The Annals of Statistics
2022, Vol. 50, No. 6, 3307–3333
https://doi.org/10.1214/22-AOS2229

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

BY XIN BING[1,a], FLORENTINA BUNEA[2,b], SETH STRIMAS-MACKEY[2,c] AND MARTEN WEGKAMP[3,d]

[1]*Department of Statistical Sciences, University of Toronto,* [a]*xin.bing@utoronto.ca*
[2]*Department of Statistics and Data Science, Cornell University,* [b]*fb238@cornell.edu,* [c]*scs324@cornell.edu*
[3]*Departments of Mathematics, and of Statistics and Data Science, Cornell University,* [d]*mhw73@cornell.edu*

This paper studies the estimation of high-dimensional, discrete, possibly sparse, mixture models in the context of topic models. The data consists of observed multinomial counts of p words across n independent documents. In topic models, the $p \times n$ expected word frequency matrix is assumed to be factorized as a $p \times K$ word-topic matrix A and a $K \times n$ topic-document matrix T. Since columns of both matrices represent conditional probabilities belonging to probability simplices, columns of A are viewed as p-dimensional mixture components that are common to all documents while columns of T are viewed as the K-dimensional mixture weights that are document specific and are allowed to be sparse.

The main interest is to provide sharp, finite sample, ℓ_1-norm convergence rates for estimators of the mixture weights T when A is either known or unknown. For known A, we suggest MLE estimation of T. Our nonstandard analysis of the MLE not only establishes its ℓ_1 convergence rate, but also reveals a remarkable property: the MLE, with no extra regularization, can be exactly sparse and contain the true zero pattern of T. We further show that the MLE is both minimax optimal and adaptive to the unknown sparsity in a large class of sparse topic distributions. When A is unknown, we estimate T by optimizing the likelihood function corresponding to a plug in, generic, estimator $\widehat{A}$ of A. For any estimator $\widehat{A}$ that satisfies carefully detailed conditions for proximity to A, we show that the resulting estimator of T retains the properties established for the MLE. Our theoretical results allow the ambient dimensions K and p to grow with the sample sizes.

Our main application is to the estimation of 1-Wasserstein distances between document generating distributions. We propose, estimate and analyze new 1-Wasserstein distances between alternative probabilistic document representations, at the word and topic level, respectively. We derive finite sample bounds on the estimated proposed 1-Wasserstein distances. For word level document-distances, we provide contrast with existing rates on the 1-Wasserstein distance between standard empirical frequency estimates. The effectiveness of the proposed 1-Wasserstein distances is illustrated by an analysis of an IMDB movie reviews data set. Finally, our theoretical results are supported by extensive simulation studies.

1. Introduction. We consider the problem of estimating high-dimensional, discrete, mixture distributions, in the context of topic models. The focus of this work is the estimation, with sharp finite sample convergence rates, of the distribution of the latent topics within the

Received July 2021; revised September 2022.
MSC2020 subject classifications. Primary 62H12, 62H30; secondary 62F10.
Key words and phrases. Adaptive estimation, high-dimensional estimation, maximum likelihood estimation, minimax estimation, multinomial distribution, mixture model, sparse estimation, nonnegative matrix factorization, topic models, anchor words.

documents of a corpus. Our main application is to the estimation of Wasserstein distances between document generating distributions.

In the framework and traditional jargon of topic models, one has access to a corpus of n documents generated from a common set of K latent topics. Each document $i \in [n] := \{1, \dots, n\}$ is modeled as a set of N_i words drawn from a discrete distribution $\Pi_*^{(i)}$ on p points, where p is the dictionary size. We observe the p-dimensional word-count vector $Y^{(i)}$ for each document $i \in [n]$, where we assume

$$Y^{(i)} \sim \text{Multinomial}_p\big(N_i, \Pi_*^{(i)}\big).$$

The topic model assumption is that the matrix of expected word frequencies in the corpus, $\mathbf{\Pi}_* := (\Pi_*^{(1)}, \dots, \Pi_*^{(n)})$ can be factorized as

$$\mathbf{\Pi}_* = A\boldsymbol{T}_* \tag{1}$$

Here, A represents the $p \times K$ matrix of conditional probabilities of a word, given a topic and, therefore, each column of A belongs to the p-dimensional probability simplex

$$\Delta_p := \big\{x \in \mathbb{R}^p \mid x \succeq \mathbf{0}, \mathbf{1}_p^\top x = 1\big\}.$$

The notation $x \succeq \mathbf{0}$ represents $x_j \geq 0$ for each $j \in [p]$, and $\mathbf{1}_p$ is the vector of all ones. The $K \times n$ matrix $\boldsymbol{T}_* := (T_*^{(1)}, \dots, T_*^{(n)})$ collects the probability vectors $T_*^{(i)} \in \Delta_K$, the simplex in $\mathbb{R}^K$. The entries of $T_*^{(i)}$ are probabilities with which each of the K topics occurs within document i, for each $i \in [n]$. Relationship (1) would be a very basic application of Bayes' theorem if A also depended on i. A matrix A that is common across documents is the topic model assumption, which we will make in this paper.

Under model (1), each distribution on words, $\Pi_*^{(i)} = AT_*^{(i)} \in \Delta_p$, is a discrete mixture of $K < p$ distributions. The mixture components correspond to the columns of A, and are therefore common to the entire corpus, while the weights, given by the entries of $T_*^{(i)}$, are document specific. Since not all topics are expected to be covered by all documents, the mixture weights are potentially sparse, in that $T_*^{(i)}$ may be sparse. Using their dual interpretation, throughout the paper we will refer to a vector $T_*^{(i)}$ as either the topic distribution or the vector of mixture weights, in document i.

The observed word frequencies are collected in a $p \times n$ data matrix $\boldsymbol{X} = (X^{(1)}, \dots, X^{(n)})$ with independent columns $X^{(i)} = Y^{(i)}/N_i$ corresponding to the ith document. Our main interest is to estimate $\boldsymbol{T}_*$ when either the matrix A is known or unknown. We allow for the ambient dimensions K and p to depend on the sizes of the samples $\{N_1, \dots, N_n\}$ and n throughout the paper.

While, for ease of reference to the existing literature, we will continue to employ the text analysis jargon for the remainder of this work, and our main application will be to the analysis of a movie review data set, our results apply to any data set generated from a model satisfying (1), for instance in biology [18, 22], hyperspectral unmixing [29] and collaborative filtering [25].

The specific problems treated in this work are listed below, and expanded upon in the following subsections:

1. The main focus of this paper is on the derivation of sharp, finite-sample, ℓ_1-error bounds for estimators $\widehat{T}^{(i)}$ of the potentially sparse topic distributions $T_*^{(i)}$, under model (1), for each $i \in [n]$. The finite sample analysis covers two cases, corresponding to whether the components of the mixture, provided by the columns of A, are either (i) known or (ii) unknown, and estimated by $\widehat{A}$ from the corpus data $\boldsymbol{X}$. As a corollary, we derive corresponding finite sample ℓ_1-norm error bounds for mixture model-based estimators of $\Pi_*^{(i)}$.

2. The main application of our work is to the construction and analysis of similarity measures between the documents of a corpus, for measures corresponding to estimates of the Wasserstein distance between different probabilistic representations of a document.

1.1. *A finite sample analysis of topic and word distribution estimators.* Finite sample error bounds for estimators $\widehat{A}$ of A in topic models (1) have been studied in [3, 5, 9, 10, 24], while the finite sample properties of estimators of $T_*^{(i)}$ and, by extension, those of mixture-model-based estimators of $\Pi_*^{(i)}$, are much less understood, even when A is known beforehand and, therefore, $\widehat{A} = A$.

When $\Pi_*^{(i)} \in \Delta_p$ is a probability vector parametrized as $\Pi_*^{(i)} = g(T)$, with $T \in \mathbb{R}^K$, $K < p$ and some *known* function g, provided that T is identifiable, the study of the asymptotic properties of the maximum likelihood estimator (MLE) of T, derived from the p-dimensional vector of observed counts $Y^{(i)}$, is over eight decades old. Proofs of the consistency and asymptotic normality of the MLE, when the ambient dimensions K and p do not depend on the sample size, can be traced back to [32, 33] and later to the seminal work of [11], and are reproduced, in updated forms, in standard textbooks on categorical data [1, 12].

The mixture parametrization treated in this work, when A is known, is an instance of these well-studied low-dimensional parametrizations. Specialized to our context, for document $i \in [n]$, the parametrization is $\Pi_{*j}^{(i)} = A_{j\cdot}^\top T_*^{(i)}$ with $T_*^{(i)} \in \Delta_K$, for each component $j \in [p]$ of $\Pi_*^{(i)}$. However, even when p and K are fixed, the aforementioned classical asymptotic results are not applicable, as they are established under the following key assumptions that typically do not hold for topic models:

(1) $0 < T_{*k}^{(i)} < 1$, for *all* $k \in [K]$,
(2) $\Pi_{*j}^{(i)} = A_{j\cdot}^\top T_*^{(i)} > 0$ for *all* $j \in [p]$.

The regularity assumption (1) is crucial in classical analyses [32, 33], and stems from the basic requirement of M-estimation that $T_*^{(i)}$ be an interior point in its appropriate parameter space. In effect, since $\sum_{k=1}^K T_{*k}^{(i)} = 1$, this is a requirement on only a $(K-1)$ subvector of it. In the context of topic models, a given document i of the corpus may not touch upon all K topics, and in fact is expected not to. Therefore, it is expected that $T_{*k}^{(i)} = 0$, for some k. Furthermore, K represents the number of topics common to the entire corpus, and although topic k may not appear in document i, it may be the leading topic of some other document j. Both presence and absence of a topic in a document are subject to discovery, and are not known prior to estimation. Moreover, one does not observe the topic proportions $T_*^{(i)}$ per document i directly. Therefore, one cannot use background knowledge, for any given document, to reduce K to a smaller dimension in order to satisfy assumption (1).

The classical assumption (2) also typically does not hold for topic models. To see this, note that the matrix A is also expected to be sparse: conditional on a topic k, some of the words in a large p-dimensional dictionary will not be used in that topic. Therefore, in each column $A_{\cdot k}$, we expect that $A_{jk} = 0$, for many rows $j \in [p]$. When the supports of $A_{j\cdot}$ and $T_*^{(i)}$ do not intersect, the corresponding probability of word j in document i is zero, $\Pi_{*j}^{(i)} = A_{j\cdot}^\top T_*^{(i)} = 0$. Since zero word probabilities are induced by unobservable sparsity in the topic distribution (or, equivalently, in the mixture weights), one once again cannot reduce the dimension p a priori in a theoretical analysis. Therefore, the assumption (2) is also expected to fail.

The analysis on the MLE of $T_*^{(i)}$ is thus an open problem with A being known even for fixed p scenarios, when the standard assumptions (1) and (2) do not hold and when the problem cannot be artificially reduced to a framework in which they do.

Finite sample analysis of the rates of the MLE of topic distributions, for known A. In Section 2.1, we provide a novel analysis of the MLE of $T_*^{(i)}$ for known A, under a sparse discrete mixture framework, in which both the ambient dimensions K and p are allowed to grow with the sample sizes N_i and n. Kleinberg and Sandler [25] refer to the assumption of A being known as the semiomniscient setting in the context of collaborative filtering and note that even this setting is, surprisingly, very challenging for estimating the mixture weights. By studying the MLE of $T_*^{(i)}$ when A is known, one gains appreciation of the intrinsic difficulty of this problem, that is present even before one further takes into account the estimation of the entire $p \times K$ matrix A.

To the best of our knowledge, the only existing work that treats the aspect of our problem is [4], under the assumptions that

(a) the support S_* of $T_*^{(i)}$ is known and $T_{\min} := \min_{k \in S_*} T_{*k}^{(i)} \geq c/s$ with $s = |S_*|$ and $c \in (0, 1]$,

(b) the matrix A is known and $\kappa = \min_{\|x\|_1 = 1} \|Ax\|_1 > 0$.

The parameter κ is called the $\ell_1 \to \ell_1$ condition number of A [25], which measures the amount of linear independence between columns of A that belong to the simplex Δ_p. Under (a) and (b), the problem framework is very close to the classical one, and the novelty in [4] resides in the provision of a finite sample ℓ_1-error bound of the difference between the restricted MLE (restricted to the known support S_*) and the true $T_*^{(i)}$, a bound that is valid for growing ambient dimensions. However, assumption (a) is rather strong, as the support of $T_*^{(i)}$ is typically unknown. Furthermore, the restriction $\sum_{k \in S_*} T_{*k}^{(i)} = 1$ implies that $T_{\min} \leq 1/s$. Hence (a) essentially requires $T_*^{(i)}$ to be approximately uniform on its a priori known support. This does not hold in general. For instance, even if the support were known, many documents will primarily cover a very small number of topics, while only mentioning the rest, and thus some topics will be much more likely to occur than others, per document.

Our novel finite sample analysis in Section 2.1 avoids the strong condition (a) in [4]. For notational simplicity, we pick one $i \in [n]$ and drop the superscripts (i) in $X^{(i)}$, $T_*^{(i)}$ and $\Pi_*^{(i)}$ within this section. In Theorem 1 of Section 2.1.1, we first establish a general bound for the ℓ_1-norm of the error $(\widehat{T}_{\mathrm{mle}} - T_*)$, with $\widehat{T}_{\mathrm{mle}}$ being the MLE of T_*. Then, in Section 2.1.2, we use this bound as a preliminary result to characterize the regime in which the Hessian matrix of the loss in (4), evaluated at $\widehat{T}_{\mathrm{mle}}$, is close to its population counterpart (see condition (16) in Section 2.1.2). When this is the case, we prove a potentially faster rate of $\|\widehat{T}_{\mathrm{mle}} - T_*\|_1$ in Theorem 2. A consequence of both Theorem 1 and Theorem 2 is summarized in Corollary 3 of Section 2.1.2 for the case when T_* is dense such that $S_* = [K]$. For dense T_*, provided that $T_{\min}^3 \geq C \log(K)/(\kappa^4 N_i)$ for some sufficiently large constant $C > 0$, $\|\widehat{T}_{\mathrm{mle}} - T_*\|_1$ achieves the parametric rate $\sqrt{K/N_i}$, up to a multiplicative factor κ^{-1}.

