Electronic Journal of Statistics
Vol. 13 (2019) 37–66
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1516

# Convergence rates of latent topic models under relaxed identifiability conditions

**Yining Wang**

*Machine Learning Department*
*Carnegie Mellon University*
*5000 Forbes Avenue, Pittsburgh PA 15213, USA*
*e-mail:* yiningwa@cs.cmu.edu

**Abstract:** In this paper we study the frequentist convergence rate for the Latent Dirichlet Allocation (Blei, Ng and Jordan, 2003) topic models. We show that the maximum likelihood estimator converges to one of the finitely many equivalent parameters in Wasserstein's distance metric at a rate of $n^{-1/4}$ without assuming separability or non-degeneracy of the underlying topics and/or the existence of more than three words per document, thus generalizing the previous works of Anandkumar et al. (2012, 2014) from an information-theoretical perspective. We also show that the $n^{-1/4}$ convergence rate is optimal in the worst case.

**Keywords and phrases:** Latent Dirichlet Allocation, topic models, maximum likelihood, rates of convergence.

Received March 2018.

## 1. Introduction

The *Latent Dirichlet Allocation (LDA)* model, first introduced by Blei, Ng and Jordan (2003), has been very influential in machine learning as a probabilistic admixture model that characterizes latent topic structures in natural language document collections. The original LDA paper (Blei, Ng and Jordan, 2003) has accumulated a total of over 20,000 citations up to the year of 2017, with many follow-up works also impactful in machine learning research (Griffiths and Steyvers, 2004; Blei, 2012; Fei-Fei and Perona, 2005; Blei and Lafferty, 2006). At a higher level, the LDA model posits the existence of $K$ latent (unknown) topic vectors, and models the generation of a document as a collection of $m$ *conditionally independent* words given a mixing topic vector for the document.

More specifically, let $V$ be the vocabulary size, $K$ be the number of topics and denote conveniently each of the $V$ words in the vocabulary as $1, 2, \cdots, V$. Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ where $\theta_k \in \Delta^{V-1} = \{\pi \in \mathbb{R}^V : \pi \geq 0, \sum_i \pi_i = 1\}$ be a collection of $K$ fixed but unknown topic word distribution vectors that one wishes to estimate. The LDA then models the generation of a document $X = (x_1, \cdots, x_m) \subseteq \{1, \cdots, V\} =: [V]$ of $m$ words as follows:

$$(x_1, \cdots, x_m)|h \overset{i.i.d.}{\sim} \text{Categorical}(h_1\theta_1 + \cdots + h_K\theta_K), \quad h \sim \nu_0. \tag{1}$$

Here Categorical($\pi$) is the categorical distribution over $[V]$ parameterized by $\pi \in \Delta^{V-1}$, meaning that $p(x = j|\pi) = \pi_j$ for $j \in [V]$, and $\nu_0$ is a known distribution that generates the "mixing vector" $h \in \Delta^{K-1}$. A likelihood model $p_{\boldsymbol{\theta}}(x)$ can then be explicitly written out as

$$\begin{aligned} p_{\boldsymbol{\theta}}(x) &= \int_{\Delta^{K-1}} p_{\boldsymbol{\theta},h}(x)\mathrm{d}\nu_0(h) = \int_{\Delta^{K-1}} \left[\prod_{i=1}^m p_{\boldsymbol{\theta},h}(x_i)\right] \mathrm{d}\nu_0(h) \\ &= \int_{\Delta^{K-1}} \left[\prod_{i=1}^m \sum_{k=1}^K h_k\theta_k(x_i)\right] \mathrm{d}\nu_0(h) \end{aligned} \tag{2}$$

for every $x = (x_1, \cdots, x_m) \in [V]^m$. In the original LDA model (Blei, Ng and Jordan, 2003) $\nu_0$ is taken to be the Dirichlet distribution, while in this paper we allow $\nu_0$ to belong to a much wider family of distributions. We also remark that the number of topics $K$ is *known* before estimation and inference.

The objective of this paper is to study rates of convergence for estimating $\boldsymbol{\theta}$ from a collection of independently sampled unlabeled documents $X_1, \cdots, X_n$. Each document is assumed to be of the same length $m$. [1] The estimation error between the underlying true model $\boldsymbol{\theta}$ and an estimator $\widehat{\boldsymbol{\theta}}$ is evaluated by their *Wasserstein's distance*:

$$d_{\mathsf{W}}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \min_{\pi:[K]\to[K]} \sum_{k=1}^K \|\theta_k - \widehat{\theta}_{\pi(k)}\|_1, \tag{3}$$

where $\pi : [K] \to [K]$ is a *permutation* on $K$. When $K$ and $V$ are fixed, the $\ell_1$-norm in the definition of Eq. (3) is not important as all vector $\ell_p$ norms are equivalent. Apart from the data (documents) $X_1, \cdots, X_n$, the vocabulary size $V$, number of topics $K$ and mixing distribution $\nu_0$ are also known. The latent topic mixing vectors $\{h_i\}_{i=1}^n$ as well as the parameter of interest $\{\theta_k\}_{k=1}^K$, on the other hand, are unknown.

When $\boldsymbol{\theta}$ satisfies certain non-degenerate conditions, such as $\{\theta_j\}_{j=1}^K$ being linear independent (Anandkumar et al., 2012, 2014) or satisfying stronger "anchor word" (Arora, Ge and Moitra, 2012) or "$p$-separability" conditions (Arora et al., 2013), computationally tractable estimators exist that recover $\boldsymbol{\theta}$ at an $n^{-1/2}$ rate measured in the Wasserstein's distance $d_{\mathsf{W}}(\cdot, \cdot)$. The general case of $\boldsymbol{\theta}$ being non-separable or degenerate, however, is much less understood. To the best of our knowledge, the only convergence result for general $\boldsymbol{\theta}$ case in the $d_{\mathsf{W}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ distance measure is due to Nguyen (2015), who established an $n^{-1/2(K+\alpha)}$ posterior contraction rate for hierarchical Dirichlet process models. We discuss in Sec. 1.1 several important differences between (Nguyen, 2015) and this paper.

We analyze the maximum likelihood estimation of the topic model in Eq. (1) and show that, with a relaxed "finite identifiability" definition, the ML estimator converges to one of the finitely many equivalent parameterizations (see

[1] Our analysis is still valid if the length of each document is sandwiched between two constants. However we decide to proceed with the assumption that each document is of equal length to simplify presentations.

Definition 2 and Theorem 1 for a rigorous statement) in Wasserstein's distance $d_{\mathsf{W}}(\cdot,\cdot)$ at the rate of at least $n^{-1/4}$ even if $\{\theta_j\}_{j=1}^K$ are non-separable or degenerate. Such rate is shown to be optimal by considering a simple "over-fitting" example. In addition, when $\{\theta_j\}_{j=1}^K$ are assumed to be linear independent, we recover the $n^{-1/2}$ parametric convergence rate established in (Anandkumar et al., 2012, 2014).

In terms of techniques, we adapt the classical analysis of rates of convergence for ML estimates in (Van der Vaart, 1998) to give convergence rates under finite identifiability settings. We also use Le Cam's method to prove corresponding local minimax lower bounds. At the core of our analysis is a binomial expansion of the total-variation (TV) distance between distributions induced by neighboring parameters, and careful calculations of the "level of degeneracy" in the TV-distance expansion of topic models, which subsequently determines the convergence rate.

### *1.1. Related work*

In the non-degenerate case where $\{\theta_j\}_{j=1}^K$ are linear independent, Anandkumar et al. (2012, 2014); Arora, Ge and Moitra (2012) applied the method of moments with noisy tensor decomposition techniques to achieve the $n^{-1/2}$ parametric rate for recovering the underlying topic vectors $\boldsymbol{\theta}$ in Wasserstein's distance. Extension and generalization of such methods are many, including supervised topic models (Wang and Zhu, 2014), model selection (Cheng, He and Liu, 2015), computational efficiency (Wang et al., 2015) and online/streaming settings (Huang et al., 2015; Wang and Anandkumar, 2016). Under slightly stronger "anchor word" type assumptions, Arora, Ge and Moitra (2012) developed algorithms beyond spectral decomposition of empirical tensors and Arora et al. (2013) demonstrated empirical success of the proposed algorithms.

Topic models are also intensively studied from a Bayesian perspective, with Dirichlet priors imposed on the underlying topic vectors $\boldsymbol{\theta}$. Early works considered variational inference (Blei, Ng and Jordan, 2003) and Gibbs sampling (Griffiths and Steyvers, 2004) for generating samples or approximations of the posterior distribution of $\boldsymbol{\theta}$. Tang et al. (2014); Nguyen (2015) considered the posterior contraction of the convex hull of topic vectors and derived an $\widetilde{N}^{-1/2}$ upper bound on the posterior contraction rate, where $\widetilde{N}=\frac{\log n}{n}+\frac{\log m}{m}+\frac{\log m}{n}$. Nguyen (2013, 2016) further considered the more difficult question of posterior contraction with respect to the Wasserstein's distance. Apart from the Bayesian treatments of posterior contraction that contrasts our frequentist point of view of worst-case convergence, one important aspect of the work of (Tang et al., 2014; Nguyen, 2015, 2013, 2016) is that the number of words per document $m$ has to grow together with the number of documents $n$, and the posterior contraction rate becomes vacuous (i.e., constant level of error) for fixed $m$ settings. In contrast, in this paper we consider $m$ being fixed as $n$ increases to infinity.

Our work is also closely related to convergence analysis of *singular* finite-mixture models. In fact, our $n^{-1/4}$ convergence rate can be viewed as a "discretized version" of the seminal result of Chen (1995), who showed that an $n^{-1/4}$ rate is unavoidable to recover mean vectors in a degenerate Gaussian mixture model with respect to the Wasserstein's distance. Difference exists, however, as topic models have a $K$-dimensional mixing vector $h$ for each observation and are therefore technically not finite mixture models. Ho and Nguyen (2016) proposed a general algebraic statistics framework for singular finite-mixture models, and showed that the optimal convergence rate for skewed-normal mixtures is $n^{-1/12}$. More generally, singular learning theory is studied in (Watanabe, 2009, 2013), and the algebraic structures of Gaussian mixture/graphical models and structural equation models are explored in (Leung, Drton and Hara, 2016; Drton, Foygel and Sullivant, 2011; Drton, 2016).

### *1.2. Limitations and future directions*

We state some limitations of this work and bring up important future directions. In this paper the vocabulary size $V$ and the number of topics $K$ are treated as fixed constants and their dependency in the asymptotic convergence rate is omitted. In practice, however, $V$ and $K$ could be large and understanding the (optimal) dependency of these parameters is important. We consider this as a *high-dimensional* version of the topic modeling problem, whose convergence rate remains largely unexplored in the literature.

Our results, similar to existing works of Anandkumar et al. (2012, 2014), are derived under a "fixed $m$" setting. In fact, the convergence rates remain nearly unchanged by uniformly sampling 2 or 3 words per document, and it is not clear how longer documents could help estimation of the underlying topic vectors under our framework. In contrast, the posterior contraction results in (Tang et al., 2014; Nguyen, 2015) are only valid under the "$m$ increasing" setting. We conjecture that the actual behavior of the ML estimator should be a combination of both perspectives: $m \geq 2$ and $n \to \infty$ are sufficient for consistent estimation, and $m$ growing with $n$ should deliver faster convergence rates.

A more general setting is when the number of topics $K$ or even the mixing prior $\nu$ is unknown. In such cases, the problem of estimating $\{\theta_k\}_{k=1}^K$ is ill-posed because one may split each $\theta_k$ into two identical vectors (thus doubling the number of topics $K$) without affecting the probabilistic model governing the generation of documents. In such settings, an alternative target of inference would be $\mathbb{E}p(h) = \mathbb{E}_{\nu_0} p(h)$, where $p(h) = \sum_{k=1}^K h_k\theta_k$ is the word distribution conditioned on $h$.

Finally, the ML estimator for the topic modeling problem is well-known to be computationally challenging, and computationally tractable alternatives such as tensor decomposition and/or non-negative matrix factorization are usually employed. In light of this paper, it is an interesting question to design *computationally efficient* methods that attain the $n^{-1/4}$ convergence rate without assuming separability or non-degeneracy conditions on the underlying topic distribution vectors.

### *1.3. Additional notations*

For two distributions $P$ and $Q$, we write $d_{\mathsf{TV}}(P;Q) = \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q| = \sup_A |P(A) - Q(A)|$ as the total variation distance between $P$ and $Q$, and $\mathrm{KL}(P\|Q) = \int \log \frac{\mathrm{d}P}{\mathrm{d}Q}\mathrm{d}P$ as the Kullback-Leibler (KL) divergence between $P$ and $Q$. For a sequence of random variables $\{A_n\}$, we write $A_n = O_{\mathbb{P}}(a_n)$ if for any $\delta \in (0,1)$, there exists a constant $C > 0$ such that $\limsup_{n\to\infty} \Pr[|A_n/a_n| > C] \le \delta$.

## 2. Assumptions, identifiability, and target of inference

### *2.1. Assumptions*

We make the following regularity assumptions on $\boldsymbol{\theta}$ and $\nu_0$:

(A1) There exists constant $c_0 > 0$ such that $\theta_j(\ell) > c_0$ for all $j \in [K]$ and $\ell \in [V]$;

(A2) $\nu_0$ is exchangeable, meaning that $\nu_0(\mathcal{A}) = \nu_0(\pi(\mathcal{A}))$ for any permutation $\pi : [K] \to [K]$; furthermore, $\mathbb{E}_{\nu_0}[h_1^2] > \mathbb{E}_{\nu_0}[h_1 h_2]$ for $K \ge 2$ and $\mathbb{E}_{\nu_0}[h_1^3] + 2\mathbb{E}_{\nu_0}[h_1 h_2 h_3] > 3\mathbb{E}_{\nu_0}[h_1^2 h_2]$ for $K \ge 3$.

The assumption (A2) only concerns the mixing distribution $\nu_0$ which is known a priori, and is satisfied by "typical" priors of $h$, such as Dirichlet distributions and the "finite mixture" prior $p_{\nu_0}(h = e_k) = 1/K$, $\forall k \in [K]$. We also remark that if $\nu_0$ is exchangeable and "non-degenerate", meaning that $\nu_0(\mathcal{A}) > 0$ for any $\mathcal{A}$ with positive Lebesgue measure, then (A2) is implied.

Condition (A1) assumes that all topic vectors $\{\theta_j\}_{j=1}^K$ in the underlying parameter $\boldsymbol{\theta}$ lie on the interior of the $V$-dimensional probabilistic simplex $\Delta^{V-1}$. This is a technical condition, which can be viewed as an analogue of the "support condition" in classical analysis of MLE where parameters in the considered parameter set $\Theta = \{\boldsymbol{\theta}\}$ give rises to the same support on observables. If (A1) is violated, then different parameterization $\boldsymbol{\theta}$ might lead to different support of observables, posing technical difficulties for our analysis. More specifically, Proposition 4 will no longer hold as $p_{\boldsymbol{\theta}}(x)$ could be arbitrarily small. We also remark that (A1) is a well-received technical condition in previous works (Nguyen, 2015; Tang et al., 2014) on convergence rates of admixture models. We use $\Theta_{c_0}$ to denote all parameters $\boldsymbol{\theta}$ that satisfies (A1).

