ability of each method to estimate these two classes of cell probabilities. In each simulation run, we compute for each estimator Π^{est} among $\widehat{\Pi}$, $\widetilde{\Pi}_A$, and $\widetilde{\Pi}_{\widehat{A}}$ the quantities $\sum_{j \in \bar{J} \setminus J} |\Pi^{\text{est}}_j - \Pi_{*j}|$ and $\sum_{j \in \bar{J}^c} |\Pi^{\text{est}}_j - \Pi_{*j}|$; we plot their average values over 100 runs as a function of N in the top right and bottom panels of Figure 4, respectively.

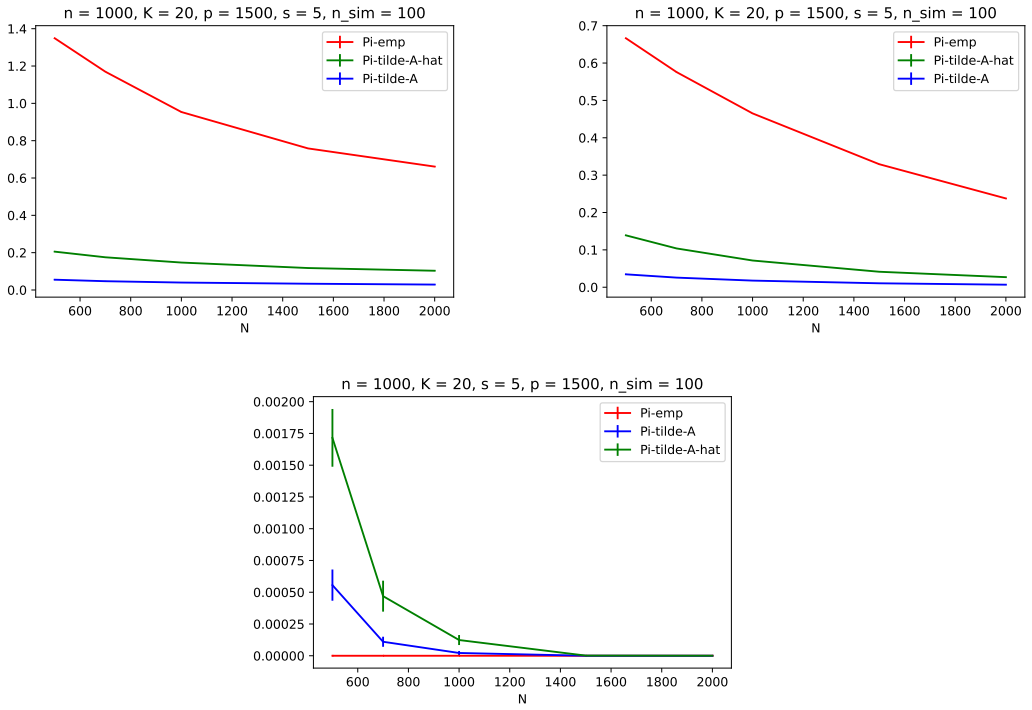

Fig 4: Top left: ℓ_1 error of the estimation of Π_* as a function of document length N. Top right: Error in estimating the cell probabilities Π_{*j} with $j \in \bar{J} \setminus J$. Bottom: Error in estimating Π_{*j} with $j \in \bar{J}^c$. Error bars are present in top plots but too small to observe.

Results. We observe from the left panel of Figure 4 that while the error of all three estimators decays with N, the error of the empirical estimator is substantially larger than that of the model-based estimator $\widetilde{\Pi}_{\widehat{A}}$, which is in turn larger than the error of model-based estimator with known A, $\widetilde{\Pi}_A$, while being very close to it. This demonstrates the basic motivation of the model-based estimation approach: by borrowing statistical strength from across the full corpus, $\widetilde{\Pi}_A$ and $\widetilde{\Pi}_{\widehat{A}}$ provide a far superior estimate of the frequencies for an individual docu-

ment. The difference between $\widetilde{\Pi}_A$ and $\widetilde{\Pi}_{\widehat{A}}$ on the other hand reflects the effect of estimating A.

In the top right panel of Figure 4, we verify that the two model-based estimators are able to estimate the non-zero cell probabilities with zero counts (Π_{*j} with $j \in \bar{J} \setminus J$) much better than the trivial estimate of $\widehat{\Pi}_j = 0$. We note for clarity that for $\Pi^{\text{est}} = \widehat{\Pi}$, $\sum_{j \in \bar{J} \setminus J} |\Pi^{\text{est}}_j - \Pi_{*j}|$ reduces to $\sum_{j:\ \Pi_{*j}>0,\ X_j=0} \Pi_{*j}$. While we see that this quantity decreases with N from the red line in the top right panel of Figure 4, this is simply due to the fact that $|\{j : X_j = 0\}|$ decreases with N.

Lastly, in the bottom panel of Figure 4, we see that while for small N ($N \leq 1500$ for $\widetilde{\Pi}_{\widehat{A}}$) the error in estimating the zero cell probabilities by the model-based estimators is non-zero, it is several orders of magnitude smaller than the overall ℓ_1 error in the top left panel, and in any case quickly decays to zero as N increases. As expected, $\widehat{\Pi}_j = 0$ for all $j \in \bar{J}$, so the error for $\widehat{\Pi}$ is exactly zero in this bottom panel.

In summary, by borrowing statistical strength across the corpus of n documents, the two model-based estimators $\widetilde{\Pi}_A$ and $\widetilde{\Pi}_{\widehat{A}}$ perform substantially better than the empirical estimator at estimating Π_* in ℓ_1 error and estimating non-zero cell probabilities with zero counts, while still having nearly the same performance as the empirical estimator at estimating the zero cell probabilities.

APPENDIX E: SEMI-SYNTHETIC SIMULATIONS TO COMPARE DOCUMENT-DISTANCE ESTIMATION RATES

We perform semi-synthetic simulations to empirically study the rate of estimation of the topic-based document distance (48) for the choice (50) of D^{topic}, by the estimator (53). We also ran the same simulations for the choice (49) of D^{topic} with the estimator (52) and found similar results, which we do not report here due to space limitations.

Data and preprocessing. We work with the NIPS bag-of-words dataset (Dua and Graff, 2017). We preprocess the data by removing stop words, removing documents with less than 150 words, and removing words that appear in less than 150 documents. We are left with 1490 documents and dictionary size $p = 1270$. From this data we estimate a loading matrix A_0 using the Sparse-TOP algorithm (Bing, Bunea and Wegkamp, 2020b) and find $K = 21$ topics. We then treat this estimated A_0 as our ground truth for semi-synthetic experiments.

Semi-synthetic data generation. We generate topic distributions $T_*^{(1)}, \ldots, T_*^{(n)}$ with $K = 21$ exactly following the procedure in Section D. In particular, $T_*^{(1)}, \ldots, T_*^{(n)}$ all have the same support size, which we denote as s. We choose $n = 2000$ for all simulations. For each $i \in [n]$, we set $\Pi_*^{(i)} = A_0 T_*^{(i)}$ and draw $X^{(i)} \sim \text{Mutlinomial}_p(N, \Pi^{(i)})$.

For each simulation, we form the estimate $\widehat{A}$ using Sparse-TOP Bing, Bunea and Wegkamp (2020b), and form estimates $\widehat{T}^{(1)}$, $\widehat{T}^{(2)}$ of the topic distributions of the first two documents using the MLE (29). From $\widehat{A}$, we compute the estimated topic-distance metric

$$\widehat{D}^{\text{topic}}_{TV}(k,l) = \frac{1}{2}\|\widehat{A}_{\cdot k} - \widehat{A}_{\cdot l}\|_1 \quad \forall k,l \in \{1,2,\ldots,21\},$$

and its population counterpart with $\widehat{A}$ replaced by A_0. We then compute the error

$$|W_1(\widehat{T}^{(1)}, \widehat{T}^{(2)}; \widehat{D}^{\text{topic}}_{TV}) - W_1(T_*^{(1)}, T_*^{(2)}; D^{\text{topic}}_{TV})|. \tag{E.1}$$

We repeat this simulation $n_{sim} = 50$ times for different values of N and s, and plot the average error in Figure 5.

Results. We see from Figure 5 that the error (E.1) grows significantly as the support size s of $T_*^{(1)}$ and $T_*^{(2)}$ increases. This is can be understood by the fact that the error in estimating $T_*^{(1)}$ and $T_*^{(2)}$ also increases with s; recall Figure 3 for an empirical demonstration of this. For all values of s, we observe the error decaying as N increases.

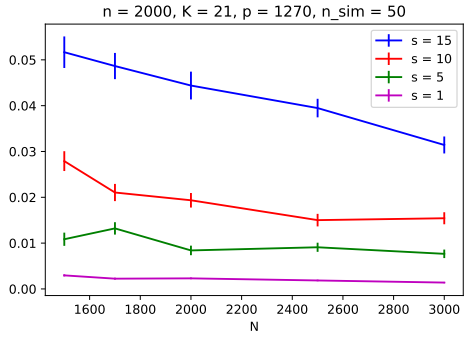

Fig 5: Error (E.1) as a function of N, for different values of support size s of the synthetically generated topic distributions $T_*^{(1)}$ and $T_*^{(2)}$.

APPENDIX F: PROOFS FOR SECTION 2.1: ESTIMATION WITH KNOWN A

Throughout the proofs, we will suppress the subscript $*$ for notational simplicity. Correspondingly, we write $S_T = S_*$ to denote its dependency on T.

F.1. Proof of Theorem 1: The general finite sample bounds of the ℓ_1 norm convergence rate of the MLE. Recall ε_j is defined in (7). Define the event

$$\mathcal{E} := \bigcap_{j=1}^{p} \{|X_j - \Pi_j| \le \varepsilon_j\} \tag{F.1}$$

which, according to Lemma I.1 in Appendix I, holds with probability at least $1 - 2p^{-1}$. On the event $\mathcal{E}$, we have

$$\underline{J} \subseteq J \subseteq \overline{J}.$$

Indeed, $\underline{J} \subseteq J$ follows by noting that, for any $j \in \underline{J}$, $X_j \ge \Pi_j - |X_j - \Pi_j| > \varepsilon_j$. The other direction $J \subseteq \overline{J}$ holds trivially since $\Pi_j = 0$ implies $X_j = 0$ for all $j \in [p]$. We work on the event $\mathcal{E}$ for the remainder of the proof.

For notational simplicity, we write $\widehat{T} = \widehat{T}_{\mathrm{mle}}$. Recall that

$$\widehat{T} := \operatorname*{arg\,max}_{T \in \Delta_K} N \sum_{j \in J} X_j \log\left(A_{j\cdot}^\top T\right).$$

From the KKT conditions of this optimization problem we have

$$N \sum_{j \in J} X_j \frac{A_{j\cdot}}{A_{j\cdot}^\top \widehat{T}} + \lambda + \mu \mathbf{1}_K = 0, \tag{F.2}$$

$$\lambda_k \ge 0, \quad \lambda_k \widehat{T}_k = 0, \ \forall k \in [K], \quad \mathbf{1}_K^\top \widehat{T} = 1. \tag{F.3}$$

After taking the inner-product with $\widehat{T}$ on both sides of (F.2), we get

$$\mu = -N \sum_{j \in J} X_j = -N.$$

Plugging this into (F.2) gives the expression

$$N \sum_{j \in J} X_j \frac{A_{j\cdot}}{A_{j\cdot}^\top \widehat{T}} + \lambda = N \, \mathbf{1}_K.$$

Next, we take the inner-product on both sides with $\Delta := \widehat{T} - T$ and use the fact that $\mathbf{1}_K^\top \Delta = 0$ to obtain

$$N \sum_{j \in J} X_j \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top \widehat{T}} + \lambda^\top \Delta = 0.$$

By adding and subtracting terms, we have

$$\begin{aligned}
0 &= N \sum_{j \in J} X_j \left(\frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top \widehat{T}} - \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top T} \right) + N \sum_{j \in J} \frac{X_j}{A_{j\cdot}^\top T} A_{j\cdot}^\top \Delta + \lambda^\top \Delta \\
&= N \sum_{j \in J} X_j \left(\frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top \widehat{T}} - \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top T} \right) + N \sum_{j \in \overline{J}} \frac{X_j}{A_{j\cdot}^\top T} A_{j\cdot}^\top \Delta + \lambda^\top \Delta \\
&= N \sum_{j \in J} X_j \left(\frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top \widehat{T}} - \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top T} \right) + N \sum_{j \in \overline{J}} \left(X_j - A_{j\cdot}^\top T \right) \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top T} + N \sum_{j \in \overline{J}} A_{j\cdot}^\top \Delta + \lambda^\top \Delta.
\end{aligned}$$

In the second equality, we used $\Pi_j = A_{j\cdot}^\top T > 0$ for $j \in \overline{J}$ and $X_j = 0$ for $j \in \overline{J} \setminus J$. Since

$$\frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top \widehat{T}} - \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top T} = -\frac{(A_{j\cdot}^\top \Delta)^2}{A_{j\cdot}^\top \widehat{T} \cdot A_{j\cdot}^\top T},$$

we conclude

$$\begin{aligned}
N \sum_{j \in J} \frac{X_j}{A_{j\cdot}^\top T} \frac{(A_{j\cdot}^\top \Delta)^2}{A_{j\cdot}^\top \widehat{T}} &= N \sum_{j \in \overline{J}} \left(X_j - A_{j\cdot}^\top T \right) \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top T} + N \sum_{j \in \overline{J}} A_{j\cdot}^\top \Delta + \lambda^\top \Delta \\
&\leq N \sum_{j \in \overline{J}} \left(X_j - A_{j\cdot}^\top T \right) \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top T} + N \sum_{j \in \overline{J}} A_{j\cdot}^\top \Delta \qquad \text{(F.4)}
\end{aligned}$$

by using $\lambda^\top \Delta = -\lambda^\top T \leq 0$ from (F.3) in the last step. For the left hand side in (F.4), use $\underline{J} \subseteq J$ to obtain

$$\sum_{j \in J} \frac{X_j}{A_{j\cdot}^\top T} \frac{(A_{j\cdot}^\top \Delta)^2}{A_{j\cdot}^\top \widehat{T}} \geq \sum_{j \in \underline{J}} \frac{X_j}{A_{j\cdot}^\top T} \frac{(A_{j\cdot}^\top \Delta)^2}{A_{j\cdot}^\top \widehat{T}} \geq \min_{j \in \underline{J}} \frac{X_j}{A_{j\cdot}^\top T} \sum_{j \in \underline{J}} \frac{(A_{j\cdot}^\top \Delta)^2}{A_{j\cdot}^\top \widehat{T}}.$$

Since $\sum_{j=1}^p A_{j\cdot}^\top \widehat{T} = 1$, we further observe that

$$\sum_{j \in \underline{J}} \frac{(A_{j\cdot}^\top \Delta)^2}{A_{j\cdot}^\top \widehat{T}} = \sum_{j \in \underline{J}} \frac{(A_{j\cdot}^\top \Delta)^2}{A_{j\cdot}^\top \widehat{T}} \left(\sum_{j \in \underline{J}} A_{j\cdot}^\top \widehat{T} + \sum_{j \in \underline{J}^c} A_{j\cdot}^\top \widehat{T} \right)$$

$$\begin{aligned}
&\geq \sum_{j\in \underline{J}} \frac{(A_{j\cdot}^\top \Delta)^2}{A_{j\cdot}^\top \widehat{T}} \sum_{j\in \underline{J}} A_{j\cdot}^\top \widehat{T} \\
&\geq \left(\sum_{j\in \underline{J}} |A_{j\cdot}^\top \Delta| \right)^2 \\
&\geq \kappa^2(A_{\underline{J}}, s) \|\Delta\|_1^2.
\end{aligned}$$

Here we use the Cauchy-Schwarz inequality in the third line and the definition (9) of the $\ell_1 \to \ell_1$ condition number $\kappa(A_{\underline{J}}, s)$ together with $\Delta \in \mathcal{C}(S_T)$ in the last line. From the inequality

$$\frac{X_j}{A_{j\cdot}^\top T} \geq \frac{\Pi_j - |X_j - \Pi_j|}{\Pi_j} \geq 1 - \frac{\varepsilon_j}{\Pi_j} \geq \frac{1}{2} \qquad \forall j \in \underline{J},$$

we can now conclude

$$\sum_{j\in J} \frac{X_j}{A_{j\cdot}^\top T} \frac{(A_{j\cdot}^\top \Delta)^2}{A_{j\cdot}^\top \widehat{T}} \;\geq\; \frac{1}{2}\kappa^2(A_{\underline{J}}, s)\|\Delta\|_1^2.$$