As mentioned earlier, since T_* is not necessarily an interior point, we cannot appeal to the standard theory of the MLE, nor can we rely on having a zero gradient of the log likelihood at $\widehat{T}_{\mathrm{mle}}$. Instead, our proofs of Theorem 1 and 2 consist of the following key steps:

- We prove that the KKT conditions of maximizing the log likelihood under the restriction that $\widehat{T}_{\mathrm{mle}} \in \Delta_K$ lead to a quadratic inequality in $(\widehat{T}_{\mathrm{mle}} - T_*)$ of the form $(\widehat{T}_{\mathrm{mle}} - T_*)^\top \widetilde{H} (\widehat{T}_{\mathrm{mle}} - T_*) \leq (\widehat{T}_{\mathrm{mle}} - T_*)^\top E$, where (the infinity norm of) E is defined in the next point, and

$$\widetilde{H} = \sum_{j : X_j > 0} \frac{X_j}{\Pi_{*j} A_{j\cdot}^\top \widehat{T}_{\mathrm{mle}}} A_{j\cdot} A_{j\cdot}^\top.$$

- We bound the linear term of this inequality by $\|E\|_\infty \|\widehat{T}_{\mathrm{mle}} - T_*\|_1$ together with a sharp concentration inequality (Lemma I.2 of Appendix I [8]) for
$$\|E\|_\infty = \max_{k\in[K]} \left| \sum_{j:\Pi_{*j}>0} \frac{A_{jk}}{\Pi_{*j}} (X_j - \Pi_{*j}) \right|.$$
- We prove that the quadratic term can be bounded from below by $(\kappa^2/2)\|\widehat{T}_{\mathrm{mle}} - T_*\|_1^2$, using the definition of the $\ell_1 \to \ell_1$ condition number of A, and control of the ratios X_j/Π_{*j} over a suitable subset of indices j such that $X_j > 0$.
- The faster rate in Theorem 2 requires a more delicate control of $\widetilde{H}$, and its analysis is complicated by the division by $A_{j\cdot}^\top \widehat{T}_{\mathrm{mle}}$. To this end, we use the bound in Theorem 1 to first prove that $A_{j\cdot}^\top \widehat{T}_{\mathrm{mle}} \le (1+c)\Pi_{*j}$, for all j with $\Pi_{*j} > 0$ and some constant $c \in (0,1)$. We then prove a sharp concentration bound (Lemma I.4 of Appendix I) for the operator norm of the matrix $H^{-1/2}(\widehat{H} - H)H^{-1/2}$ for $\widehat{H} = \sum_j X_j \Pi_{*j}^{-2} A_{j\cdot} A_{j\cdot}^\top$ and $H = \sum_j \Pi_{*j}^{-1} A_{j\cdot} A_{j\cdot}^\top$. This will lead to an improved quadratic inequality
$$\begin{aligned}(\widehat{T}_{\mathrm{mle}} - T_*)^\top H (\widehat{T}_{\mathrm{mle}} - T_*) &\le (1+c)(\widehat{T}_{\mathrm{mle}} - T_*)^\top E \\ &\le (1+c)\|H^{1/2}(\widehat{T}_{\mathrm{mle}} - T_*)\|_2 \|H^{-1/2}E\|_2.\end{aligned}$$
Finally, a sharp concentration inequality for $\|H^{-1/2}E\|_2$ gives the desired faster rates on $\|\widehat{T}_{\mathrm{mle}} - T_*\|_1$.

Minimax optimality and adaptation to sparsity of the MLE of topic distributions, for known A. In Section 2.1.3, we show that the MLE of T_* can be sparse, without any need for extra regularization, a remarkable property that holds in the topic model set-up. Specifically, we introduce in Theorem 5 a new incoherence condition on the matrix A under which $\{\mathrm{supp}(\widehat{T}_{\mathrm{mle}}) \subseteq \mathrm{supp}(T_*)\}$ holds with high probability. Therefore, if the vector T_* is sparse, its zero components will be among those of $\widehat{T}_{\mathrm{mle}}$. Our analysis uses a primal-dual witness approach based on the KKT conditions from solving the MLE. To the best of our knowledge, this is the first work proving that the MLE of sparse mixture weights can be exactly sparse, without extra regularization, and determining conditions under which this can happen. Since $\mathrm{supp}(\widehat{T}_{\mathrm{mle}}) \subseteq \mathrm{supp}(T_*)$ implies that if $T_{*k} = 0$ for some k, so is $[\widehat{T}_{\mathrm{mle}}]_k$, this sparsity recovery property further leads to a faster $\sqrt{s/N_i}$ rate (up to a logarithmic factor) for $\|\widehat{T}_{\mathrm{mle}} - T_*\|_1$ with $s = |S_*|$, as summarized in Corollaries 4 and 6 of Section 2.1.3. In Section 2.1.4, we prove that $\sqrt{s/N_i}$ in fact is the minimax rate of estimating T_* over a large class of sparse topic distributions, implying the minimax optimality of the MLE as well as its adaptivity to the unknown sparsity s.

Finite sample analysis of the estimators of topic distributions, for unknown A. We study the estimation of T_* when A is unknown in Section 2.2. Our procedure of estimating T_* is valid for any estimator $\widehat{A}$ of A with columns of $\widehat{A}$ belonging to Δ_p. For any such estimator $\widehat{A}$, we propose to plug it into the log-likelihood criterion $\sum_j X_j \log(\widehat{A}_{j\cdot}^\top T)$ for estimating T_*. While the proofs are more technical, we can prove that the resulting estimate $\widehat{T}$ of T_* by using $\widehat{A}$ retains all the properties proved for the MLE $\widehat{T}_{\mathrm{mle}}$ based on the known A in Section 2.1, provided that the error $\|\widehat{A} - A\|_{1,\infty} := \max_k \|\widehat{A}_{\cdot k} - A_{\cdot k}\|_1$ is sufficiently small. In fact, all bounds of $\|\widehat{T} - T_*\|_1$ in Theorems 8 and 9 and Corollary 11 of Section 2.2.2, have an extra additive term $\|\widehat{A} - A\|_{1,\infty}$ reflecting the effect of estimating A. In Theorem 10 of Section 2.2, we also show that the estimator $\widehat{T}$ retains the sparsity recovery property despite using $\widehat{A}$. Essentially, our take-home message is that the rate for $\|\widehat{T} - T_*\|_1$ is the same as $\|\widehat{T}_{\mathrm{mle}} - T_*\|_1$ plus the additive error $\|\widehat{A} - A\|_{1,\infty}$, provided that $\widehat{A}$ estimates A well in $\|\cdot\|_{1,\infty}$ norm, with one instance given by the estimator in [9] and fully analyzed in Section 2.2.3.

Finite sample analysis of the estimators of word distributions. In Section 2.3, we compare the mixture-model-based estimator $\widetilde{\Pi}_A = A\widehat{T}_{\mathrm{mle}}$ of Π_* with the empirical estimator $\widehat{\Pi} = X$ (we drop the document-index i), which is simply the p-dimensional observed word frequencies, in two aspects: the ℓ_1 convergence rate and the estimation of probabilities corresponding to zero observed frequencies. For the empirical estimator $\widehat{\Pi}$, we find $\mathbb{E}[\|\widehat{\Pi} - \Pi_*\|_1] \le \sqrt{\|\Pi_*\|_0/N}$ with $\|\Pi_*\|_0 = \sum_j 1\{\Pi_{*j} > 0\}$, while $\mathbb{E}[\|\widetilde{\Pi}_A - \Pi_*\|_1] \le \mathbb{E}[\|\widehat{T}_{\mathrm{mle}} - T_*\|_1] = \mathcal{O}(\sqrt{K\log(K)/N})$. We thus expect a faster rate for the model-based estimate $\widetilde{\Pi}_A$ whenever $K\log(K) = \mathcal{O}(\|\Pi_*\|_0)$. Regarding the second aspect, we note that we can have zero observed frequency ($X_j = 0$) for some word j that has strictly positive word probability ($\Pi_{*j} > 0$). The probabilities of these words are estimated incorrectly by zeroes by the empirical estimate $\widehat{\Pi}$ whereas the model-based estimator $\widetilde{\Pi}_A$ can produce strictly positive estimates, for instance, under conditions stated in Section 2.3. On the other hand, for the words that have zero probabilities in Π_* (hence zero observed frequencies), the empirical estimate $\widehat{\Pi}$ makes no mistakes in estimating their probabilities while the estimation error of $\widetilde{\Pi}_A$ tends to zero at a rate that is no slower than $\sqrt{K\log(K)/N}$. In the case that $\widehat{T}_{\mathrm{mle}}$ has correct one-sided sparsity recovery, detailed in Section 2.1.3, $\widetilde{\Pi}_A$ also estimates zero probabilities by zeroes.

1.2. *Estimates of the* 1*-Wasserstein document distances in topic models.* In Section 3, we introduce two alternative probabilistic representations of a document $i \in [n]$: via the word generating probability vector, $\Pi_*^{(i)}$, or via the topic generating probability vector $T_*^{(i)}$. We use either the 1-Wasserstein distance (see Section 3 for the definition) between the word distributions, $W_1(\Pi_*^{(i)}, \Pi_*^{(j)}; D^{\mathrm{word}})$, or the 1-Wasserstein distance between the topic distributions, $W_1(T_*^{(i)}, T_*^{(j)}; D^{\mathrm{topic}})$, in order to evaluate the proximity of a pair of documents i and j, for metrics D^{word} and D^{topic} between words and topics, defined in displays (46) and (49)–(50), respectively. In particular, in Section 3.1 we explain in detail that we regard a topic as a distribution on words, given by a column of A and, therefore, distances between topics are distances between discrete distributions in Δ_p, and need to be estimated when A is not known.

In Section 3.2, we propose to estimate the two 1-Wasserstein distances by plug-in estimates $W_1(\widetilde{\Pi}^{(i)}, \widetilde{\Pi}^{(j)}; D^{\mathrm{word}})$ and $W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}; \widehat{D}^{\mathrm{topic}})$, respectively, where $\widetilde{\Pi}^{(i)} = \widehat{A}\widehat{T}^{(i)}$ is the model-based estimator of $\Pi_*^{(i)}$ based on a generic estimator $\widehat{A}$ of A and the estimator $\widehat{T}^{(i)}$ of $T_*^{(i)}$ that uses the same $\widehat{A}$, as studied in Section 2. We prove in Proposition 12 of Section 3.2 that the absolute values of the errors of both estimates can be bounded by

$$\max_{\ell\in\{i,j\}} \|\widehat{T}^{(\ell)} - T_*^{(\ell)}\|_1 + \|\widehat{A} - A\|_{1,\infty}.$$

A main theoretical application of the ℓ_1-error bounds for the topic distributions derived in Section 2 can be used to bound the first term while the second term reflects the order of the error in estimating A and, therefore, vanishes if A is known. For completeness, we take the estimator $\widehat{A}$ proposed in [9] and provide in Corollary 13 of Section 3.2 explicit rates of convergence of both errors of estimating two 1-Wasserstein distances by using this $\widehat{A}$. The practical implications of the corollary are that a short document length (small N) can be compensated for, in terms of speed of convergence, by having a relatively small number of topics K covered by the entire corpus, whereas working with a very large dictionary (large p) will not be detrimental to the rate in a very large corpus (large n).

To the best of our knowledge, this rate analysis of the estimates of 1-Wasserstein distance corresponding to estimators of discrete distributions in topic models is new. The only related results, discussed in Section 3.1, have been established relative to empirical frequency estimators of discrete distributions, from an asymptotic perspective [35, 36] or in finite samples [37].

In Remark 8 of Section 3.2, we discuss the net computational benefits of representing documents in terms of their K-dimensional topic distributions, for 1-Wasserstein distance calculations. Using an IMBD movie review corpus as a real data example, we illustrate in Appendix B the practical benefits of these distance estimates, relative to the more commonly used earth(word)-mover's distance [27] between observed empirical word-frequencies, $W_1(\widehat{\Pi}^{(i)}, \widehat{\Pi}^{(j)}; D^{\text{word}})$, with $\widehat{\Pi}^{(i)} := X^{(i)}$, for all $i \in [n]$. Our analysis reveals that all our proposed 1-Wasserstein distance estimates successfully capture differences in the relative weighting of topics between documents, whereas the standard $W_1(\widehat{\Pi}^{(i)}, \widehat{\Pi}^{(j)}; D^{\text{word}})$ is substantially less successful, likely owing in part to the fact noted in Section 1.1 above, that when the dictionary size p is large, but the document length N_i is relatively small, the quality of $\widehat{\Pi}^{(i)}$ as an estimator of $\Pi_*^{(i)}$ will deteriorate, and the quality of $W_1(\widehat{\Pi}^{(i)}, \widehat{\Pi}^{(j)}; D^{\text{word}})$ as an estimator of (47) will deteriorate accordingly.

The remainder of the paper is organized as follows. In Section 2.1, we study the estimation of T_* when A is known. A general bound of $\|\widehat{T}_{\text{mle}} - T_*\|_1$ is stated in Section 2.1.1 and is improved in Section 2.1.2. The sparsity of the MLE is discussed in Section 2.1.3 and the minimax lower bounds of estimating T_* are established in Section 2.1.4. Estimation of T_* when A is unknown is studied in Section 2.2. In Section 2.3, we discuss the comparison between model-based estimators and the empirical estimator of Π_*. Section 3 is devoted to our main application: the 1-Wasserstein distance between documents. In Section 3.1, we introduce alternative Wasserstein distances between probabilistic representations of documents with their estimation studied and analyzed in Section 3.2. The Appendix contains the analysis of a real data set of IMDB movie reviews, all proofs, auxiliary results and all simulation results.