Suppose $X_1, \cdots, X_n \in [V]^m$ are $n$ documents i.i.d. sampled from Model (1), each with $m$ words. Let

$$p_{\boldsymbol{\theta},m}(X_i) = \int_{\Delta^{K-1}} \prod_{j=1}^{m} p_{\boldsymbol{\theta},h}(X_{ij})\mathrm{d}\nu_0(h) \tag{4}$$

be the likelihood contribution of $X_i$ with respect to parameter $\boldsymbol{\theta}$, where $p_{\boldsymbol{\theta},h}(x) = \sum_{j=1}^K h_j\theta_j(x)$. Alternatively, we also write $p_{\boldsymbol{\theta},m}(X_i) = \mathbb{E}_h[p_{\boldsymbol{\theta},h}(x)]$ where $p_{\boldsymbol{\theta},h}(x) = \prod_{j=1}^m p_{\boldsymbol{\theta},h}(X_{ij})$.

### 2.2. Finite identifiability

Our target of inference is the underlying topic vectors $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$. In the classical theory of statistical estimation, one necessary condition to consistently estimate $\boldsymbol{\theta}$ from empirical observations $\{X_i\}_{i=1}^n$ is the *identifiability* of $\boldsymbol{\theta}$, loosely meaning that different parameter in the parameter space gives rises to different distributions on the observables.

**Definition 1** (exact/classical identifiability)**.** *A distribution class* $\{p_\theta\}_{\theta\in\Theta}$ *is* identifiable *with respect to* $\Theta$ *if for any* $\theta, \theta' \in \Theta$, $d_{\mathsf{TV}}(p_\theta; p_{\theta'}) = 0$ *implies* $\theta = \theta'$.

In the context of mixture models, the classical notion of identifiability is usually too strong to hold. For example, in most cases $\theta_1, \cdots, \theta_K$ can only be estimated up to permutations, provided that $\nu_0$ is exchangeable. This motivates us to consider a weaker notion of identifiability, which we term as "finite identifiability":

**Definition 2** (finite identifiability)**.** *A distribution class* $\{p_\theta\}_{\theta\in\Theta}$ *is* finitely identifiable *with respect to* $\Theta$ *if for any* $\theta \in \Theta$, $|\{\theta' \in \Theta : d_{\mathsf{TV}}(p_\theta; p_{\theta'}) = 0\}| < \infty$.

Finite identifiability is weaker than the classical/exact notion of identifiability in the sense that two different parameterization $\theta, \theta' \in \Theta$ is allowed to have the same observable distributions (almost everywhere), making them indistinguishable from any statistical procedures. On the other hand, finite identifiability is sufficiently strong that non-trivial convergence can be studied for any infinite parameter space $\Theta$.

*Example* 1. If $d_{\mathsf{TV}}(p_\theta; p_{\theta'}) = 0$ implies $d_{\mathsf{W}}(\theta, \theta') = 0$ then $\{p_\theta\}$ is finitely identifiable. This includes a wide range of convergence results for finite mixture models (Chen, 1995; Hsu and Kakade, 2013; Ge, Huang and Kakade, 2015), in which the underlying parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ can be consistently estimated up to permutations.

### 2.3. Identifiability in topic models

For topic models, the following lemma characterizes the finite identifiability of $\boldsymbol{\theta}$ under various settings of $K$ and $m$.

**Lemma 1.** *Suppose there are* $K \geq 2$ *topics and* $m \geq 1$ *words per document.*

1. *If* $m = 1$*, then* $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ is not *finitely identifiable;*
2. *If* $m \geq 2$*, then* $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ is *finitely identifiable.*

The non-identifiability of $\boldsymbol{\theta}$ for the $m = 1$ case is easy to see, because with one word per document the distribution of each document is determined by $K^{-1}\sum_{k=1}^K \theta_k$, to which an uncountable number of parameterizations $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ exist. The finite identifiability of $\boldsymbol{\theta}$ when $m \geq 2$ words are available per document, on the other hand, is much more involved and involves connections between total-variation distance between distributions and the "order of

Table 1

*Marginal distributions of $m = 2$ words per document, parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ in Eq. (5). To calculate the marginal probability of $(j_1, j_2) \in \{\texttt{apple}, \texttt{pear}\}^2$, simply evaluate $\mathbb{E}_{\nu_0}[h_1\theta_1(j_1)\theta_1(j_2) + h_2\theta_2(j_1)\theta_2(j_2)] = \int_0^1 [x\theta_1(j_1)\theta_1(j_2) + (1-x)\theta_2(j_1)\theta_2(j_2)]\mathrm{d}x$.*

| | (apple, apple) | (apple, pear) | (pear, pear) |
|---|---|---|---|
| $\boldsymbol{\theta} = (\theta_1, \theta_2)$ | .25 | .50 | .25 |
| $\boldsymbol{\theta}' = (\theta_1', \theta_2')$ | .26 | .48 | .26 |

degeneracy" of $\boldsymbol{\theta}$ (an important concept that will be introduced in the next section). To make our presentation clean we defer the complete proof of Lemma 1 to Sec. 5.1.

Lemma 1 shows that inference of $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ is possible (in a finite identifiability sense) if and only if each document consists of at least two words. We shall thus make the following assumption throughout the rest of this paper to make the target of inference sound.

(A3) $m \geq 2$.

To better illustrate the finite identifiability of $\boldsymbol{\theta}$, we consider a toy model with $V = 2$ words, $K = 2$ topics and $m = 2$ words per document. Two potential parameterizations $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and $\boldsymbol{\theta}' = (\theta_1', \theta_2')$ are considered:

$$\boldsymbol{\theta} : \begin{cases} \theta_1 = (0.5, 0.5); \\ \theta_2 = (0.5, 0.5). \end{cases} \qquad \boldsymbol{\theta}' : \begin{cases} \theta_1 = (0.6, 0.4); \\ \theta_2 = (0.4, 0.6). \end{cases} \tag{5}$$

The mixing distribution $\nu_0$ is taken to be the Dirichlet distribution $\mathrm{Dir}(1, 1)$, meaning that for $h \sim \nu_0$, $h_1$ follows the uniform distribution on $[0, 1]$ and $h_2 = 1 - h_1$.

For notational simplicity, we use "apple" and "pear" to denote the two words in the vocabulary. Because both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ have the same "average" marginal distribution $\overline{\theta} = 0.5(\theta_1 + \theta_2) = (0.5, 0.5)$, the marginal word distribution with *one* word per document is $(\texttt{apple} : 0.5, \texttt{pear} : 0.5)$ for *both* $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, meaning that $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are not distinguishable with $m = 1$ word per document. On the other hand, with $m = 2$ words per document, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ yield different marginal distributions (see Table 1) despite the fact that they have the same average distribution. This is consistent with Lemma 1.

## 3. Order of degeneracy

We introduce a concept which we name the *order of degeneracy*, which is later used to establish finite identifiability of $\boldsymbol{\theta}$ (Lemma 1) and also to characterize the optimal local convergence rates of latent topic models.

**Definition 3** (Order of degeneracy)**.** *Let $\mathcal{X} = [V]$ be the vocabulary set and $\mu$ be the counting measure on $\mathcal{X}$. Let $\mathcal{X}^m = [V]^m$ be the product space of $\mathcal{X}$ and $\mu^m$ be the product measure of $\mu$. For any $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K) \subseteq \Delta^{V-1}$ and $1 \leq \mathfrak{p} \leq m$, the* $\mathfrak{p}$th-order degeneracy criterion $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta})$ *is defined as*

$$\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) := \inf_{\|\boldsymbol{\delta}\|_1=1,\sum_\ell \delta_j(\ell)=0} \int_{\mathcal{X}^m} \left|\mathbb{E}_h p_{\boldsymbol{\theta},h}(x) \sum_{1\leq i_1<\cdots<i_{\mathfrak{p}}\leq m} \frac{\boldsymbol{\delta}_h(x_{i_1})\cdots\boldsymbol{\delta}_h(x_{i_{\mathfrak{p}}})}{p_{\boldsymbol{\theta},h}(x_{i_1})\cdots p_{\boldsymbol{\theta},h}(x_{i_{\mathfrak{p}}})}\right| \mathrm{d}\mu_m(x), \quad (6)$$

*where* $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_K) \in \mathbb{R}^V$, $\|\boldsymbol{\delta}\|_1 := \sum_{k=1}^K \|\delta_k\|_1$ *and* $\boldsymbol{\delta}_h(x) = \sum_{k=1}^K h_k\delta_k(x)$.

The definition of $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta})$ arises from a Taylor expansion of the likelihood function at neighboring parameters $p_{\boldsymbol{\theta}',m}(x)-p_{\boldsymbol{\theta},m}(x)$, which is given in Eq. (13). While Eq. (6) appears complicated, for the purpose of convergence rates it suffices to check whether $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) > 0$ or $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) = 0$, and the exact values of $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta})$ are not important. We thus define

$$\mathfrak{p}(m;\boldsymbol{\theta}) := \min\left\{\mathfrak{p} \in \mathbb{Z}^+ : \mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) > 0\right\} \quad (7)$$

as the smallest positive integer such that $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) > 0$. (If $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) = 0$ for all $1 \leq \mathfrak{p} \leq m$ then define $\mathfrak{p}(m;\boldsymbol{\theta}) := \infty$.) The quantity $\mathfrak{p}(m;\boldsymbol{\theta})$ will be used exclusively in Theorem 1 in the next section, establishing upper and local lower bounds on the convergence rates of $\boldsymbol{\theta}$. Intuitively, the smaller $\mathfrak{p}(m;\boldsymbol{\theta})$ is, the faster an estimator converges to $\boldsymbol{\theta}$ (or one of its finite equivalents), with the special case of $\mathfrak{p}(m;\boldsymbol{\theta}) = 1$ corresponding to the classical $n^{-1/2}$ convergence rate for regular parametric models.

We next give some additional results regarding $\mathfrak{p}(m;\boldsymbol{\theta})$. We show that under assumptions (A1) through (A3), it always holds that $\mathfrak{p}(m;\boldsymbol{\theta}) \leq 2$ regardless of the number of words per document (provided that $m \geq 2$, i.e., (A3)) and the underlying parameter $\boldsymbol{\theta}$. This is shown in Lemma 5, which essentially implies finite identifiability and a general $n^{-1/4}$ convergence rate under (A1) through (A3) by Theorem 1. Furthermore, Lemma 2 shows that under additional linear independence conditions $\mathfrak{p}(m;\boldsymbol{\theta}) = 1$, yielding the classical $n^{-1/2}$ rate that is faster than $n^{-1/4}$ for general $\boldsymbol{\theta}$. We also give examples for which $\mathfrak{p}(m;\boldsymbol{\theta}) > 1$, showing that the $\mathfrak{p}(m;\boldsymbol{\theta}) \leq 2$ result in Lemma 5 cannot be improved unconditionally. Finally, we remark on how to computationally evaluate $\mathfrak{p}(m;\boldsymbol{\theta})$, even when the true $\boldsymbol{\theta}$ is unknown and only an estimate $\widehat{\boldsymbol{\theta}}$ is available.

### 3.1. *First-order identifiability*

When an underlying parameter $\boldsymbol{\theta}$ satisfies $\mathfrak{p}(m;\boldsymbol{\theta}) = 1$, we say it has *first-order identifiability*. By Theorem 1, first-order identifiability of $\boldsymbol{\theta}$ essentially implies a (local) convergence rate of $n^{-1/2}$, which is similar to convergence rates in classical parametric models (Van der Vaart, 1998). The objective of this subsection is to discuss scenarios under which first-order identifiability is present.

Our first lemma shows that, if at least $m \geq 3$ words per document are present and the underlying topic vectors $\{\theta_1, \cdots, \theta_K\}$ are *linear independent*, then first-order identifiability is guaranteed.

**Lemma 2.** *If* $\{\theta_j\}_{j=1}^K$ *are linear independent then* $\mathfrak{d}_{3,1}(\boldsymbol{\theta}) > 0$.

*Remark* 1. Lemma 2 implies that $\mathfrak{p}(3;\boldsymbol{\theta})=1$ if $\boldsymbol{\theta}$ consists of linearly independent topics. Furthermore, because $\mathfrak{p}(\cdot;\boldsymbol{\theta})$ is a monotonic function in $m$ (see Corollary 1), we have $\mathfrak{p}(m;\boldsymbol{\theta})=1$ for all $m\geq 3$.

Lemma 2 is a simple consequence of the convergence results of (Anandkumar et al., 2012, 2014) and the local minimax lower bounds established in Theorem 1 of this paper. More specifically, Anandkumar et al. (2012, 2014) explicitly constructed method-of-moments estimators that attain $n^{-1/2}$ convergence rate for $m=3$ and linearly independent $\boldsymbol{\theta}$, which would violate the local minimax lower bound in Theorem 1 if $\mathfrak{p}(3;\boldsymbol{\theta})>1$. A complete proof of Lemma 2 is given in Sec. 5.3.

Lemma 2, as well as the results of Anandkumar et al. (2012, 2014), require two conditions: that $\{\theta_j\}_{j=1}^K$ being linear independent, and that $m\geq 3$, meaning that there are at least 3 words per document. It is an interesting question whether both conditions are necessary to ensure first-order identifiability. We give partial answers to this question in the following two lemmas.

**Lemma 3.** *If $\theta_j=\theta_k$ for some $j\neq k$ then $\mathfrak{d}_{m,1}(\boldsymbol{\theta})=0$ for all $m\geq 2$.*

**Lemma 4.** *Suppose $\{\theta_k\}_{k=1}^K$ are distinct. Then $\mathfrak{d}_{2,1}(\boldsymbol{\theta})=0$ if and only if $K\geq 3$.*

Lemma 3 shows that, if duplicates exist in the $K$ underlying topics then $\boldsymbol{\theta}$ cannot have first-order identifiability, regardless of how many words are present in each document. It is proved by a careful construction of $\boldsymbol{\delta}=(\delta_1,\cdots,\delta_K)$ such that the contribution of $\delta_j$ cancels out $\delta_k$ on all $x\in\mathcal{X}^m$, exploiting the condition that $\theta_j(v)=\theta_k(v)$ for all $v\in\mathcal{X}$. A complete proof of Lemma 3 is given in Sec. 5.4.

Lemma 4 studies the first-order identifiability of $\boldsymbol{\theta}$ from a different perspective. The "IF" part of Lemma 4 shows that, as long as $K\geq 3$ topics are present, merely having $m=2$ words per document cannot lead to first-order identifiability. We prove this by constructing the $\boldsymbol{\delta}=(\delta_1,\cdots,\delta_K)$ vectors as $\delta_1\propto\theta_2-\theta_3$, $\delta_2\propto\theta_3-\theta_1$, $\delta_3\propto\theta_1-\theta_2$ and showing that $\delta_1,\delta_2,\delta_3$ cancel out each other if only $m=2$ words are present in each document. On the other hand, the "ONLY IF" part of Lemma 4 is more intriguing, which states that $m=2$ words per document is sufficient for first-order identifiability if only two distinct topic vectors are to be estimated. The proof of the only if part is however much more complicated, involving analytically verifying the full-rankness of a coefficient matrix. A complete proof of Lemma 4 is given in Sec. 5.5.