It remains to bound from above the right-hand side

$$N \sum_{j\in \overline{J}} \left(X_j - A_{j\cdot}^\top T \right) \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top T} + N \sum_{j\in \overline{J}} A_{j\cdot}^\top \Delta$$

of (F.4). The identity $\sum_{j=1}^p A_{j\cdot}^\top \Delta = \mathbf{1}_K^\top \Delta = 0$ implies

$$\sum_{j\in \overline{J}} A_{j\cdot}^\top \Delta = -\sum_{j\notin \overline{J}} A_{j\cdot}^\top \Delta \leq \sum_{j\notin \overline{J}} A_{j\cdot}^\top T = \sum_{j\notin \overline{J}} \Pi_j = 0. \tag{F.5}$$

The last equality uses the definition of $\overline{J}$. The inequality $u^\top v \leq \|u\|_1 \|v\|_\infty$ gives

$$\sum_{j\in \overline{J}} \left(X_j - A_{j\cdot}^\top T \right) \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top T} \leq \|\Delta\|_1 \max_{k\in[K]} \left| \sum_{j\in \overline{J}} \left(X_j - A_{j\cdot}^\top T \right) \frac{A_{jk}}{A_{j\cdot}^\top T} \right|.$$

By invoking Lemma I.2 in Appendix I with a union bound over $k \in [K]$ to bound the above term, we conclude that, for any $t > 0$,

$$\frac{1}{2}\kappa^2(A_{\underline{J}}, s)\|\Delta\|_1 \leq \sqrt{\frac{2\rho \log(K/t)}{N}} + \frac{2\rho \log(K/t)}{3N}$$

with probability $1 - 2t$. The proof is complete. □

F.2. Proof of Theorem 2: Fast rates of the MLE. To prove Theorem 2, we first work on the event under which Theorem 1 holds, that is,

$$\|\widehat{T}_{\mathrm{mle}} - T\|_1 \;\leq\; \frac{2}{\kappa^2(A_{\underline{J}}, s)} \left\{ \sqrt{\frac{2\rho \log(K/\epsilon)}{N}} + \frac{2\rho \log(K/\epsilon)}{N} \right\}.$$

We write $\widehat{T} = \widehat{T}_{\mathrm{mle}}$ for notational ease for the remainder of the proof. Condition (19) together with $\rho \leq (1 \vee \xi)/T_{\min}$ then ensures that

$$\max_{j\in \overline{J}} \frac{|A_{j\cdot}^\top (\widehat{T} - T)|}{A_{j\cdot}^\top T} \leq \rho \|\widehat{T}_{\mathrm{mle}} - T\|_1 \leq c$$

for some sufficiently small constant $c > 0$. As a result, we can deduce

$$(1-c)A_{j\cdot}^\top T \;\le\; A_{j\cdot}^\top \widehat{T} \;\le\; (1+c)A_{j\cdot}^\top T, \quad \forall j \in \overline{J}. \tag{F.6}$$

Recall from (F.4) that

$$\sum_{j\in J} \frac{X_j}{A_{j\cdot}^\top T} \frac{(A_{j\cdot}^\top \Delta)^2}{A_{j\cdot}^\top \widehat{T}} \le \sum_{j\in\overline{J}} \left(X_j - A_{j\cdot}^\top T\right) \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top T} + \sum_{j\in\overline{J}} A_{j\cdot}^\top \Delta.$$

By (F.6), $A_{j\cdot}^\top \widehat{T} \ge (1-c)\Pi_j > 0$ for all $j \in \overline{J} \setminus J$. Together with $X_j = 0$ for all $j \in \overline{J} \setminus J$, the above display implies

$$\sum_{j\in\overline{J}} \frac{X_j}{A_{j\cdot}^\top T} \frac{(A_{j\cdot}^\top \Delta)^2}{A_{j\cdot}^\top \widehat{T}} \le \sum_{j\in\overline{J}} \left(X_j - A_{j\cdot}^\top T\right) \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top T} + \sum_{j\in\overline{J}} A_{j\cdot}^\top \Delta,$$

which, by (F.6) again, further implies

$$\frac{1}{1+c} \sum_{j\in\overline{J}} \frac{X_j}{(A_{j\cdot}^\top T)^2} (A_{j\cdot}^\top \Delta)^2 \le \sum_{j\in\overline{J}} \left(X_j - A_{j\cdot}^\top T\right) \frac{A_{j\cdot}^\top \Delta}{A_{j\cdot}^\top T}.$$

Define

$$H = \sum_{j\in\overline{J}} \frac{1}{\Pi_j} A_{j\cdot} A_{j\cdot}^\top, \qquad \widehat{H} = \sum_{j\in\overline{J}} \frac{X_j}{\Pi_j^2} A_{j\cdot} A_{j\cdot}^\top. \tag{F.7}$$

Notice that, for any $v \in \mathbb{R}^K$,

$$v^\top H v = \sum_{j\in\overline{J}} \frac{(A_{j\cdot}^\top v)^2}{\Pi_j} = \sum_{j\in\overline{J}} \frac{(A_{j\cdot}^\top v)^2}{\Pi_j} \sum_{j\in\overline{J}} \Pi_j \ge \|A_{\overline{J}} v\|_1^2 \ge \kappa^2(A_{\overline{J}}, K)\|v\|_1^2. \tag{F.8}$$

The first inequality uses the Cauchy-Schwarz inequality. Condition (19) implies $\kappa(A_{\overline{J}}, K) > 0$, hence H is invertible. We thus have

$$\frac{\Delta^\top \widehat{H} \Delta}{1+c} \le \left\| \sum_{j\in\overline{J}} \left(X_j - A_{j\cdot}^\top T\right) \frac{H^{-1/2} A_{j\cdot}}{A_{j\cdot}^\top T} \right\|_2 \|H^{1/2}\Delta\|_2.$$

By invoking Lemma I.3 with $t = 4K\log(5)$ and Lemma I.4 in Appendix I concludes

$$\begin{aligned} \|H^{1/2}\Delta\|_2^2 &\lesssim \|H^{1/2}\Delta\|_2 \left(\sqrt{\frac{K}{N}} + \frac{(1\vee\xi)}{3\kappa(A_{\overline{J}}, K) T_{\min}} \cdot \frac{K}{N} \right) \\ &\lesssim \|H^{1/2}\Delta\|_2 \sqrt{\frac{K}{N}} \qquad \text{by (19)} \end{aligned} \tag{F.9}$$

with probability $1 - 2K^{-1} - 2e^{-K}$. Since (F.8) also implies

$$\|H^{1/2}\Delta\|_2 \ge \kappa(A_{\overline{J}}, s)\|\Delta\|_1, \tag{F.10}$$

by using $\Delta \in \mathcal{C}(S_T)$, the result follows. The proof is complete.

F.3. Proof of Corollary 4: Fast rates of the MLE when it is sparse. On the event $\mathcal{E}_{\text{supp}}$, for any $T \in \mathcal{T}(s)$ with $|S_T| = s$, we observe

$$[\widehat{T}_{\text{mle}}]_{S_T} = \operatorname*{arg\,max}_{T\in\Delta_s} N \sum_{j\in J} X_j \log\left(A_{jS_T}^\top T_{S_T}\right), \qquad [\widehat{T}_{\text{mle}}]_{S_T^c} = 0.$$

Since $\|\widehat{T}_{\text{mle}} - T\|_1 = \|[\widehat{T}_{\text{mle}} - T]_{S_T}\|_1$, the result follows immediately from Corollary 3 with $K = s$. If we take $\log(s\vee n)$ instead of $\log(s)$ in (21) and take $s\log(1/\epsilon)$ instead of $\log(s)$ in the bound, the resulting probability tail becomes $1 - 2p^{-1} - 4(s\vee n)^{-1} - 2\epsilon^s$.

F.4. Proof of Theorem 5: One-sided sparsity recovery of the MLE. For any T with $\mathrm{supp}(T) = S_T$, our proof of $\mathrm{supp}(\widehat{T}_{\mathrm{mle}}) \subseteq \mathrm{supp}(T)$ consists of two parts:

(i) we show that there exists an optimal solution $\widetilde{T}$ to (4) such that $\mathrm{supp}(\widetilde{T}) \subseteq \mathrm{supp}(T)$;
(ii) we show that if there exists any optimal solution $\bar{T}$ to (4) that is different from $\widetilde{T}$, then $\mathrm{supp}(\bar{T}) \subseteq \mathrm{supp}(T)$.

Proof of step (i). Our proof of step (i) uses the primal-dual witness approach by first constructing an oracle estimator $\widetilde{T}$ with $\mathrm{supp}(\widetilde{T}) \subseteq \mathrm{supp}(T)$, and then proving that $\widetilde{T}$ is an optimal solution.

Towards this end, we first notice that any pair $(\widehat{T}, \lambda, \mu)$ is an optimal solution to (4) if and only if it satisfies the KKT condition in (F.2) – (F.3). Having this in mind, we define $\widetilde{T}_{S_T^c} = 0$ and

$$\widetilde{T}_{S_T} = \arg\max_{\beta \in \Delta_s} N \sum_{j \in J} X_j \log\left(A_{jS_T}^\top \beta\right). \tag{F.11}$$

The KKT condition corresponding to (F.11) states

$$N \sum_{j \in J} X_j \frac{A_{jS_T}}{A_{jS_T}^\top \widetilde{T}_{S_T}} + \widetilde{\lambda}_{S_T} + \widetilde{\mu} \mathbf{1}_s = 0; \tag{F.12}$$

$$\widetilde{\lambda}_k \geq 0, \quad \widetilde{\lambda}_k \widetilde{T}_k = 0, \ \forall k \in S_T, \quad \widetilde{T}_{S_T}^\top \mathbf{1}_s = 1. \tag{F.13}$$

Note that (F.12) and (F.13) together imply $\widetilde{\mu} = -N$ by multiplying both sides of (F.12) by $\widetilde{T}_{S_T}$. We thus define

$$\widetilde{\mu} = -N, \qquad \widetilde{\lambda}_k = N\left(1 - \sum_{j \in J} X_j \frac{A_{jk}}{A_{jS_T}^\top \widetilde{T}_{S_T}}\right), \ k \in [K].$$

Clearly, $\mathrm{supp}(\widetilde{T}) \subseteq \mathrm{supp}(T)$ by definition. It remains to verify $(\widetilde{T}, \widetilde{\lambda}, \widetilde{\mu})$ satisfies (F.2) – (F.3) in lieu of $(\widehat{T}, \lambda, \mu)$. By construction, we only need to prove

$$\widetilde{\lambda}_k > 0,^{2} \qquad \forall k \in S_T^c. \tag{F.14}$$

Pick any $k \in S_T^c$. Adding and subtracting terms yields

$$\begin{aligned}
\sum_{j \in J} X_j \frac{A_{jk}}{A_{jS_T}^\top \widetilde{T}_{S_T}} &= \sum_{j \in J} X_j \left(\frac{A_{jk}}{A_{jS_T}^\top \widetilde{T}_{S_T}} - \frac{A_{jk}}{A_{jS_T}^\top T_{S_T}}\right) + \sum_{j \in J} X_j \frac{A_{jk}}{A_{jS_T}^\top T_{S_T}} \\
&= \sum_{j \in J} X_j \frac{A_{jk} A_{jS_T}^\top (T_{S_T} - \widetilde{T}_{S_T})}{A_{jS_T}^\top \widetilde{T}_{S_T} A_{jS_T}^\top T_{S_T}} + \sum_{j \in \overline{J}} X_j \frac{A_{jk}}{A_{jS_T}^\top T_{S_T}} \\
&= \sum_{j \in J} X_j \frac{A_{jk} A_{jS_T}^\top (T_{S_T} - \widetilde{T}_{S_T})}{A_{jS_T}^\top \widetilde{T}_{S_T} A_{jS_T}^\top T_{S_T}} + \sum_{j \in \overline{J}} \left(X_j - A_{jS_T}^\top T_{S_T}\right) \frac{A_{jk}}{A_{jS_T}^\top T_{S_T}} + \sum_{j \in \overline{J}} A_{jk} \\
&\coloneqq R_{1,k} + R_{2,k} + \sum_{j \in \overline{J}} A_{jk},
\end{aligned}$$

[2] We in fact only need a non-strict inequality for proving (i). The strict inequality is used to prove (ii).

where in the second step we used $\Pi_j = A_{jS_T}^\top T_{S_T} > 0$ for $j \in \overline{J}$ and $X_j = 0$ for $j \in \overline{J} \setminus J$. Since $\sum_{j=1}^p A_{jk} = 1$, it suffices to show

$$|R_{1,k}| + |R_{2,k}| \le \sum_{j \in \overline{J}^c} A_{jk}, \qquad \forall k \in S_T^c.$$

To bound $R_{1,k}$, by writing $\Delta = \widetilde{T}_{S_T} - T_{S_T}$ for simplicity, we observe

$$\begin{aligned}
|R_{1,k}| &= \left| \sum_{j \in J} X_j \frac{A_{jk} A_{jS_T}^\top \Delta}{A_{jS_T}^\top \widetilde{T}_{S_T} A_{jS_T}^\top T_{S_T}} \right| \\
&= \left| \sum_{a \in S_T} \Delta_a \sum_{j \in J} X_j \frac{A_{jk} A_{ja}}{A_{jS_T}^\top \widetilde{T}_{S_T} A_{jS_T}^\top T_{S_T}} \right| \\
&\le \|\Delta\|_1 \max_{a \in S_T} \sum_{j \in J} X_j \frac{A_{jk} A_{ja}}{A_{jS_T}^\top \widetilde{T}_{S_T} A_{jS_T}^\top T_{S_T}} \\
&\le \|\Delta\|_1 \max_{j \in \overline{J}} \frac{A_{jk}}{A_{jS_T}^\top T_{S_T}} \max_{a \in S_T} \sum_{j \in J} X_j \frac{A_{ja}}{A_{jS_T}^\top \widetilde{T}_{S_T}}.
\end{aligned}$$

From the KKT conditions (F.12) – (F.13), we deduce that

$$\max_{a \in S_T} \sum_{j \in J} X_j \frac{A_{ja}}{A_{jS_T}^\top \widetilde{T}_{S_T}} \le 1.$$

Also by $A_{jS_T}^\top T_{S_T} = \Pi_j$, we conclude

$$\max_{k \in S_T^c} |R_{1,k}| \le \|\Delta\|_1 \max_{k \in S_T^c} \max_{j \in \overline{J}} \frac{A_{jk}}{\Pi_j} \overset{(13)}{=} \|\Delta\|_1 \rho_{S_T^c}.$$

Regarding $R_{2,k}$, invoking Lemma I.2 with an union bound over $k \in S_T^c$ yields

$$\max_{k \in S_T^c} |R_{2,k}| \le \sqrt{\frac{2\rho_{S_T^c} \log((K-s)/t)}{N}} + \frac{2\rho_{S_T^c} \log((K-s)/t)}{3N}$$

with probability $1 - 2p^{-1} - 2t$. The desired result follows, provided that

$$\min_{k \in S_T^c} \sum_{j \in \overline{J}^c} A_{jk} > \rho_{S_T^c} \|\Delta\|_1 + \sqrt{\frac{2\rho_{S_T^c} \log((K-s)/t)}{N}} + \frac{2\rho_{S_T^c} \log((K-s)/t)}{3N},$$

which is ensured by (24) in Theorem 5 coupled with the rate of $\|\Delta\|_1$ in Corollary 4 and the bound $\rho_{S_T^c} \le \xi / T_{\min}$ from (13) of Remark 2.