Notation. For any positive integer d, we write $[d] := \{1, \dots, d\}$. For two real numbers a and b, we write $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For any set S, its cardinality is written as $|S|$. For any vector $v \in \mathbb{R}^d$, we write its ℓ_q-norm as $\|v\|_q$ for $0 \le q \le \infty$. For a subset $S \subset [d]$, we define v_S as the subvector of v with corresponding indices in S. Let $M \in \mathbb{R}^{d_1 \times d_2}$ be any matrix. For any set $S_1 \subseteq [d_1]$ and $S_2 \subseteq [d_2]$, we use $M_{S_1 S_2}$ to denote the submatrix of M with corresponding rows S_1 and columns S_2. In particular, $M_{S_1 \cdot}$ ($M_{\cdot S_2}$) stands for the whole rows (columns) of M in S_1 (S_2). Sometimes we also write $M_{S_1} = M_{S_1 \cdot}$ for succinctness. We use $\|M\|_{\text{op}}$ and $\|M\|_q$ to denote the operator norm and elementwise ℓ_q norm, respectively. We write $\|M\|_{1,\infty} = \max_j \|M_{\cdot j}\|_1$. The kth canonical unit vector in $\mathbb{R}^d$ is denoted by $\boldsymbol{e}_k$ while $\mathbf{1}_d$ represents the d-dimensional vector of all ones. $\boldsymbol{I}_d$ is short for the $d \times d$ identity matrix. For two sequences a_n and b_n, we write $a_n \lesssim b_n$ if there exists $C > 0$ such that $a_n \le C b_n$ for all $n \ge 1$. For a metric D on a finite set $\mathcal{X}$, we use boldface $\boldsymbol{D} := (D(a, b))_{a,b \in \mathcal{X}}$ to denote the corresponding $|\mathcal{X}| \times |\mathcal{X}|$ matrix. The set $\mathcal{H}_d$ contains all $d \times d$ permutation matrices.

2. Estimation of topic distributions under topic models. We consider the estimation of the topic distribution vector, $T_*^{(i)} \in \Delta_K$, for each $i \in [n]$. Pick any $i \in [n]$; for notational simplicity, we write $T_* = T_*^{(i)}$, $X = X^{(i)}$ and $\Pi_* = \Pi_*^{(i)}$ as well as $N = N_i$ throughout this section.

We allow, but do not assume, that the vector T_* is sparse, as sparsity is expected in topic models: a document will cover some, but most likely not all, topics under consideration. We therefore introduce the following parameter space for T_*:

$$\mathcal{T}(s) = \{T \in \Delta_K : |\text{supp}(T)| = s\},$$

with s being any integer between 1 and K. From now on, we let $S_* := \text{supp}(T_*)$ and write $|S_*|$ for its cardinality.

In Section 2.1, we study the estimation of T_* from the observed data X, generated from background probability vector Π_* parametrized as $\Pi_* = AT_*$, with known matrix A. The intrinsic difficulties associated with the optimal estimation of T_* are already visible when A is known, and we treat this in detail before providing, in Section 2.2, a full analysis that includes the estimation of A. We remark that assuming A known is not purely unrealistic in topic models used for text data, since then one typically has access to a large corpus (with n in the order of tens of thousands). When the corpus can be assumed to share the same A, this matrix can be very accurately estimated.

The results of Section 2.1 hold for any known A, not required to have any specific structure: in particular, we do not assume that it follows a topic model with anchor words (Assumption 1 stated in Section 2.2.1 below). We will make this assumption when we consider optimal estimation of T_* when A itself is unknown, in which case Assumption 1 serves as both a needed identifiability condition and a condition under which estimation of both A and T_*, in polynomial time, becomes possible. This is covered in detail in Section 2.2.

2.1. *Estimation of T_* when A is known.* When A is known and given, with columns $A_{\cdot k} \in \Delta_p$, the data has a multinomial distribution,

$$NX \sim \text{Multinomial}_p(N; AT_*), \tag{2}$$

where $T_* \in \Delta_K$ is the topic distribution vector, with entries corresponding to the proportions of the K topics, respectively. Under (2), it is natural to consider the Maximum Likelihood Estimator (MLE) $\widehat{T}_{\text{mle}}$ of T_*. The log likelihood, ignoring terms independent of T, is proportional to

$$\sum_{j=1}^{p} X_j \log(A_{j\cdot}^\top T) = \sum_{j \in J} X_j \log(A_{j\cdot}^\top T),$$

where the last summation is taken over the index set of observed relative frequencies,

$$J := \{j \in [p] : X_j > 0\}, \tag{3}$$

and using the convention that $0^0 = 1$. Then

$$\widehat{T}_{\text{mle}} := \operatorname*{arg\,max}_{T \in \Delta_K} \sum_{j \in J} X_j \log(A_{j\cdot}^\top T). \tag{4}$$

This optimization problem is also known as the log-optimal investment strategy; see, for instance [16], problem 4.60. It can be computed efficiently, since the loss function in (4) is concave on its domain, the open half-space $\bigcap_{j \in J}\{x \in \mathbb{R}^K \mid A_{j\cdot}^\top x > 0\}$, and the constraints $T \succeq 0$ and $\mathbf{1}_K^\top T = 1$ are convex.

The following two subsections state the theoretical properties of the MLE in (4), and include a study of its adaptivity to the potential sparsity of T_* and minimax optimality. In Section 2.3, we show that although $\widehat{T}_{\text{mle}}$ is constructed only from observed, nonzero, frequencies, $A_{j\cdot}^\top \widehat{T}_{\text{mle}}$ can be a nonzero estimate of Π_{*j} for those indices $j \in J^c$ for which we observe $X_j = 0$.

2.1.1. *A general finite sample bound for $\|\widehat{T}_{\text{mle}} - T_*\|_1$.* To analyze $\widehat{T}_{\text{mle}}$, we first introduce two deterministic sets that control J defined in (3). Recalling $\Pi_* = AT_*$, we collect the words with nonzero probabilities in the set

$$\overline{J} := \{j \in [p] : \Pi_{*j} > 0\}. \tag{5}$$

We will also consider the set

$$\underline{J} := \{j \in [p] : \Pi_{*j} > 2\varepsilon_j\}, \tag{6}$$

where

$$\varepsilon_j := 2\sqrt{\frac{\Pi_{*j}\log p}{N}} + \frac{4\log p}{3N}, \quad \forall 1 \le j \le p. \tag{7}$$

The sets $\overline{J}$ and $\underline{J}$ are appropriately defined such that $\underline{J} \subseteq J \subseteq \overline{J}$ holds with probability at least $1 - 2p^{-1}$ (see Lemma I.1 of Appendix I). Define

$$\rho := \max_{j \in \overline{J}} \frac{\|A_{j\cdot}\|_\infty}{\Pi_{*j}}. \tag{8}$$

We note that $\overline{J}$, $\underline{J}$ and ρ all depend on T_* implicitly via Π_*. Another important quantity is the following $\ell_1 \to \ell_1$ restricted condition number of the submatrix $A_{\underline{J}}$ of A, defined as

$$\kappa(A_{\underline{J}}, s) := \min_{S \subseteq [K]: |S| \le s} \min_{v \in \mathcal{C}(S)} \frac{\|A_{\underline{J}} v\|_1}{\|v\|_1}, \tag{9}$$

with

$$\mathcal{C}(S) := \{v \in \mathbb{R}^K \setminus \{\mathbf{0}\} : \|v_S\|_1 \ge \|v_{S^c}\|_1\}.$$

We make the following simple, but very important, observation that

$$\widehat{T}_{\mathrm{mle}} - T_* \in \mathcal{C}(S_*) \tag{10}$$

with $S_* = \mathrm{supp}(T_*)$, by using the fact that both $\widehat{T}_{\mathrm{mle}}$ and T_* belong to Δ_K. In fact, (10) holds generally for any estimator $\widehat{T} \in \Delta_K$ as

$$0 = \|T_*\|_1 - \|\widehat{T}\|_1 = \|(T_*)_{S_*}\|_1 - \|\widehat{T}_{S_*}\|_1 - \|\widehat{T}_{S_*^c}\|_1 \le \|(\widehat{T} - T_*)_{S_*}\|_1 - \|(\widehat{T} - T_*)_{S_*^c}\|_1.$$

Display (10) implies that the "effective" ℓ_1 error bound of $\widehat{T}_{\mathrm{mle}} - T_*$ arises mainly from the estimation of $(T_*)_{S_*}$. Also because of this property, we need the condition number of A to be positive only over the cone $\mathcal{C}(S_*)$ rather than the whole $\mathbb{R}^K$.

The following theorem states the convergence rate of $\|\widehat{T}_{\mathrm{mle}} - T_*\|_1$. Its proof can be found in Appendix F.1.

THEOREM 1. *Assume $\kappa(A_{\underline{J}}, s) > 0$. For any $\epsilon \ge 0$, with probability $1 - 2p^{-1} - 2\epsilon$, one has*

$$\|\widehat{T}_{\mathrm{mle}} - T_*\|_1 \le \frac{2}{\kappa^2(A_{\underline{J}}, s)} \left\{ \sqrt{\frac{2\rho\log(K/\epsilon)}{N}} + \frac{2\rho\log(K/\epsilon)}{N} \right\}. \tag{11}$$

Theorem 1 is a general result that only requires $\kappa(A_{\underline{J}}, s) > 0$. The rates depend on two important quantities: $\kappa(A_{\underline{J}}, s)$ and ρ, which we discuss below in detail. In the next section, we will show that the bound in Theorem 1 serves as an initial result, upon which one could obtain a faster rate of the MLE in certain regimes.

REMARK 1 (Discussion on $\kappa(A_{\underline{J}}, s)$). The $\ell_1 \to \ell_1$ condition number, $\kappa(A, K)$, is commonly used to quantify the linear independence of the columns belonging to Δ_p of the matrix $A \in \mathbb{R}_+^{p \times K}$ [25]. As remarked in [25], the $\ell_1 \to \ell_1$ condition number $\kappa(A, K)$ plays the role of the smallest singular value, $\sigma_K(A) = \inf_{v \ne 0} \|Av\|_2 / \|v\|_2$, but it is more appropriate for matrices with columns belonging to a probability simplex. Because of the chain inequalities

$$\frac{\kappa(A, K)}{\sqrt{p}} \le \sigma_K(A) \le \sqrt{K}\kappa(A, K),$$

and the fact that $K \ll p$, having $\kappa^{-1}(A, K)$ appear in the bound loses at most a $\sqrt{K}$ factor comparing to $\sigma_K^{-1}(A)$. But using $\sigma_K^{-1}(A)$ potentially yields a much worse bound than using $\kappa^{-1}(A, K)$: there are instances for which $\kappa(A, K)$ is lower bounded by a constant whereas $\sigma_K(A)$ is only of order $o(1)$ (see, for instance, [25], Appendix A).

The restricted $\ell_1 \to \ell_1$ condition number $\kappa(A, s)$ in (9) for $1 \le s \le K$ generalizes $\kappa(A, K)$ by requiring the condition of A over the cones $\mathcal{C}(S)$ with $S \subseteq [K]$ and $|S| \le s$. We thus view $\kappa(A, s)$ as the analogue of the restricted eigenvalue [7] of the Gram matrix in the sparse regression settings. In topic models, it has been empirically observed that the (restricted) condition number of A is oftentimes bounded from below by some absolute constant [4].

To understand why $\kappa(A_{\underline{J}}, s)$ appears in the rates, recall that the MLE in (4) only uses the words in J as defined in (3). Intuitively, only the condition number of A_J should play a role as we do not observe any information from words in $J^c := [p] \setminus J$. Since $\underline{J} \subseteq J \subseteq \overline{J}$ holds with high probability, we can thus bound $\kappa(A_J, s)$ from below by $\kappa(A_{\underline{J}}, s)$. For the same reason, ρ in (8) is defined over $j \in \overline{J}$ rather than $j \in J$.

REMARK 2 (Discussion on ρ). Define the smallest nonzero entry in T_* as

$$T_{\min} := \min_{k \in S_*} T_{*k}.$$

Recall $\Pi_{*j} = A_{j\cdot}^\top T_* = A_{jS_*}^\top T_{*S_*}$. We have $\rho = \max\{\rho_{S_*}, \rho_{S_*^c}\}$ where

$$\rho_{S_*} = \max_{k \in S_*} \max_{j \in \overline{J}} \frac{A_{jk}}{\sum_{a \in S_*} A_{ja} T_{*a}} \le \frac{1}{T_{\min}}, \tag{12}$$

$$\rho_{S_*^c} = \max_{k \in S_*^c} \max_{j \in \overline{J}} \frac{A_{jk}}{\sum_{a \in S_*} A_{ja} T_{*a}} \le \frac{1}{T_{\min}} \cdot \max_{k \in S_*^c} \max_{j \in \overline{J}} \frac{A_{jk}}{\sum_{a \in S_*} A_{ja}}. \tag{13}$$

The magnitudes of both ρ_{S_*} and $\rho_{S_*^c}$ closely depend on $T_{\min}$ while $\rho_{S_*^c}$ also depends on

$$\xi := \max_{j \in \overline{J}} \frac{\|A_{jS_*^c}\|_\infty}{\|A_{jS_*}\|_1}, \tag{14}$$

a quantity that essentially balances the entries of A_{jS_*} and those of $A_{jS_*^c}$. Clearly, when T_* is dense, that is, $|S_*| = K$, we have $\xi = 0$. In general, we have

$$\rho \le (1 \vee \xi)/T_{\min}. \tag{15}$$

We further remark that if A has a special structure such that there exists at least one anchor word for each topic $k \in S_*$, that is, for each $k \in [K]$, there exists a row $A_{j\cdot} \propto \boldsymbol{e}_k$ (see Assumption 1 in Section 2.2.1 below), it is easy to verify that the inequality for ρ_{S_*} in (12) is in fact an equality.

2.1.2. *Faster rates of* $\|\widehat{T}_{\mathrm{mle}} - T_*\|_1$. In this section, we state conditions under which the general bound stated in Theorem 1 can be improved. We begin by noting that one of the main difficulties in deriving a faster rate for $\|\widehat{T}_{\mathrm{mle}} - T_*\|_1$ is in establishing a link between the Hessian matrix (the second-order derivative) of the loss function in (4) evaluated at $\widehat{T}_{\mathrm{mle}}$ to that evaluated at T_*.