While Lemmas 3 and 4 combined show the necessity of $m\geq 3$ and additional non-degeneracy condition in Lemma 2, we remark that Lemmas 2, 3 and 4 do not cover all cases of $\boldsymbol{\theta}$ in the parameter space. One notable exception is when $m\geq 3$, $K\geq 3$ and $\{\theta_k\}_{k=1}^K$ are distinct but not linearly independent, for which none of the three lemmas apply and whether such parameterization satisfies first-order identifiability remains an open question. Nevertheless, in Sec. 3.3 we give a computational routine that determines whether $\mathfrak{p}(m;\boldsymbol{\theta})=1$ or $\mathfrak{p}(m;\boldsymbol{\theta})>1$ using any consistent estimates $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, which nicely complements the analytical results in Lemmas 2, 3 and 4.

### *3.2. Second-order identifiability*

When an underlying parameter $\boldsymbol{\theta}$ satisfies $\mathfrak{p}(m;\boldsymbol{\theta}) \leq 2$, we say it has *second-order identifiability*. By definition, if $\boldsymbol{\theta}$ satisfies first-order identifiability then it also satisfies second-order identifiability, but the reverse statement is generally not true. Hence, second-order identifiability is weaker than its first-order counterparts, which also suggests potentially slower rates of convergence in parameter estimation.

In this section we show, perhaps surprisingly, that *all* parameterization $\boldsymbol{\theta}$ have second-order identifiability under (A1) through (A3).

**Lemma 5.** *For all $\boldsymbol{\theta}$, $\mathfrak{d}(2,2)(\boldsymbol{\theta}) \geq c(\nu_0)/V^3K > 0$, where $c(\nu_0) := \mathbb{E}_{\nu_0}[h_1^2 - h_1h_2] > 0$ is a positive constant only depending on $\nu_0$.*

*Remark* 2. Lemma 5 implies that $\mathfrak{p}(2;\boldsymbol{\theta}) \leq 2$ for all $\boldsymbol{\theta}$ satisfying (A1) and (A2). By monotonicity of $\mathfrak{p}(\cdot;\boldsymbol{\theta})$ (see Corollary 1), we also have $\mathfrak{p}(m;\boldsymbol{\theta}) \leq 2$ for all $m \geq 2$.

While appears surprising, the proof of Lemma 5 is actually quite simple. The key observation is the existence of documents consisting of identical words (i.e., $x = (x_1, x_2)$ where $x_1 = x_2$), on which the $\boldsymbol{\delta}_h(x_1)\boldsymbol{\delta}_h(x_2)$ term becomes a square and equals zero only if $\boldsymbol{\delta} = 0$. A complete proof of Lemma 5 is given in Sec. 5.6.

Lemma 5 shows that, for any underlying parameter $\boldsymbol{\theta}$, if there are at least 2 words per document then $\mathfrak{p}(m;\boldsymbol{\theta}) \leq 2$. This also suggests a general $n^{-1/4}$ convergence rate of an ML estimate of $\boldsymbol{\theta}$, by Theorem 1. This conclusion holds even for the "over-complete" setting $K \geq V$, under which existing works require particularly strong prior knowledge on $\boldsymbol{\theta}$ (e.g., $\{\theta_j\}_{j=1}^K$ being i.i.d. sampled uniformly from the $V$-dimensional probabilistic simplex) for (computationally tractable) consistent estimation (Anandkumar, Ge and Janzamin, 2017; Ma, Shi and Steurer, 2016).

### *3.3. Numerical checking of $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) > 0$*

As we remarked in previous sections, Lemmas 2, 3 and 4 do not cover all cases, and there are parameters $\boldsymbol{\theta}$ whose order of degeneracy is not determined by the above lemmas. In addition, in practical applications it might be desirable to compute the order of degeneracy with only an estimate $\widehat{\boldsymbol{\theta}}$ of the underlying parameter $\boldsymbol{\theta}$. In this section we present numerical procedures that decides whether $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) > 0$. We also show that the calculation can be carried out on estimates $\widehat{\boldsymbol{\theta}}$ and show its asymptotic consistency for the special case of $\mathfrak{p} = 1$.

**Proposition 1.** *For any $\boldsymbol{\theta}$, $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) > 0$ if and only if the following polynomial system in $\{\delta_{jk}\}$, $j \in [K]$, $k \in [V]$ does not have non-zero solutions:*

$$\sum_{1\leq i_1<\cdots<i_{\mathfrak{p}}\leq m}\ \sum_{j_1,\cdots,j_{\mathfrak{p}}=1}^{K} \xi(\boldsymbol{i},\boldsymbol{j};\boldsymbol{\theta},x)\prod_{\ell=1}^{\mathfrak{p}} \delta_{j_\ell,x_{i_\ell}} = 0, \quad \forall x = (x_1,\cdots,x_m) \in [V]^m;$$

TABLE 2

*Numerical estimations of the $\ell_1$-condition number $\kappa_1(\mathbf{A}(\boldsymbol{\theta})) := \|\mathbf{A}(\boldsymbol{\theta})\|_1 \|\mathbf{A}(\boldsymbol{\theta})^{-1}\|_1$ for different $V, K, m$ and $\boldsymbol{\theta}$. The numerical estimation procedure of $\kappa_1(\mathbf{A}(\boldsymbol{\theta}))$ was given in (Hager, 1984) and adopted in MATLAB's `condest` routine. Each entry in the topic vectors are i.i.d. generated from $U[0,1]$ and then normalized so that $\|\theta_k\|_1 = 1$ for all $k \in [K]$.*

| | $V$ | $K$ | $M$ | $\kappa_1(\mathbf{A}(\boldsymbol{\theta}))$ | $\mathfrak{p}(m;\boldsymbol{\theta})$ |
|---|---|---|---|---|---|
| linear independent $\{\theta_k\}$ | 10 | 3 | 3 | $2.9 \times 10^4$ | $= 1$ |
| linear independent $\{\theta_k\}$ | 10 | 3 | 2 | $1.1 \times 10^{19}$ | $> 1$ |
| linear independent $\{\theta_k\}$ | 10 | 2 | 2 | $8.0 \times 10^2$ | $= 1$ |
| $\theta_1 = \theta_2$ | 10 | 2 | 3 | $6.0 \times 10^{17}$ | $> 1$ |
| $\theta_1 = \theta_2 \neq \theta_3$ | 10 | 3 | 4 | $2.1 \times 10^{18}$ | $> 1$ |
| $\theta_3 = 0.5(\theta_1 + \theta_2)$ | 10 | 3 | 3 | $9.7 \times 10^4$ | $= 1$ |
| $\theta_3 = 0.8\theta_1 + 0.2\theta_2$ | 10 | 3 | 3 | $4.7 \times 10^5$ | $= 1$ |

$$\sum_{k=1}^{V} \delta_{jk} = 0, \qquad \forall j \in [K].$$

*Here the coefficients $\xi(\boldsymbol{i}, \boldsymbol{j}; \boldsymbol{\theta}, x)$ is defined as*

$$\xi(\boldsymbol{i}, \boldsymbol{j}; \boldsymbol{\theta}, x) := \mathbb{E}_h \left[ \prod_{i \notin \{i_1, \cdots, i_{\mathfrak{p}}\}} p_{\boldsymbol{\theta},h}(x_i) \prod_{\ell=1}^{\mathfrak{p}} h_{j_\ell} \right].$$

*Proof.* Because $\mu_m$ in the definition of $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta})$ is a counting measure, $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) = 0$ if and only if all terms within the integral in Eq. (6) are zero. This gives the proposition. □

With Proposition 1, $\mathfrak{p}(m;\boldsymbol{\theta})$ can be determined by enumerating from $\mathfrak{p} = 1$ to $\mathfrak{p} = m$ and recording the smallest $\mathfrak{p}$ such that $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) > 0$, which is the smallest $\mathfrak{p}$ such that the polynomial system in Proposition 1 does not have non-zero solutions.

The polynomial system in Proposition 1 has maximum degree $\mathfrak{p}$. In principle, whether such a polynomial system admits non-zero solutions can be decided by converting the system under the *Gröbner basis* and apply results from computational algebraic geometry. Such an approach is however technically very complicated, and soon becomes computationally intractable when $V$ is large.

Fortunately, our result in Lemma 5 shows that $\mathfrak{p}(m;\boldsymbol{\theta}) \leq 2$ under very mild conditions. More specifically, as long as each document consists of at least $m \geq 2$ words, the task of determining $\mathfrak{p}(m;\boldsymbol{\theta})$ reduces to checking whether $\mathfrak{d}_{m,1}(\boldsymbol{\theta}) > 0$ only, as $\mathfrak{p}(m;\boldsymbol{\theta}) \leq 2$ is always correct. Furthermore, to decide whether $\mathfrak{d}_{m,1}(\boldsymbol{\theta}) > 0$ the polynomial system in Proposition 1 reduces to a *linear system*, whose existence of non-trivial solutions is easily determined by the *rank* of its design matrix. The following proposition formalizes the above discussion.

**Proposition 2.** $\mathfrak{d}_{m,1}(\boldsymbol{\theta}) > 0$ *if and only if the following linear system does not have non-zero solutions:*

$$\sum_{i=1}^{m} \sum_{j=1}^{K} \xi(i, j; \boldsymbol{\theta}, x) \delta_{j,x_i} = 0, \qquad \forall x = (x_1, \cdots, x_m) \in [V]^m;$$

$$\sum_{k=1}^{V} \delta_{jk} = 0, \qquad \forall j \in [K];$$

*where* $\xi(i, j; \boldsymbol{\theta}, x) := \mathbb{E}_h[h_j \prod_{i' \neq i} p_{\boldsymbol{\theta},h}(x_{i'})]$.

*Proof.* Immediately follows Proposition 1. □

Proposition 2 constructs a *linear* system with $VK$ variables and $(V^m + K)$ equations. The existence of a non-trivial (non-zero) solution can be determined by explicitly constructing the $(V^m + K) \times VK$ matrix $\mathbf{A}$ in the equation $\mathbf{A}\mathrm{vec}(\{\delta_k\}) = \mathbf{0}$ and checking whether $\mathbf{A}$ has full column rank.

We give in Table 2 some computational results of $\mathfrak{p}(m; \boldsymbol{\theta})$ for some representative $\boldsymbol{\theta}$ settings. Due to physical constraints of numerical precision, we use the $\ell_1$-condition number $\kappa_1(\mathbf{A}(\boldsymbol{\theta}))$ as an indication of whether $\mathbf{A}(\boldsymbol{\theta})$ has full column rank, where a large condition number suggests that $\mathbf{A}(\boldsymbol{\theta})$ is rank-deficient. The first 6 lines in Table 2 verify our results in Lemmas 2, 3 and 4. The last 2 lines in Table 2 provide additional information regarding the first-order identifiability of linearly dependent but distinct topic vectors $\{\theta_k\}$. They show that $\{\theta_k\}_k$ is first-order identifiable (i.e., $\mathfrak{p}(m; \boldsymbol{\theta}) = 1$) even if $\{\theta_k\}_k$ are linear dependent, provided that they are distinct and $m \geq 3$. It remains an open question to formally establish such first-order identifiability for distinct but linear dependent topics.

In practice, the underlying $\boldsymbol{\theta}$ is unknown and only an estimate $\widehat{\boldsymbol{\theta}}$ is available. The following proposition shows that the procedure of checking whether $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) > 0$ remains valid asymptotically if one replaces $\boldsymbol{\theta}$ with $\widehat{\boldsymbol{\theta}}$.

**Proposition 3.** *Let* $\mathbf{A}(\boldsymbol{\theta})$ *and* $\mathbf{A}(\widehat{\boldsymbol{\theta}})$ *be the* $(V^m+K)\times VK$ *matrices constructed using* $\boldsymbol{\theta}$ *and* $\widehat{\boldsymbol{\theta}}$*, respectively. Let* $\sigma_{\min}(\mathbf{A}(\boldsymbol{\theta}))$ *and* $\sigma_{\min}(\mathbf{A}(\widehat{\boldsymbol{\theta}}))$ *be the smallest singular values of* $\mathbf{A}(\boldsymbol{\theta})$ *and* $\mathbf{A}(\widehat{\boldsymbol{\theta}})$*. If* $d_{\mathsf{W}}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \xrightarrow{p} 0$ *then* $\sigma_{\min}(\mathbf{A}(\widehat{\boldsymbol{\theta}})) \xrightarrow{p} \sigma_{\min}(\mathbf{A}(\boldsymbol{\theta}))$.

*Proof.* By Weyl's inequality we know that $|\sigma_{\min}(\mathbf{A}(\widehat{\boldsymbol{\theta}})) - \sigma_{\min}(\mathbf{A}(\boldsymbol{\theta}))| \leq \|\mathbf{A}(\widehat{\boldsymbol{\theta}}) - \mathbf{A}(\boldsymbol{\theta})\|_{\mathrm{op}}$. It is easy to verify that $[\mathbf{A}(\widehat{\boldsymbol{\theta}})]_{ij} \xrightarrow{p} [\mathbf{A}(\boldsymbol{\theta})]_{ij}$ for all $i, j$ provided that $d_{\mathsf{W}}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \xrightarrow{p} 0$, because the coefficients are invariant under permutation $\pi : [K] \to [K]$ thanks to (A2). We then have $\|\mathbf{A}(\widehat{\boldsymbol{\theta}}) - \mathbf{A}(\boldsymbol{\theta})\|_{\mathrm{op}} \xrightarrow{p} 0$ because $\mathbf{A}(\cdot)$ are finite-dimensional matrices. □

Proposition shows that by substituting $\boldsymbol{\theta}$ with a consistent estimator $\widehat{\boldsymbol{\theta}}$ in the construction of the $(V^m + K) \times VK$ coefficient matrix $\mathbf{A}$ and comparing the least singular value of $\mathbf{A}$ with a small number that slowly grows to zero, we can decide consistently whether $\mathfrak{d}_{m,1}(\boldsymbol{\theta}) > 0$ using only $\widehat{\boldsymbol{\theta}}$.

## 4. Maximum likelihood estimation and its convergence rate

We consider the *Maximum Likelihood (ML)* estimator, defined as

$$\widehat{\boldsymbol{\theta}}^{\mathsf{ML}}_{n,m} \in \arg\max_{\boldsymbol{\theta}\in\Theta_{c_0}} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta},m}(X_i)$$

$$= \arg\max_{\boldsymbol{\theta}\in\Theta_{c_0}} \sum_{i=1}^{n} \log\left\{\int_{\Delta^{K-1}}\left[\prod_{j=1}^{m}\sum_{k=1}^{K} h_k\theta_k(x_{ij})\right]\mathrm{d}\nu_0(h)\right\}. \tag{8}$$

It should be noted that $\widehat{\boldsymbol{\theta}}^{\mathsf{ML}}_{n,m}$ is constrained to the parameter set $\Theta_{c_0}$, which is assumed to be known a priori.

The next theorem is the main result of this paper, which derives the convergence rate of the ML estimator $\widehat{\boldsymbol{\theta}}^{\mathsf{ML}}_{n,m}$ using the concept of order of degeneracy $\mathfrak{p}(m;\boldsymbol{\theta})$ developed in the previous section. We also prove that such convergence rates are locally minimax optimal and therefore cannot be improved.