Proof of step (ii). Suppose there exists $\bar{T} \neq \widetilde{T}$ such that $\bar{T}$ is also an optimal solution to (4). Then, the fact that both $\bar{T}$ and $\widetilde{T}$ are optimal solutions implies

$$f(\widetilde{T}) = f(\bar{T}), \qquad \text{with} \qquad f(T) = N \sum_{j \in J} X_j \log(A_{j\cdot}^\top T).$$

Let $\nabla f(\widetilde{T})$ denote the gradient of $f(T)$ at $\widetilde{T}$. By adding and subtracting terms, we obtain

$$f(\widetilde{T}) - f(\bar{T}) + \langle \nabla f(\widetilde{T}), \bar{T} - \widetilde{T} \rangle = \langle \nabla f(\widetilde{T}), \bar{T} - \widetilde{T} \rangle.$$

The concavity of $f(T)$ ensures that the left hand side of the above equality is positive. We thus have

$$\langle \nabla f(\widetilde{T}), \widetilde{T} - \bar{T} \rangle \leq 0.$$

Since $\widetilde{T}$ satisfies the KKT condition in (F.2), $\mathbf{1}_K^\top(\widetilde{T} - \bar{T}) = 0$ and $\widetilde{\lambda}^\top \widetilde{T} = 0$ from the restrictions (F.3), we further deduce that

$$\begin{aligned}
0 &= \langle \nabla f(\widetilde{T}) + \widetilde{\lambda} + \widetilde{\mu}\mathbf{1}_K, \widetilde{T} - \bar{T} \rangle \\
&= \langle \nabla f(\widetilde{T}), \widetilde{T} - \bar{T} \rangle + \langle \widetilde{\lambda}, \widetilde{T} - \bar{T} \rangle + \langle \widetilde{\mu}\mathbf{1}_K, \widetilde{T} - \bar{T} \rangle \\
&= \langle \nabla f(\widetilde{T}), \widetilde{T} - \bar{T} \rangle + \langle \widetilde{\lambda}, \widetilde{T} - \bar{T} \rangle \\
&\leq \langle \widetilde{\lambda}, \widetilde{T} - \bar{T} \rangle \\
&= -\widetilde{\lambda}^\top \bar{T} \\
&\leq 0,
\end{aligned}$$

that is, $\widetilde{\lambda}^\top \bar{T} = 0$. We conclude that since (F.14) holds, that is, $\widetilde{\lambda}_{S_T^c} \succ \mathbf{0}$, then we must have $\bar{T}_{S_T^c} = \mathbf{0}$. This shows that $\mathrm{supp}(\bar{T}) \subseteq \mathrm{supp}(T) = S_T$ and completes our proof. □

F.5. Proof of Theorem 7: Minimax lower bounds of estimating T_* in ℓ_1 norm. We start by constructing the hypotheses. Pick any $1 < s \leq K$. We choose

$$T^{(0)} = \frac{1}{s}(\mathbf{1}_s^\top, \mathbf{0}^\top)^\top.$$

For now, suppose s is even. Let $\mathcal{M} = \{0,1\}^{s/2}$. Following Tsybakov (2008, Lemma 2.9), there exists $w^{(j)} \in \mathcal{M}$ for $j = 1, \ldots, |\mathcal{M}|$ such that $w^{(0)} = \mathbf{0}$, $\log(|\mathcal{M}|) \geq s\log(2)/16$ and

$$\|w^{(j)} - w^{(i)}\|_1 \geq \frac{s}{16}, \quad \forall i \neq j.$$

For all $1 \leq j \leq |\mathcal{M}|$, let $\widetilde{w}^{(j)} = ([w^{(j)}]^\top, -[w^{(j)}]^\top, \mathbf{0}^\top)^\top$ and

$$T^{(j)} = T^{(0)} + \gamma\, \widetilde{w}^{(j)}$$

with $\gamma = \sqrt{c_0/(sN)}$ for some constant $c_0 > 0$. It is easy to see that $T^{(j)} \in \mathcal{T}'(s)$ for all $0 \leq j \leq |\mathcal{M}|$, under $s \leq cN$ for sufficiently small $c > 0$. We aim to invoke Tsybakov (2008, Theorem 2.5) by proving the following:

(a) $\mathrm{KL}(\mathbb{P}_{T^{(j)}}, \mathbb{P}_{T^{(0)}}) \leq \log(|\mathcal{M}|)/16$, for all $1 \leq j \leq |\mathcal{M}|$;
(b) $\|T^{(j)} - T^{(i)}\|_1 \geq c'\sqrt{s/n}$ for all $1 \leq i \neq j \leq |\mathcal{M}|$.

Write $\Pi^{(j)} = AT^{(j)}$ for $0 \leq j \leq |\mathcal{M}|$ and $\overline{J}^{(0)} = \{i : \Pi_i^{(0)} > 0\}$ for simplicity. To prove (a), since

$$\max_{i \in \overline{J}^{(0)}} \frac{|\Pi_i^{(j)} - \Pi_i^{(0)}|}{\Pi_i^{(0)}} = \gamma \max_{i \in \overline{J}^{(0)}} \frac{|A_{i\cdot}^\top \widetilde{w}^{(j)}|}{A_{i\cdot}^\top T^{(0)}} \leq s\gamma < 1,$$

and $\Pi_i^{(j)} > 0$ for all $i \in \overline{J}^{(0)}$, invoke Bing, Bunea and Wegkamp (2020b, Lemma 12) with $n = 1$ to obtain

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{T^{(j)}}, \mathbb{P}_{T^{(0)}}) &\leq (1 + c'')\gamma^2\, N \sum_{i \in \overline{J}^{(0)}} \frac{[A_{i\cdot}^\top \widetilde{w}^{(j)}]^2}{A_{i\cdot}^\top T^{(0)}} \\
&\leq (1 + c'')\gamma^2 N \sigma_1(G_0) \|\widetilde{w}^{(j)}\|_2^2 \\
&\leq (1 + c'') c_0\, \sigma_1(G_0) \qquad \text{by } \|\widetilde{w}^{(j)}\|_2^2 \leq s
\end{aligned}$$

where $\sigma_1(G_0)$ denotes the largest eigenvalue of G_0 with

$$G_0 = \sum_{i\in\overline{J}^{(0)}} \frac{1}{A_{i\cdot}^\top T^{(0)}} A_{iS_0} A_{iS_0}^\top.$$

Here we write $S_0 = \mathrm{supp}(T^{(0)})$. The result of part (a) then follows by showing $\sigma_1(G_0) \le s$. To this end, by using the inequality $\sigma_1(M) \le \|M\|_{\infty,1}$ for any symmetric matrix M, we have

$$\begin{aligned}
\sigma_1(G_0) &\le \max_{k\in S_0} \sum_{a\in S_0} \sum_{i\in\overline{J}^{(0)}} \frac{A_{ia}A_{ik}}{A_{i\cdot}^\top T^{(0)}} \\
&= s \max_{k\in S_0} \sum_{i\in\overline{J}^{(0)}} A_{ik} && \text{by } A_{i\cdot}^\top T^{(0)} = \frac{\|A_{iS_0}\|_1}{s} \\
&\le s && \text{by } \sum_{i=1}^p A_{ik} = 1.
\end{aligned}$$

We proceed to prove (b) by noting that

$$\|T^{(j)} - T^{(i)}\|_1 \ge \gamma \frac{s}{16} = \sqrt{\frac{c_0}{16^2}}\sqrt{\frac{s}{n}}, \quad \forall i \ne j.$$

This concludes the proof when s is even. When s is odd and $s \ge 3$, the same arguments hold by defining $\mathcal{M}' = \{0,1\}^{(s-1)/2}$.

APPENDIX G: PROOFS FOR SECTION 2.2: ESTIMATION WITH UNKNOWN A

G.1. Proof of Theorem 8: The general bound of the ℓ_1-norm convergence rate of $\widehat{T}$. We work on the intersection of events defined in (30) and (31), and defined in (F.1). Without loss of generality, we assume (30) and (31) hold for the permutation $P = \boldsymbol{I}_K$. First, we recall that

$$\text{(G.1)} \qquad \frac{1}{2}\Pi_j \le \Pi_j - |(\widehat{A}_{j\cdot} - A_{j\cdot})^\top T| \le \widehat{A}_{j\cdot}^\top T \le \Pi_j + |(\widehat{A}_{j\cdot} - A_{j\cdot})^\top T| \le \frac{3}{2}\Pi_j,$$

for all $j \in \overline{J}$, see (32). The proof resembles the proof of Theorem 1. The KKT conditions are now

$$\text{(G.2)} \qquad N\sum_{j\in J} X_j \frac{\widehat{A}_{j\cdot}}{\widehat{A}_{j\cdot}^\top \widehat{T}} + \lambda + \mu \mathbf{1}_K = 0;$$

$$\text{(G.3)} \qquad \lambda_k \ge 0, \quad \lambda_k \widehat{T}_k = 0, \ \forall k \in [K], \quad \widehat{T}^\top \mathbf{1}_K = 1.$$

Using the same reasoning in the proof of Theorem 1, we arrive at

$$\text{(G.4)}$$

$$\begin{aligned}
N\sum_{j\in J} \frac{X_j}{\widehat{A}_{j\cdot}^\top T} \frac{(\widehat{A}_{j\cdot}^\top \Delta)^2}{\widehat{A}_{j\cdot}^\top \widehat{T}} &= N\sum_{j\in J} X_j \frac{\widehat{A}_{j\cdot}^\top \Delta}{\widehat{A}_{j\cdot}^\top T} + \lambda^\top \Delta \\
&\le 2N \sum_{j\in\overline{J}} X_j \frac{\widehat{A}_{j\cdot}^\top \Delta}{\Pi_j} \\
&= 2N \sum_{j\in\overline{J}} (X_j - \Pi_j) \frac{A_{j\cdot}^\top \Delta}{\Pi_j} + 2N \sum_{j\in\overline{J}} A_{j\cdot}^\top \Delta + 2N \sum_{j\in\overline{J}} X_j \frac{(\widehat{A}_{j\cdot} - A_{j\cdot})^\top \Delta}{\Pi_j}
\end{aligned}$$

with $\Delta := \widehat{T} - T$. For the left hand side of (G.4), use $\underline{J} \subseteq J$ to obtain

$$\sum_{j\in J} \frac{X_j}{\widehat{A}_{j\cdot}^\top T} \frac{(\widehat{A}_{j\cdot}^\top \Delta)^2}{\widehat{A}_{j\cdot}^\top \widehat{T}} \geq \sum_{j\in \underline{J}} \frac{X_j}{\widehat{A}_{j\cdot}^\top T} \frac{(\widehat{A}_{j\cdot}^\top \Delta)^2}{\widehat{A}_{j\cdot}^\top \widehat{T}} \geq \min_{j\in \underline{J}} \frac{X_j}{\widehat{A}_{j\cdot}^\top T} \sum_{j\in \underline{J}} \frac{(\widehat{A}_{j\cdot}^\top \Delta)^2}{\widehat{A}_{j\cdot}^\top \widehat{T}}.$$

We argue as before in the proof of Theorem 1 to obtain

$$\sum_{j\in \underline{J}} \frac{(\widehat{A}_{j\cdot}^\top \Delta)^2}{\widehat{A}_{j\cdot}^\top \widehat{T}} \geq \kappa^2(\widehat{A}_{\underline{J}}, s)$$

and use (33) to prove $\kappa(\widehat{A}_{\underline{J}}, s) \geq \frac{1}{2}\kappa(A_{\underline{J}}, s)$, cf. (34). Since the inequality

$$\frac{X_j}{\widehat{A}_{j\cdot}^\top T} \overset{(G.1)}{\geq} \frac{2X_j}{3\Pi_j} \geq \frac{2}{3} \cdot \frac{\Pi_j - |X_j - \Pi_j|}{\Pi_j} \geq \frac{2}{3}\left(1 - \frac{\varepsilon_j}{\Pi_j}\right) \geq \frac{1}{3}$$

holds for all $j \in \underline{J}$, we conclude

$$\sum_{j\in \underline{J}} \frac{X_j}{\widehat{A}_{j\cdot}^\top T} \frac{(\widehat{A}_{j\cdot}^\top \Delta)^2}{\widehat{A}_{j\cdot}^\top \widehat{T}} \geq \frac{1}{3}\kappa^2(A_{\underline{J}}, s)\|\Delta\|_1^2. \tag{G.5}$$

Next, for the right hand side of (G.4), we have shown in the proof of Theorem 1 that

$$\left|\sum_{j\in \overline{J}} (X_j - \Pi_j)\frac{A_{j\cdot}^\top \Delta}{\Pi_j}\right| \leq \|\Delta\|_1 \left\{\sqrt{\frac{2\rho\log(K/t)}{N}} + \frac{2\rho\log(K/t)}{3N}\right\} \tag{G.6}$$

with probability $1 - 2t$, for any $t \geq 0$. Furthermore, invoking the event $\mathcal{E}$ defined in (F.1) and using the expression of ε_j in (7) gives

$$\begin{aligned}\left|\sum_{j\in \overline{J}} X_j \frac{(\widehat{A}_{j\cdot} - A_{j\cdot})^\top \Delta}{\Pi_j}\right| &\leq \|\Delta\|_1 \max_{1\leq k\leq K} \sum_{j\in \overline{J}} \left(1 + \frac{\varepsilon_j}{\Pi_j}\right)|A_{jk} - \widehat{A}_{jk}| \\ &\leq \|\Delta\|_1 \max_{1\leq k\leq K} \sum_{j\in \overline{J}} |A_{jk} - \widehat{A}_{jk}| \left(1 + 2\sqrt{\frac{\log(p)}{\Pi_j N}} + \frac{4\log(p)}{3\Pi_j N}\right).\end{aligned}$$

Recall that $\underline{J} \subseteq \overline{J}$ and $\Pi_j \geq 8\log(p)/(3N)$ for all $j \in \underline{J}$. We have, for any $k \in [K]$,

$$\begin{aligned}&\sum_{j\in \overline{J}} |\widehat{A}_{jk} - A_{jk}| \left(1 + 2\sqrt{\frac{\log(p)}{\Pi_j N}} + \frac{4\log(p)}{3\Pi_j N}\right) \\ &\leq \left[\sum_{j\in \underline{J}} |\widehat{A}_{jk} - A_{jk}| + \sum_{j\in \overline{J}\setminus \underline{J}} |\widehat{A}_{jk} - A_{jk}|\right]\left(2 + \frac{7\log(p)}{3\Pi_j N}\right) \\ &\leq 3\sum_{j\in \overline{J}} |\widehat{A}_{jk} - A_{jk}| + \sum_{j\in \overline{J}\setminus \underline{J}} |\widehat{A}_{jk} - A_{jk}| \frac{7\log(p)}{3\Pi_j N},\end{aligned}$$

implying

$$\left|\sum_{j\in \overline{J}} X_j \frac{(\widehat{A}_{j\cdot} - A_{j\cdot})^\top \Delta}{\Pi_j}\right| \leq \|\Delta\|_1 \left(3\|\widehat{A}_{\overline{J}} - A_{\overline{J}}\|_{1,\infty} + \sum_{j\in \overline{J}\setminus \underline{J}} \frac{\|\widehat{A}_{j\cdot} - A_{j\cdot}\|_\infty}{\Pi_j} \frac{7\log(p)}{3N}\right). \tag{G.7}$$

Finally, by collecting terms (G.5), (G.6) and (G.7) and using $\sum_{j\in\overline{J}} A_{j\cdot}^\top \Delta \le 0$ from (F.5), we conclude

$$\frac{1}{3}\kappa^2(A_{\underline{J}},s)\|\Delta\|_1 \le \; 2\left(3\|\widehat{A}_{\overline{J}} - A_{\overline{J}}\|_{1,\infty} + \sum_{j\in\overline{J}\setminus\underline{J}} \frac{\|\widehat{A}_{j\cdot} - A_{j\cdot}\|_\infty}{\Pi_j}\frac{7\log(p)}{3N}\right)$$
$$+ 2\sqrt{\frac{2\rho\log(K/t)}{N}} + \frac{4\rho\log(K/t)}{3N}$$

with probability $1-2t$, for any $t \ge 0$. Take $t = p^{-1}$ to complete the proof. □

G.2. Proof of Theorem 9: Fast rates of $\widehat{T}$. We work on the intersection of the events, defined in (30) and (36), and the event $\mathcal{E}$ in (F.1), so we can assume that the conclusion of Theorem 8 holds for $P = \boldsymbol{I}_K$ without loss of generality, that is,

$$\|\widehat{T} - T\|_1 \le \frac{6}{\kappa^2(A_{\underline{J}},s)}\left\{\sqrt{\frac{2\rho\log(p)}{N}} + \frac{2\rho\log(p)}{3N} + 3\left\|\widehat{A}_{\overline{J}} - A_{\overline{J}}\right\|_{1,\infty}\right.$$
$$\left. + \frac{7}{3}\sum_{j\in\overline{J}\setminus\underline{J}} \frac{\|\widehat{A}_{j\cdot} - A_{j\cdot}\|_\infty}{\Pi_j}\frac{\log(p)}{N}\right\}$$
$$\lesssim \frac{1}{\kappa^2(A_{\underline{J}},s)}\left\{\sqrt{\frac{\rho\log(p)}{N}} + \frac{\rho\log(p)}{N} + \left\|\widehat{A}_{\overline{J}} - A_{\overline{J}}\right\|_{1,\infty}\right\}.$$