To derive this link, we prove in Appendix F.1 that a relative weighted error of estimating T_* by $\widehat{T}_{\mathrm{mle}}$ stays bounded in probability, in the precise sense that

$$\max_{j \in \overline{J}} \frac{|A_{j\cdot}^\top (\widehat{T}_{\mathrm{mle}} - T_*)|}{A_{j\cdot}^\top T_*} = \mathcal{O}_{\mathbb{P}}(1). \tag{16}$$

Further, we show in Lemma I.4 in Appendix I that the Hessian matrix of (4) at T_* concentrates around its population-level counterpart, with X_j replaced by Π_{*j}. A sufficient condition under which (16) holds can be derived as follows. First, note that

$$\max_{j\in\overline{J}} \frac{|A_{j\cdot}^\top(\widehat{T}_{\mathrm{mle}} - T_*)|}{A_{j\cdot}^\top T_*} \le \max_{j\in\overline{J}} \frac{\|A_{j\cdot}\|_\infty}{\Pi_{*j}} \|\widehat{T}_{\mathrm{mle}} - T_*\|_1 \overset{(8)}{=} \rho \|\widehat{T}_{\mathrm{mle}} - T_*\|_1. \tag{17}$$

We have bounded ρ by $\rho \le (1 \vee \xi)/T_{\min}$ in (15), and have provided an initial bound on $\|\widehat{T}_{\mathrm{mle}} - T_*\|_1$ in Theorem 1. Therefore, (16) holds if these two bounds combine to show $\rho\|\widehat{T}_{\mathrm{mle}} - T_*\|_1$ is of order $\mathcal{O}_{\mathbb{P}}(1)$. This is summarized in the following theorem. Let $\kappa(A_{\overline{J}}, K)$ be defined in (9) with $s = K$ and $A_{\overline{J}}$ in place of $A_{\underline{J}}$. Recall that ξ is defined in (14). In addition, we define

$$\begin{aligned} M_1 &:= \frac{\log K}{\kappa^4(A_{\underline{J}}, s) T_{\min}^3}(1 \vee \xi^3), \\ M_2 &:= \frac{\log K}{\kappa^2(A_{\overline{J}}, K) T_{\min}^2}(1 \vee \xi)(1 + \xi\sqrt{K - s}). \end{aligned} \tag{18}$$

THEOREM 2. *For any $T_* \in \mathcal{T}(s)$ with $1 \le s \le K$, assume there exists some sufficiently large constant $C > 0$ such that*

$$N \ge C \max\{M_1, M_2\}. \tag{19}$$

Then, with probability $1 - 2p^{-1} - 4K^{-1} - 2e^{-K}$, we have

$$\|\widehat{T}_{\mathrm{mle}} - T_*\|_1 \lesssim \kappa^{-1}(A_{\overline{J}}, s)\sqrt{K/N}.$$

Condition (19) requires the sample size N to be sufficiently large relative to $T_{\min}$, ξ and the $\ell_1 \to \ell_1$ condition number of A. If $N \ge CM_1$, then the argument in (17) above implies (16), while we use $N \ge CM_2$ to prove in Appendix I that the Hessian matrix of (4) at T_* concentrates around its population-level counterpart.

Combining the bounds in Theorem 1 and Theorem 2, we immediately have the following faster rate of the MLE under (19):

$$\|\widehat{T}_{\mathrm{mle}} - T_*\|_1 = \mathcal{O}_{\mathbb{P}}\left(\min\left\{\kappa^{-2}(A_{\underline{J}}, s)\sqrt{\frac{\rho \log K}{N}}, \kappa^{-1}(A_{\overline{J}}, s)\sqrt{\frac{K}{N}}\right\}\right). \tag{20}$$

We remark that, when T_* is sparse, the first term in the minimum on the right of (20) could be smaller than the second one (see one instance under item *(a)* of Corollary 6).

However, for dense $T_* \in \mathcal{T}(K)$ such that $|S_*| = K$, the newly derived rate in Theorem 2 (the second term in (20)) is always faster than that in Theorem 1 (the first term in (20)), as summarized in the following corollary. Its proof follows immediately from Theorem 2 by replacing s by K and noting that in that case $\xi = 0$, by (14).

COROLLARY 3 (Dense T_*). *For any $T_* \in \mathcal{T}(K)$, assume there exists some sufficiently large constant $C > 0$ such that*

$$N \ge C \frac{\log K}{\kappa^4(A_{\underline{J}}, K) T_{\min}^3}. \tag{21}$$

Then we have

$$\mathbb{P}\left\{\|\widehat{T}_{\mathrm{mle}} - T_*\|_1 \lesssim \kappa^{-1}(A_{\overline{J}}, K)\sqrt{\frac{K}{N}}\right\} \ge 1 - 2p^{-1} - 4K^{-1} - 2e^{-K}.$$

Although in our current application we expect T_* to be exactly sparse, there are many other applications where T_* can only be approximately sparse. For instance, in a standard latent Dirichlet allocation model [15], T_* follows a Dirichlet distribution and is never exactly sparse. The theoretical results derived above are directly applicable to these situations.

Theorem 2 and Corollary 3 allow us to pin-point the difficulty in establishing rate adaptation to sparsity of the $\widehat{T}_{\mathrm{mle}}$ of a potentially sparse T_*, when its sparsity pattern is neither known, nor recovered. To this end, notice that although the bound in Theorem 2 is derived for sparse T_*, the rate is essentially the same as that of Corollary 3, that pertains to a dense T_* and, moreover, is established under the stronger condition (19). This condition involves the quantity ξ defined in (14), which balances entries of $A_{\overline{J}S_*^c}$ and $A_{\overline{J}S_*}$. We thus view (19) as the price to pay, compared to (21), for not knowing the support S_* of T_*. We recall that all prior existing literature on this problem, either classical [1, 12] or more recent [4] assumes that S_* is known.

The next section establishes the remarkable fact that the MLE of T_* in topic models can be exactly sparse, under conditions that we establish in this section. This property allows us to relax (19) and prove that the rate of $\|\widehat{T}_{\mathrm{mle}} - T_*\|_1$ can adapt to the unknown sparsity of T_*, when the support of $\widehat{T}_{\mathrm{mle}}$ is included in the support of T_*.

2.1.3. *The sparsity of the MLE in topic models.* We will shortly investigate and discuss conditions under which $\widehat{T}_{\mathrm{mle}}$ in topic models is sparse, a remarkable feature of the MLE since there is no explicit regularization in (4). To that end, we will show that

$$\mathcal{E}_{\mathrm{supp}} := \{\mathrm{supp}(\widehat{T}_{\mathrm{mle}}) \subseteq \mathrm{supp}(T_*)\} \tag{22}$$

holds with high probability. Therefore, when T_* has zero entries, $\widehat{T}_{\mathrm{mle}}$ will also be sparse, and have at least as many zeroes. Before stating these results more formally, we give a first implication, in Corollary 4, of the sparsity of the MLE on its ℓ_1-norm rate.

COROLLARY 4. *For any $T_* \in \mathcal{T}(s)$ with $1 \le s < K$, assume there exists some sufficiently large constant $C > 0$ such that*

$$N \ge C \frac{\log(s \vee n)}{\kappa^4(A_{\underline{J}}, s) T_{\min}^3}. \tag{23}$$

Then, for any $\epsilon \ge 0$,

$$\mathbb{P}\left[\mathcal{E}_{\mathrm{supp}} \cap \left\{\|\widehat{T}_{\mathrm{mle}} - T_*\|_1 \lesssim \kappa^{-1}(A_{\overline{J}}, s)\sqrt{\frac{s\log(1/\epsilon)}{N}}\right\}\right] \ge 1 - \frac{2}{p} - \frac{4}{s \vee n} - 2\epsilon^s.$$

To compare the rates with Theorem 1, suppose Assumption 1 in Section 2.2.1 holds and we have $\rho \ge \rho_{S_*} = 1/T_{\min}$ from Remark 2. Since $T_{\min} \le 1/s$ and $1 \ge \kappa(A_{\overline{J}}, s) \ge \kappa(A_{\underline{J}}, s)$, we conclude that the rate in Corollary 4 is no slower than that in Theorem 1.

Compared to Theorem 2 and condition (19), on the even $\mathcal{E}_{\mathrm{supp}}$, the faster rate in Corollary 4 is obtained under a weaker condition (23). This reflects the benefit of (one-sided) support recovery, $\mathrm{supp}(\widehat{T}_{\mathrm{mle}}) \subseteq \mathrm{supp}(T_*)$.

In the following theorem, we show that $\mathcal{E}_{\mathrm{supp}}$ indeed holds with high probability under an incoherence condition on A.

THEOREM 5. *For any $T_* \in \mathcal{T}(s)$ with any $1 \le s < K$, assume (23). Further assume there exists some sufficiently small constant $c > 0$ such that*

$$\left(\kappa^{-1}(A_{\underline{J}}, s)\sqrt{\frac{\xi s}{T_{\min}}} + 1\right)\sqrt{\frac{\xi \log(K)}{T_{\min} N}} \le c \min_{k \in S_*^c} \sum_{j \in \overline{J}^c} A_{jk}. \tag{24}$$

Then one has $\mathbb{P}(\mathcal{E}_{\mathrm{supp}}) \ge 1 - 2p^{-1} - 4(s \vee n)^{-1} - 4K^{-1}$.

SKETCH OF THE PROOF. We defer the detailed proof to Appendix F.4, but offer a sketch here. For any T_* with $\operatorname{supp}(T_*) \subseteq [K]$, our proof of $\operatorname{supp}(\widehat{T}_{\mathrm{mle}}) \subseteq \operatorname{supp}(T_*)$ consists in two steps. We show that:

(i) there exists an optimal solution $\widetilde{T}$ to (4) such that $\operatorname{supp}(\widetilde{T}) \subseteq \operatorname{supp}(T_*)$;

(ii) if there exists any other optimal solution $\bar{T}$ to (4) that is different from $\widetilde{T}$, we also have $\operatorname{supp}(\bar{T}) \subseteq \operatorname{supp}(T_*)$.

Since $\widehat{T}_{\mathrm{mle}}$ itself is an optimal solution to (4), combining (i) and (ii) yields the desired result.

To prove (i), we use the primal-dual witness approach based on the KKT condition of (4). Specifically, we construct the (oracle) optimal solution $\widetilde{T}$ as

$$\widetilde{T}_{S_*} = \operatorname*{arg\,max}_{\beta \in \Delta_s} N \sum_{j \in J} X_j \log\bigl(A_{jS_*}^{\top}\beta\bigr), \qquad \widetilde{T}_{S_*^c} = \mathbf{0}. \tag{25}$$

Here, $S_* = \operatorname{supp}(T_*)$ and $s = |S_*|$. On the random event,

$$\max_{k \in S_*^c} \sum_{j \in J} X_j \frac{A_{jk}}{A_{jS_*}^{\top}\widetilde{T}_{S_*}} < 1, \tag{26}$$

we prove step (i) by showing that $\widetilde{T}$ is an optimal solution to (4) via its KKT condition. Also on the event (26), we prove step (ii) by using the concavity of the loss function in (4) together with some intermediate results from proving (i). Finally, we show that the random event (22) holds with the specified probability in Theorem 5 under condition (24). □

For completeness, in Appendix A, we show that for a certain class of topic models $\widehat{T}_{\mathrm{mle}}$ is not only sparse, but can also consistently estimate the zero entries in T_*. Other examples are possible, but we restrict our attention to topic models (1) with anchor words, satisfying Assumption 1 stated in Section 2.2.1, for which we show that we also have $\operatorname{supp}(T_*) \subseteq \operatorname{supp}(\widehat{T}_{\mathrm{mle}})$ with high probability. Combination with Theorem 5 proves consistent support recovery of $\widehat{T}_{\mathrm{mle}}$, in this class of topic models, a fact also confirmed by our simulations in Appendix D.

EXAMPLE 1. We argued above that, when A has a certain configuration, if T_* has zero entries, so will $\widehat{T}_{\mathrm{mle}}$. We provide below a simple but illuminating example of this fact. Assume that all words are anchor words: each topic uses its own dedicated words, and there is no overlap between words per topic. We collect the respective word indices, per topic, in the set $\{I_1, \ldots, I_K\}$ which forms a partition of $[p]$. In this case, the columns of A have disjoint supports, and by inspecting the displays (F.12)–(F.13) in the proof of Theorem 1, one can deduce that $\widehat{T}_{\mathrm{mle}}$ has the following closed-form expression:

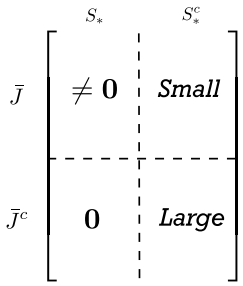
$$[\widehat{T}_{\mathrm{mle}}]_k = \sum_{i \in I_k} X_i, \quad \forall k \in [K].$$

Indeed, the above expression can be understood by noting that $Z \sim \mathrm{Multinomial}_K(N; T_*)$ where $Z_k = N \sum_{i \in I_k} X_i$ for each $k \in [K]$. Therefore, when $T_{*k} = 0$ for some $k \in [K]$, we immediately have $Z_k = 0$, with probability one. Thus, $[\widehat{T}_{\mathrm{mle}}]_k = 0$, and $\operatorname{supp}(\widehat{T}_{\mathrm{mle}}) \subseteq \operatorname{supp}(T_*)$. For a more general A, the phenomenon $\operatorname{supp}(\widehat{T}_{\mathrm{mle}}) \subseteq \operatorname{supp}(T_*)$ still remains under the incoherence condition (24) that we explain in detail in the following remark.

REMARK 3. Condition (24) can be interpreted as an incoherence condition on the submatrices $A_{\cdot S_*}$ and $A_{\cdot S_*^c}$. To see this, recall from Remark 2 that ξ controls the largest ratio of $\|A_{jS_*^c}\|_\infty$ to $\|A_{jS_*}\|_1$ over all $j \in \overline{J}$. Since

$$\overline{J} = \bigl\{j \in [p] : A_{jS_*} \neq \mathbf{0}\bigr\} \quad \text{and} \quad \overline{J}^c = \bigl\{j \in [p] : A_{jS_*} = \mathbf{0}\bigr\},$$

the left-hand side of (24) controls from above the magnitude of the entries $A_{jS_*^c}$ for the rows with $A_{jS_*} \neq \mathbf{0}$, whereas the right-hand side bounds from below $A_{\cdot S_*^c}$ on the rows with $A_{jS_*} = \mathbf{0}$. To aid intuition, the following figure illustrates the restriction on A where the submatrix $A_{\bar{J}S_*^c}$ is required to have relatively small entries, while the submatrix $A_{\bar{J}^c S_*^c}$ needs to have relatively large entries. Generally speaking, the more incoherent $A_{\cdot S_*}$ and $A_{\cdot S_*^c}$ are, the more likely condition (24) holds.