**Theorem 1.** *Fix $K \geq 2$, $m \geq 2$, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K) \in \Theta_{c_0}$. Let $\widetilde{\Theta}_{c_0}(\boldsymbol{\theta}) := \{\widetilde{\boldsymbol{\theta}'} \in \Theta_{c_0} : d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}) = 0\}$ be the equivalent parameter set with respect to $\boldsymbol{\theta}$, which is finite thanks to Lemma 1. Let $\mathfrak{p}(m;\boldsymbol{\theta})$ be defined as in Eq. (13), and suppose $\mathfrak{p}(m;\boldsymbol{\theta}) < \infty$.*

1. (Global convergence rate of the MLE).
$$\min_{\widetilde{\boldsymbol{\theta}}\in\widetilde{\Theta}_{c_0}(\boldsymbol{\theta})} d_{\mathsf{W}}(\widetilde{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}^{\mathsf{ML}}_{n,m}) = O_{\mathbb{P}}(n^{-1/2\mathfrak{p}(m;\boldsymbol{\theta})}) \tag{9}$$
*under $p_{\boldsymbol{\theta},m}$ (or equivalently $p_{\widetilde{\boldsymbol{\theta}},m}$), where in $O_{\mathbb{P}}(\cdot)$ we hide dependency on $\nu_0, m$ and $\boldsymbol{\theta}$;*
2. (Local minimax rate). *Then there exists a constant $r_{\boldsymbol{\theta}} > 0$ depending only on $\nu_0, m$ and $\boldsymbol{\theta}$ such that*
$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}'\in\Theta_n(\boldsymbol{\theta})} \mathbb{E}_{\boldsymbol{\theta}'}\left[d_{\mathsf{W}}(\boldsymbol{\theta}', \widehat{\boldsymbol{\theta}})\right] = \Omega(n^{-1/2\mathfrak{p}(m;\boldsymbol{\theta})}), \tag{10}$$
*where $\Theta_n(\boldsymbol{\theta})$ is a shrinking neighborhood of $\boldsymbol{\theta}$ defined as $\{\boldsymbol{\theta}' \in \Theta_{c_0} : d_{\mathsf{W}}(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq r_{\boldsymbol{\theta}} \cdot n^{-1/2\mathfrak{p}(m;\boldsymbol{\theta})}\}$, and $a_n = \Omega(b_n)$ means $\lim_{n\to\infty} |a_n|/|b_n| > 0$.*

*Remark* 3. Our proof for the lower bound part of Theorem 1 actually proves the stronger statement that, for any $\boldsymbol{\theta}' \in \Theta_n(\boldsymbol{\theta})$, there exists constant $\tau > 0$ such that no procedure can distinguish $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ with success probability smaller than $\tau$, as $n \to \infty$. Note that Eq. (10) is a direct corollary of this testing lower bound by Markov's inequality.

Theorem 1 characterizes the convergence rates of MLE *locally* at parameters $\boldsymbol{\theta} \in \Theta_{c_0}$, with the convergence rates dependent on $\mathfrak{p}(m;\boldsymbol{\theta}) \in \mathbb{N}$. While convergence rates depending on $\boldsymbol{\theta}$ might seem like a weak result, we argue that such convergence is probably the best one can hope for, and the "local convergence" results still provide much valuable information about the statistical estimation problem of latent topic models. In particular, we have the following observations:

1. It is arguable that convergence rates depending on the underlying parameter $\boldsymbol{\theta}$ (or its close neighborhoods) are the best one can hope for. Because if the worst-case convergence rates are considered over all $\boldsymbol{\theta} \in \Theta_{c_0}$, by our Theorem 1 and Lemma 4 the only reasonable convergence rate is $n^{-1/4}$

which is slow; on the other hand, by restricting ourselves to "local" convergence we can hope to get much faster rates like $n^{-1/2}$ for certain parameter settings;

2. By deriving $\boldsymbol{\theta}$-specific convergence rates, we obtain more information about the structure of the statistical estimation problem in latent topic models. In particular, our results show that when topic vectors are linearly independent, the convergence rate is much faster than cases when duplicate topic vectors are present. This is an interesting observation and is largely unknown in previous research on latent topic models.
3. One common difficulty with $\boldsymbol{\theta}$-specific rates is the challenge of deriving matching lower bounds, because if $\boldsymbol{\theta}$ is known the trivial estimator of outputting $\boldsymbol{\theta}$ always has zero measure. We get around this issue by considering a "close neighborhood" of $\boldsymbol{\theta}$ and derive "local" minimax rates for any statistical procedure, which match the convergence rates of the ML estimator. Such a "local analysis" was used, for example in Van der Vaart (1998) to show the optimality of $I(\theta_0)^{-1}$ of MLE under classical settings. Our analysis, on the other hand, focuses on "local rates of convergence" as the Fisher's information $I(\theta_0)$ in our model is not necessarily invertible, under which case rates worse than $n^{-1/2}$ is unavoidable.

The upper bound on convergence rates of MLE in Theorem 1 is proved by adapting the classical analysis of (Van der Vaart, 1998) and considering higher order of binomial approximation depending on $\mathfrak{p}(m;\boldsymbol{\theta})$. The local minimax lower bound is proved by considering two hypothesis $\boldsymbol{\theta},\boldsymbol{\theta}'$ and applying the Le Cam's inequality. The $n^{-1/2\mathfrak{p}(m;\boldsymbol{\theta})}$ term arises in the upper bound of TV-distance between distributions induced by $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, which is again bounded by higher-order binomial approximations. The complete proof of Theorem 1 is given in Sec. 5.2.

## 5. Proofs

In this section we prove the main results of this paper. To simplify presentation, we use $C>0$ to denote any constant that only depends on $V,K,m,\nu_0$ and $c_0$. We also use $C_{\boldsymbol{\theta}}>0$ to denote constants that further depends on $\boldsymbol{\theta}\in\Theta_{c_0}$, the underlying parameter that generates the observed documents. Neither $C$ nor $C_{\boldsymbol{\theta}}$ will depend on the number of observations $n$.

We first state and prove a key lemma that connects the defined degeneracy criterion with the total-variation (TV) distance between measures corresponding to neighboring parameters. Some corollaries are also given after the proof of Lemma 6.

**Lemma 6.** *Suppose $\boldsymbol{\theta}\in\Theta_{c_0}$, $m\geq 2$ and $\mathfrak{p}(m;\boldsymbol{\theta})<\infty$. Then for any $0<\epsilon\leq\epsilon_0<1/2$,*

$$\inf_{\epsilon\leq d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')\leq\epsilon_0} d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m};p_{\boldsymbol{\theta}',m})\geq\left[\mathfrak{d}_{m,\mathfrak{p}(m;\boldsymbol{\theta})}(\boldsymbol{\theta})-\frac{V^m\epsilon_0}{1-\epsilon_0}\right]\cdot\epsilon^{\mathfrak{p}(m;\boldsymbol{\theta})};\tag{11}$$

$$\sup_{d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')\leq\epsilon} d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m};p_{\boldsymbol{\theta}',m})\leq\frac{V^m}{1-\epsilon}\cdot\epsilon^{\mathfrak{p}(m;\boldsymbol{\theta})}.\tag{12}$$

*Proof.* We first prove Eq. (11). Let $\widetilde{\boldsymbol{\delta}} = \boldsymbol{\theta}'_\pi - \boldsymbol{\theta}$ under appropriate permutation $\pi : [K] \to [K]$ such that $\widetilde{\epsilon} := \|\widetilde{\boldsymbol{\delta}}\|_1 = d_{\mathsf{W}}(\boldsymbol{\theta}, \boldsymbol{\theta}') \in [\epsilon, \epsilon_0]$. We then have (without loss of generality let $\pi(k) \equiv k$)

$$\begin{aligned}
&p_{\boldsymbol{\theta}',m}(x) - p_{\boldsymbol{\theta},m}(x)\\
&\quad= \mathbb{E}_h\left[p_{\boldsymbol{\theta}',h}(x) - p_{\boldsymbol{\theta},h}(x)\right]\\
&\quad= \mathbb{E}_h\left\{p_{\boldsymbol{\theta},h}(x)\left[\frac{\prod_{i=1}^m p_{\boldsymbol{\theta}',h}(x_i)}{\prod_{i=1}^m p_{\boldsymbol{\theta},h}(x_i)} - 1\right]\right\}\\
&\quad= \mathbb{E}_h\left\{p_{\boldsymbol{\theta},h}(x)\left[\prod_{i=1}^m\left(1 + \frac{p_{\boldsymbol{\theta}',h}(x_i) - p_{\boldsymbol{\theta},h}(x_i)}{p_{\boldsymbol{\theta},h}(x_i)}\right) - 1\right]\right\}\\
&\quad= \mathbb{E}_h\left\{p_{\boldsymbol{\theta},h}(x)\left[\prod_{i=1}^m\left(1 + \frac{\sum_{j=1}^K h_j\theta'_j(x_i) - \sum_{j=1}^K h_j\theta_j(x_i)}{p_{\boldsymbol{\theta},h}(x_i)}\right) - 1\right]\right\}\\
&\quad= \mathbb{E}_h\left\{p_{\boldsymbol{\theta},h}(x)\left[\prod_{i=1}^m\left(1 + \frac{\widetilde{\boldsymbol{\delta}}_h(x_i)}{p_{\boldsymbol{\theta},h}(x_i)}\right) - 1\right]\right\}\\
&\quad=: \sum_{\mathfrak{p}'=1}^m r_{\mathfrak{p}'}(x),
\end{aligned}\tag{13}$$

where $\widetilde{\boldsymbol{\delta}}_h(x_i) = \sum_{j=1}^k h_j\delta_j(x_i)$, $\delta_j(x_i) = \theta'_j(x_i) - \theta_j(x_i)$ and

$$r_{\mathfrak{p}'}(x) := \mathbb{E}_h\left\{p_{\boldsymbol{\theta},h}(x)\sum_{1\le i_1<\cdots<i_{\mathfrak{p}'}\le m}\frac{\widetilde{\boldsymbol{\delta}}_h(x_{i_1})\cdots\widetilde{\boldsymbol{\delta}}_h(x_{i_{\mathfrak{p}'}})}{p_{\boldsymbol{\theta},h}(x_{i_1})\cdots p_{\boldsymbol{\theta},h}(x_{i_{\mathfrak{p}'}})}\right\}.$$

By definition of $\mathfrak{p}(m;\boldsymbol{\theta})$ and $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta})$ we know that $r_{\mathfrak{p}'}(x') = 0$ for all $1 \le \mathfrak{p}' < \mathfrak{p}(m;\boldsymbol{\theta})$ and $x' \in \mathcal{X}^m$; therefore

$$\int_{\mathcal{X}^m}|r_{\mathfrak{p}'}(x)|\mathrm{d}\mu_m(x) = 0.$$

For $\mathfrak{p}' = \mathfrak{p}(m;\boldsymbol{\theta})$, integrating over all $x \in \mathcal{X}^m$ with respect to the counting measure we have

$$\begin{aligned}
&\int_{\mathcal{X}^m}|r_{\mathfrak{p}'}(x)|\mathrm{d}\mu_m(x)\\
&= \int_{\mathcal{X}^m}\left|\mathbb{E}_h\left\{p_{\boldsymbol{\theta},h}(x)\sum_{1\le i_1<\cdots<i_{\mathfrak{p}'}\le m}\frac{\widetilde{\boldsymbol{\delta}}_h(x_{i_1})\cdots\widetilde{\boldsymbol{\delta}}_h(x_{i_{\mathfrak{p}'}})}{p_{\boldsymbol{\theta},h}(x_{i_1})\cdots p_{\boldsymbol{\theta},h}(x_{i_{\mathfrak{p}'}})}\right\}\right|\mathrm{d}\mu_m(x)\\
&= \|\widetilde{\boldsymbol{\delta}}\|_1^{\mathfrak{p}'}\cdot\int_{\mathcal{X}^m}\left|\mathbb{E}_h\left\{p_{\boldsymbol{\theta},h}(x)\sum_{1\le i_1<\cdots<i_{\mathfrak{p}'}\le m}\frac{\widetilde{\boldsymbol{\delta}}_h(x_{i_1})\cdots\widetilde{\boldsymbol{\delta}}_h(x_{i_{\mathfrak{p}'}})}{p_{\boldsymbol{\theta},h}(x_{i_1})\|\boldsymbol{\delta}\|_1\cdots p_{\boldsymbol{\theta},h}(x_{i_{\mathfrak{p}'}})\|\boldsymbol{\delta}\|_1}\right\}\right|\mathrm{d}\mu_m(x)\\
&\ge \|\widetilde{\boldsymbol{\delta}}\|_1^{\mathfrak{p}'}\cdot\inf_{\substack{\|\boldsymbol{\delta}\|_1=1\\ \sum_{\ell=1}^V\delta_j(\ell)=0}}\int_{\mathcal{X}^m}\left|\mathbb{E}_h\left\{p_{\boldsymbol{\theta},h}(x)\sum_{1\le i_1<\cdots<i_{\mathfrak{p}'}\le m}\frac{\boldsymbol{\delta}_h(x_{i_1})\cdots\boldsymbol{\delta}_h(x_{i_{\mathfrak{p}'}})}{p_{\boldsymbol{\theta},h}(x_{i_1})\cdots p_{\boldsymbol{\theta},h}(x_{i_{\mathfrak{p}'}})}\right\}\right|\mathrm{d}\mu_m(x)
\end{aligned}$$

$$= \|\widetilde{\boldsymbol{\delta}}\|_1^{\mathfrak{p}'} \cdot \mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) = \mathfrak{d}_{m,\mathfrak{p}'}(\boldsymbol{\theta})[d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')]^{\mathfrak{p}'}.$$

Here the third line holds because $\boldsymbol{\delta} := \widetilde{\boldsymbol{\delta}}/\|\widetilde{\boldsymbol{\delta}}\|_1$ satisfies $\|\boldsymbol{\delta}\|_1 = 0$ and $\sum_{\ell=1}^{V} \delta_j(\ell) = 0$. For $\mathfrak{p}(m;\boldsymbol{\theta}) \le \mathfrak{p}' \le m$ and all $x' \in \mathcal{X}^m$, it holds that

$$\begin{aligned}
&\int_{\mathcal{X}^m} |r_{\mathfrak{p}'}(x')| \mathrm{d}\mu_m(x) \\
&\quad \le \int_{\mathcal{X}^m} \left| \mathbb{E}_h \left\{ \sum_{1 \le i_1 < \cdots < i_{\mathfrak{p}'} \le m} \prod_{i' \notin \{i_1,\cdots,i_{\mathfrak{p}'}\}} p_{\boldsymbol{\theta},h}(x_{i'}) \prod_{j=1}^{\mathfrak{p}'} \widetilde{\boldsymbol{\delta}}_h(x_{i_j}) \right\} \right| \mathrm{d}\mu_m(x) \\
&\quad \le V^m \cdot \|\widetilde{\boldsymbol{\delta}}\|_1^{\mathfrak{p}'} = V^m [d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')]^{\mathfrak{p}'}.
\end{aligned}$$