The last step also uses (30) and $|\overline{J}\setminus\underline{J}| \le C'$ to collect terms. Similar to the arguments of proving Theorem 2, we notice that (30), (35) and (36) guarantee that

$$(1-c)\Pi_j \le A_{j\cdot}^\top \widehat{T} \le (1+c)\Pi_j, \qquad \forall\, j \in J \subseteq \overline{J}. \tag{G.8}$$

From (G.4) and (F.5), we have

$$N\sum_{j\in J}\frac{X_j}{\widehat{A}_{j\cdot}^\top T}\frac{(\widehat{A}_{j\cdot}^\top\Delta)^2}{\widehat{A}_{j\cdot}^\top\widehat{T}} \le 2N\sum_{j\in\overline{J}}(X_j-\Pi_j)\frac{A_{j\cdot}^\top\Delta}{\Pi_j} + 2N\sum_{j\in\overline{J}}X_j\frac{(\widehat{A}_{j\cdot}-A_{j\cdot})^\top\Delta}{\Pi_j}.$$

By the arguments in the proof of Theorem 2, one can deduce that

$$\left|\sum_{j\in\overline{J}}(X_j-\Pi_j)\frac{A_{j\cdot}^\top\Delta}{\Pi_j}\right| \le \left|\sum_{j\in\overline{J}}(X_j-\Pi_j)\frac{A_{j\cdot}^\top H^{-1/2}H^{1/2}\Delta}{\Pi_j}\right|$$
$$\le \|H^{1/2}\Delta\|_2\left\|\sum_{j\in\overline{J}}\frac{X_j-\Pi_j}{\Pi_j}H^{-1/2}A_{j\cdot}\right\|_2$$
$$\lesssim \|H^{1/2}\Delta\|_2\sqrt{\frac{K\log(p)}{N}}$$

with probability $1-2p^{-1}$. With the same probability, by using (G.7) together with (F.10), we conclude

(G.9)

$$\sum_{j\in J}\frac{X_j}{\widehat{A}_{j\cdot}^\top T}\frac{(\widehat{A}_{j\cdot}^\top\Delta)^2}{\widehat{A}_{j\cdot}^\top\widehat{T}}$$

$$\lesssim \|H^{1/2}\Delta\|_2\sqrt{\frac{K\log(p)}{N}} + \frac{\|H^{1/2}\Delta\|_2}{\kappa(A_{\overline{J}},s)}\left(\|\widehat{A}_{\overline{J}} - A_{\overline{J}}\|_{1,\infty} + \sum_{j\in\overline{J}\setminus\underline{J}}\frac{\|\widehat{A}_{j\cdot} - A_{j\cdot}\|_\infty}{\Pi_j}\frac{\log(p)}{N}\right).$$

We proceed to bound from below the left hand side. Since (36) together with (G.8) implies

$$\begin{aligned}(1/2-c)\Pi_j &\le A_{j\cdot}^\top\widehat{T} - \|\widehat{A}_{j\cdot} - A_{j\cdot}\|_\infty\|\widehat{T}\|_1\\ &\le \widehat{A}_{j\cdot}^\top\widehat{T}\\ &\le A_{j\cdot}^\top\widehat{T} + \|\widehat{A}_{j\cdot} - A_{j\cdot}\|_\infty\|\widehat{T}\|_1 \le (3/2+c)\Pi_j \qquad \text{(G.10)}\end{aligned}$$

and

$$\Pi_j/2 \le \widehat{A}_{j\cdot}^\top T \le 3\Pi_j/2, \qquad \text{(G.11)}$$

for all $j \in J \subseteq \overline{J}$. We have

$$\sum_{j\in J}\frac{X_j}{\widehat{A}_{j\cdot}^\top T}\frac{(\widehat{A}_{j\cdot}^\top\Delta)^2}{\widehat{A}_{j\cdot}^\top\widehat{T}} = \sum_{j\in\overline{J}}\frac{X_j}{\widehat{A}_{j\cdot}^\top T}\frac{(\widehat{A}_{j\cdot}^\top\Delta)^2}{\widehat{A}_{j\cdot}^\top\widehat{T}} \gtrsim \Delta^\top\widetilde{H}\Delta$$

where we write

$$\widetilde{H} = \sum_{j\in\overline{J}}\frac{X_j}{\Pi_j^2}\widehat{A}_{j\cdot}\widehat{A}_{j\cdot}^\top.$$

Recall the definition of $\widehat{H}$ from (F.7). It follows that

$$\Delta^\top\widetilde{H}\Delta \ge \Delta^\top\widehat{H}\Delta - \sum_{j\in\overline{J}}X_j\frac{|\widehat{A}_{j\cdot}^\top\Delta|}{\Pi_j}\frac{|(\widehat{A}_{j\cdot}-A_{j\cdot})^\top\Delta|}{\Pi_j} - \sum_{j\in\overline{J}}X_j\frac{|A_{j\cdot}^\top\Delta|}{\Pi_j}\frac{|(\widehat{A}_{j\cdot}-A_{j\cdot})^\top\Delta|}{\Pi_j}.$$

By using

$$\max_{j\in\overline{J}}\frac{\|\widehat{A}_{j\cdot}\|_\infty}{\Pi_j} \le \rho + \max_{j\in\overline{J}}\frac{\|\widehat{A}_{j\cdot}-A_{j\cdot}\|_\infty}{\Pi_j} \le \rho + \frac{1}{2} \le 2\rho,$$

we first have

$$\begin{aligned}&\sum_{j\in\overline{J}}X_j\frac{|\widehat{A}_{j\cdot}^\top\Delta|}{\Pi_j}\frac{|(\widehat{A}_{j\cdot}-A_{j\cdot})^\top\Delta|}{\Pi_j}\\ &\le 2\rho\|\Delta\|_1^2\max_{k\in[K]}\sum_{j\in\overline{J}}\frac{X_j}{\Pi_j}|\widehat{A}_{jk}-A_{jk}|\\ &\le 2\rho\|\Delta\|_1^2\max_{k\in[K]}\sum_{j\in\overline{J}}\left(1+\frac{|X_j-\Pi_j|}{\Pi_j}\right)|\widehat{A}_{jk}-A_{jk}|\\ &\le 4\rho\|\Delta\|_1^2\max_{k\in[K]}\sum_{j\in\overline{J}}\left(1+\frac{7\log(p)}{3\Pi_j N}\right)|\widehat{A}_{jk}-A_{jk}|\\ &\lesssim \rho\|\Delta\|_1^2\left(\|\widehat{A}_{\overline{J}}-A_{\overline{J}}\|_{1,\infty} + \sum_{j\in\overline{J}\setminus\underline{J}}\frac{\|\widehat{A}_{j\cdot}-A_{j\cdot}\|_\infty}{\Pi_j}\frac{\log(p)}{N}\right)\\ &\lesssim \frac{\rho}{\kappa(A_{\overline{J}},s)}\left(\|\widehat{A}_{\overline{J}}-A_{\overline{J}}\|_{1,\infty} + \frac{\log(p)}{N}\right)\|H^{1/2}\Delta\|_2^2\end{aligned}$$

where the last line uses (F.10) and $|\overline{J} \setminus \underline{J}| \le C'$. A similar argument also gives the same upper bound for

$$\sum_{j \in \overline{J}} X_j \frac{|A_{j\cdot}^\top \Delta|}{\Pi_j} \frac{|(\widehat{A}_{j\cdot} - A_{j\cdot})^\top \Delta|}{\Pi_j}.$$

Under condition (35) and (36), and also by invoking Lemma I.4, we readily have

$$\mathbb{P}\left\{ \Delta^\top \widetilde{H} \Delta \gtrsim \|H^{1/2}\Delta\|_2^2 \right\} \ge 1 - 2K^{-1},$$

which together with (G.9) gives

$$\|H^{1/2}\Delta\|_2 \lesssim \sqrt{\frac{K\log(p)}{N}} + \frac{1}{\kappa(A_{\overline{J}}, s)} \left(\|\widehat{A}_{\overline{J}} - A_{\overline{J}}\|_{1,\infty} + \sum_{j \in \overline{J} \setminus \underline{J}} \frac{\|\widehat{A}_{j\cdot} - A_{j\cdot}\|_\infty}{\Pi_j} \frac{\log(p)}{N} \right).$$

Invoke (F.10) and use (35) and $|\overline{J} \setminus \underline{J}| \le C'$ to simplify the expression to complete the proof. □

G.3. Proof of Theorem 10: One-sided sparsity recovery of $\widehat{T}$. The arguments resemble the proof of Theorem 5. The proof of step (ii) follows exactly from the same argument by replacing $A_{j\cdot}$ by $\widehat{A}_{j\cdot}$. We therefore only prove step (i): there exists an optimal solution $\widetilde{T}$ to (29) such that $\mathrm{supp}(\widetilde{T}) \subseteq \mathrm{supp}(T)$. Similarly, we define $\widetilde{T}_{S_T^c} = 0$ and

$$\widetilde{T}_{S_T} = \arg\max_{\beta \in \Delta_s} N \sum_{j \in J} X_j \log\left(\widehat{A}_{jS_T}^\top \beta \right). \tag{G.12}$$

The KKT condition corresponding to (G.12) states

$$N \sum_{j \in J} X_j \frac{\widehat{A}_{jS_T}}{\widehat{A}_{jS_T}^\top \widetilde{T}_{S_T}} + \widetilde{\lambda}_{S_T} + \widetilde{\mu} \mathbf{1}_s = 0; \tag{G.13}$$

$$\widetilde{\lambda}_k \ge 0, \quad \widetilde{\lambda}_k \widetilde{T}_k = 0, \ \forall k \in S_T, \quad \widetilde{T}_{S_T}^\top \mathbf{1}_s = 1. \tag{G.14}$$

By similar reasoning as the proof of Theorem 5, we define

$$\widetilde{\mu} = -N, \quad \widetilde{\lambda}_k = N \left(1 - \sum_{j \in J} X_j \frac{\widehat{A}_{jk}}{\widehat{A}_{jS_T}^\top \widetilde{T}_{S_T}} \right), \qquad \forall k \in [K].$$

The verification that $(\widetilde{T}, \widetilde{\lambda}, \widetilde{\mu})$ satisfies (G.2) – (G.3) in lieu of $(\widehat{T}, \lambda, \mu)$ boils down to show

$$\widetilde{\lambda}_k > 0, \qquad \forall k \in S_T^c. \tag{G.15}$$

We show this on the event defined in Theorem 10. Notice that conditions in Theorem 10 imply that both (G.10) and (G.11) hold.

Pick any $k \in S_T^c$. Adding and subtracting terms yields

$$\begin{aligned}
\sum_{j \in J} X_j \frac{\widehat{A}_{jk}}{\widehat{A}_{jS_T}^\top \widetilde{T}_{S_T}} &= \sum_{j \in J} X_j \frac{\widehat{A}_{jk} - A_{jk}}{\widehat{A}_{jS_T}^\top \widetilde{T}_{S_T}} + \sum_{j \in J} X_j \frac{A_{jk} \widehat{A}_{jS_T}^\top (T_{S_T} - \widetilde{T}_{S_T})}{\widehat{A}_{jS_T}^\top \widetilde{T}_{S_T} \widehat{A}_{jS_T}^\top T_{S_T}} \\
&\quad + \sum_{j \in J} X_j \frac{A_{jk} (A_{jS_T} - \widehat{A}_{jS_T})^\top T_{S_T}}{\widehat{A}_{jS_T}^\top T_{S_T} A_{jS_T}^\top T_{S_T}} + \sum_{j \in J} (X_j - \Pi_j) \frac{A_{jk}}{\Pi_j} + \sum_{j \in J} A_{jk} \\
&= R_{1,k} + R_{2,k} + R_{3,k} + R_{4,k} + \sum_{j \in J} A_{jk}.
\end{aligned}$$

We bound each term separately. For $R_{1,k}$, using (G.10), $J \subseteq \overline{J}$ and the proof of Theorem 9 gives, on the event $\mathcal{E}$ in (F.1),

$$\begin{aligned}
|R_{1,k}| &\lesssim \sum_{j\in J} \frac{X_j}{\Pi_j} |\widehat{A}_{jk} - A_{jk}| \\
&\lesssim \|\widehat{A}_{\overline{J}} - A_{\overline{J}}\|_{1,\infty} + \sum_{j\in \overline{J}\setminus \underline{J}} \frac{\|\widehat{A}_{j\cdot} - A_{j\cdot}\|_\infty}{\Pi_j} \frac{\log(p)}{N} \\
&\lesssim \|\widehat{A}_{\overline{J}} - A_{\overline{J}}\|_{1,\infty} + \frac{\log(p)}{N} \qquad \text{by } |\overline{J} \setminus \underline{J}| \le C'. \qquad \text{(G.16)}
\end{aligned}$$

To bound $R_{2,k}$, by writing $\Delta = \widetilde{T}_{S_T} - T_{S_T}$ for simplicity and using similar arguments in the proof of Theorem 5, we have

$$|R_{2,k}| = \left| \sum_{j\in J} X_j \frac{A_{jk} \widehat{A}_{jS_T}^\top \Delta}{\widehat{A}_{jS_T}^\top \widetilde{T}_{S_T} \widehat{A}_{jS_T}^\top T_{S_T}} \right| \le \|\Delta\|_1 \max_{j\in\overline{J}} \frac{A_{jk}}{\widehat{A}_{jS_T}^\top T_{S_T}} \max_{a\in S_T} \sum_{j\in J} X_j \frac{\widehat{A}_{ja}}{\widehat{A}_{jS_T}^\top \widetilde{T}_{S_T}}.$$

From (G.13) – (G.14), we deduce that

$$\max_{a\in S_T} \sum_{j\in J} X_j \frac{\widehat{A}_{ja}}{\widehat{A}_{jS_T}^\top \widetilde{T}_{S_T}} \le 1.$$

Also, by using (G.11) together with

$$\|\Delta\|_1 \lesssim \sqrt{\frac{s\log(p)}{N}} + \|\widehat{A}_{\overline{J}} - A_{\overline{J}}\|_{1,\infty}$$

deduced from Theorem 9 with $K = s$ and $\kappa^{-1}(A_{\overline{J}}, s) \le C''$, we conclude that

$$|R_{2,k}| \lesssim \rho_{S_T^c} \left(\sqrt{\frac{s\log(p)}{N}} + \|\widehat{A}_{\overline{J}} - A_{\overline{J}}\|_{1,\infty} \right). \qquad \text{(G.17)}$$

holds with probability at least $1 - 8p^{-1}$. For $R_{3,k}$, by the arguments of bounding $R_{1,k}$ and $R_{2,k}$, it is easy to see that

$$|R_{3,k}| \le \rho_{S_T^c} \max_{a\in S_T} \sum_{j\in J} \frac{X_j}{\Pi_j} |\widehat{A}_{ja} - A_{ja}| \lesssim \rho_{S_T^c} \left(\|\widehat{A}_{\overline{J}} - A_{\overline{J}}\|_{1,\infty} + \frac{\log(p)}{N} \right). \qquad \text{(G.18)}$$

Regarding $R_{4,k}$, invoking Lemma I.2 with $t = 1/p$ and taking a union bounds over $k \in S_T^c$ yields

$$\max_{k\in S_T^c} |R_{4,k}| \le \sqrt{\frac{2\rho_{S_T^c}\log(p)}{N}} + \frac{2\rho_{S_T^c}\log(p)}{3N} \qquad \text{(G.19)}$$

with probability $1 - 2p^{-1}$. Finally, since

$$1 - \sum_{j\in J} A_{jk} \ge \sum_{j\in\overline{J}^c} A_{jk},$$

by collecting terms in (G.16), (G.17), (G.18) and (G.19), the desired result follows provided that

$$\min_{k\in S_T^c} \sum_{j\in\overline{J}^c} A_{jk} \gtrsim \rho_{S_T^c} \sqrt{\frac{s\log(p)}{N}} + \sqrt{\frac{\rho_{S_T^c}\log(p)}{N}} + (1 + \rho_{S_T^c}) \|\widehat{A}_{\overline{J}} - A_{\overline{J}}\|_{1,\infty} + \frac{\log(p)}{N}$$

which is ensured by the condition in Theorem 10 coupled with the fact $\rho_{S_T^c} \le \xi / T_{\min}$. The proof is then complete. □

G.4. Proof of Corollary 11. The result follows from Theorem 9 and Theorem 10 after we verify its conditions (30), (35), (36) and (37). Condition (35) simplifies to (40). The rate on $\|\widehat{A}_{j\cdot} - (AP)_{j\cdot}\|_\infty$ in (39), condition (41) and the inequality $\max_{j\in\overline{J}} \|A_{j\cdot}\|_\infty/\Pi_{*j} \leq \rho \leq (1 \vee \xi)/T_{\min} \lesssim 1/T_{\min}$ imply (30). The rate on $\|\widehat{A} - AP\|_{1,\infty}$ in (38) and the bounds $\kappa^{-1}(A_{\overline{J}}, K) = \mathcal{O}(1)$ and $\xi = \mathcal{O}(1)$ together with conditions (41) and (42) imply (36) and (37). □

APPENDIX H: PROOF OF PROPOSITION 12 IN SECTION 3.2

We prove (54) and (55) – (56) separately in this section. We collect technical lemmas that are used in the proofs at the end of this section.