In particular, condition (24) always holds if $A_{\cdot S_*}$ and $A_{\cdot S_*^c}$ have disjoint supports. Another favorable situation for (24) is when there exist anchor words in the dictionary (see Assumption 1 in Section 2.2.1). Specifically, when there exist at least m anchor words for each of the topics indexed by S_*^c, and their nonzero entries in the corresponding rows of A are lower bounded by $\delta \in (0, 1/m]$ (recall that columns of A sum up to one), the right-hand side of (24) is no smaller than $c(m\delta)$.

To conclude our discussion of the fast rates of the MLE, we remark that the rate in Theorem 1 per se could be as fast as $\sqrt{s \log K/N}$ under additional conditions and if we restrict ourselves to the following subspace of $\mathcal{T}(s)$:

$$\mathcal{T}'(s) := \mathcal{T}'(s, c_\star) = \{T \in \mathcal{T}(s) : T_{\min} \geq c_\star/s\}. \tag{27}$$

Here, $c_\star \in (0, 1]$ is some absolute constant. The following corollary summarizes the conditions that we need to simplify the rates in Theorem 1 and combines them with the conditions in Corollary 4 and Theorem 5 to yield a faster rate of the MLE when T_* is sparse.

COROLLARY 6. *For any $T_* \in \mathcal{T}(s)$ with $1 \leq s < K$,*

(a) *if $T_* \in \mathcal{T}'(s)$, $\xi = \mathcal{O}(1)$, $\kappa^{-1}(A_{\underline{J}}, s) = \mathcal{O}(1)$ and $s \log(K) = \mathcal{O}(N)$, then*

$$\|\widehat{T}_{\text{mle}} - T_*\|_1 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{s \log(K)}{N}}\right);$$

(b) *if conditions (23)–(24) and $\kappa^{-1}(A_{\bar{J}}, s) = \mathcal{O}(1)$ hold, then*

$$\|\widehat{T}_{\text{mle}} - T_*\|_1 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{s}{N}}\right).$$

We note that the bound in case (a) from Theorem 1 is slower by a factor $\sqrt{\log(K)}$, which is the price to pay for not recovering the support of T_*. In Section 2.1.4, we benchmark the fast rate $\sqrt{s/N}$ in Corollary 6 and show that it is minimax rate optimal, by establishing the minimax lower bounds of estimating $T_* \in \mathcal{T}(s)$ for any $1 \leq s \leq K$.

REMARK 4 (Comparison with existing work). For known A, and when S_* is also known, [4] analyzes the estimator $\widetilde{T}$ as defined in (25). Note that $\widetilde{T}$ is not the MLE in general and $\|\widetilde{T} - T_*\|_1 = \|(\widetilde{T} - T_*)_{S_*}\|_1$ holds by definition. Under $\kappa^{-1}(A_{\overline{J}}, s) = \mathcal{O}(1)$ and a condition similar to (23), [4], Theorem 5.3, proves $\|(\widetilde{T} - T_*)_{S_*}\|_1 = \mathcal{O}_{\mathbb{P}}(\sqrt{s/N})$ *only* for $T_* \in \mathcal{T}'(s)$. Therefore, the result of [4] is only comparable to ours when T_* is dense with $S_* = [K]$. Even in this case, our result (see, for instance, Corollary 3) is more general in the sense that we do not require $T_{\min} \geq c/K$ to obtain the same rate. More generally, when S_* is unknown, our result in Corollary 6 shows that the MLE can still have fast rates in many scenarios. Moreover, we prove that the MLE is actually sparse and consistently estimate the zero entries of T_* under the incoherence condition (24).

2.1.4. *Minimax lower bounds and the optimality of the MLE.* To benchmark the rate of $\widehat{T}_{\text{mle}}$ in Corollary 6, we now establish the minimax lower bound of estimating T_* over $\mathcal{T}'(s)$ for any $1 < s \leq K$. Notice that such a lower bound is also a minimax lower bound over $T_* \in \mathcal{T}(s)$, a larger parameter space.

The following theorem states the ℓ_1-norm minimax lower bound of estimating T_* in (2), from data X.

THEOREM 7. *Under* (2), *assume* $1 < s \leq cN$ *for some small constant* $c > 0$. *Then there exists some absolute constants* $c_0 > 0$ *and* $c_1 \in (0, 1]$, *depending on* c *only*, *such that*

$$\inf_{\widehat{T}} \sup_{T_* \in \mathcal{T}'(s)} \mathbb{P}\left\{ \|\widehat{T} - T_*\|_1 \geq c_0 \sqrt{\frac{s}{N}} \right\} \geq c_1.$$

The infimum is taken over all estimators $\widehat{T} \in \Delta_K$.

Different from the standard ℓ_1-norm minimax rate, $s/\sqrt{N}$, of estimating an s-dimensional unconstrained vector from N i.i.d. observations, for instance the regression coefficient vector in linear regression, Theorem 7 shows that the ℓ_1-minimax rate of estimating the probability vector $T_* \in \mathcal{T}'(s)$ is of order $\sqrt{s/N}$.

In view of Theorem 7, under the conditions of Corollary 6, the MLE is minimax optimal for $T_* \in \mathcal{T}'(s)$. In fact, Corollary 6 also shows that under conditions therein, the optimal rate can be still achieved by the MLE on a larger space $\mathcal{T}(s)$. Furthermore, the derived rates in the minimax lower bounds in Theorem 7 are sharp.

It is also worth mentioning that in contrast to the sparse linear regression setting where the minimax optimal rates of estimating a p-dimensional vector with at most s nonzero entries contain a $\log(ep/s)$ term, the minimax optimal rates in our context do not contain an additional $\log(eK/s)$ term, an advantage of support recovery of the MLE.

REMARK 5 (Method of moments and least squares estimators). The method of moments is a natural alternative to MLE-based estimation. It would correspond to estimating T_* by the solution $X = AT_*$. Since this solution may not lie in the probability simplex Δ_K, one can consider instead the restricted least squares estimator (RLS) that regresses X onto A over the probability simplex Δ_K. However, this method is not optimal, as it does not take into account the heteroscedasticity of the data X. We confirmed this via our simulation study in Appendix D. An iterative weighted RLS could be used to improve the performance of the RLS. It is well known that in the classical setting with K and p fixed, this technique is asymptotically (as $N \to \infty$) equivalent with the MLE (in fact, both are efficient); see, for instance, [12] and [1]. We confirmed this in our simulation studies in Appendix D, but found that it never improved upon the MLE, and furthermore, had a significantly greater computational time than the MLE.

REMARK 6. Since our target T_* lies in a probability simplex, we view the ℓ_1 norm as a natural metric for quantifying the estimation error. Nevertheless, our analysis readily gives the error bounds of estimating T_* in ℓ_2-norm, as stated in Appendix L.

2.2. *Estimation of T_* when A is unknown.* When A is unknown, we propose to estimate A first. The estimation of A has been well understood in the literature of topic models, as reviewed in Section 2.2.1. Our procedure of estimating T_* for unknown A is valid for any estimator of A, and is stated and analyzed in Section 2.2.2. In Section 2.2.3, we illustrate our general result by applying it to a particular estimator of A.

2.2.1. *Estimation of A.* The estimation of A under topic models has been originally studied within a Bayesian framework [15, 23], and variational-Bayes type approaches were further proposed to accelerate the computation of fully Bayesian approaches. We refer to [14] for an in-depth overview of this class of techniques.

More recently, [2, 3, 5, 9, 10, 20, 24] studied provably fast algorithms for estimating A from a frequentist point of view. The common thread of these works, both theoretically and computationally, is the usage of the following separability condition.

ASSUMPTION 1. For each $k \in [K]$, there exists $j \in [p]$ such that $A_{jk} \neq 0$ and $A_{jk'} = 0$ for all $k' \in [K] \setminus \{k\}$.

Assumption 1 is also known as the anchor word assumption as it translates into assuming the existence of words that are only related to a single topic. It has been empirically shown in [19] that Assumption 1 holds in most large corpora for which the topic models are reasonable modeling tools. Assumption 1, coupled with a mild regularity condition on the topic matrix $\boldsymbol{T}_* \in \mathbb{R}^{K \times n}$, also serves as an identifiablity condition on model (1), in that it can be shown that the matrix A can be uniquely recovered from the expected frequency matrix $\boldsymbol{\Pi}_*$. See, [3, 13] for the case when K is known and more recently, [9], for the case when K is unknown. Since K can be consistently estimated when it is unknown (see, for instance, [9]), in the sequel we focus on estimators of A that have K columns and belong to the space

$$\mathcal{A} = \{A \in \mathbb{R}^{p \times K} : A_{\cdot k} \in \Delta_p, \forall k \in [K]\}. \tag{28}$$

Our results of estimating T_* in Section 2.2.2 below will apply to any estimator $\widehat{A} \in \mathcal{A}$ that is sufficiently close to A in the matrix norms $\|\cdot\|_\infty$ and $\|\cdot\|_{1,\infty}$.

2.2.2. *Estimation of T_*.* Our theory for estimating T_* in this section holds for any estimator $\widehat{A} \in \mathcal{A}$. We therefore state them as such, and offer an example of the theory applied with a particular estimator at the end of this section. Motivated by (4), given any estimate $\widehat{A} \in \mathcal{A}$, we propose to estimate T_* by

$$\widehat{T} = \operatorname*{arg\,max}_{T \in \Delta_K} N \sum_{j \in J} X_j \log(\widehat{A}_{j\cdot}^\top T). \tag{29}$$

Note that, in contrast to $\widehat{T}_{\mathrm{mle}}$ in (4) for known A, the above $\widehat{T}$ depends on $\widehat{A}$ and is not the MLE in general for unknown A.

Since one can only identify and estimate A up to some permutation of columns, the following theorem provides the convergence rate of $\|\widehat{T} - P^\top T_*\|_1$ with $P \in \mathcal{H}_K$ being some $K \times K$ permutation matrix. Its proof can be found in Appendix G.1. Recall that the sets $\underline{J}$ and $\overline{J}$ are defined in (5) and (6), and the quantity ρ is defined in (8).

THEOREM 8. *Suppose the events*

$$\bigcap_{j\in\overline{J}}\left\{\|\widehat{A}_{j\cdot}-(AP)_{j\cdot}\|_\infty\le\frac{1}{2}\Pi_{*j}\right\} \tag{30}$$

and

$$\left\{\|\widehat{A}_{\underline{J}}-(AP)_{\underline{J}}\|_{1,\infty}\le\frac{1}{2}\kappa(A_{\underline{J}},s)\right\} \tag{31}$$

hold with probability $1-\alpha$, *for some permutation matrix* $P\in\mathcal{H}_K$. *Then we have*, *with probability* $1-4p^{-1}-\alpha$,

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

The restrictions (30) and (31) and the last two terms in the bound above reflect both the requirement and the effect of estimating A on the overall ℓ_1-convergence rate of $\widehat{T}$. Note that by using condition (30) the last term in the bound can be simply bounded from above by

$$\frac{7}{\kappa^2(A_{\underline{J}},s)}\frac{|\overline{J}\setminus\underline{J}|\log(p)}{N}.$$

This term originates from words that have very small probability of occurrence, $\Pi_{*j}=\mathcal{O}(\log(p)/N)$, but have nonzero observed frequencies, $X_j>0$. For ease of presentation, we assume in the sequel that the number of such words is bounded, that is, $|\overline{J}\setminus\underline{J}|\le C$ for some finite constant $C>0$. Still, our analysis allows one to track their presence throughout the proof.

To provide intuition of the first requirement (30), suppose $P=\boldsymbol{I}_K$ and note that this event guarantees that, for all $T_*\in\Delta_K$ and for all $j\in\overline{J}$,

$$\widehat{A}_{j\cdot}^\top T_*\in\big[A_{j\cdot}^\top T_*\pm|\widehat{A}_{j\cdot}^\top T_*-A_{j\cdot}^\top T_*|\big]\subseteq\big[\Pi_{*j}\pm\|\widehat{A}_{j\cdot}-A_{j\cdot}\|_\infty\big]\subseteq\left[\frac{1}{2}\Pi_{*j},\frac{3}{2}\Pi_{*j}\right] \tag{32}$$

so that $\widehat{A}_{j\cdot}^\top T_*$ and Π_{*j} are the same up to a constant factor. In particular, $\Pi_{*j}>0$ implies $\widehat{A}_{j\cdot}^\top T_*>0$, ensuring that T_* lies in the domain of the log-likelihood function $N\sum_{j\in J}X_j\log(\widehat{A}_j^\top T)$.

The second restriction (31) allows us to replace the $\ell_1\to\ell_1$ condition number of the random matrix $\widehat{A}_{\underline{J}}$ by that of $A_{\underline{J}}$. Since

$$\begin{aligned}\kappa(\widehat{A}_{\underline{J}},s)&=\min_{S\subseteq[K]:|S|\le s}\min_{v\in\mathcal{C}(S)}\frac{\|\widehat{A}_{\underline{J}}v\|_1}{\|v\|_1}\\&\ge\kappa(A_{\underline{J}},s)-\max_{S\subseteq[K]:|S|\le s}\max_{v\in\mathcal{C}(S)}\frac{\|(\widehat{A}_{\underline{J}}-A_{\underline{J}})v\|_1}{\|v\|_1}\\&\ge\kappa(A_{\underline{J}},s)-\|\widehat{A}_{\underline{J}}-A_{\underline{J}}\|_{1,\infty},\end{aligned} \tag{33}$$

the bound in (31) immediately yields

$$\kappa(\widehat{A}_{\underline{J}},s)\ge\frac{1}{2}\kappa(A_{\underline{J}},s). \tag{34}$$

Similar to the case of A known, treated in Section 2.1.2, when $\widehat{T}$ lies in the vicinity of T_* in the sense of (16), the rate of $\|\widehat{T} - P^\top T_*\|_1$ can be improved. The following result is an analogue of Theorem 2 for unknown A. Its proof can be found in Appendix G.2. Recall that M_1 and M_2 are defined in (18).