Subsequently, using the fact that $d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}') \in [\epsilon, \epsilon_0]$ and $\epsilon_0 < 1/2$, we have

$$\begin{aligned}
d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; d_{\boldsymbol{\theta}',m}) &= \int_{\mathcal{X}^m} \left| \sum_{\mathfrak{p}'=1}^{m} r_{\mathfrak{p}'}(x) \right| \mathrm{d}\mu_m(x) \\
&\ge \int_{\mathcal{X}^m} |r_{\mathfrak{p}(m;\boldsymbol{\theta})}(x)| \mathrm{d}\mu_m(x) - \sum_{\mathfrak{p}'=\mathfrak{p}(m;\boldsymbol{\theta})+1}^{m} \int_{\mathcal{X}^m} |r_{\mathfrak{p}'}(x)| \mathrm{d}\mu_m(x) \\
&\ge \mathfrak{d}_{m,\mathfrak{p}(m;\boldsymbol{\theta})}(\boldsymbol{\theta})[d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')]^{\mathfrak{p}(m;\boldsymbol{\theta})} - V^m \cdot \sum_{\mathfrak{p}'=\mathfrak{p}(m;\boldsymbol{\theta})+1}^{m} [d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')]^{\mathfrak{p}'} \\
&\ge \mathfrak{d}_{m,\mathfrak{p}(m;\boldsymbol{\theta})}(\boldsymbol{\theta})[d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')]^{\mathfrak{p}(m;\boldsymbol{\theta})} \\
&\quad - \frac{V^m}{1 - d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')} \cdot [d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')]^{\mathfrak{p}(m;\boldsymbol{\theta})+1} \\
&\ge \left[ \mathfrak{d}_{m,\mathfrak{p}(m;\boldsymbol{\theta})}(\boldsymbol{\theta}) - \frac{V^m \epsilon_0}{1-\epsilon_0} \right] \cdot \epsilon^{\mathfrak{p}(m;\boldsymbol{\theta})}.
\end{aligned}$$

We next prove Eq. (12). Let again $\widetilde{\boldsymbol{\delta}} := \boldsymbol{\theta}'_\pi - \boldsymbol{\theta}$ and $\widetilde{\epsilon} := \|\widetilde{\boldsymbol{\delta}}\|_1 \le \epsilon$ for all $\boldsymbol{\theta}' \in \Theta_{c_0}$ such that $d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}') \le \epsilon$. Then

$$\begin{aligned}
d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}) &\le \sum_{\mathfrak{p}'=\mathfrak{p}(m;\boldsymbol{\theta})}^{m} \int_{\mathcal{X}^m} |r_{\mathfrak{p}'}(x)| \mathrm{d}\mu_m(x) \\
&\le V^m \sum_{\mathfrak{p}'=\mathfrak{p}(m;\boldsymbol{\theta})}^{m} \widetilde{\epsilon}^{\mathfrak{p}'} \\
&\le V^m \sum_{\mathfrak{p}'=\mathfrak{p}(m;\boldsymbol{\theta})}^{m} \epsilon^{\mathfrak{p}'} \le \frac{V^m}{1-\epsilon} \cdot \epsilon^{\mathfrak{p}(m;\boldsymbol{\theta})}. \qquad \square
\end{aligned}$$

**Corollary 1** (Monotonicity of $\mathfrak{p}(m;\boldsymbol{\theta})$). *$\mathfrak{p}(m';\boldsymbol{\theta}) \le \mathfrak{p}(m;\boldsymbol{\theta})$ for all $m' \ge m$.*

*Proof.* If $\mathfrak{p}(m;\boldsymbol{\theta}) = \infty$ then the inequality automatically holds. Suppose $\mathfrak{p}(m;\boldsymbol{\theta}) = \mathfrak{p}$ and assume by way of contradiction that $\mathfrak{p}(m';\boldsymbol{\theta}) = \mathfrak{p}' > \mathfrak{p}$ for some $m' > m$. Invoking Lemma 6 and the data processing inequality, we know that for all $0 < \epsilon < 1/4$,

$$\sup_{d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')\leq 2\epsilon} d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}) \leq \sup_{d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')\leq 2\epsilon} d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m'}; p_{\boldsymbol{\theta}',m'}) \leq \frac{V^m 2^{\mathfrak{p}'}}{1-2\epsilon}\cdot \epsilon^{\mathfrak{p}'}. \quad (14)$$

On the other hand, because $\mathfrak{p}(m;\boldsymbol{\theta}) = \mathfrak{p}$, we know that for all $0 < \epsilon \leq \epsilon_0 < 1/2$,

$$\begin{aligned}\inf_{\epsilon\leq d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')\leq \epsilon_0} d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}) &\geq \left[\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) - \frac{V^m\epsilon_0}{1-\epsilon_0}\right]\cdot \epsilon^{\mathfrak{p}} \\ &= \frac{1}{\epsilon^{\mathfrak{p}'-\mathfrak{p}}}\left[\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) - \frac{V^m\epsilon_0}{1-\epsilon_0}\right]\cdot \epsilon^{\mathfrak{p}'}. \quad (15)\end{aligned}$$

Eqs. (14) and (15) clearly contradict each other by considering $\epsilon_0 > 0$ moderately small such that $\mathfrak{d}_{m,\mathfrak{p}}(\boldsymbol{\theta}) \geq 2V^m\epsilon_0/(1-\epsilon_0)$, and $\boldsymbol{\theta}'$ sufficiently close to $\boldsymbol{\theta}$ such that such that $\epsilon \leq d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}') \leq 2\epsilon$ and letting $\epsilon \to 0^+$. Thus, we conclude that $\mathfrak{p}(m';\boldsymbol{\theta}) \leq \mathfrak{p}(m;\boldsymbol{\theta})$. □

By Lemma 5 and Corollary 1, we immediately have the following claim:

**Corollary 2** (Finiteness of $\mathfrak{p}(m;\boldsymbol{\theta})$)**.** *For any $\boldsymbol{\theta} \in \Theta_{c_0}$ and $m \geq 2$, $\mathfrak{p}(m;\boldsymbol{\theta}) \leq 2$.*

### *5.1. Proof of Lemma 1*

We first show if $K \geq 2$ and $m = 1$ then $\boldsymbol{\theta}$ is not finitely identifiable. In this case, any parameterization $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ with the same "average" word distribution $\overline{\theta} = \frac{1}{K}\sum_{k=1}^K \theta_k$ yields the same distribution of documents, and for any $\boldsymbol{\theta}$ there are infinitely many $\boldsymbol{\theta}'$ that matches exactly its average distribution $\overline{\theta}$. Therefore $\boldsymbol{\theta}$ is not finitely identifiable.

We next show the finite identifiability of $\boldsymbol{\theta}$ when $K \geq 2$ and $m \geq 2$. By data processing inequality we know that $d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}) \geq d_{\mathsf{TV}}(p_{\boldsymbol{\theta},2}; p_{\boldsymbol{\theta}',2})$ for $m \geq 2$. Therefore, we only need to prove finite identifiability for $\{p_{\boldsymbol{\theta},2}\}_{\boldsymbol{\theta}\in\Theta_{c_0}}$, i.e., $m = 2$.

We first consider the case of $K = 2$ and let $\boldsymbol{\theta} = (\theta_1, \theta_2)$ be the underlying topics. Let $\boldsymbol{\theta}' = (\theta_1', \theta_2')$ be one of its equivalent parameterization such that $d_{\mathsf{TV}}(p_{\boldsymbol{\theta},2}; p_{\boldsymbol{\theta}',2}) = 0$. By the data processing inequality, we must have $d_{\mathsf{TV}}(p_{\boldsymbol{\theta},1}; p_{\boldsymbol{\theta},1}) = 0$ and therefore

$$\mathbb{E}_{\nu_0}[h_1]\theta_1(x) + \mathbb{E}_{\nu_0}[h_2]\theta_2(x) = \mathbb{E}_{\nu_0}[h_1]\theta_1'(x) + \mathbb{E}_{\nu_0}[h_2]\theta_2'(x), \qquad \forall x \in \mathcal{X}.$$

Because $\nu_0$ is exchangeable, the above identity implies that

$$\theta_1(x) + \theta_2(x) = \theta_1'(x) + \theta_2'(x) \qquad \forall x \in \mathcal{X}. \quad (16)$$

We now consider document $X = (x_1, x_2)$ consisting of identical words $x_1 = x_2 = x \in \mathcal{X}$. Because

$$p_{\boldsymbol{\theta}}(X) = \mathbb{E}_{\nu_0}\left[(h_1\theta_1(x) + h_2\theta_2(x))^2\right] = p_{\boldsymbol{\theta}'}(X),$$

using the exchangeability of $\nu_0$ we have that

$$\mathbb{E}_{\nu_0}[h_1^2]\left[\theta_1(x)^2+\theta_2(x)^2\right]+2\mathbb{E}_{\nu_0}[h_1h_2]\theta_1(x)\theta_2(x)$$
$$=\mathbb{E}_{\nu_0}[h_1^2]\left[\theta_1'(x)^2+\theta_2'(x)^2\right]+2\mathbb{E}_{\nu_0}[h_1h_2]\theta_1'(x)\theta_2'(x),\qquad \forall x\in\mathcal{X}.$$

Subtracting $\mathbb{E}_{\nu_0}[h_1]^2(\theta_1(x)+\theta_2(x))^2$ on both sides of the above identity and invoking Eq. (16) that $\theta_1(x)+\theta_2(x)=\theta_1'(x)+\theta_2'(x)$, we have

$$2\mathbb{E}_{\nu_0}[h_1h_2-h_1^2]\theta_1(x)\theta_2(x)=2\mathbb{E}_{\nu_0}[h_1h_2-h_1^2]\theta_1'(x)\theta_2'(x)\qquad \forall x\in\mathcal{X}.$$

Because $\mathbb{E}_{\nu_0}[h_1h_2-h_1^2]>0$ thanks to assumption (A2), we have

$$\theta_1(x)\theta_2(x)=\theta_1'(x)\theta_2'(x)\qquad \forall x\in\mathcal{X}. \tag{17}$$

When $\boldsymbol{\theta}=(\theta_1,\theta_2)$ is fixed, Eqs. (16,17) form a quadratic system of $\theta_1'(x),\theta_2'(x)$ for every $x\in\mathcal{X}$, which has at most two solutions. Therefore, $|\{\boldsymbol{\theta}':d_{\mathsf{TV}}(p_{\boldsymbol{\theta},2};p_{\boldsymbol{\theta}',2})=0\}|\leq 2^V<\infty$, and the finite identifiability is proved.

We next consider the case of $K\geq 3$ and $m=2$. We know that $\mathfrak{d}_{2,1}(\boldsymbol{\theta})=0$ and $\mathfrak{d}_{2,2}(\boldsymbol{\theta})\geq c(\nu_0)/V^3K$ for all $\boldsymbol{\theta}\in\Theta_{c_0}$, thanks to Lemmas 4 and 5. By choosing $\epsilon_0:=c(\nu_0)/[2V^5K+c(\nu_0)]$, by Lemma 6 we have

$$d_{\mathsf{TV}}(p_{\boldsymbol{\theta},2};p_{\boldsymbol{\theta}',2})>0,\qquad \forall d_{\mathsf{W}}(\boldsymbol{\theta},\boldsymbol{\theta}')\leq\epsilon_0. \tag{18}$$

For arbitrary $\boldsymbol{\theta}\in\Theta_{c_0}$ let $\widetilde{\Theta}_{c_0}(\boldsymbol{\theta}):=\{\boldsymbol{\theta}'\in\Theta_{c_0}:d_{\mathsf{TV}}(p_{\boldsymbol{\theta},2};p_{\boldsymbol{\theta}',2})=0\}$ be the set of all its equivalent parameterizations. By Eq. (18), $\widetilde{\Theta}_{c_0}$ forms a *packing* of $\Theta_{c_0}$ with radius $\epsilon_0$ with respect to $d_{\mathsf{W}}(\cdot,\cdot)$. Because $\epsilon_0>0$ is a positive constant depending only on $\nu_0,V,K$ and $\Theta_{c_0}$ is compact, we conclude that $|\widetilde{\Theta}_{c_0}(\boldsymbol{\theta})|<\infty$.

### *5.2. Proof of Theorem 1*

We use a multi-point variant of the classical analysis of maximum likelihood (Van der Vaart, 1998, Sec. 5.8) to establish the rate of convergence for MLE, and Le Cam's method to prove corresponding (local) minimax lower bounds.

*Proof of upper bound.* Let $\boldsymbol{\theta}\in\Theta_{c_0}$ be the underlying parameter that generates the data. Define

$$\widetilde{\Theta}_{c_0}(\boldsymbol{\theta}):=\left\{\widetilde{\boldsymbol{\theta}}\in\Theta_{c_0}:d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m};p_{\widetilde{\boldsymbol{\theta}},m})=0\right\}$$

as the set of its equivalent parameterizations, which is guaranteed to be finite thanks to Lemma 1. For $\epsilon>0$, define

$$\overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta}):=\left\{\boldsymbol{\theta}'\in\Theta_{c_0}:d_{\mathsf{W}}(\boldsymbol{\theta}',\widetilde{\boldsymbol{\theta}})\geq\epsilon,\forall\widetilde{\boldsymbol{\theta}}\in\widetilde{\Theta}_{c_0}(\boldsymbol{\theta})\right\}$$

as the set of all parameters that are at least $\epsilon$ away from any equivalent parameterization in $\widetilde{\Theta}_{c_0}(\boldsymbol{\theta})$ in Wasserstein's distance $d_{\mathsf{W}}(\cdot,\cdot)$. The following technical proposition and corollary shows that $d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m};p_{\boldsymbol{\theta}';m})$ is uniformly lower bounded from below for all $\boldsymbol{\theta}'\in\overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$.