Proof of (54). Using the triangle inequality for W_1 (Lemma H.2 below),

$$W_1(\widetilde{\Pi}^{(i)}, \widetilde{\Pi}^{(j)}; D^{\text{word}}) \leq W_1(\widetilde{\Pi}^{(i)}, \Pi_*^{(i)}; D^{\text{word}}) + W_1(\Pi_*^{(i)}, \Pi_*^{(j)}; D^{\text{word}}) + W_1(\Pi_*^{(j)}, \widetilde{\Pi}^{(j)}; D^{\text{word}}),$$

and thus

$$W_1(\widetilde{\Pi}^{(i)}, \widetilde{\Pi}^{(j)}; D^{\text{word}}) - W_1(\Pi_*^{(i)}, \Pi_*^{(j)}; D^{\text{word}}) \leq \sum_{k\in\{i,j\}} W_1(\widetilde{\Pi}^{(k)}, \Pi_*^{(k)}; D^{\text{word}}).$$

Combining this with a second application of the triangle inequality with the roles of $\widetilde{\Pi}^{(k)}$ and $\Pi^{(k)}$ switched for $k \in \{i, j\}$, we find

$$\begin{aligned} \left| W_1(\widetilde{\Pi}^{(i)}, \widetilde{\Pi}^{(j)}; D^{\text{word}}) - W_1(\Pi_*^{(i)}, \Pi_*^{(j)}; D_w) \right| &\leq \sum_{k\in\{i,j\}} W_1(\widetilde{\Pi}^{(k)}, \Pi_*^{(k)}; D^{\text{word}}) \\ &\leq \|\boldsymbol{D}^{\text{word}}\|_\infty \frac{1}{2} \sum_{k\in\{i,j\}} \|\widetilde{\Pi}^{(k)} - \Pi_*^{(k)}\|_1, \end{aligned} \tag{H.1}$$

where the second step follows from Lemma H.2 below. For $k \in \{i, j\}$, we find

$$\begin{aligned} \|\widetilde{\Pi}^{(k)} - \Pi_*^{(k)}\|_1 &= \|\widehat{A}\widehat{T}^{(k)} - AT_*^{(k)}\|_1 \\ &= \|\widehat{A}\widehat{T}^{(k)} - A\widehat{T}^{(k)} + A\widehat{T}^{(k)} - AT_*^{(k)}\|_1 \\ &\leq \|(\widehat{A} - A)\widehat{T}^{(k)}\|_1 + \|A(\widehat{T}^{(k)} - T_*^{(k)})\|_1 \\ &\leq \max_{l\in[K]} \|\widehat{A}_{\cdot l} - A_{\cdot l}\|_1 \|\widehat{T}^{(k)}\|_1 + \max_{l\in[K]} \|A_{\cdot l}\|_1 \|\widehat{T}^{(k)} - T_*^{(k)}\|_1 \qquad \text{(H.2)} \\ &= \max_{l\in[K]} \|\widehat{A}_{\cdot l} - A_{\cdot l}\|_1 + \|\widehat{T}^{(k)} - T_*^{(k)}\|_1, \end{aligned}$$

where (H.2) follows from the fact that for any $v \in \mathbb{R}^K$,

$$\|Av\|_1 = \sum_{i=1}^p \left| \sum_{k=1}^K A_{ik} v_k \right| \leq \sum_{k=1}^K |v_k| \sum_{i=1}^p |A_{ik}| \leq \max_{l\in[K]} \|A_{\cdot l}\| \|v\|_1,$$

and in the final step we use that $A_{\cdot l} \in \Delta_p$ for all $l \in [K]$ and $\widehat{T}^{(k)} \in \Delta_K$. Plugging this into (H.1) we find

$$\begin{aligned} &\left| W_1(\widetilde{\Pi}^{(i)}, \widetilde{\Pi}^{(j)}; D^{\text{word}}) - W_1(\Pi_*^{(i)}, \Pi_*^{(j)}; D_w) \right| \\ &\leq \|\boldsymbol{D}^{\text{word}}\|_\infty \left\{ \max_{l\in[K]} \|(\widehat{A} - A)\boldsymbol{e}_l\|_1 + \frac{1}{2} \sum_{k\in\{i,j\}} \|\widehat{T}^{(k)} - T_*^{(k)}\|_1 \right\}. \end{aligned} \tag{H.3}$$

Finally, note that for any $P \in \mathcal{H}_K$, $PP^\top = I_K$, so for $k \in \{i, j\}$,

$$\Pi_*^{(k)} = AT_*^{(k)} = APP^\top T_*^{(k)}.$$

Furthermore, $AP \in \Delta_p$ and $P^\top T_*^{(k)} \in \Delta_K$. Thus, (H.3) holds when A and $T_*^{(k)}$ are replaced by AP and $P^\top T_*^{(k)}$, respectively, for any $P \in \mathcal{H}_K$. We can thus take the maximum over $P \in \mathcal{H}_K$, which completes the proof of (54). □

Proof of (55) and (56). We will prove the bound

$$\begin{aligned}&\left|W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}; \widehat{D}^{\text{topic}}) - W_1(T_*^{(i)}, T_*^{(j)}; D^{\text{topic}})\right| \\ &\le 2 \max_{k\in[K]} d(\widehat{A}_{\cdot k}, A_{\cdot k}) + \|\boldsymbol{D}^{\text{topic}}\|_\infty \frac{1}{2} \sum_{k\in\{i,j\}} \|\widehat{T}^{(k)} - T_*^{(k)}\|_1, \end{aligned} \tag{H.4}$$

where d is any metric on Δ_p, and

$$\widehat{D}^{\text{topic}}(k,l) := d(\widehat{A}_{\cdot k}, \widehat{A}_{\cdot l}), \quad D^{\text{topic}}(k,l) := d(A_{\cdot k}, A_{\cdot l}) \quad \forall k, l \in [K]. \tag{H.5}$$

Combining this with Lemma H.1 below, (H.4) yields

$$\begin{aligned}&\left|W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}; \widehat{D}^{\text{topic}}) - W_1(T_*^{(i)}, T_*^{(j)}; D^{\text{topic}})\right| \\ &\le \max_{P\in\mathcal{H}_K} \left\{ 2 \max_{k\in[K]} d(\widehat{A}_{\cdot k}, (AP)_{\cdot k}) + \|\boldsymbol{D}^{\text{topic}}\|_\infty \frac{1}{2} \sum_{k\in\{i,j\}} \|\widehat{T}^{(k)} - P^\top T_*^{(k)}\|_1 \right\}. \end{aligned} \tag{H.6}$$

Equation (56) follows immediately from (H.6) using $d(a,b) = \frac{1}{2}\|a-b\|_1$ for $a, b \in \Delta_p$, and noting that for this choice of d, $\|\boldsymbol{D}^{\text{topic}}\|_\infty \le 1$. To prove (55), choose $d(a,b) = W_1(a, b; D^{\text{word}})$ for $a, b \in \Delta_p$, and note that by Lemma H.2,

$$W_1(\widehat{A}_{\cdot k}, (AP)_{\cdot k}; D^{\text{word}}) \le \|D^{\text{word}}\|_\infty \frac{1}{2} \|\widehat{A}_{\cdot k} - (AP)_{\cdot k}\| \quad \forall k \in [K], \tag{H.7}$$

and for this choice of d,

$$\|\boldsymbol{D}^{\text{topic}}\|_\infty = \max_{k,l\in[K]} W_1(A_{\cdot k}, A_{\cdot l}; D^{\text{word}}) \le \|D^{\text{word}}\| \max_{k,l\in[K]} \|A_{\cdot k} - A_{\cdot l}\| \le \|D^{\text{word}}\|. \tag{H.8}$$

Combining (H.7) and (H.8) with (H.6) proves (55).

Proof of (H.4). We first find

$$\begin{aligned} W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}; \widehat{D}^{\text{topic}}) &= \inf_{w\in\Gamma(\widehat{T}^{(i)}, \widehat{T}^{(j)})} \text{tr}(w\widehat{\boldsymbol{D}}^{\text{topic}}) \\ &= \inf_{w\in\Gamma(\widehat{T}^{(i)}, \widehat{T}^{(j)})} \left\{ \text{tr}(w\boldsymbol{D}^{\text{topic}}) + \text{tr}(w[\widehat{\boldsymbol{D}}^{\text{topic}} - \boldsymbol{D}^{\text{topic}}]) \right\} \\ &= \inf_{w\in\Gamma(\widehat{T}^{(i)}, \widehat{T}^{(j)})} \text{tr}(w\boldsymbol{D}^{\text{topic}}) + \|\widehat{\boldsymbol{D}}^{\text{topic}} - \boldsymbol{D}^{\text{topic}}\|_\infty \qquad \text{(H.9)} \\ &= \|\widehat{\boldsymbol{D}}^{\text{topic}} - \boldsymbol{D}^{\text{topic}}\|_\infty + W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}; D^{\text{topic}}), \qquad \text{(H.10)} \end{aligned}$$

where in (H.9) we use that for any $w \in \Gamma(\widehat{T}^{(i)}, \widehat{T}^{(j)})$,

$$\text{tr}(w[\widehat{\boldsymbol{D}}^{\text{topic}} - \boldsymbol{D}^{\text{topic}}]) = \sum_{t,l=1}^K w_{tl}(\widehat{\boldsymbol{D}}_{lt}^{\text{topic}} - \boldsymbol{D}_{lt}^{\text{topic}})$$

$$\leq \|\widehat{\boldsymbol{D}}^{\text{topic}} - \boldsymbol{D}^{\text{topic}}\|_\infty \cdot \sum_{t,l=1}^{K} w_{tl} \qquad \text{since } w_{lt} \geq 0 \text{ for } t,l \in [K]$$

$$= \|\widehat{\boldsymbol{D}}^{\text{topic}} - \boldsymbol{D}^{\text{topic}}\|_\infty. \tag{H.11}$$

Using the triangle inequality for W_1 (Lemma H.2), we find

$$W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}; D^{\text{topic}}) \leq W_1(\widehat{T}^{(i)}, T_*^{(i)}; D^{\text{topic}}) + W_1(T_*^{(i)}, T_*^{(j)}; D^{\text{topic}}) + W_1(\widehat{T}^{(j)}, T_*^{(j)}; D^{\text{topic}}).$$

Plugging this into (H.10) we find

$$W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}; \widehat{D}^{\text{topic}}) - W_1(T_*^{(i)}, T_*^{(j)}; D^{\text{topic}})$$

$$\leq \|\widehat{D}^{\text{topic}} - D^{\text{topic}}\|_\infty + \sum_{k\in\{i,j\}} W_1(\widehat{T}^{(k)}, T_*^{(k)}; D^{\text{topic}}). \tag{H.12}$$

Using the triangle inequality again,

$$W_1(T_*^{(i)}, T_*^{(j)}; D^{\text{topic}}) \leq W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}; D^{\text{topic}}) + \sum_{k\in\{i,j\}} W_1(\widehat{T}^{(k)}, T_*^{(k)}; D^{\text{topic}})$$

$$\leq \|\widehat{\boldsymbol{D}}^{\text{topic}} - \boldsymbol{D}^{\text{topic}}\|_\infty + W_1(\widehat{T}_i, \widehat{T}_j; \widehat{D}^{\text{topic}}) + \sum_{k\in\{i,j\}} W_1(\widehat{T}_k, T_k; D^{\text{topic}}),$$

where in the second line we used the same argument as in (H.10) with the roles of $\widehat{\boldsymbol{D}}^{\text{topic}}$ and $\boldsymbol{D}^{\text{topic}}$ reversed. Combining this with (H.12), we find

$$\left|W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}; \widehat{D}^{\text{topic}}) - W_1(T_*^{(i)}, T_*^{(j)}; D^{\text{topic}})\right|$$

$$\leq \|\widehat{\boldsymbol{D}}^{\text{topic}} - \boldsymbol{D}^{\text{topic}}\|_\infty + \sum_{k\in\{i,j\}} W_1(\widehat{T}_k, T_k; D^{\text{topic}})$$

$$\leq \|\widehat{\boldsymbol{D}}^{\text{topic}} - \boldsymbol{D}^{\text{topic}}\|_\infty + \|\boldsymbol{D}^{\text{topic}}\|_\infty \frac{1}{2} \sum_{k\in\{i,j\}} \|\widehat{T}_*^{(k)} - T_*^{(k)}\|_1, \tag{H.13}$$

where we use Lemma H.2 in the last line.

Next, we find by the triangle inequality and (H.5) that

$$\widehat{\boldsymbol{D}}_{tl}^{\text{topic}} \leq d(\widehat{A}_{\cdot t}, A_{\cdot t}) + d(A_{\cdot t}, A_{\cdot l}) + d(A_{\cdot l}, \widehat{A}_{\cdot l}),$$

and

$$\boldsymbol{D}_{tl}^{\text{topic}} \leq d(A_{\cdot t}, \widehat{A}_{\cdot t}) + d(\widehat{A}_{\cdot t}, \widehat{A}_{\cdot l}) + d(\widehat{A}_{\cdot l}, A_{\cdot l}),$$

which together give

$$\|\widehat{\boldsymbol{D}}^{\text{topic}} - \boldsymbol{D}^{\text{topic}}\|_\infty \leq 2 \max_{t\in[K]} d(A_{\cdot t}, \widehat{A}_{\cdot t}).$$

Plugging this into (H.13) completes the proof of (H.4).

□

We use the following simple lemma in the proof of (55) and (56).

LEMMA H.1. *For any metric $d : \Delta_p \times \Delta_p \to \mathbb{R}$, $T, T' \in \Delta_K$, and $A \in \mathbb{R}^{p\times K}$ with columns in Δ_p, and any $P \in \mathcal{H}_K$,*

$$\inf_{w\in\Gamma(T,T')} \sum_{k,l=1}^{K} w_{kl} d(A_{\cdot k}, A_{\cdot l}) = \inf_{w\in\Gamma(P^\top T, P^\top T')} \sum_{k,l=1}^{K} w_{kl} d((AP)_{\cdot k}, (AP)_{\cdot l}).$$

PROOF. Fix $P \in \mathcal{H}_K$ and let $\pi : [K] \to [K]$ be the associated bijection, so $P_{kl} = 1[k = \pi(l)]$ for all $k, l \in [K]$. From this it follows that

$$(P^\top T)_k = T_{\pi(k)}, \quad (P^\top T')_k = T'_{\pi(k)}, \quad (AP)_{\cdot k} = A_{\cdot \pi(k)}. \tag{H.14}$$

Let $w \in \Gamma(T, T')$, and define w^π by $w^\pi_{kl} = w_{\pi(k)\pi(l)}$. Then for $l \in [K]$,

$$\begin{aligned}
\sum_{k=1}^K w^\pi_{kl} &= \sum_{k=1}^K w_{\pi(k)\pi(l)} \\
&= T'_{\pi(l)} && \text{since } w \in \Gamma(T, T') \\
&= (P^\top T')_l. && \text{by } (H.14)
\end{aligned}$$

A similar calculation shows $\sum_l w^\pi_{kl} = (P^\top T)_k$, and we thus conclude that $w^\pi \in \Gamma(P^\top T, P^\top T')$.