THEOREM 9. *Assume there exists a sufficiently large constant $C > 0$ such that*

$$N \geq C \frac{\log(p)}{\log(K)} \max\{M_1, M_2\}. \tag{35}$$

Further assume $|\overline{J} \setminus \underline{J}| \leq C'$ for some constant $C' > 0$. Suppose the events (30) and

$$\left\{ \rho \|\widehat{A}_{\overline{J}} - (AP)_{\overline{J}}\|_{1,\infty} \leq \frac{1}{24} \kappa^2(A_{\underline{J}}, s) \right\} \tag{36}$$

hold with probability $1 - \alpha$, for some permutation matrix $P \in \mathcal{H}_K$. Then we have, with probability $1 - 8p^{-1} - \alpha$,

$$\|\widehat{T} - P^\top T_*\|_1 \lesssim \frac{1}{\kappa(A_{\overline{J}}, s)} \sqrt{\frac{K \log(p)}{N}} + \frac{1}{\kappa^2(A_{\overline{J}}, s)} \|\widehat{A}_{\overline{J}} - (AP)_{\overline{J}}\|_{1,\infty}.$$

Condition (35) only differs from condition (19) for known A by a $\log(p)$ term. Compared to the restrictions (30) and (31) in Theorem 8, Theorem 9 replaces (31) by the stronger requirement (36) on $\|\widehat{A}_{\overline{J}} - (AP)_{\overline{J}}\|_{1,\infty}$ by a factor $\rho/\kappa(A_{\underline{J}}, s)$.

Regarding the support recovery of $\widehat{T}$, we also have an analogue of Theorem 5 for unknown A. The following theorem states the one-sided support recovery of $\widehat{T}$ in (29) when A is unknown and estimated by $\widehat{A} \in \mathcal{A}$. Its proof can be found in Appendix G.3.

THEOREM 10. *Assume there exists some positive constants C, C', C'' such that $N \geq C \log(p)/T_{\min}^3$, $|\overline{J} \setminus \underline{J}| \leq C'$ and $\kappa^{-1}(A_{\underline{J}}, s) \leq C''$. Suppose the intersection of events (30), (36) and*

$$\begin{aligned} &\sqrt{\frac{\xi \log(p)}{T_{\min} N}} \left(1 + \sqrt{\frac{\xi s}{T_{\min}}}\right) + \frac{\log(p)}{N} + \left(1 + \frac{\xi}{T_{\min}}\right) \|\widehat{A}_{\overline{J}} - (AP)_{\overline{J}}\|_{1,\infty} \\ &\quad \leq c \min_{k \in S_*^c} \sum_{j \in \overline{J}^c} A_{jk} \end{aligned} \tag{37}$$

holds with probability at least $1 - \alpha$, for some permutation matrix $P \in \mathcal{H}_K$ and some sufficiently small constant $c > 0$. Then

$$\mathbb{P}\{\operatorname{supp}(\widehat{T}) \subseteq \operatorname{supp}(T_*)\} \geq 1 - 10p^{-1} - \alpha.$$

Comparing to (24) in Theorem 5, condition (37) is stronger by the factor $\log(p)/N + (1 + \xi/T_{\min})\|\widehat{A}_{\overline{J}} - (AP)_{\overline{J}}\|_{1,\infty}$ due to the error of estimating A. Theorem 10 in conjunction with Theorem 9 immediately implies that, under the conditions therein,

$$\|\widehat{T} - P^\top T_*\|_1 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{s \log p}{N}} + \|\widehat{A}_{\overline{J}} - (AP)_{\overline{J}}\|_{1,\infty}\right).$$

Theorem 10 provides the one-sided support recovery of the estimator $\widehat{T}$ based on an estimated A that satisfies (30), (36) and (37). Similar to the results we established for $\widehat{T}_{\text{mle}}$ in Section 2.1.3, the support of $\widehat{T}$ can also consistently recover the support of T_*, over a certain class of topic models, as discussed in Appendix A.2.

REMARK 7. Our estimation of T_* uses a plug-in estimator $\widehat{A}$ of A in (29). The estimation error naturally depends on how well $\widehat{A}$ estimates A. Alternatively, if one is willing to assume additional structure on $\boldsymbol{T}_*$, then there exist approaches that directly estimate $\boldsymbol{T}_*$ without estimating A first. See, for instance, [6] and [26].

2.2.3. *Application with the estimator proposed in* [9]. Our results in Section 2.2.2 hold for any estimator $\widehat{A} \in \mathcal{A}$ provided that the rate of $\widehat{A}$ satisfies certain requirements. In this section, we illustrate these general results by taking $\widehat{A}$ as the estimator proposed in [9] and by providing concrete conditions for the aforementioned requirements on $\widehat{A}$.

Since [9] studies the estimation of A under Assumption 1, we denote by I_k the index set of anchor words in topic k for each $k \in [K]$. We write $|I_{\max}| = \max_{k\in[K]} |I_k|$ and $I = \bigcup_{k=1} I_k$ with its complement set $I^c = [p] \setminus I$. Let $M := n \vee p \vee N$. Under conditions stated in Appendix K.1, [9] establishes the following guarantees on $\widehat{A}$:

$$\min_{P\in\mathcal{H}_K} \|\widehat{A} - AP\|_{1,\infty} = \mathcal{O}_{\mathbb{P}}\Bigg(\sqrt{\frac{K(|I_{\max}| + |I^c|)\log(M)}{nN}}\Bigg). \tag{38}$$

The above rate of convergence in $\|\cdot\|_{1,\infty}$ norm is useful to apply Theorem 9 and is further shown to be minimax optimal, up to the factor $\log(M)$, in [9] under Assumption 1. To validate condition (30) in Theorem 9, one also needs a control of $\|\widehat{A}_{j\cdot} - (AP)_{j\cdot}\|_\infty$ for $j \in \overline{J}$, which is not studied in [9]. We establish a new result on the rate of convergence of $\|\widehat{A}_{j\cdot} - (AP)_{j\cdot}\|_\infty$, that is,

$$\min_{P\in\mathcal{H}_K} \big\|\widehat{A}_{j\cdot} - (AP)_{j\cdot}\big\|_\infty \lesssim \sqrt{\|A_{j\cdot}\|_\infty \frac{K\log(M)}{nN}}\big(1 \vee \sqrt{p\|A_{j\cdot}\|_\infty}\big) \tag{39}$$

holds uniformly over $j \in \overline{J}$ with probability at least $1 - \mathcal{O}(M^{-1})$. We defer its precise statement and proof to Theorem K.1 of Appendix K.1. Equipped with the guarantees on $\widehat{A}$ in (38) and (39), for the estimator $\widehat{T}$ of T_* that uses $\widehat{A}$ as the estimator of A, the following corollary provides the rate of convergence of $\|\widehat{T} - T_*\|_1$ and its one-sided support recovery. Set $\Pi_{\min} := \min_{j\in\overline{J}} \Pi_{*j}$.

COROLLARY 11. *Assume that the quantities* $\kappa^{-1}(A_{\underline{J}}, s)$, $\kappa^{-1}(A_{\overline{J}}, K)$, ξ *and* $|\overline{J} \setminus \underline{J}|$ *are bounded*,

$$N \geq C\frac{\log(p)}{T_{\min}^2}\max\Bigg\{\frac{1}{T_{\min}}, 1 + \sqrt{K - s}\Bigg\} \tag{40}$$

and

$$T_{\min} \gtrsim \sqrt{\frac{pK\log(M)}{nN}}, \qquad \Pi_{\min} T_{\min} \gtrsim \frac{K\log(M)}{nN}. \tag{41}$$

Then the estimator $\widehat{T}$ *from* (29) *based on* $\widehat{A}$ *satisfies*

$$\min_{P\in\mathcal{H}_K} \|\widehat{T} - P^\top T_*\|_1 = \mathcal{O}_{\mathbb{P}}\Bigg(\sqrt{\frac{K\log(p)}{N}} + \sqrt{\frac{K(|I_{\max}| + |I^c|)\log(M)}{nN}}\Bigg).$$

Furthermore, *if*

$$\frac{1}{T_{\min}}\sqrt{\frac{s\log(p)}{N}} + \frac{1}{T_{\min}}\sqrt{\frac{K(|I_{\max}| + |I^c|)\log(M)}{nN}} \leq c\min_{k\in S_*^c} \sum_{j\in\overline{J}^c} A_{jk}, \tag{42}$$

holds for some sufficiently small constant $c > 0$, *then with probability tending to one as* $p \to \infty$, *we have* $\operatorname{supp}(\widehat{T}) \subseteq \operatorname{supp}(T_*)$ *and*

$$\min_{P \in \mathcal{H}_K} \|\widehat{T} - P^\top T_*\|_1 \lesssim \sqrt{\frac{s \log(p)}{N}} + \sqrt{\frac{K(|I_{\max}| + |I^c|)\log(M)}{nN}}.$$

The result of Corollary 11 requires that

(a) A is well behaved in that the quantities $\kappa^{-1}(A_{\underline{J}}, s)$, $\kappa^{-1}(A_{\overline{J}}, K)$ and ξ are bounded,
(b) there are only finitely many very small probability words ($|\overline{J} \setminus \underline{J}|$ stays bounded),
(c) the sample size N is large enough to guarantee (40),
(d) the corpus size n and sample size N are large enough and both topic probabilities and word probabilities need to satisfy mild signal strength conditions to guarantee (41), and
(e) A is incoherent, to satisfy (42) for one-sided support recovery.

The final bound for $\min_P \|\widehat{T} - P^\top T_*\|_1$ involves two terms. Provided

$$n \gtrsim K\big(|I_{\max}| + |I^c|\big)\log(M)/s, \tag{43}$$

the rate $\sqrt{s\log(p)/N}$ dominates and compared to Corollary 6 and Theorem 7, Corollary 11 implies that the estimator $\widehat{T}$ that uses $\widehat{A}$ in [9] has the same optimal convergence rate as $\widehat{T}_{\mathrm{mle}}$ that uses the true A, up to a $\log(p)$ factor. By using $|I_{\max}| + |I^c| < p$, one set of sufficient conditions for (43) is $n \gg p\log(M)$ and $K \asymp s$. In many topic model applications, the number of documents n is typically much larger than the vocabulary size p and the number of topics remains small. For instance, in the IMDB movie reviews in Appendix B, we have $p \approx 500$ while $n \approx 20{,}000$ with the estimated K being 6.

2.3. *Estimation of* Π_* *in topic models*. We compare the model-based estimator of Π_* with the empirical estimator in two aspects: the ℓ_1 convergence rate and the estimation of probabilities corresponding to zero observed frequencies.

2.3.1. *Improved convergence rate*. We begin our discussion for known A. Let $\widetilde{\Pi}_A = A\widehat{T}_{\mathrm{mle}}$ be the model-based estimator of Π_* with $\widehat{T}_{\mathrm{mle}}$ obtained in (4) of Section 2.1. Recall that $\widehat{\Pi} = X$ is the empirical estimator of Π_*. Further recall $\overline{J} = \{j : \Pi_{*j} > 0\}$ from (5) and write $\overline{p} = |\overline{J}|$. Consider $s = K$, for simplicity.

For $\widehat{\Pi}$, it is easy to see, using the fact that each component of $N\widehat{\Pi}$ has a Binomial distribution and the Cauchy–Schwarz inequality (twice), that

$$\mathbb{E}\|\widehat{\Pi} - \Pi_*\|_1 \le \sum_{i \in \bar{J}} \sqrt{\frac{\Pi_{*i}}{N}} \le \sqrt{\frac{\overline{p}}{N}} \tag{44}$$

holds. Furthermore, the bound (44) is also sharp (one instance is when $\Pi_{*i} \asymp 1/\overline{p}$). On the other hand, Corollary 3 together with $\|A\|_{1,\infty} = 1$ implies

$$\mathbb{E}\|\widetilde{\Pi}_A - \Pi\|_1 \le \mathbb{E}\|\widehat{T}_{\mathrm{mle}} - T\|_1 \lesssim \sqrt{\frac{K}{N}},$$

provided that $\kappa^{-1}(A, K)$ is bounded. This rate is faster than the rate (44) for $\widehat{\Pi}$ by a factor $\sqrt{K/\overline{p}}$. In the high-dimensional setting where $p \ge \overline{p} \gg N$, the bound in (44) does not converge to zero unless the summability condition $\sum_{i \in \overline{J}} \sqrt{\Pi_{*i}} = \mathcal{O}(1)$ holds. In contrast, consistency of $\widetilde{\Pi}_A$ is guaranteed as long as $K = o(N)$.

When A is unknown, the rate of the empirical estimator can still be improved by the model-based estimator $\widetilde{\Pi}_{\widehat{A}} = \widehat{A}\widehat{T}$ with $\widehat{T}$ obtained from (29) by using an accurate estimator $\widehat{A} \in \mathcal{A}$

of A. Specifically, provided that $\kappa^{-1}(A_{\overline{J}}, s)$ and ξ are bounded, the error due to estimating A plays the following role in estimating Π_*:

$$\mathbb{E}\|\widetilde{\Pi}_{\widehat{A}} - \Pi_*\|_1 \le \min_P \{\mathbb{E}\|\widehat{A}_{\overline{J}} - (AP)_{\overline{J}}\|_{1,\infty} + \mathbb{E}\|\widehat{T} - P^\top T_*\|_1\}$$

$$\lesssim \min_P \mathbb{E}\|\widehat{A}_{\overline{J}} - (AP)_{\overline{J}}\|_{1,\infty} + \sqrt{\frac{K\log p}{N}},$$

where we used $\|A\|_{1,\infty} = 1$ and $\|\widehat{T}\|_1 = 1$ in the first line and invoked Theorem 9 to derive the second line. For the estimator $\widehat{A}$ studied in Section 2.2.3, we have

$$\mathbb{E}\|\widetilde{\Pi}_{\widehat{A}} - \Pi_*\|_1 \lesssim \sqrt{\frac{K(|I_{\max}| + |I^c|)\log(M)}{nN}} + \sqrt{\frac{K\log(p)}{N}}.$$

If $(|I_{\max}| + |I^c|)\log(M) \le n$, the above rate simplifies to $\sqrt{K\log(p)/N}$. Moreover, as long as

$$n \ge K(|I_{\max}| + |I^c|)\log(M)/\overline{p}$$

and $K\log p \le \overline{p}$, the estimate $\widetilde{\Pi}_{\widehat{A}}$ improves upon $\widehat{\Pi}$ (in the ℓ_1 norm).