**Proposition 4.** *For every fixed $\boldsymbol{\theta} \in \Theta_{c_0}$, $d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m})$ is continuous in $\boldsymbol{\theta}'$ with respect to $\|\cdot\|_2$, meaning that for every $\varepsilon > 0$, there exists $\delta > 0$ such that $|d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}) - d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}'',m})| \leq \varepsilon$ for all $\boldsymbol{\theta}', \boldsymbol{\theta}'' \in \Theta_{c_0}$ such that $\|\boldsymbol{\theta}' - \boldsymbol{\theta}''\|_2 \leq \delta$, where $\|\boldsymbol{\theta}' - \boldsymbol{\theta}''\|_2 := \sqrt{\sum_{i=1}^K \sum_{j=1}^V |\theta_i'(j) - \theta_i''(j)|^2}$.*

Proposition 4 can be easily proved by explicitly expanding the total variation between distributions parameterized by two parameters $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_{c_0}$. We give its complete proof in the appendix. As a consequence of Proposition 4, we have the following corollary:

**Corollary 3.** *For any $0 < \epsilon < 1/2$, $\inf_{\boldsymbol{\theta}' \in \overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})} d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}) > 0$.*

*Proof.* We first show that $\overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$ is *compact* under the general topology of $\mathbb{R}^{VK}$ by treating each $\boldsymbol{\theta}$ as a $VK$-dimensional vector. $\overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$ is obviously bounded (with respect to $\|\cdot\|_2$), because $\Theta_{c_0}$ is bounded and $\overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta}) \subseteq \Theta_{c_0}$. In addition, $\overline{\Theta}_{c_0,\epsilon}$ can be written as

$$\overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta}) = \bigcap_{\boldsymbol{\theta}' \in \widetilde{\Theta}_{c_0}(\boldsymbol{\theta})} \Theta_{c_0} \backslash \left\{\boldsymbol{\theta}'' \in \mathbb{R}^{VK} : \|\boldsymbol{\theta}'' - \boldsymbol{\theta}'\|_2 < \epsilon\right\}. \tag{19}$$

Note that we have replaced $d_{\mathsf{W}}(\cdot,\cdot)$ with the $\|\cdot\|_2$ norm, which remains correct because all permutations of a parameterization $\boldsymbol{\theta}' \in \widetilde{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$ are also contained in $\widetilde{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$. Because $\Theta_{c_0}$ is closed, $\{\boldsymbol{\theta}'' \in \mathbb{R}^{VK} : \|\boldsymbol{\theta}'' - \boldsymbol{\theta}'\|_2 < \epsilon\}$ is open, and any intersection of closed sets are closed, we conclude that $\overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$ is closed. Therefore $\overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$ is compact. Also, because $\epsilon < 1/2$, $\overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$ is clearly non-empty. By the extreme value theorem [2] and the fact that $d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m})$ is continuous in $\boldsymbol{\theta}'$ with respect to $\|\cdot\|_2$ (Proposition 4), $d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m})$ attains its minimum on $\overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$. The corollary is then proved by noting that $d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}) > 0$ for all $\boldsymbol{\theta}' \in \overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$. $\square$

For any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_{c_0}$, and $X_1, \cdots, X_n \in \mathcal{X}^m$ i.i.d. sampled from the underlying distribution $p_{\boldsymbol{\theta},m}$, define the "empirical KL-divergence" $\widehat{\mathrm{KL}}(p_{\boldsymbol{\theta},m} \| p_{\boldsymbol{\theta}',m})$ as

$$\widehat{\mathrm{KL}}_n(p_{\boldsymbol{\theta},m} \| p_{\boldsymbol{\theta}',m}) := \frac{1}{n} \sum_{i=1}^n \log \frac{p_{\boldsymbol{\theta},m}(X_i)}{p_{\boldsymbol{\theta}',m}(X_i)}.$$

By definition of the ML estimator, we know $\inf_{\widetilde{\boldsymbol{\theta}} \in \widetilde{\Theta}_{c_0}(\boldsymbol{\theta})} d_{\mathsf{W}}(\widehat{\boldsymbol{\theta}}^{\mathsf{ML}}_{n,m}, \widetilde{\boldsymbol{\theta}}) \leq \epsilon$ provided that

$$\widehat{\mathrm{KL}}_n(p_{\boldsymbol{\theta},m} \| p_{\boldsymbol{\theta}',m}) > 0 \quad \text{for all } \boldsymbol{\theta}' \in \overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta}). \tag{20}$$

Furthermore, we know that the "population" version of Eq. (20) must be correct: $\inf_{\boldsymbol{\theta}' \in \overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})} \mathrm{KL}(p_{\boldsymbol{\theta},m} \| p_{\boldsymbol{\theta}',m}) > 0$, because the KL-divergence is lower bounded by the total-variation distance, which is further uniformly bounded away from

[2] For any compact set $K \subset \mathbb{R}^d$ and continuous function $f : K \to \mathbb{R}$, $f$ attains its minimum and maximum on $K$.

below by Corollary 3. Therefore, to prove convergence rate of the MLE it suffices to upper bound the perturbation between empirical and population KL-divergence and lower bounds the population divergence for all $\boldsymbol{\theta}' \in \overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$.

We first consider the simpler task of bounding the perturbation between $\widehat{\mathrm{KL}}_n(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m})$ and its population version $\mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m})$. Note that $\widehat{\mathrm{KL}}_n(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m})$ is a sample average of i.i.d. random variables. Using classical empirical process theory, we have the following lemma that bounds the uniform convergence of $\widehat{\mathrm{KL}}_n$ towards KL; its complete proof is given in the appendix.

**Lemma 7.** *There exists $C_{\boldsymbol{\theta}} > 0$ depending only on $\boldsymbol{\theta}, c_0, m, \nu_0$ such that*

$$\mathbb{E}_{\boldsymbol{\theta}} \sup_{\boldsymbol{\theta}' \in \Theta_{c_0}} \frac{\left|\widehat{\mathrm{KL}}_n(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m}) - \mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m})\right|}{\sqrt{\mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m})}} \leq \frac{C_{\boldsymbol{\theta}}}{\sqrt{n}}.$$

As a corollary, by Markov's inequality we know that for all $\delta \in (0,1)$, with probability $1-\delta$

$$\widehat{\mathrm{KL}}_n(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m}) \geq \mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m}) - \sqrt{\mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m})} \cdot \frac{C_{\boldsymbol{\theta}}}{\delta\sqrt{n}}, \quad \forall \boldsymbol{\theta}' \in \Theta_{c_0}.$$

Subsequently, with probability $1-\delta$

$$\inf_{\boldsymbol{\theta}' \in \overline{\Theta}_{\epsilon,c_0}} \widehat{\mathrm{KL}}_n(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m}) > 0 \quad \Longleftarrow \quad \inf_{\boldsymbol{\theta}' \in \overline{\Theta}_{\epsilon,c_0}} \mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m}) > \frac{C_{\boldsymbol{\theta}}^2}{\delta^2 n}. \tag{21}$$

We next establish a lower bound on $\mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m})$ for all $\boldsymbol{\theta}' \in \overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$. By Pinsker's inequality, we have that for any $\boldsymbol{\theta}' \in \overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})$,

$$\mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m}) \geq 2d_{\mathsf{TV}}^2(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}).$$

Define

$$\epsilon_0(\boldsymbol{\theta}) := \min\left\{\frac{1}{4}, \frac{\mathfrak{d}_{m,\mathfrak{p}(m;\boldsymbol{\theta})}(\boldsymbol{\theta})}{2V^m + \mathfrak{d}_{m,\mathfrak{p}(m;\boldsymbol{\theta})}(\boldsymbol{\theta})}\right\}.$$

Invoking Lemma 6 and noting that

$$\mathfrak{d}_{m,\mathfrak{p}(m;\boldsymbol{\theta})}(\boldsymbol{\theta}) - \frac{V^m \epsilon_0(\boldsymbol{\theta})}{1-\epsilon_0(\boldsymbol{\theta})} \leq \frac{1}{2}\mathfrak{d}_{m,\mathfrak{p}(m;\boldsymbol{\theta})}(\boldsymbol{\theta}),$$

we have for all $0 < \epsilon < \epsilon_0(\boldsymbol{\theta})$ that

$$\inf_{\boldsymbol{\theta}' \in \overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta}) \backslash \overline{\Theta}_{c_0,\epsilon_0(\boldsymbol{\theta})}(\boldsymbol{\theta})} \mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m}) \geq \frac{1}{2}[\mathfrak{d}_{m,\mathfrak{p}(m;\boldsymbol{\theta})}(\boldsymbol{\theta})]^2 \cdot \epsilon^{2\mathfrak{p}(m;\boldsymbol{\theta})} =: \gamma_{\boldsymbol{\theta}} \epsilon^{2\mathfrak{p}(m;\boldsymbol{\theta})},$$

where $\gamma_{\boldsymbol{\theta}} > 0$ is a positive constant independent of $n$ or $\epsilon$. Furthermore, by Corollary 3 and the Pinsker's inequality we know that $\inf_{\boldsymbol{\theta}' \in \overline{\Theta}_{c_0,\epsilon_0(\boldsymbol{\theta})}(\boldsymbol{\theta})} \mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m})$ $> 0$. Because $\epsilon_0(\boldsymbol{\theta})$ does not depend on $\epsilon$ or $n$, this infimum must be bounded away from below by a constant depending only on $\boldsymbol{\theta}, c_0, \nu_0$ and $m$. Subsequently, for sufficiently small $\epsilon > 0$ we have

$$\inf_{\boldsymbol{\theta}' \in \overline{\Theta}_{c_0,\epsilon}(\boldsymbol{\theta})} \mathrm{KL}(p_{\boldsymbol{\theta},m} \| p_{\boldsymbol{\theta}',m}) \geq \gamma'_{\boldsymbol{\theta}} \epsilon^{2\mathfrak{p}(m;\boldsymbol{\theta})}, \tag{22}$$

where $\gamma'_{\boldsymbol{\theta}}$ is a positive constant depending only on $\boldsymbol{\theta}, c_0, \nu_0$ and $m$.

Combining Eqs. (20), (21) and (22) with $\epsilon \asymp n^{-1/2\mathfrak{p}(m;\boldsymbol{\theta})}$ we complete the proof of convergence rate of the ML estimator.

*Proof of lower bound.* Let $n$ be sufficiently large such that $r_{\boldsymbol{\theta}} n^{-1/2\mathfrak{p}(m;\boldsymbol{\theta})} < 1/2$, where $r_{\boldsymbol{\theta}}$ is the positive constant in the definition of $\Theta_n(\boldsymbol{\theta})$ that is independent of $n$. Invoking Lemma 6 we have that

$$\sup_{\boldsymbol{\theta}' \in \Theta_n(\boldsymbol{\theta})} d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}) \leq 2V^m \cdot r_{\boldsymbol{\theta}}^{\mathfrak{p}(m;\boldsymbol{\theta})} n^{-1/2}, \tag{23}$$

In addition, for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_{c_0}$ the following proposition upper bounds their KL-divergence using TV distance:

**Proposition 5.** *There exists a constant $C > 0$ depending only on $V, K, \nu_0, c_0$ and $m$ such that, for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_{c_0}$,*

$$\mathrm{KL}(p_{\boldsymbol{\theta},m} \| p_{\boldsymbol{\theta}',m}) \leq C \cdot d^2_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}).$$

At a higher level, Proposition 5 can be viewed as an "exact" reverse of the Pinsker's inequality with matching upper and lower bounds for the KL divergence. It is not generally valid for arbitrary distributions, but holds true for our particular model with $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_{c_0}$ because both $p_{\boldsymbol{\theta},m}$ and $p_{\boldsymbol{\theta}',m}$ are supported and bounded away from below on a finite set. We give the complete proof of Proposition 5 in the appendix.

Let $\boldsymbol{\theta}'$ be an arbitrary parameterization in $\Theta_n(\boldsymbol{\theta})$, and let $p_{\boldsymbol{\theta},m}^{\otimes n} = p_{\boldsymbol{\theta},m} \times \cdots \times p_{\boldsymbol{\theta},m}$ be the $n$-times product measure of $p_{\boldsymbol{\theta},m}$, Using Eq. (23), Proposition 5 and the fact that the KL-divergence is additive for product measures, we have

$$\mathrm{KL}(p_{\boldsymbol{\theta},m}^{\otimes n} \| p_{\boldsymbol{\theta}',m}^{\otimes n}) \leq n \cdot \mathrm{KL}(p_{\boldsymbol{\theta},m} \| p_{\boldsymbol{\theta}',m}) \leq 2V^m \cdot r_{\boldsymbol{\theta}}^{2\mathfrak{p}(m;\boldsymbol{\theta})}.$$

Subsequently, using Pinsker's inequality we have

$$d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}^{\otimes n}; p_{\boldsymbol{\theta}',m}^{\otimes n}) \leq \sqrt{2V^m \cdot r_{\boldsymbol{\theta}}^{2\mathfrak{p}(m;\boldsymbol{\theta})}}.$$

By choosing $r_{\boldsymbol{\theta}} := [8V^m]^{-2\mathfrak{p}(m;\boldsymbol{\theta})}$ we can upper bound the right-hand side of the above inequality by $1/2$. Applying Le Cam's inequality we conclude that no statistical procedure can distinguish $\boldsymbol{\theta}$ from $\boldsymbol{\theta}'$ using $n$ observations with success probability higher than $3/4$. The lower bound is thus proved by Markov's inequality.

### 5.3. Proof of Lemma 2

This lemma is essentially a consequence of (Anandkumar et al., 2014), which developed a $\sqrt{n}$-consistent estimator for linear independent topics via the method of moments. More specifically, the main result of (Anandkumar et al., 2014) can be summarized by the following theorem:

**Theorem 2.** *Suppose* $2 \leq K \leq V$, $m = 3$ *and consider the parameter subclass* $\Theta_{\sigma_0,c_0} := \{\boldsymbol{\theta} \in \Theta_{c_0} : \sigma_{\min}(\boldsymbol{\theta}) \geq \sigma_0\}$, *where* $\sigma_{\min}(\boldsymbol{\theta}) := \inf_{\|w\|_2=1} \|\sum_{j=1}^K w_j\theta_j\|_2$ *is the least singular value of the topics vectors, and* $\sigma_0 > 0$ *is a positive constant. Then there exists a (computationally tractable) estimator* $\widehat{\boldsymbol{\theta}}_n$ *such that for all* $\boldsymbol{\theta} \in \Theta_{\sigma_0,c_0}$,

$$d_{\mathsf{W}}(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}) \leq C_{\sigma_0} \cdot O_{\mathbb{P}}(n^{-1/2}),$$

*where* $C_{\sigma_0} > 0$ *is a constant that only depends on* $V, K, \nu_0$ *and* $\sigma_0$.

We remark that the original paper of (Anandkumar et al., 2014) only considered the case where $\nu_0$ is the Dirichlet distribution. However, our assumption (A2) is sufficient for the success of their proposed algorithms and analysis.

Next consider any $\boldsymbol{\theta} \in \Theta_{c_0}$ such that $\{\theta_k\}_{k=1}^K$ are linear independent. Define $\sigma_{\boldsymbol{\theta}} := \sigma_{\min}(\boldsymbol{\theta})/2 > 0$. The "shrinking neighborhood" $\Theta_n(\boldsymbol{\theta})$ defined in Theorem 1 is then contained in $\Theta_{\sigma_{\boldsymbol{\theta}},c_0}$ for sufficiently large $n$. Let $\boldsymbol{\theta}' \in \Theta_n(\boldsymbol{\theta}) \subseteq \Theta_{\sigma_{\boldsymbol{\theta}},c_0}$ be such that $d_{\mathsf{W}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \Omega(n^{-1/2\mathfrak{p}(3;\boldsymbol{\theta})})$. If $\mathfrak{p}(3;\boldsymbol{\theta}) = 1$ we already proved $\mathfrak{d}_{3,1}(\boldsymbol{\theta}) > 0$. On the other hand, if $\mathfrak{p}(3;\boldsymbol{\theta}) > 1$ we know that $d_{\mathsf{W}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \Omega(n^{-1/4})$. By Theorem 2, there exists a statistical procedure that can distinguish $\boldsymbol{\theta}$ from $\boldsymbol{\theta}'$ with success probability arbitrarily close to 1 for sufficiently large $n$, which violates the lower bound in Theorem 1 (Remark 3). Thus, it is concluded that $\mathfrak{p}(3;\boldsymbol{\theta}) = 1$ and therefore $\mathfrak{d}_{3,1}(\boldsymbol{\theta}) > 0$.