Next note that

$$\begin{aligned}
\sum_{k,l=1}^K w_{kl} d(A_{\cdot k}, A_{\cdot l}) &= \sum_{k,l=1}^K w_{\pi(k)\pi(l)} d(A_{\cdot \pi(k)}, A_{\cdot \pi(l)}) && \text{since } \pi \text{ is a bijection} \\
&= \sum_{k,l=1}^K w^\pi_{kl} d((AP)_{\cdot k}, (AP)_{\cdot l}) && \text{by } (H.14) \\
&\geq \inf_{w \in \Gamma(P^\top T, P^\top T')} \sum_{k,l=1}^K w_{kl} d((AP)_{\cdot k}, (AP)_{\cdot l}). && \text{since } w^\pi \in \Gamma(P^\top T, P^\top T')
\end{aligned}$$

Since this holds for all $w \in \Gamma(T, T')$, we find

$$\inf_{w \in \Gamma(T, T')} \sum_{k,l=1}^K w_{kl} d(A_{\cdot k}, A_{\cdot l}) \geq \inf_{w \in \Gamma(P^\top T, P^\top T')} \sum_{k,l=1}^K w_{kl} d((AP)_{\cdot k}, (AP)_{\cdot l}). \tag{H.15}$$

Applying (H.15) with P, T, T' and A replaced by $P^\top$, $P^\top T$, $P^\top T'$, and AP, respectively, gives the opposite inequality. Combined with (H.15), this completes the proof. □

We also use the following standard results on the 1-Wasserstein distance in the proofs in this section (see, for example, (Gibbs and Su, 2002; Villani, 2003)).

LEMMA H.2. *Let D be a metric on a finite, non-empty, set $\mathcal{X}$. Then,*

1. *$W_1(\cdot, \cdot; D)$ is a metric on $\Delta_{|\mathcal{X}|}$.*
2. *For any $a, b \in \Delta_{|\mathcal{X}|}$,*

$$W_1(a, b; D) \leq \max_{x,y \in \mathcal{X}} D(x, y) \cdot \frac{1}{2} \|a - b\|_1.$$

APPENDIX I: TECHNICAL LEMMAS

LEMMA I.1. *For any $t \geq 0$, with probability $1 - 2pe^{-t/2}$,*

$$|X_j - \Pi_j| \leq \sqrt{\frac{\Pi_j t}{N}} + \frac{2t}{3N}, \qquad \textit{uniformly over } 1 \leq j \leq p.$$

PROOF. The proof follows by a simple application of the Bernstein's inequality for bounded random variables (see, for instance, the proof of Lemma 15 in Bing, Bunea and Wegkamp (2020a)). □

LEMMA I.2. *Pick any $k \in [K]$. For any $t \geq 0$, with probability $1 - 2e^{-t/2}$,*

$$\left| \sum_{j \in \overline{J}} \frac{X_j - \Pi_j}{\Pi_j} A_{jk} \right| \leq \sqrt{\frac{\rho_k t}{N}} + \frac{2\rho_k t}{3N},$$

with $\rho_k = \max_{j \in \overline{J}} A_{jk}/\Pi_j$.

PROOF. For any $j \in \overline{J}$, notice that

$$X_j - \Pi_j = \frac{1}{N} \sum_{i=1}^{N} (B_{ij} - \Pi_j)$$

where $B_{ij} \sim \text{Bernoulli}(\Pi_j)$ and $(B_{i1}, \ldots, B_{ip})^\top \sim \text{Multinomial}(1, \Pi_j)$ for $i \in [N]$. Let

$$\sum_{j \in \overline{J}} \frac{X_j - \Pi_j}{\Pi_j} A_{jk} = \frac{1}{N} \sum_{i=1}^{N} Z_i$$

with

$$Z_i = \sum_{j \in \overline{J}} (B_{ij} - \Pi_j) \frac{A_{jk}}{\Pi_j}$$

such that $\mathbb{E}[Z_i] = 0$, $|Z_i| \leq 2 \max_{j \in \overline{J}} A_{jk}/\Pi_j = 2\rho_k$ and

$$\mathbb{E}[Z_i^2] = \text{Var}\left(\sum_{j \in \overline{J}} \frac{A_{jk}}{\Pi_j} B_{ij} \right) \leq \sum_{j \in \overline{J}} \frac{A_{jk}^2}{\Pi_j} \leq \rho_k \sum_{j \in \overline{J}} A_{jk} \leq \rho_k.$$

Then an application of Bernstein's inequality gives

$$\mathbb{P}\left\{ \frac{1}{N} \left| \sum_{i=1}^{n} Z_i \right| \geq \sqrt{\frac{\rho_k t}{N}} + \frac{2\rho_k t}{3N} \right\} \leq 2e^{-t/2}, \quad \forall t \geq 0,$$

which completes the proof. □

Recall that

$$H = \sum_{j \in \overline{J}} \frac{A_{j\cdot} A_{j\cdot}^\top}{\Pi_j}.$$

LEMMA I.3. *For any $t \geq 0$, with probability $1 - 2e^{-t/2 + K \log 5}$,*

$$\left\| \sum_{j \in \overline{J}} \frac{X_j - \Pi_j}{\Pi_j} H^{-1/2} A_{j\cdot} \right\|_2 \leq 2\sqrt{\frac{t}{N}} + \frac{2(1 \vee \xi)}{3\kappa(A_{\overline{J}}, K) T_{\min}} \cdot \frac{t}{N}.$$

PROOF. First note that

$$\left\| \sum_{j \in \overline{J}} \frac{X_j - \Pi_j}{\Pi_j} H^{-1/2} A_{j\cdot} \right\|_2 = \sup_{v : \|v\|_2 = 1} \left| \sum_{j \in \overline{J}} \frac{X_j - \Pi_j}{\Pi_j} A_{j\cdot}^\top H^{-1/2} v \right|.$$

Let $\mathcal{N}$ be a minimal $(1/2)$-net of $\{v : \|v\|_2 = 1\}$. By definition and the property of $(1/2)$-net, we have

$$\left\| \sum_{j \in \overline{J}} \frac{X_j - \Pi_j}{\Pi_j} H^{-1/2} A_{j\cdot} \right\|_2 \leq 2 \sup_{v \in \mathcal{N}} \left| \sum_{j \in \overline{J}} \frac{X_j - \Pi_j}{\Pi_j} A_{j\cdot}^\top H^{-1/2} v \right|.$$

Pick any $v \in \mathcal{N}$. By similar reasoning as in the proof of Lemma I.2, we have

$$\sum_{j \in \overline{J}} \frac{X_j - \Pi_j}{\Pi_j} A_{j\cdot}^\top H^{-1/2} v = \frac{1}{N} \sum_{i=1}^N Z_i$$

with

$$Z_i = \sum_{j \in \overline{J}} \frac{B_{ij} - \Pi_j}{\Pi_j} A_{j\cdot}^\top H^{-1/2} v.$$

Note that $\mathbb{E}[Z_i] = 0$ and

$$\begin{aligned}
|Z_i| &\leq \max \left\{ \max_{j \in \overline{J}} \frac{|A_{j\cdot}^\top H^{-1/2} v|}{\Pi_j}, \ \left| \sum_{j \in \overline{J}} A_{j\cdot}^\top H^{-1/2} v \right| \right\} \\
&\leq \max \left\{ \max_{j \in \overline{J}} \frac{|A_{j\cdot}^\top H^{-1/2} v|}{\Pi_j}, \ \max_{j \in \overline{J}} \left| \frac{A_{j\cdot}^\top H^{-1/2} v}{\Pi_j} \right| \sum_{j \in \overline{J}} \Pi_j \right\} \\
&\leq \max_{j \in \overline{J}} \frac{\|A_{j\cdot}\|_\infty \|H^{-1/2} v\|_1}{\Pi_j} \\
&\leq \rho \, \|H^{-1/2} v\|_1.
\end{aligned}$$

Since, for any $u \in \mathbb{R}^K$, one has

$$\begin{aligned}
u^\top H u &= \sum_{j \in \overline{J}} \frac{(A_{j\cdot}^\top u)^2}{\Pi_j} \\
&= \sum_{j \in \overline{J}} \frac{(A_{j\cdot}^\top u)^2}{\Pi_j} \sum_{j \in \overline{J}} \Pi_j && \text{by } \sum_{j \in \overline{J}} \Pi_j = 1 \\
&\geq \left(\sum_{j \in \overline{J}} |A_{j\cdot}^\top u| \right)^2 \\
&\geq \kappa^2(A_{\overline{J}}, K) \|u\|_1^2, && \text{by (9)}
\end{aligned}$$

from which, we deduce that

$$\|H^{-1/2} u\|_1 \leq \kappa^{-1}(A_{\overline{J}}, K) \|u\|_2, \qquad \forall u \in \mathbb{R}^K. \tag{I.1}$$

We thus conclude $|Z_i| \leq \rho \kappa^{-1}(A_{\overline{J}}, K)$. Furthermore, observe that

$$\mathbb{E}[Z_i^2] = \mathrm{Var} \left(\sum_{j \in \overline{J}} \frac{B_{ij}}{\Pi_j} A_{j\cdot}^\top H^{-1/2} v \right) \leq v^\top H^{-1/2} \sum_{j \in \overline{J}} \frac{A_{j\cdot} A_{j\cdot}^\top}{\Pi_j} H^{-1/2} v = 1.$$

An application of the Bernstein's inequality gives

$$\mathbb{P}\left\{\frac{1}{N}\left|\sum_{i=1}^{n} Z_i\right| \geq \sqrt{\frac{t}{N}} + \frac{\rho t}{3\kappa(A_{\overline{J}}, K)N}\right\} \leq 2e^{-t/2}, \quad \forall t \geq 0.$$

The proof follows immediately by using $\rho \leq (1 \vee \xi)/T_{\min}$ and taking a union bound over $v \in \mathcal{N}$ together with $|\mathcal{N}| \leq 5^K$. □

Recall that

$$\widehat{H} = \sum_{j \in \overline{J}} \frac{X_j}{\Pi_j^2} A_{j\cdot} A_{j\cdot}^\top.$$

The following lemma provides a concentration inequality of $H^{-1/2}(\widehat{H} - H)H^{-1/2}$ via an application of the Matrix Bernstein inequality (Tropp, 2015).

LEMMA I.4. *For any $t \geq 0$, one has*

$$\mathbb{P}\left\{\left\|H^{-1/2}(\widehat{H} - H)H^{-1/2}\right\|_{\text{op}} \leq \sqrt{\frac{2Bt}{N}} + \frac{Bt}{3N}\right\} \geq 1 - 2Ke^{-t/2},$$

with

$$B = \frac{1 \vee \xi}{\kappa^2(A_{\overline{J}}, K)T_{\min}^2}\left(1 + \xi\sqrt{K - s}\right).$$

Moreover, if

$$N \geq CB \log K$$

for some sufficiently large constant $C > 0$, then, with probability $1 - 2K^{-1}$, one has

$$\lambda_{\min}(\widehat{H}) \geq c\lambda_{\min}(H)$$

for some constant $c > 0$.

PROOF. By similar arguments in the proof of Lemma I.2, we have

$$H^{-1/2}(\widehat{H} - H)H^{-1/2} = \frac{1}{N}\sum_{i=1}^{N} Z_i$$

with

$$Z_i = \sum_{j \in \overline{J}} \frac{B_{ij} - \Pi_j}{\Pi_j^2} H^{-1/2} A_{j\cdot} A_{j\cdot}^\top H^{-1/2}.$$

Notice $\mathbb{E}[Z_i] = 0$. To apply the Matrix Bernstein inequality, we first find the bound for $\|Z_i\|_{\text{op}}$ as

$$\begin{aligned}\|Z_i\|_{\text{op}} &\leq \max\left\{\max_{j \in \overline{J}}\left\|\frac{H^{-1/2}A_{j\cdot}A_{j\cdot}^\top H^{-1/2}}{\Pi_j^2}\right\|_{\text{op}}, \left\|\sum_{j \in \overline{J}}\frac{H^{-1/2}A_{j\cdot}A_{j\cdot}^\top H^{-1/2}}{\Pi_j}\right\|_{\text{op}}\right\} \\ &\leq \max\left\{\max_{j \in \overline{J}}\frac{A_{j\cdot}^\top H^{-1} A_{j\cdot}}{\Pi_j^2},\ 1\right\}.\end{aligned}$$

Since

$$A_{j\cdot}^\top H^{-1} A_{j\cdot} \le \|A_{j\cdot}\|_\infty \|H^{-1/2}\|_{1,\infty} \|H^{-1/2} A_{j\cdot}\|_1,$$

by using (I.1), we obtain

$$A_{j\cdot}^\top H^{-1} A_{j\cdot} \le \|A_{j\cdot}\|_\infty \|A_{j\cdot}\|_2 \kappa^{-2}(A_{\overline{J}}, K).$$

Since $\Pi_j \ge T_{\min} \|A_{jS_T}\|_1$, we conclude

$$\begin{aligned}
\max_{j\in\overline{J}} \frac{A_{j\cdot}^\top H^{-1} A_{j\cdot}}{\Pi_j^2} &\le \rho \kappa^{-2}(A_{\overline{J}}, K) \max_{j\in\overline{J}} \frac{\|A_{j\cdot}\|_2}{\Pi_j} \\
&\le \frac{\rho}{\kappa^2(A_{\overline{J}}, K) T_{\min}} \max_{j\in\overline{J}} \left(\frac{\|A_{jS_T}\|_2}{\|A_{jS_T}\|_1} + \frac{\|A_{jS_T^c}\|_2}{\|A_{jS_T}\|_1} \right) \\
&\le \frac{\rho}{\kappa^2(A_{\overline{J}}, K) T_{\min}} \left(1 + \xi \sqrt{|S_T|^c} \right) \\
&\le \frac{1 \vee \xi}{\kappa^2(A_{\overline{J}}, K) T_{\min}^2} \left(1 + \xi \sqrt{K - s} \right) = B. \qquad \text{(I.2)}
\end{aligned}$$

We used the definition (14) in the penultimate step and used $\rho \le (1 \vee \xi)/T_{\min}$ in the last step. Thus, $\|Z_i\|_{\text{op}} \le B$. For the second moment of Z_i, we have

$$\begin{aligned}
\mathbb{E}[Z_i Z_i^\top] &= \mathbb{E}\left[\sum_{j\in\overline{J}} \frac{B_{ij} - \Pi_j}{\Pi_j^2} H^{-1/2} A_{j\cdot} A_{j\cdot}^\top H^{-1/2} \sum_{j\in\overline{J}} \frac{B_{ij} - \Pi_j}{\Pi_j^2} H^{-1/2} A_{j\cdot} A_{j\cdot}^\top H^{-1/2} \right] \\
&= \mathbb{E}\left[\sum_{j\in\overline{J}} \frac{B_{ij}}{\Pi_j^2} H^{-1/2} A_{j\cdot} A_{j\cdot}^\top H^{-1/2} \sum_{j\in\overline{J}} \frac{B_{ij}}{\Pi_j^2} H^{-1/2} A_{j\cdot} A_{j\cdot}^\top H^{-1/2} \right] \\
&\quad + \mathbb{E}\left[\sum_{j\in\overline{J}} \frac{1}{\Pi_j} H^{-1/2} A_{j\cdot} A_{j\cdot}^\top H^{-1/2} \sum_{j\in\overline{J}} \frac{1}{\Pi_j} H^{-1/2} A_{j\cdot} A_{j\cdot}^\top H^{-1/2} \right] \\
&\preceq \mathbb{E}\left[\sum_{j\in\overline{J}} \frac{B_{ij}}{\Pi_j^2} H^{-1/2} A_{j\cdot} A_{j\cdot}^\top H^{-1/2} \right] \max_{j\in\overline{J}} \frac{H^{-1/2} A_{j\cdot} A_{j\cdot}^\top H^{-1/2}}{\Pi_j^2} + \boldsymbol{I}_K \\
&= \boldsymbol{I}_K \max_{j\in\overline{J}} \frac{H^{-1/2} A_{j\cdot} A_{j\cdot}^\top H^{-1/2}}{\Pi_j^2} + \boldsymbol{I}_K.
\end{aligned}$$

Since

$$\max_{j\in\overline{J}} \frac{H^{-1/2} A_{j\cdot} A_{j\cdot}^\top H^{-1/2}}{\Pi_j^2} = \max_{j\in\overline{J}} \frac{A_{j\cdot}^\top H^{-1} A_{j\cdot}}{\Pi_j^2},$$

by (I.2), we conclude

$$\left\| \mathbb{E}[Z_i Z_i^\top] \right\|_{\text{op}} = \left\| \mathbb{E}[Z_i^\top Z_i] \right\|_{\text{op}} \le 1 + B \le 2B.$$

The first result then follows from an application of the Matrix Bernstein inequality. The second result follows immediately by using the first result with $t = 2 \log K$ and noting that

$$\lambda_{\min}(\widehat{H}) \ge \lambda_{\min}(H) \lambda_{\min}(H^{-1/2} \widehat{H} H^{-1/2}) \ge \lambda_{\min}(H) \left(1 - \|H^{-1/2}(\widehat{H} - H) H^{-1/2}\|_{\text{op}} \right).$$

We use Weyl's inequality in the second step. □

APPENDIX J: ALGORITHM OF ESTIMATING THE WORD-TOPIC MATRIX A

We recommend the following procedure for estimating the word-topic matrix A under Assumption 1. It consists of two parts: (a) estimation of the partition of anchor words, and (b) estimation of the word-topic matrix A. Step (a) uses the procedure proposed in Bing, Bunea and Wegkamp (2020a), stated in Algorithm 1 while step (b) uses the procedure proposed in Bing, Bunea and Wegkamp (2020b), summarized in Algorithm 2.