Our model-based estimation of $\mathbf{\Pi}_*$ uses the topic model assumption (1) and is to some extent related to other works, such as [17, 38], where the estimation of $\mathbf{\Pi}_*$ is studied under a low-rank structure of $\mathbf{\Pi}_*$.

2.3.2. *Estimating word probabilities corresponding to zero observed frequencies.* One distinct aspect of the model-based mixture estimator compared to the empirical estimator lies in the estimation of the cell probabilities Π_{*j} with $j \in J^c = \{j : X_j = 0\}$.

We distinguish between two situations: (i) $j \in J^c$ and $\Pi_{*j} > 0$ and (ii) $j \in J^c$ and $\Pi_{*j} = 0$. We discuss them separately. For ease of reference to the results of the previous sections, recall that $\overline{J} = \{j : \Pi_{*j} > 0\}$.

In case (i), the empirical estimator always estimates Π_{*j} by $\widehat{\Pi}_j = X_j = 0$, while the mixture estimator $\widetilde{\Pi}_A$ may produce nonzero estimates, as it is designed to combine the strength of the mixture components. For instance, if condition (16) holds, then $[1 - o_{\mathbb{P}}(1)]\Pi_{*j} \le \widetilde{\Pi}_{A,j} \le [1 + o_{\mathbb{P}}(1)]\Pi_{*j}$, for all $j \in \overline{J}$, that is,

$$|\widetilde{\Pi}_{A,j} - \Pi_{*j}| = o_{\mathbb{P}}(\Pi_{*j}) = o_{\mathbb{P}}\big(|\widehat{\Pi}_j - \Pi_{*j}|\big) \quad \forall j \in \overline{J} \cap J^c,$$

showing that, indeed, $\widetilde{\Pi}_{A,j}$ is a nonzero estimator of a nonzero Π_{*j}, and has smaller estimation error than $\widehat{\Pi}_j$.

In case (ii), for any j such that $\Pi_{*j} = X_j = 0$, the empirical estimator makes no mistake while the model-based estimator $\widetilde{\Pi}_{A,j} = A_{j\cdot}^\top \widehat{T}_{\mathrm{mle}}$ could be nonzero. However, we remark that the total error of estimating $j \in \overline{J}^c$ made by $\widetilde{\Pi}_A$ is at most $\|(\widehat{T}_{\mathrm{mle}} - T_*)_{S_*^c}\|_1$, which converges to zero no slower than $\sqrt{K\log(K)/N}$ as shown in Section 2.1. Indeed, by the fact that $A_{jS_*} = \mathbf{0}$ for $j \in \overline{J}^c$,

$$\sum_{j\in\overline{J}^c} |\widetilde{\Pi}_{A,j} - \Pi_{*j}| = \sum_{j\in\overline{J}^c} A_{jS_*^c}^\top (\widehat{T}_{\mathrm{mle}})_{S_*^c}$$

$$\le \max_{k\in S_*^c} \sum_{j\in\overline{J}^c} A_{jk} \big\|(\widehat{T}_{\mathrm{mle}} - T_*)_{S_*^c}\big\|_1$$

$$\le \big\|(\widehat{T}_{\mathrm{mle}} - T_*)_{S_*^c}\big\|_1.$$

In particular, if $\mathrm{supp}(\widehat{T}_{\mathrm{mle}}) \subseteq \mathrm{supp}(T_*)$ holds, $\|(\widehat{T}_{\mathrm{mle}} - T_*)_{S_*^c}\|_1 = 0$ and $\widetilde{\Pi}_A$ makes no mistake of estimating Π_{*j} for $j \in \overline{J}^c$.

Summarizing, on the one hand, we expect the model-based estimator to outperform the empirical estimator for estimating the cell probabilities in (i). On the other hand, the model-based estimator is no worse than the empirical estimator for estimating the cell probabilities in (ii) when A satisfies an incoherence condition (for instance, condition (24)). We verify these two points in our simulation studies in Appendix D.

3. The 1-Wasserstein distance between documents in topic models. We now turn to the main application of the results of Section 2. By abuse of terminology, we refer to the 1-Wasserstein distance between probabilistic representations of documents as the distance between documents. This section is devoted to the theoretical evaluation of the Wasserstein distance between appropriate discrete distributions, in topic models, and to the illustration of our proposed methods and theory to the analysis of a real data set.

Consider two discrete distributions γ, ρ on $\mathcal{X} := \{x_1, \ldots, x_\ell, \ldots, x_L\}$, with $x_\ell \in E$, where E is a general, abstract space, and for some $L \geq 1$. Let D be a metric on $\mathcal{X}$ and denote by $\boldsymbol{D} := (D(x_a, x_b))_{1 \leq a,b \leq L}$ the $L \times L$ matrix that collects pairwise distances between the elements in $\mathcal{X}$. The W_1 distance between γ and ρ with respect to the metric D is defined as

$$W_1(\gamma, \rho; D) := \inf_{w \in \Gamma(\gamma, \rho)} \operatorname{tr}(w\boldsymbol{D}), \tag{45}$$

where $\Gamma(\gamma, \rho)$ is the set of discrete distributions w on $\mathcal{X} \times \mathcal{X}$ with marginals γ and ρ, respectively. In the above notation, w is a doubly-stochastic $L \times L$ matrix.

3.1. *The 1-Wasserstein distance between probabilistic representations of documents at the word and topic level.* We consider two alternative probabilistic representations of a document i: (1) as a probability vector on p words, $\Pi_*^{(i)}$, or (2) as a probability vector on K topics, $T_*^{(i)}$.

In view of our data example in Appendix B, we regard words as vectors in $\mathbb{R}^d$, for some d. Pretrained embeddings of words [30], sentences [34] and documents [28] have become a popular general approach in natural language processing [31], and in particular allow one to define metrics between words as metrics between their Euclidean vector representations. Specifically, let $\mathcal{X}_{\text{word}} := \{x_1, \ldots, x_a, \ldots, x_p\}$, so $x_a \in \mathbb{R}^d$ is a vector representing word a in the dictionary via an embedding in $\mathbb{R}^d$. Then, with $\|\cdot\|_2$ denoting the Euclidean distance on $\mathbb{R}^d$, we define

$$D^{\text{word}}(a, b) := \|x_a - x_b\|_2 \tag{46}$$

as the distance between words a and b for $a, b \in [p]$. The 1-Wasserstein distance between two discrete distributions $\Pi_*^{(i)}$ and $\Pi_*^{(j)}$ supported on these words, for any $i, j \in \{1, \ldots, n\}$, is

$$W_1\big(\Pi_*^{(i)}, \Pi_*^{(j)}; D^{\text{word}}\big) := \inf_{w \in \Gamma(\Pi_*^{(i)}, \Pi_*^{(j)})} \operatorname{tr}\big(w\boldsymbol{D}^{\text{word}}\big). \tag{47}$$

Alternatively, viewing the corpus as an ensemble, and under model (1), document differences can be explained in terms of 1-Wasserstein distances between what can be regarded as sketches of the documents, the topic distributions $\boldsymbol{T}_*$ in (1). For each document $i \in [n]$, the topic proportion $T_*^{(i)}$ is a discrete distribution supported on K topics. Analogous to (47), we define a population-level distance between topic distributions in document i and j, based on the 1-Wasserstein distance, by

$$W_1\big(T_*^{(i)}, T_*^{(j)}; D^{\text{topic}}\big) = \inf_{\alpha \in \Gamma(T_*^{(i)}, T_*^{(j)})} \operatorname{tr}\big(\alpha\boldsymbol{D}^{\text{topic}}\big), \tag{48}$$

where $\boldsymbol{D}^{\text{topic}} \in \mathbb{R}_+^{K \times K}$ is a metric matrix on K topics.

To define D^{topic}, we view a topic as being itself a distribution, on words. Specifically, for every $k \in [K]$, topic k is a distribution on the p words of the dictionary, with mass corresponding to $A_{\cdot k} \in \Delta_p$. We recall that the topic model specifies A_{jk} as the probability of word j given topic k. We therefore let $\mathcal{X}_{\text{topic}} = \{A_{\cdot 1}, \ldots, A_{\cdot k}, \ldots, A_{\cdot K} : A_{\cdot k} \in \Delta_p \text{ for } k \in [K]\}$. With this view, metrics between two topics k and l are distances between discrete distributions $A_{\cdot k}$ and $A_{\cdot \ell}$ in Δ_p, with supports in $\mathcal{X}_{\text{word}}$.

In this work, we focus on two closely related such metrics. The first one is itself a 1-Wasserstein distance:

$$D_W^{\text{topic}}(k, \ell) := W_1\big(A_{\cdot k}, A_{\cdot \ell}, D^{\text{word}}\big), \quad \forall k, \ell \in [K], \tag{49}$$

the calculation of which requires optimization in p dimensions and employs input $\boldsymbol{D}^{\text{word}}$ which, in the context of text analysis, is obtained from domain knowledge, as explained above, and further discussed in Appendix B. The second metric is the total variation, TV distance:

$$D_{\text{TV}}^{\text{topic}}(k, \ell) := \frac{1}{2}\|A_{\cdot k} - A_{\cdot \ell}\|_1, \quad \forall k, \ell \in [K], \tag{50}$$

which is optimization free, and independent of the domain knowledge required by (49).

We note that the space $\mathcal{X}_{\text{topic}}$ is bounded with respect to both metrics (49) and (50). In particular, the total variation distance is always bounded by 1, and hence, $\|\boldsymbol{D}_{\text{TV}}^{\text{topic}}\|_\infty \le 1$. Furthermore, by Lemma H.2 in Appendix H, for any $k, \ell \in [K]$,

$$D_W^{\text{topic}}(k, \ell) = W_1\big(A_{\cdot k}, A_{\cdot \ell}, D^{\text{word}}\big) \le \big\|\boldsymbol{D}^{\text{word}}\big\|_\infty \frac{1}{2}\|A_{\cdot k} - A_{\cdot \ell}\|_1 \le \big\|\boldsymbol{D}^{\text{word}}\big\|_\infty,$$

and thus, $\|\boldsymbol{D}_W^{\text{topic}}\|_\infty \le \|\boldsymbol{D}^{\text{word}}\|_\infty$. As noted in Remark 8 below, $\|\boldsymbol{D}^{\text{word}}\|_\infty$ is typically bounded; in practice, word embeddings are often normalized to unit-length, in which case $\|\boldsymbol{D}^{\text{word}}\|_\infty \le 2$.

3.2. *Finite sample error bounds for estimates of the* 1*-Wasserstein distance between documents*. The theoretical analysis of estimates of the 1-Wasserstein distance $W_1(\gamma, \rho; D)$ between discrete probability measures γ and ρ supported on a metric space $\mathcal{X}$ endowed with metric D has been restricted, to the best of our knowledge, to estimates $W_1(\widehat{\rho}^{(i)}, \widehat{\gamma}^{(j)}; D)$ corresponding to observed empirical frequencies $\widehat{\rho}^{(i)}$, $\widehat{\gamma}^{(j)}$, respectively, observed on samples i and j, of sizes N_i and N_j.

We drop the superscripts and subscripts in the next few paragraphs, for ease of presentation, to give a brief overview of the one-sample related results.

When L is fixed and $(\mathcal{X}, D)$ has bounded diameter, [35] showed that $\sqrt{N} W_1(\widehat{\rho}, \rho; D)$ converges in distribution, while [36] showed that when $p = \infty$ and their summability condition (3) holds, $\sqrt{N} W_1(\widehat{\rho}, \rho; D)$ converges weakly over the set of probability measures with finite first moment with respect to D, defined in their Section 2.1.

Finite sample rates of convergence for $W_1(\widehat{\rho}, \rho; D)$ when $L = L(N)$ are less studied, with the exception of [37], who showed that they are of the order $\sqrt{L/N}$, for $L < N$, when $(\mathcal{X}, D)$ has bounded diameter, and obtained this result as a particular case of a general theory.

When $(\mathcal{X}, D)$ has bounded diameter, the rate of $W_1(\widehat{\rho}, \rho; D)$ can be obtained directly from a bound on $\|\widehat{\rho} - \rho\|_1$, via the basic inequalities $c\|\widehat{\rho} - \rho\|_1 \le W_1(\widehat{\rho}, \rho; D) \le C\|\widehat{\rho} - \rho\|_1$ [21], where $c = \min_{x \ne y \in \mathcal{X}} D(x, y)$ and $C = \max_{x, y \in \mathcal{X}} D(x, y)$. Therefore, when $\rho, \widehat{\rho} \in \Delta_L$, and $\widehat{\rho}$ are observed frequencies, the rate $W_1(\widehat{\rho}, \rho; D) \lesssim \sqrt{L/N}$, with high probability, is therefore immediate, and is small when $L < N$. Furthermore, $W_1(\widehat{\rho}, \rho; D) \lesssim \sqrt{1/N}$ when $\sum_{j=1}^{L} \sqrt{\rho_j} < \infty$, for any L, allowed to depend on N and be larger than N, matching the rate established for $L = \infty$ in [36].