### 5.4. Proof of Lemma 3

Consider $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_K)$ with $\delta_j = \frac{1}{4}(e_1 - e_2)$, $\delta_k = \frac{1}{4}(e_2 - e_1)$ and $\delta_\ell = 0$ for all $\ell \neq j, k$, where $e_1 = (1, 0, \cdots, 0)$ and $e_2 = (0, 1, 0, \cdots, 0)$ are standard basis vectors in $\mathbb{R}^V$. Clearly $\|\boldsymbol{\delta}\|_1 = 1$ and $\sum_{\ell=1}^V \delta_j(\ell) = 0$ for all $j \in [K]$. Define $p_{\boldsymbol{\theta},h}(x_{-i}) := \prod_{j\neq i} p_{\boldsymbol{\theta},h}(x_j)$. We then have, for arbitrary $x = (x_1, \cdots, x_m) \in \mathcal{X}^m$,

$$\begin{aligned}\left|\mathbb{E}_h p_{\boldsymbol{\theta},h}(x) \sum_{i=1}^m \frac{\boldsymbol{\delta}_h(x_i)}{p_{\boldsymbol{\theta},h}(x_i)}\right| &= \left|\sum_{i=1}^m \sum_{\ell=1}^k \delta_\ell(x_i)\mathbb{E}_h\left[h_\ell p_{\boldsymbol{\theta},h}(x_{-i})\right]\right| \\ &\leq \frac{1}{2}\sum_{i=1}^m \mathbf{1}_{[x_i\in\{1,2\}]}\big|\mathbb{E}_h[h_j p_{\boldsymbol{\theta},h}(x_{-i})] - \mathbb{E}_h[h_k p_{\boldsymbol{\theta},h}(x_{-i})]\big|.\end{aligned}$$

Because $\nu_0$ is exchangeable and $\theta_j = \theta_k$, we have that $\mathbb{E}_h[h_j p_{\boldsymbol{\theta},h}(x_{-i})] = \mathbb{E}_h[h_k p_{\boldsymbol{\theta},h}(x_{-i})]$ for all $x_{-i} \in \mathcal{X}^{m-1}$. Thus, $\mathfrak{d}_{m,1}(\boldsymbol{\theta}) = 0$.

### 5.5. *Proof of Lemma 4*

*Proof of the "IF" part.* Let $\theta_1, \theta_2$ and $\theta_3$ be any three topic vectors in $\boldsymbol{\theta}$. We assume $\theta_1, \theta_2, \theta_3$ are distinct, because otherwise $\mathfrak{d}_{2,1}(\boldsymbol{\theta}) = 0$ is already implied by Lemma 3. Consider $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_K)$ defined as

$$\begin{aligned}
\delta_1 &:= (\theta_2 - \theta_3)/6;\\
\delta_2 &:= (\theta_3 - \theta_1)/6;\\
\delta_3 &:= (\theta_1 - \theta_2)/6;\\
\delta_k &:= 0, \quad \forall 3 < k \leq K.
\end{aligned}$$

It is easy to verify that $\|\boldsymbol{\delta}\|_1 = 1$ and $\sum_{\ell=1}^{V} \delta_k(\ell) = 0$ for all $k \in [K]$. We then have, for any $x = (x, y) \in \mathcal{X}^2$,

$$\mathbb{E}_h p_{\boldsymbol{\theta},h}(x,y) \left[ \frac{\boldsymbol{\delta}_h(x)}{p_{\boldsymbol{\theta},h}(x)} + \frac{\boldsymbol{\delta}_h(y)}{p_{\boldsymbol{\theta},h}(y)} \right] = \mathbb{E}_h[\boldsymbol{\delta}_h(x) p_{\boldsymbol{\theta},h}(y)] + \mathbb{E}_h[\boldsymbol{\delta}_h(y) p_{\boldsymbol{\theta},h}(x)] \quad (24)$$

By definition of $\boldsymbol{\delta}$, we have that $6\boldsymbol{\delta}_h(x) = \theta_1(x)(h_3 - h_2) + \theta_2(x)(h_1 - h_3) + \theta_3(x)(h_2 - h_1)$. Define $\beta := (\mathbb{E}_{\nu_0}[h_1^2] - \mathbb{E}_{\nu_0}[h_1 h_2])/6$. We then have

$$\begin{aligned}
&\mathbb{E}_h[\boldsymbol{\delta}_h(x) p_{\boldsymbol{\theta},h}(y)]\\
&\quad = \beta\theta_1(x)(\theta_3(y) - \theta_2(y)) + \beta\theta_2(x)(\theta_1(y) - \theta_3(y)) + \beta\theta_3(x)(\theta_2(y) - \theta_1(y)).
\end{aligned} \quad (25)$$

Similarly,

$$\begin{aligned}
&\mathbb{E}_h[\boldsymbol{\delta}_h(y) p_{\boldsymbol{\theta},h}(x)]\\
&\quad = \beta\theta_1(y)(\theta_3(x) - \theta_2(x)) + \beta\theta_2(y)(\theta_1(x) - \theta_3(x)) + \beta\theta_3(y)(\theta_2(x) - \theta_1(x)).
\end{aligned} \quad (26)$$

Comparing Eqs. (25,26) we note that

$$\mathbb{E}_h[\boldsymbol{\delta}_h(x) p_{\boldsymbol{\theta},h}(y)] = -\mathbb{E}_h[\boldsymbol{\delta}_h(y) p_{\boldsymbol{\theta},h}(x)]$$

for all $(x, y) \in \mathcal{X}^2$, which means that the right-hand side of Eq. (24) is always 0. Therefore, $\mathfrak{d}_{2,1}(\boldsymbol{\theta}) = 0$.
*Proof of the "ONLY IF" part.* We show that if $K = 2$ and $\theta_1 \neq \theta_2$ then $\mathfrak{d}_{2,1}(\boldsymbol{\theta}) > 0$. Define $\beta := \mathbb{E}_{\nu_0}[h_1 h_2]$ and $\gamma := \mathbb{E}_{\nu_0}[h_1^2] - \mathbb{E}_{\nu_0}[h_1 h_2]$. By (A2) we have that $\gamma > 0$. We then have

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\delta}_h(x) p_{\boldsymbol{\theta},h}(y)] &= \mathbb{E}[(h_1\delta_1(x) + h_2\delta_2(x))(h_1\theta_1(y) + h_2\theta_2(y))]\\
&= \delta_1(x)[\beta\overline{\theta}(y) + \gamma\theta_1(y)] + \delta_2(x)[\beta\overline{\theta}(y) + \gamma\theta_2(y)],
\end{aligned}$$

where $\overline{\theta}(y) := \theta_1(y) + \theta_2(y)$. Similarly,

$$\mathbb{E}[\boldsymbol{\delta}_h(y) p_{\boldsymbol{\theta},h}(x)] = \delta_1(y)[\beta\overline{\theta}(x) + \gamma\theta_2(x)] + \delta_2(y)[\beta\overline{\theta}(x) + \gamma\theta_2(x)].$$

We can then simplify Eq. (24) as

$$\begin{aligned} T_{\boldsymbol{\theta},x,y}(\boldsymbol{\delta}) &:= \mathbb{E}_h p_{\boldsymbol{\theta},h}(x,y)\left[\frac{\boldsymbol{\delta}_h(x)}{p_{\boldsymbol{\theta},h}(x)} + \frac{\boldsymbol{\delta}_h(y)}{p_{\boldsymbol{\theta},h}(y)}\right] \\ &= \delta_1(x)[\beta\overline{\theta}(y) + \gamma\theta_1(y)] + \delta_2(x)[\beta\overline{\theta}(y) + \gamma\theta_2(y)] \\ &+ \delta_1(y)[\beta\overline{\theta}(x) + \gamma\theta_1(x)] + \delta_2(y)[\beta\overline{\theta}(x) + \gamma\theta_2(x)]. \end{aligned}$$

Assume by way of contradiction that $\mathfrak{d}_{2,1}(\boldsymbol{\theta}) = 0$, which implies the existence of $\boldsymbol{\delta} \neq 0$, $\sum_{\ell=1}^{V} \delta_j(\ell) = 0$ such that $T_{\boldsymbol{\theta},x,y}(\boldsymbol{\delta}) = 0$ for all $x, y \in [V]$. We then have

$$\mathbf{B}_1\delta_1 + \mathbf{B}_2\delta_2 = 0, \tag{27}$$

where $\mathbf{B}_1 = (b_{11}, \cdots, b_{1V})$ and $\mathbf{B}_2 = (b_{21}, \cdots, b_{2V})$ are $K \times (V^2 + 2)$ matrices. Furthermore, $b_{j\ell}$ for $j \in \{1, 2\}$ and $\ell \in [V]$ can be explicitly formed as

$$b_{j\ell} = (\beta\overline{\theta} + \gamma\theta_j)(\boldsymbol{e}_{\ell\cdot} + \boldsymbol{e}_{\cdot\ell}) + \mu_{j\ell}\boldsymbol{e}_{\ell\ell}$$

where $\mu_{j\ell} = \beta\overline{\theta}(\ell) + \gamma\theta_j(\ell)$ and $\{\boldsymbol{e}_{\ell\ell'}\}_{\ell,\ell'=1}^{V}$ denotes the $V^2$ components of $b_{j\ell}$. Subsequently,

$$\begin{aligned} \mathbf{B}_1\delta_1 + \mathbf{B}_2\delta_2 &= \sum_{\ell=1}^{V} \delta_1(\ell) b_{1\ell} + \sum_{\ell=1}^{V} \delta_2(\ell) b_{2\ell} \\ &= \sum_{\ell=1}^{V} \boldsymbol{e}_{\ell\cdot}\left[\sum_{j=1,2} \delta_j(\ell)(\beta\overline{\theta} + \gamma\theta_j) + \mu_{j\ell}\delta_j\right]; \end{aligned}$$

therefore,

$$\sum_{j=1,2} \delta_j(\ell)(\beta\overline{\theta} + \gamma\theta_j) + \mu_{j\ell}\delta_j = 0, \qquad \forall \ell \in [V]. \tag{28}$$

We next state a technical proposition that will be proved in the appendix, which shows that $\delta_1$ and $\delta_2$ can be expressed as linear combinations of $\beta\overline{\theta} + \gamma\theta_1$ and $\beta\overline{\theta} + \gamma\theta_2$:

**Proposition 6.** *There exists $\xi_{11}, \xi_{12}, \xi_{21}, \xi_{22} \in \mathbb{R}$ such that $\delta_1 = \xi_{11}(\beta\overline{\theta} + \gamma\theta_1) + \xi_{12}(\beta\overline{\theta} + \gamma\theta_2)$ and $\delta_2 = \xi_{21}(\beta\overline{\theta}_1 + \gamma\theta_2) + \xi_{22}(\beta\overline{\theta}_1 + \gamma\theta_2)$.*

Substituting the expression of $\delta_1$ and $\delta_2$ in Proposition 6 into Eq. (28), we have

$$\sum_{j=1,2} (\beta\overline{\theta} + \gamma\theta_j)\left[\sum_{k=1,2} \mu_{k\ell}(\xi_{jk} + \xi_{kj})\right] = 0, \qquad \forall \ell \in [V]. \tag{29}$$

Because $\beta\overline{\theta} + \gamma\theta_1$ and $\beta\overline{\theta} + \gamma\theta_2$ are linear independent if $\gamma > 0$ and $\theta_1 \neq \theta_2$, it must hold that $\sum_{k=1,2} \mu_{k\ell}(\xi_{jk} + \xi_{kj}) = 0$ for all $j = 1, 2$ and $\ell \in [V]$. Recall that $\mu_{k\ell} = \beta\overline{\theta}(\ell) + \gamma\theta_k(\ell)$. Subsequently, for $j = 1, 2$ we have

$$\sum_{k=1,2} (\xi_{jk} + \xi_{kj})(\beta\overline{\theta} + \gamma\theta_k) = 0.$$

Using again the fact that $\beta\overline{\theta} + \gamma\theta_1$ and $\beta\overline{\theta} + \gamma\theta_2$ are linear independent, we conclude $\xi_{jk} + \xi_{kj} = 0$ for all $k = 1, 2$. Thus, $\xi_{11} = \xi_{22} = 0$ and $\xi_{12} = -\xi_{21}$. On the other hand, because $\mathrm{sum}(\delta_1) = \mathrm{sum}(\delta_2) = 0$ and $\mathrm{sum}(\beta\overline{\theta} + \gamma\theta_1) = \mathrm{sum}(\beta\overline{\theta}+\gamma\theta_2) = \beta+\gamma > 0$, where $\mathrm{sum}(z) := \sum_{\ell=1}^{V} z(\ell)$, we must have $\xi_{11}+\xi_{12} = \xi_{21}+\xi_{22} = 0$, and hence Eq. (27) only has the trivial solution $\delta_1 = \delta_2 = 0$. Thus, $\mathfrak{d}_{2,1}(\boldsymbol{\theta}) = 0$.

### 5.6. Proof of Lemma 5

For any $\ell \in [V]$ consider $x = (x_1, x_2) \in [V]^2$ where $x_1 = x_2 = \ell$. Because of (A1), $p_{\boldsymbol{\theta},h}(x) > 0$ for all $h \in \Delta^{K-1}$. Subsequently,

$$\begin{aligned}
\mathbb{E}_h p_{\boldsymbol{\theta},h}(x) \frac{\boldsymbol{\delta}_h(x_1)\boldsymbol{\delta}_h(x_2)}{p_{\boldsymbol{\theta},h}(x_1)p_{\boldsymbol{\theta},h}(x_2)} &= \mathbb{E}_h\left[\boldsymbol{\delta}_h(x_1)\boldsymbol{\delta}_h(x_2)\right] = \mathbb{E}_h\left[\left(\sum_{j=1}^{k} h_j\delta_j(\ell)\right)^2\right] \\
&= \mathbb{E}[h_1h_2]\left(\sum_{j=1}^{k}\delta_j(\ell)\right)^2 + (\mathbb{E}[h_1^2] - \mathbb{E}[h_1h_2])\sum_{j=1}^{k}\delta_j(\ell)^2 \\
&\geq c(\nu_0)\sum_{j=1}^{k}\delta_j(\ell)^2.
\end{aligned}$$

Here in the last line we use the fact that $\nu_0$ is exchangeable and the definition that $c(\nu_0) = \mathbb{E}_{\nu_0}[h_1^2 - h_1h_2] > 0$. Subsequently, for every $\boldsymbol{\delta}$ satisfying $\|\boldsymbol{\delta}\|_1 = 1$, it holds that

$$\begin{aligned}
\mathfrak{d}_{2,2}(\boldsymbol{\theta}) &\geq V^{-2}\sum_{\ell=1}^{V}\left|\mathbb{E}_h p_{\boldsymbol{\theta},h}(\ell,\ell)\frac{\boldsymbol{\delta}_h(\ell)^2}{p_{\boldsymbol{\theta},h}(\ell)^2}\right| \\
&\geq V^{-2}\cdot c(\nu_0)\sum_{j=1}^{k}\sum_{\ell=1}^{V}\delta_j(k)^2 \\
&\geq V^{-2}\cdot c(\nu_0)\cdot\frac{(\sum_{j=1}^{k}\sum_{\ell=1}^{V}|\delta_j(k)|)^2}{VK} = \frac{c(\nu_0)\|\boldsymbol{\delta}\|_1^2}{V^3K} = \frac{c(\nu_0)}{V^3K}.
\end{aligned}$$

## Appendix A: Missing proofs

We give missing proofs of technical lemmas in this appendix.