Recall that $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(n)})$ with N_i denoting the length of document i. Define

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

and

$$\widehat{R} = D_X^{-1}\widehat{\Theta}D_X^{-1} \tag{J.2}$$

with $D_X = n^{-1}\mathrm{diag}(\boldsymbol{X}\mathbf{1}_n)$.

J.1. Estimation of the index set of the anchor words, its partition and the number of topics. We write the set of anchor words as $I = \cup_{k\in[K]} I_k$ and its partition $\mathcal{I} = \{I_1, \ldots, I_K\}$ where

$$I_k = \{j \in [p] : A_{jk} > 0,\ A_{\ell k} = 0,\ \forall\, \ell \neq j\}.$$

Algorithm 1 estimates the index set I, its partition $\mathcal{I}$ and the number of topics K from the input matrix $\widehat{R}$. The choice $C_1 = 1.1$ is recommended and is empirically verified to be robust in Bing, Bunea and Wegkamp (2020a). A data-driven choice of $\delta_{j\ell}$ is specified in Bing, Bunea and Wegkamp (2020a) as

$$\widehat{\delta}_{j\ell} = \frac{n^2}{\|\boldsymbol{X}_{j\cdot}\|_1\|\boldsymbol{X}_{\ell\cdot}\|_1}\left\{\widehat{\eta}_{j\ell} + 2\widehat{\Theta}_{j\ell}\sqrt{\frac{\log M}{n}}\left[\frac{n}{\|\boldsymbol{X}_{j\cdot}\|_1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_{ji}}{N_i}\right)^{\frac{1}{2}} + \frac{n}{\|\boldsymbol{X}_{\ell\cdot}\|_1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_{\ell i}}{N_i}\right)^{\frac{1}{2}}\right]\right\} \tag{J.3}$$

with $M = n \vee p \vee \max_i N_i$ and

$$\widehat{\eta}_{j\ell} = 3\sqrt{6}\left(\|\boldsymbol{X}_{j\cdot}\|_\infty^{\frac{1}{2}} + \|\boldsymbol{X}_{\ell\cdot}\|_\infty^{\frac{1}{2}}\right)\sqrt{\frac{\log M}{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_{ji}\boldsymbol{X}_{\ell i}}{N_i}\right)^{\frac{1}{2}} + \\ + \frac{2\log M}{n}\left(\|\boldsymbol{X}_{j\cdot}\|_\infty + \|\boldsymbol{X}_{\ell\cdot}\|_\infty\right)\frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i} + 31\sqrt{\frac{(\log M)^4}{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_{ji}+\boldsymbol{X}_{\ell i}}{N_i^3}\right)^{\frac{1}{2}} \tag{J.4}$$

J.2. Estimation of the word-topic matrix A with a given partition of anchor words. Given the estimated partition of anchor words $\widehat{\mathcal{I}} = \{\widehat{I}_1, \ldots, \widehat{I}_{\widehat{K}}\}$ and its index set $\widehat{I} = \cup_{k\in[\widehat{K}]}\widehat{I}_k$, Algorithm 2 below estimates the matrix A.

Bing, Bunea and Wegkamp (2020b) recommends to set $\lambda = 0$ whenever $\widehat{M}$ is invertible and otherwise choose λ large enough such that $\widehat{M} + \lambda \boldsymbol{I}_K$ is invertible. Specifically, Bing, Bunea and Wegkamp (2020b) recommends to choose λ as

$$\lambda(t^*) = 0.01 \cdot t^* \cdot K\left(\frac{K\log(n\vee p)}{[\min_{i\in\widehat{I}}(D_X)_{ii}]n}\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i}\right)^{1/2}. \tag{J.5}$$

Algorithm 1 Estimate the partition of the anchor words $\mathcal{I}$ by $\widehat{\mathcal{I}}$

Require: matrix $\widehat{R} \in \mathbb{R}^{p\times p}$, C_1 and $Q \in \mathbb{R}^{p\times p}$ such that $Q[j,\ell] := C_1\delta_{j\ell}$

procedure FINDANCHORWORDS($\widehat{R}$, Q)
  initialize $\widehat{\mathcal{I}} = \varnothing$
  for $i \in [p]$ **do**
    $\widehat{a}_i = \arg\max_{1\le j\le p} \widehat{R}_{ij}$
    set $\widehat{I}^{(i)} = \{\ell \in [p] : \widehat{R}_{i\widehat{a}_i} - \widehat{R}_{il} \le Q[i,\widehat{a}_i] + Q[i,\ell]\}$ and ANCHOR(i) = TRUE
    for $j \in \widehat{I}^{(i)}$ **do**
      $\widehat{a}_j = \arg\max_{1\le k\le p} \widehat{R}_{jk}$
      if $\left|\widehat{R}_{ij} - \widehat{R}_{j\widehat{a}_j}\right| > Q[i,j] + Q[j,\widehat{a}_j]$ **then**
        ANCHOR(i) = FALSE
        break
    if ANCHOR(i) **then**
      $\widehat{\mathcal{I}} =$ MERGE$(\widehat{I}^{(i)}, \widehat{\mathcal{I}})$
  return $\widehat{\mathcal{I}} = \{\widehat{I}_1, \widehat{I}_2, \ldots, \widehat{I}_{\widehat{K}}\}$

procedure MERGE($\widehat{I}^{(i)}, \widehat{\mathcal{I}}$)
  for $G \in \widehat{\mathcal{I}}$ **do**
    if $G \cap \widehat{I}^{(i)} \neq \varnothing$ **then**
      replace G in $\widehat{\mathcal{I}}$ by $G \cap \widehat{I}^{(i)}$
      return $\widehat{\mathcal{I}}$
  $\widehat{I}^{(i)} \in \widehat{\mathcal{I}}$
  return $\widehat{\mathcal{I}}$

where

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

Algorithm 2 Sparse Topic Model solver (STM)

Require: frequency data matrix $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ with document lengths $N_1, \ldots, N_n$; the partition of anchor words $\{I_1, \ldots, \widehat{I}_{\widehat{K}}\}$ and its index set $\widehat{I} = \cup_{k\in[\widehat{K}]}\widehat{I}_k$, the tuning parameter $\lambda \ge 0$

procedure
  compute $D_X = n^{-1}\mathrm{diag}(\boldsymbol{X}\mathbf{1}_n)$, $\widehat{\Theta}$ from (J.1) and $\widehat{R}$ from (J.2)
  compute $\widehat{B}_{\widehat{I}\cdot}$ by $\widehat{B}_{i\cdot} = \boldsymbol{e}_k$ for each $i \in \widehat{I}_k$ and $k \in [\widehat{K}]$
  compute $\widehat{M} = \widehat{B}^{+}_{\widehat{I}\cdot}\widehat{R}_{\widehat{I}\widehat{I}}\widehat{B}^{+\top}_{\widehat{I}\cdot}$ and $\widehat{H} = \widehat{B}^{+}_{\widehat{I}\cdot}\widehat{R}_{\widehat{I}\widehat{I}^c}$ with $\widehat{B}^{+}_{\widehat{I}\cdot} = (\widehat{B}^{\top}_{\widehat{I}\cdot}\widehat{B}_{\widehat{I}\cdot})^{-1}\widehat{B}^{\top}_{\widehat{I}\cdot}$ and $\widehat{I}^c = [p] \setminus \widehat{I}$
  solve $\widehat{B}_{\widehat{I}^c\cdot}$ from

$$\widehat{B}_{j\cdot} = 0, \qquad \text{if } (D_X)_{jj} \le \frac{7\log(n\vee p)}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i}\right),$$

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

  for each $j \in \widehat{I}^c$, with $\widehat{h}^{(j)}$ being the corresponding column of $\widehat{H}$.
  compute $\widehat{A}$ by normalizing $D_X\widehat{B}$ to unit column sums
  return $\widehat{A}$

APPENDIX K: SOME EXISTING RESULTS ON ESTIMATION OF A

For completeness, we state the upper bounds of the estimator $\widehat{A}$ of A proposed in Bing, Bunea and Wegkamp (2020a) as well as the conditions under which $\widehat{A}$ is optimal in the minimax sense, up to a logarithmic factor.

Let $N = N_i$ for all $i \in [n]$ for simplicity and write $M = n \vee p \vee N$. Under Assumption 1, recall that I denotes the index set of anchor words and $I^c = [p] \setminus I$. From Corollary 8 of Bing, Bunea and Wegkamp (2020a), under the conditions stated in Appendix K.1, the following holds with probability at least $1 - 8M^{-1}$,

$$\min_{P \in \mathcal{P}_K} \|\widehat{A} - AP\|_{1,\infty} \lesssim \sqrt{\frac{(|I| + K|I^c|)\log M}{nN}}, \tag{K.1}$$

$$\min_{P \in \mathcal{P}_K} \|\widehat{A} - AP\|_{1} \lesssim K\sqrt{\frac{(|I| + K|I^c|)\log M}{nN}}, \tag{K.2}$$

On the other hand, the minimax lower bounds in Theorem 6 of Bing, Bunea and Wegkamp (2020a) further imply that the rates in (K.1) – (K.2) are minimax optimal, up to the $\log(M)$ factor.

K.1. Conditions under which (K.1) – (K.2) hold. Let $\boldsymbol{T} := \boldsymbol{T}_*$ and $\boldsymbol{\Pi} := \boldsymbol{\Pi}_*$. Define

$$\nu := n\zeta_i\zeta_j \left[\frac{\zeta_i}{\zeta_j} \wedge \frac{\zeta_j}{\zeta_i} - \cos(\angle(\boldsymbol{T}_{i\cdot}, \boldsymbol{T}_{i\cdot}))\right]$$

with $\zeta_i = \|\boldsymbol{T}_{i\cdot}\|_2 / \|\boldsymbol{T}_{i\cdot}\|_1$. This quantity quantifies the incoherence between rows of $\boldsymbol{T}$.

(1) The matrix A satisfies
   a) the anchor word assumption in Assumption 1,
   b) the balancing condition $\max_{i\in I} \|A_{i\cdot}\|_\infty \asymp \min_{i\in I} \|A_{i\cdot}\|_\infty$ and

$$\frac{1}{|I^c|} \sum_{j \in I^c} \frac{\|A_{j\cdot}\|_\infty}{\max_{i\in I} \|A_{i\cdot}\|_\infty} \lesssim 1,$$

   c) the separation condition between anchor and non-anchor words

$$\min_{i\in I, j\in I^c} \|\widetilde{A}_{i\cdot} - \widetilde{A}_{j\cdot}\|_1 \geq 8\delta/\nu$$

   where $A = D_\Pi^{-1} A D_T$ with $D_\Pi = \mathrm{diag}(\boldsymbol{\Pi}\mathbf{1}_n)$ and $D_T = \mathrm{diag}(\boldsymbol{T}\mathbf{1}_n)$ and the expression of δ is stated below.

(2) The matrix $\boldsymbol{T} := \boldsymbol{T}_*$ satisfies
   a) $\mathrm{rank}(\boldsymbol{T}) = K$,
   b) the incoherence condition $\nu > 4\delta$,
   c) the balancing condition $\max_{k\in[K]} \sum_{i=1}^n \boldsymbol{T}_{ki} \asymp \min_{k\in[K]} \sum_{i=1}^n \boldsymbol{T}_{ki}$,
   d) the weak dependency condition $\sum_{k'\neq k} \sqrt{C_{kk'}} \lesssim \sqrt{C_{kk}}$ for all $k \in [K]$ with $C = n^{-1}\boldsymbol{T}\boldsymbol{T}^\top$;

(3) The matrix $\boldsymbol{\Pi} := \boldsymbol{\Pi}_*$ satisfies

$$\min_{j\in[p]} \frac{1}{n} \sum_{i=1}^n \boldsymbol{\Pi}_{ji} \geq \frac{2\log M}{3N}, \qquad \min_{j\in[p]} \max_{1\leq i\leq n} \boldsymbol{\Pi}_{ji} \geq \frac{(3\log M)^2}{N}.$$

For detailed interpretation and justification of the above conditions, we refer the reader to Remarks 2, 3, 9, 10 & 11 of Bing, Bunea and Wegkamp (2020a). The quantity δ mentioned

above represents the noise level in the context of estimating A, defined as $\delta = \max_{j,\ell\in[p]} \delta_{j\ell}$, where

$$\delta_{j\ell} := \frac{p^2\eta_{j\ell}}{\mu_j\mu_\ell} + \frac{p^2\Theta_{j\ell}}{\mu_j\mu_\ell}\left(\sqrt{\frac{p}{\mu_j}} + \sqrt{\frac{p}{\mu_\ell}}\right)\sqrt{\frac{\log M}{nN}}. \tag{K.3}$$

with

$$\eta_{j\ell} = \sqrt{\frac{\Theta_{j\ell}\log M}{nN}}\sqrt{\frac{m_j+m_\ell}{p} \vee \frac{\log^2 M}{N}} + \frac{2(m_j+m_\ell)}{p}\frac{\log M}{nN}$$

$$+ \sqrt{\frac{\log^4 M}{nN^3}}\sqrt{\frac{\mu_j+\mu_\ell}{p} \vee \frac{\log M}{N}}. \tag{K.4}$$

Here, $\Theta = n^{-1}\mathbf{\Pi}\mathbf{\Pi}^\top$,

$$\frac{m_j}{p} = \max_{1\le i\le n} \mathbf{\Pi}_{ji}, \qquad \frac{\mu_j}{p} = \frac{1}{n}\sum_{i=1}^n \mathbf{\Pi}_{ji}, \quad \forall\, j\in[p].$$

Let $\widehat{A}$ be the estimator obtained the procedure proposed in Bing, Bunea and Wegkamp (2020a). Write $|I_{\max}| = \max_{k\in[K]}|I_k|$ and recall that $M = n\vee p\vee N$.

THEOREM K.1. *Under conditions in (1) – (3) stated in Appendix K.1, assume*

$$(|I_{\max}| + K|I^c|)\log(M) \le cnN \tag{K.5}$$

for some absolute constant $c\in(0,1)$. Then there exists some permutation matrix $P\in\mathcal{P}_K$ such that, with probability $1-8M^{-1}$, we have

$$\|(\widehat{A}P)_{j\cdot} - A_{j\cdot}\|_\infty \lesssim \sqrt{\|A_{j\cdot}\|_\infty \frac{K\log(M)}{nN}}\left(1\vee\sqrt{p\|A_{j\cdot}\|_\infty}\right), \qquad \forall\, j\in[p].$$

PROOF. The proof repeatedly uses the results and proofs in the supplement of Bing, Bunea and Wegkamp (2020a). We only go through the major steps here and refer the reader to Bing, Bunea and Wegkamp (2020a) for detailed notation and formal statements.