We complement this literature by constructing and analyzing alternate estimates of the 1-Wasserstein distance between discrete distributions generated according to a topic model (1). After obtaining any estimate $\widehat{A} \in \mathcal{A}$ and the estimate $\widehat{T}^{(\ell)}$ from (29) by using this $\widehat{A}$ and $X^{(\ell)}$, for each $\ell \in \{i, j\}$, we propose to estimate the word-level document distance (47) by

$$W_1\big(\widetilde{\Pi}^{(i)}, \widetilde{\Pi}^{(j)}; D^{\text{word}}\big), \quad \text{with } \widetilde{\Pi}^{(\ell)} = \widehat{A}\widehat{T}^{(\ell)}, \forall \ell \in \{i, j\}. \tag{51}$$

For the Wasserstein distance between topic distributions in (48) with the two choices of D^{topic} in (49) and (50), we propose to estimate $W_1(T_*^{(i)}, T_*^{(j)}; D_W^{\text{topic}})$ and $W_1(T_*^{(i)}, T_*^{(j)}; D_{\text{TV}}^{\text{topic}})$, respectively, by

$$W_1\big(\widehat{T}^{(i)}, \widehat{T}^{(j)}; \widehat{D}_W^{\text{topic}}\big), \quad \text{with } \widehat{D}_W^{\text{topic}}(k, \ell) = W_1\big(\widehat{A}_{\cdot k}, \widehat{A}_{\cdot \ell}; D^{\text{word}}\big), \forall k, \ell \in [K]; \tag{52}$$

$$W_1\big(\widehat{T}^{(i)}, \widehat{T}^{(j)}; \widehat{D}_{\text{TV}}^{\text{topic}}\big), \quad \text{with } \widehat{D}_{\text{TV}}^{\text{topic}}(k, \ell) = \frac{1}{2}\|\widehat{A}_{\cdot k} - \widehat{A}_{\cdot \ell}\|_1, \forall k, \ell \in [K]. \tag{53}$$

The following proposition shows how error rates of the various Wasserstein distance estimates depend on the estimation of A and $T_*^{(\ell)}$. Its proof can be found in Appendix H. Recall that $\|M\|_{1,\infty} = \max_j \|M_{\cdot j}\|_1$ for any matrix M. Define

$$R\big(\widehat{A}, \widehat{T}^{(i)}, \widehat{T}^{(j)}\big) := \min_{P \in \mathcal{H}_K} \Big\{ \|\widehat{A} - AP\|_{1,\infty} + \frac{1}{2} \sum_{\ell \in \{i,j\}} \|\widehat{T}^{(\ell)} - P^\top T_*^{(\ell)}\|_1 \Big\}.$$

PROPOSITION 12. *For any estimator $\widehat{A} \in \mathcal{A}$ and the estimators $\widehat{T}^{(i)}$, $\widehat{T}^{(j)}$ from (29) based on this $\widehat{A}$, we have*:

$$\big|W_1\big(\widetilde{\Pi}^{(i)}, \widetilde{\Pi}^{(j)}; D^{\text{word}}\big) - W_1\big(\Pi_*^{(i)}, \Pi_*^{(j)}; D^{\text{word}}\big)\big| \le \|\boldsymbol{D}^{\text{word}}\|_\infty R\big(\widehat{A}, \widehat{T}^{(i)}, \widehat{T}^{(j)}\big); \tag{54}$$

$$\big|W_1\big(\widehat{T}^{(i)}, \widehat{T}^{(j)}; \widehat{D}_W^{\text{topic}}\big) - W_1\big(T_*^{(i)}, T_*^{(j)}; D_W^{\text{topic}}\big)\big| \le \|\boldsymbol{D}^{\text{word}}\|_\infty R\big(\widehat{A}, \widehat{T}^{(i)}, \widehat{T}^{(j)}\big); \tag{55}$$

$$\big|W_1\big(\widehat{T}^{(i)}, \widehat{T}^{(j)}; \widehat{D}_{\text{TV}}^{\text{topic}}\big) - W_1\big(T^{(i)}, T^{(j)}; D_{\text{TV}}^{\text{topic}}\big)\big| \le R\big(\widehat{A}, \widehat{T}^{(i)}, \widehat{T}^{(j)}\big). \tag{56}$$

We provide supporting simulations in Appendix E to study the rate of estimation of document distances, focusing on the estimator (53) as an illustrative example.

COROLLARY 13. *Under conditions of Corollary 11, for the estimator $\widehat{A}$ proposed in* [9], *and estimators $\widehat{T}^{(i)}$, $\widehat{T}^{(j)}$ from (29) based on this $\widehat{A}$, with probability tending to one, the bounds given in Proposition 12 hold with*

$$R\big(\widehat{A}, \widehat{T}^{(i)}, \widehat{T}^{(j)}\big) \lesssim \sqrt{\frac{\max\{\|T_*^{(i)}\|_0, \|T_*^{(j)}\|_0\}\log(p)}{N}} + \sqrt{\frac{K(|I_{\max}| + |I^c|)\log(M)}{nN}}.$$

REMARK 8. We make the following remarks:

1. All error upper bounds given by Proposition 12 are of the same order, when $\|\boldsymbol{D}^{\text{word}}\|_\infty \le C$, for some constant $C > 0$. In practice, word embedding vectors are often normalized to unit length when used to define D^{word}, in which case $\|\boldsymbol{D}^{\text{word}}\|_\infty \le 2$.

2. The first two error bounds are the same, but in the first the estimation of both A and T_* play a role in the estimation of Π_*, whereas the second bound is influenced by the estimation of A via the estimation of the distance metric. Although the error bounds are the same, computing the LHS of (54) involves an optimization in dimension p, whereas the LHS of (55) is in the much lower dimension K. Although the distance metric in (55) does require the computation of $K(K-1)/2$ Wasserstein distances in dimension p, as in (54), all $n(n-1)/2$ pairwise distances between the documents in the corpus can be computed by only

a K-dimensional Wasserstein distance; this results in a substantial computational gain for K small and n and p large, the typical case in topic modeling (in our example in Appendix B, $n = 20{,}605$ and $p = 500$, whereas $K = 6$). We note that approximations to the W_1 distance can be considered to reduce computational complexity at the cost of accuracy, as in [27]; we instead focus on exact calculation of the W_1 distance, but in a reduced dimension (K).

3. The LHS in (56) is once again an optimization in dimension K, with input independent of $\|\boldsymbol{D}^{\text{word}}\|_\infty$ and, therefore, its bound is also independent of this quantity. Furthermore, $\widehat{D}_{\text{TV}}^{\text{topic}}$ is computed from simple ℓ_1 norms of the columns of $\widehat{A}$, so avoids the computational issues of the Wasserstein distance entirely.

4. We will shortly illustrate the advantage of our Wasserstein distance estimates in Appendix B below, where we analyze an IMBD movie review corpus. To exploit the geometry of the word embeddings, [27] was the first to suggest using the 1-Wasserstein distance (also known as the Earth Mover's Distance) between the word frequency vectors $\widehat{\Pi}^{(i)}$, $\widehat{\Pi}^{(j)}$. The benefit of using the Wasserstein distance, relative to the previously used ℓ_2 or TV distances, is that it takes into account the relative distance between words, as captured by D^{word}, so documents with similar meaning can have a small distance even if there is little overlap in the exact words they use.

The analysis of a corpus of movie reviews, presented in Appendix B, illustrates on the same data set that the three newly proposed document-distance estimates, $W_1(\widetilde{\Pi}^{(i)}, \widetilde{\Pi}^{(j)}; D^{\text{word}})$ and $W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}; \widehat{D}^{\text{topic}})$, for estimates of $\widehat{D}^{\text{topic}}$ of the two metrics defined in (49) and (50), are competitive. In particular, $W_1(\widehat{T}^{(i)}, \widehat{T}^{(j)}; \widehat{D}_{\text{TV}}^{\text{topic}})$ yields qualitatively similar results, relative to our other two proposed distances, while having the net benefit of involving optimization only in K dimensions, and $K \ll p$, typically by several orders of magnitude. Furthermore, it obviates the need for pretrained word embeddings. Our analysis further reveals that all our proposed distance estimates capture well topical differences between the documents, while the standard $W_1(\widehat{\Pi}^{(i)}, \widehat{\Pi}^{(j)}; D^{\text{word}})$ between observed document frequencies is substantially less successful.

Acknowledgments. We thank the Editor, the Associate Editor and two referees for their detailed reviews, which helped to improve the paper substantially.

Funding. Bunea is supported in part by NSF Grant DMS-2015195 and DMS-2210563. Wegkamp is supported in part by NSF Grants DMS-2015195 and DMS-2210557.

SUPPLEMENTARY MATERIAL

Supplement to "Likelihood estimation of sparse topic distributions in topic models and its applications to Wasserstein document distance calculations" (DOI: 10.1214/22-AOS2229SUPP; .pdf). The supplement contains the analysis of an IMDB data set, all proofs, auxiliary results and all simulation results.

REFERENCES

[1] AGRESTI, A. (2013). *Categorical Data Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR3087436

[2] ANANDKUMAR, A., FOSTER, D. P., HSU, D. J., KAKADE, S. M. and LIU, Y.-K. (2012). A spectral algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems* 25 (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) 917–925. Curran Associates, Red Hook.

[3] 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.

[4] ARORA, S., GE, R., KOEHLER, F., MA, T. and MOITRA, A. (2016). Provable algorithms for inference in topic models. In *Proceedings of the* 33*rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.). *Proceedings of Machine Learning Research* **48** 2859–2867. PMLR, New York, New York, USA.

[5] ARORA, S., GE, R. and MOITRA, A. (2012). Learning topic models—going beyond SVD. In 2012 *IEEE* 53*rd Annual Symposium on Foundations of Computer Science—FOCS* 2012 1–10. IEEE Computer Soc., Los Alamitos, CA. MR3185945

[6] BANSAL, T., BHATTACHARYYA, C. and KANNAN, R. (2014). A provable SVD-based algorithm for learning topics in dominant admixture corpus. *Adv. Neural Inf. Process. Syst.* **27**.

[7] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[8] BING, X., BUNEA, F., STRIMAS-MACKEY, S. and WEGKAMP, M. (2022). Supplement to "Likelihood estimation of sparse topic distributions in topic models and its applications to Wasserstein document distance calculations." https://doi.org/10.1214/22-AOS2229SUPP

[9] BING, X., BUNEA, F. and WEGKAMP, M. (2020). A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli* **26** 1765–1796. MR4091091 https://doi.org/10.3150/19-BEJ1166

[10] BING, X., BUNEA, F. and WEGKAMP, M. (2020). Optimal estimation of sparse topic models. *J. Mach. Learn. Res.* **21** 177. MR4209463

[11] BIRCH, M. W. (1964). A new proof of the Pearson–Fisher theorem. *Ann. Math. Stat.* **35** 817–824. MR0169324 https://doi.org/10.1214/aoms/1177703581

[12] BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (2007). *Discrete Multivariate Analysis*: *Theory and Practice*. Springer, New York. With the collaboration of Richard J. Light and Frederick Mosteller, Reprint of the 1975 original. MR2344876

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

[14] BLEI, D. M. (2012). Introduction to probabilistic topic models. *Commun. ACM* **55** 77–84.

[15] BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* 993–1022.

[16] BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575 https://doi.org/10.1017/CBO9780511804441

[17] CAO, Y., ZHANG, A. and LI, H. (2020). Multisample estimation of bacterial composition matrices in metagenomics data. *Biometrika* **107** 75–92. MR4064141 https://doi.org/10.1093/biomet/asz062

[18] CHEN, S., RIVAUD, P., PARK, J. H., TSOU, T., CHARLES, E., HALIBURTON, J. R., PICHIORRI, F. and THOMSON, M. (2020). Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign. *Proc. Natl. Acad. Sci. USA* **117** 28784–28794.

[19] DING, W., ISHWAR, P. and SALIGRAMA, V. (2015). Most large topic models are approximately separable. In 2015 *Information Theory and Applications Workshop* (*ITA*) 199–203.

[20] DING, W., ROHBAN, M. H., ISHWAR, P. and SALIGRAMA, V. (2013). Topic discovery through data dependent and random projections. In *Proceedings of the* 30*th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.). *Proceedings of Machine Learning Research* **28** 1202–1210. PMLR, Atlanta, GA, USA.

[21] GIBBS, A. L. and SU, F. E. (2002). On choosing and bounding probability metrics. *Int. Stat. Rev.* **70** 419–435.

[22] GONZÁLEZ-BLAS, C. B., MINNOYE, L., PAPASOKRATI, D., AIBAR, S., HULSELMANS, G., CHRISTIAENS, V., DAVIE, K., WOUTERS, J. and AERTS, S. (2019). cisTopic: Cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16** 397–400. https://doi.org/10.1038/s41592-019-0367-1

[23] GRIFFITHS, T. L. and STEYVERS, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **101** 5228–5235.

[24] KE, T. Z. and WANG, M. (2017). A new SVD approach to optimal topic estimation. Available at arXiv:1704.07016.

[25] KLEINBERG, J. and SANDLER, M. (2008). Using mixture models for collaborative filtering. *J. Comput. System Sci.* **74** 49–69. MR2364181 https://doi.org/10.1016/j.jcss.2007.04.013

[26] KLOPP, O., PANOV, M., SIGALLA, S. and TSYBAKOV, A. (2021). Assigning topics to documents by successive projections. ArXiv preprint. Available at arXiv:2107.03684.

[27] KUSNER, M., KOLKIN, Y. S. N. I. and WEINBERGER, K. Q. (2015). From word embeddings to document distances. http://proceedings.mlr.press/v37/kusnerb15.pdf.

[28] LE, Q. V. and MIKOLOV, T. (2014). Distributed Representations of Sentences and Documents.

[29] MA, W.-K., BIOUCAS-DIAS, J. M., CHAN, T.-H., GILLIS, N., GADER, P., PLAZA, A. J., AMBIKAPATHI, A. and CHI, C.-Y. (2013). A signal processing perspective on hyperspectral unmixing: Insights from remote sensing. *IEEE Signal Process. Mag.* **31** 67–81.

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

[31] QIU, X., SUN, T., XU, Y., SHAO, Y., DAI, N. and HUANG, X. (2020). Pre-trained Models for Natural Language Processing: A Survey.

[32] RAO, C. R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā* **18** 139–148. MR0105183

[33] RAO, C. R. (1958). Maximum likelihood estimation for the multinomial distribution with infinite number of cells. *Sankhyā* **20** 211–218. MR0107334

[34] REIMERS, N. and GUREVYCH, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

[35] SOMMERFELD, M. and MUNK, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 219–238. MR3744719 https://doi.org/10.1111/rssb.12236

[36] TAMELING, C., SOMMERFELD, M. and MUNK, A. (2019). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *Ann. Appl. Probab.* **29** 2744–2781. MR4019874 https://doi.org/10.1214/19-AAP1463

[37] WEED, J. and BACH, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* **25** 2620–2648. MR4003560 https://doi.org/10.3150/18-BEJ1065

[38] ZHU, Z., LI, X., WANG, M. and ZHANG, A. (2021). Learning Markov models via low-rank optimization. *Oper. Res.*