### A.1. Proof of Lemma 7

For any $\boldsymbol{\theta}' \in \Theta_{c_0}$ define a $V^m$-dimensional random vector $v_{\boldsymbol{\theta}'}$ as $v_{\boldsymbol{\theta}'}(x) := \log\frac{p_{\boldsymbol{\theta},m}(x)}{p_{\boldsymbol{\theta}',m}(x)}$ for $x \in [V]^m$. We then have that $\widehat{\mathrm{KL}}_n(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m}) = \frac{1}{n}\sum_{i=1}^{n} v_{\boldsymbol{\theta}'}(X_i)$

and $\mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m}) = \mathbb{E}_{\boldsymbol{\theta}}[v_{\boldsymbol{\theta}'}(X)]$. By a simple re-scaling argument, we have that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}} \sup_{\boldsymbol{\theta}'\in\Theta_{c_0}} \frac{|\widehat{\mathrm{KL}}_n(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m}) - \mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m})|}{\|v_{\boldsymbol{\theta}'}\|_2} \\
&\quad\le \mathbb{E}_{\boldsymbol{\theta}} \sup_{\|v\|_2=1} \left|\frac{1}{n}\sum_{i=1}^n v(X_i) - \mathbb{E}_{\boldsymbol{\theta}}[v(X)]\right|. \qquad (30)
\end{aligned}$$

Consider the unit $V^m$-dimensional $\ell_2$ ball $\mathbb{B}_2(V^m) := \{z \in \mathbb{R}^{V^m} : \|z\|_2 \le 1\}$. Using standard empirical process theory (e.g., (Van der Vaart, 1998, Lemma 19.36), (Talagrand, 1994, Theorem 1.1)) we have

$$\mathbb{E}_{\boldsymbol{\theta}} \sup_{\|v\|_2\le 1} \sqrt{n}\left|\frac{1}{n}\sum_{i=1}^n v(X_i) - \mathbb{E}_{\boldsymbol{\theta}}[v(X)]\right| \le C, \qquad (31)$$

where $C > 0$ is a constant that only depends on $V$ and $m$. In addition, because $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_{c_0}$ we know that both $p_{\boldsymbol{\theta},m}$ and $p_{\boldsymbol{\theta}',m}$ are lower bounded by $c_0^m$ uniformly on $[V]^m$; hence, for any $\boldsymbol{\theta}' \in \Theta_{c_0}$, using second-order Taylor expansion of the logarithm we have

$$\begin{aligned}
\|v_{\boldsymbol{\theta}'}\|_2 &\le V^{m/2} \max_{x\in[V]^m} \left|\log \frac{p_{\boldsymbol{\theta},m}(x)}{p_{\boldsymbol{\theta}',m}(x)}\right| \le V^{m/2} \max_{x\in[V]^m} 2c_0^{-2m}\big|p_{\boldsymbol{\theta},m}(x) - p_{\boldsymbol{\theta}',m}(x)\big| \\
&\le 2V^{m/2}c_0^{-2m} \cdot d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}) \le \sqrt{2}V^{m/2}c_0^{-2m} \cdot \sqrt{\mathrm{KL}(p_{\boldsymbol{\theta},m}\|p_{\boldsymbol{\theta}',m})}. \qquad (32)
\end{aligned}$$

Here the last inequality holds by Pinsker's inequality. Combining Eqs. (30,31,32) we complete the proof of Lemma 7.

### *A.2. Proof of Proposition 4*

By definition, for fixed $c_0, m$ and any two $\boldsymbol{\theta}, \theta' \in \Theta_{c_0}$, we have

$$\begin{aligned}
&d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}) \\
&= \int_{\mathcal{X}^m} \big|p_{\boldsymbol{\theta}}(x) - p_{\boldsymbol{\theta}'}(x)\big| \mathrm{d}\mu_m(x) \\
&\le V^m \cdot \max_{x\in\mathcal{X}^m} \big|p_{\boldsymbol{\theta}}(x) - p_{\boldsymbol{\theta}'}(x)\big| \\
&= V^m \cdot \max_{x\in\mathcal{X}^m} \left|\int_{\Delta^{K-1}} [p_{\boldsymbol{\theta},h}(x) - p_{\boldsymbol{\theta}',h}(x)]\mathrm{d}\nu_0(h)\right| \\
&\le V^m \cdot \max_{x\in\mathcal{X}^m} \sup_{h\in\Delta^{K-1}} \big|p_{\boldsymbol{\theta},h}(x) - p_{\boldsymbol{\theta}',h}(x)\big| \\
&= V^m \cdot \max_{x\in\mathcal{X}^m} \sup_{h\in\Delta^{K-1}} \left|\prod_{i=1}^m \left(\sum_{j=1}^K h_j\theta_j(x_i)\right) - \prod_{i=1}^m \left(\sum_{j=1}^K h_j\theta'_j(x_i)\right)\right|
\end{aligned}$$

$$
\begin{aligned}
&\le V^m \cdot \max_{x\in\mathcal{X}^m} \sup_{h\in\Delta^{K-1}} \sum_{j_1,\cdots,j_m=1}^{K} h_{j_1}\cdots h_{j_m} \\
&\quad \times \left|\theta_{j_1}(x_1)\cdots\theta_{j_m}(x_m) - \theta'_{j_1}(x_1)\cdots\theta'_{j_m}(x_m)\right| \\
&\le V^m \cdot \max_{x\in\mathcal{X}^m} \sup_{h\in\Delta^{K-1}} \max_{j_1,\cdots,j_m\in[K]} \left|\theta_{j_1}(x_1)\cdots\theta_{j_m}(x_m) - \theta'_{j_1}(x_1)\cdots\theta'_{j_m}(x_m)\right| \\
&\quad \cdot \left(\sum_{j_1,\cdots,j_m=1}^{K} h_{j_1}\cdots h_{j_m}\right) \\
&= V^m \cdot \max_{x\in\mathcal{X}^m} \sup_{h\in\Delta^{K-1}} \max_{j_1,\cdots,j_m\in[K]} \left|\theta_{j_1}(x_1)\cdots\theta_{j_m}(x_m) - \theta'_{j_1}(x_1)\cdots\theta'_{j_m}(x_m)\right|.
\end{aligned}
$$

Here the last inequality holds because $\sum_{j=1}^{K} h_j = 1$. Furthermore, because $\theta_j(x), \theta'_j(x) \in (0,1]$, we have

$$
\begin{aligned}
&\max_{j_1,\cdots,j_m\in[K]} \left|\theta_{j_1}(x_1)\cdots\theta_{j_m}(x_m) - \theta'_{j_1}(x_1)\cdots\theta'_{j_m}(x_m)\right| \\
&\le \max_{j_1,\cdots,j_m\in[K]} \sum_{\ell=1}^{m} \binom{m}{\ell} \left(\max_{j\in[K],x\in[V]} |\theta_j(x) - \theta'_j(x)|\right)^{\ell} \\
&\le 2^m \cdot \max_{j\in[K],x\in[V]} |\theta_j(x) - \theta'_j(x)| \\
&\le 2^m \|\boldsymbol{\theta}' - \boldsymbol{\theta}'\|_2.
\end{aligned}
$$

Therefore, we have for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_{c_0}$ that

$$
d_{\mathsf{TV}}(p_{\boldsymbol{\theta},m}; p_{\boldsymbol{\theta}',m}) \le (2V)^m \cdot \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2,
$$

and the proposition is proved, because both $V$ and $m$ are fixed quantities independent of $\boldsymbol{\theta}$ or $\boldsymbol{\theta}'$.

### *A.3. Proof of Proposition 5*

We prove a more general statement: if $P$ and $Q$ are distributions uniformly lower bounded by a constant $c > 0$ on a finite domain $\mathcal{D}$, then there exists a constant $C > 0$ depending only on $c$ such that $\mathrm{KL}(P\|Q) \le C \cdot d^2_{\mathsf{TV}}(P;Q)$. This implies Proposition 5 because for any $\boldsymbol{\theta} \in \Theta_{c_0}$, $p_{\boldsymbol{\theta},m}$ is uniformly lower bounded by $c_0^m$ on $\mathcal{X}^m$.

Let $\mu$ be the counting measure on $\mathcal{D}$. Using the definition of KL divergence and second-order Taylor expansion of the logarithm, we have

$$
\begin{aligned}
\mathrm{KL}(P\|Q) &= \int_{\mathcal{D}} P \log\frac{P}{Q}\mathrm{d}\mu = \int_{\mathcal{D}} P\log\left(1 + \frac{P-Q}{Q}\right)\mathrm{d}\mu \\
&\le \int_{\mathcal{D}} \frac{P^2}{Q}\mathrm{d}\mu - 1 + \int_{\mathcal{D}} \frac{P(P-Q)^2}{2Q^2}\mathrm{d}\mu \\
&= \int_{\mathcal{D}} \frac{P^2 - Q^2}{Q}\mathrm{d}\mu + \int_{\mathcal{D}} \frac{P(P-Q)^2}{2Q^2}\mathrm{d}\mu
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{\mathcal{D}} \frac{(P-Q)^2 + 2PQ - 2Q^2}{Q} \mathrm{d}\mu + \int_{\mathcal{D}} \frac{P(P-Q)^2}{2Q^2} \mathrm{d}\mu \\
&= \int_{\mathcal{D}} \frac{(P-Q)^2}{Q} \mathrm{d}\mu + \int_{\mathcal{D}} \frac{P(P-Q)^2}{2Q^2} \mathrm{d}\mu \\
&\leq (1/2c^2 + 1/c) \cdot \int_{\mathcal{D}} (P-Q)^2 \mathrm{d}\mu.
\end{aligned}
$$

On the other hand, $d_{\mathsf{TV}}(P;Q) = \int_{\mathcal{D}} |P-Q| \mathrm{d}\mu \geq \sqrt{\int_{\mathcal{D}} (P-Q)^2 \mathrm{d}\mu}$. Therefore, $\mathrm{KL}(P\|Q) \leq (1/2c^2 + 1/c) \cdot d_{\mathsf{TV}}^2(P;Q)$.

### *A.4. Proof of Proposition 6*

We prove that $\mathrm{span}\{\delta_j\}_{j=1}^2 \subseteq \mathrm{span}\{\beta\overline{\theta} + \gamma\theta_j\}_{j=1}^2$, which would then imply the proposition. Re-arranging terms in Eq. (28) we have

$$
\sum_{j=1,2} \mu_{j\ell}\delta_j = -\sum_{j=1,2} \delta_j(\ell)(\beta\overline{\theta} + \gamma\theta_j), \qquad \forall \ell \in [V].
$$

Comparing both sides of the above identity it is clear that $\mathrm{span}\{\sum_{j=1,2}\mu_{j\ell}\delta_j\}_{\ell=1}^V \subseteq \mathrm{span}\{\beta\overline{\theta}+\gamma\theta_j\}_{j=1}^2$. It remains to prove $\mathrm{span}\{\delta_j\}_{j=1}^2 \subseteq \mathrm{span}\{\sum_{j=1,2}\mu_{j\ell}\delta_j\}_{\ell=1}^V$.

Recall that $\mu_{j\ell} = \beta\overline{\theta}(\ell) + \gamma\theta_j(\ell)$. Because $\beta\overline{\theta}+\gamma\theta_1$ and $\beta\overline{\theta}+\gamma\theta_2$ are linear independent, we know that $\dim\mathrm{span}\{\beta\overline{\theta} + \gamma\theta_j\}_{j=1}^2 = 2$ and hence $\dim\mathrm{span}\{(\mu_{1\ell},\mu_{2\ell})\}_{\ell=1}^V = 2$, because the row rank and the column rank of a matrix are equal. Thus, for any $(u,v) \in \mathbb{R}^2$, there exists real coefficients $\{w_\ell\}_{\ell=1}^V$ such that $(u,v) = \sum_{\ell=1}^V w_\ell(\mu_{1\ell},\mu_{2\ell})$. This implies $\mathrm{span}\{\delta_j\}_{j=1}^2 \subseteq \mathrm{span}\{\sum_{j=1,2}\mu_{j\ell}\delta_j\}_{\ell=1}^V$, which completes the proof.

## References

Anandkumar, A., Ge, R. and Janzamin, M. (2017). Analyzing tensor power method dynamics in overcomplete regime. *Journal of Machine Learning Research* **18** 1–40. MR3634889

Anandkumar, A., Foster, D. P., Hsu, D. J., Kakade, S. M. and Liu, Y.-K. (2012). A spectral algorithm for latent dirichlet allocation. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*. MR3385380

Anandkumar, A., Ge, R., Hsu, D. J., Kakade, S. M. and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research* **15** 2773–2832. MR3270750

Arora, S., Ge, R. and Moitra, A. (2012). Learning topic models–going beyond SVD. In *Proceedings of the IEEE Annual Symposium on Foundations of Computer Science (FOCS)*. MR3185945

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y. and Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the International Conference on Machine Learning (ICML)*. MR3388262

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM* **55** 77–84.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the international conference on Machine learning (ICML)*.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research* **3** 993–1022.

Chen, J. (1995). Optimal Rate of Convergence for Finite Mixture Models. *The Annals of Statistics* **23** 221–233. MR1331665

Cheng, D., He, X. and Liu, Y. (2015). Model Selection for Topic Models via Spectral Decomposition. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Drton, M. (2016). Algebraic Problems in Structural Equation Modeling. *arXiv preprint arXiv:1612.05994*.

Drton, M., Foygel, R. and Sullivant, S. (2011). Global identifiability of linear structural equation models. *The Annals of Statistics* **39** 865–886. MR2816341

Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ge, R., Huang, Q. and Kakade, S. M. (2015). Learning mixtures of gaussians in high dimensions. In *Proceedings of the annual ACM symposium on Theory of computing (STOC)*. MR3388256

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences* **101** 5228–5235.

Hager, W. W. (1984). Condition estimates. *SIAM Journal on scientific and statistical computing* **5** 311–316. MR0740850

Ho, N. and Nguyen, X. (2016). Singularity structures and impacts on parameter estimation in finite mixtures of distributions. *arXiv preprint arXiv:1609.02655*.

Hsu, D. and Kakade, S. M. (2013). Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the conference on Innovations in Theoretical Computer Science (ITCS)*. MR3385380

Huang, F., Niranjan, U., Hakeem, M. U. and Anandkumar, A. (2015). Online tensor methods for learning latent variable models. *Journal of Machine Learning Research* **16** 2797–2835. MR3450524

Leung, D., Drton, M. and Hara, H. (2016). Identifiability of directed Gaussian graphical models with one latent source. *Electronic Journal of Statistics* **10** 394–422. MR3466188

Ma, T., Shi, J. and Steurer, D. (2016). Polynomial-time tensor decompositions with sum-of-squares. In *Proceedings of the IEEE Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics* **41** 370–400. MR3059422

Nguyen, X. (2015). Posterior Contraction of the population polytope in infinite admixture models. *Bernoulli* **21** 618–646. MR3322333

Nguyen, X. (2016). Borrowing strength in hierarchical Bayes Posterior concentration of the Dirichlet base measure. *Bernoulli* **22** 1535–1571. MR3474825

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *The Annals of Probability* **22** 28–76. MR1258865

Tang, J., Meng, Z., Nguyen, X., Mei, Q. and Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Van der Vaart, A. W. (1998). *Asymptotic statistics* **3**. Cambridge university press. MR1652247

Wang, Y. and Anandkumar, A. (2016). Online and differentially-private tensor decomposition. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

Wang, Y. and Zhu, J. (2014). Spectral methods for supervised topic models. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

Wang, Y., Tung, H.-Y., Smola, A. J. and Anandkumar, A. (2015). Fast and guaranteed tensor decomposition via sketching. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*. Cambridge University Press. MR2554932

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research* **14** 867–897. MR3049492