We work on the event $\mathcal{E} := \mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_2$ defined in page 8 of the supplement of Bing, Bunea and Wegkamp (2020a). Recall that $\widehat{A} = T^{-1}\sum_{i=1}^T \widehat{A}^i$. It suffices to prove the desired result for any $i\in[T]$. We follow the arguments in the proof of Theorem 7 of Bing, Bunea and Wegkamp (2020a) and write $\widehat{A} = \widehat{A}^i$ for simplicity.

We first recall from Theorem 7 of Bing, Bunea and Wegkamp (2020a) that, by assuming the identity permutation without loss of generality, $\widehat{K} = K$ and $\widehat{I}_k = I_k$ for all $k\in[K]$ hold on $\mathcal{E}$. As a result, we have $\widehat{L} = L$ with $L = \{i_1,\dots,i_K\}$ and $i_k\in I_k$ for each $k\in[K]$. From page 28 of the supplement of Bing, Bunea and Wegkamp (2020a), we have, for any $j\in[p]$ and $k\in[K]$,

$$\begin{aligned}|\widehat{A}_{jk} - A_{jk}| &\le \frac{\left|\|\widehat{B}_{\cdot k}\|_1 - \|B_{\cdot k}\|_1\right|}{\|B_{\cdot k}\|_1}\widehat{A}_{jk} + \frac{|\widehat{B}_{jk}-B_{jk}|}{\|B_{\cdot k}\|_1}\\ &\le \frac{\|\widehat{B}_{\cdot k}-B_{\cdot k}\|_1}{\|B_{\cdot k}\|_1}\left(A_{jk} + |\widehat{A}_{jk}-A_{jk}|\right) + \frac{|\widehat{B}_{jk}-B_{jk}|}{\|B_{\cdot k}\|_1}\end{aligned}$$

and $\|B_{\cdot k}\|_1 = p/\alpha_{i_k}$, where, following Bing, Bunea and Wegkamp (2020a), we define

$$\text{(K.6)} \qquad \alpha_j := p \max_{1\le k\le K} A_{jk}, \qquad \gamma_k := \frac{K}{n}\sum_{i=1}^{n} W_{ki}, \qquad \text{for each } j\in[p],\ k\in[K].$$

Since Corollary 8 of Bing, Bunea and Wegkamp (2020a) ensures that

$$\max_{k\in[K]} \frac{\|\widehat{B}_{\cdot k} - B_{\cdot k}\|_1}{\|B_{\cdot k}\|_1} \lesssim \sqrt{\frac{(|I_{\max}| + K|I^c|)\log(M)}{nN}}$$

with $|I_{\max}| = \max_k |I_k|$, condition (K.5) implies

$$(1-c)|\widehat{A}_{jk} - A_{jk}| \le A_{jk}\sqrt{\frac{(|I_{\max}| + K|I^c|)\log(M)}{nN}} + \frac{\alpha_{i_k}}{p}|\widehat{B}_{jk} - B_{jk}|,$$

hence

$$\text{(K.7)} \qquad \|\widehat{A}_{j\cdot} - A_{j\cdot}\|_\infty \lesssim \|A_{j\cdot}\|_\infty \sqrt{\frac{(|I_{\max}| + K|I^c|)\log(M)}{nN}} + \max_k \frac{\alpha_{i_k}}{p}|\widehat{B}_{jk} - B_{jk}|,$$

It thus remains to bound the second term. We distinguish two cases:

(i) If $j \in I_k$ for some $k \in [K]$, then by using the fact that $|\widehat{B}_{jk} - B_{jk}| = \|\widehat{B}_{Ik} - B_{Ik}\|_1$ whenever $|I_k| = 1$, invoking the bound of $\|\widehat{B}_{Ik} - B_{Ik}\|_1$ in display (58) of the supplement of Bing, Bunea and Wegkamp (2020a) with $I_k = \{j\}$ yields

$$|\widehat{B}_{jk} - B_{jk}| \ \lesssim\ \frac{\alpha_j}{\alpha_{i_k}\sqrt{\underline{\alpha}_I \gamma_k}}\sqrt{\frac{pK\log(M)}{nN}}$$

where

$$\underline{\alpha}_I = \min_{i\in I}\alpha_i, \qquad \overline{\alpha}_I = \max_{i\in I}\alpha_i.$$

As a result, by using $\gamma_k \asymp 1$ which is implied by condition (2) c) in Appendix K.1, we have

$$\max_k \frac{\alpha_{i_k}}{p}|\widehat{B}_{jk} - B_{jk}| \ \lesssim\ \frac{\alpha_j}{\sqrt{\underline{\alpha}_I\gamma_k}}\sqrt{\frac{K\log(M)}{npN}} = \|A_{j\cdot}\|_\infty\sqrt{\frac{pK\log(M)}{nN}}$$

where we also use $\underline{\alpha}_I \asymp \alpha_j$ for $j \in I$ from condition (1) b).

(ii) If $j \in I^c$, then following the arguments in page 27 of the supplement of Bing, Bunea and Wegkamp (2020a), one can deduce

$$\begin{aligned}
|\widehat{B}_{jk} - B_{jk}| &\le \|\widehat{\omega}_k\|_1\|\widehat{\Theta}_{jL} - \Theta_{jL}\|_\infty + \|B_{j\cdot}\|_\infty\left(\|\widehat{\Theta}_{LL}\widehat{\omega}_k - \boldsymbol{e}_k\|_1 + \|\widehat{\omega}_k\|_1\|\widehat{\Theta}_{LL} - \Theta_{LL}\|_{\infty,1}\right)\\
&\le C_0\|\omega_k\|_1 \max_{i\in L}\eta_{ij} + 2\|B_{j\cdot}\|_\infty\|\omega_k\|_1\lambda\\
&\lesssim \frac{p^2}{\alpha_{i_k}\underline{\alpha}_I}\|C^{-1}\|_{\infty,1}\left(\max_{i\in L}\eta_{ij} + \frac{p}{\underline{\alpha}_I}\|A_{j\cdot}\|_\infty\max_{i\in L}\sum_{j\in L}\eta_{ij}\right)\\
&\lesssim \frac{p^2K}{\alpha_{i_k}\underline{\alpha}_I}\left(\max_{i\in L}\eta_{ij} + \frac{p}{\underline{\alpha}_I}\|A_{j\cdot}\|_\infty\sqrt{\frac{\overline{\alpha}_I^3\overline{\gamma}\log M}{Knp^3N}}\right)
\end{aligned}$$

with $\overline{\gamma} = \max_k \gamma_k$, where in the penultimate step we have used $\|C^{-1}\|_{\infty,1} \lesssim K$ and the bound for $\max_{i\in L}\sum_{j\in L}\eta_{ij}$ in the proof of Corollary 8 of Bing, Bunea and Wegkamp (2020a). Then

by $\max_k \gamma_k \asymp 1$, $\overline{\alpha}_I \asymp \underline{\alpha}_I$, we obtain

$$\max_k \frac{\alpha_{i_k}}{p}|\widehat{B}_{jk} - B_{jk}| \lesssim \frac{pK}{\overline{\alpha}_I}\max_{i\in L}\eta_{ij} + \|A_{j\cdot}\|_\infty\sqrt{\frac{pK\log M}{\underline{\alpha}_I nN}}.$$

Finally, to bound $\max_{i\in L}\eta_{ij}$, recalling from display (57) of the supplement of Bing, Bunea and Wegkamp (2020a), we have, for any $i \in I_a$ and $a \in [K]$,

$$\eta_{ij} \lesssim \sqrt{\frac{1}{n}\langle \boldsymbol{T}_{a\cdot}, \boldsymbol{\Pi}_{j\cdot}\rangle}\sqrt{\frac{\alpha_i(\alpha_i+\alpha_j)\log M}{np^2N}} + \frac{(\alpha_i+\alpha_j)\log M}{npN} + \sqrt{\frac{(\alpha_i+\alpha_j)(\log M)^4}{npN^3}}.$$

Since

$$\langle \boldsymbol{T}_{a\cdot}, \boldsymbol{\Pi}_{j\cdot}\rangle = A_{j\cdot}^\top \frac{1}{n}\boldsymbol{T}\boldsymbol{T}_{a\cdot} \le \|A_{j\cdot}\|_\infty \frac{1}{n}\sum_{t=1}^n \boldsymbol{T}_{at} = \|A_{j\cdot}\|_\infty \frac{\gamma_a}{K} \lesssim \frac{\|A_{j\cdot}\|_\infty}{K},$$

we have

$$\begin{aligned}\frac{pK}{\overline{\alpha}_I}\max_{i\in L}\eta_{ij} \lesssim{}& \sqrt{\|A_{j\cdot}\|_\infty\left(1+\frac{\alpha_j}{\underline{\alpha}_I}\right)}\sqrt{\frac{K\log(M)}{nN}} + \left(1+\frac{\alpha_j}{\underline{\alpha}_I}\right)\frac{K\log(M)}{nN} \\ &+ \sqrt{\frac{\overline{\alpha}_I+\alpha_j}{\underline{\alpha}_I^2}}\sqrt{\frac{pK^2\log^4(M)}{nN^3}}.\end{aligned}$$

By using the same arguments in the proof of Lemma 13 of Bing, Bunea and Wegkamp (2020a), one can show the last two terms are smaller in order than the first term under condition (3) in the Appendix K.1. Therefore, we conclude

$$\frac{pK}{\overline{\alpha}_I}\max_{i\in L}\eta_{ij} \lesssim \sqrt{\|A_{j\cdot}\|_\infty\left(1+\frac{\alpha_j}{\underline{\alpha}_I}\right)}\sqrt{\frac{K\log(M)}{nN}} \lesssim \|A_{j\cdot}\|_\infty\sqrt{\frac{pK\log(M)}{(\alpha_j\wedge\underline{\alpha}_I)nN}}.$$

Since condition (2) b) implies

$$1 \le \frac{1}{p}\sum_{i=1}^p \alpha_i \lesssim \frac{1}{p}\left(|I^c|+|I|\right)\underline{\alpha}_I = \underline{\alpha}_I$$

and $\|A_{j\cdot}\|_\infty = \alpha_j/p$, we conclude

$$\max_k \frac{\alpha_{i_k}}{p}|\widehat{B}_{jk} - B_{jk}| \lesssim \max\left\{\|A_{j\cdot}\|_\infty\sqrt{\frac{pK\log(M)}{nN}}, \sqrt{\|A_{j\cdot}\|_\infty\frac{K\log(M)}{nN}}\right\}$$

for any $j \in I^c$, which together with case (i), display (K.7) and the fact that $|I_{\max}| + K|I^c| \le pK$ completes the proof. □

APPENDIX L: ERROR BOUNDS FOR $\widehat{T}_{\text{mle}} - T_*$ IN ℓ_2 NORM

In this section we state the results on $\|\widehat{T}_{\text{mle}} - T_*\|_2$ with $\widehat{T}_{\text{mle}}$ defined in (4) for known A.

Assume the conditions in Theorem 2. The display (F.9) in the proof of Theorem 2 yields the following ℓ_2 norm convergence rate of $T_{\min} - T^*$:

$$\|T_{\min} - T^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sigma^{-1}(I,s)\sqrt{\frac{K}{N}}\right) \tag{L.1}$$

where

$$\sigma^2(I,s) = \min_{S\subseteq[K],|S|\le s} \sup_{v\in\mathcal{C}(S)} \frac{v^\top I v}{\|v\|_2^2}, \qquad \text{with} \quad I = \sum_{j\in\overline{J}} \frac{A_{j\cdot}A_{j\cdot}^\top}{\Pi_j}.$$

On the event $\mathcal{E}_{supp}$, (L.1) could be improved to

$$\|T_{\min} - T^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sigma^{-1}(I,s)\sqrt{\frac{s}{N}}\right) \tag{L.2}$$

When $\sigma^{-1}(I,s) = \mathcal{O}(1/\sqrt{s})$, the rate in (L.2) is minimax optimal according to Theorem 7. Indeed, since $\widehat{T} - T_* \in \mathcal{C}(s)$ for any $T_* \in \mathcal{T}'(s)$ and $\widehat{T} \in \Delta_K$, we have $\|\widehat{T} - T_*\|_1 \le 2\|[\widehat{T} - T_*]_{S_*}\|_1 \le 2\sqrt{s}\|\widehat{T} - T_*\|_2$ with $S_* = \mathrm{supp}(T_*)$ and $|S_*| = s$. Consequently, Theorem 7 implies

$$\inf_{\widehat{T}} \sup_{T_*\in\mathcal{T}'(s)} \mathbb{P}\left\{\|\widehat{T} - T_*\|_2 \ge c_0\sqrt{\frac{1}{N}}\right\} \ge c_1.$$

REMARK L.1 (Discussion on $\sigma(I,s)$). To understand the magunitude of $\sigma(I,s)$, it is helpful to consider $s = K$, in which case $\sigma^2(I,K)$ is simply the smallest eigenvalue of I. We then have

$$\sigma^2(I,K) \le \min_k \sum_{j\in\overline{J}} \frac{A_{jk}^2}{\Pi_j} \le \min_k \sum_{j\in\overline{J}} \frac{A_{jk}^2}{A_{jk}T_k} \le \max_k \frac{1}{T_k} \le K,$$

where the last step uses $\max_k T_k \ge 1/K$. We also note that this upper bound is attainable (in terms of rates), for instance, when all words are anchor words and the numbers of anchor words of all topics are the same order. Immediately, when $\sigma(I,K) \asymp \sqrt{K}$, (L.1) yields

$$\|\widehat{T}_{\mathrm{mle}} - T^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{1/N}\right).$$

Next, we connect $\sigma(I,s)$ to a quantity that is only related with A and s. By (F.8), we have

$$\sigma(I,s) \ge \kappa_2(A_{\overline{J}}, s)$$

where

$$\kappa_2(A_{\overline{J}}, s) = \min_{S\subseteq[K],|S|\le s} \sup_{v\in\mathcal{C}(S)} \frac{\|A_{\overline{J}}v\|_1}{\|v\|_2}.$$

Since

$$\frac{\kappa_2(A_{\overline{J}}, s)}{\sqrt{s}} \le \kappa(A_{\overline{J}}, s) \le \kappa_2(A_{\overline{J}}, s),$$

the ideal case is $\kappa_2(A_{\overline{J}}, s) \asymp \sqrt{s}\,\kappa(A_{\overline{J}}, s)$, whence

$$\|\widehat{T}_{\mathrm{mle}} - T^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(\kappa^{-1}(A_{\overline{J}}, s)\sqrt{1/N}\right).$$

REFERENCES

AGRESTI, A. (2012). *Categorical Data Analysis, 3rd Edition. Wiley Series in Probability and Statistics*. Wiley.

ARORA, S., GE, R., HALPERN, Y., MIMNO, D. M., MOITRA, A., SONTAG, D., WU, Y. and ZHU, M. (2013). A Practical Algorithm for Topic Modeling with Provable Guarantees. In *ICML (2)* 280–288.

BING, X., BUNEA, F. and WEGKAMP, M. (2020a). A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli* **26** 1765–1796.

BING, X., BUNEA, F. and WEGKAMP, M. (2020b). Optimal estimation of sparse topic models. *Journal of Machine Learning Research* **21** 1–45.

BISHOP, Y. M., FIENBERG, S. E. and HOLLAND, P. W. (2007). *Discrete Multivariate Analysis Theory and Practice*. Springer, New York Originally published by MIT Press, 1975.

BITTORF, V., RECHT, B., RE, C. and TROPP, J. A. (2012). Factoring nonnegative matrices with linear programs. *arXiv:1206.1270*.

CAO, Y., ZHANG, A. and LI, H. (2020). Multisample estimation of bacterial composition matrices in metagenomics data. *Biometrika* **107** 75–92.

DUA, D. and GRAFF, C. (2017). UCI Machine Learning Repository.

GIBBS, A. L. and SU, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review* **70** 419–435.

MAAS, A. L., DALY, R. E., PHAM, P. T., HUANG, D., NG, A. Y. and POTTS, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 142–150. Association for Computational Linguistics, Portland, Oregon, USA.

MIKOLOV, T., CHEN, K., CORRADO, G. and DEAN, J. (2013). Efficient Estimation of Word Representations in Vector Space.

TROPP, J. A. (2015). *An Introduction to Matrix Concentration Inequalities*. *Foundations and trends in machine learning*. Now Publishers.

TSYBAKOV, A. B. (2008). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer New York.

VILLANI, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society, Providence.

ZHU, Z., LI, X., WANG, M. and ZHANG, A. (2021). Learning Markov models via low-rank optimization. *Operations Research*.